PROCEEDINGS

Redžúr [r E dZ U: r] 2011

5th International Workshop on Multimedia and Signal Processing

Edited by: Juraj Vojtko
Gregor Rozinaj
Anna Kondelová
Ján Tóth
Juraj Londák

5th International Workshop on Multimedia and Signal Processing, May 12, 2011, Bratislava, Slovak Republic

PROCEEDINGS
Redžúr 2011

5th International Workshop on Multimedia and Signal Processing
May 12, 2011, Bratislava, Slovak Republic

EDITED BY:

Juraj Vojtko
Gregor Rozinaj
Anna Kondelová
Ján Tóth
Juraj Londák

Slovak University of Technology in Bratislava
Faculty of Electrical Engineering and Information Technology
Institute of Telecommunications
Ilkovičova 3
812 19 Bratislava
SlovakRepublic

Published by:

Nakladateľstvo STU Bratislava
in cooperation with
Združenie používateľov telekomunikácií Slovenska

ISBN 978-80-227-3506-3

General Chair

Gregor Rozinaj

Slovak University of Technology, Bratislava, Slovak Republic

PROGRAM COMMITTEE

Žarko Čučej

University of Maribor, Slovenia

Gerhard Doblinger

ViennaUniversity of Technology Austria

Mislav Grgic

University of Zagreb, Croatia

Gerhard Gruhler

Heilbronn University, Germany

Juraj Kačur

Slovak University of Technology, Slovakia

Jarmila Pavlovičová

Slovak University of Technology, Slovakia

Pavol Podhradský

Slovak University of Technology, Slovakia

Bhanu Prasad

Florida A&M University, USA

Markus Rupp

ViennaUniversity of TechnologyAustria

Martin Turi Nagy

Slovak University of Technology, Slovakia

REVIEW COMMITTEE

Beniak M., Slovak Republic

Čučej Ž., Slovenia

Doblinger G., Austria

Grgic M., Croatia

Gruhler G., Germany

Kačur J., Slovak Republic

Kondelová A., Slovak Republic

Kőrösi J., Slovak Republic

Londák J., Slovak Republic

Polec J., Slovak Republic

Rozinaj G., Slovak Republic

Rupp M., Austria

Rybárová R., Slovak Republic

Turi Nagy M., Slovak Republic

Tóth J., Slovak Republic

Vargic R., Slovak Republic

Vojtko J., Slovak Republic

ORGANIZING COMMITTEE

Organizing Committee Chair

JurajVojtko

SlovakUniversity of Technology, Bratislava, Slovak Republic

Kondelová A., Slovak Republic

Londák J., Slovak Republic

Tóth J., Slovak Republic

Turi Nagy M., Slovak Republic

Preface

Welcome to the 5th International Workshop on Multimedia and Signal Processing Redžúr 2011. Welcome to Bratislava.

The Redžúr 2011 has been organized by the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology, Bratislava and Telecommunications Users Group of Slovakia, under auspices of EURASIP.

After reviewing process, the International Program and Review Committees have selected 26 papers. We would like to express our thanks to all members of the Review Committee for their effort and valuable time within the review process. This year for the first time, we have selected a list of papers for the oral presentation. All participants present their works within a poster session, too.

We are proud this year to have an invited speaker, Jelena Božek from the University of Zagreb, Croatia.

The conference is organized especially for students. For them is this event often a first active contact with a scientific conference. We express our gratitude to all authors for their contribution and confidence to the conference Redžúr 2011. The great effort has been expended by the Organizing Committee to prepare all details surrounding this conference.

We wish to all conference participants a successful participation on the event from scientific point of view but we hope you will enjoy the possibility for informal contacts, as well.

Gregor Rozinaj,
Chairman of Redžúr 2011

Contents

Mammographic Image Analysis

(Invited Talk)

Jelena Bozek

University of Zagreb, Faculty of Electrical Engineering, Department of Wireless Communications
Unska 3/XII, HR-10000 Zagreb, Croatia
jelena.bozek@fer.hr

Abstract. *Mammography is one of the best examination procedures for early detection of breast cancer in screening programs. Due to the huge amount of mammographic images acquired in screening programs, aid of computers in the evaluation of mammographic images can be of great help for radiologists. Computer-aided detection algorithms can serve as a second reader and help radiologists to reveal breast abnormalities such as calcifications and masses, as well as subtle signs of breast cancer such as architectural distortion and bilateral asymmetry. Breast abnormalities are often indistinguishable from the surrounding tissue and are defined with wide range of features which makes computer-aided detection a challenging task. In this paper main issues in mammographic image analysis are pointed out and some of the developed algorithms for computer-aided detection of breast cancer are presented.*

Keywords

Breast cancer, digital mammography, mammographic image, computer-aided detection and diagnosis.

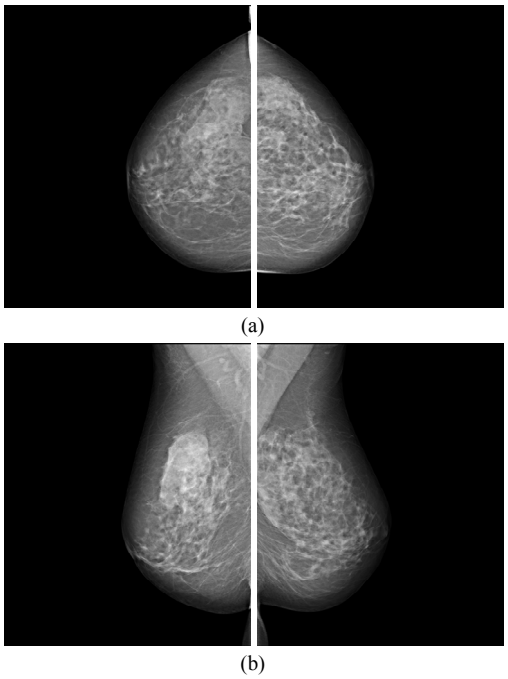
1. Introduction

Mammography is an effective method for early detection of breast cancer in screening programs. However, the success of screening and early detection depends on the interpretation of the images and detection of abnormalities. Big amount of images acquired through screening programs is evaluated independently by two expert radiologists. Use of computers in the mammographic image analysis could help radiologists as second reading, assure them in their diagnosis or indicate lesions that they might have missed.

Most computer-aided detection (CAD) algorithms analyze features that define change or distortion from normal healthy tissue. Breast abnormalities are defined with wide range of features that can indicate malignant changes but can also be a part of benign changes. Most of the features such as shape, margin, distribution, size etc. can be detected by using developed algorithms. However, there are some problems in detection and diagnosis of breast abnormalities. Some of the problems are poor visibility of the abnormality, possibility to differ it from the surrounding tissue and appropriate classification of the change as malignant or benign. Terms for describing lesion features are developed by the American College of Radiology (ACR) and suggested through standardized Breast Imaging Reporting and Data System (BI-RADS) [1].

Mammographic image is taken in two standard views: cranio-caudal (CC) and mediolateral-oblique (MLO) view. An example of mammographic image of left and right breast in CC and MLO view is shown in Fig. 1 (a) and (b), respectively. Radiologists use both views and compare right and left breast image in the search for possible changes in the breast.

Short overview and main characteristics of breast abnormalities are presented in Section 2. In Section 3 are presented basic methods and algorithms developed for computer-aided detection of breast cancer. Finally, Section 4 brings a conclusion.



(a)

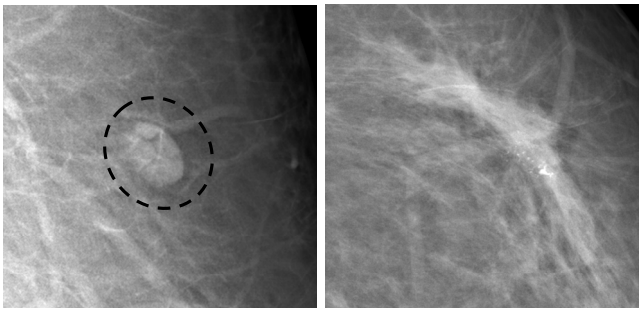
(b)

Fig. 1. Left and right breast in: (a) CC view and (b) MLO view.

2. Breast Abnormalities

Breast abnormalities that are most common sign of breast cancer are masses and calcifications. Subtle changes in breast tissue such as architectural distortion and bilateral asymmetry may also indicate breast cancer in an early stage.

According to BI-RADS a mass is defined as a space occupying lesion seen in at least two different projections [1]. Masses have different density, different margins and different shape. Benign mass is usually round with smooth and circumscribed margin while masses with spiculated, rough and blurry margins are usually a sign of malignancy. A benign round mass is shown in Fig. 2 (a) and malignant spiculated mass is shown in Fig. 2 (b).



(a) (b)

Fig. 2. An example of: (a) round mass, (b) spiculated mass

Calcifications are deposits of calcium in the breast tissue and are usually very bright compared to the surrounding tissue. Due to their small size, sometimes they can be easily mistaken for noise. Malignant calcifications tend to be numerous, clustered, small, varying in size and shape, angular, irregularly shaped and branching in orientation [2]. An example of malignant calcifications is shown in Fig. 3. Benign calcifications are usually larger than calcifications associated with malignancy. They are coarser, often round with smooth margins, smaller in number, more diffusely distributed, more homogeneous in size and shape and are much more easily seen on a mammographic image.

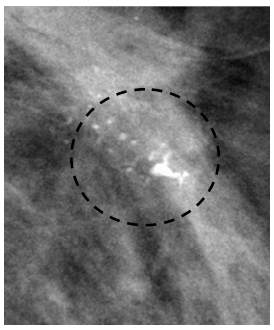


Fig. 3. Example of fine pleomorphic clustered calcifications

The third most common mammographic sign of cancer is architectural distortion. It is defined as distortion of the normal architecture with no definite mass visible, including spiculations radiating from a point and focal retraction or distortion at the edge of the parenchyma [1]. Architectural distortion of breast tissue can indicate malignant changes especially when integrated with visible lesions such as mass, asymmetry or calcifications. Architectural distortion can be classified as benign when there is a scar and soft-tissue damage due to trauma. Mammographic image with architectural distortion is shown in Fig. 4.

Asymmetry of breast parenchyma between corresponding regions in left and right breast is useful sign for detecting primary breast cancer. Asymmetric breast tissue can be expected in approximately 3% of the population [3]. Asymmetric breast tissue is usually benign, but an asymmetric area may indicate a developing mass or an underlying cancer [4]. Thus, asymmetrical breasts could be reliable indicators of future breast disease in women and this factor should be considered in a woman's risk profile [5]. Bilateral asymmetries of concern are those that are changing or enlarging or new, those that are palpable and those that are associated with other findings, such as microcalcifications or architectural distortion [2]. Pair of left and right mammographic image with bilateral asymmetry visible in left breast is shown in Fig. 5.

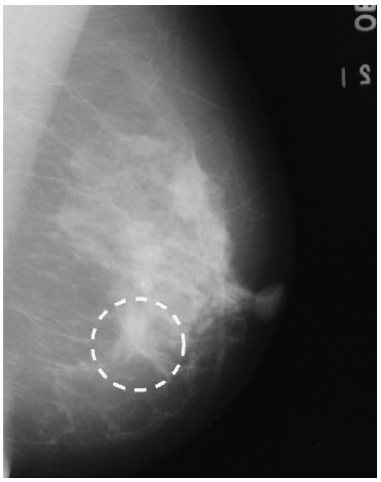
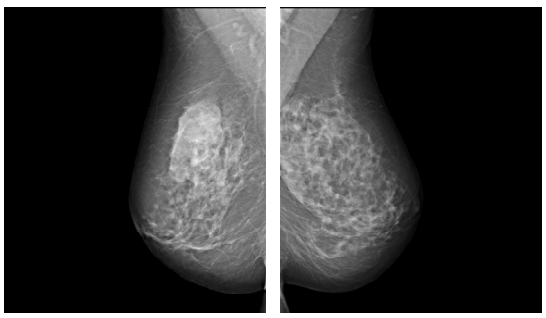


Fig. 4. Mammographic image with architectural distortion (dashed circle) [6]



(a) (b)

Fig. 5. Mammographic image pair with bilateral asymmetry: (a) left breast with asymmetry (b) right breast

3. Computer-Aided Detection

The goal of CAD algorithms is to indicate suspicious breast abnormalities thereby assuring accurate diagnosis by a radiologist. Findings in a number of studies have demonstrated that CAD has the ability to detect and prompt mammographic signs of cancer with the potential to increase cancer detection rates by approximately 20% [2]. There are several typical steps in mammographic image analysis that are depicted in Fig. 6.

Depending on the image quality, in the preprocessing step noise is removed and image is enhanced. Part of the preprocessing is segmentation of the background and pectoral muscle from the breast area and limiting analysis to region of interest (ROI) where suspicious abnormalities may occur. In the feature extraction step the features are calculated from the characteristics of the whole breast or from the segmented region of interest. Critical issue in algorithm design is the feature selection step where the best set of features is selected for the final classification step. In the classification step the number of false positives is reduced and abnormality is classified based on the selected features.

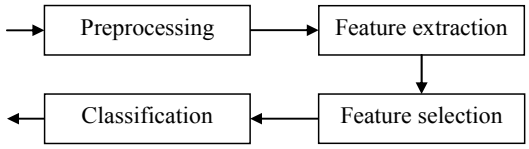


Fig. 6. Main steps in the CAD algorithm

There are many algorithms developed for various steps of CAD system. Since features are specific for a certain abnormality, different algorithms are developed for detecting a specific abnormality.

Since malignancy of a mass is defined with its morphological characteristics and texture, researchers have extracted and combined different sets of morphological and textural features for mass detection [7-12]. Some researchers used temporal features obtained from mammographic images taken at consecutive exams for detecting new or developing masses [13] as well as features obtained by combining two mammographic projections [14].

A number of different approaches have been applied for the detection of calcifications. Calcifications represent high spatial frequencies in the image. Thus, one approach to the calcification detection task is to localize the high spatial frequencies of the image using wavelet transform [15-19]. Other non-wavelet-based methods use the fact that calcifications have much higher intensity values than the surrounding tissue in a mammogram [20-23]. The problem with intensity based methods is that calcifications may be missed if the breast tissue is dense.

Methods for detection of architectural distortion are often included in mass detection algorithms. However, methods designed exclusively for the detection of architectural distortion can achieve better performance than the application of methods for the detection of spiculated masses, which may rely on the presence of a central mass [6]. In order to detect architectural distortion some methods are based on the detection of spiculated lesions [24], on the detection of architectural distortion around the skin line and within the mammary gland [25] and some are texture-based [26]. Accurate detection of architectural distortion could be the key to efficient detection of early breast cancer, at pre-mass-formation stages.

The evaluation of bilateral breast asymmetry based on density, shape and size is usually the first stage in the mammographic evaluation process [27]. Before performing asymmetry analysis it is necessary to apply some kind of alignment of the breasts. However, alignment procedures applied to mammograms have to confront many difficult problems such as the natural asymmetry of the breasts, absence of good corresponding points between left and right breast images to perform matching and distortions inherent to breast imaging [2]. Some of the developed methods for detection of bilateral asymmetry are texture-based [8]. Others are based on measures of shape, topology and distribution of brightness in the fibroglandular disk [28] or are based on measures of brightness, roughness and directionality [29]. More methods are desirable in this area to analyze asymmetry from multiple perspectives, including pattern asymmetry in the fibroglandular tissue as well as morphological and density measures related to the breast and the fibroglandular disk [6].

4. Conclusion

Mammography is one of the best examination procedures for early detection of breast cancer. To help radiologists in evaluating huge amount of mammographic images acquired in screening programs algorithms for automated detection of breast abnormalities are being developed. Wide range of features that describe abnormalities make the computer-aided detection a challenging task. Sensitivity of commercial CAD systems for calcification detection is up to 98.5%, but the sensitivity of CAD systems for mass detection is still not satisfactory, only 89.2% [30]. Further improvement of algorithms for early breast cancer detection is necessary in order to achieve better sensitivity and accuracy although there are already promising results achieved by several different research groups.

Acknowledgements

The work in this paper was conducted under the research project "Intelligent Image Features Extraction in Knowledge Discovery Systems" (036-0982560-1643), supported by the Ministry of Science, Education and Sports of the Republic of Croatia.

References

[1] AMERICAN COLLEGE OF RADIOLOGY (ACR), *ACR Breast Imaging Reporting and Data System, Breast Imaging Atlas*. 4th Edition, Reston, VA. USA, 2003.

[2] SURI, J.S., RANGAYYAN, R.M. *Recent Advances in Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer*. SPIE, Bellingham, Washington, USA, 2006

[3] KOPANS, D.B., SWANN, C.A., WHITE, G., MCCARTHY, K.A., HALL, D.A., BELMONTE, S.J., GALLAGHER, W. Asymmetric breast tissue. *Radiology*, 1989, vol. 171, no. 3, p. 639-643

[4] SAMARDAR, P., DE PAREDES, E.S., GRIMES, M.M., WILSON, J.D. Focal asymmetric densities seen at mammography: us and pathologic correlation. *Radiographics*, 2002, vol. 22, p. 19-33

[5] SCUTT, D., LANCASTER, G.A., MANNING, J.T. Breast asymmetry and predisposition to breast cancer. *Breast Cancer Research*, 2006, vol. 8, R 14, available at: breast-cancer-research.com/content/8/2/R14S

[6] RANGAYYAN, R.M., AYRES, F.J., DESAUTELS, J.E.L. A review of computer-aided diagnosis of breast cancer: toward the detection of subtle signs. *Journal of the Franklin Institute*, 2007, vol. 344, issues 3-4, p. 312-348

[7] SAHINER, B., CHAN, H.-P., PETRICK, N., HELVIE, M.A., HADJIISKI, L.M. Improvement of mammographic mass characterization using spiculation measures and morphological features. *Medical Physics*, 2001, vol. 28, no 7, p. 1455-1465

[8] RANGAYYAN, R.M., GULIATO, D., DE CARVALHO, J.D., SANTIAGO, S.A. Feature extraction from the turning angle function for the classification of contours of breast tumors. *IEEE Special Topic Symposium on Information Technology in Biomedicine*, Iaonnina (Greece), 2006, 4 pages on CDROM

[9] BELLOTTI, R., DE CARLO, F., TANGARO, S., GARGANO, G., MAGGIPINTO, G., CASTELLANO, M., MASSAFRA, R., CASCIO, D., FAUCI, F., MAGRO, R., RASO, G., LAURIA, A., FORNI, G., BAGNASCO, S., CERELLO, P., ZANON, E., CHERAN, S. C., LOPEZ TORRES, E., BOTTIGLI, U., MASALA, G.L., OLIVA, P., RETICO, A., FANTACCI, M. E., CATALDO, R., DE MITRI, I., DE NUNZIO, G. A completely automated CAD system for mass detection in a large mammographic database. *Medical Physics*, 2006, vol. 33, no. 8, p. 3066-3075

[10] VARELA, C., TAHOCES, P.G., MÉNDEZ, A.J., SOUTO, M., VIDAL, J.J. Computerized detection of breast masses in digitized mammograms. *Computers in Biology and Medicine*, 2007, vol. 37, p. 214 – 226

[11] YUAN, Y., GIGER, M.L., LI, H., SENNETT, C. Correlative feature analysis of FFDM images. *Medical Imaging 2008: Computer-Aided Diagnosis*, edited by Maryellen L. Giger, Nico Karssemeijer, Proc. of SPIE, 2008, vol. 6915

[12] SAHINER, B., HADJIISKI, L.M., CHAN, H.P., PARAMAGUL, C., NEES, A., HELVIE, M., SHI, J. Concordance of computer-extracted image features with BI-RADS descriptors for mammographic mass margin. *Medical Imaging 2008: Computer-Aided Diagnosis*, edited by Maryellen L. Giger, Nico Karssemeijer, Proc. of SPIE, 2008, vol. 6915

[13] TIMP, S., KARSSEMEIJER, N. Interval change analysis to improve computer aided detection in mammography. *Medical Image Analysis*, 2006, vol. 10, p. 82–95

[14] VAN ENGELAND, S., KARSSEMEIJER, N. Combining two mammographic projections in a computer aided mass detection method. *Medical Physics*, 2007, vol. 34, no. 3, p. 898-905

[15] STRICKLAND, R.N., HAHN, H.I. Wavelet transforms for detecting microcalcifications in mammograms. *IEEE Transactions on Medical Imaging*, 1996, vol. 15, no. 2, p. 218-229

[16] YOSHIDA, H., DOI, K., NISHIKAWA, R.M., GIGER, M.L., SCHMIDT, R.A. An improved computer-assisted diagnostic scheme using wavelet transform for detecting clustered microcalcifications in digital mammograms. *Academic Radiology*, 1996, vol. 3, no. 8, p. 621-627

[17] ZHANG, W., YOSHIDA, H., NISHIKAWA, R.M., DOI, K. Optimally weighted wavelet transform based on supervised training for detection of microcalcifications in digital mammograms. *Medical Physics*, 1998, vol. 25, no. 6, p. 949-956

[18] QIAN, W., KALLERGI, M., CLARKE, L.P., LI, H.D., VENUGOPAL, P., SONG, D., CLARK, R.A. Tree structured wavelet transform segmentation of microcalcifications in digital mammography. *Medical Physics*, 1995, vol. 22, no. 8, p. 1247-1254

[19] GURCAN, M.N., YARDIMCI, Y., CETIN, A.E., ANSARI, R. Detection of microcalcifications in mammograms using higher order statistics. *IEEE Signal Processing Letters*, 1997, vol. 4, no. 8, p. 213-216

[20] CHAN, H.P., DOI, K., GALHOTRA, S., VYBORNY, C.J., MACMAHON, H., JOKICH, P.M. Image feature analysis and computer-aided diagnosis in digital radiography. I. Automated detection of microcalcifications in mammography. *Medical Physics*, 1987, vol. 14, no. 4, p. 538-548

[21] CHAN, H.P., LO, S.C., SAHINER, B., LAM, K.L., HELVIE, M.A. Computer-aided detection of mammographic microcalcifications: pattern recognition with an artificial neural network. *Medical Physics*, 1995, vol. 22, no. 10, p. 1555-1567

[22] DAVIES, D.H., DANCE, D.R. Automatic computer detection of clustered calcifications in digital mammograms. *Physics in Medicine and Biology*, 1990, vol. 35, no. 8, p. 1111-1118

[23] NISHIKAWA, R.M., JIANG, Y., GIGER, M.L., SCHMIDT, R.A., VYBORNY, C.J., ZHANG, W., PAPAIOANNOU, J, BICK, U., NAGEL, R., DOI, K. Performance of automated CAD schemes for the detection and classification of clustered microcalcifications. *Digital Mammography*, editors A.G. Gate et al., Elsevier, Amsterdam, 1994

[24] SAMPAT, P., WHITMAN, G.J., MARKEY, M.K., BOVIK, A.C. Evidence based detection of spiculated masses and architectural distortion. In *Proceedings of SPIE Medical Imaging 2005: Image Processing*, 2005, vol. 5747, p. 26-37

[25] MATSUBARA, T., ICHIKAWA, T., HARA, T., FUJITA, H., KASAI, S., ENDO, T., IWASE, T. Automated detection methods for architectural distortions around skinline and within mammary gland on mammograms. In *Proceedings of the 17th International Congress and Exhibition on Computer Assisted Radiology and Surgery*, Elsevier, London, UK, 2003, p. 950-955

[26] MUDIGONDA, N.R., RANGAYYAN, R.M. Texture flow-field analysis for the detection of architectural distortion in mammograms. In *Proceedings of Biovision*, Bangalore, India, 2001, p. 76-81

[27] KINOSHITA, S.K., DE AZEVEDO-MARQUES, P.M., PEREIRA JR., R.R. Content-based retrieval of mammograms using visual features related to breast density patterns. *Journal of Digital Imaging*, 2007, vol. 20, no. 2, p. 172-190

[28] Miller, P., Astley, S. Automated detection of breast asymmetry using anatomical features. In: *State of the Art in Digital Mammographic Image Analysis*, vol. 9 of Series in Machine Perception and Artificial Intelligence, World Scientific, River Edge, NJ, 1994, p. 247-261

[29] Lau, T.K., Bischof, W.F. Automated detection of breast tumors using the asymmetry approach, *Computers and Biomedical Research*, 1991, vol. 24, no. 3, p. 273-295

[30] ASTLEY, S.M. Computer-based detection and prompting of mammographic abnormalities. *The British Journal of Radiology*, 2004, vol. 77, p. 194-200

Speaker Identification

Matúš PETEJA[1], *Juraj KAČUR*[1]

[1] Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
matus.peteja@gmail.com

Abstract. The aim of this paper was the area of speaker identification. Paper includes brief description of speech production, the main characteristics of it, and moreover the extraction of its attributes. Practical part deals with already mentioned speaker identification based on voice sample, which is matched with another voice samples from database. For classification purposes a simple but powerful method of KNN (K-nearest neighbor) was used. We investigated the effect of different parameterization methods and various settings of KNN on the accuracy. Very encouraging results were achieved, although not all of them confirmed theoretical assumption.

Keywords

Speaker identification, filter banks, MFCC, KNN.

Introduction

Speaker recognition is process of automatic recognition of person, who is speaking based on the information obtained in speech. Recognition is making by identification and verification. Speaker identification is process of person's identification based on his/her voice and samples enrolled in the database. The size of a database defines the decision-making alternatives, so the more speakers in database, the less speaker identification accuracy.

1. Speech characteristics

Speech is one of the basic communication methods. Human speech is continuous, time variable process that is a carrier of information from speaker to listener. It is coded and is presented by the acoustic waves. Speech is produced when the breathed out air stream from lungs is influenced by the voice apparatuses, from vocal chords to lips.

Human speech is characteristic by its acoustic structure, linguistic structure and exhibition of speaker's personality. Fundamental attribute of sound, thus speech as well, is the intensity (loudness), pitch and color. Basic tone of human voice is characterized by the frequency vibration of vocal chords. Adult male has this frequency between 90 and 150 Hz, female from 130 to 300 Hz, and children over 300 Hz. When the voice is traveling in the vocal apparatuses, there are resonances in the oral, nasal and pharynx cavities. These resonances intensify some parts of sound spectrum and thus produce formants. While the basic frequency indicates the pitch formants make resultant acoustic feeling.

2. Speech features

In speech processing, we need to convert speech into sequences of feature vectors. These vectors should contain relevant information of sounds during utterance, which is necessary for recognition. There is no consensus which method is optimal for obtaining these features, but most uses extraction of spectral features [3].

2.1 Frequency spectrum

Frequency spectrum of signal is actually representation of this signal in frequency area. It can be computed by Fourier transformation and the output attributes can be amplitudes or phases.

DFT (Discrete Fourier Transformation)

$$\mathrm{X(k)} = \frac{1}{\mathrm{N}} \times \sum_{n=0}^{N-1} x(n) \times e^{-j \times \frac{2\pi}{N} \times n \times k} \quad (1)$$

IDFT (Inverse Discrete Fourier Transformation)

$$\mathrm{x(n)} = \sum_{k=0}^{N-1} X(k) \times e^{j \times \frac{2\pi}{N} \times n \times k} \quad (2)$$

Fourier transformation is a function, which describes amplitude and phase of ever sinusoid that is adequate to specific frequency.

Put it simply, for human, it is necessary just to listen, because ear makes this calculation automatically.

2.2 Logarithmic spectrum

Samples with the higher amplitude show off in logarithmic frequency spectrum, whereas samples with lower amplitude are bottle up. As a result the useful information is showed off. It is assumed that human auditory system performs similar operation.

2.3 Filter banks

Filter bank is one of the elementary means of speech signal analysis. The aim of this analysis is to detect the energy of a signal in particular frequency bands.

2.3.1 Linear filter bank

It is the basic type of filter bank. The width of particular bands is equal and central frequencies of these bands are equally distributed at frequency ax.

2.3.2 Nonlinear filter bank

Human ear has more selective ability for lower frequencies, what means that it more distinguish sounds at lower frequencies. That is the reason why the linear distribution is not optimal.

The most used method for making nonlinear filter bank is to convert frequency spectrum to Mel spectrum by formula (3)

$$\text{mel} = 1127{,}01048 \times \ln(1 + \frac{f}{700}) \text{ (mel)} \quad (3)$$

where f is frequency in Hz. Then the equally distribution of bands is made, and after that, the backward conversion to frequency spectrum is done by formula (4)

$$\text{f} = 700 \times (\frac{mel}{e^{1127{,}01048}} - 1) \text{ (Hz)} \quad (4)$$

That is the way how the filter bank is made, where the energy of bands at lower frequencies is calculated from shorter frequency bands than at the higher frequencies.

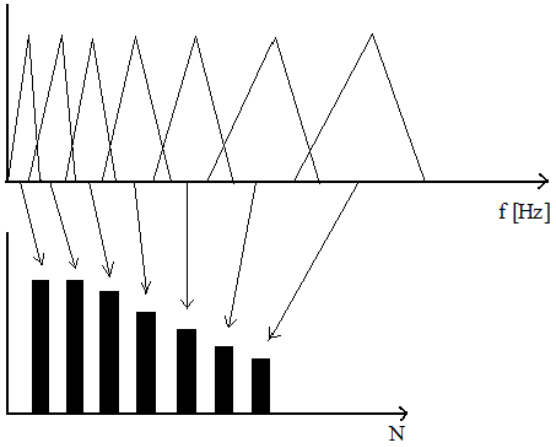


Fig. 1. Nonlinear filter bank.

2.4 MFCC (Mel-frequency cepstral coefficients)

In modern systems for speaker identification it is most common to use MFCC [1] features. The reason is, as it was mentioned in nonlinear filter bank (2.3.2), to make computer signal processing as close to human auditory system as possible. First steps to get Mel-frequency cepstral coefficients are described in (2.3.2). When we have Mel-scaled filter banks that represent energy in particular bands, then we calculate logarithm of these energies. Finally, the DCT (discrete cosine transformation) is applied to this log filter bank. This process is described by equation

$$C_{mfcc}(k) = \sum_{m=1}^{M} \log X(m) \times \cos(k(m - \frac{1}{2})\frac{\pi}{M}) \quad (5)$$

where X(m) is a coefficient of filter bank, M is number of filter banks and k=0,1,…M

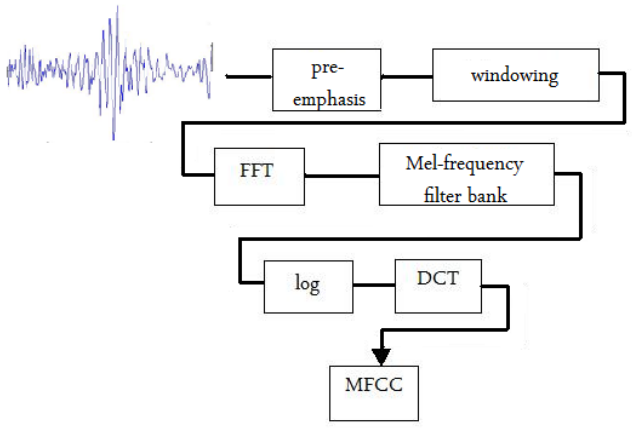


Fig. 2. MFCC process execution scheme

3. KNN (K nearest neighbor)

K nearest neighbor is one of the most fundamental and intuitive technique in area of machine learning [2], [4]. It is nonparametric method, marking of new data point, by finding the closest point from training data. To find the closest point, the similarity based on distance measure is used. As it was mentioned, k-nearest neighbor method is a simple and efficient, because the training with a training group is not necessary, and known pattern doesn't need advanced training, what makes this method more flexible. Computational cost is commensurable to amount of training group. Disadvantage of k-nearest neighbor method is in its enormous computing complicacy. It is necessary to calculate distances among every sample that needs to be classified, and every known samples, to get the nearest neighbor. Often used solution for this deficiency is removal of parts of signal that have no effect on classification. For example, the long time parts of silence should be removed. Another disadvantage of KNN is its disability of classifying

two or more classes that have the same number of nearest neighbor samples.

Classifying by KNN is primary described by number of neighbor. This parameter defines some kind of identification efficiency, or accuracy. It is not easy to define it and for different applications is good to use different number of neighbor. However in general, higher value of K means better immunity against noise.

3.1 Distance – weighted Nearest Neighbor

One of the modifications of classic KNN method is to use weight of each neighbor. This weight defines the distance of a neighbor to the query sample. The bigger distance between these samples, the less weight is used. Summation of weights for the same class neighbors is the basis for the decision-making.

$$\hat{\vartheta}(x_k) \leftarrow \arg\max_{\upsilon \in V} \sum_{l=1}^{m} w_l \delta(\upsilon, \vartheta(y_l)) \qquad (6)$$

where

$$w_l = \frac{1}{d(x_k, y_l)^2} \qquad (7)$$

Where x_k is test sample, y_l is training sample, $\delta(a,b) = 1$ if $a = b$ and $\delta(a,b) = 0$ otherwise.

We assign $\hat{\vartheta}(x_k)$ to be $\vartheta(y_l)$ in the case of $d(x_k, y_l)^2 = 0$ what means that the test sample, respectively query sample x_k exactly matches one of the training sample.

3.2 Local distances

Speaker identification by KNN is in principal based on local distance, the distance among particular features of speech.

3.2.1 Euclid distance

KNN method usually uses Euclid distance for expressing of similarities. It simply says: the shorter distance between two samples is, the more similar they are. This distance is computed by Euclid formula

$$d(x_k, y_l) = \sqrt{\sum_{i=1}^{N} (x_{ki} - y_{li})^2} \qquad (8)$$

where x_k is test sample, y_l is training sample, x_{ki} is i-th feature of vector x_k and y_{li} is i-th feature of vector y_l.

3.2.2 Weighted Euclid distance

Similarity measure depending on all features with the same weight is often misleading, because features are scale variant and they may not carry the same amount of information. Thus weighting contribution of each component differently may ensure that each element will have different effect in classification. Formula for weighted Euclid distance is

$$d(x_k, y_l) = \sqrt{\sum_{i=1}^{N} \left[(x_{ki} - y_{li})^2 \times \frac{1}{w_i} \right]} \qquad (9)$$

where w_i is weight of particular feature.

3.2.3 Mahalanobis distance

This metric involves the relations between unknown test samples and training samples that are described by covariance matrix. In this way any possible correlations are removed and all scales are unified thus all elements may contain the same dispersion. From this point of view Euclidean distance is better, but on the other hand it's much more time-consuming. The formula is

$$d(x_k, y_l) = \sqrt{(x_k - y_l)^T \times \Sigma^{-1} \times (x_k - y_l)} \qquad (10)$$

Where Σ is the covariance matrix of the training data.

4. Experiments

The database of records consist from 13 (+2 different records of the same speaker in "real test") persons of both sexes, however there is only one female. Records were obtained by stochastic choice from another database, so we can not say what surroundings they were made in and what terms were during recording. All of them are stored in WAV file format with sampling frequency 22.05 kHz and 16 bits per sample. They last from 15 to 17 seconds and have no longer pauses or damaged parts that should be removed.

Each record was partitioned into frames with length of 20 ms and shift 10 ms. This means the frame overlap the previous one. Then the frequency spectrum by Fourier transformation (1) was obtained frame by frame.

The aim of this first test was to determine the error rate for the whole speaker database, i.e. to verify if KNN method is relevant for identification. Each voice record was split so that 70% was putted into training part and 30% into testing part. In principle the overall distance between training and testing records was calculated and the smallest distance marks strong relevancy for particular speaker, see tab 1.

spectrum	Euclid distance	Filter bank	Number of neighbors KNN	Error rate in %
Magnitude	weighted	Raw spectrum	1	7,7
Magnitude	Non weighted	Raw spectrum	1	0
Magnitude	weighted	Raw spectrum	2	7,7
Magnitude	Non weighted	Raw spectrum	2	7,7
magnitude	weighted	Raw spectrum	3	7,7
magnitude	Non weighted	Raw spectrum	3	7,7
Logarithmic	weighted	linear	1	0
Logarithmic	Non weighted	linear	1	0
Logarithmic	weighted	linear	2	0
Logarithmic	Non weighted	linear	2	0
Logarithmic	weighted	Non linear	1	0
Logarithmic	Non weighted	Non linear	1	0
Logarithmic	weighted	Non linear	2	0
Logarithmic	Non weighted	Non linear	2	0

Tab. 1. Achieved identification errors using different settings and the overall distance (of closest neighbors).

	MFCC, Euclid distance			MFCC, Mahalanobis distance		
	Number of neighbors KNN			Number of neighbors KNN		
	1	3	5	1	3	5
Recognition accuracy	90,45 %	95,56 %	96,18 %	86,24 %	92,87 %	94,27 %

Tab. 2. Achieved identification accuracy using different settings and the nearest neighbor majority criteria.

The remaining testing is aimed at speaker identification in conditions getting near to the real system. Thus Mel-frequency cepstral coefficients were used as speech parameters and distance-weighted KNN classification method was used when more than one neighbor were considered. It is necessary to mention, that whereas in the previous test the local distance was the decision-making parameter, here the total number of closest speakers is used (majority criteria), tab. 2.

5. Conclusions

Sometimes we come to surprising conclusions that not come up to expectations. Each set of parameters should cause different outcomes. But as we can see there are just two levels of error rates in the first experiment, tab. 1. It could be caused by small amount of speakers in the database (with rising speaker database, the identification time is raising so fewer tests can be performed). From our outcomes and considering the theoretical background we can claim that the best parameters for speaker identification are logarithmic spectrum, weighted Euclid distance, non linear filter bank and 2 neighbors.

Tests aimed at speaker identification using majority rule instead of the total distance, confirmed theoretical claims that Mel-frequency cepstral coefficients are very good parameters used in speaker identification. Surprising conclusion were observed regarding Euclid and Mahalanobis distances, where we expected better results to be produced by Mahalanobis distance which was not confirmed. However, the test shows similarities in the number of neighbors, the more neighbors we use, the better result in the recognition we get. So the best settings for speaker identification system are MFCC, Euclid distance and 5 neighbors.

Acknowledgements

This article was supported by VEGA – 1/0718/09 and FP7-ICT-2011-7 HBB-Next

References

[1] HOSSAN, M. A. ; MEMON, S. ; GREGORY, M. A. A Novel Approach for MFCC Feature Extraction. In Signal Processing and Communication Systems (ICSPCS), 2010

[2] T. M. Mitchell, Machine Learning, McGraw-Hill, ISBN 0-07-042807-7, 1997

[3] FURUI, SADAOKI. *Digital speech processing, synthesis, and recognition*: Second Edition, Revised and Expanded. NY: Marcel Dekker, Inc.. 2001

[4] WANG, QINGMIAO – JU, SHIGUANG. A Mixed Classifier Based on Combination of HMM and KNN. In *Natural Computation, 2008. ICNC '08.* 2008. p. 39-40, [online]. <http://www.ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4667244>. ISBN 978-0-7695-3304-9.

Improving video quality evaluation by mutual information

Michal MARDIAK[1], *Jaroslav Polec*[1]

[1] Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
mardiak@ktl.elf.stuba.sk, polec@ktl.elf.stuba.sk

Abstract. *In this paper we present the improvement of video quality evaluation by mutual information. In the first stage of the metric calculation the sequences are pre-processed by the Human Visual System. In the second stage we calculate mutual information which has been utilized as the quality evaluation criteria. To prove reliability of our metric we compare it with some commonly used objective methods for measuring the video quality. The results show that presented improvement of video quality evaluation provides relevant results in comparison with other objective methods so it is suitable candidate for measuring the video quality.*

Keywords

Objective video quality, HVS, mutual information, VQM, SSIM, PSNR.

1. Introduction

The recent century became a golden age in the area of technical innovations. One of the most widespread innovation is video in all its variations like cinema, television, videoconference etc. As the popularity of the video grows the requirements for providing video grows too. The most reliable results provide subjective video quality metrics which anticipate more directly the viewer's reactions [1]. However the quality evaluation of the video by subjective methods is expensive and too slow to be used in real-time applications. Therefore the objective methods start to be used. The main goal in the objective quality assessment research is to design metric which can provide sufficient quality evaluation regarding to the subjective results [2]. For better approximation of viewers visual perception in the terms of video quality the Human Visual System (HVS) models has been implemented. Various types of HVS have been used in the objective video quality evaluation.

We compare proposed method with several objective methods which have been used for the quality assessment of video sequence i.e. SSIM, PSNR and VQM. In the Section 2 different models of HVS are described and the Section 3 contains the calculation of mutual information. Then the results of our metric are presented and concluded at the end of the paper.

2. Human visual system

The purpose of Human Visual System is to simulate human visual perception of the video and it utilization should lead in general to a better quality of the reconstructed image [3] even if the simplified HVS is used [4]. We present some of HVS here:

The HVS is more sensitive in dark areas than in light so the spatial frequency sensitivity of the HVS decreases for high frequencies. The frequency sensitivity should be simulated by low-pass filter [3]. In our paper we choose following simple low-pass Gaussian filter:

$$HVS1 = \frac{1}{16} \cdot \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix} \tag{1}$$

Another way how to simulate HVS can be by using the band-pass filters [3, 4]. One of these filters can be expressed by the transfer function in the polar coordinates [4]:

$$H(\rho) = \begin{cases} 0{,}05e^{\rho^{0.554}} & \text{for } \rho < 7 \\ e^{-9(\left|\log_{10}\rho - \log_{10} 9\right|)^{2,3}} & \text{for } \rho \geq 7 \end{cases} \tag{2}$$

where $\rho = (u^2+v^2)^{0.5}$. Operator $U\{\cdot\}$ can be used for image processed by the transfer function $H(\rho)$ and afterwards transformed by the inverse discrete cosine transformation (DCT) as follows [4]:

$$U\{x(i,j)\} = DCT^{-1}\left\{H\left(\sqrt{u^2+v^2}\right) X_{DCT}(u,v)\right\} \tag{3}$$

where $x(i,j)$ is multispectral pixel vector of image, $X_{DCT}(u,v)$ represents the 2D DCT of the image and DCT^{-1} stands for 2D inverse DCT [4].

LoG filters can emulate the fact that HVS is more sensitive to the angular resolution and not to the image resolution [5,6]. We used two LoG filters with the size *7x7* and parameter $\sigma = 1$ and $\sigma = 1.2$.

The last HVS model presented in this paper is based on the two temporal filters which are also used in the JND metric.

These filters are defined by the following impulse response functions [7]:

$$h_1(t) = a \cdot e^{-at} \cdot u(t) \tag{4}$$

$$h_2(t) = b\{\frac{(bt)^3}{3!} - \frac{(bt)^5}{5!}\}e^{-bt}u(t) \tag{5}$$

3. Mutual information

In the presented method we calculate mutual information for the original and test sequence in RGB color space. Let us assume that pixel of k-th component $x_k(i,j)$ has value $x_k(i,j) \in \langle 0,G \rangle$. The values of intensity level are l and l'. The $P^k_{x,y}(l'/l)$ represents the count of changes from the intensity level l in the frame x from original sequence to the intensity level of l' in the corresponding frame y from the test sequence for the k-th component regarding to the total amount of the pixel in the frame. Parameters $P^k_x(l)$ and $P^k_y(l)$ stand for count of the intensity level l in the frame from original sequence and the intensity level of l' in the corresponding frame from the test sequence regarding to the total amount of the pixel in the image [8].

Total mutual information is defined as:

$$I_{x,y} = \sum_{k=1}^{K}\sum_{l=0}^{G}\sum_{l'=0}^{G} P^k_{x,y}\left(l'/l\right) log_2 \frac{P^k_{x,y}\left(l'/l\right)}{P^k_x(l)\cdot P^k_y\left(l'\right)} \tag{6}$$

4. Results

In this paper we choose for the test purpose of our metric the 'Foreman' video sequence. The test sequence has been at CIF resolution (352 x 288 pixels) coded by H.264 codec using the CABAC entropy coding method. The calculation of the presented objective video quality metric consists from two stages. In the first stage HVS is applied to the test and original video sequence. We use four different HVS mentioned above. In the second stage we calculate mutual information according to equation (6) between frame from the original sequence and corresponding frame from the test sequence.

Fig. 1 shows comparison between SSIM metric and the mutual information with mentioned HVS. From the beginning of sequence the mutual information evaluation of quality is slightly increasing and decreasing. By the applying of the Gaussian filter the run of mutual information become smoother. During the movement of foreman head the Gibbs phenomenon appears on the edges and the structural component of the SSIM metric change. That is the reason why also the SSIM quality assessment in this part of the sequence varies. Applying the LoG filters cause that changes in the quality of overall run at the start of the sequence correspond more to the SSIM. However not all of the peaks or bottoms occur at the same place. The first bigger notable improvement of quality SSIM indicates when the fast move of the hand blurs the part of the frame in the original sequence. The Gaussian filter implementation reacts on this fact also by improving the quality but after 20 frames later. On the other hand the applying of the LoG filters cause that the mutual information evaluates this blurring in the frame as a degradation of the quality.

The major ascent of the quality is indicated by SSIM when the camera moves and blurs the major part of the frame which contains the background with the few colors. However the run of mutual information with the Gaussian filter falling down when this happens. Implementation where LoG filters are used reacts also by decreasing the quality even if there is the peak when the quality is falling down. On the other side the mutual information in case where the second and fourth HVS model is used, indicates some quality improvement at this part of the sequence.

After major SSIM quality ascents in the frame 192, the quality has decreasing trend with some peaks. The mixture of colors at frame 231 causes the change of the contrast and structural component in SSIM metric. The run of the mutual information with Gaussian filter rise and slightly vary in quality with no bigger peaks or bottoms. The mutual information together with LoG filters also indicates improvement of quality but the peaks and the bottoms are more noticeable. The overall run of mutual information with second HVS is rising and falling during the whole sequence.

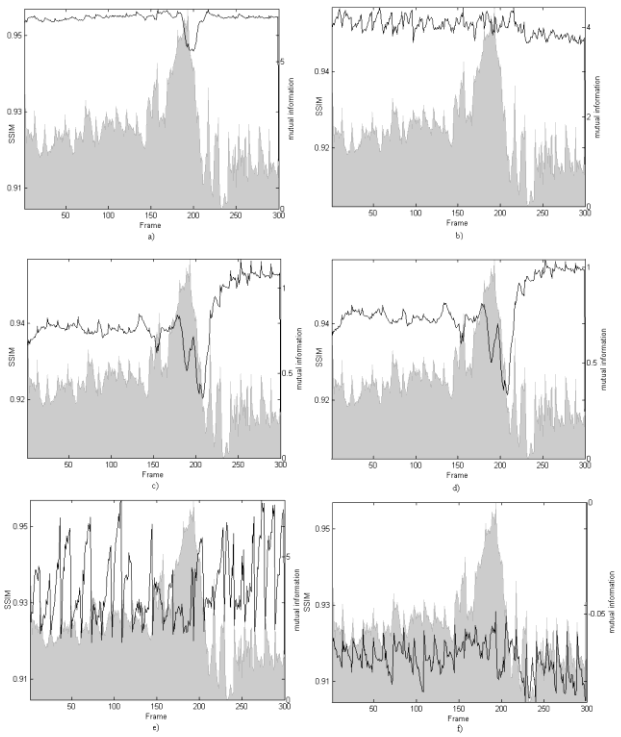


Fig. 1. The left vertical axis and gray curve correspond to the SSIM. The right vertical axis and the black curve correspond to the mutual information with HVS: (a) Gaussian filter, (b) second HVS model, (c) LoG filter $\sigma=1$, (d) LoG filter with $\sigma=1.2$, (e) HVS simulated by $h(1)$, (f) HVS simulated by $h(2)$.

The final correlation is higher even if there is no noticeable improvement or degradation of quality. The mutual information with last HVS model indicates quality very different, lots of peaks and bottoms despite of the SSIM run even if the implementation with $h_2(t)$ has smoother run in comparison with $h_1(t)$.

Comparison between VQM and mutual information pre-processed with the HVS is shown on Fig. 2. At the start of the sequence the run of VQM has some oscillation in terms of the quality. Moving of the head cause changes in the local contrast due to fact that the face is darker area and the helmet is brighter. This appears as oscillations in the VQM quality. Mutual information with each HVS also contains some peaks and bottom in this part of the sequence. The Gaussian filter reduces differences between quality oscillations which do not correspond to the VQM run. In the implementation with LoG filters, the mutual information indicates little improvement of the quality at the beginning but then the peaks better corresponds with VQM. The second HVS model causes that run of the mutual information is rising and falling down in the same frames but in the reverse order. It means that if the VQM indicates improvement of the quality, mutual information indicates its degradation. That is the reason why the final correlation coefficient in the Tab. 1 is negative. When movement of hand blurs just the part of the image the VQM run does not show any evident change, only a very slight improvement regarding the previous oscillations.

In the frame 192 VQM has rapid quality grow. This is caused by moving the camera and thus blurring the frame in the original sequence. Mutual information together with filters for simulate HVS (Gaussian and LoG filters) indicate degradation of the quality at this point. LoG filters have one exception from decreasing trend but this peak shows just slightly quality grow and does not affect overall descending character of the run. The second HVS model has two peaks at this part of sequence and one of them is corresponding with the reducing VQM quality. After major improvement of quality in the frame 192, the overall quality is falling down. Mutual information with first three HVS models start to raise from the frame 192. First and second HVS model reach approximately the same quality as before. However the third HVS model indicates better quality in comparison with the beginning of the sequence. Mutual information with the last HVS model contains many peaks where some of them occur at same place as the VQM peaks and some of them occur where VQM has bottoms.

Fig. 3 shows the comparison between the peak signal-to-noise ratio and mutual information with implemented every HVS model. The quality peaks and bottoms alternate from the beginning of the sequence. As mentioned before mutual information has some quality oscillation for every HVS. By applying the Gaussian filter mutual information run become smoother so the correlation is not very high. In the case where third HVS model is used the changes in the quality corresponds well with the run of PSNR. The second and fourth HVS model cause mutual information to vary more in the quality. Some of the peaks in the second HVS

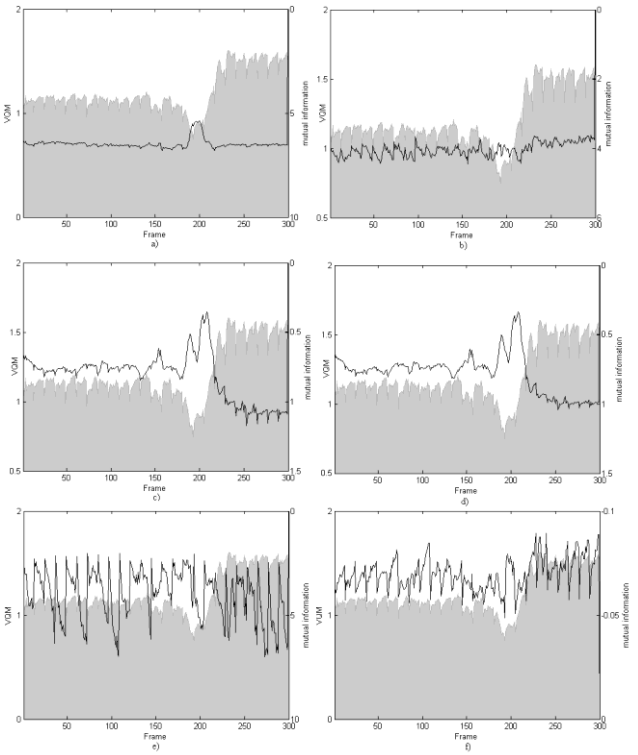


Fig. 2. The left vertical axis and gray curve correspond to the VQM. The right vertical axis and the black curve correspond to the mutual information with HVS: (a) Gaussian filter, (b) second HVS model, (c) LoG filter $\sigma=1$, (d) LoG filter with $\sigma=1.2$, (e) HVS simulated by $h(1)$, (f) HVS simulated by $h(2)$.

model correspond with peaks of the metrics. Slightly improvement of quality in the place where part of the frame is blurred indicates PSNR metric. This change corresponds with the run of mutual information pre-processed by the second and the third HVS model. However in case of LoG filters implementation this change is inverse to the improvement indicated by the PSNR. The mutual information combined by the Gaussian filter does not indicate any noticeable improvement.

Different behavior appears when the whole frame is blurred because of the camera movement. The PSNR quality indicates improvement of the quality in this part of sequence. All mutual information implementations indicate degradation of quality so the correlation between them and PSNR is not well enough except in the case when $h_2(t)$ is used. From this point up the end of the sequence the quality oscillates. PSNR has the peaks and bottoms more visible due to changes of intensity of pixels belongings to wall. Peaks and bottoms of LoG filters implementation best correlate with the PSNR changes. The smoother run of the Gaussian filter does not contain any essential peaks, while the fourth HVS model contains a lot of peaks especially implementation with $h_2(t)$.

Following Tab. 1 shows the correlation coefficient between reference metrics and mutual information with particular HVS implementation:

	metric	SSIM	VQM	PSNR
HVS model	no HVS	0.1424	-0.2196	0.2593
	Gaussian filter	-0.1125	0.0978	-0.0656
	second HVS	0.262	-0.4408	0.4354
	LoG 1.0	-0.4799	0.8085	-0.7843
	LoG 1.2	-0.474	-0.7863	-0.7628
	h1	-0.3387	0.364	-0.3574
	h2	0.4026	-0.4822	0.4746

Tab. 1. The normalized correlation coefficient for test sequence.

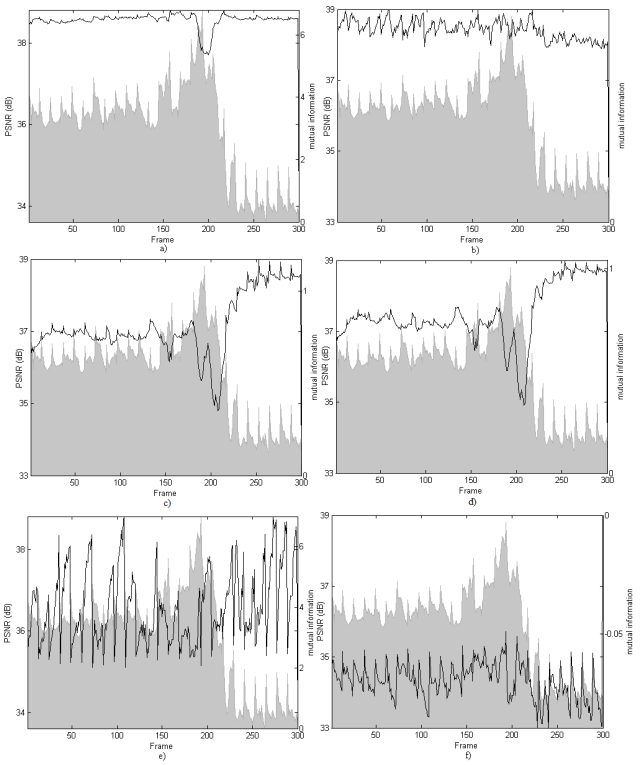


Fig. 3. The left vertical axis and gray curve correspond to the PSNR. The right vertical axis and the black curve correspond to the mutual information with HVS: (a) Gaussian filter, (b) second HVS model, (c) LoG filter $\sigma=1$, (d) LoG filter with $\sigma=1.2$, (e) HVS simulated by *h(1)*, (f) HVS simulated by *h(2)*.

5. Conclusion

In this paper we present improvement of video quality evaluation by mutual information. To prove the reliability of the results provided by our metric we compare it with other three objective metrics. By implementing the simple low-pass Gaussian filter as simulation of human visual perception the run of the mutual information become smoother. In this case the overall results show only little correlation between mutual information and objective methods used for comparison. The second HVS based on band-pass filter and DCT provides better results as the Gaussian filter. However the correlation of the second HVS model is still small. Run of *LoG12* filter is characterized by less peaks and smoother rising unlike the filter with $\sigma=1$ where the crossing between bottoms and peaks is more rapid. By using two different impulse response functions in the last HVS model the number of oscillations in the video quality grow rapidly. From the comparison with other objective methods it can be seen that this model of HVS is not suitable to be used with mutual information even if the correlation coefficient is not the smallest. The best results are provided by LoG filter with parameter $\sigma=1$ where the correlation between the mutual information and the VQM metric is above 0.8. It seems that the mutual information is sensitive on the massive blurriness in the frame and reacts on this fact by the degradation of the quality. The results show that our metric could be useful in objective evaluation of quality. For future work, we would like to run a more complex set of experiments with different video sequences to prove the relevance of the proposed method. Furthermore subjective testing will be necessary to run and also the region of interest can possible improve the obtain results.

Acknowledgements

Research described in the paper was financially supported by the Slovak Research Grant Agency (VEGA) under grant No. 1/0602/11.

References

[1] ITU-R BT.500-11, Recommmendation ITU-R, 2002.

[2] MARTINEZ, J. L., CUENCA, P., DELICADO, F., QUILES, F. Objective video quality metrics: A performance analysis. In *Proceedings of the 14th European Signal Processing Conference*, Florence (Italy), 2006.

[3] WESTEN, S. J. P., LAGENDIJK, R. L., BIEMOND, J. Perceptual image quality based on a multiple channel HVS model. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Detroit (USA), 1995, p. 2351-2354.

[4] AVCIBAŞ, I., SANKUR, B., SAYOOD, K. Statistical evaluation of image quality measures. *J. of Electronic Imaging*, 2002, vol. 11, no. 2, p. 206-223.

[5] TIWARI, R. B., YARDI, A. R. Dental X-ray Image Enhancement Based on Human Visual System and Local Image Statistics. In *Proceedings of International Conference on Image Processing, Computer Vision and Pattern Recognition*, Las Vegas (USA), 2006, p. 100-108.

[6] MINOO, K., NGUYEN, T. Q. A perceptual metric for blind measurement of blocking artifacts with applications in transform-block-based image and video coding. In *Proceedings of 15th IEEE International Conference on Image Processing*, San Diego (USA), 2008, p. 3152-3155.

[7] YU, Z., WU, H. R. Human Visual System based objective digital video quality metrics. In *Proceedings of 5th International Conference on Signal Processing*, Beijing (China), 2000, p. 1088-1095.

[8] CERNEKOVA, Z. *Temporal video segmentation and video summarization*. Bratislava: Dept. App. Inf., Comenius University, 2009.

Comparison of MFCC and PLP feature extraction for speech recognition purpose

Róbert Kožička, Tibor Trnovský, Juraj Kačur

Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia

Abstract. *Speech recognition process consists of several different but very important steps. One of the most important task is the extraction of suitable parameters from highly redundant speech signal by using appropriate methods. In this article we will focus on differences between mel-frequency cepstral coefficients (MFCC) and perceptual linear predictive (PLP) feature extraction. Also very basic ideas about Hidden Markov Model will be presented. Both signal analysis methods were tested on isolated commands and digits tasks using MOBILDAT-SK database. For this purpose we use HTK 3.4.1 with MASPER training procedure. Results of these experiment are also presented.*

Keywords

Mel-frequency cepstral coefficients, Perceptual linear predictive, Hidden Markov models, HTK, MASPER, speech recognition

1. Introduction

Signal processing using suitable feature extraction is the first step to successful trained models of speech recognizer and this operation is very important in terms of gaining parameters describing the speech signal as closely as possible while reducing its high redundancy. This initial operation is important in implementing recognizers, because on the used feature extraction of speech signal depends their success, and thus affects the result of recognition.

2. Mel-Frequency Cepstral Coefficients (MFCC)

In the current speech recognition systems it is preferred to use a modification of homomorfic analysis presented by mel-frequency ceppstral coefficients. Such a speech parameterization is designed to maintain the characteristic of human sound perception. Compensation for non-linear perception of frequency is implemented by the bank of triangular band filters with the linear distribution of frequencies along so called mel-frequency range. Linear deployment of filters to mel-frequency axis results in a non-linear distribution for the standard frequency axis in hertz. Definition of the mel-frequency range is described by the following equation

$$f_m = 2595 \log_{10}(1 + \frac{f}{700}), \tag{1}$$

where f is frequency in linear range and f_m is the corresponding frequency in nonlinear mel-frequency range.

The procedure by which the mel-frequency cepstral coefficients are obtained consists of several steps. Input of the system is supplied with signal samples $s(k)$ and the signal pre-emphasis is done. It is filtration to emphasis the higher frequencies of speech signals which show a greater attenuation. The main frequency components of the speech spectrum are amplified too. Pre-emphasis of the speech signal is realized with this simple FIR filter

$$H(z) = 1 - az^{-1} \tag{2}$$

where a is from interval [0.9,1].

On the next step the signal segmentation with Hamming window is done and window length is set to 10 - 30ms. The exact time length of window (number of samples in the frequency of sampling F_v) is chosen equal to the power of 2 due to the subsequent processing of fast Fourier transformation [1].

In the next step the fast Fourier transform (FFT) is used to calculate components of magnitude spectrum $|S(f)|$ of analyzed signal.

The most important step in this signal processing is mel-freqency transformation. The algorithm is carried out by the bank of triangular band filters with a uniform distribution of center frequencies of each triangular filter along the frequency axis in mel-frequency range.

The next step is to calculate the logarithm of the outputs of individual filters, which affects the dynamics of the signal.

The final step in calculation of the MFCC coefficients is, instead of inverse Fourier transform, the application of the discrete cosine transform (DCT)

which is defined:

$$c_{mf}(n) = \sqrt{\frac{2}{K}} \sum_{j=1}^{K} \log\, m_j \, \cos[n(j-0.5)\frac{\pi}{K}], \quad (3)$$

where n is the number of mel-frequency cepstral coefficients and K is the number of mel-frequency band filters in the bank of filters. Discrete cosine transformation tend to produce non-correlated coefficients for wide range of signals $c_{mf}(n)$, which has a very positive impact on simplification of some steps in the design of the classifiers based on continuous Hidden Markov models [1].

Coefficients of MFCC can be classified as static because all the vector items are obtained from the current weighted microsegment of the signal defined by window functions (Hamming). But there are so called dynamic parameters known as delta (Δc_m) and delta-delta ($\Delta^2 c_m$). Speech signal is time variant therefore with the help of coefficients describing the differences between adjacent parameters in the vectors it is possible to get more accurately characterized speech signal with new and useful information.

3. Perceptual Linear Predictive (PLP)

An alternative to the Mel-Frequency Cepstral Coefficients is the use of Perceptual Linear Prediction (PLP) coefficients. Standard LPC analysis doesn't maintain properties of hearing sound by human ear. For transformation of the energy spectrum of speech signal to the corresponding auditory signal this method combine some fact about psychoacoustic of hearing: critical band of spectral sensitivity, equal-loudness curves and the relationship expressing the dependence between the intensity of sound and its loudness.

There are few important steps of PLP analysis. At the beginning is performed the calculation of short-time energy spectrum of signal for each microsegment $s(k)$. Speech signal is weighted by Hamming window and there are calculated samples of signal spectrum $S(\omega)$ at each microsegment. Short-time energy spectrum of speech signal $P(\omega)$ is defined as

$$P(\omega) = |S(\omega)|^2 = [Re\ S(\omega)]^2 + [Im\ S(\omega)]^2. \quad (4)$$

Modelling of phenomenon such as logarithmic perception of sound or the so-called masking of sounds in the critical bandwidths which size may varies with frequency is in PLP analysis carried out by transforming the frequency axis to axis measured in barks according relation

$$\Omega(\omega) = 6\ ln\left(\frac{\omega}{1200\pi} + \sqrt{\left(\frac{\omega}{1200\pi}\right)^2 + 1}\right), \quad (5)$$

where $\omega = 2\pi f$ is in radians and $\Omega(\omega)$ is in barks, further design of linear distributed bandpass filters as masking curves to simulate the critical bands of hearing. Bandpass filter characteristic and the amplitude frequency response is described in detail in [1]. According to the width of filtered band B_w [Hz] or B_{bw} [bark] are also recommended values for the number of filters and their deployment step.

Next important step is adaptation of bandpass filters to equal-loudness curve. For adjustment of the energy spectrum $P(\omega)$ to properties of human ear, firstly is necessary to carry out pre-emphasis to discrete samples from curves simulating bandpass filter m-th critical zone and the corresponding values approximating equal-loudness curves $E(\omega)$

$$\Phi_m\big(\Omega(\omega)\big) = E(\omega)\ \Psi\big(\Omega(\omega) - \Omega_m\big), \quad (6)$$

where Ω_m[bark] is the mean frequency of m-th critical bandpass filter $m = 0, ..., M-1$. Function $E(\omega)$ is suggested as a rough approximation of equal sensitivity of human hearing for a variety of frequencies. Equation of transfer function for different volume levels is described in [1].

Next step is weighted spectral summarization of energy spectrum samples. The values of the energy spectra $P(\omega)$ after passing an m-th critical bandpass filter suited to the equal-loudness curves volume can be expressed

$$\Xi(\Omega_m) = \sum_{\Omega=\Omega_m-2,5}^{\Omega_m+1,3} P(\Omega)\Phi_m(\Omega), \quad (7)$$

or

$$\Xi(\Omega_m) = \sum_{\omega=\omega_{md}}^{\omega_{mh}} P(\omega)\Phi_m(\Omega(\omega)). \quad (8)$$

Summation limits can be determined from the inverse relationship to equation 5.

In the next step a conversion from the intensity to a perceivable measure of loudness is done according to equation

$$\xi(\Omega_m) = \Big(\Xi(\Omega_m)\Big)^{0,3} \quad (9)$$

also guaranteeing a reduction the amplitude variability at output of critical bandpass filters.

Last step is all-pole spectrum approximation. To implement this approximation is necessary to formulate a relations of linear predictive analysis for frequency spectrum described in [1].

4. Basic characteristic of Hidden Markov model

Hidden Markov model is a model of dynamic stochastic process, which can be seen as a probabilistic finite automat, which in discrete time generates a random sequence of observations $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots . \mathbf{o}_T\}$. In each step the model changes its state according to transition probabilities a_{ij}. State s_j, that models particular speech unit, generates the observation vector $\mathbf{o}_t$, according to the output probability distribution $b_j(\mathbf{o}_t)$ [1].

HMM λ is characterized by the following triplet probabilities

$$\lambda = (\pi, A, B), \tag{10}$$

where π is a vector of initial state probabilities, A is a matrix of transition probabilities and B is a matrix of probabilities of generating observation.

The initial probability π indicates the probability of each state at time t_0 . It holds

$$\sum_{i=0}^{N} \pi_i = 1, \tag{11}$$

where N is the number of states in model.

Transition probability determines how likely the model transit from the state i, at the time t to state j at time $t + 1$. In the speech recognition systems left-right HMM models are used for speech units.

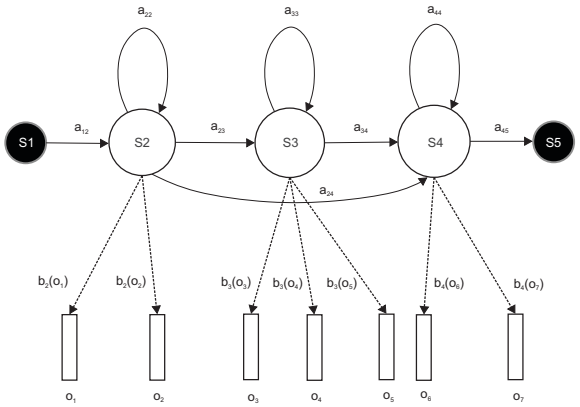


Figure 1: Hidden Markov model

This model allows the transition to the next state or remains in the same state. Transition probability matrix has non-zero values only at the main and secondary diagonal. The condition for the total probability is

$$\sum_{j=0}^{N} a_{ij} = 1, \tag{12}$$

where N is the number of HMM states. Matrix of transition probabilities for the Fig. 1 looks like

$$A = \begin{bmatrix} 0 & a_{12} & 0 & 0 & 0 \\ 0 & a_{22} & a_{23} & a_{24} & 0 \\ 0 & 0 & a_{33} & a_{34} & 0 \\ 0 & 0 & 0 & a_{44} & a_{45} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$b_j(\mathbf{o}_t)$ is a function of probability density generating the vector $\mathbf{o}_t$ at time t in the state j. We use mixed Gaussian probability density given by

$$b_j(\mathbf{o}_t) = \sum_{m=1}^{M_r} C_{jm} \mathcal{N}(\mathbf{o}_t, \mu_{jm}, \Sigma_{jm}) \tag{13}$$

where M_r is the number of mixtures, C_{jm} is weight m-th component and $\mathcal{N}(\mathbf{o}_t, \mu_{jm}, \Sigma_{jm})$ is multidimensional Gaussian distribution with mean μ and covariance matrix Σ. Where

$$\mathcal{N}(\mathbf{o}, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{\frac{1}{2}(\mathbf{o}-\mu)^T \Sigma^{-1} (\mathbf{o}-\mu)}, \tag{14}$$

where n is a dimension of vector $\mathbf{o}$.

We assume that each sequence of observed speech vectors $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots . \mathbf{o}_T\}$ was generated by hidden Markov model λ. The probability that the observation vector $\mathbf{O}$ was generated by hidden Markov models λ shown in the Fig.1 we can calculate as follows

$$P(\mathbf{O}, S|\lambda) = a_{12} b_2(\mathbf{o}_1) a_{22} b_2(\mathbf{o}_2) a_{23} b_3(\mathbf{o}_3) a_{33} \ldots \tag{15}$$

Practicaly we know only the sequence of observations $\mathbf{O}$, states of the model are hidden. That is why this model is called hidden Markov model. Probability generating sequence $\mathbf{O}$ with model λ must be calculated as sum over all possible sequence of the model states

$$P(\mathbf{O}|\lambda) = \sum_{X} a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(\mathbf{o}_t) a_{x(t)x(t+1)}, \tag{16}$$

where $x(0)$ is the input state of model a $x(T+1)$ is the output state of model. These steates don't generate any observation; they are non-emitting states.

5. Experiment and Results

MASPER training procedure needs to set several configuration files before it can be lunched. We focus on that files which have something to do with extraction of parameters. Files located in `CONFIG` folder are *extfea.cfg* (used in process of analysis of files from speech database) and *reest.cfg* (used in training process). Type of feature extraction is in line `TARGETKIND`. We must run the whole training procedure twice following the necessary run of test scripts for isolated application words and digits. Once with the `TARGETKIND=MFCC_0_D_A_Z` and

once with `TARGETKIND=PLP_0_D_A_Z`. We leave the number of filters in both runs of script at the same value of 12 and do not investigate how influence have number of filters in MFCC or PLP feature analysis.

Although the training process includes initialization models in the process of creating a recognizer for us is essential an models of monophones and tied triphones, therefore we are focus only on them.

6. Conclusions

As can be seen from the results at figure 2 and 4 at the begining of training monophones PLP has little bit worse word error rate as MFCC but with the increasing number of Gaussian mixtures followed by two steps of reestimation the PLP-trained monophones are slightly better at the final monophone models than MFCC. Whereas tied triphones at figure 3 and 5 are based on well trained monophones therefore the lack of training data phenomenom take place here, which causes an inaccurate estimate of model parameters. As we can see this lack of data is more reflected in tied triphone models. This is because the triphones are much more specific than phonemes and thus they require more training data for „better" estimation. Therefore we must find another methods how to improve the overall training process.

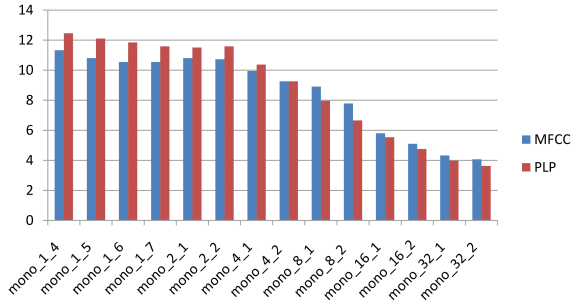


Figure 2: Comparison of using MFCC vs. PLP feature analysis for isolated application words recognition on phoneme models

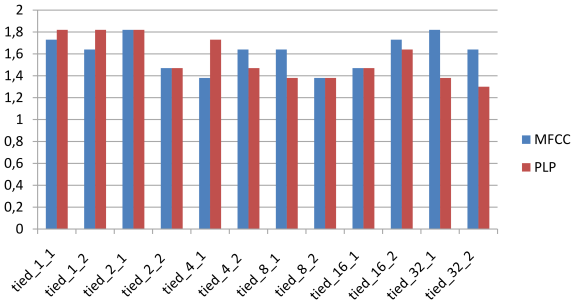


Figure 3: Comparison of using MFCC vs. PLP feature analysis for isolated application words recognition on tied triphone models

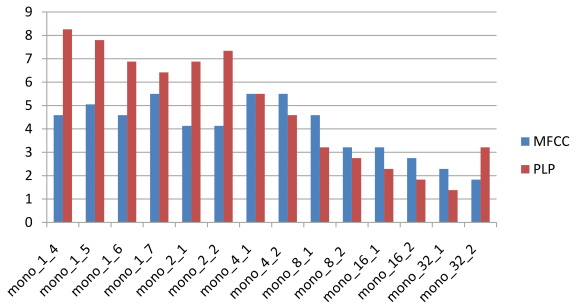


Figure 4: Comparison of using MFCC vs. PLP feature analysis for isolated digits recognition on phoneme models

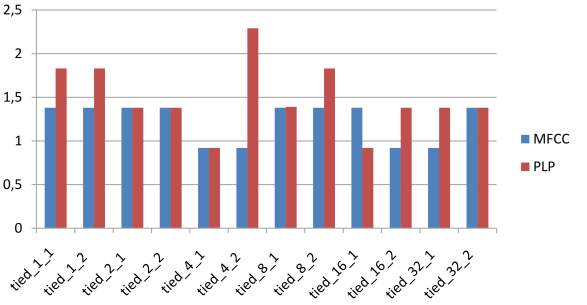


Figure 5: Comparison of using MFCC vs. PLP feature analysis for isolated digits recognition on tied triphone models

Acknowledgement

This article was supported by VEGA - 1/0718/09.

References

[1] J. Psutka, L. Müller, J. Matoušek a V. Radová, *Mluvíme s počítačem česky.* Praha, Česká republika: Academia, 2006.

[2] S. Young, G. Evermann, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D.Ollason, D. Povey, V. Valtehev a P. Woodland, *The HTK Book v. 3.4.1.* Cambridge, England: Cambridge University, 2009.

[3] J.G. Proakis, D.G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications* New Jersey, USA: Prentice-Hall Inc., 1996.

Graphical User Interface of Speaker Dependent Detector for Slovak Phonemes

Lukáš BORIK[1], Ján KŐRÖSI[2]

Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
99thproblem@gmail.com(1) korosi@ktl.elf.stuba.sk(2)

Abstract. *This paper described the creation of simple language phoneme models, which is used in speech recognizer based on measuring of distance of obtained features vectors. For this purpose we designed GUI based application in Matlab. This system allows us select the combination of parametrizations and their options, from which we create the models. Finally, the results are presented in the graph with additional information and also they can be stored into the table.*

Keywords

MFCC, PLP, LPC, Euclidean distance, Mahalanobis distance, Speech recognition

1. Introduction

The most used continual speech recognizers are usually too expensive for computation power, what is caused by large dictionary containing thousands of words. For elimination of this disadvantage there can be some approximation provided. Our system is designed as a utility, which should help to determine some speech characteristics, which allow us to find some suboptimal combination of speech features and it should also shows user the way how where some approximation can be used. We focused on basics type of features extractions, which we want to combine with the state of art types of parametrizations.

2. Creating models of phonemes

We used the database which contains 468 WAV files, whose duration is between one to five seconds. This database is recorded by one speaker, but we planned to use bigger database with 100 speakers. Now it is used just in development phase. This cause, that some phonemes have very small numbers of occurrences. For example phonemes like l=, Q-, l=~, r=~ occur up to 5 times, which is quite low for creation of better model. The reason why we decide to use this database is the present of annotation, which is handmade. This allows us better evaluation of results, which is big advantage in development of systems of this kind.

Each occurrence of phoneme in all WAV files is segmented and we obtained the features vectors from which we compute statistical characteristics, which we used in recognition process as acoustic models. In these phase we determine only phoneme models, but it is also possible to use triphone models.

In recognition process we decide to use several types of measure distances as a criteria function. The input speech signal is segmented by the setting from GUI interface, which can be different from the segmentation used in training of models. For each segment we should calculate the subset of parameterization used in models. From obtained feature vectors we evaluate the distance and determined the best model and the accuracy of recognition.

3. Description of designed system

The aim of our work is to create the user friendly system, which make facilitates the work with the program, which creates acoustic models and calculate, which phoneme have the highest probability to be in segment. This system should be helpfull for people, who are not so familiar with speech recognition and speech processing.

We made graphics form of this system in wizard (Guide) and then we wrote for each object a piece of code that defines the operations with the object. We divide the system into two parts. First one is used for creation of the models and second one for recognition of speech signal. In this system constant segmentation is used, with the possibility to set size of blocks in millisecond and shift of blocks also in millisecond. It is possible to set different segmentation for creation of models and for input speech signal, which will be recognized. We know a lot of kinds of distances measure, which calculates distance between two vectors or among vectors and space of points, for example: Euclidean distance, Log-spectral measure, Mahalanobis distance, Cepstral distance measure.

In our system only two kinds of distance is used. First one is Euclidean distance, which is calculated by formula [6]:

$$D(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (1)$$

Second one is Mahalanobis distance, which is calculated by the formula [6]:

$$D_M(x) = \sqrt{(x-u)^T D^{-1}(x-u)} \quad (2)$$

In this system we can set the parameterization or combination of parameterizations, which are used in process of creation of acoustic models and in recognition of phonemes. We can select from these parameterizations: energy and intensity of signal, zero-crossings rate, PLP, LPC, and MFCC.

Energy of signal is calculated by [5]:

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m) \quad (3)$$

Intensity of signal is calculated by [5]:

$$I_n = \sum_{m=-\infty}^{\infty} |x(m)|\, h(n-m) \quad (4)$$

Zero-crossings rate is calculated by [5]:

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]|\, w(n-m) \quad (5)$$

For LPC we can set the number of LPC coefficients. Transfer function of vocal tract is described by the following formula [5], where a_i are LPC coefficients:

$$H(z) = \frac{1}{1 + a_1 z^{z-1} + \ldots + a_p z^{-p}} \quad (6)$$

For MFCC coefficients, we can set twelve parameters: number of coefficients, exponent for lifting, sum power flag, application of pre-emphasis filter, dither, the lowest frequency, highest frequency, number of warped spectral bands, width of the audio spectral filter, type of DCT, type of frequency waves and model order. MFCC coefficients are calculated according to the formula:

$$mfcc(k) = \sum_{m=1}^{M} log\{|X(m)|\} cos\left(k\left(m - \frac{1}{2}\right)\frac{\pi}{M}\right) \quad (7)$$

The last parameterizations is PLP coefficients, for which can choose the following parameters: number of coefficients, exponent for liftering, sum power, applies pre-emphasis filter, dither, the lowest frequency, highest frequency, number of warped spectral bands, width of the audio spectral filter, type of DCT, type of frequency waves and model order.

Next step in the part of creation of models is design of structure of models. We decided for the folders structure. The first folder is defines what distance will use with models. We use two kinds of distances: Euclidean distance or Mahalanobis distance. Models for each distance are different. Model for Euclidean distance consists of average values of parameterizations and model for Mahalanobis distance consists of the average values of parameterizations and covariance matrix of parameterizations. The second folder defines the segmentation of model. Type of segmentation is written in the folder name. The first number is the size of block in millisecond and the second number is the size of shift in millisecond. In this folder are models for parameterization, which have not more parameters, such as: energy of signal, zero-crossings rate and intensity of signal. Parameterization with more parameters (LPC, MFCC and PLP) is in other folders, which is called as the name of parameterization. In this folder are the models in the structure form. Models are stored in the format of MAT files, native format of Matlab. Figure 1 shows the folder structure of the models.

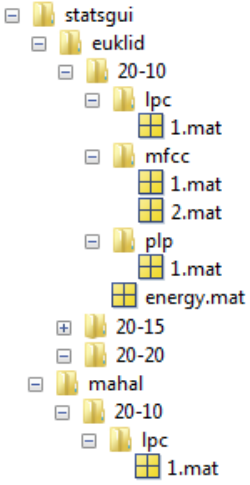


Fig. 1. The structure of the models

One from option, which the user can set in the graphical system is kind of input speech signal in recognition of phonemes. The first possibility is that the speech signal has been loaded from WAV file. This option has the advantage that if WAV file has annotation, so then we can see the success of recognition. The second option is using the microphone as input. For this we can set time of recording in seconds and the sampling frequency in Hz.

In the current system some support parameterizations can be used. These parameterizations may increase success of recognition and reduce processing time. As supported parameterizations we use autocorrelation function. It is calculated by the following formula [5]:

$$R(i) = \frac{1}{N}\sum_{n=0}^{N-1-i} s(n)s(n-1) \quad (8)$$

Autocorrelation function detects which segments are voiced or unvoiced. Recognizer does not calculate the distance for voiced phonemes in voiceless segment and vice versa. Acceleration of calculation can be seen in indicators of the CPU time.

Recognition phonemes can be presented in the graph, which are the input speech signal and annotation. Here we can see if recognition phonemes are agree with annotation. Example of graph we can see on Fig. 1.

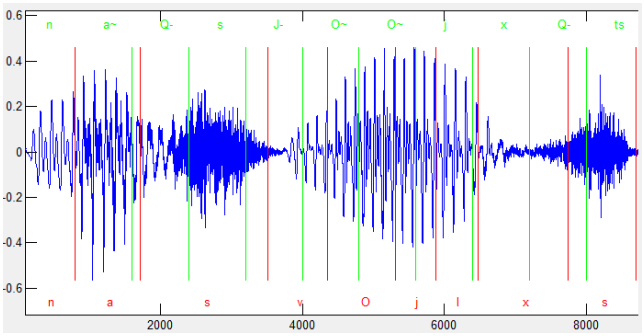


Fig. 1. On figure is part of input speech signal from WAV file. Green is displayed recognition phonemes and red is displayed annotation. In recognition is used Euclidean distance, size of block is 20 ms, shift of block is 10 ms. The same segmentation is using in acoustic model. As parameterizations are using energy of signal, eight LPC coefficients and eight PLP coefficients.

On Fig. 1, we can see that some phonemes are match with phonemes form annotation.

If we use as input speech signal WAV file with annotation, then graphical system display us the success of recognition. Program calculates two kind of success: phonemes recognition and group of phonemes recognition. As recognition phoneme in the segment is the phoneme with the smallest distance. Success phonemes recognition is defined as that, if recognition phoneme is match with the annotation phoneme, then recognition is successful. Another success is the group recognition. Base group (Vowels, Consonants, Vowels Open etc.) are predefined. Other groups can be created. Success of group recognition is defined as if recognition phoneme is the same group as the annotation phoneme, and then recognition is successful.

4. The success of detection

Finally, we tested the success of detection on the one man speaked database. This database has been handmade annotated, which is consider as a big advantage. We create from results the tables with the Euclidean distance (in Tab.1) and for Mahalanobis (in Tab. 2).

Combination of the parameterizations	Segmen-tation ms	Phoneme %	Group %	CPU time
PLP 10	20-10	27,55	77,89	102,2
MFCC 12	30-15	30,75	77,41	37,1
PLP 10	40-40	27,19	75,92	36,8
MFCC 12	20-10	27,74	75,63	57,3
E, MFCC 10	30-15	32,95	75,58	42,1
LPC 8	30-15	32,48	75,45	39,6
E, LPC 8, MFCC 8	30-15	32,48	75,45	39,5
MFCC 8	40-40	30,71	74,90	19,5
LPC 10 , MFCC 8	20-10	31,61	74,82	59,4
E, MFCC 8, PLP 8	20-10	31,48	74,81	143,8
I	20-10	3,71	72,51	10,0
LPC 12	20-10	16,41	71,37	15,0
E	40-40	3,87	71,17	7,0
LPC 8	25-12.5	18,86	70,87	11,5
E	20-10	4,08	69,42	13,4
E, ACF	20-10	5,37	66,90	36,5
CZ	20-10	10,80	59,74	09,5

Tab. 1. Result success of detector for Euclidean distance.

In the first column is a combination of the parameterizations, which were used in the detector, where:

- *E is energy of signal*
- *I is intensity of signal*
- *CZ is zero-crossings rate*
- *ACF is autocorrelation function*
- *LPCxxx, MFCCxxx and PLPxxx are the well-known parametrizations with number of used coefficients*

The second column is the segmentation of models and the segmentation in the detector. In the third column is the success of detection phonemes and in the fourth is success of detection group. As group we chose vowels. The last column is CPU time in seconds.

In Tab. 1, we can see that the largest success have MFCC and PLP coefficients. And when we used largest segmentation, the CPU time is reduced more than half and success is only slightly reduced. Autocorrelation function did not change the success and CPU time is increased. Autocorrelation function increase the time of calculation about one hundred seventy percent and reduced the percentage of recognition about three percent. Autocorrelation function is not suitable when ingested

Euclidean distance. Energy of signal and intensity of signal have good results for detection of vowels, but for detect phonemes are very badly. Zero-crossings rate has the worst results of detection. The combination of parameterization is not good idea. The combination of parameterization only increase the CPU time.

Combination of the parameterization	Segmen-tation ms	Phoneme %	Group %	CPU time
MFCC 8, PLP 10, ACF	40-40	37,40	79,19	56
LPC 10, MFCC 8, ACF	30-15	37,86	79,00	259
PLP 10, ACF	30-15	37,36	78,63	149
MFCC 10, ACF	30-15	34,54	77,64	112
LPC 8, PLP 10, ACF	40-40	38,66	77,32	70
MFCC 12	20-10	38,95	75,54	412
LPC 10	20-10	30,22	73,68	290
E, LPC 10, MFCC 8, PLP 10, ACF	20-10	30,68	72,90	1211
LPC 10, ACF	20-10	23,89	72,75	191
PLP 10, ACF	20-10	32,05	72,59	281
I	20-10	2,41	72,32	49
MFCC 12, ACF	20-10	27,49	71,51	286
E	20-10	3,98	70,33	54
CZ	20-10	8,87	59,34	49

Tab. 2. Result success of detector for Mahalanobis distance.

Also in the Tab. 2 we can see that the largest success have MFCC and PLP coefficients and a combination of this parameterization. Autocorrelation function does not increase the percentage, but reduced the CPU time almost half. Autocorrelation function reduce the time of calculation about thirty percent and reduced the percentage of recognition about two percent. Autocorrelation function is suitable for detection with Mahalanobis distance. Energy of signal, intensity of signal and zero-crossings rate has the same results as in Euclidean distance. The combination of parameterization is good, but grows up the CPU time.

Mahalanobis distance has a higher success rate for detection of phonemes than Euclidean distance, but Euclidean distance has less CPU time than Mahalanobis distance. In the future, we will use cepstrum for detection of voiced segments and formants to better detection of vowels. These methods will be accelerate the calculation of recognition and increase the percentage of recognition

When we use the Euclidean distance, twelve MFCC coefficients, size of segment is 30 ms and shift of segment is 15 ms, so we get enough information about vowels position Our solution allow us reduce looked for state space, what increase the speed of recognition process. The less good method is energy of signal with Euclidean distance. The percentage reduced by seven percent but CPU time is reduced to one third. Energy of signal is good method for detection of vowels in cases, when we need a little computation power. Mahalanobis distance is useless in real time for his high computation power.

Acknowledgements

The authors hereby declare that the article was founded by the grant: VEGA 1/0718/09 and FP7-ICT-2011-7.

References

[1] PSUTKA, J: Komunikace s počítačem mluvenou řečí, ACADEMIA PRAHA, 1995, ISBN 80-200-0203-0

[2] QINQ, REN.: This function is to calculate MFCC coefficients. [online]. Updated 25-3-2008 [cit. 4-3-2010]. Available at internet: <http://read.pudn.com/downloads105/sourcecode/speech/430698/mfcc.m__.htm>

[3] ELLIS, D.: PLP and RASTA (and MFCC, and inversion) in Matlab using melfcc.m and invmelfcc.m. [online]. Updated 3-7-2006 [cit. 29-3-2010]. Available at internet: < http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/>

[4] HERMANSKY, H: Perceptual linear predictive (PLP) analysis of speech. [online]. Updated 21-8-1989 [cit. 3-4-2010]. Available at internet: <http://148.204.64.201/paginas%20anexas/voz/articulos%20interesantes/front%20end/PLP/PLP.pdf>

[5] RABINER, L. R., SCHAFER, R. W. Digital processing of speech signals. Prentice-Hall, Inc., Englewood Cliffs, New Jersey. 1978.

[6] DEZA, E, DEZA, M.M.: Encyclopedia of Distances. Springer-Verlag Berlin Heidelberg. 2009

[7] Kőrösi, Ján - Vojtko, Juraj – Rozinaj, Gregor: Statistic Evaluation of Various Speech Parameters for Phonemes in Slovak Language, In: 52nd International Symposium ELMAR-2010. Zadar, Croatia, 2010, pp. 375-378

[8] Divičan, Martin - Vojtko, Juraj - Kőrösi, Ján: Recognition of Slovak Phonemes using Support Vector Machines, In: 4th International Workshop on Speech and Signal processing Redžúr 2010, May 14, 2010, Bratislava, Slovak Republic.

[9] Kačur, Juraj - Vojtko, Juraj: Efficient Adaptations of the SphinxTrain Procedure for Building a Robust ASR System in Slovak In: 15th International Conference on Systems, Signals and Image Processing IWSSIP 2008, June 25-28 2008, Bratislava, Slovak Republic, pp. 1-4, ISBN 978-80-227-2856-0, 978-80-227-2880-5, IEEE Catalog Number CFP0855E-PRT, CFP0855E-CDR

A Novel Technique of Frames' Comparison for Video Cut Detection

Lenka Krulikovská[1], *Jaroslav Polec*[1]

[1] Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
krulikovska@ktl.elf.stuba.sk, polec@ktl.elf.stuba.sk

Abstract. *In this paper we present a novel technique of frames' comparison for the abrupt cut detection. An abrupt cut is the most frequent kind of transition between shots, therefore their detection is a very important task in the field of video analysis. The majority of existing methods compare pairs of successive frames. We compare actual frame with its motion estimated prediction. We have chosen Pearson correlation coefficient for evaluating of frames' similarity. We also propose a novel method of adaptive threshold to evaluate the accuracy of shot cut detection. The effectiveness of proposed method was verified through test experiments and the obtained results were compared with those achieved by existing logic of frames' comparison. The proposed method gives better results, what is demonstrated by calculated recall and precision. Other advantage of presented method is it can be used in video encoding process without significant increase of computational complexity to enable the use of GOP structure adaptable to video content. Video encoding based on adaptive GOP provides higher coding efficiency and save bandwidth needed for video transmission.*

Keywords

Shot cut detection, Pearson correlation coefficient, motion estimation, threshold.

1. Introduction

Progress in the multimedia compression technology and computer performance has led to the widespread availability of digital video. There is a corresponding growth in the need for methods to reliably detect shot boundaries within the video sequence. The detection of shot boundaries provides a base for nearly all video abstraction and high-level video segmentation approaches. Therefore, solving the problem of shot-boundary detection is one of the major prerequisites for revealing higher level video content structure. Moreover, other research areas can profit considerably from successful automation of shot-boundary detection processes as well.

There are a number of different types of transitions or boundaries between shots [1]. A cut is an abrupt shot change that occurs in a single frame. A fade is a slow change in brightness usually resulting in or starting with a solid black frame. A dissolve occurs when the images of the first shot get dimmer and the images of the second shot get brighter, with frames within the transition showing one image superimposed on the other. A wipe occurs when pixels from the second shot replace those of the first shot in a regular pattern such as in a line from the left edge of the frames. Of course, many other types of gradual transition are possible.

Different approaches have been proposed to extract shots. The major techniques used for the shot boundary detection are pixel differences, statistical differences, histogram comparisons [2], edge differences, compression differences and motion vectors [3, 4, 5].

There are various possibilities for improving on the basic methods. The variety of basic methods opens up the possibility of combining several of them into a multiple expert framework, explored in [6, 7, 8]. Also, one can use an adaptive threshold setting, by using statistics of the dissimilarity measure within a sliding window [9, 10, 11].

The majority of proposed and published techniques use comparison of two successive frames. We propose a novel technique of comparison, where the actual frame is compared with its motion compensated prediction. Proposed method was verified through test experiments and evaluated with frequently used frame by frame based comparison. In general, abrupt transitions are much more common than gradual transitions, accounting for over 99% of all transitions found in video [12]. Therefore, we focus only on the detection of an abrupt cut.

The paper is structured as follows: in the second section a proposed method of shot cut detection is described. Results are displayed in the third section. All results are summarized and discussed in conclusion.

2. Shot cut detection

The novelty of presented method is in the evaluation of the positions of abrupt cut. The most of existing methods calculate similarity of two consecutive frames by chosen metric and determine the position of cut based on obtained values.

Our proposed method compares the actual frame with its prediction. Thus, this method can be performed during video encoding process. For evaluating the similarity of the

frame and prediction we have chosen Pearson correlation coefficient as a representative of correlation metrics.

Fig. 1 illustrates the differences among the proposed methods and existing ones. The arrows indicate which frames are compared.

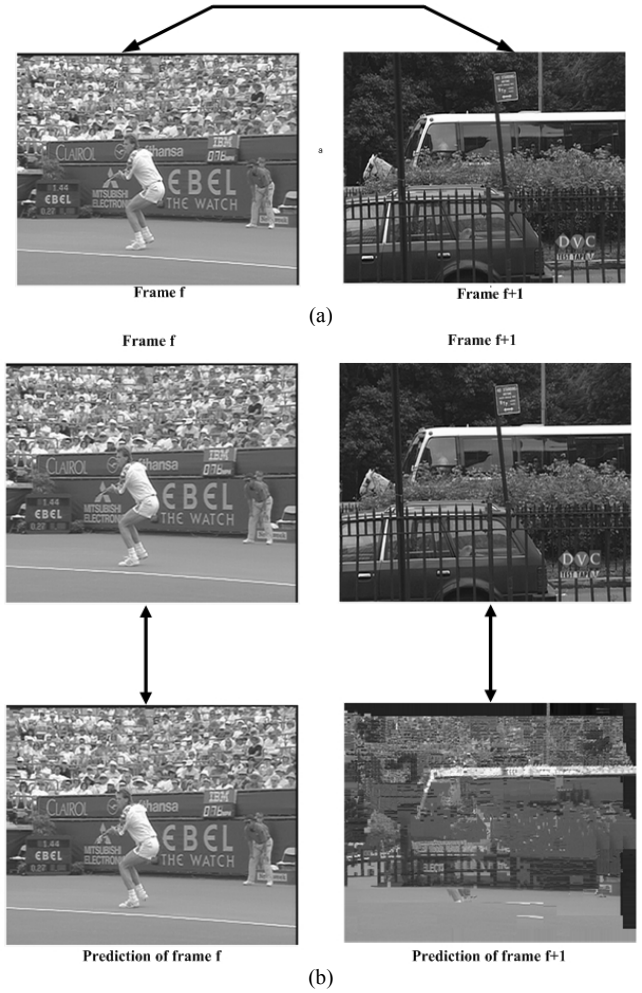


Fig. 1. The principle of comparison in shot cut detection process for the a) majority of published methods and b) proposed method.

2.1 Pearson correlation coefficient

In statistics, the Pearson's correlation coefficient typically denoted by r (sometimes also referred to as the Pearson product-moment correlation coefficient) has been widely employed to measure the correlation (or strength of linear dependence) between two variables X and Y [13]. The value for a Pearson correlation coefficient can fall between 0 (no correlation) and 1 (perfect correlation). Generally, correlations above 0.80 are considered as really high. Therefore the lowest will be determined as cuts.

The Pearson correlation coefficient for 2D signals like video sequences is expressed as follows [13]:

$$r = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(f(i,j)-f^{m})(f_p(i,j)-f_p^{m})}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(f(i,j)-f^{m})^2(f_p(i,j)-f_p^{m})^2}} \quad (1)$$

where M and N stand for dimension of frames f and its prediction f_p. $f(i,j)$ and $f_p(i,j)$ express the pixel intensity for $(i,j)th$ element of frames. f^m is the mean pixel intensity of the frame f and f_p^m is the mean pixel intensity of its prediction f_p.

3. Experimental results

We confirmed the effectiveness of proposed method through a test experiment. The obtained results are compared with results of method using consecutive frames' comparison.

For test purposes we created a video yuv sequence (1989 frames) at CIF resolution (352 x 288 pixels) with 7 abrupt cuts sampled at rate of 30 frames per second. The test video sequence consists of eight standard test sequences: akyio, foreman, hall, flower, mobile, mother-daughter, stephan and bus. For prediction of frames we have employed motion estimation scheme used in H.264 video encoding standard. The Pearson correlation coefficient is calculated for each component Y, U and V. The total value for YUV is computed as an average of components' values.

Fig. 2. shows results for abrupt cut detection. All shots were detected by both methods, they are represented by a significant decrease in the value of the Pearson correlation coefficient. Proposed method reached values higher than 0,8 for all non cuts frames. In addition the proposed method suppressed local extremes caused by huge motion in test sequences. This assures higher robustness to object or camera motion and decreases the probability of false detections.

The use of threshold is needed for automatic shot boundary detection. We can use fixed or adaptive threshold. For fixed threshold it is needed to select an appropriate value, otherwise the shot boundary detection would achieve poor results. The efficiency of shot cut detection algorithm can be evaluated by recall, precision and F1 score.

The recall measure, also known as the positive true function or sensitivity, corresponds to the ratio of correct experimental detections over the number of all true detections. The precision measure is defined as the ratio of correct experimental detections over the number of all experimental detections. F1 score measure s a combined measure that results in high value if, and only if, both precision and recall result in high values.

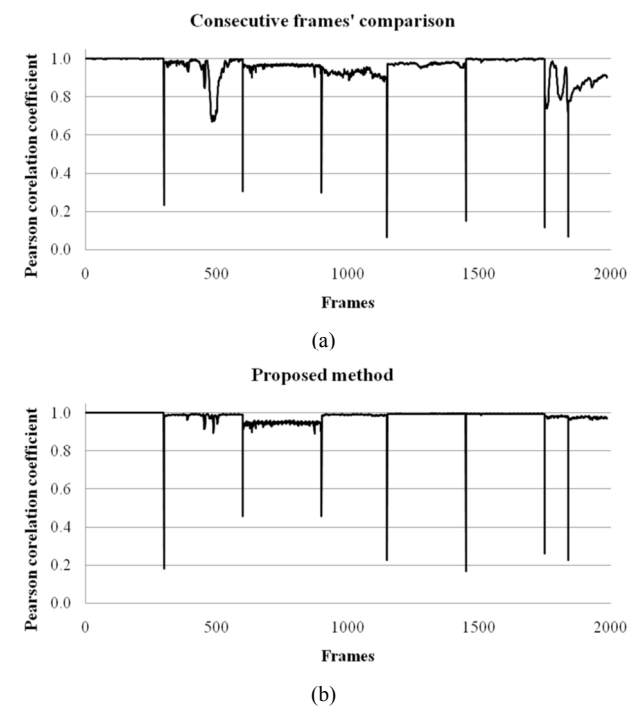


Fig. 2. Plot of Pearson correlation coefficient for a) majority of published methods and b) proposed method.

To show the impact of threshold selection we ran simulation with the values of threshold from 0,001 to 1 with step 0,001. Every value under selected threshold is classified as shot cut. The dependency of recall, precision and F1 score to threshold are displayed on Fig.3. – Fig.5 for proposed method and method using existing logic of frame comparison.

As we can see, the method based on existing approach in frames' comparison reached the value 1 for recall faster. Existing method reached lower values for positions of cuts, so all of the cuts are detected with lower thresholds.

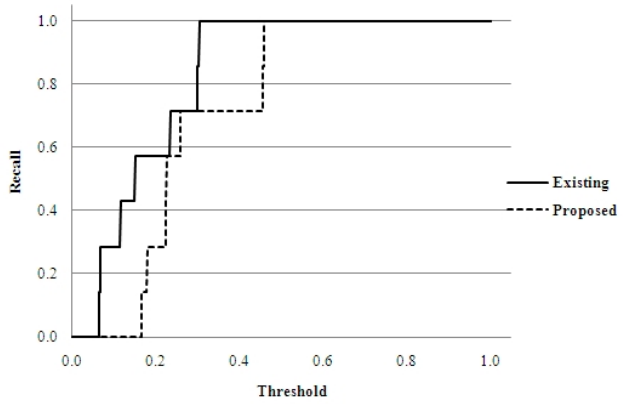


Fig. 3. Recall for majority of published methods and proposed method.

Similar situation can be observed for precision measure. Existing method reached the maximum value 1 with lower threshold, but the value starts to decrease also in lower threshold's value. This method bahaves in this way due to local extremes caused by motion activity in video sequences, what lead to false cut detection. The proposed method is more stable, it keeps the precision value 1 for about 70% of threshold's range (the existing approach achieve about 50%).

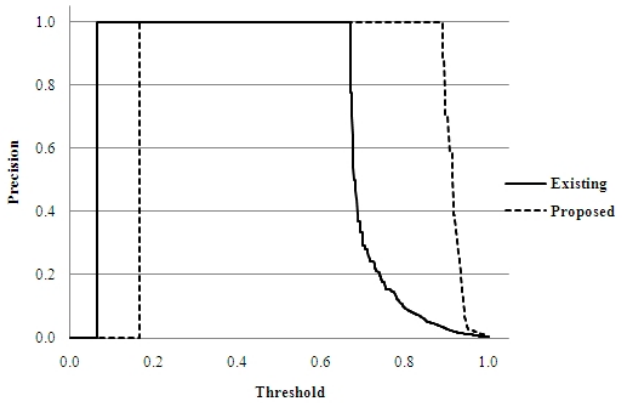


Fig. 4. Precision for majority of published methods and proposed method.

F1 score takes into account both the precision and recall measure. The proposed method holds the highest possible value (1) for about 50% of threshold's range in contrast to less than 40% achieved by method based on the existing approach. It proves that the presented method is more stable according to threshold selection. However, too high or too low chosen value of threshold would cause the decrease of detection accuracy for both methods.

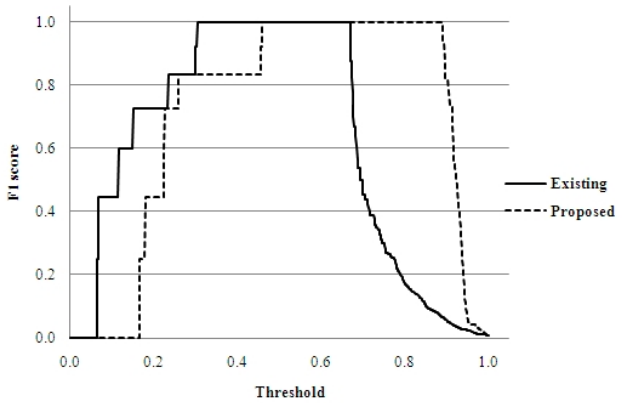


Fig. 5. F1 score for majority of published methods and proposed method.

We have also proposed two versions of adaptive threshold to provide automatic shot boundary detection. The initial value of threshold is set to 0,5 and after detecting the first cut it is set to value of found cut multiplied three and four times respectively.

Tab. 1. illustrates the results obtained for automatic shot boundary detection with proposed adaptive threshold measured by precision, recall and F1 score. Proposed method achieved value 1 for all measures for both version of adaptive threshold. The method based on existing frames' comparison reached F1 score 0,45161 for first version of adaptive threshold and 0,027237 for second one.

Threshold multiple	Shot boundary method	Recall	Precision	F1 score
3	Existing	1	0,29167	0,45161
	Proposed	1	1	1
4	Existing	1	0,013807	0,027237
	Proposed	1	1	1

Tab. 1. The results of shot boundary detection with proposed adaptive threshold.

4. Conclusion

This paper presents a novel technique of frames' comparison for video shot boundary detection. The majority of existing methods compares successive frames, our approach is to compare actual frame with its motion estimated prediction. We have chosen Pearson correlation coefficient for evaluating the similarity of compared frames.

The efficiency of proposed method was verified through test experiments and compared with results of method based on existing approach. The proposed method suppresses local extremes caused by motion activity, which are visible for existing methods, and could lead to false cut detections.

We ran analyses for fixed threshold in the range of all values reachable by Pearson correlation coefficient. These analyses were evaluated by recall, precision and F1 score measures. The results show the proposed method is more stable and holds the maximal accuracy for more values of threshold.

We also propose two versions of adaptive threshold. The proposed methods achieve significantly better results in comparison to commonly used technique.

The next advantage of proposed method is in its simplicity and it can be used for determining adaptive GOP structure during video encoding process without any significant increase of the computational complexity. Using adaptive GOP structure improves coding efficiency and helps to save bandwidth needed for video transmission. Finally we would like to examine the impact of the proposed method on video traffic prediction combined with methods [14, 15, 16].

Acknowledgements

Research described in the paper was financially supported by the Slovak Research Grant Agencies: KEGA under grant No. 119-005TVU-4/2010 and VEGA under grant No. 1/0602/11.

References

[1] CERNEKOVA, Z. *Temporal Video Segmentation and Video Summarization*, Ph.D. dissertation, Dept. App. Inf., Comenius Univ., Bratislava, SK, 2009.

[2] AMIRI, A., FATHY, M. Video shot boundary detection using QR decomposition and Gaussian transition detection. *EURASIP Journal on Advances in Signal Processing,* Volume 2009, Article ID 509438.

[3] HANJALIC, A. Shot-boundary detection: unraveled and resolved?. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 2, pp. 90–105, 2002.

[4] BORECZKY, J. S., ROWE, L. A. Comparison of video shot boundary detection techniques. In *Storage and Retrieval for Still Image and Video Databases IV*, Proc. SPIE 2664, pp. 170-179, 1996.

[5] LIENHART, R. Comparison of automatic shot boundary detection algorithms. In *Storage and Retrieval for Image and Video Databases VII,* vol. 3656 of Proceedings of *SPIE*, pp. 290–301, San Jose, Ca, USA, 1999.

[6] NAPHADE, M. R., MEHROTRA, R., FERMAN, A. M., WARNICK, J., HUANG, T. S., TEKALP, A. M. A high-performance shot boundary detection algorithm using multiple cues. In *Proc. IEEE Int. Conf. on Image Proc.*, volume 2, pages 884–887, 1998.

[7] TASKIRAN, C., DELP, E. J. Video scene change detection using the generalized sequence trace. In *Proc. IEEE Int. Conf. on Image Proc.*, pages 2961–2964, 1998.

[8] YUSOFF, Y., KITTLER, J., CHRISTMAS W. Combining multiple experts for classifying shot changes in video sequences. In *Proc. 6th Int. Conf. on Multimedia Comp. and Systems (ICMCS)*, volume 2, pages 700–704, Florence, Italy, 1999.

[9] DUGAD, R., RATAKONDA, K., AHUJA, N. Robust video shot change detection. In *IEEE Workshop on Multimedia Signal Processing*, 1998.

[10] YEO, B. L., LIU, B. Rapid scene analysis on compressed video. In *IEEE Trans. On Circuits and Systems for Video Technology*, 5(6):533–544, 1995.

[11] ZABIH, R., MILLER, J., MAI, K. A feature-based algorithm for detecting and classifying production effects. *ACM Multimedia Systems*, 7(2):119–128, 1999.

[12] PASCHALAKIS, S., SIMMONS, D. (2008, April 24), Detection of gradual transitions in video sequences [Online]. Available: http://www.wipo.int/pctdb/en/wo.jsp?WO=2008046748&IA=EP2007060594&DISPLAY=STATUS.

[13] EUGENE, Y. K., JOHNSTON, R. G. *The Ineffectiveness of the Correlation Coefficient for Image Comparisons*, Technical Report LA-UR-96-2474, Los Alamos, 1996.

[14] ORAVEC, M., PETRÁŠ, M., PILKA, F. Video Traffic Prediction Using Neural Networks, *Acta Polytechnica Hungarica*, Budapest, Hungary, ISSN 1785-8860, Vol.5, No.4, 2008, pp.59-78

[15] MRAČKA, I., ORAVEC, M. Classification of Traffic of Communication Networks by Multilayer Perceptron, *Proc. of International Conference New Information and Multimedia Technologies NIMT–2008*, September 18-19, 2008, Brno, Czech Republic, ISBM 978-80-214-3708-1, pp. 46-49

[16] PROCHASKA, J., VARGIC, R.: Using Digital Filtration for Hurst Parameter Estimation. *Radioengineering*, Vol. 18, No.2, June 2009, pp. 238-241

Performance of Principal Component Analysis in Different Applications

Peter VISZLAY, Jozef JUHÁR

Dept. of Electronics and Multimedia Communications, Technical University of Košice, Park Komenského 13, 041 20 Košice, Slovakia
peter.viszlay@tuke.sk, jozef.juhar@tuke.sk

Abstract. *In this paper the current state of our work in application of Principal Component Analysis (PCA) in different recognition tasks is presented. PCA is a well-known data processing and linear dimension reduction method. It is used to refine the basic feature extraction process in many speech recognition systems. This paper describes five applications of PCA divided into two categories. The first one is the application of PCA in the TIMIT phoneme recognition task under different conditions and modifications. The second category presents a PCA-based feature extraction method in acoustic events detection (AED) system. For all applications the experimental conditions are described and the results of experiments are listed. In the conclusions a possible proposals to improve the used method are outlined. The whole paper is finished with description of future work.*

Keywords

Acoustic event, Dimension, Feature vector, Linear transformation, Matrix, Principal Component, Supervector.

1. Introduction

Linear feature transformations (LTs) are widely used in most automatic speech recognition systems. LTs can improve the recognition performance of the baseline system by transforming and reducing the feature vectors. This step can also speed up the acoustic models training and recognition process. LTs are applied in the feature extraction step, which is an important part of the whole recognition process. The goal of LT is to determine an optimal transformation matrix with respect to some optimization criterion. The most popular transformations are Linear Discriminant Analysis (LDA) [1] and Principal Component Analysis (PCA) [2] used also in our experiments.

This paper is organized as follows. The next section deals with related work. Section 3 describes the PCA. Section 4 presents the experimental results of TIMIT phoneme recognition. In Section 5 the acoustic events detection using PCA is presented. The paper is finished with conclusions and proposals to future work.

2. Related works

In our experiments the TIMIT speech database [3] was used to evaluate the speech recognition system. Several works, concerning PCA [4], [5] and [6] are focused to applications evaluated especially on this database with using some additional methods. The concatenated vector approach (used also in our work) was used in [5] and it is a fundamental operation in LDA application [1] in speech recognition. In [5] the PCA was applied to the multi-frame context windows. In [6] the researchers have applied LP transformation (concatenating the LDA and PCA coefficients into one vector) to transform the base feature vectors.

In context of acoustic events detection (AED) it can be noted that this is a relatively new research domain compared to speech recognition. Several works are focused to propose different kinds of feature extraction methods in order to represent the signal in detection system. For example, the works [7] and [8] deal with MPEG-7 low level descriptors and zero crossing rates. For AED are also used other popular speech acoustic analysis methods such as Mel-Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC) or Perceptual Linear Prediction (PLP). In our work, we have applied the PCA to extract features in AED system. We have classified two different acoustic events and the background.

3. Principal Component Analysis

PCA is a popular data processing and dimension reduction method applied also in feature extraction in speech recognition. PCA maps the n-dimensional input data to m-dimensional, $m<n$, with respect to their variability. The method is based on the assumption that most information about classes is contained in the directions, along which the variations are the largest [6], but there is no guarantee that variability explained by PCA

is useful for speech recognition [2]. PCA transforms the data by principal components – PCs (uncorrelated and ordered variables). Usually, using the first few PCs can be represented about 80% variability of the basic vectors.

According to [9] suppose that we have M feature vectors $x_1,x_2,...,x_M$ corresponding to speech signals in the training set. The input data have to be centered before the PCA processing:

$$\Phi_i = x_i - \bar{x} = x_i - \frac{1}{M}\sum_{i=1}^{M} x_i, \quad (1)$$

where Φ_i is the i-th centered vector and $\bar{x}$ is the mean. From these vectors are then created the centered data matrix $A = [\Phi_1 \Phi_2 \ldots \Phi_M]$. Principal components can be given by K leading eigenvectors of the global covariance matrix C resulting from its eigendecomposition:

$$Cu_i = \lambda_i u_i, \quad i \in 1,...,N, \quad (2)$$

where $u_1,u_2,\ldots,u_N$ are the eigenvectors and $\lambda_i, i \in \langle 1;N \rangle$ are the eigenvalues of the covariance matrix:

$$C = \frac{1}{M-1}\sum_{n=1}^{M} \Phi_n \Phi_n^T = \frac{1}{M-1}\sum_{n=1}^{M} (x_i - \bar{x})(x_i - \bar{x})^T. \quad (3)$$

The dimensionality reduction step is performed by keeping only the eigenvectors corresponding to the K largest eigenvalues ($K<N$) and put them into matrix $U_K=[u_1u_2...u_K]$, where $\lambda_1 > \lambda_2 > \ldots > \lambda_K$. Finally, the linear transformation $\mathbf{R}_N \rightarrow \mathbf{R}_K$ is computed as:

$$y_i = U_K^T x_i, \quad (4)$$

where y_i represents the transformed vector and U_K is the reduced rank PCA transformation matrix. The determination of K can be done via comparative criterion with threshold $T \in \langle 0.9;0.95 \rangle$

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{N} \lambda_i} > T. \quad (5)$$

4. TIMIT recognition task using PCA

As was mentioned before, in our experiments the TIMIT speech database [3] we used to evaluate the ASR system. This database is divided into train and test set, with 4620 and 1680 recordings, respectively.

There are 61 distinct phones, which, for evaluation purposes, were reduced to an inventory of 40 symbols using a phoneme mapping structure according to [10]. All closure phonemes (‘bcl‘, ‘dcl‘, ‘gcl‘, ‘kcl‘, ‘pcl‘, ‘tcl‘) were merged in the training phase with previous phoneme segment.

4.1 Feature extraction and HMM topology

The feature extraction for the models was performed in common way, which is used as standard in MFCC-based feature extraction. The input speech signal is preemphasized and windowed using Hamming window. The window size was 25*ms* and the step size was 10*ms*. Short time spectral analysis by fast Fourier transform was applied to the windowed segments. In the next step, the Mel-filterbank analysis was applied to the segments followed by logarithm application to the linear filter outputs, which resulted in LMFE features. These ones are used in experiments described in Sections 4.2 and 4.5. In Section 4.4 we used the classical MFCC features with different dimensions, which were obtained applying the discrete cosine transform (DCT) to LMFE (Logarithm Mel-Filter Energy) vectors. A specific feature combination was used in experiment described in Section 4.4.

The parameters of the three-state HMMs for 40 phonemes were trained. 1-256 Gaussian mixtures were used to model the HMM states. In order to test the models a backoff bigram language model was built from whole database [11]. The feature extraction, acoustic models training and testing by HTK (Hidden Markov Model Toolkit) were carried out. All PCA processing in MATLAB environment was done.

4.2 PCA-based LMFE decorrelation

In this section we compare two different decorrelation techniques used in HMM-based acoustic modeling. The first one is the Discrete Cosine Transformation (DCT) used in conventional MFCC parametrization (application to logarithmic Mel-filterbank outputs). The second one is the mentioned PCA, which was applied instead of DCT in the following manner.

Each speech signal in the train corpus is represented after the parametrization process by one separate data matrix with 26-dimensional LMFE features. By pooling these LMFE data matrices, the training set can be represented by one big data matrix. This matrix was then used as the input for the main PCA according to mathematical description in Section 3. The PCA analysis resulted in an optimal full rank transformation matrix. However, for the final transformation the reduced rank version of transformation matrix was used to transform the train and test set independently. The reduced matrix consisted only of those eigenvectors that corresponded to the largest eigenvalues of the global covariance matrix.

During the statistical PCA analysis the optimal dimension 6 was determined for the input training data. At the training and testing process, the delta and acceleration coefficients were automatically computed by HTK, which finally resulted in 18-dimensional PCA vectors. In order to compare the base and PCA-based models it was necessary to build an MFCC model with the same dimension.

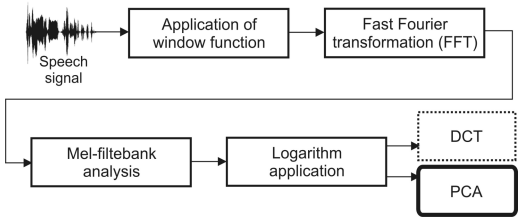


Fig. 1. Block diagram of speech signal preprocessing.

Num. of mixtures	MFCC model	PCA model
128	70.87%	72.12%
256	71.21%	72.47%

Tab. 1. Accuracy levels for baseline MFCC and PCA-based models.

The accuracy of the ASR system for 1-256 Gaussian mixtures was evaluated, but only the cases for 128 and 256 mixtures are listed. The experimental results listed in Tab. 1 show that the accuracies for PCA-based acoustic model are approximately about one percent higher than the accuracies for the baseline model at the same dimension. Especially, the best results have been achieved in case of using 128 and 256 Gaussian mixtures. More detailed information about this application can be found in [10]. Notice that the accuracies in the evaluation process were computed as the ratio of number of all word matches to number of reference words.

4.3 Dimension reduction of multi-frames

In addition, we have also applied the PCA to speech multi-frames created by concatenating the basic successive feature vectors resulting in so-called *supervectors*. In this way also the contextual information was taken into account for the transformation. As the basic vectors the MFCC vectors were used with various dimensions. The PCA in this experiment was done in similar way as was mentioned before, but the input data matrix was created from the concatenated feature vectors. We used two lengths of context *C*; *C*=3 and *C*=5.

The experimental results show that in some cases (lower model dimensions) this approach achieved higher accuracies than the baseline MFCC parametrization (approximately 0.5 – 1%). Notice that the shorter context *C*=3 was more effective than *C*=5 (longer contexts are suitable for larger speech databases). The results for 128 mixtures in the Tab. 2 are listed.

Dimension	15	24	30	33	36	39
Acc. [%] MFCC	69.69	73.58	75.44	76.32	76.70	77.05
Acc. [%] PCA, C=3	69.99	73.93	75.56	75.64	76.44	76.56
Acc. [%], PCA, C=5	70.33	74.02	75.48	75.61	76.11	76.49

Tab. 2 Accuracies for MFCC and PCA model with different lengths of context.

4.4 PCA applied to combined features

We wanted to investigate the interaction of PCA and some non-standard feature compositions. We have tried a feature combination of popular basic feature extraction methods to compose a 39-dimensional "MFPLPC" feature vector from 13 MFCC, 13 PLP and 13 LPC coefficients, respectively. These features were considered as the input for PCA. All conditions of experiments are identical with previous ones. The Tab. 3 shows the accuracies for 128 Gaussian mixtures and different model dimensions. The results are also compared to results of standard MFCC models. The MFPLPC-based model achieved similar accuracies, the difference is approximately 0.5 – 1%.

Dim.	27	30	33	36	39	42	45
MFCC	75.04	75.44	76.32	76.70	77.05	77.37	77.50
MFPLPC	74.03	74.65	75.15	75.54	75.91	76.37	76.40

Tab. 3. Accuracies for MFCC and MFPLPC models.

4.5 Partial training of transformation matrix

In case of large training corpus (thousands of recordings and more) it may occur a problem with training of PCA matrix. Especially, the constructing of the big data matrix for PCA is memory intensive. Therefore, we were interested in investigating the influence of the amount of training data for PCA to the transformation performance. Several initial experiments we have done in order to demonstrate the dependency of PCA training and the amount of training data. We found that for optimal learning of transformation matrix is not necessary to use the whole training data, but it is enough to use a part of them. In our experiments we randomly created several partial training databases from the whole one according to Tab. 4. As it can be seen from it, with shrinking amount of training data, the recognition accuracy was increasing. This idea could be refined by some algorithm that would find the right feature vectors according to their statistical properties to form the big data matrix. We say that for optimal training are probably sufficient only the statistically significant data. This approach can radically speed up the PCA training and transformation. The results of experiments are listed in the Tab. 4 for 256 Gaussian mixtures and optimal dimension 5 with dynamic coeffs.

Part of database	*100%*	*80%*	*50%*	*20%*	*10%*	*8%*
Num. of recordings	4620	3696	2310	924	462	370
Accuracy [%]	71.13	71.16	71.06	70.75	70.95	71.12
Part of database	*5%*	*3%*	*2%*	*1%*	*0.5%*	*0.1%*
Num. of recordings	231	139	92	46	23	5
Accuracy [%]	71.33	71.01	71.45	71.44	71.25	72.42

Tab. 4. Accuracies for partial training of PCA transformation matrix.

5. PCA-based acoustic events detection

In this experiment, we supplemented the basic feature extraction process by PCA. The AED system is proposed for two types of audio events – gun shot and breaking glass. The LMFE features were computed by HTK on 20*ms* windows in similar was described in Section 4.1. These ones from all sounds into one big data matrix were arranged according to Section 4.2. This big matrix was considered as the input matrix for PCA (matrix *A* labeled in Section 3). During the PCA analysis the optimal dimensions according to criterion (5) for each sound class were determined independently. The train and test sets were transformed using optimal reduced rank transformation matrix. For the classes *glass, shot* and *background* the optimal dimensions 3, 4 and 7 were determined, respectively. Thus, each class was represented by 3, 4 and 7 dimensional PCA vectors. The PCA vectors at the training and recognition were extended to 9, 12 and 21 dimensions with the first and second order derivatives. The train corpus consisted of 187 sounds (28, 31 and 128 for glass, shots and background, respectively) and the test set consisted of 46 sounds (5, 10 and 31 for glass, shots and background, respectively). The transformed sets were used to train the acoustic models based on three states HMMs modeled with 1-256 Gaussians. The averaged accuracies for optimal dimensions of each class were obtained from confusion matrix and are listed in Tab. 5.

Average Acc. [%]	Dim. 9	Dim. 12	Dim. 21
Glass	100	99,26	94,81
Shot	90,37	94,81	96,3
Background	84,71	86,74	88,89

Tab. 5. Average accuracies for sound classes at optimal dimensions.

For the class *glass* and *background* the highest average accuracies were achieved, as was expected. The class *shot* did not follow this tendency. Its optimal dimension was 12 but the highest accuracy for dimension 21 was achieved. Detailed information about this experiment can be found in [12].

6. Conclusions and future work

PCA applied to different feature kinds achieved higher accuracies compared to baseline models (in our cases for lower dimensions), but PCA applied to combined features did not. In AED, PCA also works well, when a separate acoustic model for each class was created. Our intention in the future is to refine the present PCA method and to combine PCA with LDA to achieve better efficiency. In case of partial training of PCA matrix we would like to develop an algorithm, which could find the statistically significant data (features) from the training database that will be used to learn the PCA matrix.

Acknowledgements

This work is the result of the project implementation: Development of the Center of Information and Communication Technologies for Knowledge Systems (ITMS project code: 26220120030) supported by the Research & Development Operational Program funded by the ERDF.

References

[1] HAEB-UMBACH, R., NEY, H. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Proceedings of IEEE ICASSP*, San Francisco, CA, 1992, Mar 23-26, p. 13-16.

[2] ABBASIAN, H., NASERSHARIF, B. A. Class-Dependent PCA Optimization Using Genetic Programming for Robust MFCC Extraction. *In: 3rd Conference on Information and Knowledge technology (IKT)*, 2007, Mashhad, Iran.

[3] GAROFOLO, J., LAMEL, L., FISHER, W., FISCUS, J., PALLETT, D., DAHLGREN, N. The DARPA TIMIT acoustic-phonetic continuous speech corpus cd-rom. Technical report, 1993.

[4] ERRITY, A., McKENNA, J., KIRKPATRICK, B. Manifold Learning-Based Feature Transformation for Phone Classification. *LNCS*, 2007, Volume 4885, Springer.

[5] SOMERVUO, P. Experiments With Linear And Nonlinear Feature Transformations In HMM Based Phone Recognition. *In Proceedings of ICASSP*, 2003, p. 52-55.

[6] WANG, X., O'SHAUGHNESSY, D. Improving the efficiency of automatic speech recognition by feature transformation and dimensionality reduction. *In: Eurospeech*, September 1-4, 2003, Geneva, p. 1025-1028.

[7] MUHAMMAD, G., ALGHATHBAR, K. Environment recognition from audio using MPEG-7 features. In *Proc. EN-Com,* 2009, ISBN: 978-1-4244-4995-8.

[8] NTALAMPIRAS, S., POTAMITIS, I., FAKOTAKIS, N. On acoustic surveillance of hazardous situations. *In Proc. ICASSP*, 2009, p. 165-168.

[9] BEBIS, G. Principal Component Analysis. Online: www.cse.unr.edu/~bebis/MathMethods/PCA/lecture.pdf

[10] VISZLAY, P., PLEVA, M., JUHÁR, J. Comparison of different feature decorrelation techniques used in HMM-based acoustic models. *In: Digital technologies 2010: 7th International workshop on digital technologies, circuits, system and signal processing*, 2010, Žilina, p. 1-4.

[11] STAŠ, J., HLÁDEK, D., PLEVA, M., JUHÁR, J. Slovak Language Model from Internet Text Data. *In A. Esposito et al. (Eds.): Towards Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues*, LNCS 6456, 2011, Springer-Verlag, p. 352—358.

[12] VOZÁRIKOVÁ, E. et al. Acoustic events detection via PCA-based feature extraction. *In: Journal of Computer Science and Control Systems*, 2010, Vol. 3, p. 99 – 102.

Unequal Error Control for Image with ROI

Tomáš HIRNER[1], *Jaroslav POLEC*[1]

[1] Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
tomas.hirner@gmail.com, polec@ktl.elf.stuba.sk

Abstract. *A new method of unequal error control for image with ROI is described. Region of interest (ROI) technique is important in applications where certain parts of an image are of a higher importance than the rest of the image. Image with ROI is realized by using EZW and SPIHT methods. Unequal coding of image is provided by ROI and subband image coding (SBC). In this paper, coding techniques are proposed allowing unequal error control of individual bands according to certain rules. In our proposed method, image is divided into two parts, ROI and background (BG). Transmitted bytes are divided into two groups, more important bytes (MIB) and less important bytes (LIB). The result of our proposal is to demonstrate comparison of image transmission with ROI technique and without ROI technique over multiple noisy channels.*

Keywords

Region of Interest (ROI), unequal error control (UEC), subband coding (SBC), most important bytes (MIB), less important bytes (LIB).

1. Introduction

With the growth of multimedia application and the spread of Internet, the access of digital image becomes effortless. Hence, the image content-based retrieval is essential for digital image libraries and databases. Image compression has become necessary in storage and transmission application. The objective of image compression is to decrease the bit rate for transmission or storage while maintaining an acceptable fidelity or image quality. Some examples of image compression techniques are embedded zerotree wavelet (EZW) and set partitioning in hierarchical trees (SPIHT) [8].

Recently, much attention has been paid to the ROI coding since the functionality of ROI is suitable for many applications in which certain parts of an image are more meaningful than the other parts of the image [1, 2, 3].

In subband coding of images the image frequency band is split up into subbands after which each subband is encoded separately using a coder and bit rate accurately matched to the statistics of that particular band [4, 5].

The areas of ROI and SBC have already been described in a number of sources. Some of new ROI coding methods were described by Liu and Fan in [1] and by Wang and Bovik in [2]. In detail SBC of images was described by Westerink [4]. Description of unequal error protection codes for image transmission was presented by Le and Liyana-Pathirana [5]. Some of basic information about SBC was described by Mihalík and Smith with Eddins [8, 9].

This article presents new method of unequal error control for image with ROI. The main idea of our method is UEC image information for transmission over discrete channel with noise by using ROI. We will analyze comparison of image transmission with ROI technique and without ROI technique over noisy channel by using UEC.

The paper is organized as follows. First, basic information about UEC is described in Section II. Basis of ROI is referred to in Section III. Some information about SBC will be mentioned briefly in Section IV. In Section V we will in detail describe our method of unequal error control for image with ROI. Then we will show some important results of our method in Section VI. Finally, some concluding remarks will be given in Section VII.

2. Unequal error control

Masnick and Wolf first introduced the concept of UEC codes in 1969. Their approach influenced different techniques of protection of codeword symbols, restricting the known facts to systematic codes. The structure of codes with UEC differs fairly from the ordinary code. In the case of UEC the bits of the code words are protected in order of importance. The necessity for UEC arises in applications where the transmitted data is a coded signal such as speech, audio, image or video [5, 7].

3. Region of Interest (ROI)

Region of interest (ROI) coding is important in applications where certain parts of an image are of a higher importance than the rest of the image. In these cases the ROI is encoded with higher quality than the background. Example applications include client/server applications or face images [1, 2, 3].

4. Subband coding (SBC)

Subband representation of image [8, 9] the image frequency band is split up into subbands [8]. Subband image coding has been shown to be an effective technique for high quality coding at low bit rates. The subband coding system may be viewed as having two basic components: the subband analysis/synthesis subsystem, which is composed of filter banks, and the coding subsystem, which may employ some form of differential or vector quantization. Both components must be carefully designed [4, 9].

A four-band coding is illustrated in Figure 1, where band splitting is carried out alternately in the horizontal and vertical directions. In the figure, *L* and *H* represent the low pass and high pass filters with a 2:1 down – sampling, respectively [6].

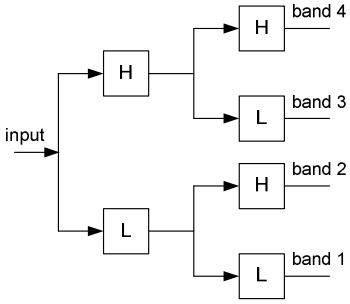


Fig. 1. Block diagram of four-band coding [6]

5. Description of method

The idea of our method is in UEC image information for transmission over discrete channel with noise by using ROI. In this method image is divided into two parts: region of interest (ROI) and background (BG).

In our analysis we will focus on the transmission of region of interest (ROI). Background of image (BG) in this analysis is not important. Image compression is realized by using EZW and SPIHT methods.

In this method, the first, we will choose suitable ROI of image. Subsequently compression techniques EZW and SPIHT are implemented on the image. These compression techniques will be implemented in the ROI and BG of image. In the next step we will decompose the image into individual bands (subimages) using subband coding (SBC).

In this step of the method is to need implement various error controls for each subband (subimages) of the transmitted images with defines ROI. In our experiment, the allocation of bits is used on the basis of variances of the coefficients from individual subimages. We will use four-band decomposition within SBC and RS codes for protection of information bytes. The Reed-Solomon codes utilized here are block based error correcting codes and are widely used for channel coding. The RS codes correct the symbol (or byte) error and not the bit error; lengths in terms of symbols. In method individual information bytes are divided into two importance groups. Let

- *MIB* – More Important Bytes,
- *LIB* – Less Important Bytes.

The block diagram of the proposed system is shown in Figure 2.

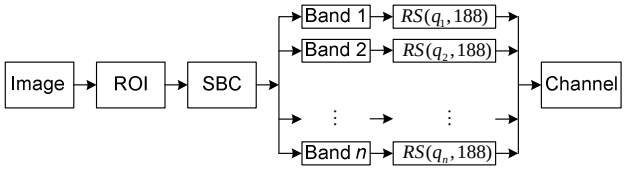


Fig. 2. Block diagram of proposed system

Our results we will compare with method of unequal error control by SBC without using ROI in image.

6. Experimental results

In this section we will examine at what level the symbol error rate (*SER*) can still transmit images through a channel with noise, but the basic result of the analysis will be compare *PSNR* values for method with using ROI technique a without using ROI technique.

The obtained results are shown in a chart, depending of *PSNR* (Peak Signal – to – Noise Ratio) on *SER* (Symbol Error Rate). Symbol is equal 8 bits in this case. Following RS codes were selected for our method:

- C_1: RS (216, 188) correcting maximum 14 errors,
- C_2: RS (196, 188) correcting maximum 4 errors.

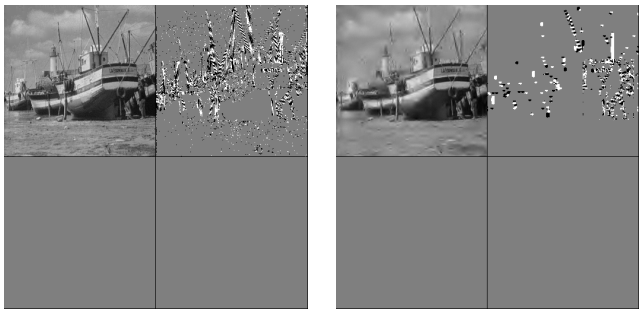
The number of information bytes chosen is 188, but for the image transfer we will use 187 bytes. One byte is intended for synchronization.

We will analyze an image “Boat” of size 512 x 512 pixels. Image with a defined ROI obtained by means of two compressions techniques EZW and SPIHT. Technique EZW is set with the parameter Maxloop on value of 12, and technique SPIHT with Maxloop equals 10. Maxloop means the maximum number of steps for the compression algorithm.

Image with draft of ROI is shown in Figure 3. Obtained images (with ROI and without ROI) from individual subbands using SBC is shown in Figure 4.

As shown in Figure 4, for both cases (No ROI, ROI), only the first and second band (subimage) is transmitted through the channel for particular values of bit per pixel (bpp). This means that in this case, we will only deal with those bands. The last two bands of decomposition are uninteresting for us in this particular analysis. The bytes in the first band will be considered as *MIB* and bytes in the second band will be considered as *LIB* (Fig. 5b).

Fig. 3. Image "Boat" with draft of ROI

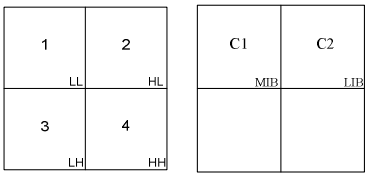


a) b)

Fig. 4. Images "Boat" obtained from individual subbands a) method without using ROI (0,74 bpp), b) method with using ROI (0,75 bpp)

In the first band, for both cases (No ROI, ROI), we will implement Code 1. In the second, we will implement Code 2 (Fig. 5b). Bit rates for ROI image and no ROI image are shown in Table 1.

Table 2 shows the numerical values of *SER* for Codes $C_1 - C_2$. Table 3 shows the numerical values of *PSNR* for individual decoded images. Experimental results for calculation of numbers of bytes and bits per pixels are shown in Table 4. Figure 6 represent original image "Boat" and decoded images corresponding to various *SER*.



a) b)

Fig. 5. a) Four-band split, b) Error protection of images related to the individual bands (subimages) using different RS codes

Subband	No ROI	ROI
	Bitrates	
1	1,81	1,92
2	1,14	1,06
3	0	0
4	0	0

Tab. 1. Calculation of bitrates (boat 512 x 512) for four-band coding without ROI (0,74 bpp) and with ROI (0,75 bpp)

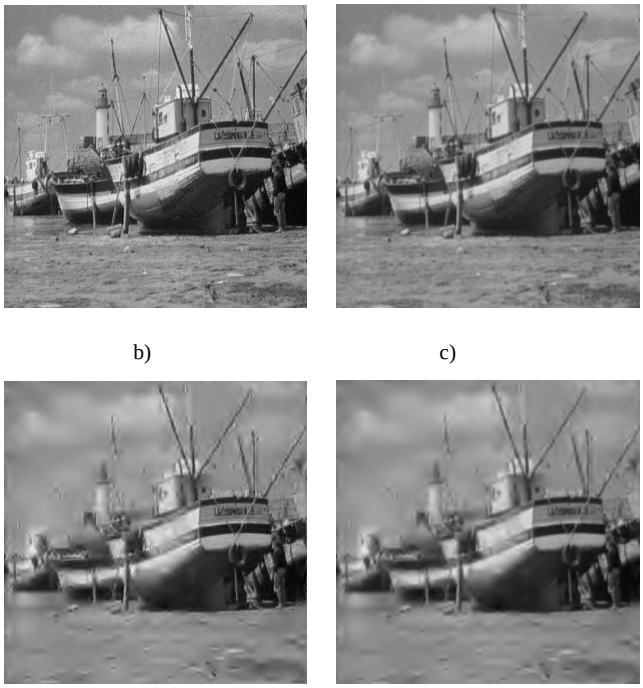
Codes	*SER*
C_1 – RS (216, 188)	0,06482
C_2 – RS (196, 188)	0,02041

Tab. 2. Calculation of *SER* for various coding techniques

Decoded images	*PSNR* [dB]
Fig. 6/b	30,50
Fig. 6/c	22,92
Fig. 6/d	36,57
Fig. 6/e	27,13

Tab. 3. Calculation of *PSNR* for various decoded images

a)



b) c)

d) e)

Fig. 6. Results on the test image "Boat" 512 x 512 for four band decomposition: a) original image, b) image after decoding without using ROI (0,74 bpp, *PSNR*=30,50dB), c) image after decoding without using ROI to ensure *MIB* using RS(216,188) (0,82 bpp, *PSNR*=22,92dB), d) image after decoding with using ROI (0,75 bpp, *PSNR*=36,57dB), e) image after decoding with using ROI to ensure *MIB* using RS(216,188) (0,83 bpp, *PSNR*=27,13dB)

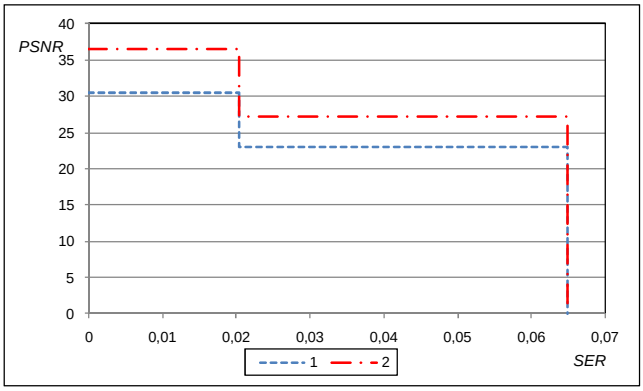


Fig. 7. Dependence of *PSNR* on *SER* for image "Boat": 1 – No ROI (Fig. 6/b, c), 2 – ROI (Fig. 6/d, e). Image after decoded where is implemented RS (216,188) code for subimage which contain *MIB* and RS (196,188) code for subimage which contain *LIB*. For $SER \leq 0{,}02041$, it can transmission entire reconstruction image (Fig. 6/b or Fig. 6/d). For $0{,}02041 < SER \leq 0{,}06482$, it can transmission only subimage which contain *MIB* (Fig. 6/c or Fig. 6/e).

No ROI	**26 914,88 B**	**0,8214 bpp**
ROI	**27 270,00 B**	**0,8322 bpp**

Tab. 4. Experimental results

In Figure 7, RS (216,188) code is implemented for subimages which contains *MIB* and RS (196,188) code for subimages which contains *LIB*. Curve 1 shows image transmission without using ROI technique. Entire reconstruction image (both subimages) can be transmitted if values of *PSNR* and *SER* are: $PSNR = 30{,}5\ dB$ and $SER \leq 0{,}02041$. Subimage which contain *MIB* can be transmitted for $PSNR = 22{,}92\ dB$ and $0{,}02041 < SER \leq 0{,}06481$. In the case of curve 2, image with using ROI can be transmitted provided that $PSNR = 36{,}57\ dB$ for $SER \leq 0{,}02041$ and $PSNR = 27{,}13\ dB$ for $0{,}02041 < SER \leq 0{,}06481$.

From Figure 7 we can see that, after protecting *MIB* with the strongest RS code, we have the ability to transfer images with these features in multiple noisy channels. Subimages (representing individual bands) with a higher importance of bytes will be transmitted even when the channel error rate is higher. For ROI analysis, we were dealing with only some part of image, defined as region of interest. After decoding, the main information is in the ROI and, BG is uninteresting for us. In our proposed method of unequal error control for image with ROI in combination with SBC, transmitted image achieve better results of *PSNR* for whole range of *SER* with comparison with transmitted image without using ROI.

That leads to the important conclusion that in case of transmission of the image information with some more important part for us, it is advantageous to use UEC for individual bands of image by using ROI and SBC.

7. Conclusion

In this article we tried to describe briefly and concisely a new method of unequal error control for image with ROI. Main idea of our method is UEC image information for transmission over discrete channel with noise by using ROI. We were analyzing an image "Boat" of size 512 x 512 pixels. In our analysis we were focus on the transmission of ROI in image. ROI and BG are distinguished by different compression techniques. Subsequently we were decomposing the image into individual bands (subimages) by SBC. For protection individual subimages with *MIB* and *LIB* we were using RS codes.

Our results are compared with method of unequal error control by SBC without using ROI in image. In case of transmission of the image information with some more important part, it is advantageous to use unequal error control for individual bands of image by using ROI and SBC.

Acknowledgements

Research described in the paper was financially supported by the Slovak Research Grant Agency VEGA under grant No. 1/0602/11.

References

[1] LIU, L. AND FAN, G. A New JPEG 2000 Region-of-Interest Image Coding Method: Partial Significant Bitplanes Shift. In*: IEEE Signal Processing Letters*, 2003, vol. 10, no. 2, p. 35-38.

[2] WANG, Z. AND BOVIC, A. C. Bitplane-by-Bitplane Shift (BbBShift) - A Suggestion for JPEG 2000 Region of Interest Coding. In: *IEEE Signal Processing Letters*, 2002, vol. 9, no. 5, p. 160-162.

[3] IDRIS, F. AND ATEF, F. An Efficient Method for Region of Interest Coding in JPEG 2000. In: *5th WSEAS International Conference on Signal Processing*, 2006, p. 65-69.

[4] WESTERINK, P. H. *Subband coding of images*, PhD thesis. Technische Universiteit Delft, 1989.

[5] LE, M. H. AND LIYANA-PATHIRANA, R. Unequal error protection codes for image transmission over noisy channels. In *First International Conference on Information Technology and Applications 2002 (ICITA 2002); IEEE Image Processing*, 2002.

[6] GHANBARI, M. *Video coding – an introduction to standard codecs*. The Institution of Electrical Engineers. London, United Kingdom, 1999.

[7] MASNICK, B. AND WOLF, J. On linear unequal error protection codes. In *IEEE Trans. Inform. Theory*, 1967, vol. 13, p. 600 – 607.

[8] MIHALÍK, J. *Picture encoding in video communications*. Merkury – Smékal. Košice, 2001.

[9] SMITH, M. J. T. AND EDDINS, S. L. Analysis/synthesis techniques for subband image coding. In *IEEE Trans. Acoust., Speech, and Signal Proc*. Atlanta, USA, 1990, vol. 38, no. 8, p. 1446 – 1456.

Error Concealment for Shape Transform Coding

Sandra ONDRUŠOVÁ[1]

[1] Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
ondrusova@ktl.elf.stuba.sk

Abstract. *In this paper, spatial shape error concealment techniques to be used for object-based image in error-prone environments are proposed. It is assumed that the shape of the corrupted object is in the form of a binary alpha plane. Some of the shape data is missing due to the channels error. We consider a geometric shape representation consisting of the object boundary, which can be extracted from the α-plane. Three different approaches are used to replace a missing boundary segment: Bezier interpolation, Bezier approximation and NURBS approximation. Experimental results on object shape with different concealment difficulty demonstrate the performance of the proposed methods. Comparisons with proposed methods are also presented*

Keywords

Error concealment, Shape coding, Object-based image, NURBS, Bezier curves .

1. Introduction

In wired and wireless networks the transmission of images may lead to loss. As retransmission of lost or damaged packets may incur delay, error resiliency methods have been developed to detect and correct transmission errors. Therefore error detection and concealment is used.

The MPEG-4 object-based audiovisual coding standard [1] opened up the way for new video services, where scenes are understood as a composition of objects; this approach may have advantages in terms of coding efficiency as well as in terms of additional functionalities. However, to make these object-based services available in error-prone environments, such as mobile networks or the Internet, with an acceptable quality, appropriate error concealment techniques dealing with both shape and texture data are necessary.

Depending on the data that is used, the decoder error concealment techniques can be divided in three major categories:

- **Spatial error concealment** – only data from the current time instance is used to perform the concealment. The corrupted areas are recovered by interpolating or approximating the data from the surrounding corrected decoded areas.

- **Temporal error concealment** – data from other time instance is used to perform the concealment. Although the most common approach is to use data from the immediately preceding time instance.

- **Spatio-temporal error concealment** – this approach is a combination of spatial and temporal error concealment. Some parts of the image might be concealed using spatial concealment, others by using temporal concealment and still others by using a little bit of all.

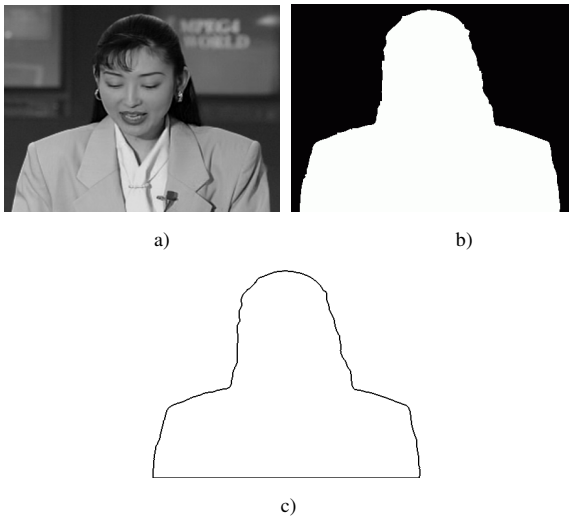


Fig. 1. An Object and its transmitted components: a) original image; b) image α-plane; c) reference contour

In this paper, original spatial shape error concealment techniques will be presented. All techniques assume that the shape of the corrupted object is in the form of binary alpha plane and that some of the shape data is missing due to channel errors. Due to the missing some of the contour segments contour will be broken. The idea is to approximate the missing contours using NURBS. Many different approaches can be found in [2], [3], [4], [5], [6]. We propose a spatial error concealment technique based on a contour representation of the object shape, i.e., the boundary of its texture (Fig. 1).

Error concealment includes the construction of a new curve that successfully replaces the missing boundary parts and joins smoothly with the received parts. The existing

geometric concealment approaches build a polynomial concealment curve based on

2. Spatial Shape Error Concealment Scheme

In object-based coding system, objects are encoded independently, although they together build a scene. In general error concealment in object-based system can be structured in two levels: object level concealment and scene level concealment. The first one deals with error detection and concealment for each single object. Each object uses only its own data and no data from other surrounding objects. Scene level concealment deals with error detection and concealment for the whole scene.

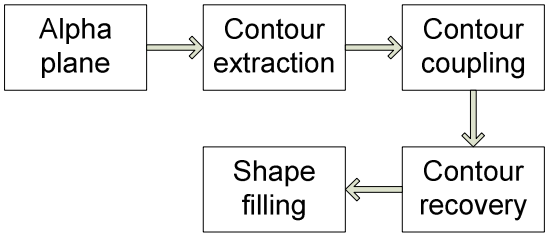


Fig. 2. Proposed shape-concealment process

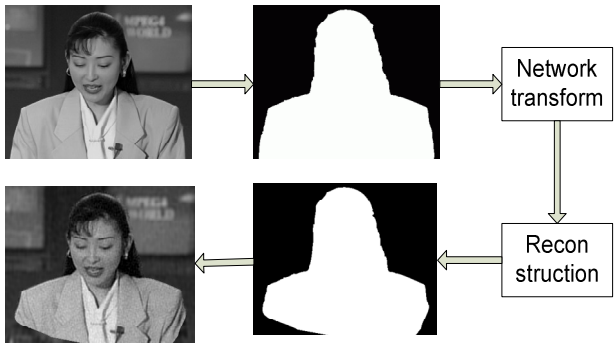


Fig. 3. Proposed schematic process

The technique proposed in this paper is a spatial technique, in the sense that it does not rely on information from other temporal instants. Fig. 2 illustrates block diagram for the shape concealment technique. The input is corrupted alpha plane. Our proposed schematic process is shown in Fig. 3.

The four steps that have to be followed in order to conceal the corrupter alpha plane are [2]:

• Contour extraction – extract the contour from the corrupted shape data is the first step to conceal the lost blocks in the alpha plane. The contour will be broken if some of the lost shape blocks were border blocks. An example of broken contour, extracted from the shape of the Akiyo video object, is shown in Fig. 4 b). If the lost blocks do not correspond to the border blocks, the contour extracted from the corrupted shape is in no way affected.

• Contour recovery - the missing contours are approximated with Bezier curves and NURBS. Bezier and NUBS curves are easier to manipulate and much faster to compute which is especially important when video applications are considered.

For each pair of contour endings, the recovery of the contour inside the missing area with NURBS and Bézier curves is performed by following three steps.

1) Determination of the four points that fully specify the NURBS and Bézier curves relevant for the contour in question.

2) Definition of an analytical continuous expression of the NURBS and Bézier curves from the points determined in the previous step.

3) Finally, computation of a discrete representation of the continuous NURBS and Bézier curves determined in the previous step, in order to obtain the edge representation of the interpolating contour inside the lost area.

• Shape filling - after the lost contour parts have been recovered a new closed contour is obtained. The corresponding shape is then filled with shape level information, i.e. black [7].

3. Performance Evaluation

In order to evaluate the proposed shape-concealment technique, several MPEG-4 bitstreams have been tested, each bitstream containing one video object (according to the Core Visual Object Type [1]) encoded at a given bit rate. The exact bit rate value used is unimportant because it does not influence the quality of the shape since lossless shape coding was used.3 On the other hand, the quality of the texture data is highly dependent on the used bit rate. Since users are very sensitive to errors in the shape data, shape is typically coded in intra mode for every time instant. This way, after the concealment is done, if a given VOP still has some shape artifacts, these will not propagate to the following VOPs.

To evaluate the performance of the proposed technique in terms of shape recovery, a shape quality metric is needed. During the development of the MPEG-4 standard, a shape quality metric was used within MPEG to evaluate the performance of several proposed lossy shape coding techniques [8]. This metric is based on the ratio between the number of shapels that are different in the original and reconstructed alpha planes and the total number of original shapels as shown in equation:

$$D_n = \frac{diff\_shapel}{all\_shapel}$$

Where $diff\_shapel$ is number of different shapels in the original and reconstructed alpha plane and

all_shapel is number of the opaque shapels in the original alpha plane.

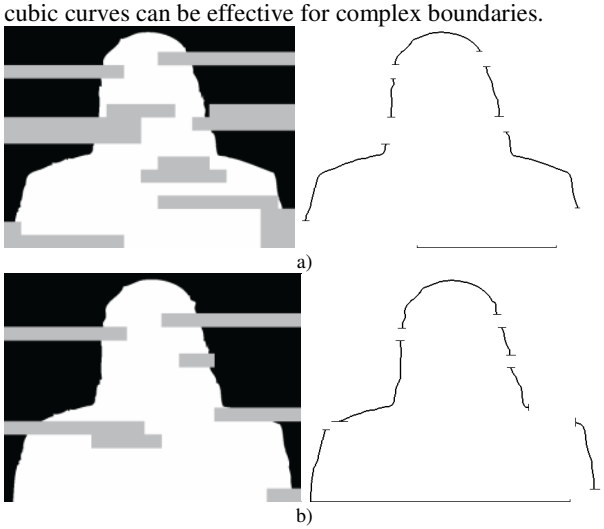
This metric can also be expressed as a percentage:

$$D_n = 100xD_n[\%]$$

4. Results

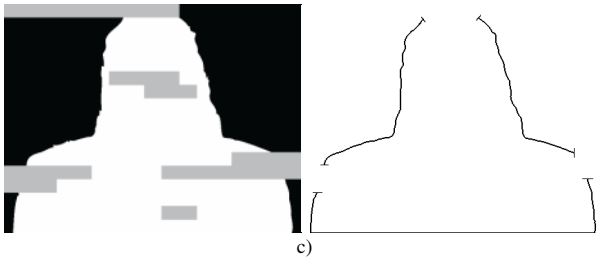
To test the proposed concealment method a number of experiments were performed, some of which are presented here. Examples showing the visual outcome are show besides numerical results. In order to quantify the performance of the proposed concealment method, we will use a relative measure, the ratio D_n of the number of different pixels in the original and reconstructed α-plane divided by the total number of object pixels in the original α-plane. In MPEG-4 is used this quality metric to evaluate shape coding techniques. We will compare our method to the error concealment method proposed by Soares and Pereira [2]. Finally, reconstructed α-planes and restored images will be illustrated for subjective evaluation.

There is one object shape used in our experiments, namely Akiyo. For this boundary we assumed a missing segment consisting of various points and applied the proposed method. After boundary reconstruction corresponding α-plane was extracted. The D_n values associated with every object are shown in Table 1. As can be seen in Table 1, only a small percentage of the reconstructed object pixels differ from the original ones. In most cases, such small differences are hardly visible. Comparing the results in Table I, Bezier interpolation gives the best results. NURBS improvement is greater in the case of smoother boundary. This is explained by the fact that cubic curves can be effective for complex boundaries.



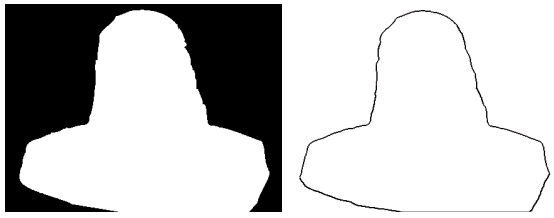
a)

b)

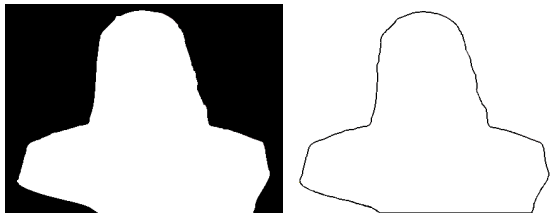


c)

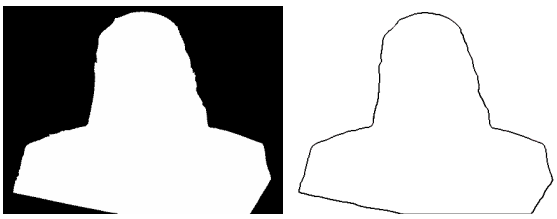
Fig. 4 Examples of corrupted shape of alpha plane and corrupted contour: a) error pattern 1; b) error pattern 2; c) error pattern 3



a)

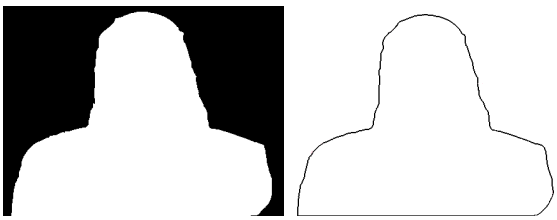


b)

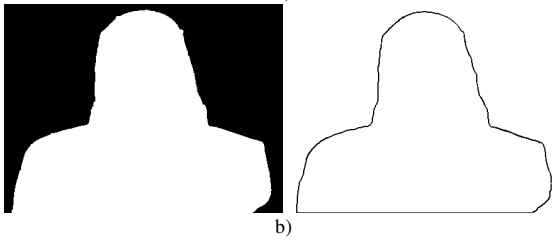


c)

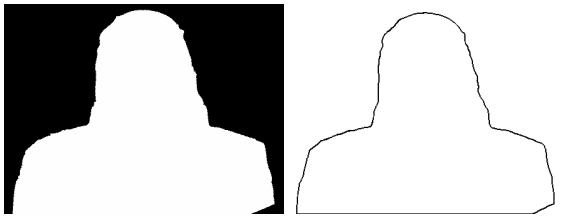
Fig. 5 Examples of reconstructed contour and recovered alpha plane (pattern 1 – Fig 4 a) using; a) NURBS; b) Bezier approximation; c) Bezier interpolation



a)

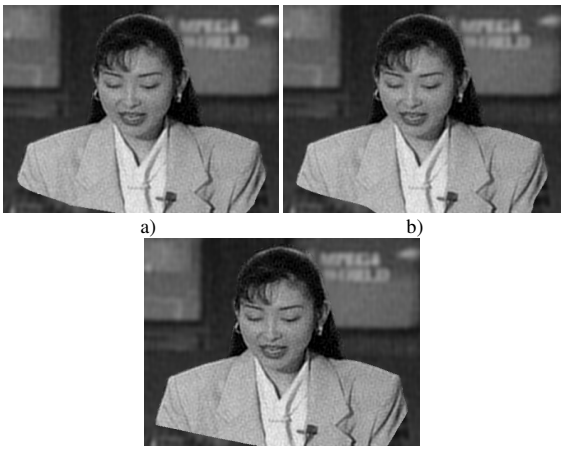


b)



c)

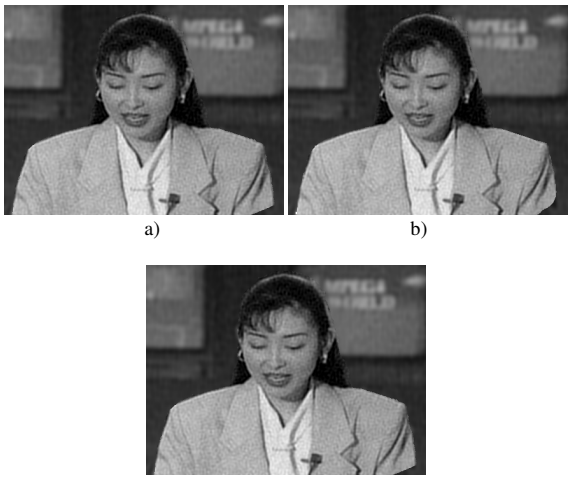
Fig. 6 Examples of reconstructed contour and recovered alpha plane (pattern 2 – Fig. 4 b) using; a) NURBS; b) Bezier approximation; c) Bezier interpolation



a) b)

c)

Fig. 8 Examples of reconstructed decoded VOP [0,4bpp] (pattern 1 – Fig. 4 a) using; a) NURBS; b) Bezier approximation; c) Bezier interpolation



a) b)

c)

Fig. 9 Examples of reconstructed decoded VOP [0,4bpp] (pattern 2 – Fig. 4 b) using; a) NURBS; b) Bezier approximation; c) Bezier interpolation

In Table II comparison using objective criterion is shown. PSNR values are presented in decibels.

	NURBS	App_Bezier	Int_Bezier
pattern 1	3,25	3,27	2,88
pattern 2	0,62	0,72	0,43
pattern 3	0,67	0,69	0,51

Tab. 1. D_n [%] values of for Akiyo

	NURBS	App_Bezier	Int_Bezier
pattern 1	26,77	26,57	27,18
pattern 2	34,07	34,05	34,52
pattern 3	36,05	35,85	36,34

Tab. 2. PSNR values of for Akiyo [0,4bpp]

5. Conclusion

In this paper, a technique was proposed to conceal shape errors in the binary alpha planes. NURBS and Bezier curves are used for reconstruction of corrupted image boundaries. Methods are based on the interpolation and approximation of the received boundary in a way that can represent its complexity level and preserve its direction at the connecting points. The key idea of this study is to use NURBS to represent the portion of image data without corruption. The concealment curve is a cubic B–spline curve having the same direction at the connecting points. Our method leads to better objective and subjective results than the current state of the art.

As follow from experiment, Non-Uniform Rational B-Spline can be used for shape reconstruction and gain very good results.

Acknowledgements

Research described in the paper was financially supported by the Slovak Research Grant Agency: VEGA under grant No. 1/0602/11.

References

[1] ISO/IEC 14496-2, Information Technology – Coding of Audio-Visual Objects, Part 2: Visual, December 1999.

[2] SOARES, L. D., PEREIRA, F. Spatial Shape Error Concealment for Object-based Image and Video Coding, *IEEE Trans. on Image Proc*., Vol. 13, No. 4, pp. 586-599, April 2004.

[3] HRUŠOVSKÝ, B., MOCHÁČ, J., MARCHEVSKÝ, S. Extended Error Concealment Algorithm for Intra-frames in H.264/AVC, *Acta Electrotechnica et Informatica,* Vol. 10, No. 4, 2010, pp.59-63, December 2010

[4] SHIRANI, S., EROL, B., KOSSENTINI, F., A concealment method for shape information in MPEG-4 coded video sequences, *IEEE Trans. Multimedia*, vol. 2, pp. 185–190, Sept. 2000

[5] SOHEL, A.H., GOUR, K.C., LAURENCE, D.S. Image Dependent Spatial Shape Error Concealment for Multiple Shapes. *Signal-Image Technology & Internet-Based Systems (SITIS),* 10.1109/SITIS.2009.36, Marrakesh, December 2009

[6] SCHUSTER, G. M., LI, X., KATSAGGELOS,A.K. Shape Error Concealment Using Hermite Splines, *IEEE Trans. on Image Proc*., Vol. 13, No. 6, pp. 808-820, June 2004

[7] KAUP, A., AAACH, T. Coding of segmented images using shape-independent basis functions, *IEEE Transaction on Image Processing*, Vol.7, No.7, 1998, pp. 937-

Intelligibility of Single-Handed and Double-Handed Finger Alphabets

Petra Heribanová[1], Jaroslav Polec [2], Angela Mordelová[2], Ján Poctavek[2]

[1] Dept. of Algebra, Geometry and Didactics of Mathematics, Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Mlynska dolina,842 48 Bratislava, Slovakia petra.heribanova@fmph.uniba.sk
[2] Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
jaroslav.polec@stuba.sk, poctavek@ktl.elf.stuba.sk

Abstract. *This paper describes the new technique of evaluating the quality of encoded video signals based on logatom recognizability using so-called sign logatomes and the result in single-handed and double-handede finger alphabets. Proposed method was used for video quality measurement in h.264 encoded realtime distributed video of hearing-impaired children and adults.*

Keywords

Cued Speech, Intelligibility, Logatom, Video

1. Introduction

Evolving technologies and advanced processing techniques in TV, internet, or telecommunications raise their standards of image quality and sound. But high quality video also requires considerable volume of data that needs to be transferred (and paid). Therefore, we always try to find the best compromise between acceptable video quality and cost.

Subjective tests show that sound tends to reduce people's ability to recognize video image degradation. Deaf people however are not affected by sound, so their subjective video quality evaluation can differ from hearing people. Actually, the biggest difference of video of cued speech is its purpose - it is the equivalent of sound channel in normal audiovisual recordings. Hearing-impaired people doesn't rely that much on video quality, as the most important thing to them is whether they are able to understand the meaning.

The main difference between the terms quality and intelligibility is that the term "quality" describes the appearance of decoded video signal ("how" the viewer sees it) and the "intelligibility" is just one aspect of quality saying if the received information gives any sense ("what" the viewer sees in it). High-quality video signal is likely to be intelligible. Conversely, of course it may or may not apply. Anyway, unintelligibility is an indicator of poor quality. In the acoustics, intelligibility threshold is defined as a point, after which one does hear, but one does not understand [1].

Our aim is to find criteria for video signal quality encoded in various bit-rates, to achieve full intelligibility of Slovak (or other) cued speech and finger alphabet.

2. Cued Speech and Finger Alphabet

Cued speech is the primary communication tool of hearing impaired or hard of hearing people. It is visual and spatial language with its own grammar and gesture vocabulary. It has visual-motile modality and it is independent of spoken language. But it is not international. It uses three-dimensional space (the gesture space) for communication, which is defined horizontally and vertically. In gesture languages, we have two types of meaning carriers:

- manual = position, shape and movement of hands
- non-manual = facial expression, position of eyes, head, upper body, mouth movement

The basic communication element is gesture. It is given by configuration (shape and placement) of the hands in gesture space, by palm and finger orientation, and also by hand movements themselves. It is quite difficult to learn the gestures from books or static images, because even slight difference in movement and location of the hand can change the meaning. Hence, personal demonstration, or understandable video preview is needed.

Finger alphabet was not created naturally and spontaneously by deaf people. It was adapted from monasteries. It is a system of finger and movement configurations that represent alphabetic characters. The number of characters is related to the number of speech sounds (phonemes) of the language. It is commonly used for purposes of clarification, such as unfamiliar words, names of persons, geographical names, or with words, for whose the asking person doesn't know the appropriate gesture. An advantage of the finger alphabet is that its adoption is not difficult or time-consuming. It helps to express the words in correct grammatical form and thus it is the tool for obtaining a richer vocabulary. In the world,

there are two widely used systems of the finger alphabet [2]:

- Single-handed method (also "finger-spelling") (Fig. 1)
- Double-handed method. (Fig. 2)

Single-handed finger alphabet (dactyl) is used to teach children at schools for students with hearing impairment. It is more widespread in the world. On international meetings, the only used finger-spelling alphabet is the one approved by The World Federation of The Deaf.

Double-handed finger alphabet tends to be used by older people, because it is slower. Despite its slowness, it is also used at lectures and seminars because of its better intelligibility and visibility [3].

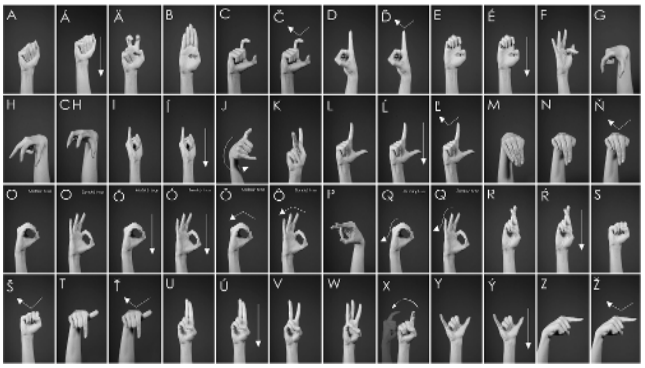


Fig. 1. Example of single-handed finger alphabet.

Fig. 2. Example of double-handed finger alphabet.

3. Subjective and Objective Methods for the Quality and Intellibility

kSubjective evaluations are based on comparing the original and processed video signal by a group of hearing impaired volunteers that evaluate (by their subjective feelings) the quality and intelligibility of the stream, according to a defined scale.

Objectively, intelligibility is measured by statistical methods. In the simplest case, it is the percentage of correctly recognized elements. For sentence intelligibility, recognition is considered successful, when the reproduced sentence has correct context and makes sense. Logatom recognizability is expressed as the percentage of correct consonants and vowels from all speech sounds in transmitted logatomes. Resulting from this, it is clear that logatom based recognizability is much more demanding than sentence or word based one, because the meaning cannot be guessed from the context [1].

4. The Intelligibility (Recognizability)

In acoustics, the intelligibility of the language (Z) defines the percentage of correctly received elements or parts of speech (a) divided by their total number (b):

$$Z = \frac{a}{b} \cdot 100\ \% \tag{1}$$

We distinguish consonant, logatom, word, and sentence based intelligibility. Logatomes are artificial words designed to look alike words of given language, but they do not have the meaning. The term recognizability is used in recognition of speech sounds (phonemes) and logatomes, as one can either recognize or not recognize them, but there is nothing to be understood [4].

Similarly, we can explore the intelligibility of video recordings: sentence and word intelligibility using gestures, while logatom and consonant recognizability using the finger alphabet. One sign in finger alphabet represents one speech sound in logatom. Thereby it is possible to create a sort of "sign logatomes" for the deaf.

4.1 Sentence intelligibility

In [5] there is a new evaluation methodology of video signal quality in transmissions of gesture language in videoconferencing.

4.2 Logatom recognizability

In logatom recognizability evaluation we use artificial monosyllabic words without meaning (logatoms) to mitigate people's tendency to correct the incorrectly understood consonants or words according to the meaning. We create so-called "sign logatoms" – every speech sound in logatom is represented by an appropriate sign from Slovak one-handed or two-handed alphabet. It is a new evaluation methodology of video signal quality in transmissions of gesture language in videoconferencing.

This work shows a new objective method for examining the logatom recognizability (as used in telephonometry for speech sound articulation) with a use of subjective ACR method (full categorical evaluation). This methodology is based on the intelligibility according to variable transmission channel capacity. The aim is to determine video degradation threshold, at which the signs of alphabet (single-handed and double-handed) are still correctly understood, the degree of degradation

of particular alphabet signs and, alternatively, mutual sign exchangeability.

5. Experiment

a

b

Fig. 3. Picture taken from the experiment (cut): a) H.264 decoded frame with parameter QP=40 (640x360); b) H.264 decoded frame with parameter QP=50 (640x360).

Based on this methodology we created the following experiment.

5.1 Single-handed finger alphabet

We produced 2 video previews with seven different logatoms in Slovak single-handed finger alphabet (one with 41 consonants, one with 42 consonants). The length of the video previews is about one minute. For the whole experiment we used four different video formats of 1280x720, 640x360, 320x180 and 160x90 pixels per frame with 25 frames per second. Subsequently, these recordings were encoded by the H.264 codec in various bit rates (QP = 30, 40, 50 that corresponds to rates from 390 kbit/s to 4.5 kbit/s respectively).

12 created samples were shown to 8 elementary school pupils. Testing was realized according to subjective ACR method. A random sequence of consonants is quite hard to remember; therefore some sequences were shown multiple times to the same people (in different bit-rate and/or video format) without mentioning it in advance. Additionally, there was need for another person (as interpreter), because the children were not able to watch the video and write down the meaning at the same time; so they were just showing (in finger alphabet) what did they see and the interpreter person was writing it into the answer sheet. Then the pupils evaluated the subjective video quality according to Tab 1.

The whole test consists of two parts:

1. Subjective, where the video was evaluated according to given voting options shown in Tab 1.

1	*Completely understandable*
2	*Understandable*
3	*Sporadically inapprehensible*
4	*Inapprehensible*

Tab. 1. Proposed voting options for consonant intelligibility testing

2. Objective, where the respondent had to rewrite the consonants organized into logatomes to the letters of the Slovak alphabet. While the sentence intelligibility evaluation was based on subjective rating, the logatom recognizability expresses the correctness of all consonants in logatom in percents.

Resolution	QP	30	40	50
160x90	Objective evaluation [%]	90,24	71,95	0
	Subjective evaluation	2	3	4
	Bitrate [kbit/s]	18,3	7,2	4,5
320x180	Objective evaluation [%]	93,90	76,19	45,23
	Subjective evaluation	2	3	4
	Bitrate [kbit/s]	51,6	17	7,5
640x360	Objective evaluation [%]	96,49	83,30	50,00
	Subjective evaluation	1	2	3
	Bitrate [kbit/s]	149,2	42,1	18,4
1280x720	Objective evaluation [%]	95,23	95,12	93,90
	Subjective evaluation	1	1,5	2
	Bitrate [kbit/s]	389,4	126,9	56,2

Tab. 2. Results from performed experiment

The results of the intelligibility evaluation of single-handed Slovak finger alphabet fulfilled our anticipation, as they are clearly dependent on image quality. Minimal

video transfer speed depends on video format settings. At the bit-rate of 35 kbit/s and video format 640x360 the respondent is able to recognize 50% of consonants, while at the same bit-rate, but in 160x90 the respondent recognizes nearly 90%. With decreasing recognizability there was an increasing number of consonant interchanges, mostly between 'a' and 's', 'o' and 'f', and there was also higher frequency of missed or extra added consonants. Light conditions, camera settings, and background color have big impact on overall intelligibility, as well as on visibility and ability to interpret the signs.

5.2 Double-handed finger alphabet

We produced one video preview with five different logatoms in Slovak double-handed finger alphabet (together 26 consonants). Simillarily as with single-handed alphabet, we performed the test. 12 created samples were shown to 10 adult persons in age from 20 to 33 years. Video Resuts are shown in Tab.3.

a

b

Fig. 14 Picture taken from the experiment (cut): a) H.264 decoded frame with parameter QP=30 (640x360); b) H.264 decoded frame with parameter QP=50 (640x360).

Resolution	QP	30	35	40
160x90	Objective evaluation [%]	97,3	88,5	80,8
	Subjective evaluation	2	2,5	3
	Bitrate [kbit/s]	18,9	11,2	7,1
320x180	Objective evaluation [%]	100	93,8	87,5
	Subjective evaluation	1	1,5	2
	Bitrate [kbit/s]	52,7	26.6	17,8
640x360	Objective evaluation [%]	100	100	100
	Subjective evaluation	1	1	1
	Bitrate [kbit/s]	151,3	66.4	42,1

Tab. 3. Results from performed experiment

6. Conclusion

This paper describes the new technique of evaluating the quality of video signals based on logatom recognizability using so-called sign logatomes and the result in single-handed and double-handede finger alphabets.

In our next work we will further investigate the methodology of evaluating the quality of video signals based on selected augmentative and alternative communication methods.

Acknowledgements

Research described in the paper was financially supported by the Slovak Research Grant Agencies: KEGA under grant No. 119-005TVU-4/2010.

References

[1] GRANAT, M., Objective methods for evaluation of audio signal quality (in Slovak), Brno: Brno University of Technology, 2009.

[2] TARCSIOVÁ, D., Pedagogics of hearing-impaired (in Slovak), Bratislava: MABAG spol. s r. o., 2008.

[3] HEFTY, M., Finger alphabet (in Slovak), In *Organization Myslím – development of thinking not only for hearing-impaired (in Slovak),* 2009, www.zzz.sk

[4] MAKÁŇ, F., Elektroacoustics (in Slovak), Bratislava: Vydavateľstvo STU, 1995.

[5] POLEC,J., ONDRUŠOVÁ, S., MORDELOVÁ,A., FILANOVÁ,J. , "New Objective Method of Evaluation Cued Speech Recognition in Videoconferences," In *Proceedings Redžúr 2010: 4th International Workshop on Speech and Signal Processing.* Bratislava (Slovak Republic), 2010. - Bratislava: STU v Bratislave FEI, 2010.

Speech Recognition

Matúš ŠTRBÁŇ[1]

[1] Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
matus.strban@gmail.com

Abstract. *Speech recognition can be realized in many ways. This article describes a basic training process of hidden Markov models and a validation procedure as is realized via Masper procedure that is based on HTK (Hidden Markov Toolkit) system. Evaluation is performed on a set of application words for Slovak language.*

Keywords

Speech Recognition, HMMs, MFCC, Masper.

Introduction

Speech recognition is process of automatic recognition of said speech. Hidden Markov models (HMMs) have proved to be an appropriate solution for complex systems that are able to operate with large vocabulary and complex grammar, and also to work independently from speaker. Therefore, much attention is concentrated on research in HMMs. New solutions are been found to improve efficiency, increase success and reduce the computational time of recognition process. First of all must be carried training process that sets HMM parameters which represent recognized speech units, in this case phonemes, using databases containing training speech recordings. After it, the recognition process may start. The output of training process determines the success of the recognition rate in specific application.

1.Principle of HMMs

Speech recognition systems generally assume that the speech signal is a realization of some message encoded as a sequence of one or more symbols [1]. Source of speech signal is a state model, which is in each moment located in specific state. In other moment it can remain in this state or move to another state.

Human speech is characteristic by its acoustic structure, linguistic structure and exhibition of speaker's personality. Fundamental attribute of sound, thus speech as well, is the intensity (loudness), pitch and color. Basic tone of human voice is characterized by the frequency vibration of vocal chords. Adult male has this frequency between 90 and 150 Hz, female from 130 to 300 Hz, and children over 300 Hz. When the voice is traveling in the vocal apparatuses, there are resonances in the oral, nasal and pharynx cavities. These resonances intensify some parts of sound spectrum and thus produce formants. While the basic frequency indicates the pitch formants make resultant acoustic feeling.

Imagine that we have a finite number of sounds generated by the vocal tract, for example, individual phonemes always pronounced in the same way. The speech signal consists of stocks, representing sounds (phonemes), their time limits are unknown. Therefore we divide the signal into short segments, so that each of them will be generating only one sound (the short section of the order of tens of milliseconds assume stationary speech signal). The condition is that each segment is modelled by one state. Therefore, we have a sequence of states (which we do not see) and the sequence of features (sounds) that we observe. Now, we have a bigger number of different fetures, while we are not sure, which feature is generated by which state. We know only probability density $b_i(o_j)$ which means generation of the speech vectors o_j by the particular states of i. Now while knowing observed sequence of speech vectors we can identify only probability of particular sequence of states. Let´s imagine now that we have higher number of such models (different words). The task of speech recognition is to identify which model has generated the observed sequence of speech vectors with the highest probability – which word was said. This principle represents the simplest case – the recognition of isolated words simulated by models of whole words.

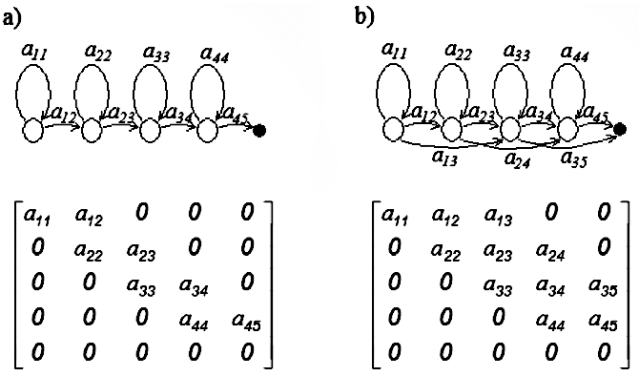


Fig. 1 Examples of left/right models with 4 emitting states and non-emitting exit state

1.1 Mathematical Representation

Probabilities of transitions are expressed in the transition matrix $\underline{A}=[a_{ij}]$ for $1 \le i \le N$, $1 \le j \le N$, where a_{ij}

reflects the probability of transition from state *i* to state *j*. For modelling processes arising at the time are used left-right models, where each of the state can only go into a state with a higher or the same index. The simplest and most common model with possibility of transition only into the next state, but models are also used to skipping one or even several states as indicated in the figures below:

The probability that model at the beginning will be in a particular state *i* reflects the initial probability π_i. For the whole model is then the vector $\pi=[\pi_i]$ for $1 \le i \le N$. The most commonly used simple example where $\pi_1=1$, $\pi_i=0$ for $2 \le i \le N$. Instead of using the vector π, the one non-emitting state is often added to the beginning of the model. Vector of initial probabilities thus becomes the first line of matrix $\underline{A}$ (expect a_{11} which is equal to 0). In addition to the transition matrix and vector of initial probabilities, the hidden Markov model is expressed also by probabilities of densities. I've dealt with the tied-mixtures models.

1.2 Tied-mixtures models

After the process of parameterisation of the speech signal from one time segment we get the one vector of parameters. Each of these parameters can acquire a continuous range of rates. The tied-mixture model is then determined by the parameters $\underline{A}$, π, b. $\underline{A}$ represents transition matrix, π represents vector of initial probabilities. $b = [bi(o)]$ for $1 \le i < N$ represents vector of probabilities of densities for particular states *i*. By introducing the probability density function as the sum of Gaussian mixtures we got expressions of tied-mixture hidden Markov model using a set of parameters $\underline{A}$, π, $\underline{C}$, $\underline{\mu}$, $\underline{U}$, where $\underline{\mu}$ is the mean matrix and $\underline{U}$ is the covariance matrix. Probabilities are expressed as multi-dimensional Gaussian mixtures, because of the diversity of the human voice (women, men and children).

1.3 Compute the probability of model

For calculation is used *forward-backward algorithm* which consists of the forward probabilities $\alpha_j(t)$ and backward probabilities $\beta_j(t)$. $\alpha_j(t)$ reflects the probability that at time *t*, the model is in state *i* and until time *i* was generated sequence of signs $\{o_1, o_2, \dots, o_t\}$. $\beta_j(t)$ reflects probability that from time t +1 was generated sequence $\{o_{t+1}, o_{t+2}, ..., o_T\}$ if the model at time *t* is in state *i*. This way we can determine the probability of generating the observed sequence with given model, and thus determine the model for which this probability is maximal.

1.4 Training and testing HMMs

Training process involves efforts to find the λ model parameters, to maximize the probability $P(O \mid \lambda)$, that model representing a certain speech unit (eg word) generates a sequence of O, which was created by parameterisation of said unit (word). I reached this by *Baum-Welch algorithm* which is based on "maximum likelihood" criterion.

Recognition is a process which seeks an optimal sequence of states of the model, that the probability of generating observed sequence is greatest. This task can be reached by Viterbi algorithm. $\delta_i(t)$ is variable which reflects probability (rate) of the most optimal path ending at time *t* in state *i*.

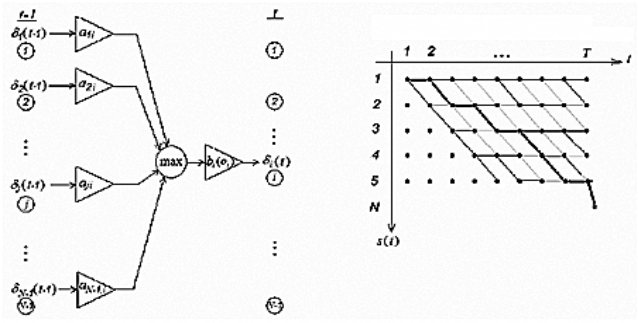


Fig. 2 Illustration of the iterative calculation δi (t) using the Viterbi algorithm (left) and an example of optimal paths in the model with 5 emitting states (right)

2.HTKs and MASPER

All the theoretical knowledge could be tested in the training process Masper by recognition of application words (commands). The system HTK (Hidden Markov modelling toolkit) consists of a large number of elementary modules for self-training and recognition. I have been using train and test scripts for the process Masper prepared for the Slovak language and for database MobilDat-SK which was available for testing. Scripts are written in Perl for Unix operating system. Therefore, I used program Cygwin, Unix emulator under MS Windows. The actual procedure is then summarized in a few steps (not counting the number of database operations, such as creating indexes, change into a readable dictionary for HTK and below).

2.1 Training

Training is realized through a script MonoTrain.pl, without input parameters, which calls necessary subscripts and subroutines in HTK format during the procedure:

2.1.1 Parameterisation

At the beginning of training all audio files from all folders of speech database are parameterised and corresponding sets of parameters (features) are created.

2.1.2 Initialisation and training of monophones

The prototype of model is initialized by the "flat-start" method. Several processes of Baum-Welch Re-Estimation follow until the 32 mixtures model is reached. In this type of training are modelling monophones (models without time limits).

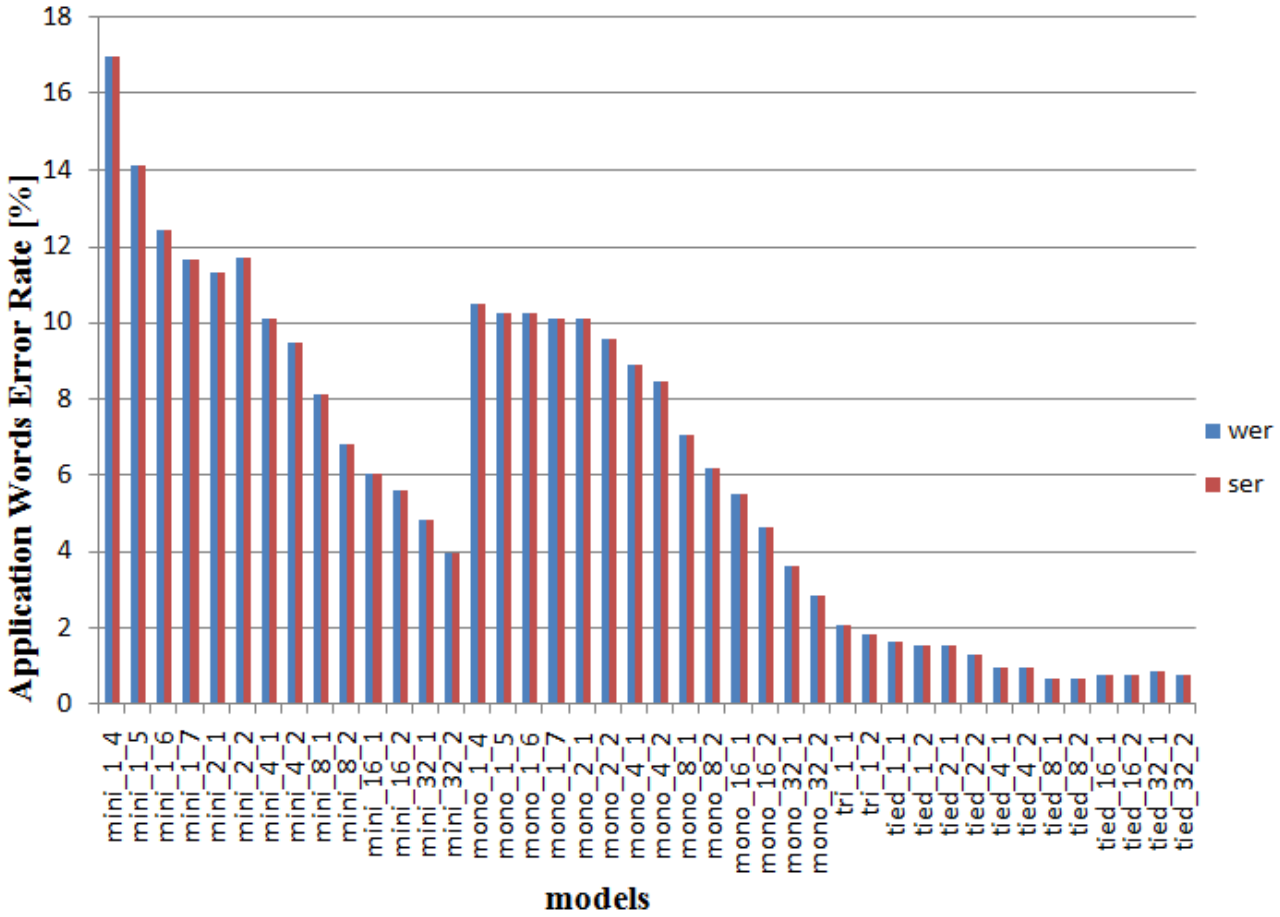


Fig. 3 Results of application words Recognition

```
====================== HTK Results Analysis =======================
  Date: Mon Mar 28 18:28:00 2011
  Ref : nworkdir_train/testset.mlf
  Rec : nresults_train/tied_32_2/rec.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=99.22 [H=1147, S=9, N=1156]
WORD: %Corr=99.22, Acc=99.22 [H=1147, D=0, S=9, I=0, N=1156]
------------------------ Confusion Matrix -------------------------
       slovo  skončí  ponuka  pomoc  stop  spojka  opakuj  zavolaj  vytoč  adresa  zoznam  predaj  další  poslať  predaj  zmeniť  vymazať  uložiť  prehrať  nahrať  poslať  prog
slov     40       0       0      0     0       0       0        0      0       0       0       0       0       0       0       0        0       0        0       0       0     0
skon      0      93       0      0     0       0       0        0      0       0       0       0       0       0       0       0        0       0        0       0       0     0
ponu      0       0      35      0     0       0       0        0      0       0       0       0       0       0       0       0        0       0        0       0       0     0
pomo      0       0       0     35     0       0       0        0      0       0       0       1       0       0       0       0        0       0        0       0       0     0
stor      0       0       0      0    93       0       0        0      0       0       0       0       0       0       0       0        0       0        0       0       0     0
stop      0       0       0      1     0      46       0        0      0       0       0       0       0       0       0       0        0       0        0       0       0     0
pokr      0       0       0      0     0       0      55        0      0       0       0       0       0       0       0       0        0       0        0       0       0     0
opak      0       0       0      0     0       0       0       36      0       0       0       0       0       0       0       0        0       0        0       0       0     0
spoj      0       0       0      0     0       0       0        1     39       0       0       0       0       0       0       0        0       0        0       0       0     0
zavo      0       0       0      0     0       0       0        0      0      41       0       0       0       0       0       0        0       0        0       0       0     0
vyto      0       0       0      0     0       0       0        0      0       0      42       0       0       0       0       0        0       0        0       0       0     0
adre      0       0       0      0     0       0       0        0      0       0       0      46       0       0       0       0        0       0        0       0       0     0
zozn      0       0       0      0     0       0       0        0      0       0       0       0      89       0       0       0        0       2        0       0       0     0
pred      0       0       0      0     0       0       0        0      0       0       0       0       0      43       0       0        0       0        0       0       0     0
dalš      0       0       0      0     0       0       0        0      0       0       0       0       0       0      45       0        0       0        0       0       0     0
posl      0       0       0      0     0       0       0        0      0       0       0       0       0       0       0      33        0       0        0       0       0     0
prid      0       0       0      0     0       0       0        0      0       0       0       0       0       0       0       0       39       0        0       0       0     0
zmen      0       0       0      0     0       0       0        0      0       0       0       0       0       0       0       0        0      46        0       0       0     0
vyma      0       0       0      0     0       0       0        0      0       0       0       0       1       0       0       0        0       0       43       0       0     0
ulož      0       0       0      0     0       0       0        0      0       0       0       0       0       0       0       0        0       0        0      39       0     0
preh      0       0       0      0     0       0       0        0      0       0       0       0       3       0       0       0        0       0        0       0      39     0
nahr      0       0       0      0     0       0       0        0      0       0       0       0       0       0       0       0        0       0        0       0       0     0
posl      0       0       0      0     0       0       0        0      0       0       0       0       0       0       0       0        0       0        0       0       0     0
prog      0       0       0      0     0       0       0        0      0       0       0       0       0       0       0       0        0       0        0       0       0     0
Ins       0       0       0      0     0       0       0        0      0       0       0       0       0       0       0       0        0       0        0       0       0     0
====================================================================
```

Fig. 4 Output from file *rec.res* for tied model with 32 mixtures

2.1.3 Time warping and training of monophones with time limits

Model that has been trained is again re-used for time warping of phoneme level. These transcripts are later used to create transcripts of triphones. Records with non-probable transcripts are excluded. Now we can initialize each model separately and train as a "single-model" on pursuance of own records. Consequently the model of silence is established and several processes of embedded training are done. Mixtures of model are increased on 1 to 32 (models *mono_2_1*, ... , *mono_32_2*).

2.1.4 Creating and training of context-dependent phonemes (triphones)

Based on the alignment of the beginning of the previous step transcripts are created on the level of triphones. So-called internal expansion is used - models like *SIL* (silence) and *SP* (short pause) do not create contextual dependences. At the edges of words rise bi-phones or monophones. Two processes of Baum Welch training follow (models *tri_1_1* and *tri_1_2*). Statics of particular occurrences are generated with these Re-Estimations too.

2.1.5 Creating and training state clustered triphones

According to statistics from the previous step, and according to rules of phonetics of the Slovak language, some states are tied and then are trained simultaneously. Again two processes of Baum Welch Re-Estimations follow (models tied_1_1 and tied_1_2). Training is concluded by increasing number of mixtures gradually on 2,4,8,16,32. After each increase two processes of Baum Welch Re-Estimations follow (models tied_2_1, ... , tied_32_2). Now is the training of triphones with tied states finished. Re-Training mixture monophones follow and it is also the end of process script *MonoTrain*.

2.2 Recognition of application words

It is realised through *sviptest-apwords* procedure and through *MonoSVIP.pl* script which are part of Masper. The recognition script takes as input parameters the name of the directory holding the trained models. Before running the script we need to edit two test configuration files in *config* subdirectory:

- *vocab.lis* – this file define the test vocabulary, including semantic mapping
- *testccds.lis* – this is the list of corpus codes participating in the test [1]

In the training procedure parameterisation MFCC_0_D_A_Z was used. The output of script in Cygwin is WER (word error rate) and SER (sequence error rate) for all trained models. In my case, each of these two rates was the same because application words consisted of one-word sequence. Graph showing the results of error rates is on the next page. Of the information we can logically deduced percentage of correct recognition:

Correctly Identified Words = 100% – Words Error Rate

Detailed recognition results are stored in the subdirectory *nresults_train* for all models in files *rec.res.* Example of recognition results for *tied_32_2* models is shown on the next page.

3.Conclusions

As we see from results recognition (on the next page), Hidden Markov Models is meaningful because of low error rate which is being decreased by addition of useful information especially models with 32 Gaussian mixtures.

Each set of parameters should cause different outcomes. Sometimes we come to surprising conclusions that not come up to expectations. In my case I expected that model with 32 mixtures would have better rates than model with 8 mixtures. But error rate for tied model with 8 mixtures reached 0,69% and 0,78% for tied model with 32 mixtures. I can said that speech recognition per HMMs with 5 emitting states and MFCC_0_D_A_Z parameterization reached excellent results bellow 1% of error rate. This technology can offer many options for the people around the world, so it is very useful.

Acknowledgements

This article was supported by VEGA – 1/0718/09 and FP7-ICT-2011-7 HBB-Next.

References

[1] Young, S. et al.: The HTK Book (for HTK Version 3.2.1). Cambridge University Engineering Department, 2002

[2] Juhar, Jozef - Cizmar, Anton - Rusko, Milan - Trnka, Marian - Rozinaj, Gregor – Jarina, Roman:Voice Operated Information System in Slovak, In: Computing and Informatics, Volume 26, 2007, Number 6, ISSN 1335-9150

Adaptive ARQ/HARQ for H.264 Video Streaming Over Wireless Channels with Variable Error Rate

Ján POCTAVEK [1], *Jaroslav POLEC*[1], *Kvetoslava KOTULIAKOVÁ*[1]

[1] Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
poctavek@ktl.elf.stuba.sk, polec@ktl.elf.stuba.sk, kkotul@ktl.elf.stuba.sk

Abstract. *In this paper, we construct a new adaptive ARQ/HARQ algorithm for H.264 video streaming over wireless channels. The algorithm takes into account extensive changes in channel bit error rate and according to its current state, the appropriate transmit scheme is chosen. Our scheme switches between hybrid ARQ using RS codes, when channel is in high error state, and pure ARQ method when in low error state to save the unnecessary throughput and reduce the delay.*

Throughput performance of the proposed algorithm has been analyzed by simulating the transmission over land mobile satellite channel. Furthermore, we performed an optimization of scheme parameters to obtain best possible throughput for each channel conditions.

Keywords

H.264, adaptive ARQ/HARQ, Reed-Solomon, land mobile satellite, simulation.

1.Introduction

In the next generation wireless communication systems, there is an increasing demand for wireless multimedia services. However, successful multimedia transmission still poses important challenges that deserve special attention. One of the main challenges is the difficulty of reliable transmission over time varying and lossy wireless links. High and variable error rates of the wireless channel causes packet erasures. Multimedia data is especially vulnerable to channel errors due to the predictive coding techniques used in multimedia compression schemes such as H.264 and MPEG-4 [1]. These compression schemes divide continuous media into frames, and encode the frames so that they are dependent on each other. Due to this dependency an error that occurs in a frame may propagate to other frames. Moreover, multimedia data is usually time-sensitive. A frame should be received and decoded prior to a deadline; otherwise, the frame will be discarded and the succeeding frames will also be affected.

When transmitting through channels with errors, the most common solution is to implement standard automatic repeat request (ARQ) strategies (considering feedback channel is available), such as Stop and Wait (SW), Go-Back-N (GBN), or Selective Repeat (SR). These algorithms however are effective only in channels with low error rate because any single errored bit causes whole packet to be discarded and re-sent.

Hybrid ARQ methods (HARQ) extend the communication ability of ARQ in considerably lower levels of signal to noise ratio (SNR) employing forward error correction codes (FEC). There are three types of HARQ schemes (HARQ type I, II and III) [2].

2.New Adaptive Scheme

The process of developing adaptive ARQ scheme has two main stages:

- Choosing the type of transfer scheme used in high error state.
- Designing the adaptation algorithm that manages switching between pure ARQ and the other scheme.

2.1 HARQ Scheme Used

In our approach, HARQ type-I with RS (Reed-Solomon) code is chosen as FEC. The reason for using RS is that in common environments channel errors occur in bursts and often are not mutually independent. Under such circumstances, the RS code is much more efficient because it can handle error bursts very well [3].

Main structure of RS codes is shown in Fig. **1** .

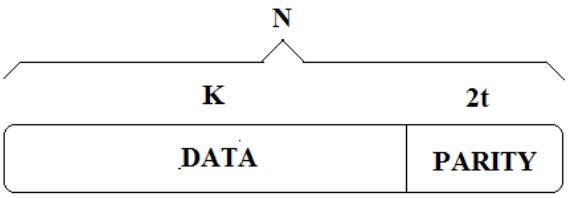


Fig. 1. Code word of RS code

RS code [N, K, t]

N – Length of code word (in RS symbols)

K – Count of information symbols (data)

t – Count of repairable symbols (2t = redundancy)

$$t = \frac{(N-K)}{2} \quad (1)$$

Code word is created by generating polynomial:

$$g(x) = (x+\alpha^{i}) \cdot (x+\alpha^{i+1}) \cdot \ldots \cdot (x+\alpha^{i+2t-1}) \quad (2)$$

for i = 0,1,…,2t-1

2.2 Considered ARQ Strategies

Our main focus is on GBN and SR transfer modes. The main advantage of GBN is a good throughput at low implementation complexity since it does not require buffering on the receiver side. SR mode, on the other hand, has the best throughput, but at the cost of a higher complexity. We also analyze the SW mode, but only marginally, as with increasing delay its throughput quickly deteriorates beyond usable values (shown in Fig. 7).

Anyway, our presented methodology for optimizing an adaptive throughput is independent of a transfer mode.

2.3 The Adaptation Algorithm

The design of the adaptation algorithm is a very important part of the whole scheme as the final throughput is dependent on its performance.

We consider the forward channel to have two states; L state (low error rate) and H state (high error rate), as shown in Fig. 2 [5].

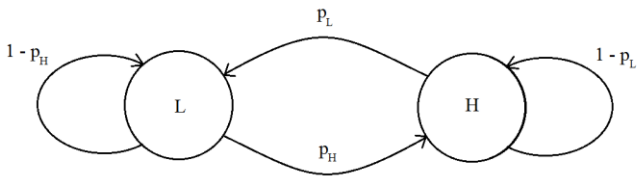


Fig. 2. Channel model with 2 states: low error and high error state with probabilities of changing or not changing the state

From this Gilbert model, based on assumptions made by Sastry in [4] and modified by Yao in [5], the switching scheme can be derived (Fig. 3). Big advantage of this model is that it does not require the knowledge of instantaneous packet error probability [5]. This channel state model is used to describe the way of switching between transfer methods.

We have modified the original Yao's model to work with pure ARQ and HARQ type I transfer mode using RS coding (Fig. 3).

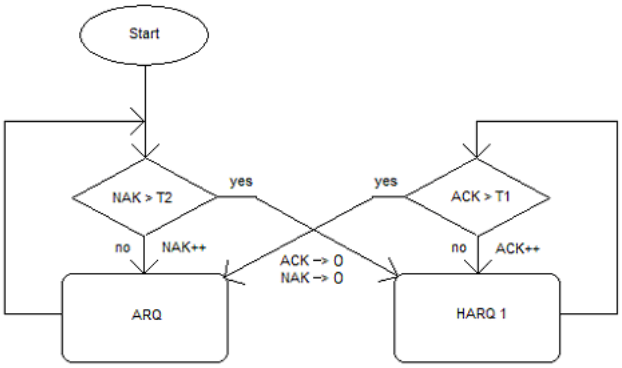


Fig. 3. Proposed Adaptive ARQ/HARQ Scheme

We continued further by introducing an objective performance criterion, using which we are able to investigate the throughput performance of proposed scheme and find optimal parameter settings for every channel conditions. Furthermore, it is possible to automatize the search, so we are able to determine the best parameter values for any transmission setup. We also consider the other ARQ schemes and by attaching the appropriate error detection (CRC) code to the packets in pure ARQ mode, we create robust detection-correction system.

2.4 Scheme throughput

Corresponding to the two channel states, there are two operation modes in the proposed scheme.

In channel state *L*, the transmitter follows the pure GBN or pure ISR ARQ and the throughput can be expressed as [6]:

$$\eta_{B-GBN} = \frac{1-P_e}{1+S \cdot P_e} \quad (3)$$

$$\eta_{B-ISR} = 1-P_e \quad (4)$$

where P_e is the block error probability which can be expressed as:

$$P_e = 1-(1-P_b)^n \quad (5)$$

where P_b is a bit error probability in BSC, n is the block length and S is the delay expressed in data blocks.

In channel state *H*, the transmitter works in HARQ transmission mode. It encodes the data before transmission using RS *[N, K, t]* code. As stated in Fig. 1, the RS code used is able to repair t symbol errors. Its code word length is *N* symbols and its information word length is *K* symbols with code rate *K/N*. Its throughput can be expressed as [6]:

$$\eta_{H-GBN} = \frac{1-P_{He}}{1+S\cdot P_{He}} \quad (6)$$

$$\eta_{H-ISR} = 1-P_{He} \quad (7)$$

where P_{He} is block error probability for channel state *H*.

2.5 Parameters Description

In Fig. 3, you can see two parameters – threshold T_1 and T_2. These parameters are used as a limit for positive and negative acknowledgements respectively. When the amount of consecutive NAKs in low error state overreaches the threshold T_2, the transmission mode is switched from ARQ mode to HARQ. In high error state, when the amount of consecutive ACKs overreaches the threshold T_1, the transmission is switched back from HARQ to ARQ. The ACK/NAK counters are reset to zero after every mode switch.

3. Construction of the Simulator

After specifying all parameters for transmission and schemes, we are ready to build a simulator to find optimal threshold parameters.

We have constructed a simulator that can work with all common ARQ strategies and its parameters are described below.

3.1 Modes of Operation

The simulator has two modes of operation:

- No-RS
- Full-RS

When Full-RS mode is requested, FEC recovery data from RS code is appended to the end of the packet, extending it to the length of *N*. In No-RS mode, the data packet of length *K+CRC* is sent, where *CRC* value means error detection code length. In our simulations, we use 32-bit CRC code, which is the same as used in ethernet frames that contain comparable number of bits.

Similar overall approach of changing packet length is used also in wireless broadcast [7].

3.2 Delay

Round-trip delay is always considered in the simulation. At the start of the simulation, it is expressed in multiples of data length (*K*), because this is the only value that is present in all modes. Delay in case of ARQ/HARQ scheme cannot be expressed exactly in number of slots, because the packet length changes over time (as RS gets involved). The delay is, therefore, always evaluated to be of the same length (in bits); no matter the current transfer mode (all simulation results presented in this paper have delay set to 5 times *K*).

4. Throughput analysis

We evaluate the throughput performance of the proposed ARQ scheme under the following assumptions:

1. The feedback channel of the system is error-free. As described in [8], feedback errors degrade the throughput of the scheme, but their effect on the parameter optimization turns out to be almost negligible.
2. Errors in consecutive packets occur independently. We assume that mutually independent bit errors are distributed according to selected channel model. RS codes achieve the best performance when repairing error bursts [3]. Therefore, independent errors are considered the worst-case scenario for the throughput performance. (It is also possible to simulate the real channel with error bursts and packet losses [9], but for our purpose it does not affect the main outcome of this paper.)
3. Channel errors are always detected. We assume that error detection code is robust enough to detect all channel errors in both ARQ and HARQ (because RS code is error detection code too).

In Figures 4 - 7, we provide sample simulation results with parameters optimized for various channel settings.

The optimal threshold values for two main ARQ schemes using two different RS code settings are shown in Tab 1.

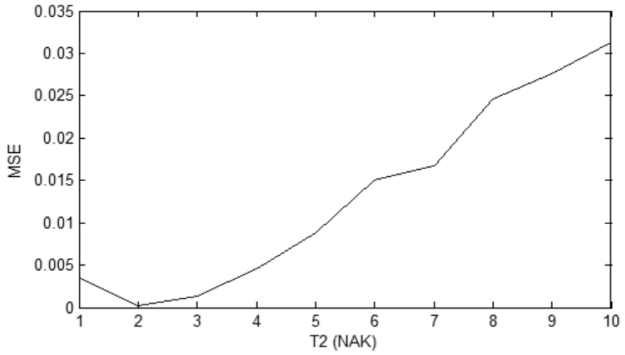


Fig. 4. Optimization of parameter T2 in rural environment, GBN ARQ mode, RS[511, 383] and T1=150. You can see that for this environment optimal parameter value for T2 is 2 (with MSE(2)=0.00019).

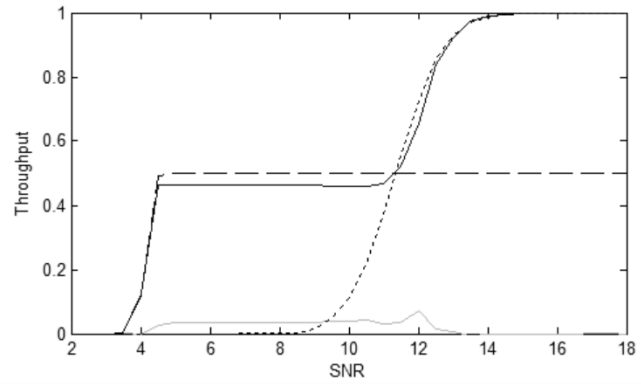


Fig. 5. Scheme throughput with optimized parameter T1; RS [511, 255], solid black - ACK=80, NAK=2; solid grey - throughput difference from adaptive scheme and ideal ARQ/HARQ; dashed - ideal HARQ; dotted - ideal ARQ

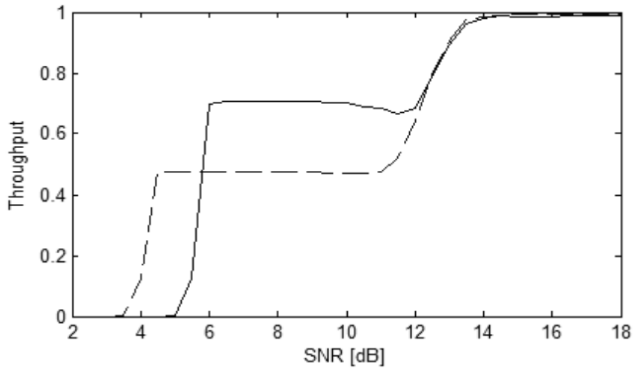


Fig. 6. Throughput in dependence of SNR for two optimized schemes with different RS code in rural environment, GBN; solid black - RS[511, 383], ACK=150, NAK=2; dashed - RS[511, 255], ACK=80, NAK=2

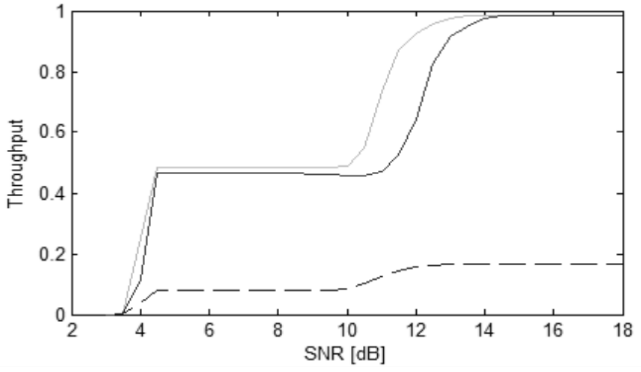


Fig. 7. Throughput in dependence of SNR for all major ARQ schemes in rural env, RS[511, 255]; solid grey - Selective repeat, ACK=130, NAK=5; solid black - GBN, ACK=80, NAK=2; dashed - S&W, ACK=20, NAK=3

	GBN	SR
RS [511, 255]	ACK=80 NAK=2	ACK=130 NAK=5
RS [511, 383]	ACK=150 NAK=2	ACK=100 NAK=5

Tab. 1. Optimal values of T_1 and T_2 for rural environment

5. Conclusion

In this paper, we proposed and analyzed a new adaptive ARQ/HARQ scheme and the adaptation rule that estimates channel state by counting ACKs and NAKs. The described model provides a very good throughput performance when threshold parameters are properly used and set up. Another advantage of this scheme is its low computational complexity involving only two feedback counters. Also the quick optimization of performance parameter allows us to find the best scheme setup for various environments and transmission modes. Next attention can be focused on scheme's marginal error detection capability to properly determine the best combinations of ARQ detection and HARQ correction codes and possibly to set up switching between more than two operation modes.

Acknowledgements

Research described in the paper was financially supported by the Slovak Research Grant Agency (VEGA) under grant No. 1/0602/11, No. 1/0243/10.

References

[1] ITU-T Rec. H.264 ISO/IEC 14496-10 AVC, Advanced video coding for generic audiovisual services. *ITU-T*, 2003

[2] KALLEK, S.: Analysis of a type II hybrid ARQ scheme with code combining, *IEEE Trans. Commun*., vol 38, pp. 1133-1137, 1999

[3] PURSER, M.: Introduction to Error-Correcting Codes, Artech House, 1995

[4] SASTRY, A. R. K., Improving Automatic-Repeat-Request (ARQ) Performance on Satellite Channels under High Error Rate Conditions. *IEEE Transaction on Communications*, Vol. COM-23, No. 4, pp. 436-439, April 1975

[5] YAO, Y.D., An Effective Go-Back-N ARQ Scheme for Variable-Error-Rate Channels. *IEEE Trans. Commun*., Vol. 43, No. 1, pp. 20-23, Jan. 1995

[6] POLEC, J., KARLUBÍKOVÁ, T.: Stochastic models in telecommunication (in Slovak) *1. FABER* 1999

[7] TAVARES, J.; NAVARRO, A.: Optimal IP packet length for DVB-T transmission, *Proceedings of the Ninth International Symposium on Consumer Electronics*, ISBN: 0-7803-8920-4, pp. 385-390, 2005

[8] LIINAHARJAA, M., CHAKRABORTY S. S.: Analysis and Optimization of an Adaptive Selective-Repeat Scheme for Time-Varying Channels with Feedback Errors, *International Journal of Electronics and Communications*, Vol. 56, Issue 3, pp. 177-186, 2002

[9] POLEC, J., POCTAVEK, J., PAVLOVIČ, J., KRULIKOVSKÁ, L.: Cascade of Markov Models for Packet Loss and Subsequent Bit Error Description; In: *Proceedings ELMAR 2010*. Zadar, Croatia, pp. 285-288

Building IP-based television systems using open-source software (April, 2011)

Andrej BINDER, Ivan KOTULIAK

Faculty of Informatics and Information Technologies, Slovak University of Technology,
Ilkovičova 3, 812 19 Bratislava, Slovakia
andrej@binder.sk

Abstract. *This paper serves as an overview of available open-source software that can be used to build a complete IP television system on an IP network. It defines a set of basic components that have to be present in such a system to be fully functional. It then analyzes the available software solutions that can fill the role of each component. The ultimate goal of this paper is to provide a viable starting point for anyone that wishes to develop a fully featured IPTV solution.*

Keywords

IP-based television, multimedia broadcasting, open-source software, specific communication network application

1. Introduction

This document analyzes the basic aspects of a technology used for multimedia broadcasting over IP Networks. It describes the basic structure of a system designed for efficient delivery of multimedia content from its source to the customer using IP based technologies. Since this kind of a system is of a highly modular nature, it begins by listing the basic technological components needed for its operation and then describes their relation to the rest of the system. Once the relations between the components are clearly explained, it proceeds with a detailed description of each of the components. The analysis is finalized by presenting a generic topology of a fully featured IP television system.

2. Common IP television system components

Every network subsystem consists of several independent modules, each of which has its distinct role and position in the system. This section describes each of the components of the ITPV network subsystem, explains its role in the system, and defines its position in relation to other components directly connected to it.

2.1 Content gateways

Content gateways play a crucial role in every multimedia subsystem. Their purpose is to obtain multimedia streams from various outside sources and convert their signaling and encapsulation into the form required by the network on which the system is built.

Our definition of a content gateway does not include re-encoding of the multimedia streams. They are supposed to capture the streams as they are delivered to them and make only those modifications to the stream that are required for it to be transported through the selected network. Ideally there should be no information loss caused by this conversion but sometimes it is inevitable (for example when the source is of analog nature).

They are usually getting their data directly from an outside source and outputting it either directly into the network or into a re-encoder, which converts it and then injects it to the transport network in a re-encoded form.

2.2 Re-Encoders

Re-Encoders serve a special role in an IP television system. Their purpose is to receive a stream with a specific set of parameters and convert it into a stream with a different set of parameters. The parameters influenced by the conversion include bandwidth requirements, resolution, codec choice and others. This process often results in a loss of data, but it is not always a rule. Some codec's are of a lossless nature allowing them to decrease the bandwidth requirement of a stream by increasing its computational complexity.

They can be placed between a content gateway and the transport network. This is desirable if the whole transport network is homogenous, and the receivers are all compatible with a common stream, but the stream outputted by the content gateway does not match the criteria required by either the transport network or the receivers used.

When the network is not homogenous it is often reasonable to place the re-encoders to the borders dividing the network into sections with different parameters.

2.3 Transport networks

A transport network is responsible for reliable and effective transport of the data stream from its source (a content gateway, an re-encoder or a third party) to its destination (an receiver, an re-encoder or a third party). The physical topologies and technologies used in a transport network are defined by historical decisions, geographical limitations and various economic influences that are out of scope of this document.

The most common technology used in IPTV broadcast is basic layer 2 ethernet switching and IP multicast.

2.4 Service Directories

Service directories are special components of most IPTV systems. Their sole purpose is to distribute information about available multimedia streams to the end point receivers.

There are multiple methods of delivering the information from the service directory to the endpoint receiver dynamically. The simplest one is for a service directory to simply broadcast (or multicast) the information on a common address that the receivers listen on.

2.5 Network Controllers

Network controllers are often nicknamed the "brains" of IPTV subsytems. They may not be present in all designs but they often turn into a requirement with growing network complexity and the desire to provide advanced services to diverging groups of customers. Their main role is to act as a center for the management, accounting, and access control for the network while also providing many additional services.

They are named network controller because their most important role is to control all the active devices in the network that need to be dynamically managed. This control involves the configuration of various quality of service parameters of the transport network that is often required in heavy loaded networks that need to reliably broadcast multimedia streams.

2.6 Receivers

Receivers are the end points of IPTV networks. This work will use the generic term "receiver" as a common name for any device receiving multimedia streams from the IPTV subsystem. These devices may range from a specialized software installed on a common PC, through basic hardware devices like set-top boxes connected to televisions, to various other devices that do not even need to be the end-points for the stream. A gateway to a different network that is used as a border router for multimedia stream exchange can be called a receiver if the stream traversing it is coming from inside our network into some other.

They are usually connected directly to the transport network that they rely on for accessing the multimedia stream data, obtaining service lists from service directories and communicating with the network controllers. In some rare cases it is possible to separate the transport network used for receiving the inbound multimedia stream data from the connections used for obtaining service lists and communicating with the network controllers.

3. Generic IP television system topology

Most IP television system designs have a very similar basic topology consisting of a set of content gateways, a single or multiple transport network segments and a set of receivers. In order to satisfy any advanced service requirements, most designs also contain a specific set of optional components. These optional components include re-encoders, network controllers, service directories and others.

A very basic example of a general IP television system topology is visualized in Figure 1 – Generic IP television system topology. This example includes two transport network segments divided by a re-encoder. It also includes all the optional components incorporated into the network. The links between the components represent the flow of

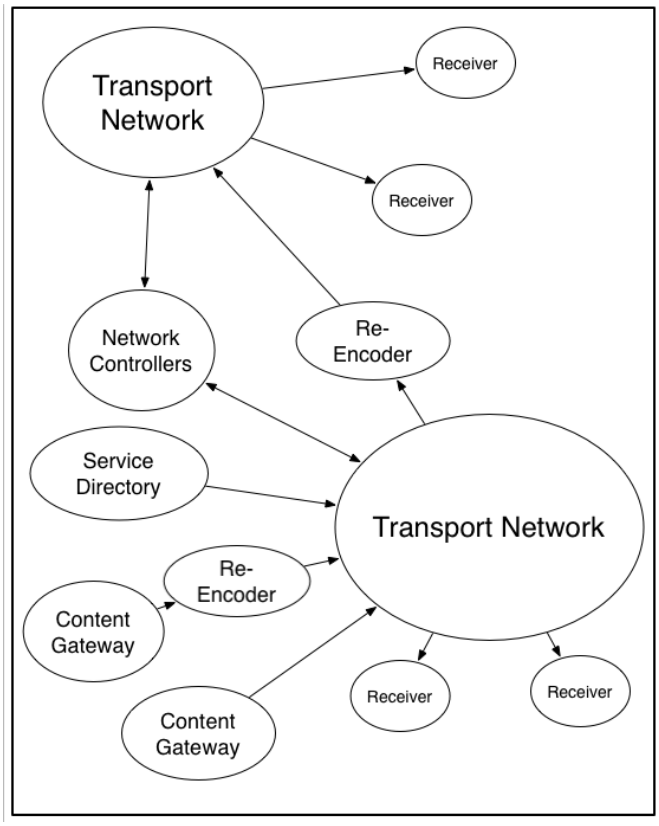


Fig. 1. Generic IP television subsystem topology

communications and the arrows represent its type (uni or bi-directional).

4. Available open source solutions

Because highly modular nature is a characteristic property of almost every IP television system, it is possible to flexibly combine various software solutions in order to create a fully featured system that can match the requirements of almost any project regardless of its scale. This section briefly describes each of the available open source solutions that should be considered when designing an open source IP television system. This information was obtained as a result of the author`s diploma thesis, whose goal was to design a fully functional IP television system in a laboratory and production environment.

4.1 Content gateway software: Linux-dvb project

The main role of a common content gateway is to obtain multimedia information from various external sources that are based on standards like DVB-T, DVB-S, DVB-C or oven those that are based on completely analog signals. This role inherently requires the use of a special hardware that is designed for this very purpose. This hardware has to include a tuner that is capable to tune to the frequency required by a chosen standard and a set of circuits that can assist in decoding or decrypting the streams obtained. There is a very promising open source project called linux-dvb that shelters all the low-level kernel drivers related to this kind of devices under a single framework featuring a single common application interface. The existence of this project makes linux the best environment for deployment of a content gateway. All of the content gateway software that will be discussed later relies on this framework.

4.2 Content gateway software: Dvblast

Dvblast is lightweight MPEG-TS demultiplexing and streaming application designed specifically for implementation of a production-ready content gateway. It can accept input from any DVB-ASI or linux-dvb supported card. It can also accept UDP or RTP streams. Its main role is to do PID-based or service-based demultiplexing of a DVB-S, DVB-S2, DVB-C or DVB-T transponders. It supports hardware or software PID filtering and descrambling via a CAM device. EPG information pass-through is also supported.

Together with the dvb-linux project framework it has proven to be the most reliable content gateway solution. It does not consume almost any system resources and thus can be scalable indefinitely as long as there is network bandwidth and PCI bus capacity available.

4.3 Content gateway software: getstream, MuMuDVB, DVBStreamer, VLS, ffserver

These projects are all very similar to dvblast but their versions available at the time of writing proved to be inferior to the dvblast solution mentioned earlier. They either did not have all the features that a fully featured content gateway software has to have, were not production ready or their development has ceased completely. They are still worth looking into because some of them may develop into viable alternatives to dvblast in the future.

4.4 Content gateway software: VLC

VLC is very universal software that can, among other things; act as a fully featured content gateway. Its problem is that its universality makes it so heavyweight and resource intensive that it is not recommended for production environment. It is capable of fulfilling the role of a good content gateway without a problem but the system resources required while doing it are several orders of magnitude higher than those of more specialized solutions mentioned earlier.

4.5 Re-Encoder software: ffmpeg, MEncoder, Transcode, VLC

Re-encoder software normally consists of just two basic parts. One of them is the demultiplexing part that is responsible for encapsulating the bare multimedia stream in a proper container and the second one is the actual multimedia codec used to encode the stream. While the demultiplexing part differs between the solutions mentioned, the codec part that does all the heavy work is usually the same for each of them as it is just an externally linked library that is often identical. This means that the only mayor difference between the solutions is the way that they are managed, not the performance of their workload processing. This makes it impossible to clearly recommend a specific solution. It is needles to say, that each of the solutions listed can act as a fully featured re-encoder component of an IP television system.

4.6 Service directory software: minisapserver

Service directory module of a IP television system is its least standardized part. This means that almost all of the production solutions used today are proprietary and developed specifically with the parameters of the chosen receiver in mind. There is one exception. The minisapserver application is a simple daemon implementing the session announcement protocol in order to inform receivers about available multimedia streams. It is directly supported by VLC and because of its simple XML based nature can be easily implemented in almost any custom receiver.

4.7 Network Controller software: OpenIMS Core

There is no comprehensive network controller software that is directly developed with IP television applications in mind. There is an architecture that strives to become the common control mechanism for all multimedia content including not only IP television but also IP telephony and many other multimedia services. This architecture is called by the name Internet Multimedia Subsystem and the best example of a real open source project based on it is the OpenIMS Core project. The problem is that the IP television section of it is not fully standardized yet and the only fully featured solution is only compatible with the UCT IMS Client receiver application which does not fully adhere to the IMS standards.

4.8 Network Controller software: VLMA

VLMA is an application that is specifically developed to manage a set of VLC-based content gateway servers. It provides centralized management and workload distribution for a set of content gateways. It is a java-based application managed through a web interface. It controls the VLC servers using the telnet protocol. It is a good example of a centralized management component present in an IP television system; even if it can control only a small part of the whole system.

4.9 Receiver software: VLC

VLC was originally developed as a multimedia stream receiver and was later enhanced with a lot of other functionality that can make it act as a content gateway or an re-encoder. While it is still great in its role of a simple receiver, the added complexity present because of its universality makes it hard to use by novice users. It is the perfect tool for an administrator because both its good debugging capability and its ability to cooperate with almost any protocol that known in the IP television field. Using it as a customer solution would probably require creating a special version of it with limited functionality and a simple user interface.

4.10 Receiver software: Kaffeine

Kaffeine is very similar to VLC in its role of a receiver. It is not burdened by all the complexity present in VLC because of its ability to fulfill other roles. This makes it a more suitable application for the end user.

4.11 Receiver software: UCT IMS Client

UCT IMS Client is a good example of a solution that was developed specifically as a receiver in a bigger, centrally managed IP television subsystem. It is based on an architecture called UCT Advanced IPTV, which is in turn based on the IMS architecture. This means that the UCT IMS Client can be used in any IMS based network and can make use of its centralized authentication, accounting and provisioning features.

4.12 Receiver hardware

There is no open source receiver hardware available but there is a lot of products that can work flawlessly in a IP television system based purely on open source hardware. Good examples of such hardware are the multimedia players produced by Amino, HDI Dune and others. Most of the players available are based on chips produced by Sigma Designs. Sigma Designs also builds specialized development kits that can be used in order to build a custom fully featured receiver set-top box that can be based purely on open source software.

5. Conclusion

The content of this paper proves that the IP television subsystem is an idea that is well embraced by the open source community. All the mayor components are available in fully featured and production ready flavors. This means that building a complete and well scalable production ready IP television system does not require much more than minor modifications to the solutions available. The fact that there are many alternatives available for each of the components together with the ability to freely modify any of them since they are all open source means that it is possible to easily and cheaply build a custom solution that is well suited to fit almost any requirements.

Acknowledgement

This paper has been supported by the VEGA project 1/0243/10.

References

[1] VideoLAN website: http://www.videolan.org/

[2] UCT IMS website: http://uctimsclient.berlios.de/

[3] OpenIMS Core website: http://openimscore.org/

[4] LinuxTV project website: http://www.linuxtv.org/

[5] Maisonneuve, J., Deschanel, M.: An Overview of IPTV Standards Development. IEEE Transactions On Broadcasting, vol. 55, č.2, jún 2009, p. 315-328

[6] Saman, J.: Streaming networks with VLC [online], http://m2x.nl/mambo/packages/abstract_nluug2006.pdf

[7] Tomek, R., Kadlic, R., Mikóczy, E., Podhradsky, P.: IPTV applications in the NGN environment. 50th International Symposium ELMAR-2008, 10-12 September 2008, p. 549-552

[8] Camarillo, G., Garcia-Martin, M. A.: The 3G IP Multimedia Subsystem. 2. rel. 2006.

Shot Boundary Detection Based on H.264 Compressed Domain

Tomáš Máťuš[1], *Lenka Krulikovská*[1], *Jaroslav Polec*[1]

[1] Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
tmatus15@gmail.com, polec@ktl.elf.stuba.sk

Abstract. *In this paper we propose a method to detect abrupt cut changes in H.264 coded video that operates directly in the compressed domain. The proposed algorithm is fast and simple and it is suitable for the real-time implementation. This method is based on monitoring the number of I macroblocks in frames P and B. The ability to find cuts using this method for different GOP structures was analyzed in the experiment. Analysis was focused on sensitivity of this method to various evaluation thresholds that determine if the cut was occurred. The evaluation was performed on the base of three metrics: precision, recall and F1-measure.*

Keywords

Shot transition detection, H.264/AVC, GOP structure.

1. Introduction

The extensive usage of digital video material gives rise to the need of improving the accessibility to video content by the users. A fundamental and initial step of the applications which provide this is, naturally, to structure the videos into shorter elementary units, i.e., to perform a temporal segmentation of the video. Among the possible types of elementary units, there is the shot which has been considered an appropriate elementary unit for this kind of applications and has been used by a great majority of them; a shot consists of a series of interrelated consecutive pictures taken contiguously by a single camera and representing a continuous action in time and space. Due to the importance of shot transition detection in this application context, shot transition detection tools have been an extensively researched and reported in the relevant literature [1, 2, 3, 4].

However, digital video content is nowadays made available in a compressed format to reduce its storage and transmission requirements. Recently, more and more applications of H.264/AVC standard [5] call for a set of new methods that can effectively organize, present, index, and search H.264/AVC bit streams. Thus a method for shot transition detection in compressed domain is needed.

This paper is structured as follows: in the second section a proposed method of shot cut detection is described. Results are displayed in the third section. All results are summarized and discussed in conclusion.

Fig. 1. An example of abrupt video cut.

2. Proposed method

We propose an algorithm to automatically identify video shot boundaries using compressed domain features of H.264. In this method the compressed domain features of H.264 video obtained during encoding the video data is used to identify the shot boundaries. The feature used in our approach is the number of I macroblocks (MBs) in P and B frames. For the evaluation of proposed method we have chosen following measures: precision, recall and F1 score. The precision measure is defined as the ratio of correct video cut detections over the number of all video cut detections [6].

$$Precission = \frac{|Det \cap GT|}{|Det|} \tag{1}$$

where GT denotes the correct cut detection and DET denotes all detected (correct and false) cuts.

The Recall measure is defined as the ratio of correct video cut detections over the number of all correct video cut detections [6].

$$Recall = \frac{|Det \cap GT|}{|GT|} \tag{2}$$

The F1 score combines precision and recall and is defined as the two times ratio of precision times recall over precision plus recall [7].

$$F1 = 2 \cdot \frac{Precission \cdot Recall}{Precission + Recall} \tag{3}$$

3. Experimental results

We confirmed the effectiveness of proposed method through a test experiment. For test purposes we created a video sequence (1989 frames) at CIF resolution (352 x 288 pixels) with 7 abrupt cuts sampled at rate of 30 frames per second. The test video sequence consists of eight standard test sequences: akyio, foreman, hall, flower, mobile, mother-daughter, stephan and bus.

We have simulated three GOP structures for H.264 video encoding – IPPPP, IPBPB and IPBBP. We have employed fixed threshold for determining if examined P or B frame is an abrupt cut. The threshold takes values from 0 to 400 I MBs with step 50. The obtained results for each GOP structure are evaluated by precision, recall and F1 score.

3.1 IPPPP GOP structure

Fig. 2 illustrates the shot detection for IPPPP GOP structure. All seven cuts can be clearly identified as there is sufficient difference among cuts and non-cuts frames.

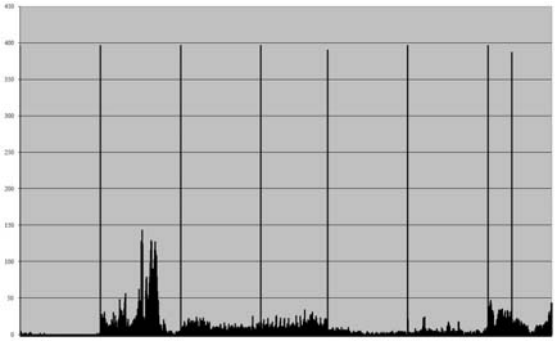


Fig. 2. Plot of shot detection for IPPPP GOP structure (x-axis: frames, y-axis: number of I MB in frame).

Tab. 1 shows the obtained results for detection with simulated thresholds and the results of evaluation measures.

I MBs	GT	Det	R	P	F1
0	7	1989	1	0,004	0,007
50	7	56	1	0,125	0,222
100	7	23	1	0,304	0,467
150	7	7	1	1,000	1,000
200	7	7	1	1,000	1,000
250	7	7	1	1,000	1,000
300	7	7	1	1,000	1,000
350	7	7	1	1,000	1,000
400	0	0	0	0,000	0,000

Tab. 1. The results obtained by shot cut detection for IPPPP GOP structure.

The maximum value 1 for recall (R), precision (P), and thus also for F1 score, was reached for selected threshold from 150 I MBs to 350 I MBs in P or B frames.

3.2 IPBPB GOP structure

Fig. 3 displays the shot detection for IPBPB GOP structure. All seven cuts are visible, but non-cuts frames reach high values for second shot.

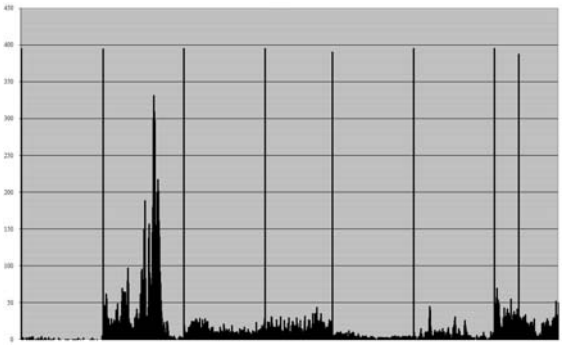


Fig. 3. Plot of shot detection for IPBPB GOP structure (x-axis: frames, y-axis: number of I MB in frame).

Tab. 2 contains the obtained values for shot cut detection with evaluation measures. The maximum value 1 for recall (R), precision (P), and thus also for F1 score, was reached only for threshold of 350 I MBs. It is caused by higher values in second video sequence, what lead to false detections.

I MBs	GT	Det	R	P	F1
0	7	1989	1	0,004	0,007
50	7	63	1	0,111	0,200
100	7	27	1	0,259	0,412
150	7	21	1	0,333	0,500
200	7	14	1	0,500	0,667
250	7	11	1	0,636	0,778
300	7	10	1	0,700	0,824
350	7	7	1	1,000	1,000
400	0	0	0	0,000	0,000

Tab. 2. The results obtained by shot cut detection for IPBPB GOP structure.

3.3 IPBBP GOP structure

Fig. 4 illustrates the shot detection for IPBBP GOP structure. All seven cuts show high amount of I MBs, but it is needed to notice that the non cuts frames in second video shot reach nearly the same values.

The results for shot cut detection are in Tab. 3. As we can see, we weren't able to reach value 1 for both recall and precision for any selected threshold. Precision never

reached the maximum value, because of high number of I MBs in non-cuts frame in second sequence.

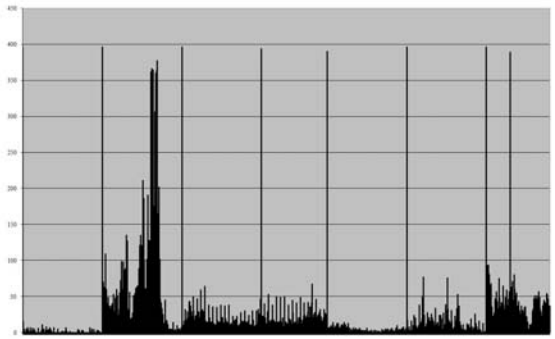


Fig. 4. Plot of shot detection for IPBBP GOP structure (x-axis: frames, y-axis: number of I MB in frame).

I MBs	GT	Det	R	P	F1
0	7	1989	1	0,004	0,007
50	7	103	1	0,068	0,127
100	7	34	1	0,206	0,341
150	7	21	1	0,333	0,500
200	7	17	1	0,412	0,583
250	7	15	1	0,467	0,636
300	7	15	1	0,467	0,636
350	7	14	1	0,500	0,667
400	0	0	0	0,000	0,000

Tab. 3. The results obtained by shot cut detection for IPBBP GOP structure.

3.4 Comparison of selected GOP structures

Fig. 5. – Fig. 7 shows the comparison of recall, precision and F1 measures for simulated GOP structures.

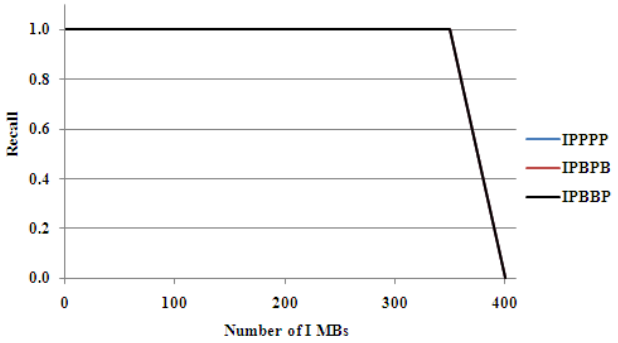


Fig. 5. Comparison of recall for selected GOP structures.

The reached values of recall are same for all GOP structure, because this measure for the performance evaluation corresponds to the ratio of correct experimental detections over the number of all true detections and the number of detected cuts is the same for every threshold's values and simulated GOP structure.

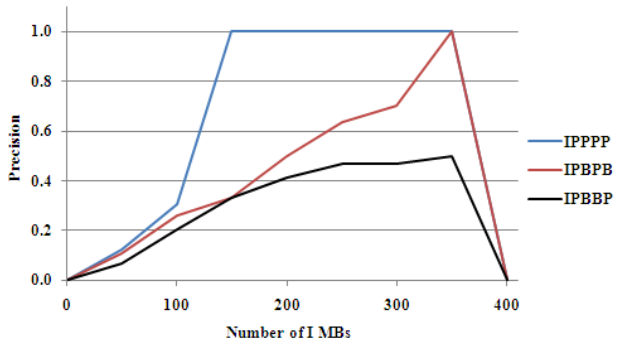


Fig. 6. Comparison of precision for selected GOP structures.

IPBBP structure didn't reach the highest precision due to many false detections in second video sequence (Foreman) caused by huge numbers of I MBs for non-cuts frames. IPBPB structure reached the value 1 only for threshold set to 350 I MBs in frames. The reason is the same as for IPBBP structure. IPPPP held the value 1 for 50% of threshold range, thus this structure gives the best result for proposed method of cut detection.

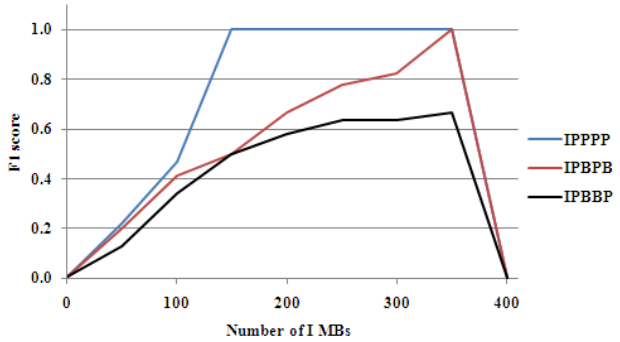


Fig. 7. Comparison of F1 score for selected GOP structures.

The comparison for F1 score looks nearly the same as the one for precision. It is caused by identical recall values for selected GOP structure (as F1 score measure is a combined measure that results in high value if and only if, both precision and recall result in high values). We can notice the result is 0 for threshold equal to 400 I MBs in frames. This threshold is too high, thus no cut is detected.

4. Conclusion

This paper presents a novel method for abrupt cut detection in H.264 coded video that operates directly in the compressed domain. Proposed method is based on the number of I macroblocks in frames P and B. Due to approach in compressed domain this algorithms is fast and simple.

The performance of proposed method was proved through test experiment for three GOP structures: IPPPP, IPBPB and IPBBP. The accuracy of shot detection was evaluated by recall, precision and F1 score.

According to obtained results, the presented method is the most effective for IPPPP structure (100% accuracy was achieved for 50% of range of simulated fixed thresholds). The structures IPBPB and IPBBP suffer for a lot of false detection cause by local motion in shots and reached much worse accuracy.

For future work, it would be needed to employ more compressed domain features of H.264 video obtained during encoding the video data in shot detection to improve efficiency of method for GOP structures using B frames. Then we would like to analyze the influence of the presented method to video trafic prediction for example in connection with methods [8, 9, 10].

Acknowledgements

Research described in the paper was financially supported by the Slovak Research Grant Agency: VEGA under grant No. 1/0602/11.

References

[1] YUAN, J. et al. A formal study of shot boundary detection. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 2, 2007, pp. 168-186.

[2] HANJALIC, A. Shot-boundary detection: unraveled and resolved?. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 2, 2002, pp. 90–105.

[3] DE BRUYNE, S. et al. A compressed-domain approach for shot boundary detection on H.264/AVC bit streams. *Signal Processing: Image Communication*, vol. 23, no 7, 2008, pp. 473-489.

[4] LIU, Y. et al. A novel compressed domain shot segmentation algorithm on H.264/AVC. *International Conference on Image Processing 2004*, Singapore, 2004.

[5] JVT Editors (WIEGAND, T., SULLIVAN, G., LUTHRA, A.). Draft ITUT Recommendation and final draft international standard of joint video specification (ITU-T Rec.H.264 |ISO/IEC 14496-10 AVC). JVT-G050r1, Geneva, 2003.

[6] CERNEKOVA, Z. *Temporal Video Segmentation and Video Summarization*, Ph.D. dissertation, Dept. App. Inf., Comenius Univ., Bratislava, SK, 2009.

[7] COTSACES, C., GAVRIELIDES, M., PITAS, I. A Survey of Recent Work in Video Shot Boundary Detections. *In Proc. of 2005 Workshop on Audio-Visual Content and Information Visualization in Digital Libraries (AVIVDiLib '05)*, Cortona, Italy, 4-6 June, 2005.

[8] ORAVEC, M., PETRÁŠ, M., PILKA, F. Video Traffic Prediction Using Neural Networks, *Acta Polytechnica Hungarica*, Budapest, Hungary, ISSN 1785-8860, Vol.5, No.4, 2008, pp.59-78

[9] MRAČKA, I., ORAVEC, M. Classification of Traffic of Communication Networks by Multilayer Perceptron, *Proc. of International Conference New Information and Multimedia Technologies NIMT–2008*, September 18-19, 2008, Brno, Czech Republic, ISBM 978-80-214-3708-1, pp. 46-49

[10] PROCHASKA, J., VARGIC, R.: Using Digital Filtration for Hurst Parameter Estimation. *Radioengineering*, Vol. 18, No.2, June 2009, pp. 238-241

Application of Psychoacoustic Principles on a Sinusoidal Model

Ivan MINÁRIK, Martin TURI NAGY [1]

[1] Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
xminariki@stuba.sk, turi@ktl.elf.stuba.sk

Abstract. *This paper deals with the psychoacoustic principles applied on the parameters obtained by analysis of speech signal with sinusoidal and noise model (the SN model). An SN (sinusoids plus noise) model is a spectral model, in which the periodic components of the sound are represented by sinusoids with time-varying frequencies, amplitudes and phases and the non-periodic components are represented by noise. In this article, we focused on the fact that for speech, the number of detected sinusoidal parameters can be reduced by the application of psychoacoustic principles on the sinusoidal parameters.*

Keywords

Psychoacoustics, masking, sinusoidal modeling, SN model.

1. Introduction

In this article, we aimed on a problem that belongs to the topic of speech processing – extending the existing SN model with the psychoacoustic model. In the SN model [1], the sinusoidal part represents the periodic components of the audio signal and the noise part represents the stochastic components of the audio signal. The psychoacoustic model allows us to reduce the number of sinusoidal parameters needed for the reconstruction of the speech without sensible decrease of the quality of the synthesized speech signal.

Other similar systems to the SN model used in speech processing are based on HNM (harmonic plus noise) model [2]. HNM assumes the speech signal is composed of a harmonic and a stochastic part. The harmonic part describes the quasiperiodic components of the speech signal while the noise part describes the nonperiodic components. Although HNM model is more commonly used in speech processing, we chose the SN model because it saves the exact frequencies of the sinusoids, not only the harmonic ones, and we assume better quality of speech when discarding the inessential sinusoids by our psychoacoustic model.

This paper is divided into following sections. First, the SN model is described. Then, the basics of psychoacoustics are depicted. After that, our application of psychoacoustic model to detected sinusoidal parameters is described. At last, the results are shown and the conclusions are discussed.

2. Overview of SN model

In the past, sinusoidal modeling was used in the speech compression and in the audio analysis/transformation/synthesis. In the computer processing of the audio signals, the sinusoids alone were not considered as a sufficient model for the modeling of a wideband audio. Serra [3] was the first who came with an improvement – the residual noise model that models the non-sinusoidal part of the signal as a time-varying noise source. These systems are called sinusoids plus noise systems (SN).

Sounds that are produced by musical instruments or other systems can be modeled as a sum of the deterministic and the stochastic part or, in other words, as a set of sinusoids plus the noise residual [3]. Sinusoidal components are produced by a vibrating system and they are usually harmonic. The residual contains the energy produced by an excitation mechanism and by other components that are not results of periodic vibration.

In the standard SN model the deterministic part is represented as a sum of sinusoids with time-varying parameters. Each sinusoid is a component with time-varying frequencies, amplitudes and phases. The stochastic part is represented by the residual. The whole signal can be written as

$$x(t) = \sum_{i=1}^{N} a_i(t)\cos(\theta_i(t)) + r(t), \tag{1}$$

where α_i and θ_i are amplitude and phase of the sinusoid and *r(t)* is the noise residual, which is represented by the stochastic model. We assume that sinusoids are locally stable. This means, that their amplitudes are not changing too fast and that phases are locally linear. The whole signal is modeled either with an utilization of the sinusoidal and the stochastic model. The residual *r(t)* contains all the components that are not represented by the sinusoidal model.

The human perception is sensitive neither to details of the sound spectral shape, nor to the phase of non-periodic

signals. In assumption that the residual contains only stochastic components, it can be represented by a filtered white noise. The instant amplitude and phase are not saved, but they are modeled by the time-varying filter of the spectral shape, or by short-time energies of fixed frequency bands, i.e. Bark bands, as used in our system.

The design of STN model is shown in Fig. 1. First, the input signal is analyzed, to take time-varying frequencies, amplitudes and phases. Then, the sinusoids are synthesized and subtracted from the original signal to take the residual. Then, residual is analyzed using the stochastic analysis. Afterward short-time energies of Bark bands are computed. In synthesis, the stochastic signal has to be synthetized and added to the synthetized sinusoids to obtain the whole signal.

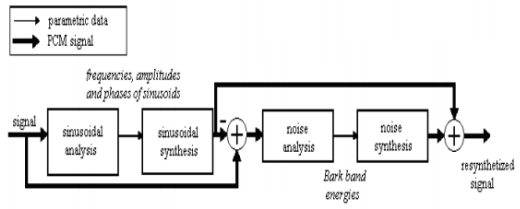


Fig. 1. Design of SN model.

The most complicated part of the system is the sinusoidal analysis. The input signal has to be divided into overlapped and windowed frames. Then the short-time spectrum of the frame is obtained, with an utilization of STFT. The spectrum is then analyzed and peaks are detected. After that, the parameters of the peaks are estimated (frequency, amplitude and phase).

In the sinusoidal synthesis, each frame is synthesized using corresponding sinusoids. The frames are added together using overlap-add synthesis.

We can obtain the residual in the time domain by subtracting the synthetized sinusoids from the original signal. This residual can be represented by the filtered white noise. Because the human ear is not sensitive to variations of energy inside the Bark bands for quasi-stationary signals, the exact spectral shape is not needed. The only information needed are the short-time energies inside the Bark bands.

In the noise synthesis, the complex spectrum is created and random phases for amplitudes (obtained from energies) are generated. Adjacent frames are combined with an utilization of the overlap-add synthesis.

3. Psychoacoustic Principles

In an attempt to reduce the size of parametric file produced by the sinusoidal coder we employ several aspects of psychoacoustics, i.e. the Absolute Threshold of Hearing, theory of the Critical Bands and Simultaneous Masking.

3.1 Absolute Threshold of Hearing

Absolute Threshold of Hearing (ATH) represents minimal level of energy of the pure tone to be audible to the listener in the noiseless environment. The level of energy depends on the frequency of the pure tone. Terhardt [4] has given an empirical function of frequency measured in SPL:

$$T_q(f) = 3.64\left(\frac{f}{1000}\right)^{-0,8} - 6.5e^{-0,6\left(\frac{f}{1000}-3.3\right)^2} + 10^{-3}\left(\frac{f}{1000}\right)^4 [dBSPL] \quad (2)$$

The expression above leads to a graphical representation of the ATH. It can be seen from the Fig. 2 that human ear is most sensitive to frequencies around 4 kHz, while with the frequency rising or decreasing, the threshold rises significantly.

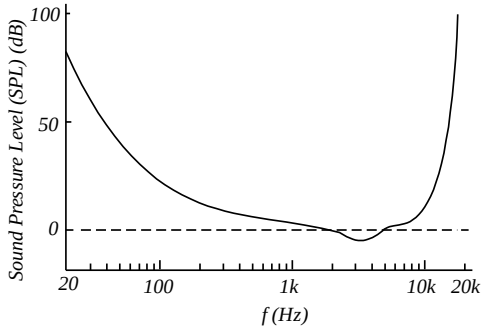


Fig. 2. Absolute Threshold of Hearing across the human ear frequency range

The value of 0 dB SPL is widely used in audio coding algorithms (MPEG L3 for instance) to be associated with +/- 1 bit amplitude (i.e. the smallest possible non-zero amplitude).

3.2 Critical Bands

Several observations of the way the human ear processes sound waves lead to an assumption that human inner ear can be considered, from the signal processing point of view, as a bank of highly overlapping bandpass filters [1]. The filters have non-linear bandwidth: the bandwidth grows with the center frequency of the particular filter. Moreover, the ear is also level-dependent which means subjectively same volume of tones is achieved with different intensity of the tone at different frequencies.

Critical bandwidth is a band in which the energy of the sound (represented by the noise) remains (relatively) constant. This means that if there are more distinct sounds present in one critical band they won't be audible unless they exceed the energy level. Outside of the band level changes will be audible.

3.3 Masking

Masking in signal processing is referred to as an effect in which one sound appears inaudible due to

presence of another sound. This effect appears both in time (non-simultaneous) and frequency (simultaneous) domain.

Simultaneous masking occurs when a frequency with low amplitude exists near frequency with considerably higher amplitude. The masking is frequency dependent with the width of the band in which masking occurs raising towards higher frequencies. This is because human ear is more sesitive to frequency changes in the lower parts of the frequency range.

3.4 Masking Threshold

According to Zölzer [6], masking threshold performs masking (covering) of frequency components that occur below the masking threshold rendering the components inaudible. A principal description is given in Fig..

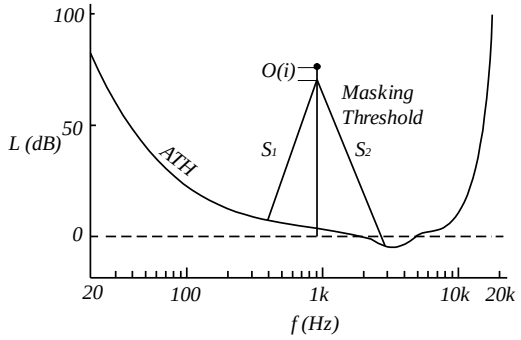


Fig.3. Masking Threshold

The slopes of the threshold are calculated depending on the masking frequency component's intensity and frequency. The lower slope S_1 and upper slope S_2 are given by equations

$$S_1 = 27\left[\frac{dB}{Bark}\right], \tag{3}$$

$$S_1 = 24 + 0.23\cdot\left(\frac{f_{c_i}}{kHz}\right)^{-1} - 2\cdot\frac{S_p(i)}{dB}\left[\frac{dB}{Bark}\right], \tag{4}$$

where f_{c_i} is the center frequency and $S_p(i)$ is the SPL of the frequency in the band i, further described later on.

4. Application of Psychoacoustic Model

In this section we will describe the application of the mentioned psychoacoustic principles to the parametric file created by the sinusoidal coder.

All of the techniques described in Chapter 3 are usually applied to signal transformed by FFT into the frequency domain. However, individual frequency bins represent specific frequencies with the same distance between each other, and their count is given by how many points does FFT use. This observation gives us means to assume that our set of frequencies with their amplitudes and phases in a frame are actually bins similar to ones given by the FFT, only not equidistant. Thus, all of the psychoacoustic principles may be applied to our representation of the frequency spectrum.

The parametric file contains information about sinusoids detected in each frame, one sinusoid per row, in the following format:

"Sine" Frequency [Hz] *Amplitude* <0 – 1> *Phase* $\left[-\frac{\pi}{2} to \frac{3}{2}\pi\right]$

Each frame is separated using defined keywords.

Since the psychoacoustic model is being developed in Matlab environment, the parametric file is first loaded into the memory. Each sinusoid's amplitude is then compared to value of the ATH at the amplitude's frequency. All amplitudes (X) are converted to dB scale in order to be used in the next computations using the scale

$$X_{dB} = 20\cdot\log_{10}\left(X\cdot 2^{16}\right) \text{ [dB]} \tag{5}$$

Any amplitude that falls below the ATH threshold (1) is set to 0, making it inaudible.

After application of ATH, remaining amplitudes are processed in terms of masking. The algorithm works as follows. At the beginning of each iteration frequency with the highest amplitude is found. For this frequency, a critical bandwidth is calculated from which the power of the band, $S_p(i)$, is computed by counting all frequencies' squared amplitudes within the critical band according to equation

$$S_p(i) = 10\cdot\log_{10}\left(\sum_{f=f_{l_i}}^{f_{u_i}} X^2(f)\right) \text{ [dB]}, \tag{6}$$

where f_{u_i} and f_{l_i} are the upper and lower boundaries of the critical band and $X(f)$ is amplitude at the frequency f.

Then, masking threshold is calculated for the critical band's center frequency. Since all the calculations are performed in units of Barks, we employed a frequency-to-bark scale converter based on Zwicker's approximation

$$Z_b(f) = 13\cdot\arctan(0.00076\cdot f) + 3.5\cdot\arctan\left(\left(\frac{f}{7500}\right)^2\right). \tag{7}$$

Next, top of the masking threshold is calculated as an offset between the center frequency's amplitude and the highest point of the masking threshold (point where the two slopes meet at the center frequency)

$$T_T = 10^{\frac{S_p(i)-O(i)}{10}} \text{ [dB]}, \tag{8}$$

where

$$O(i) = \alpha\cdot(14.5 + X_{dB}(f_c)) + (1-\alpha)\cdot 5.5 \tag{9}$$

is an offset between the actual amplitude and top of the threshold. The tonality index α determines whether the

masking signal is noise-like (α = 0) or tone-like (α = 1). Obviously, for the sinusoids we use the latter.

The equations (3) and (4) lead to analytical expressions

$$S_1(f) = 27 \cdot (X_{dB}(f) - X_{dB}(f_c)) + T_T + X_{dB}(f) \quad (10)$$

for the lower slope and

$$S_1(f) = (24 + 0.23 \cdot f_c) \cdot (X_{dB}(f_c) - X_{dB}(f)) + T_T + X_{dB}(f) \quad (11)$$

for the upper slope.

After masking threshold is calculated we can proceed to comparison of the masking threshold with amplitudes that lay within the frequency band. Any amplitude that rises below the value of the masking threshold at the amplitude's frequency is discarded (set to 0). Thus, we only keep sinusoids that are either masking or rise above the masking threshold in the critical band.

Iteration continues until all of the sinusoids in the frame are marked masking or above masking threshold or set to 0.

5. Results

The utilization of psychoacoustic model was evaluated on several speech signals. Besides subjective listening tests, objective evaluation was considered, too. The Tab.1 below shows the difference of ODG (Objective Difference Grade) between the sinusoidal singal and the sinusoidal signal with applied psychoacoustic model. Tab. 2 shows results of EHS (Harmonic Structure of Error), NMR (Noise to Mask Ratio) and ADB (Average Disturbed Frames).

Tab. 1. Difference in ODG between sinusoidal and pyschoacoustic-sinusoidal signal

	ODG difference (sinusoidal vs. psychoacoustic-sinusoidal)
Voice 1	0.32
Voice 2	0.33
Voice 3	0.24

Tab. 2. EHS, NMR and ADB

	EHS	NMR	ADB
Voice 1 (orig. vs. sinusoidal)	0,235	-0,7787	2,6889
Voice 1 (orig. vs. psychoac.-sin.)	0,1537	-0,4561	2,7712
Voice 2 (orig. vs. sinusoidal)	0,1657	-0,7118	2,6735
Voice 2 (orig. vs. psychoac.-sin.)	0,1387	-0,3921	2,7396
Voice 3 (orig. vs. sinusoidal)	0,1874	-0,8547	2,7262
Voice 3 (orig. vs. psychoac.-sin.)	0,0884	-0,4386	2,8332

6. Conclusion

The advantage of applying psychoacoustic model to the sinusoidal parameters has been discussed in this article. Frequency masking was used to decrease considerably the number of sinusoidal parameters needed for the proper reconstruction of the signal. The subjective and objective listening tests have shown small degradation of quality of the signal reconstructed by the set of sinusoids chosen by psychoacoustic model. In the subjective listening tests, there was very slight difference between the sinusoidal and psychoacoustic-sinusoidal signal.

Acknowledgements

This work has been supported by the projects VEGA 1/0718/09 and FP7-ICT-2011-7 HBB-Next.

References

[1] LEVINE, S.N. Audio Representations for Data Compression and Compressed Domain Processing, PhD thesis, Stanford University, 1999

[2] STYLIANOU, Y. Applying the harmonic plus noise model in concatenative speech synthesis, *IEEE Trans. on Speech and Audio Processing*, vol. 9, no.1, pp.21-29, Jan. 2001

[3] SERRA, X. Musical Sound Modeling with Sinusoids plus Noise. Musical signal processing. 1997, Roads C.& Pope S. & Picialli G. & De Poli G., Swets & Zeitlinger Publishers.

[4] TERHARDT, E.: Calculating virtual pitch, Hearing Research, vol. 1, pp. 155–182, 1979.

[5] SPANIAS, A., PAINTER, T., VENKATRAMAN, A.: Audio Signal Processing and Coding, John Wiley & Sons, Inc., 2007, ISBN 978-0-471-79147-8.

[6] ZÖLZER, U.: Digital Audio Signal Processing, Second Edition, John Wiley & Sons, Inc., 2008, ISBN 978-0-470-99785-7.

SIP PROTOCOL BASED INTELLIGENT SPEECH COMMUNICATION INTERFACE

Gregor ROZINAJ, Róbert GRENČÍK, Lukáš HAJDU, Marián HLAVATÝ, Martin HLUZIN, Patrik HOLLÝ

Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
rozinaj@ktl.elf.stuba.sk, *xgrencikr@stuba.sk, xhajdul1@stuba.sk, xhlavatym1@stuba.sk, xhluzin@stuba.sk, xhollyp1@stuba.sk*

Abstract. *The Intelligent Speech Communication Interface (IRKR) is audio based platform providing dialog services in telecommunication network. System IRKR is built on Galaxy Communicator (GC) platform which is a platform providing the ability to communicate to the management layer for distributed systems and use W3C standard named VoiceXML for creating dialogs.*

Keywords

Intelligent Speech Communication Interface, IRKR, dialog service, SIP, Automated Speech Recognition, Text to Speech, VoiceXML, RTP

1. IRKR system

Today we are experiencing a great expansion of modern technologies of communication through mobile networks and the Internet. This development is progressing well and our faculty is trying to keep pace with the times. Therefore, we are working on a system IRKR, which is a platform providing audio service dialog. To create these dialogues used recommendation specifically called VoiceXML. VoiceXML is a W3C standard designed for creating audio dialogs that support synthesized speech, digitized audio, recognition of spoken and DTMF input, recording of spoken input and telephony. Allows an analogous development of voice applications, such as HTML for visual applications. Like HTML documents are interpreted by visual web browser, VoiceXML documents are so interpreted by voice browsers. Standard architecture is to deploy a group voice browsers connected to the PSTN so that users can interact using a phone with voice applications. VoiceXML uses tags to instruct the voice browser to provide speech synthesis, automatic speech recognition, dialogue management and playback of audio signals. [1]

2. Architecture of IRKR

IRKR system is built on a platform of Galaxy Communicator [2], which development has been currently suspended. Galaxy Communicator offers many features and conveniences but IRKR do not use all of them because they are unnecessary for its functioning. All these disadvantages could be removed by the new concept of IRKR [3]. The main requirement for a new architecture is to propose a direct, native support for VoIP and thereby achieve its easy integrability and interoperability with these systems. This requirement is simple achieved if the proposed architecture is using open standards, which are currently used in VoIP. For internal communication in IRKR system are used VoIP standards to retain a distribution system for servers that will perform certain functionality, thus maintaining system modularity. The advantage of the modularity of the system is also possibility to simply add more servers, which can add further services, or would provide the possibility of using these services to multiple users simultaneously. It follows that the architecture used in the Galaxy Communicator, therefore the presence of one central node and more specific nodes, should be kept.

3. Modules of IRKR

IRKR system uses a star topology [see Fig. 1], which consists of individual modules. They communicate with each other through the hub module and perform different tasks. In the following paragraphs we describe the individual modules [4].

3.1 TTS server

This module is designed to convert text to speech (TTS = text to speech). It is a speech synthesis corpus-based methods. TTS contains a database of quantities for prerecorded samples in WAV format, from which it is based on the input requirements of prosodic speech properties select the most suitable sections of recordings, which are then bundled and exported as a WAV recording on the final output of the server. It is a synthesis of speech in advance of your text.

TTS server communicates with the users through gateway with RTP, but messages and information gets from others modules through hub module. TTS server sends RTP stream.

3.2 ASR server

ASR (Automated Speech Recognition) is a server module for automatic speech recognition. Obtained from the Audio Server patterned speech in the form of 16-bit PCM samples with a sampling frequency of 8000 Hz. The main task is to evaluate these samples to text.

ASR server also communicates directly with users through gateway with RTP. ASR server receive RTP stream. No one other module communicates directly with users.

3.3 HUB

This module is used as a central element for all other system modules IRKR. It connects the individual modules and directs control messages between modules, according to precise rules. Any module cannot communicate with another module without using the HUB.

When module is started it is registering itself on hub module and hub module knows how to communicate with each module. If module cannot register on hub module, it is displayed error message.

3.4 Gateway

This module is responsible for the overall communication with users. He is responsible for management and exchange of useful information. Is able to accept requests to establish a connection with the system IRKR and further sends control information to the hub. Gateway also exchange RTP stream between ASR and TTS module for processing.

3.5 Dialog Manager

The main task of this module is a dialogue with the rules defined in the VoiceXML files. The Dialog Manager receives information from modules that communicate with the user and ensures that on the output to the user is the correct answer. The output of the commands for the modules is defined in the VoiceXML file. Dialogue manager is the brain of the system that collects inputs and issuing orders.

3.6 Backend

This module is designed to obtain information from external sources, mainly from the websites. This information is then provided to application defined in the system IRKR. Backend consists of a main application that takes care of communication with other parts of the system and Dynamic link libraries (DLL) representing particular services, such as arrival and departure of trains, respectively weather forecast. It is important that the server knows to deal with small changes to the pages that happen very often. Of course, those great and fundamental changes cannot be processed and will be always incorporate the following changes to the program. Theoretically, it would be possible to make a set of parameters that should be evaluated and can be easily changed.

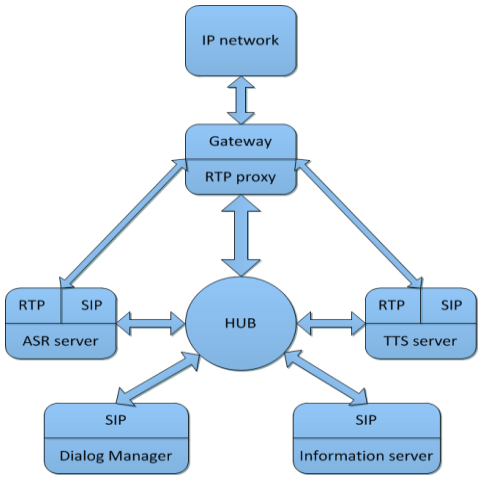


Fig. 1. Star topology of IRKR

4. IRKR and SIP protocol

The Session Initiation Protocol (SIP) is an IETF-defined signaling protocol, widely used for controlling multimedia communication sessions such as voice and video calls over Internet Protocol (IP) [6].

SIP is text-oriented application-layer protocol, based on operating requirements and response. Application relates from the client to the server or to another client and answers are in response to a specific requirement. The format of the SIP messages is based on HTTP1.1. SIP employs many headers and the same number of the response codes as the HTTP. SIP is not extending of the HTTP protocol. SIP messages consist of two parts, the header part and the body part. Header section contains important information for the processing of SIP requests and responses. The SIP message body may define any text.

SIP main responsibilities are follows:

- SIP provides the basic signaling between participants to set up the session.
- SIP uses the Session Description Protocol (SDP) to classify the nature of the communication utilize within session.
- SIP uses the suitable protocol to convey information in the session.

4.1 SIP functions

The biggest advantage of SIP is his flexibility which allow developers and other vendors to satisfy their special needs.

SIP provides five key functions [7]:

- ***user location*** - SIP determines user locations by a registration process. When a SIP client is activated on a PC or laptop, it sends out a registration to the SIP server announcing availability to the communications

network. Voice over-IP (VoIP) phones, cellular phones, or even complete teleconferencing systems can be registered as well. Depending on the registration point chosen, there may be several different locations registered simultaneously.

- ***user availability*** - User availability is simply a method of determining whether or not a user would be willing to answer a request to communicate. If you "call" and no one answers, SIP determines that a user is not available. A user can have several locations registered, but might only accept incoming communications on one device. If that is not answered, it transfers to another device, or transfers the call to another application, such as voice mail.
- ***user capabilities*** - With all the various different methods and standards of multimedia communications, something is needed to check for compatibility between the communications and the users' capabilities. For example, if a user has an IP phone on their desk, a white-board conference via that device would not work. This function also determines which encryption/decryption methods a user can support.
- ***session setup*** - SIP establishes the session parameter for both ends of the communications - more specifically, where one person calls and the other answers. SIP provides the means to setup and/or establish communications.
- ***session management*** - This function provides the greatest amount of user awe. Provided a device is capable, a user could transfer from one device to another - such as from an IP-based phone to a laptop - without causing a noticeable impact. A user's overall capabilities would change - such as being able to start new applications such as white-board sharing - perhaps affecting the voice quality temporarily as SIP re-evaluates and modifies the communications streams to return the voice quality. With SIP session management, a user can also change a session by making it a conference call, changing a telephone call to a video conference, or opening an in-house developed application. And finally, SIP terminates the communications.

4.2 SIP components

- *User Agent (UA)* - SIP network terminals (phones, gateways). May act as a server (User Agent Server - UAS) and also as a client (User Agent Client - UAC).
- *Registration server* - a SIP server, which only accepts registration requirements posted by the user. Registration server don't send further requirements.
- *Localization server* - a server that provides information to proxy redirect server about possible current locations of users.
- *Redirect server* - SIP server, which provides an address service mapping. Redirect server accepts calls and don't send requirements.
- *Proxy server* - a server that acts as a user agent server with forwarding SIP requests and as a client server to other SIP servers with the translation of forwarded requests.

4.3 SIP messages

Communication using SIP comprises series of messages. Messages can be transported independently by the network [8]. Usually they are transported in a separate UDP datagram each. Each message consist of "first line", message header, and message body. The first line identifies type of the message. There are two types of messages - requests and responses[link]. Requests are usually used to initiate some action or inform recipient of the request of something. Replies are used to confirm that a request was received and processed and contain the status of the processing.

4.3.1 SIP requests

- *INVITE* – message used to establish a session. This message also carry several important information such as: IP address and ports of caller and callee, dialog identifier to differentiate other sessions, description of the media type accepted by the sender and encoded in SDP, etc.
- *ACK* – message acknowledges receipt of a final response to INVITE. Establishing of a session utilizes 3-way hand-shaking due to asymmetric nature of the invitation. Callee's user agent periodically retransmits a positive final response until it receives an ACK (which indicates that the caller is still there and ready to communicate).
- *BYE* – these messages are used to tear down multimedia sessions. A party wishing to tear down a session sends a BYE to the other party.
- *CANCEL* – message is used to cancel not yet fully established session (typically when a callee doesn't respond for some time).
- *REGISTER* - Purpose of REGISTER request is to let registration server know of current user's location. Information about current IP address and port on which a user can be reached is carried in REGISTER messages.

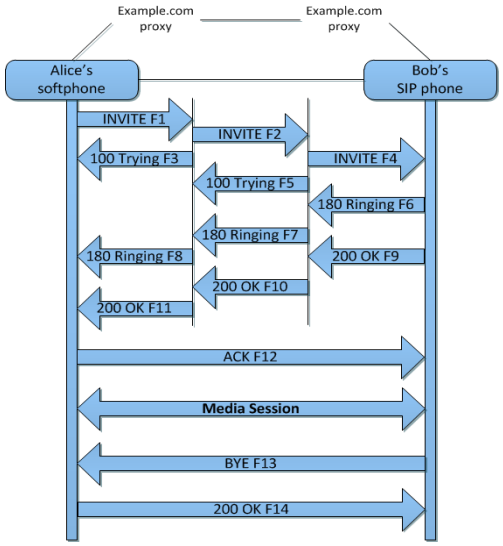


Fig. 2. Example of establishing SIP session.

4.3.2 SIP responses

When a user agent or proxy server receives a request it send a reply. Each request must be replied except ACK.
The reply code is an integer number from 100 to 699 and indicates type of the response. There are 6 classes of responses:

- **1xx** are provisional responses. A provisional response is response that tells to its recipient that the associated request was received but result of the processing is not known yet.
- **2xx** responses are positive final responses. A final response is the ultimate response that the originator of the request will ever receive. Therefore final responses express result of the processing of the associated request. Final responses also terminate transactions.
- **3xx** responses are used to redirect a caller. A redirection response gives information about the user's new location or an alternative service that the caller might use to satisfy the call. Redirection responses are usually sent by proxy servers.
- **4xx** are negative final responses. a 4xx response means that the problem is on the sender's side. The request couldn't be processed because of bad syntax or cannot be fulfilled at that server.
- **5xx** means that the problem is on server's side. The request is apparently valid but the server failed to fulfill it. Clients should usually retry the request later.
- **6xx** reply code means that the request cannot be fulfilled at any server.

5. Use of IRKR

Usefulness of this system is in many areas of human life. We can divide the biggest advantage of this system into two main areas. First one is to help simplify human-computer communication and second is to provide useful information to human communicating with this system.
First goal, simplify human-computer communication, will help people with different disabilities such as loss of vision or movement disorders.
For these people IRKR will be an invaluable tool in performing common activities during the day. They will be able to control their computers by their voice and give them commands to order for example pizza or some kind of other goods. Using IRKR they can fully benefit from advantages which todays world give us. They will be able to use all other services which they couldn't because of limitation of their medical condition.

The second goal, provide useful information, can improve accessibility of information to people which can't look up these information by theirselves. This service can be accessible on general known phone number. After dialing this number IRKR will ask some general question to find out what information caller needs. The system must be clear and easy to use. In addition to all the functionality of the resulting synthesis is important and also its clarity. This creates a database which contains a set of frequently used words and sounds, of which is synthesis generated. In creating a database we must be sure to select the right speaker. His voice must be clear and understandable.

Overall, this system will be used frequently in the future and its needs grow.

Acknowledgement

This work has been supported by the projects VEGA 1/0718/09 and FP7-ICT-2011-7 HBB-Next.

References

[1] COPJAN, P., DUJSIN, J., FLOREK, I., MAKOVINYI, P. Integration of multimedia services in NGN SIP IRKR. *Team project. 2010.*

[2] SOURCEFORGE. Galaxy Communicator Documentation. [online]. [20.3.2011]. Actualized 24-9-2003. Available on the Internet: *<http://communicator.sourceforge.net/sites/MITRE/distributions/GalaxyCommunicator/docs/manual/index.html>*.

[3] VLASAK, J. Utilization of SIP protocol for IRKR communication. *Diploma thesis, 2008.*

[4] JANIK, M. Integration of SIP into the modules of IRKR system. *Diploma thesis, 2008*.

[5] SINNREICH , H., JOHNSTON, B. A. Internet Communications Using SIP. *Indiana: Wiley Publishing, Inc., 2006.*

[6] NETWORK WORKING GROUP, Request for Comments: 3261, SIP: Session Initiation Protocol. [online]. [5.4.2011]. Actualized 06-2002. Available on the Internet: *<http://tools.ietf.org/html/rfc3261>*.

[7] MAGALHAES, M. R. Session Initiation Protocol (SIP) and Its Functions. [online].[20.3.2011]. Actualized 22-2-2005. Available on the Internet: *<http://www.windowsnetworking.com/articles_tutorials/session-initiation-protocol-functions.html>*.

[8] IPTEL.ORG. SIP messages. [online].[5.4.2011]. Actualized 2007. Available on the Internet: *<http://www.iptel.org/sip/intro/messages>*

Deterministic and statistical self-similarity

Martin BUNČÁK, Radoslav VARGIC[1]

[1] Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia, vargic@ktl.elf.stuba.sk

Abstract. *This paper describes and compares the terms deterministic and statistical self-similarity. The paper encompasses also basic theory like aggregation and lists properties of statistically self-similar processes. It proposes procedure how to produce process with self-similar properties based on shape of Koch curve. It provides method how to create process which is deterministic, visually self-similar and more visually similar to stochastic process than harmonic signal. For comparison Hurst parameter is estimated for trivial deterministic (harmonic) signal and also for fractional Gaussian noise..*

Keywords

Hurst parameter, self-similarity, Koch curve

1. Introduction

Self-similarity is a phenomenon which can be observed in nature, technology and in various areas of human push. It has multiple forms. Basically we can divide self-similarity as deterministic and statistical.

Deterministic self-similarity is represented by geometric shapes with certain properties. It is not possible to describe these shapes by classical Euclidean geometry. They have that is to say as from definition is implied infinite segmentation. They are thus in contrast to geometrically smooth shapes. Such segmented shapes are referred to as fractals. A stand-alone scientific discipline the fractal geometry deals with the fractal objects. This discipline has intensively developed circa since 1960s. As its founder is considered mathematician Benoit B. Mandelbrot, who as a first man defined the term fractal.

Koch curve [4] (the part of Koch snowflake) and Sierpinski triangle are typical and relatively simple fractals. With use of complex count it is possible to generate iteratively Mandelbrot and/or Julia set. These types of sets are examples of self-similar shapes with optically nice appearance, therefore they are best known too.

Statistical self-similarity [5] is represented in the area of time series, processes and signals most often originating from telecommunication data networks. This self-similarity has gained its name because it describes bursty time behavior of data traffic. It is hardly possible to determine which scale is in matter by pure look at the graph of data traffic. Thus, data traffic is statistically self-similar as long as it is scale invariant like deterministic self-similar objects.

Power of statistical self-similarity can be quantified. As a quantifier serves us a parameter of self-similarity - so called Hurst parameter (denoted as H). Hurst parameter can be estimated by various means. Hurst parameter for self-similar signal should belong into interval <0.5, 1). However some methods can estimate values even higher for higher degree self-similarity. Example of data traffic is shown at Fig. 1. We can see burstiness of data even at large scale specifically at time segment longing 1000 s. When we zoom the data the shape of traffic signal is similar. The traffic resembles at itself on other (smaller) scale - so it is self-similar

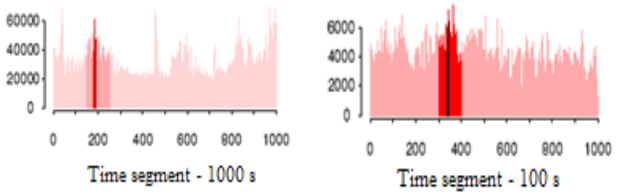


Fig. 1 Data traffic captured at long period of time (277 hours) – left is overall view, right is zoomed the darker part.

1.1 Aggregation of process

For better understanding the procedures depicted in the article we present the well-known definition of aggregated process.

Let $X = \{X_i, i \in Z\}$ be a stationary process. The m-aggregated time series $X^{(m)}$ is defined by averaging the initial series over nonoverlapping intervals with m size and substituting its mean value for each interval, i.e.

$$X_i^{(m)} = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_i \qquad (1)$$

where $k \in Z$.

1.2 Formal definition of statistical self-similarity

For discrete statistically exactly self-similar process **X**(i) with Hurst parameter H holds [5]:

$$\mathbf{X}(i) \overset{D}{=} m^{1-H}\mathbf{X}^{m}(i) \qquad (2)$$

where m = 1,2,3… and sign $\overset{D}{=}$ means equality of distribution functions and/or probability functions. Self-similar processes have certain statistical properties i.e. [5]

I. Uncountable autocorrelation function
II. Infinite power spectral density at 0
III. Hurst effect
IV. Slowly decaying variance

2. Objective

By comparison of both types of self-similarity rises up the question if both types are somehow related. Thus if can be found mathematical function which would map deterministic self-similar objects to bursty behavior of self-similar signal. Our asset is based in work, that we chose simple self-similar shape - Koch curve. We were inspired by its iterative creation. But we did not copy and evaluate the shape by numbers strictly. We create our own way how to introduce pattern of self-similarity into proposed process. Measure of self-similarity of such process should be estimated by common methods such as aggregated variance or wavelet method. And as a last step estimation of other trivial signal will be presented. We present comparison between fractional Gaussian noise, harmonic signal and process designed by us. Fractional Gaussian noise (fGn) [6] is conventionally used model for generating artificial self-similar process.

3. Exploration

Koch curve consists of bigger or smaller shapes, which resemble to equilateral triangles. Koch curve arises out iteratively. The more iterations the more complicated the Koch curve is and consists of triangles of greater count of scales.

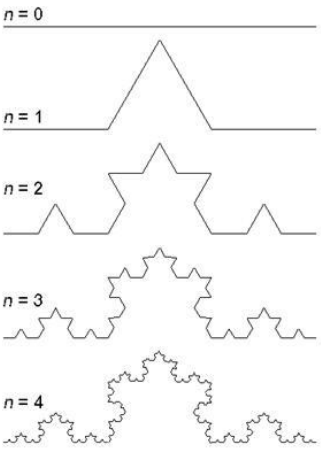


Fig. 2 Iterations of Koch curve

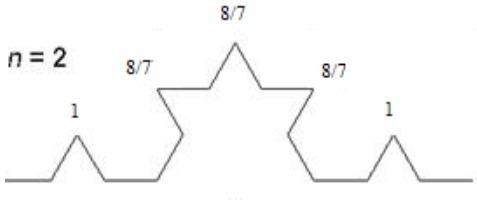


Fig. 3 Assignment of values for each top at 2. iteration

Triangles are created by sides. What a great area will triangle have depends on size of the sides. Each iteration adds sides which length is 1/3 of side length of previous iteration. Between sizes of sides which are indeed abscissas holds multiplicative relation.

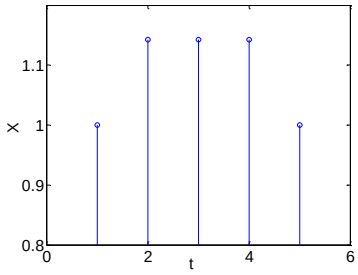


Fig. 4 Basic pattern - second iteration of process inspired by Koch curve

3.1 Procedure of generation our process

We started from the second iteration. Our basic pattern is at **Fig. 4**.and is related to 2. iteration. We were inspired by 2. iteration of Koch curve. Tops of the curve were evaluated as is shown at **Fig. 3**. Then basic pattern was concatenated three times. Then we multiplied each value of three concatenated patterns by term $(8/7)^{iter-1}$ where *iter* means the number of iteration. We denoted this three times concatenated and multiplied pattern as middle of the 3rd iteration. Before the middle of the 3rd iteration is the beginning and after is the end of 3rd iteration. The beginning of 3rd iteration is the vector of 2nd iteration i.e. our basic pattern. The end of 3rd iteration is also the vector of 2nd iteration i.e. our basic pattern. The 3rd iteration vector is shown at **Fig. 5**.

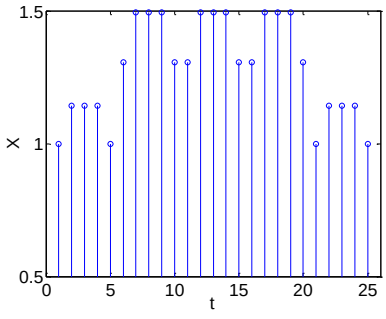


Fig. 5 Third iteration of process inspired by Koch curve

Whole iterative procedure can be summarized in x steps:

I. The vector of the k^{th} iteration is concatenated three times and multiplied by term $(8/7)^{k-1}$ so we get the middle of $(k+1)^{th}$ iteration
II. The beginning of vector of the $(k+1)^{th}$ iteration is the vector of the k^{th} iteration
III. The end of vector of the $(k+1)^{th}$ iteration is the vector of the k^{th} iteration
IV. The beginning , the middle and the end of the $(k+1)^{th}$ iteration are concatenated so we got the vector of the $(k+1)^{th}$ iteration

By this procedure we obtained value affinity across all scales. We wanted to as accurately as possible simulate data traffic where the differences in network infrastructure components loads are not so great differences in process values as they would be great if we, for instance, use multiplying integer constant. Sequence of values obtained by above mentioned procedure was further assessed as a stochastic process. This process can be easily made zero-mean process by subtracting the mean value.

Process is nothing more than sequence of numbers regularly identified in time. Process generated by this procedure can be made stochastic by initialization – we will start the realization at randomly chosen coefficient and use the following coefficients as process samples. This procedure produces wide sense stationary data.

Autocorrelation function (ACF) of such process is slowly decaying with regular cycles induced by process determinism. The periodic shape of ACF reveals the deterministic process. The used methods for determination of degree of self-similarity (Hurst parameter value) are unable to distinguish between the stochastic and/or deterministic input process. They assume certain statistical properties and do not test them. They assume second order stationarity of process. For self-similar processes holds following equation.

$$r^{(m)}(k) = r(k) \tag{3}$$

for all m=1, 2, 3… (k = 1, 2, 3). It means that ACF of aggregated process by level of *m* and ACF of original process are equal. In case of process created by us the ACFs are not equal so the formula (3) does not hold. The situation can be seen at **Fig. 6** By this way we can conclude that proposed process is not statistically self-similar.

3.2 Estimations

But are processes generated using Koch curve more self-similar according mentioned methods than some trivial nonstationary process or signal? We choose discrete harmonic sinusoidal signal as a test data. The harmonic signal is not stochastic rather deterministic. And we estimated Hurst parameter for such signal and for one realization of fGn which was generated by Paxson generator [6]. Results are in **Tab. 1**

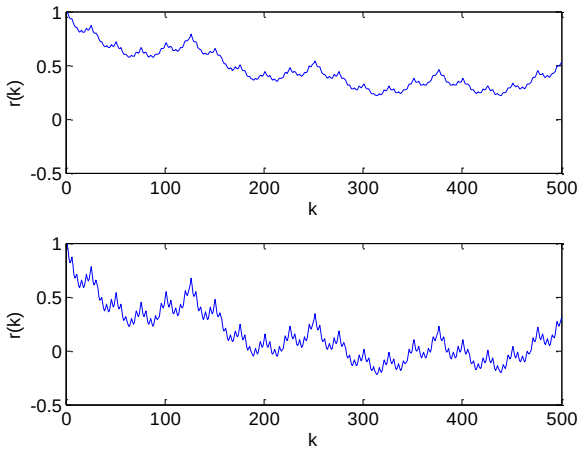


Fig. 6 Beginning of ACF of process generated by 7 iterations (ipper image) and aggregated by 5 (lower image). Displayed first 500 coefficients.

Method	Kettani	VT	AV
Harmonic	1.0000	0.9998	2.8420
fGn	0.8854	0.8793	0.9102

Tab. 1 Hurst parameter estimation of harmonic signal

Hurst parameter was estimated for generated process. For estimation we used 3 various methods aggregated variance [2], wavelet method by Abry and Veitch [1] and Kettani method [3]. Koch curve generated by seven iterations provided a process which was estimated as highly self-similar (H > 0.9 by all three methods) as we can see in **Tab. 2**.

Iteration	5.	6.	7.
Kettani	0.9759	0.9864	0.9911
VT	0.9656	0.9809	0.9877
AV	1.1488	1.2080	1.2019

Tab. 2 Hurst parameter values for several iterations of our process.

We see that Hurst parameters gained by estimating harmonic signal are even higher that Hurst parameters gained by estimating process generated by us. The realization of fGn process is estimated with lowest Hurst parameter of all examined signals.

All three signals and corresponding ACFs and power spectral densities (PSDs) are displayed at **Chyba! Nenašiel sa žiaden zdroj odkazov.**, **Fig. 7** and **Fig. 8**. We can see that our proposed process has visual appearance and properties somewhere between strictly deterministic harmonic signal and stochastic exactly self-similar fGn signal.

We would like to mention that iteration of proposed process is in some terms dual operation to aggregation. For example 8. iteration of our process aggregated by 5 has the same length as 7. iteration of the same process. Shape is

also the same but values of aggregated process of higher iteration are higher. ACFs of both processes displayed at **Fig. 10** are very similar, but picture on top is somewhat noisier.

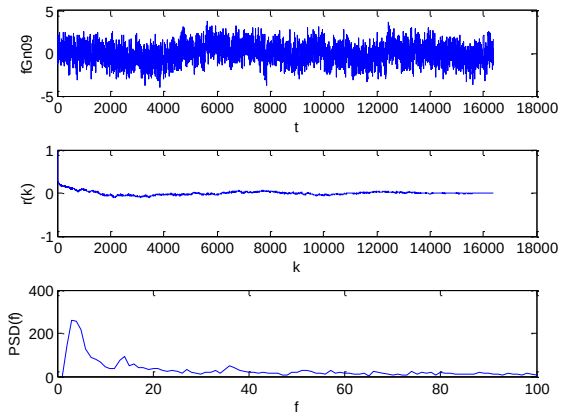


Fig. 7 FGn, its ACF and its PSD

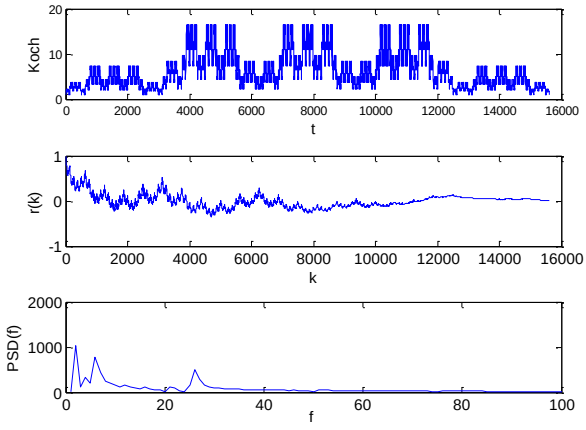


Fig. 8 Our process in time, its ACF and its PSD

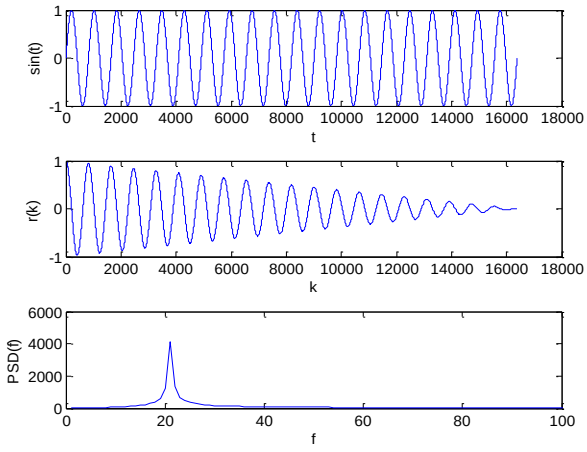


Fig. 9 Harmonic signal, its ACF and its PSD

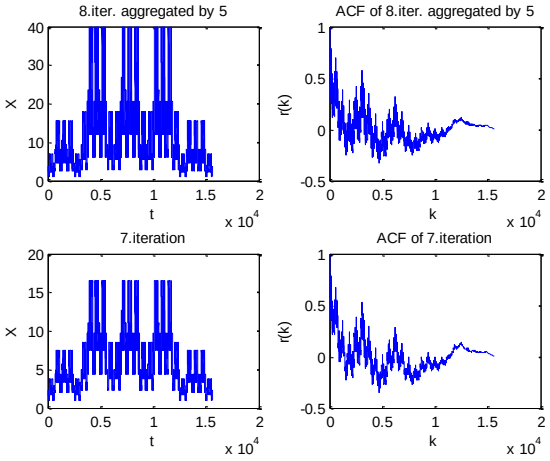


Fig. 10 Partial dualism between aggregation by 5 and iteration

4. Conclusion

If we want to create process which would be estimated by many accessible methods as self-similar, we can be inspired by Koch curve. There is one advantage by using our procedure for generating processes inspired by Koch curve - the shape. ACF and PSD of such process are more similar to stochastic fGn than harmonic signal is similar to fGn. We can confirm that determinism is a property which makes estimators indicate significantly higher Hurst parameter than would be provided by estimation of regular stochastic signal like fGn.

Acknowledgements

Research described in the paper was financially supported by the Slovak Research Grant Agency (VEGA) under grant 1/0720/09 and 1/0214/10.

References

[1] VEITCH, D., ABRY, P. A wavelet-based joint estimator of the parameters of long-range dependence: *IEEE Transactions on Information Theory*, 1999, vol. 45, no. 3, p. 878-897.

[2] TAQQU, M., TEVEROVSKY, V. On Estimating the Intensity of Long-range Dependence in Finite and Infinite Variance Time Series. Published in ADLER, R.., FELDMAN, R.. *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*. Birkhauser, Cambridge MA, 1998, ISBN 0-8176-3951-9, p. 177-217.

[3] KETTANI, H. A Novel Approach to the Estimation of the Long-Range Dependence Parameter. Dissertational thesis, University of Wisconsin – Madison, 2002

[4] MANDELBROT, B. B. *The Fractal Geometry of Nature*. W. H. Freeman and Company,1982, ISBN 0-7167-1186-9.

[5] LELAND, W., TAQQU, M., WILLINGER, W. On the Self-similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Transactions on Networking*, 1994, vol. 2, no. 1, p. 1-15.

[6] PAXSON, V., Fast, Approximate Synthesis of Fractional Gaussian Noise for Generating Self-Similar Network Traffic. *Computer Communications Review*, 1997, vol. 27, no. 5, p. 5-18.s

Usage of method “double spectrogram” for detection and identification of tones in acoustic signals

Peter GRAMBLIČKA, Radoslav VARGIC [1]

[1] Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
bc.gramblicka.peter@gmail.com, vargic@ktl.elf.stuba.sk

Abstract *This article contributes to tonal detection and identification in audio signals. The goal is to write the input acoustic signal to the notes. Notation is today very actual theme, since it was published a number of methods to convert input acoustic signal into musical score. In this contribution was proposed a method "Double spectrogram, which also solves the problem of detection and identification of tones. A comparison of double spectrogram method with selected reference methods is given.*

Keywords

Tone detection, Short-Time Fourier Transform. Double spectrogram

1. Introduction

Since the discovery of Fourier Transform [1,3] much time has passed and now is a lot of different methods (from simple to complex application solutions) [8] . Using these methods are extracted notes from the input acoustic signal. The article will mention only four scientific reference methods. These methods have been published relatively recently. Moreover, these methods use different approaches for extracting notes. For better lucidity, we use the author names of these methods: Dixon [4], Raphael [5] and Marolt [7], Monti and Sandler [6]. Section 3 contains a proposal method for detection and identification of tones. The entry of this method is the acoustic signal and the output is time-frequency graph. Graph can be converted to MIDI format. The method uses two spectrograms and therefore was named "Double spectrogram method”. Section 4 discusses methodologies for evaluation of the reference methods and the method Double spectrogram. The last section of the article is devoted to results and comparison of these methods.

2. Reference methods for detection and identification tones

In this section, only the four latest methods are mentioned. These methods are not similar and use the different notation systems.

2.1 Dixon

Dixon [4] used a standardized approach in the signal processing. The first phase of the process of detection is low filtering and the subsampling signal (12kHz). The second phase is to create time-frequency representation using the STFT [1,3] and obtain power spectrum and from power spectrum also spectral peak extraction. Adaptive peak-picking algorithm was used.

2.2 Raphael

This approach [5] is based on HMM (Hidden Markov Model) trained with the likelihood model. HMM serves on Statistical pattern recognition and machine learning for structures.

2.3 Monti and Sandler

This polyphonic note recognition system [6] uses a Fuzzy Inference System (FIS) as part of the Knowledge Sources (KSs) in a Blackboard system. Blackboard model arrangement containing hierarchy of data abstraction level and KSs dictate advancement and is activated by Scheduler.

2.4 Marolt

As the fourth reference approach, based on neural network, given in this presentation, is article by Marolt [7]. New model based on networks of adaptive oscillators was proposed to partial tracking and note recognition.

3. Double spectrogram method

Detection and identification of tone is a complex problem. Not sufficient to determine the size of notes (frequency), but also is important when the tone starts and when it ends. Both of these problems, I have solved the proposed method Double spectrogram. Method uses two spectrograms. Two, because as the uncertainty principle states, that the tone can not locate precisely in time and in frequency. Therefore, we use a first spectrogram with a short window to locate the

tone and the second spectrogram with a long window to accurate detection of tone.

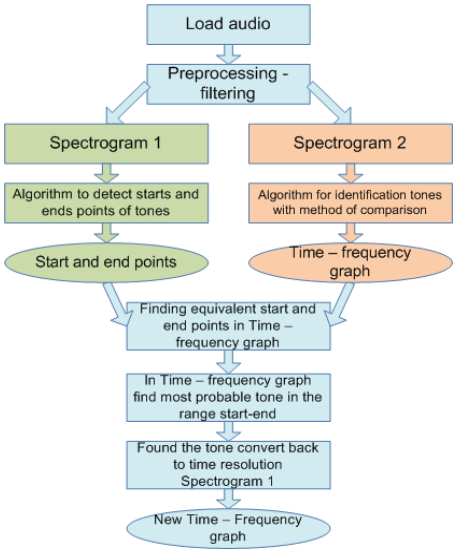


Fig. 1. Main steps in the Double Spectrogram Method

3.1 Loading and preprocessing audio signal

Double spectrogram algorithm starts downloading wav signal. Preprocessing includes filtering band-pass filter with band redundancies 27 Hz - 3980 Hz, which is an ordinary piano (from A2 to h4) for temporal tuning [2].

3.2 Algorithm to detect starts and ends points of tones

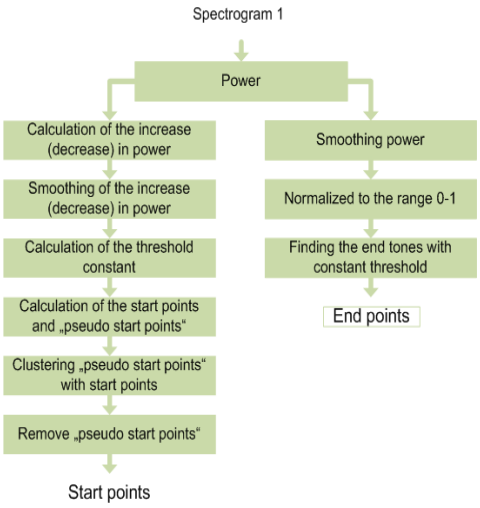


Fig. 2. Algorithm to detect starts and ends points of tones

Input to the algorithm is the spectrogram STFT (Short-Time Fourier Transform) [1,3] and its notation is as follows:

$$STFT(\omega,\tau)=\sum_{n=-\infty}^{\infty}x(n)w(n-\tau)e^{-jk\omega} \quad (1)$$

Using Spectrogram calculated for each time slot n power using the formula:

$$E(n)=\sum_{k=1}^{N/2}\left|STFT_{n,k}\right|^2 \quad (2)$$

where n is the nth frame of the STFT spectral matrix in absolute value, N is the number of samples and k are elements in the column. The result is a vector that reflects the power at each point in time (in each frame).

3.2.1 Part for start points

This section is shown on Fig. 2 in the "left branch". Essence of the algorithm is based on measuring the increase or decrease in power. This increase is simply calculated by the formula:

$$I_{increase}(n)=E(n)/E(n-1) \quad (3)$$

where I is the ratio between the actual value of the power E(n) of the previous power value E(n-1).

After part of the Calculation of the power increase is at least second part and that is the smoothing of the increase in power. Smoothing function is a convolution previous result with the Hamming window. This method is adaptive and therefore followed the third part: Calculation of the Threshold constant. This constant is calculated by $cons=1{,}2*median(v)$. Another part is: Calculation of the start points and pseudo start points. These points are obtained if they meet the condition that the first value in function $I_{increase}$ is greater than threshold constant. Now it is necessary to remove pseudo start points using part: Clustering pseudo start points with start points. Clustering is performed by convolution previous result with rectangular window. The last step is to Remove pseudo start points. From the group are selected only samples that correspond to the size of a half- length rectangular window.

3.2.2 Part for end points

Primary method of detection end points is already used results from the detection start points, using logic for mono. If the tone began at the time n, the previous tone finished in time n-1. The problem occurs, when a break between tones is. This method is solution for this problem. Part Power is the same as for the start points of tones. End points detection is different and its algorithm is shown in "right branch" on Fig. 2. Firstly is part Smoothed Power, where we use convolution to smooth over the energy. Second, normalization is applied to the previous result and output values are in the range from 0 to 1. The last step is to use non-adaptive threshold constants, which indicates that the value in a normalized vector falls below the threshold, while in the previous point n-1 is greater than threshold, the point n is declared as the end of the tone.

3.3 Algorithm for identification tones with method of comparison

This algorithm was designed based on the analysis of spectra of different instruments. In this analysis it was found that the spectra are different from instrument to instrument. Most of the current recognition methods used to detect the tones peak-picking algorithm, which monitors

the maximum amplitude in spectrum. In this algorithm, the often occur octave errors, because the highest amplitude in the spectrum doesn´t means that the amplitude of this tone was really played. A nice example of such errors is violin. Therefore, has been proposed algorithm, which worked with the profiles of the various instruments and compared them with the input signal.

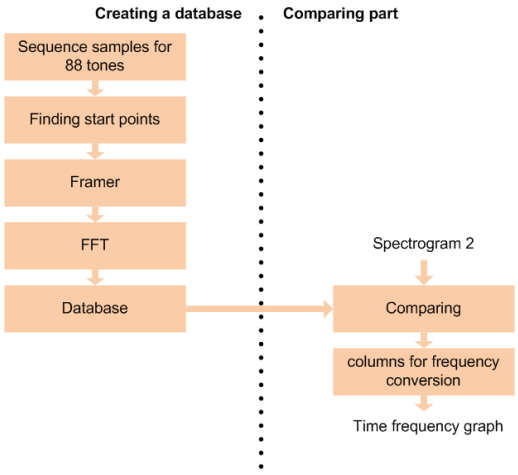


Fig. 3. Algorithm for identification tones with method of comparison

3.3.1 Database Creation

This section is intended to produce a spectral database for example piano. Load sequence of all tones (all 88 keys), from the lowest tone (27.5 Hz) to the highest tone (4186 Hz). On the sequence of 88 tones was performed algorithm to detect the start points. After evaluation, we get the start points every single tone. Now it loads from each start point of N samples, which perform the function of framing. The result of the framing is an output matrix which has N rows and 88 columns tones plus a column of silence. Subsequently, the matrix will be transferred from time to frequency domain using FFT. I can also create a database of spectral components for other instruments.

3.3.2 Comparing part

We create the STFT spectrogram using a long window. Now we will compare STFT spectrogram with the database instruments. Comparison is performed by the method of least squares using the formula:

$$MLS(k) = \sum_{k} \min \left\{ \sum_{j} \left[\sum_{i} \left(|STFT(i,k)| - |database(i,j)| \right)^2 \right] \right\} \quad (4)$$

Where i is the number of rows in the database and STFT the matrix, k is the number of columns in the STFT, j is the number of columns in the database. STFT columns are compared with database columns and calculated is the square deviation between each row. The smallest output value is declared as played tone. The last step is the easiest, because the columns are converted to the corresponding frequency. Now we get the output time-frequency graph.

3.3.3 Finding equivalent start and end points in Time frequency graph

The start and end points of tones, obtained using the algorithm for detecting the start and end tones, are mapped to the closest values that were found in first-time-frequency graph. And so, that we are looking for the nearest equivalent for time-frequency graph. The same goes for the end of the tones.

3.3.4 In Time frequency graph find most probable tone in the range start-end

Another approach is of range start - end of the tone to select only a single tone. Deciding what tone to be selected is based on the number of tones in each scale, which are weighted by the bell function (Hamming window with a length of range).

3.3.5 Found the tone convert back to time resolution Spectrogram 1

Using reverse mapping are tones, which were obtained from the previous step mapped to the resolution of first spectrogram. Through this process we get the exact frequency and time resolution.

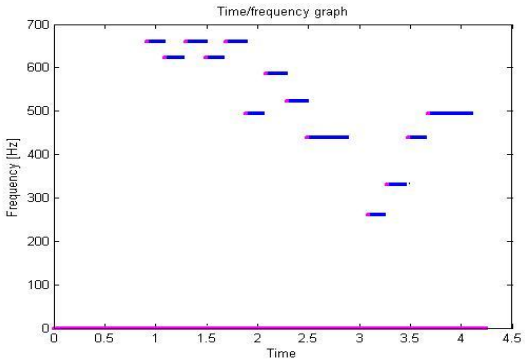


Fig. 4. Output time-frequency graph, we get the double spectrogram method. Blue lines are tones, magenta points are start points

4. Evaluation methods

4.1 Evaluation method for reference methods

For reference methods Dixon and Monti and Sandler have used the same evaluation formula:

$$score = \frac{N}{FP + FN + N} \quad (5)$$

where N=number of correctly transcribed notes; FP(False positive)=number of transcribed notes that weren´t played in original MIDI; FN(False Negative)=number of not transcribed notes.

Other method used by Raphael is error rate from speech evaluation of "Word Error Recognition Rate"

$$ErrorRate = 100 * \frac{Inspertions + Deletions + Subtitutions}{TotalWordsInTruthSentence} \quad (6)$$

An evaluation method for Marolt is unknown.

4.2 Evaluation method for Double spectrogram

	Double spectrogram method
number of notes played with one key	132
correctly detected one note	130
undetected note	2
false duplicity notes	2
number of chords played with two keys	9
correctly detected at least one note	9
undetected note	0
false duplicity notes	3
number of chords played with three keys	43
correctly detected at least one note	42
undetected note	2
false duplicity notes	2
total number of notes and chords	184
total number of undetected notes	4
total number of false duplicity notes	7
total number of notes detected	187

Tab. 1. detailed analysis of the results for the method of double spectrogram

Evaluation of detection is measured by three different music tracks from various artists (Beethoven - For Elizabeth, Mozart - A Little Night Music, Haydn - Deutschlandlied) played an acoustic grand piano. The evaluation method used in Double spectrogram is taken with (6). For mono signal is error rate:

$$ErrorRate_{mono} = 100 \times \frac{4+7}{184} = 5{,}98\% \quad (7)$$

And for poly signal is error rate:

$$ErrorRate_{poly} = 100 \times \frac{4+7+96}{279} = 38{,}35\% \quad (8)$$

Where number are from Tab. 1 and other number are explanation here: 96 is number undetected notes (keys), 279 is total number played notes (keys) for poly signal.

5. Results and comparison with reference methods

Tab. 2. contains the results achieved by different methods. The results of reference methods are obtained from the scientific articles [4], [5], [6] and [7]. Various approaches have been proposed for the notation of audio signal: standard S.P. (Signal Processing) techniques, HMM; blackboard algorithm, neural networks and double spectrogram. Common mistakes are octave, rapid passages, and quiet notes. It was very difficult to compare different approaches, which include various evaluation techniques. Difficulties were also lack the standard set of test examples and evaluation function.

methods	succes tone detection [%]	poly	mono
Dixon	70 - 80	✓	
Marolt	80 - 95	✓	
Raphael	61	✓	
Monti and Sandler	74	✓	
Double Spectrogram	61,65 for poly 94,02 for mono		✓

Tab. 2. Table success tone detection for methods

Success of detection was achieved a simple relationship:

$$Success[\%] = 100 - ErrorRate[\%] \quad (9)$$

Double spectrogram method is designed primarily for monophonic music. If the input signal is polyphonic, Double spectrogram method achieves higher error rate. The best method in comparison is Marolt. When the input signal is polyphonic Double spectrogram method achieves the same error rate as a method of Raphael. When the input signal is monophonic Double spectrogram method achieves the same error rate as a method of Marolt.

Acknowledgements

Research described in the paper was financially supported by the Slovak Research Grant Agency (VEGA) under grant– VEGA 1/0718/09 and 1/0720/09.

References

[1] VARGIC, R. Wavelety a banky filtrov, STU, Bratislava, 2004.

[2] GEIST, B. Akustika – Jevy a souvislosti v hudební teorii a praxi, Muzikus, Praha, 2005.

[3] OSTERTAG, P. Detekcia a identifikácia tónov v zvukových signáloch, Diploma thesis, STU, Bratislava, 2008

[4] Dixon, S. 2000. On the Computer Recognition of Solo Piano Music. Australasian Computer Music Conference. 31-7.

[5] Raphael, C. 2002. Automatic Transcription of Piano Music. Proceedings of the International Conference on Music Information Retrieval.

[6] Monti, G, and M. Sandler. 2002. Automatic Polyphonic Piano Note Extraction Using Fuzzy Logic in a Blackboard System. Proceedings of the International Conference on Digital Audio Effects. 39-44.

[7] Marolt, M. 2004. A connectionist approach to automatic transcription of polyphonic piano music. IEEE Transactions on Multimedia 6, no. 3 (June): 439-49

[8] Klapuri, A., Davy, M. - Signal Processing Methods for Music Transcription, Springer, 2006

The creation of a speech database for a diphone speech synthesizer

Ivan OBERT[1], *Gregor ROZINAJ*[1]

[1] Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
ivan.obert@gmail.com, rozinaj@ktl.elf.stuba.sk

Abstract. *This work has found a solution for the possibility of using pronunciation of whatever human voice in a diphone speech synthesizer. We managed to propose and execute the system which is able to create a speech database from the words comprising the speech corpus. The speech synthesizer will then be able to speak with a voice recorded by a user.*

In the first part of my work I will attend to basic principles of text-to-speech synthesis. I explain what theme of recordings is and what volume of text is needed for the best quality of synthesis. In the next part I will introduce database creation system architecture and describe particular modules comprising the system. DTW method is used for computerized segmentation of recorded spoken text that is needed to work compatibly with synthesizer. Finally I compile problems solved by DTW and describe possibility of solving current problems.

Keywords

speech synthesis, dynamic time warping, DTW, speech database, artificial voice, concatenation synthesis,

1. Introduction

A speech synthesizer is a device produced artificial human voice. An aim of the high-class synthesizer is that the generated voice should not be distinguishable from a real human voice. Synthesis systems are commonly evaluated in terms of three characteristics. Firstly, accuracy of rendering the input text, intelligibility of the resulting voice message and perceived naturalness of the resulting speech.

Present, the most used synthesis systems are concatenation synthesizer. Such synthesizer chains short or longer speech units. Concatenation of phonemes is not used very often, because of so-called coarticulation effect. More used systems are that works via diphones. Here there is no problem with coarticulation. All diphones that exist in given language, the diphone synthesizer possess.

2. Speech Synthesis from Textual Input

The aim for computer speech synthesis from either textual or conceptual input is to imitate the characteristics of the typical human speaking process well enough to produce synthetic speech that is acceptable to human listeners. Synthesis from text should be able to apply the rules used by a good reader in interpreting written text and producing speech.

Most work on speech synthesis has concentrated on text-to-speech (TTS) conversion, and TTS will form the main focus for this chapter

3. Converting from text to speech

The generation of synthetic speech from text is often characterized as a two-stage analysis-synthesis process, as illustrated in Figure 1. The first part of this process involves analysis of the text to determine underlying linguistic structure. This abstract linguistic description will include a phoneme sequence and any other information, such as stress pattern and syntactic structure, which may influence the way in which the text should be spoken.

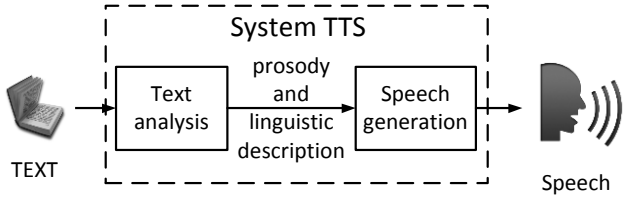


Fig. 1. The conversion from text to speech as an analysis-synthesis process.

The second part of the TTS conversion process generates synthetic speech from the linguistic description.

3.1 TTS system architecture

Both the analysis and synthesis processes of TTS conversion involve a number of processing operations, and most modern TTS systems incorporate these different operations within a modular architecture such as the one illustrated in Figure 7.2. When text is input to the system, each of the modules takes some input related to the text,

which may need to be generated by other modules in the system, and generates some output which can then be used by further modules, until the final synthetic speech waveform is generated. However, all information within the system passes from one module to another via a separate processing 'engine' and the modules do not communicate directly with each other. The processing engine controls the sequence of operations to be performed, stores all the information in a suitable data structure and deals with the interfaces required to the individual modules.

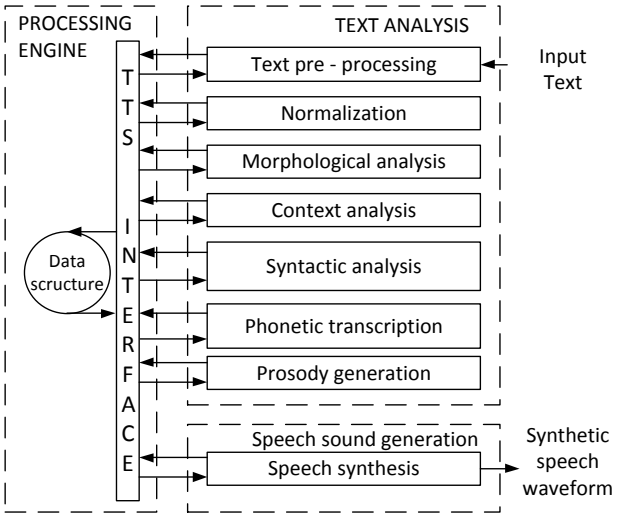


Fig. 2. Block diagram showing a modular TTS system architecture.

A major advantage of this type of architecture is the ease with which individual modules can be changed or new modules added. The only changes that are required are in the accessing of the modules in the TTS processing engine; the operation of the individual modules is not affected. In addition, data required by the system (such as a pronunciation dictionary to specify how words are to be pronounced) tend to be separated from the processing operations that act on the data. This structure has the advantage that it is relatively straightforward to tailor a general TTS system to a specific application or to a particular accent, or even to a new language.

4. Solution suggestion for database creation

We are dealing with concatenation synthesis so final speech is created by chaining particular speech units. As project of diphone synthesizer at Department of Telecommunication is semi-finished (diphones are used as speech units), the focus of this work is associated with uniform units – diphones. It means that synthesizer is not able to choose different length units from database to suit the particular circumstances, but only diphones (or phonemes if diphones do not occur there).

The database may be processed manually or automatically by computer algorithms. In the next section we will consider the advantages and disadvantages of those methods.

4.1 Automatic approach to speech database processing

A computer application controls every aspect of the database creation and user obeys orders displayed on-screen. At the beginning he says words into a microphone according to list on-screen and in certain parts of process can be requested to enter commands, controlling some steps in the algorithm.

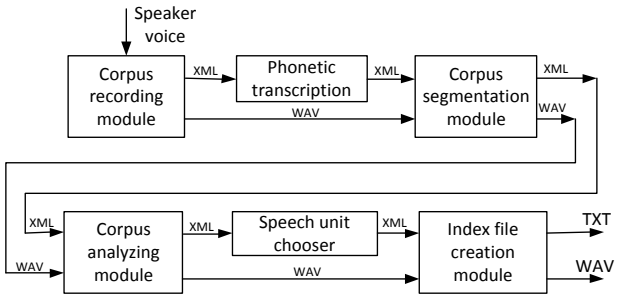


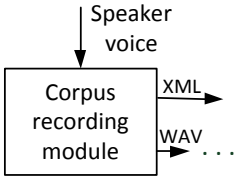
Fig. 3. Block diagram illustrating automatic speech database creation.

The whole process of preparing of the speech database into the form that the synthesizer could use is quite lengthy and therefore it is suitable to split it into several sections. Each module expects data input from the previous one, to then process them and send the changed data as input to the following module. The process continues this way until the last module in the sequence launches the complete speech database. Block diagram illustrating the speech database creation system is depicted in Figure 3. The advantage of separating the algorithm into different modules can be seen in simplicity of changing, upgrading or extending one block without any need to change another. Note the necessity of unified transfer form between modules.

4.2 The corpus recording module

First form in sequence is responsible for controlling the recording of words that will be a base for database. It has an access to a list of words, displayed on a screen for user to repeat. Thereafter, the user's utterance is recorded by microphone and saved to memory unit. The corpus recording may be done professionally in recording studio or might be accessible to larger group of people through the internet via small application.

One output is WAVE files containing recorded words, and one is XML files describing exactly what words occur. All these files are passed on to next modules to process. Structure of XML file is described in Figure 4.

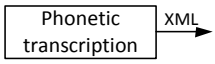



```
<?xml version="1.0"?>
<indexes>
. . .
 <file path="dievca0772.wav">
  <word>
        matka
  </word>
 </file>
 <file path="dievca0773.wav">
  <word>
        vlnka
  </word>
 </file>
 <file path="dievca0774.wav">
  <word>
        štvrtý
  </word>
 </file>
. . .
</indexes>
```

Fig. 4. XML structure demonstration occurring on the output of the corpus recording module**.**

4.3 The phonetic transcription module

This module receives the XML file where all corpus words are written and translates them into phonetic representation, i.e. how they were uttered. A class "PhoneticTranscription" of diphone synthesizer coded in C# is used. All elements <pho> are written hierarchically under the element <word>. One such element is shown in Figure 5.

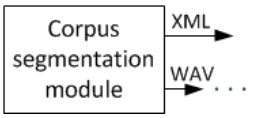



```
<?xml version="1.0"?>
<indexes>
. . .
 <file path="dievca0772.wav">
  <word>
   <pho>
      sil
   </pho>
   <pho>
      m
   </pho>
   <pho>
      a
   </pho>
. . .
  </word>
. . .
</indexes>
```

Fig. 5. XML structure demonstration occuring on the output of the phonetic transcription module.

4.4 The corpus segmentation module

The corpus segmentation module estimates the time of the cutting point between each pair of phoneme. It expects an XML file with phonemes as input and appends information about the starting and ending point of phoneme in audio file.



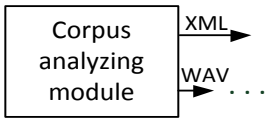

```
<?xml version="1.0"?>
<indexes>
. . .
 <file path="dievca0772.wav">
  <word>
   <pho>
      sil
    <bounds>
     <start>0</start>
     <end>0.0613318</end>
    </bounds>
   </pho>
   <pho>
      m
    <bounds>
     <start>0.0613318</start>
     <end>0.1673467</end>
    </bounds>
   </pho>
. . .
  </word>
. . .
</indexes>
```

Fig. 6. XML structure demonstration occurring on the output of the corpus segmentation module**.**

Phoneme limits are estimated by method based on template matching, where two vectors are compared. Dynamic Time Warping (DTW) is used to find out the similarity between reference template and second pattern where phoneme limits are found. The output XML file, part of which is shown in Figure 6, is enriched by phoneme borders.

4.5 The corpus analyzing module

This module is responsible for calculating some noticeable properties of particular phonemes. Typical features that can be useful for the next process are pitch energy and pitchmarks. The script of PRAAT program, focusing on speech processing and specializes in phonetic analyses and sound manipulations, is used to estimate those features.




```
<?xml version="1.0"?>
<indexes>
. . .
 <file path="dievca0772.wav">
  <word>
   <pho>
      sil
    <bounds>
     <start>0</start>
     <end>0.0613318</end>
    </bounds>
    <energy>
     <firstpart>0</firstpart>
  <secondpart>0</secondpart>
    </energy>
   </pho>
. . .
  </word>
. . .
</indexes>
```

Fig. 7. XML structure demonstration occurring on the output of the corpus analyzing module.

4.6 The speech unit chooser

Module choosing suitable speech units from corpus works upon properties and features obtained from previous modules.

In whole corpus there are lots of units of each phoneme. It is necessary to choose one utterance of each phoneme according to one of its properties because some phonemes can be unsuitable due to its disproportionate intensity in corpus. If synthesizer chooses such phoneme, a listener would notice the problem. Therefore this module is highly important in the database creation process. There is one more feature of the module – choosing the most suitable diphones as one unit of two consecutive phonemes. In the output XML file, as Figure 8. illustrates, one utterance of each phoneme and diphone is written.

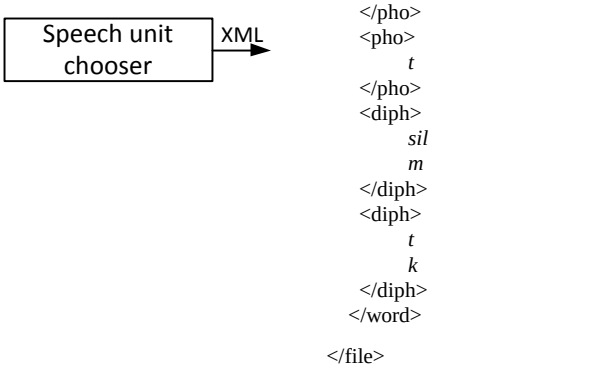


Fig. 8. XML structure demonstration occurring on the output of the speech unit chooser module.

5. Conclusion

In my work we have proposed and executed a speech database creation system. It consists of the six modules, each responsible for particular process.

DTW method is certainly a good tool for automatic segmentation of speech recordings. After applying the basic principle of DTW algorithm and a number of improvements, we have come to the conclusion that the diphone boundaries in recorded words, determined automatically simply by using DTW are very accurate. In the words starting with sibilant letters the algorithm does not always correctly identify diphone boundaries. The solution that seemed most effective was trimming spaces before and behind each word. Automated crop gaps could be resolved by comparing the energy gap which would be cut gradually until a human voice appears in the recording.

Acknowledgement

This work has been supported by the projects VEGA 1/0718/09 and FP7-ICT-2011-7 HBB-Next.

References

[1] HOLMES, J., HOLMES, W., S *Speech Synthesis and Recognition II.*, New York : Taylor & Francis, 2003,

[2] AGGOUN, A., SORIN, C., EMERARD, F., STELLA, M. Prosodic Knowledge in the Rule-Based Synthex Expert System for Speech Synthesis. *New Systems and Architectures for Automatic Speech Recognition and Synthesis*, Springer-Verlag, Berlin, 1985.

[3] RABINER, L., SCHAFER, R. Digital Processing of Speech Signals. *Prentice-Hall, Inc.*, New Jersey, 1978

[4] BICKLEY, C., SYRDAL, A., SCHROETER, J. ‘‘Speech Synthesis,’’ in the Acoustics of Speech Communication, *J.M. Picket, Ed*, Boston, NY: Allyn and Bacon, 1998.

[5] RICHARD, G., LIU, M., SINDER, D., DUNCAN, H., LIN, Q., FLANAGAN, J., LEVINSON, S., DAVIS, D., SLIMON, S. Numerical simulations of fluid flow in the vocal tract, *Proc. of Eurospeech*,, Madrid, Spain, 1995.

[6] Macchi, M., Altom, M.J., Kahn, D., Singhal, S., Spiegel, M. Intelligibility as a function of speech coding method for template-based speech synthesis, *Proc. Eurospeech'93*, Berlin, Germany, 1993.

[7] BLACK, W., CAMPBELL, N., Optimising Selection of Units from Speech Database for Concatenative Synthesis. *ATR Interpreting Telecomunications Research Laboratories*, Kyoto, Japan,

Simulator and Synthesizer for Feedback Sounds of Rotary Control Elements

Alexander TREIBER[1], *Gerhard GRUHLER*[1], *Gregor ROZINAJ*[2]

[1] Heilbronn University, Max-Planck-Str. 39, 74081 Heilbronn, Germany
[2] Dept. of Telecommunications, Slovak University of Technology, Ilkovičová 3, 812 19 Bratislava, Slovakia
treiber@photophil.net

Abstract. *This paper describes the development and testing of a method for the automatic prediction of the acceptance of the acoustic feedback of rotary switches. After a brief description of the tools and methods used to acquire acoustic measurements and individually optimized target sounds it is explained how the gathered data to identify both desirable and undesirable sounds from a large data pool. Furthermore, the benefits of well-designed and pleasant acoustic feedback to the user experience and the overall usability of human-machine-interfaces are explained.*

Keywords

Non-speech auditory feedback, benchmarking, prototyping, input device, human-machine-interface

1. Introduction

Due to the increased functionality of modern cars in terms of driver assistance systems, connectivity and infotainment on one hand and the increasing demands regarding usability and joy of use on the other, almost all car manufacturers offer menu-based user interfaces throughout almost their entire model ranges.

Usually these interfaces consist of a display unit close to the driver's line of sight as well as a control panel which is mounted relatively low so that it can be easily reached. This design principle is also recommended by the European Statement of Principles on Human Machine Interaction (EsoP).

Most systems use a combination of one central rotary encoder and several buttons. The rotary encoder is the main control element and therefore has to be operated frequently. Current high-end systems furthermore often feature a touch-pad for text input. For the acceptance and perceived quality of such a user interface it is necessary that the output of the systems (typically complex menu structures displayed on a TFT display) is logically arranged and visually appealing and the user input devices are ergonomically and aesthetically well-designed and well-built so that every user input is precisely and reliably registered.

Moreover, the system input and output also affect the security of operation: An easy to understand display allows the driver to concentrate on the road and precise input devices help to enter information into the system faster [1]. A crucial aspect is to provide the user with appropriate feedback during operation.

Feedback in general can be given to the user in three different ways:

- Visually, for example by highlighting something on a display or by control LEDs
- Tactile, typically as a sudden change of force or torque [2]
- Acoustically, by either electroacoustic stimuli or mechanically generated sounds

2. Description of the feedback signal

Acoustic and tactile feedback in conventional electromechanical control elements such as tact switches of rotary encoders are often caused by the same phenomenon and hence at the same time: the sudden change of force or momentum accelerates certain internal parts of the control element.

As a consequence, this parts impact on other parts within the device. In the rotary encoders used in this project this events typically happen within 1-2 ms, the resulting click sound typicalls lasts less than 10 ms.

The following two figures show the movement of a spring inside a rotary encoder which occurs at the time when the electric output of the device is triggered.

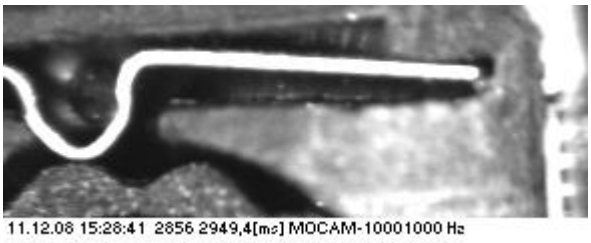


Fig. 1. Spring in it's left rest position

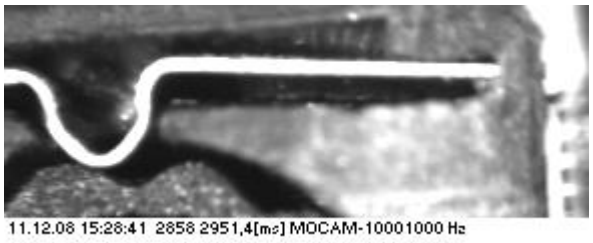


Fig. 2. Spring at it's right rest position, the displacement of the spring is 0,4 mm

As can be seen in the figures, the spring movement which causes the acoustic feedback occurs is a cast plastic enclosure which provides only 0,4 mm displacement. This value among others varies due to tolerances in the manufacturing process which results in significant changes in the acoustic feedback of the encoder between two samples.

Several thousand individual click sounds have been sampled on an integrated measurement unit which has been developed during the course of this project. A detailed description of unit can be found in [3].

The feedback signal typically shows a broadband spectrum with not very distinct peaks. It can therefore roughly be described a uniform noise modulated with a 10 ms amplitude envelope.

3. Signal synthesis, methodology and summary of early findings

The basic methodology used in this project was to iteratively improve both the synthesis and the analysis of the acoustic feedback of electromechanical control elements. The aim was to develop a synthesis algorithm which is capable of generating realistically sounding click sounds based on a relatively small set of signal parameters on the one hand and an analysis method capable of extracting the same set of parameters from a recorded natural sound.

The concept was to use the recorded sounds for a first experiment in which subjects had to rank a selection of typical sounds according to their individual tastes. Subsequently, parameters affecting the acceptance of the sound were to be isolated based on this ranking of this sounds in combination with analysis results. The gained knowledge can subsequently used to improve the synthesis- and analysis methods for a more precise identification of parameters in the next iteration step of the subjective experiments.

Since sounds are subjectively perceived differently when they are taken out of context [4] a device for realistic sound playback had to be developed for this purpose [5]. The device consists of a rotary encoder which looks like an ordinary encoder in an automotive user interface but features no inherent acoustic or tactile feedback. It is connected to a system which triggers artificial acoustic feedback which can be either recorded or synthesized in real-time.

The first experiment which used only recorded sounds showed a strong rejection of sounds which were considered to be too loud. Sounds which were very quiet were preferred by a small number of subjects while the majority rated sounds in the middle of the amplitude range to be preferable. The effect of the amplitude masked all other influences by temporal structure or spectrum. [6]

For this reason, synthetic amplitude-normalized sounds were used in the next trial. All sounds used an identical amplitude-envelope with linear attack and decay times of 1 and 12 ms respectively. This envelope was used to modulate different harmonic and noise-like waveforms. The sound which was favored in the first trial was used as a reference. The study showed that subjects generally rejected harmonic signals and preferred the ones based on noise. [7]

The insight that harmonic feedback signals are generally disliked led to the adaption of a synthesis algorithm for impact sounds originally proposed by Gaver [8]. While his original algorithm uses sine waves as oscillators modulated with exponential decays the first proposed modification is the usage of digital wavetable oscillators instead [9]. The wavetables are loaded with 3rd-octave bandpass filtered noise.

Secondly, as explained in the works of MacAdams [10] and Lutfi [11], the attack phase of a transient sound is crucial for identification. For this reason, a logarithmic attack phase was added to the amplitude envelope.

As analysis of the recorded sounds showed, the spectrum of typical clicks shows no significant spectral components below 800 Hz, therefore 15 3rd-octave bands are sufficient to cover the audible spectrum of a click sound.

Following a procedure described in [5], the synthesizer requires only three input values by the user:

- Peak Amplitude
- Decay Time / Damping Coefficient
- Spectral bias (emphasis of low or high frequencies)

4. Target sound definition and automatic prediction of acceptance

This final experiment for was divided into two parts:

In the first part the subject's task is to design a feedback signal which is perceived pleasant by the subject. This is done using an adaptive procedure for each parameter of the sound. Each parameter was adjustable by means of a pair of buttons as shown in Fig. 3:

The markers above each pair of buttons indicated when the final step size of the adaptive procedure was reached. As the three parameters influence each other (i.e. a longer sound with the same peak amplitude is perceived louder than a shorter one), the setting phase of the sound was not automatically ended as soon as all three parameters reached the final step size but the subjects were allowed so spend as long as they wanted to set up their individual stimuli.

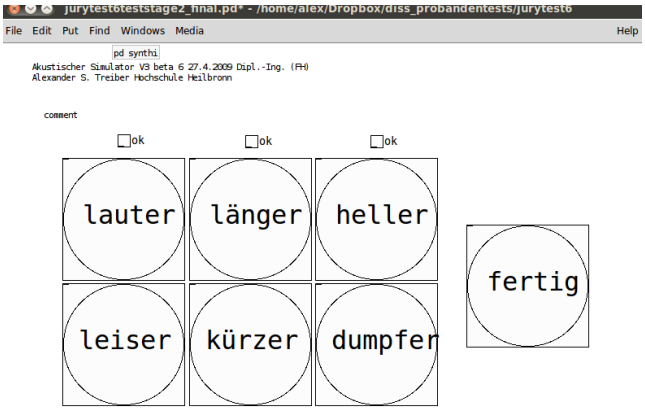


Fig. 3. Control panel to dial in the desired sound

Once the subject is satisfied with the sound, the confirm button stores the current settings and finishes this part of the experiment. The target sound definition is followed by calculation and analysis. In the calculation step the stimulus which was designed by the subject is analyzed and compared to the stimuli in the database using the methods explained in the previous chapter. The program selects a total of 29 sounds from the database, which are selected due to the following criteria:

- Increasing Correlation Coefficient of the Damping Coefficients (10 stimuli)
- Increasing Correlation Coefficient of the Peak Amplitudes (10 stimuli)

Since the sounds which are stored in the database are all relatively similar due to the fact that they are all based on physical control elements which are constructed in a similar way, their typical correlation coefficients were similar as well. Fig. 4 shows an example of a typical distribution of correlation coefficients:

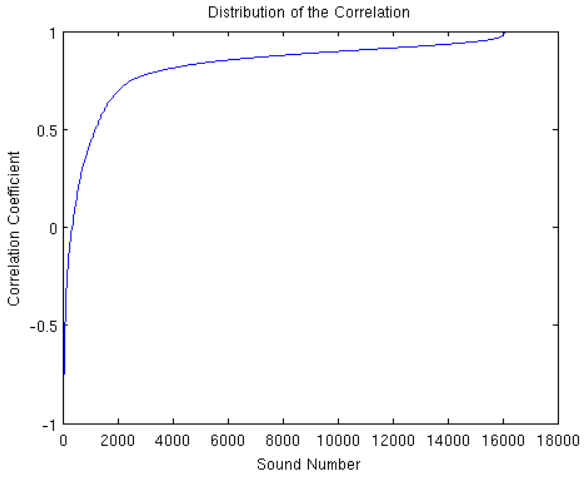


Fig. 4. Note that over 75% of the sounds feature very high correlation coefficients

The figure shows the correlation coefficients for the peak amplitude parameter for the entire database. It can be seen that the curve is relatively steep for the first 2000 sounds and then becomes relatively flat, more than 75% of the sounds show high to very high correlation with the reference value in this example. The same is true if the damping coefficients are correlated.

The analysis and subsequent correlation and data extraction can be done in less than a minute so that the subjects can continue with the second part of the experiment right anway, since the time used for calculation is used to brief the probands with their task for the second part.

In this part of the experiment it is the subject's task to rate the sounds which have been selected by the system in the step before. The subjects do this using a seven-point scale on the interface. The subjects could listen to the sound as long as they wished. Once the subject came up with a conclusion, he could enter the rating of the sound into the seven-point scale and confirm the result. The next selected stimulus would be played back using the simulator. The order in which the stimuli were presented to the subjects was randomized for every subject. The 20 sounds which are selected because of their correlation coefficients are normalized in terms of peak amplitude in order to prevent any influence through this parameter.

5. Analysis and Discussion of the Results for Acceptance Prediction

The acceptance has been predicted based on three different methods:

- Correlation of the Damping Coefficients
- Correlation of the Peak Amplitudes

The proposals for a prediction method which are based on correlation of signal parameters are supposed to have an effect on the acceptance of the stimuli, i.e. high correlation coefficients tend to higher acceptance ratings than low correlation values. A comparison of both methods of selection is shown in figure 5:

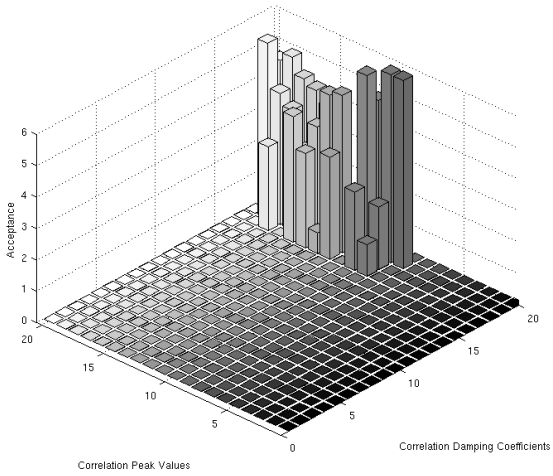


Fig. 5. This plot shows the average acceptance values achieved with stimuli which showed the according combinations of the correlation coefficients of peak amplitudes per band and damping coefficients per band.

Since the variations in the actual correlation coefficients which have been calculated for each sound are very small the values have been quantized into 20 ranges and averaged within each range. The result is shown in the above figure. x- and y-axes denote the correlation coefficients of the damping coefficients and the peak amplitudes respectively. The z-axis denotes the average acceptance within a certain correlation coefficient bin. One would expect that the highest z-values occur in the upper ends of both x- and y-axis.

This is true for the correlation coefficient of the damping coefficients, in which relatively small deviations from the highest possible values lead to a significant reduction in the acceptance values. Regarding the correlation coefficients of the peak values it can be seen that even relatively low values lead to high acceptance ratings as long as the damping coefficients correlate well with the reference.

A possible explanation for this is that the human sense of hearing works as an integrator and as long as the damping coefficients correlate well, the integral of emitted sound per band should correlate relatively well, too, even if the peak amplitude in this band did not match the desired values.

6. Conclusions and Outlook

The experiment has shown that at least one of the two proposed methods for the improvement of the prediction of the acoustic feedback in fact leads to higher acceptance values in the subjectively rated sounds. In combination with the aforementioned big effect of the overall amplitude is is possible to automatically identify sounds which are likely to suit an individual listener's taste.

By obtaining a large number of individually optimized stimuli it is possible to acquire an average target sound for a specific application and use the described method to identify an acoustically suitable electromechanical component by comparing recordings of the said component with the target. Furthermore, the synthesizer/simulator can be used as a rapid-prototyping tool on it's own for defining target sounds which do not necessarily have to be generated as a side effect of an electromechanical device. An obvious example for this application is the rapidly increasing use of touch-sensitive devices in automotive user interfaces, points of sale and – above all – mobile devices.

As a study during the course of this project showed, well designed acoustic feedback does not only make a rotary control more pleasant to use but improves the security and speed of operation as well [1].

Also, new questions are raised which provide future challenges for more research regarding the role of acoustic feedback in user interfaces. The validity of the results in conjunction with tactile feedback has yet to be examined. From research regarding mobile touch-screen devices it is known that the graphic design (shape, size, color) of virtual buttons affects the expectations of how the acoustic feedback is supposed to be. Transferred to the goals of this project a possible future challenge would be to examine the effects of shape, weight, surface texture and material of a physical knob.

Acknowledgment

This work has been supported by the projects VEGA 1/0718/09 and FP7-ICT-2011-7 HBB-Next.

References

[1] TREIBER, A.S., GRUHLER, G., ROZINAJ, G., Effects of the acoustic feedback of rotary control elements on task performance. *Proceedings of REDZUR 2010*

[2] REISINGER, J., WILD, J., MAUTER, G., BUBB, H., Haptical Feeling of Rotary Switches, *Proceedings of Eurohaptics 2006*

[3] TREIBER, A.S., GRUHLER, G. Measurement and Optimization of acoustic feedback of control elements in cars. *Proceedings of AES 122*, AES Convention Paper 7135 (2007)

[4] BLAUERT, J., GUSKI, R., Critique of Pure Psychoacoustics, Proceedings of NAG / DAGA 2009

[5] TREIBER, A.S., GRUHLER, G., HÄUPTLE, P., ROZINAJ, G., Simulator and Synthesizer for Feedback Sounds of Rotary Control Elements. *Proceedings of REDZUR 2009*

[6] TREIBER, A.S., GRUHLER, G., Subjektive Bewertung der Schaltgeräusche von Bedienelementen im Kraftfahrzeug. *Proceedings of DAGA 2008*

[7] TREIBER A.S., GRUHLER G. Psychoacoustic Evaluation of Rotary Switches, *Proceedings of IWSSIP 15*, IEEE (2008)

[8] GAVER, W.W. Synthesizing Auditory Icons, *Proceedings of INTERCHI 93*

[9] ROADS, C., STRAWN, J., The Computer Music Tutorial, *MIT press, 2009*

[10] MACADAMS, S. Audition: Cognitive Psychology of Music, *In the Mind Brain Continuum*, MIT Press (1996)

[11] LUTFI, R., *Classication and identication of recorded and synthesized impact sounds by practiced listeners, musicians and nonmusicians*, JASA Vol. 118, p. 393-404, 2005

Selected security threats in VoIP IMS architecture

Juraj Londák[1], *Pavol Podhradský*[2]

[1] Institute of Telecommunications, Slovak University of Technology, Bratislava, Slovakia
[2] Department of European Programms, Slovak University of Technology, Bratislava, Slovakia
[juraj.londak, pavol.podhrad]@stuba.sk

Abstract. *The paper is focused on NGN/IMS network and its security. IMS is the standard of converged network architecture, combining different types of networks and services, providing them in one network using IP and other relevant protocols. The work contains an overview of important points and principles of this architecture and its security risks. Paper offers overview of attacks that threaten such networks and proposals how to avoid them.*

Keywords

IP Multimedia Subsystem, Security, DoS, Toll fraud, eavesdropping

1. Introduction

The move toward all IP architectures for service delivery appears to be a strong trend. In this context, customers seem to desire an access to personalized interactive, multimedia services, on any device and anywhere. This trend introduces new requirements for network infrastructures. The IP Multimedia Subsystem (IMS) is seen as a promising solution for fulfilling these expectations.

The IP Multimedia Subsystem (IMS) is at the core of the upcoming next generation of telecommunication services. Developed by the 3rd Generation Partnership Project (3GPP), the IMS is based on Session Initiation Protocol (SIP) [2] signaling and the Internet Protocol (IP). The IMS architecture represents a significant change in the way telecommunication services are implemented and deployed. With these changes there comes a new set of challenges to provide a secure and trusted set of services.

Given that the IMS is based on the SIP and the IP, it inherits numerous known security challenges with these protocols. There have been created interesting and exhaustive works in recent years on both the problems and the solutions for the SIP-based VoIP security [3,4].

The first section of the paper is dedicated to common IMS overview. Its aim is to provide a coherent view of principles, history and development of the architecture. The subsequent section builds on this overview description enlists and explains security threads related to the whole VoIP realm. Detailed understanding of underway attacks is a necessary background for a serious discussion and proposition of relevant solutions.

2. IP Multimedia Subsystem

The IMS refers to a functional architecture for multimedia service delivery, based upon Internet protocols. Its aim is to merge Internet and cellular worlds, in order to enable rich multimedia communications. It is specified in the 3rd Generation Partnership Project (3GPP) [5].

The IMS was introduced in the UMTS release 5 and 6 [6]. In its first version, the focus was taken facilitating the development and deployment of new services in mobile networks. It was later extended by the European Telecommunication Standards Institute (ETSI), in the scope of its work on Next Generation Networks (NGNs). A standardization body of the ETSI, called Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN) standardizes the IMS as a subsystem of NGNs.

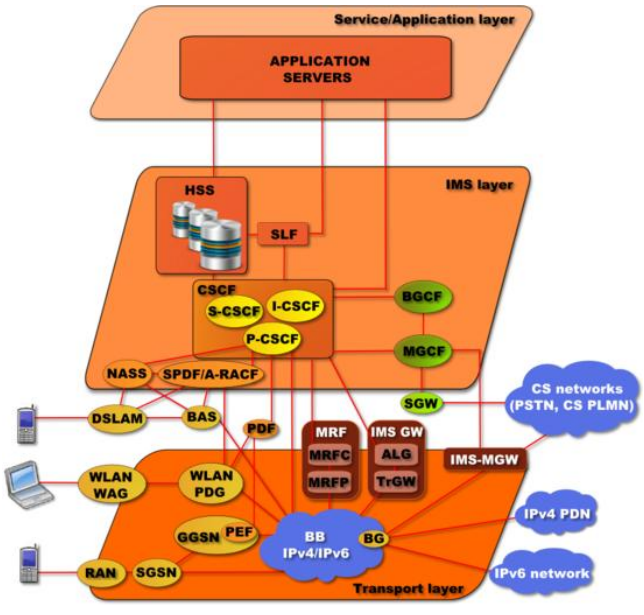


Fig. 1: 3GPP/TISPAN IMS Architectural Overview

3. Security threats

In the previous chapter, we discussed in detail the motivation, principles, architecture and threads associated with the IMS. This chapter provides an overview of threats and attacks associated not only with the IMS but with VoIP environment as a whole. Distribution of attacks in subchapters is taken from the book [10], which also served as the main but not the sole basis for writing this chapter. Like the previous chapter, this one should too serve as a sufficiently comprehensive but clear source, serving as an introduction to security of the whole VoIP issue [11].

It must be said that the used division of attacks is not completely consecutive. Attacks are often complex procedures, which are very similar in some features or they link to each other. Therefore, the division should be understood from the the target point of view rather than the one of the way the attack is done.

The most common methods of attacks that are with major or minor variations used in different targets are DoS (Denial of Services), eavesdropping and toll frauds. Therefore, in this chapter they will given a wider space for their detailed description. Each subsection represents a category of threat. The group is here explained in terms of motivation, history and other important features. It also mentioned some of the methods or measures that help prevent or defend against the given attacks.

3.1 Ecosystem attacks

The first large group of the attacks directed against communications system as a whole. Often these attacks may not even be targeted and can be caused by unforeseen collective behavior of its users.

At a time when the individual systems were separated, often operated by its own physical infrastructure, it was enough to ensure each system separately. In the case of IMS network the situation is quite different. As previously mentioned, one of the main goals of IMS is convergence of several different networks and services into one network. This linking of more services, however, causes a real security nightmare. Interconnection and dependency of services from each other generates a huge amount of possibilities and situations that are almost impossible to anticipate and avoid. Using an IP and a shared infrastructure only expands and complicates the problem.

As already stated, probably generally the most serious risk of converged networks is the problem of availability of services. The problem may not be necessarily the result of a DoS attack. A problem in the network can also be caused by users' behavior. Toll fraud has the potential to be a threat with the greatest economic impact for the operator. In order to provide the highest comfort to its customers a provider can often make security compromises. The risk of Exposure of Information is based on the same principle as the previous case. Complexity and coherence of the system and its individual components produces a large number of places and situations where it is necessary to ensure authentication. A case when the confidential customer information is available to the public is a very uncomfortable experience.

3.2 Insecure endpoints

Times when terminals were just simple plastic boxes are long gone. SIP phones actually are small computers. With the increasing complexity also a number of security risks increase.

Basic and most simple attacks against endpoints are grouped in DoS attacks. The basic attack strategy is overloading the device through the "*ping–f*" command. Though there are other more advanced means available to the public. Another SIP specific attack is sending a SIP *BYE* message, which causes termination of on-going calls. If we go beyond the basic attacks, it is necessary obtain accurate identification data of endpoint device for attacks. For this section there are also lots of great resources available on the Internet. One of them is the *SIPVicious*. After searching for possible targets, it is necessary to gain control over the device. It is a very common phenomenon, that a vast majority of users use default passwords, which is a big security risk.

3.3 Media channel attacks

We have discussed attacks harmful for global system and defense strategies against them. In the following chapters more specific attacks will be addressed in details. It is a well-known fact that the SIP signaling and media themselves are divided into separate channels and often travel network along various ways. Therefore, we can divide these attacks in attacks on media channel and signaling or control channel attacks. In this chapter we will address attacks on media channel in details.

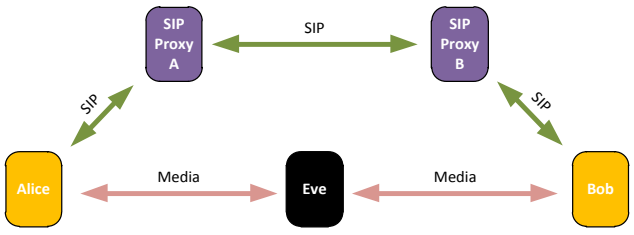


Fig. 2: An Attacker, Eve, needs to get to media path of communication

Mentioned attacks can be divided into eavesdropping and modification attacks. It is necessary to know the most important difference. Eavesdropping attacks [7] are simpler and always passive. This means that an attacker or a program just sits somewhere in a network and listens to the communication that passes through the network. An attacker does not enter actively into the process of service

initialization nor affects the routing data in any way. A program can archive intercepted data and an attacker needs to pick up the data only once in a while. On contrary, modification attacks are active attacks where a program is actively engaged in communication, what makes this type of attacks more complex. A program may alter words in IM communication during attack or even alter words in case of voice communication [8].

In the modification attack attackers placed a program among communicating terminals, as shown in the *Fig. 2*. Program forwards all communication that does not necessarily need to be modified for a successful attack. There are several methods of how an attacker/program could achieve to place itself between communicating parties. One of the best known and most widely used methods is called. *"ARP spoofing".*

The basic prevent strategy against eavesdropping and modification attacks is encryption.

3.4 Signalization channel attacks

In the previous chapter were discussed methods of how to eavesdrop and modify communication through an attack on a media channel. The same procedures and equipment can be applied to the interception of a control channel. Such attacks, however, have their own characteristic features, which will be outlined in this chapter.

Similar to eavesdropping attacks against media channels, an attacker only needs to get to the network segment where control channel traffic is occurs. An attacker can then capture all traffic on the network segment and analyse the traffic at some later time. By analysing the control channel traffic, an attacker can potentially learn important confidential information:

An attacker in the middle of a control channel can of course also cause end of denial-of-service (DoS) attacks. An attacker could simply drop certain control messages, such as endpoint registration messages, resulting in endpoints being unable to interact with the system. There is, though, a range of other DoS attacks that do not require a man-in-the middle attack to be effective.

Network flooding attacks have long been a standard part of an attacker's toolbox for denying service. The basic idea is to send a massive amount of traffic to a specific network segment with aim to create so much network congestion that authentic traffic cannot reach its target, or to generate a massive amount of traffic at a particular server causing exhaustion of all its resources trying to respond to the bogus traffic [9].

There are also SIP specific DoS attacks. One representative example is the SIP INVITE attack. The bombardment aimed to a SIP server which can result either in the server not being able to accept further sessions or in some cases server rebooting or otherwise ceasing operation. A different example is the BYE attack when attacker floods network or an IP range with BYE SIP messages causing termination of all enduring SIP sessions.

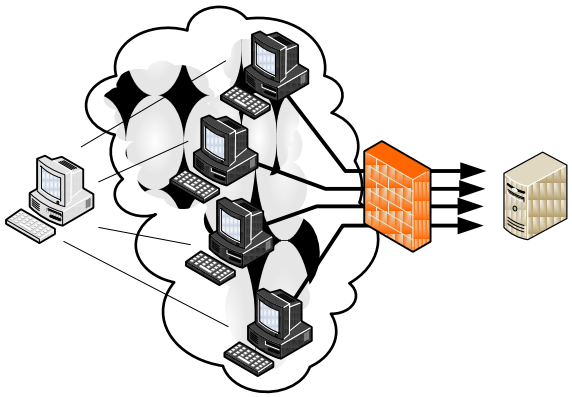


Fig. 3: Illustration of DDoS attack

Obviously, if an attacker is launches the flooding type of attack, they may want to hide the source of the attack so that they cannot be simply shut down by blocking one or just a couple of addresses. An attacker could use a botnet to launch attacks (see the *Fig. 3Fig.*). In a botnet, there are hundreds or thousands of computers out there that have been compromised and have a "bot" installed on their system waiting for commands.

Similarly to defending against media channel attacks, the best way how to prevent control channel attacks also is encryption. A typical protocol or mechanism used for encryption of a control channel is either TLS or IPsec.

3.5 Identity attacks

As we have moved from the traditional PSTN to the new IP communications networks, the challenge is that we are now in a space where "identity" can be spoofed even more easily than it was possible before. Yet, at the same time, we have a user population that is accustomed to trusting the caller identification information and is therefore currently more susceptible to abuses of the identity process.

At the SpeechTEK 2008 conference in New York, a renowned hacker Kevin Mitnick demonstrated in his keynote the way he could publicize a phone number for a bank that, when called, appeared to be from the bank and in fact did access the bank's real IVR system. However, the phone number was actually for his attack system that was sitting in the middle, relaying the bank's audio to the caller and relaying (and logging) the caller's key presses to the bank's system. After someone called this number and interacted with the bank's system, he then had all the information necessary to call back to the bank and identify himself as the caller.

4. Conclusion

This paper was aimed to summarize our understanding and knowledge of the issues of NGN/IMS network and their security. One of the objectives of the study was also to serve as a comprehensive introduction to this area.

The paper is also dedicated to attacks on VoIP networks in general [10]. The chapter is divided according to targets of attacks and their philosophy even if such a distinction is often ambiguous and intertwined with each other because of their complexity in real life. The chapter also contains detailed descriptions of the most common attacks such as DoS and eavesdropping. Consequences and implications of such attacks to network and its users as well as proposals for possible preventive measures are discussed here.

In view of this issue, it is clear that a large part of security risks of networks is caused by laxity of users and even network administrators. The reason is no use or use of very weak or default passwords by users as well as network administrators. Another case is a very slow reaction of administrators to released security patches and updates of deployed systems.

Acknowledgment

This paper also presents some of the results and acquired experience from various research and educational projects such as NGNlab project, Leonardo da Vinci projects: InTeleCT and MLARG, project FP7-ICT-2011-7 HBB-Next, Slovak National basic research project VEGA No. 1/0720/09.

References

[1] LONDÁK, J., "*Methods of VoIP IMS architecture protection from attacks* " Faculty of Electrical Engineering and Information Technology, Project for Dissertation exam, 2011.

[2] ROSENBERG, J., et al., "RFC 3261: Session Initiation Protocol (SIP)," Internet Engineeging Task Force, 2002.

[3] HUNTER, M. T., CLARK, R. J., PARK, F. S. "Security Issues with the IP Multimedia Subsystem (IMS)," Georgia Institute of Technology White Paper, 2007.

[4] PORTER, T., et al., *Practical VoIP Security*. Rockland, MA, United States of America: Syngress Publishing, 2006.

[5] 3GPP. (2010, Dec.) IP Multimedia Subsystem (IMS); Stage 2 (Release 10). Techical Report TS 23.228 V10.3.0.

[6] 3GPP. (2003) Technical Specifications and Technical Reports for a UTRAN-based 3GPP system (Release 5). Technical Specification TS 21.101 V5.4.0.

[7] UNUTH, N. (2010) Security Threats in VoIP. [Online]. http://voip.about.com/od/security/a/SecuThreats.htm

[8] ENDLER, D., COLLIER, M., *Hacking Exposed VoIP: Voice Over IP Security Secrets & Solutions*. New York, USA: McGraw-Hill, 2007.

[9] REBAHI, Y., SHER, M., MAGEDANZ, T., "Detecting Flooding Attacks Against IP Multimedia Subsystem (IMS) Networks," Fraunhofer Institut Fokus White Paper.

[10] YORK, D., *Seven Deadliest Unified Communications Attacks*. Burlington, MA, USA: Elsevier Inc., 2010.

[11] LÁBAJ, O., PODHRADSKÝ, P., KOTULIAK, I., Zabezpečenie komunikácie NGN testovacej platformy, Elektrotechnika a informatika 2007, konferencia v rámci medzinárodného veľtrhu ELOSYS, 16. – 19. október 2007, Trenčín.

LTS letter-specific tree rules

Matúš VASEK, Gregor ROZINAJ[*]

Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
rozinaj@ktl.elf.stuba.sk
matusvasek@gmail.com

Abstract. *This article is focused on process of LTS core creation. It is required automatic rules training system based on a database for Slovak language. The first requirement is the correctly formatted database. In following article data files, a descriptor file and finally generated rules trees are described. According to this manual final product can be reached, which is suitable for speech synthesizer and reflect actual requirements.*

Keywords

Speech synthesis, LTS rules, Wagon, data file, descriptor file.

1. Introduction

In view of how costly it is to develop LTS rules, particularly for a new language, attempts have been made recently to automate the creation of LTS conversion rules. Methods for the creation of LTS conversion rules are based on assumption that given set of words with correct phonetic transcriptions (the offline dictionary) and an automated learning system could lead to significant generalizations. Among them, classification and regression trees (CART) have been demonstrated to give satisfactory performances for letter-to-sound conversion [1].

In most general terms, the purpose of the analyses via tree-building algorithms is to determine a set of *if-then* logical (split) conditions that permit accurate prediction or classification of cases [2].

The topic of this article is closely connected with related articles about a topic of LTS trees creation process based on tree-building algorithm.

2. Training procedure

The process of training procedure involves following steps:

- A lexicon pre-processing into suitable training set
- Definition of sets of allowed letter/phoneme pairs. (It's intended to do this fully automatically in future versions).
- Determination of probabilities of each letter/phoneme pair.
- Aligning letters to an equal set of phonemes/_epsilons_.
- Extracting the suitable data for training by the letter.
- Building CART models (decision trees) for analyzed letter and its context [3].

3. Training process letter-by-letter

After the training database is prepared one comes to the core of training process. Result of preparing process of database is an object, which is able to join the training process. Rules are stored in the tree structure. For each letter, generally said for each proceeded object, is appropriate to create a separate tree. This steps results into a group of parallel ordered trees. Independent tree creating supposes more than one training set. We need as many training sets, as many tree files we would like to make. Separated independent data files we extract from original, universal database. We choose appropriate training vectors taking into consideration criteria mentioned in following text. Thus we produce trees for letters, e.g. "a", "á", "b" etc.

3.1. Data file

Vectors contained selected letter are cute dot from training set and written separately. At first, database has to be sorted. The 7th parameter is used as leading element in sorting process. Data ordering leads to temporary change of elements position in training vector. The 7th element comes first, other elements follow. Sorting is now executed. Vectors with the same character are grouped together. The alphabetical ordering is used for each letter. After that elements in vectors are put back into their original position. Whole database should be divided into separated ".data" files. Output is represented with the group of files, named "letter.data" e.g. "a.data".

[*] pedagogic supervisor

3.2. Descriptor file

The descriptor file together with data file creates inevitable couple for training process. The descriptor file manages and defines allowed values for each variable. Each element of the descriptor vector can have set of acceptable characters (phonemes or letters) so there should not be letter misinterpretation. Allowed values were discussed in related works [8]. The complete interpretation of each letter into SAMPA (all possible values) is mentioned in Tab. 2.

a	**a** a: { E I_^a [_.]a [_.]a:	ň	**J** n
á	**a:** a U: [_.]I: [_.]a:	o	**O** O: [_.]O a
ä	**{** E a	ó	**O:** O [_.]O [_.]O:
b	**b** p I O v x	ô	**U_^O** O O: v
c	**ts** [k_>] [ts_>] dz k g tS x z s	p	**p** [p_>] b
č	**tS** [tS_>] dZ dz tS ts	q	**k**
d	**d** ? E I_^ J\ [J_>] [d_>] c dz g t ts	r	**r** ? L r[O;] ŕ[O;]
ď	**J** ? t c [sp] [J_>]	ŕ	**ŕ[O;]** r[O;]
e	**E** ? E: I I: I_^ I_^E O U U: U_^ [_.]E [_.]I [_.]I: [_.]O a a: j {	s	**s** ? S z [ts_>] ts [s_>]
é	**E:** E I_^E U: [_.]E:	š	**S** s [tS_>] Z [S_>]
f	**f** ? I_^ J\ U_^ [c_>] b c f_v	t	**t** ? [c_>] c [t_>] d ts
g	**g** Z dZ [g_^] k	ť	**c** ? J\ [c_>] t
h	**h** E g b x z G	u	**U:** I_^U\ U U_^ [_.]U a v
i	**I** [_.]I I: I_^?	ú	**U:** U [_.]U:
í	**I:** [_.]I: [_.]I I_^ I	v	**v** U U_^ I_^ f_v f
j	**j** I_^E ? I I_^ dZ Z h x	w	**v** U U_^
k	**k** ? [k_>] g x	x	**k-s** g k z
l	**l** ? L I_^ l[O;] Í[O;]	y	**I** I_^ I: j y
Í	**Í[O;]** L	ý	**I:** I ý
ľ	**L** l	z	**z** ? S Z [dz_>] [ts_>] [z_>] dz s
m	**m** ? F [m_>] n	ź	**[z=:]** ?
n	**n** ? J N J/ [J_>] U [N\] [n_>] a m	ž	**Z** ? S [dZ_>] s z

Tab. 2.: Table of possible SAMPA equivalents

In case of large number (more than 10 occurrences) of phoneme, which wasn't included in descriptor file but occurs in Abel Kral dictionary, phoneme should be directly added to the list.

8.1. Raw tree

If data and descriptor files are prepared, program Wagon executes them with input arguments such as Fig.1.

```
Wagon –data a.data –desc a.desc –stop 10 > a.tree
```

Fig. 1. Input arguments

Output is redirected to the file „a.tree". File „a.tree" includes generated tree for the letter „a" and statistic information. In each row of generated tree structure are written probability for each phoneme mentioned as group of allowed values in the descriptor file. The probability depends on a context and the position in the tree. The classification tree vector ends with repeatedly written SAMPA character. It represents the phoneme with the greatest probability and is elected from the others. Probability is calculated as number of vectors supporting this rule divided with number of all vectors connected with this occurrence. Element example is displayed in Fig.2.

```
((n.n.n.name is a)
(((? 0) ([c_>] 0.0133333) (t 0.0533333) (c 0.933333) ([t_>] 0) (d 0) (ts 0) c))
```

Fig. 2. Raw tree element

Input argument „stop" enables to manage tree dimension and amount of branches. With growing „N" is reduced the tree structure because it is required higher number of occurrences in the training vector set which support the related rule. If the number grows over the defined value it is reason for new rule node creation. In this training process were used values 10, 20 and 30. The default value is 10. Higher values reduce extremely large trees.

8.2. Reduced and separated trees

In the complex system of text phonetic transcript are used LTS rules, generated by Miloš Cerniak. New rules format is the same as previous one. It ensures a full compatibility. It is possible to plug them in a program with no algorithm modification. This standard is based on a syntactic structure of LISP. Miliš Cerniak's rules do not suppose the probability information for individual phonemes [4]. For the raw tree generating (training) according to afore mentioned method are discarded all probability values.

```
((n.n.n.name is a)
((c))
```

Fig. 3. Reduced tree element

The chosen phoneme is, of course, presented and written in SAMPA. (See Fig. 3.) The tree including probability parameters (raw tree) can be in the future used in advanced transcription process. The second, third, etc. possible phonemes could be thus defined, taking into consideration phonemes with the biggest value of probability. This process should be activated by an error occurrence. . Actually it is not inevitable to implement this method, because it is suggested to make the error correction using a dictionary of exceptions. The dictionary of exceptions is altogether powerful and effective tool for error correction. Following chapters will discuss this topic in more detail.

8.3. Getting separated trees together

Finally, new LTS rules are built by iterative method, tree by tree. The original rules created by Miloš Cerniak are needed. Old trees are step-by-step replaced with new ones. After each step the request is sent to transcription server to test, weather the system is able to operate with changed part. The transcription server encapsulates transcript algorithm [5]. Upgrading all trees at once means high probability of error occurrence in syntax.

Considering amount of nodes it would be very difficult to find and fix errors. There are more than 2500 rows of LTS rules. Fig. 4. describes this problem.

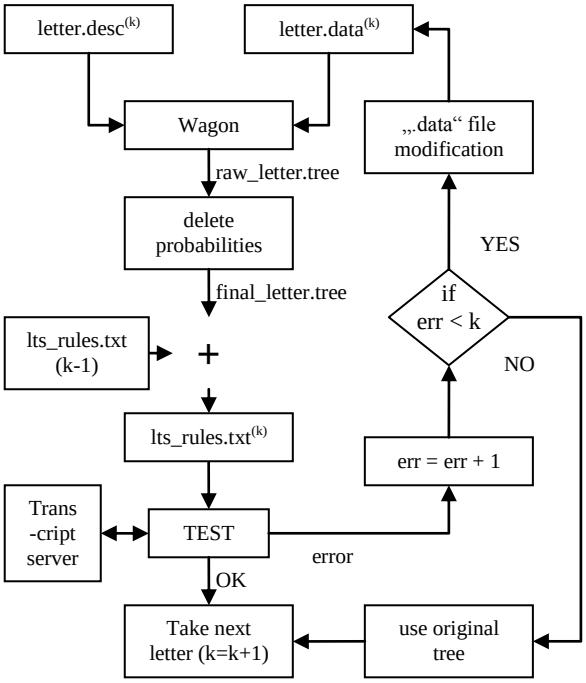


Fig. 4. Iterative concept of rules creation process

9. Statistic data analysis

The most of trees operate with more than 90% cases correctly (90% of training sample vectors) (see Tab. 2.).

Perplexity is the geometric mean of the number of words which may follow any given word for a certain lexicon and grammar [7].

letter	entropy	perplexity	correct [%]
a	0,015	1,010	**98,677**
á	0,007	1,005	**99,353**
ä	0,627	1,544	**53,699**
b	0,029	1,021	**97,665**
c	0,138	1,100	**89,525**
č	0,123	1,089	**90,940**
d	0,083	1,059	**93,392**
ď	0,301	1,232	**77,940**
e	0,013	1,009	**98,819**
é	0,034	1,024	**96,678**
f	0,044	1,031	**96,418**
g	0,046	1,033	**96,433**
h	0,098	1,070	**93,672**
i	0,022	1,015	**98,014**
í	0,024	1,017	**97,741**
j	0,091	1,065	**92,725**
k	0,194	1,144	**85,798**
l	0,052	1,037	**95,396**
Ĺ	0,025	1,017	**97,546**
ľ	0,002	1,002	**99,772**
m	0,004	1,003	**99,595**
n	0,053	1,038	**95,501**
ň	0,021	1,014	**97,970**
o	0,004	1,003	**99,579**
ó	0,009	1,006	**99,081**
ô	0,056	1,039	**94,892**
p	0,017	1,012	**98,727**
q	0,000	1,000	**100,000**
r	0,006	1,004	**99,401**
ŕ	0,008	1,006	**99,156**
s	0,119	1,086	**91,075**
š	0,072	1,051	**94,199**
t	0,093	1,066	**92,478**
ť	0,605	1,521	**56,241**
u	0,029	1,020	**97,454**
ú	0,010	1,007	**99,053**
v	0,090	1,065	**93,273**
w	0,281	1,215	**77,049**
x	0,012	1,008	**98,862**
y	0,019	1,013	**98,259**
ý	0,014	1,010	**98,574**
z	0,039	1,028	**96,947**
ź	0,000	1,000	**100,000**
ž	0,091	1,065	**93,082**

Tab. 2. Table of statistic data for letter-specific sub trees

Values rapidly fall for several letters (see Fig. 5.).

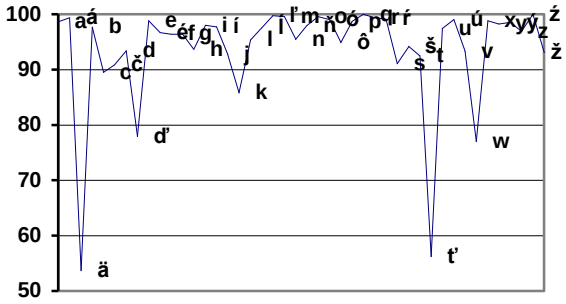


Fig. 5. Correct rate [%]

This fact influences entropy values (see Fig. 6.) because entropy is related with perplexity.

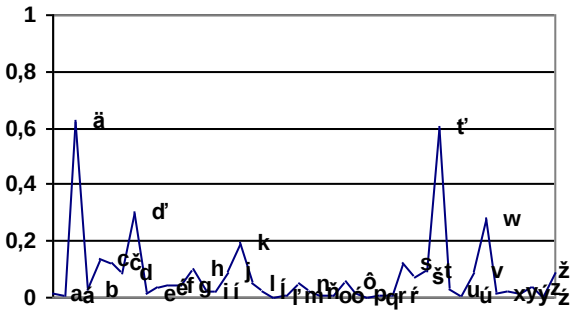


Fig. 6. Entropy

It is caused by multiple assignments in training set. That means, one word can be rewritten into more than one correct form. Training algorithm takes all this possibilities with the same weight. The letter “ä” can be interpreted by these SAMPA characters as well: “E”, “{„ (see Fig. 7.).

bábä	b a: b **a**
bábä	b a: b **E**
bábä	b a: b **{**
bábätko	b a: b **a** t k O
bábätko	b a: b **{** t k O
bábätko	b a: b **E** t k O
absolvovať	a p s O l v O v a **c**
absolvovať	a p s O l v O v a **J**

Fig. 7.: Different interpretation of the same letter in SAMPA

Both options are suitable and cannot be classified as a fault. Actually we tend to use pronunciation „E“. Similar effect occurred in case of letters „ť“ and „ď“, which pronunciation can be interpreted in two ways. Letter „w“ has luck of training sample elements (about 60). In Slovak language this letter is not generally used. These markers do not mean real problem for transcription quality.

9.1. Results

To interpret results is problematic because determining what makes a meaningful test set is difficult. On one hand, there is a paradox of motivation in that a large number of words for which a unique pronunciation can easily be specified, and so are present in the lexicon, are unlikely to need their pronunciation to be predicted by LTS. This implies whatever test set we specify is at best a poor approximation to what the LTS model would actually be required to do in practice. Furthermore, words which are not generally found in pronunciation lexica, for example foreign names, are often not pronounced consistently by human speakers anyway [6].

10. Conclusion

Several features of Wagon-based LTS rules training process were introduced. This article has described a method of trees creation process. One has been working with input database, called training set or sample as well. One of steps requires tree-by-tree replacing of original rules to come to new rules. Important issue is to ensure full compatibility with complex system for phonetic transcription of text for speech synthesis process. The simple plug and go system is guaranteed based on a modular architecture. The modular concept introduces effective solution which supports synthesizer development. Each tree is generated primary in raw format and valuated with a table of statistic data. Focusing on this information tables in data analysis leads us to creating the last part of the article. Some of the mentioned data can serve as loopback and signify that chosen procedure is able to fulfill demands, to give the tool for transcription– LTS rules.

Acknowledgements

This work was created with support of projects VEGA 1/0718/09 and FP7-ICT-2011-7 HBB-Next.

References

[1] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, Spoken Language Processing: A Guide to Theory, Algorithm and System Development, Prentice Hall, New Jersey, 2001

[2] StatSoft, Inc. (2010). Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/.

[3] Building Synthetic Voices. Lexicons. http://festvox.org/bsv/x1469.html

[4] Cerňak, M., Use of objective measurement of quality in the corpus of speech synthesis, PhD thesis, Department of Telecommunications, FEI STU, Bratislava 2004

[5] Vasek, M., Transcription of the input text to speech synthesizer, thesis, Department of Telecommunications, FEI STU, Bratislava, 2009

[6] Korin Richmond, Robert A. J. Clark, Sue Fitt, Robust LTS rules with the Combilex speech technology lexicon, Centre for Speech Technology Research,University of Edinburgh, UK, 2009. Available: http://www.cstr.ed.ac.uk/downloads/publications/2009/IS090308.pdf, Accessed: February 27, 2011.

[7] Dictionary.com, "perplexity," in *The Free On-line Dictionary of Computing*. Source location: Denis Howe. http://dictionary.reference.com/browse/perplexity. Available: http://dictionary.reference.com. Accessed: February 21, 2011.

[8] Gonšor, J., Methodology repair defective speech synthesis, thesis, Department of Telecommunications, FEI STU, Bratislava 2010

PRONUNCIATION OF NUMERALS IN SPEECH SYNTHESIS

Marek VANČO[1], *Gregor ROZINAJ*[1]
[1] Institute. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
marek.vanco@poslimi.to, gregor@ktl.elf.stuba.sk

Abstract. *This article deals with the question of numeral processing in a speech synthesis. My work is divided into two basic parts. First part is a module for grammatical category determination and second part is a number translator.*

Keywords

Numerals processing , Slovak language, determination of grammatical category, translation of numbers to numerals.

1. Introduction

The speech is the most basic and the most natural instrument for human communication. Human tried to create the communication „man - machine“ by many ways. At the beginning it was a speech with help of bags and pipes. Later, signals started to be used because of technology and science progress. The main aim was to generate the most identical sound signals based on their features. We encounter speech synthesis every day without even noticing. Speech synthesiser has become a favourite tool for train departure enouncing, infolines, electronic translators etc.

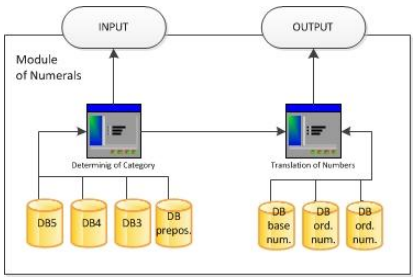
Numerals are one of extensive problems within speech synthesis and numerals are a key subject of my article. Since the numerals have not been explored within institute of telecommunication, the objective of this article is to establish the base for its future study. The linguistics sources presented the numerals as the most extensive and the most specific area. Numerals are substantial for speech synthesiser without which text parts would lose meaning. When we have to read this sentence: „Idem s 5 chlapcami.“. Synthesiser reads the sentence like this: „Idem s chlapcami.“, where the signification of quantity has been lost. When we use the module for translation, we get: „Idem s päť chlapcami.“. In this sentence, the quantity is presented but the pronunciation is not correct. When we use module for grammatical category determination: „Idem s piatimi chlapcami.“ . Therefore, I will be interested in the proposal of the module for transcription of numerals into the verbal form and in the problem of determination of grammatical categories in the sentence for inflection of numerals. I will deal with gender (in the case of male gender, vitality too), singular/plural and grammatical case.

2. The module for numerals

As I have mentioned in the introduction, the module consists of two parts:

- determining of grammatical categories,
- translation of number into the numeral.

Every module works individually, and as it can be seen in the Picture No. 1., they have their own databases. We will deal with the databases in the next part.



Picture No. 1. The module of numerals and a database

3. Determining of grammatical categories

3.1. Databases

When determining grammatical categories, there are a lot of methods how to reach required target. I have decided to use all the research results from our institute and use them for my research. For determining grammatical categories I have used a learned database of suffixes of Slovak words from Slovenský národný korpus (Slovak National Corpus), which consists of 3 parts:

- a database of 3-letter suffixes (DB3),
- a database of 4-letter suffixes (DB4),
- a database of 5-letter suffixes (DB5).

The advantage of this database is that it can be added and changed according to the actual needs and without knowledge of given program language. The databases include more than 55,000 suffixes and it can be seemed as a sufficient amount of data. Every suffix in the database has a lot of possibilities to be given to the different grammatical category. I will mention the example of determining the

grammatical categories of one word. I will use a frequent word „človek". When searching in the database of 3-letter suffixes (DB3), the software will find the following harmony of line:

vek Sms1 0,526 P 0,122 Pns4 0,049 Pis2 0,026 Sis1 0,021 Pfs1 0,021

Only first 6 possibilities of „vek" are shown. The probabilities are marshalled from the highest to the lowest. In this case, we can see the grammatical category „Sms1" with the probability of 52,6%. If we try to use the word „človek" in the database of 4-letter suffixes (DB4), we will get the following harmony:

ovek Sms1 0,994 Sis1 0,004 Sis4 0,002

Now, we can see that the amount of possibilities of suffices has decreased by more than a half and the probability of the grammatical category "Sms1" is 99,4%. In the database of 5-letter suffixes (DB5) the suffix "lovek" does not appear. Therefore, the probability of 99,4% "must" be enough for the category of Sms1 (noun, masculine gender – vital, nominative). This is one of the most ideal cases to get such a high probability when processing a given word. Sometimes, this probability of 99,4% does not have to be enough to guarantee the correct determination of grammatical categories of one word and not the part of a sentence in which the numeral occurs. As we know from morphology, another good method to determine grammatical categories are prepositions. They play a very important role and in the combination with the theory for suffices, we get even stronger tool. Thus, I have created a recorded database of several numerals which will help to increase the percentage success of determination.

3.2. Algorithm used for searching in the databases

We have 4 databases available to determine the grammatical categories for individual words. Searching for words in the database is set up to be absolutely logical. We can claim that if some harmony appears in the database of 5-letter (DB5) with the probability of 50% and if it occurs in the database of 4-letter (DB4) with the probability of 50%, we will consider the probability of 50% from DB5 as more trustworthy than the probability of 50% from DB4.We get the priorities from these individual databases. For increasing the speed of searching, we should add a database of prepositions at the beginning of all the databases of suffixes. As soon as the harmony appears in the database of prepositions, the application does not have to search the other databases.

The module tries to find out whether the sentence, which has arrived as input, includes a number. If it does not include the number, it is automatically redirected into the output of the module. If the number is shown, the sentence is divided into words and each word is analysing individually and the databases according to priorities. If the harmony appears, it will save acquired data in the field (Scheme No. 1.). The example: ***"Som na prechádzke s mojimi 3 bratmi."***

Scheme No. 1. The field of acquired probabilities of grammatical categories

	0	1	2	3	4	5	6
0	som	na	prechádzke	s	mojimi	3	bratmi
1	1	2	3	4	5	6	7
2	som	na	ádzke	s	ojimi		ratmi
3	V	E6	Sfs6	E7	Pfp7		Smp7
4	1	0,5	0,824	1	0,319		1
5		E4	Sfs3		Pip7		
6		0,5	0,176		0,284		
7					Pmp7		
8					0,25		
9					Pnp7		
10					0,147		
...	...	...	...	...	...	...	...
39	null	null	null	null	null	null	null

We can start to determine the grammatical categories with this field.

3.3. Algorithm used for determination of grammatical categories

I have divided the determination into several main parts according to the analyses about which principle and procedure to choose:

- to determine the position of the number in the sentence,
- to determine, which words are linked with the number,
- searching for the method for the common grammatical case,
- searching for the method for gender(vitality) and singular/plural.

3.3.1. Determining the position of the number in the sentence

This step is very important to be able to work correctly with words before the number or after the number, eventually to analyse correctly the whole sentence.

3.3.2. Algorithm of continuity

Not every word and preposition in the sentence has to be linked with the numeral. If we separate out the words which are connected with the numeral, we can get the particular grammatical categories for determination of numerals. This block refers the column of words, which are related with the number occurring in the sentence, to the sign "1" in the line 39. As we can see in the Scheme No. 2, the sentence includes two prepositions while the preposition "na" refers to the word "prechádzke" and the preposition "s" to the number. Of course, the preposition does not have to stand closely before the number and there might be some words between them to keep or not to keep the continuity of the case with the preposition. The example:

S mojimi 3 sestrami. – the case when the number is connected with the preposition

S kamarátom a mojimi 3 sestrami. – the case when the number is not directly related with the preposition

This is a simple example when the continuity influenced by word classes. Of course, there are a lot of combinations of when and why, but we have to find the best method to cover a big amount of cases. I have found individual relations between word classes in my study and I have created the algorithm of continuity.
It is divided in two parts:

- searching before a number,
- searching after a number.

Searching before a number is concentrated on prepositions, adjective, pronoun and adverb which can stand before a number and give enough information about grammatical categories of the number while searching after a number is concentrated on searching for the object of the sentence – the noun, which mostly gives information about the gender and vitality. Vitality is one of more difficult problems when determining grammatical categories. It often plays an important role. Vital numerals are pronounced differently from not vital numerals in different word cases. Thus, it is very important to find the maximum amount when searching for words connected with the numeral, but we have to be careful that there are no unfavourable words in the aggregate.
They can cause an unfavourable result while determining.
When the procedure of determining the continuity of words with the number is finished and all the needed columns are marked, we accede to the step when useless information in columns is deleted. It means that all the categories of unmarked columns in the line 39 are deleted. We can see the result in the Scheme No. 2. – columns unmarked by orange colour.

Scheme No. 2. Process of determining grammatical categories

	0	1	2	3	4	5	6
0	som	na	prechádzke	s	mojimi	3	bratmi
1	1	2	3	4	5	6	7
2				s	ojimi		ratmi
3				E7	Pfp7		Smp7
4				1	0,319		1
5					Pip7		
6					0,284		
7					Pmp7		
8					0,25		
9					Pnp3		
10					0,147		
...	...	...	...	...	...	...	...
39	null	null	null	1	1	null	1

3.3.3. Algorithm of searching of the way for a word case

We are trying to find a through way to announce the harmony in one word case in this part. In the most ideal case, as it is seen in the Scheme No. 2 – red data cells, we find a complete way. Not every column has to include determined word case. That´s why it was needed to create tolerance for accepting the disharmony and we start to search with the zero tolerance of the fault. We gradually increase the tolerance to the highest point. When determining the maximum tolerance, it should not be set too high or too low. The low tolerance causes in some cases that it will not be able to determine the word case even if the tolerance exists there. On the other hand, the high tolerance can find more harmonies that it is correct. I have found during testing: if the tolerance emerges from the formula (1), we will get values for tolerance, from which the most successful one is 70%

max_error = a number of words * (1-success) **(1)**

The error is rounded on the integral number. We got the amount of correct words for given success of determining. Finally, unfavourable information, which do not include a determined word case (the 7th word case in our example), is cancelled in the scheme. See Scheme No. 2 – blue cells.

3.3.4. Algorithm of searching for way for gender and singular/plural

In the last step of determination grammatical categories, we accede to the determination of common gender and singular/plural. It is determined similarly as the word case. We have to find the through way again to pass all the words, perhaps the way of tolerance. We can see steps to determine the gender and singular/plural in the Scheme No. 2 – red cells.

4. Translation of numbers

When translating the number into the inflected numeral, the module expects the reached grammatical categories on input. If determining was not set correctly, defaulted grammatical categories would be set.

4.1. Databases

This block deals with databases which were hand-made on the basis of research of morphology and suggested algorithms.
We have 3 databases available:

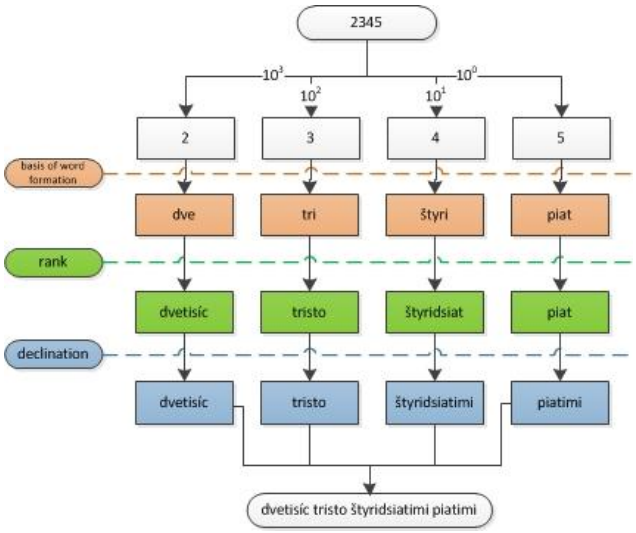
- cardinal numerals,
- ordinal numerals – singular,
- ordinal numerals – plural.

These are composed of a word-forming etyma, suffices and infinitive of words. There are only numbers from 0 to 19 in the databases including the numbers a hundred, a thousand and a million. With the help of these 23 numbers, we are able to create every number to a milliard (an open interval from the right). There is a word-forming etyma at the beginning of each line and all the marshalled suffixes of a grammatical gender and word cases follow it.

4.2. Creation of numeral

The information about acquired grammatical categories are coming into the module of input. Firstly, the application will find out the type of numeral (cardinal or ordinal numeral) because there are different rules and databases for each type. Consequently, the word in the sentence is divided into individual regulations 10^N and every regulation will be seen individually. If the harmony is found in the line, it moves according to acquired gender and case. The gender and the case also determine the shift in the line by given amount of fields. Therefore, we can move absolutely elegantly along the field with suffixes.

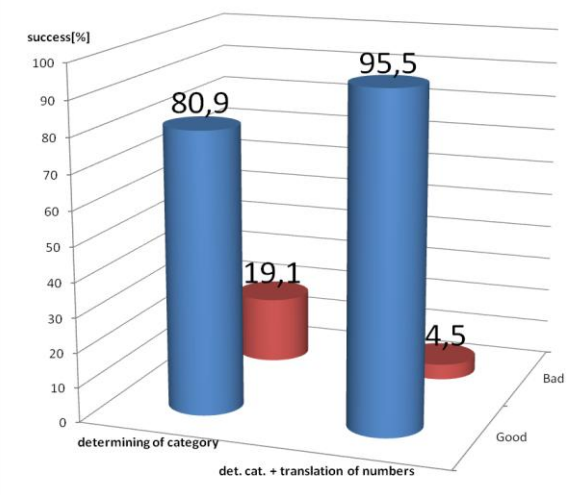
The concrete example is added as the translation takes place in the sentence: "S 2345 ženami.". We get the grammatical categories "fp7" from the previous module, it means f-female, p-plural, 7-instrumental case).



Pics. No. 4 The process of translation of the number into the numeral.

5. Testing the software

While testing reached result, I divided the test into two parts. Firstly, I only tested the module of determining grammatical categories where the following criteria were set: the correct numeral = the numeral, which has all the grammatical categories determined correctly. If one of the categories is determined wrong, the whole determining is considered to be wrong. The test consisted of 110 sentences (only one number exists in a sentence), while the sentences were constructed to combine all the grammatical genders, word cases and singular/plural. Consequently, I checked and evaluated the results manually. In the next phase, I used determined grammatical categories to change the number into the numeral, where the percentage of success was even high because the percentage of wrong determinations consisted only of vitality in a male gender. It was not expressed in the translation itself because the most of vital and not vital suffixes of word cases are in a harmony.



Picture No. 5. The results of testing the modules.

6. Summary

Speech synthesis is a very interesting branch for me because of the science and technology it includes and because of a huge challenge to do the next step in the development of a institute speech synthesizer. The main problem of determining the numerals in a telecommunication branch is not so explored. Every improvement in this branch means a big gain for the next studies in the near future. Numerals as a part of language are inseparable component of everyday communication. The idea of verbalization without numbers is very difficult in some cases. Therefore, I tried to do some improvement in this area in term of research on principles and techniques for processing numerals in a Slovak language and to create a suitable application.

Thanks

This work has been supported by the projects VEGA 1/0718/09 and FP7-ICT-2011-7 HBB-Next.

Reference literature

[1]. Psutka, J., „Mluvíme s počítačem česky", Academia, 2006, s. 746, 80-200-1309-1

[2]. Pauliny, E., „Slovenská gramatika", Slovenské pedagogické nakladateľstvo, 1981, s. 323, Š-7066/1980-32.

[3]. Ondrus, P., „Kapitoly zo slovenskej Morfológie", Slovenské pedagogické nakladateľstvo, 1978, s. 192.

[4]. Uhlíř, J., „Technologie hlasových komunikací", České vysoké učení technické v Praze, 2007, s. 276, 978-80-01-03888-8.

[5]. SAV - Jazykovedný ústav Ľ. Štúra, Slovenský národný korpus, [Online] [Dátum: 27. január 2011], Dostupné z http://korpus.juls.savba.sk

[6]. Oravec, J., „Súčastný slovenský spisovný jazyk, Morfológia", Slovenské pedagogické nakladateľstvo, 1984, s. 232, 67-167-84

[7]. Dvonč, L., „Morfológia slovenského jazyka", Slovenská akadémia vied, 1966, s. 896, 71-024-66

Simulation of Prosody Contours with Embedded Signal Generator

Ján Tóth[1], *Anna Kondelová*[1], *Peter Guzmický*[2]

[1] Institute of telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
[2] Institute of Control and Industrial Informatics, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
jan.toth@stuba.sk, anna.kondelova@stuba.sk, peter.guzmicky@stuba.sk

Abstract. *This work deals with simulation of prosody contours in any Speech Synthesizer. Signal generator is able to generate melody contour on demand. Sentence as an input of Speech Synthesizer is described with many parameters. On this level is worked with phones and their own values (frequencies). These values will be simulated into one continuous melody time progress. The results can be processed in picture form (graph) or sound form.*

Keywords

embeded signal generator; harmonic and non-harmonic signals; prosody contours; PSOLA

1.Introduction

The main goal of this work was description of possibilities existing generator of harmonic and non-harmonic signals with shape of prosody contour. The shape of prosody contour will be defined as a function values in concrete times.

In second chapter is situated description of used generator properties, inner data structure and mathematic description of generated harmonic and non-harmonic signals.

Following chapter deals with prosody. What kind of sentences we know, what's the character of sentence's time progress and how exactly prosody can be changed.

Last chapter includes evaluation of simulation prosody contours with signal generator.

2.Signal Generator

2.1 Generator of process variables

Existed embedded generator is designed to predict and generate analog and digital values and harmonic and non-

harmonic signal functions. The generator which was developed in C++ language works in three steps: (1) generator expects

inputs from users via the graphic user interface (required types and values of process variables), (2) data defined by user are processed and optimized on a background, and (3) user determines time needed for duration of simulation. Generator can be seen on Fig. 1. [1]

Fig.1. Block diagram of Generator.

On Fig. 2 is displayed inner data structure of generator (columns corresponds to the individual steps).

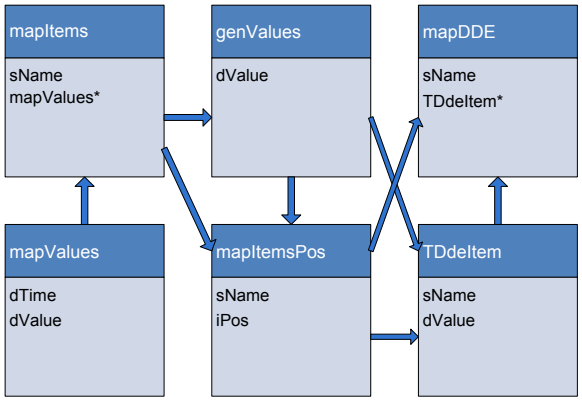


Fig.2. Inner data structure.

Generator process is separated on three parallel threads (see Fig. 3) for optimal application run. MainThread is parent of the other two threads and it is responsible for generator process and graphic user interface. Timer1Thread has timer function and generates output values for individual DDE (Dynamic Data Exchange) variables. *Timer2Thread* has also timer function and ensure presenting actual DDE variables values in graphic user interface.

Fig.3. Block diagram of parallel running threads.

2.2 Generator Realization

It's used DDE communication protocol created by Microsoft. On mutual communication participate client and server application. On Fig. 4 is displayed window of graphic user interface. [5]

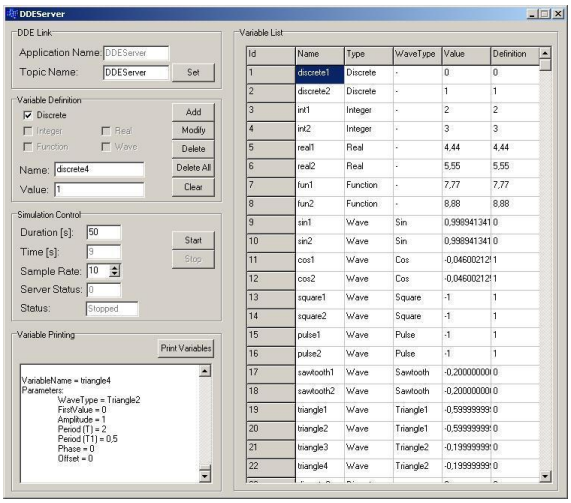


Fig.4. DDE screen.

In next figure (see Fig. 5) can be seen defining of static and dynamic variables.

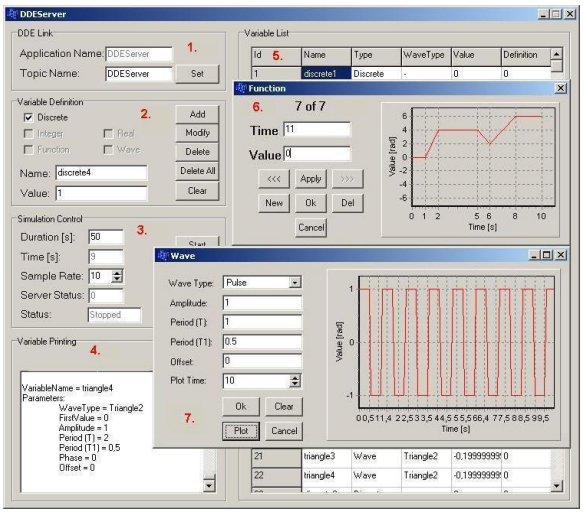


Fig.5. DDEServer Generator with numbered blocks screen.

2.3 Generator of Signal Variables

In Fig. 6 is displayed function part, which is able to define and correct parameters of signal variables. [1]

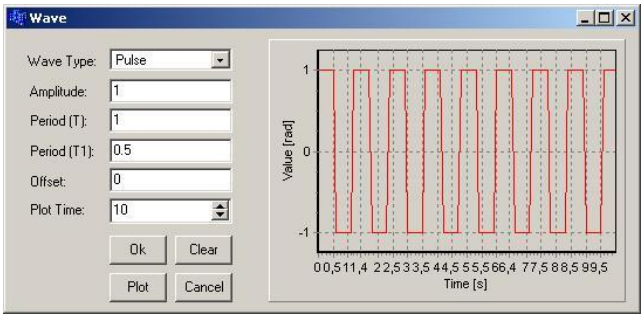


Fig.6. Parameters definition and correction of signal variables.

Sinus

Sinus is a mathematical function defined in (1) and it's used in mathematics, physics, signal processing, electrical engineering.

$$y(t) = A\sin(\omega t + \varphi) + D \qquad (1)$$

where: A – amplitude

ω – angular frequency [rad.s^{-1}]

φ – phase

D – offset, nonzero amplitude middle (DS signal component) [7]

Cosinus

Cosine (2) is also harmonic function. It could be expressed like sinus with shifted phase π/2.

$$y(t) = A\cos(\omega t + \varphi) + D \qquad (2)$$

Square signal

It's non-harmonic periodical signal (also called Rademacher function) compounded from prompt change between 2 levels (see Fig. 7).

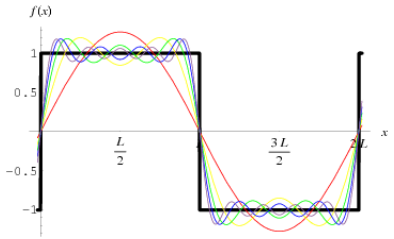


Fig.7. Graph of Square signal.

Pulse signal

Pulse signal on Fig. 8 is used to information transmission via communication channel. It's special type of square signal.

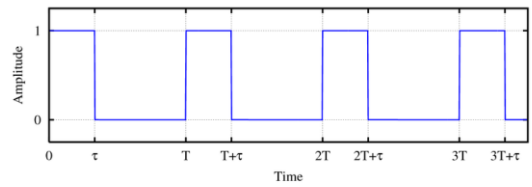


Fig.8. Graph of Pulse signal.

Saw-tooth signal

On Fig. 9 can be seen one period of Saw-tooth signal.

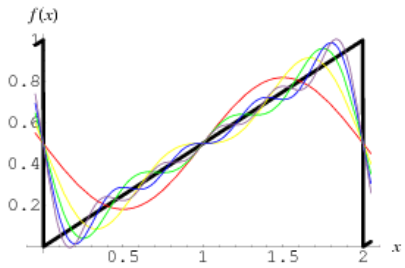


Fig.9. Graph of Saw-tooth signal.

Triangular signal

On Fig. 10 is displayed symmetric triangular signal. It's odd function.

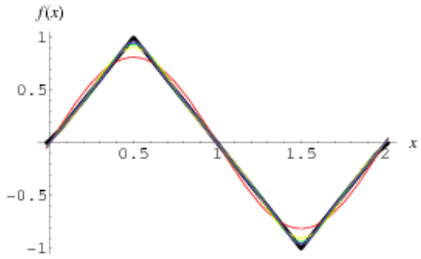


Fig.10. Graph of Triangular signal.

3.Prosody

Prosody is basically melody of continuous speech. It's needed to include prosody information except information about synthesized text in text synthesizer. Only with prosody can be expected better result as unnatural voice. [6]

It's the biggest problem to estimate the right shape (it's called melody contour). Every person has his own fundamental frequency and such melody contour is than modulated on concrete fundamental frequency. [4]

Melody of sentence has generally decreasing character. As below in sentence:"We came through the break. on Fig. 11.

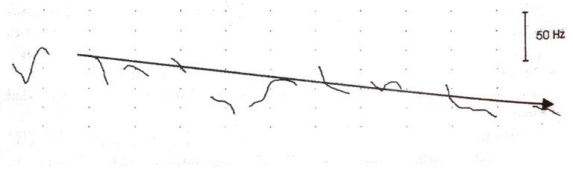


Fig.11. Time progress of melody.

But it's necessary to tell that questions have increasing character as on Fig. 12 what is short question:"Yes?".

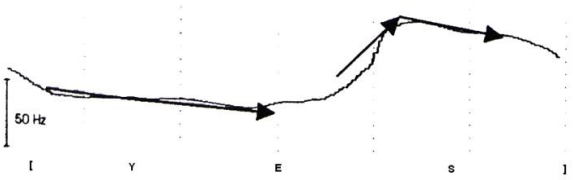


Fig.12. Time progress of question's melody.

Five different types of melody contours in Slovak were detected in short analysis, which have to be somehow simulated. [3]

The prosody after predicted melody contour has to be changed to achieve new prosody, correct prosody. The changing can be made with many possibilities. Here below is described PSOLA algorithm, which one can be used. [8]

PSOLA

This technique is often used in speech processing to change the pitch of a speech signal without affecting its duration. A very simple technique to modify the pitch would be by changing the duration of the speech signal, lengthening it to decrease the pitch and shortening it to increase the pitch.
In PSOLA, the speech waveform is first divided into several small overlapping segments and the segments are then moved closer or apart depending on whether to increase or decrease the pitch (see Fig. 13). [2]

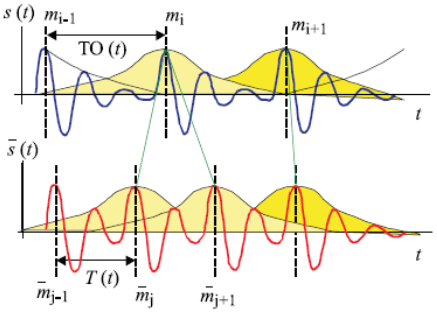


Fig.13. Sketch of moving segments closer to each other.

4.Simulation prosody contours

The research shows that the most natural prosody has two intonation peaks: one at the beginning of sentence and the other at the end. The prosody contour is drawn with blue color on Fig. 14. [9]

Prosody changing executes in two steps: melody contour's simulation on demands and itself prosody changing (PSOLA). Melody contour is defined with each phone's value (frequency) in sentence. So if the sentence has (see Fig. 14) 18 phones the signal generator will have 18 demands, 18 values. Important is at first to define the type of sentence, to make it simply, let's say question or declarative sentence. Very important is here to know the punctuation.

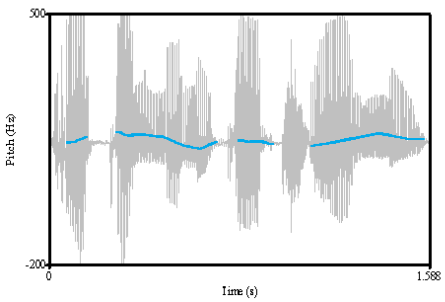


Fig. 14. Time progress of melody in question: „Is this monitor black?"

Existing signal generator will be extended with generating shape of prosody contour by defining function values in concrete times. The results will be displaying shape of prosody contour in the chart and playing sound sample of the final contour.

5. Conclusion

Subject of this paper is to implement generating of prosody contour by embedded signal generator. Signal generator will be able to create prosody contours by putting together all necessary signal types. Idea was created, when we made algorithm for changing outgoing prosody from synthesizer. We tried to predict prosody contour based on knowledge obtained in past. This method connects theoretical knowledge and experience gained in practice.

Acknowledgements

This work has been supported by the Grant Agency of the Slovak Republic, VEGA 1/0718/09, VEGA 1/1105/11, DAAD and HBB-Next.

References

[1] P. Guzmický, "Simulator of technology processes for visualization tool InTouch", diploma thesis, Bratislava, May 2010.

[2] A. Mousa "Voice Conversion using Pitch Shifting", Journal of ELECTRICAL ENGINEERING, VOL. 61, NO. 1, 2010, pp. 57-61.

[3] A. Kondelová, "Analysis of prosody features in Slovak", diploma thesis, Bratislava, May 2010.

[4] J. Čepko, M. Turi Nagy, G. Rozinaj, "Low-level synthesis of Slovak speech in S2 synthesizer", In Proc. of 5th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services, 29 June - 2 July 2005, Smolenice, Slovak Republic.

[5] Š. Kozák, "Linear digital systems 1", STU, Bratislava, 1991.

[6] J. Tóth, "Phonetic transcription of abbreviation for speech synthesis", diploma thesis, Bratislava, May 2010.

[7] R. Lasser, "Introduction to Fourier series", M. Dekker, New York, 1996.

[8] A. Kondelova, J. Toth, G. Rozinaj, "Analysis of prosody features in Slovak", In Proc. of the 52nd International Symposium ELMAR-2010, Zadar, 2010.

[9] G. Rozinaj, J. Vrabec, J. Čepko, R. Talafová, "Terminals for the Smart Information Retrieval", In Ismail Khalil Ibrahim (Ed.): Handbook of Research on Mobile Multimedia, Second Edition, Chapter XIX, IGI Global Publishing, 2008, ISBN: 978-1-60566-046-2, pp. 451

Modular Speech Synthesizer

Anna KONDELOVÁ[1], *Ján TÓTH*[1], *Ivan DROZD*[1], *Tomáš HORVÁTH*, *Matej SEMBER*[1], *Michal SERAFÍN*[1], *Matúš ŠOKA*[11]

[1] Institute of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
anna.kondelova@stuba.sk, jan.toth@stuba.sk

Abstract. *This article is concerned with the topic of the creation/production of a modular speech synthesizer. It concentrates on the possibilities of the individual modules needed for the creation of a functional modular synthesizer, possibilities of extending the synthesizer. Furthermore it looks at the requirement of their mutual communication, communication with the operating module and communication of the user with the user interface.*

Keywords

Speech synthesis, TTS synthesizer, modular synthesizer, XML-RPC communication, TCP server

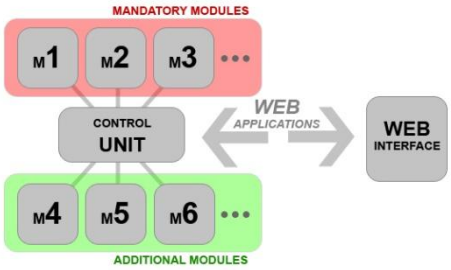
1. Introduction

The creation of a modular synthesizer is a team work project at department of telecommunication. Modular synthesizer is a programme designed for speech synthesis, in which a number of modules work together to ensure the individual parts of the process of creating the speech signal. (see Fig. 1.)

The success of the project lies in the creation of a system with good communication abilities between the modules and the operating interface. The end result should be further extensible to more modules that will increase the quality of synthesis to the required level.

2. Modules

For our project we have worked with the modules created at department of telecommunication under the bachelor and master projects. The fundamental modules are notably the transcription and synthesis modules. The other modules ensure the processing of specific cases of text such as abbreviations or numbers. Some modules need further modules for the use of Slovak language, for example the module for rewriting of abbreviations requires one for the determination of word class, correct word form and thus for the best synthesis possible. Slovak language is very sensitive to the use of the correct word forms.



.

Fig. 1 Block diagram of modular architecture of speech synthesizer.

2.1 Web Interface

Below web interface we can imagine an application run in a Web browser, whose task is to send requests for the synthesis and resynthesis, a direct response from the control module and in an appropriate way to interpret them. Web interface is an application that provides an interface between human and synthesizer (Flash application, java applet etc.). After loading the application from the web server and entering the text on the synthesis, is made TCP connection to port forward configured between web application and a control module, through which they exchange messages in these formats. For testing purpose, we developed a Java application that allows us to suffice this purpose.

```
<?xml version="1.0" encoding="utf-8"?>
<query type="new">
  <sentence value="Príjemný" />
</query>
```

Fig. 2 Format requirements for the synthesis of sent Web-based management module.

```
<?xml version="1.0" encoding="utf-8"?>
<query type="update">
  <sentence duration="2.603" wavurl="http://<IP>:<port>/id0.wav" value="...">
    <syllable text="pri" position="0" regPoint="0.6" value="115" />
    <syllable text="je" position="83" regPoint="0.3" value="106" />
    <syllable text="mny" position="123" regPoint="0.9" value="101" />
  </sentence>
</query>
```

Fig. 2 The answer format sent to the management server management module.

The above formats (see Fig. 2 and Fig. 3) are taken from existing project in the Institute of Telecommunications.

Our control module is still able to process only synthesis because the new format of resynthesis needs diphone database (with prosodic features), which we do not have available.

2.2 The Transcription Module

This module carries out one of the most important tasks in our system, namely the rewriting of the given text into SAMPA alphabet. The synthesis module can work well with this alphabet as opposed to the plain text. It is the last module to interfere with the xml file which is being exchanged between the individual modules via the operating module. Its output xml file that creates a tree-like hierarchical structure of the individual sentences received for processing from the web interface (see Figure 4.). [3]

```
<?XML VERSION="1.0" ENCODING="UTF-8"?>
<TEXT>
 <SENTENCE TEXT="PRÍJEMNÝ" DELIM=".">
  <WORD VAL="PRÍJEMNÝ">
   <TRANSKRIPT>DICTIONARY</TRANSKRIPT>
   <SYL>
    <PHO>P</PHO>
    <PHO>R</PHO>
    <PHO>I:</PHO>
    <PHO>J</PHO>
    <PHO>E</PHO>
    <PHO>M</PHO>
    <PHO>N</PHO>
    <PHO>I:</PHO>
   </SYL>
  </WORD>
 </SENTENCE>
</TEXT>
```

Fig. 4 XML file format of transcription module output.

2.3 The Synthesis Module

There is a diphone TTS synthesizer, the output of which is a sound track in the WAV format. The module of the speech synthesizer is the XML-RPC server, receiving the incoming xml file in the body of the http request using the XML-RPC protocol. The synthesis module processes the requests and returns xml file with the phoneme boundaries and basic prosodic characteristics for basic voice frequency modification, energy and accordingly others, the functionality of which can be brought into the system using additional modules, as output. Apart from the prosodic characteristics this xml-file also contains the relative way to the synthesized recording that the web interface is later to process. The synthesis module is based on the principle of selecting speech segments from the diphone database and their subsequent connection (see Fig. 5). [1]

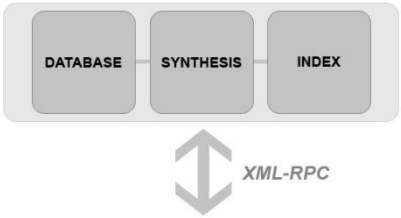


Fig. 3 Speech synthesis module block diagram.

Format of the outgoing xml-file from the synthesis module is the following one:

```
<?XML VERSION="1.0" ENCODING="UTF-8"?>
<TEXT WAVURL="C:\SYNTHSERVER\OUTPUT\ID0.WAV">
 <SENTENCE TEXT="PRÍJEMNÝ" DELIM=".">
  <WORD VAL="PRÍJEMNÝ">
   <TRANSKRIPT>DICTIONARY</TRANSKRIPT>
   <SYL>
    <PHO>P
     <BOUNDS>
      <START>16603</START>
      <END>13545</END>
     </BOUNDS>
    </PHO>
    <PHO>R
     <BOUNDS>
      <START>13545</START>
      <END>14902</END>
     </BOUNDS>
    </PHO>
    <PHO>I:
     <BOUNDS>
      <START>14902</START>
      <END>26365</END>
     </BOUNDS>
    </PHO>
    .
    .
    .
   </SYL>
  </WORD>
 </SENTENCE>
</TEXT>
```

Fig. 4 XML file format of speech synthesis module output.

As it can be seen from the example (see Fig. 6), the synthesis module fills in the incoming xml-file and the way where the synthesized recording and phoneme boundaries in the synthesized recording are. [2]

2.4 The Speech Prosody changing Module

An important part of the creation of synthesized speech is its prosodic arrangement. Prosody is that part of speech which adds the sentence melodies their declarative value. In most cases in Slovak it depends on the type of sentence (question, answer, announcing sentence). It is based on the modification of some speech parameters such as vocal cords frequency, rhythm, volume, length of pauses between words. We need to modify these parameters in particular for a good speech synthesis. Following such speech modification we should obtain the result of naturally sounding human speech. We are planning to add the prosodic modification module to our system only later. These are some solutions that extend the lifelikeness of the whole modular synthesizer. They include two parts. The first is text analysis and setting of the correct sentence melodic characteristics. The second one is the possibility of adjusting the prosody by the user himself using web flash interface. Whilst the first is a necessary part of a good speech synthesizer, the second is its extension which allows the users to change the parameters in such way as to adjust their synthesized text further according to their requirements. [4]

3. The Control Module

The Control Module or Multiclient is a heart of this project, because is the part which was designed only by us. The task is to receive requests from user and communicate with individual modules, to receive, resend and elavuate their requests. If you design this module, you have to deal with bigger number of requests, to create a sophisticated way how to manage requests to use functional modules of modular synthesizer most effectively.

The control module has to recognize between individual modules because there are two types of requests, which could the control module receive (synthesize and resynthesize). The control module has exactly to evaluate, which module deals with specific request.

3.1 Operating principle

After starting up the control module is this module in passive mode and it is necessary to configure it. In module is needed to set the port, where module can receive the requests on synthesis, maximal length of request, maximal and minimal number of underthreds, which manipulate the requests.

After configuration is necessary to initiate the registration server. When this server starts up, the other individual modules register on the control module. Under registration we understand sending necessary information on XML-RPC client establishment. When all individual modules are registered is the control module able to use their functions and can be enabled the request control. Under enabling request control we can understand disconnection of registration server and connecting TCP server through which directly receives requests on synthesis.

3.2 Graphic interface

Graphic interface of our modular synthesizer tries to be user friendly, but on the other hand interface has to allow possibility of the system for many conditions (see Fig. 7). [5]

3.2.1 Parts of graphic interface

- Tray – After application initiation is created instance of the class SystemTray resulting creation of application running on background and minimlised on a system tray.
- Status – Status window, where the program inserts information about correct or incorrect operation or about mistake, which could appear in communication between individual modules. After tab switching the status window displays current connected modules.
- Settings – Interface whereby it is easy and intuitively to set registration port, port where TCP server will expect requests, maximal number of requests in queue, maximal number of underthreads on request control and minimal number of underthreads on request management.

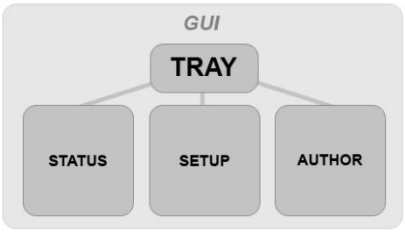


Fig. 5 Block diagram of user interface.

4. Inter modules communication

Our main effort was to create universal system not having hard programmed communication between modules. Program was designed with mind to create system with easy adaptation on demands without necessity to interfere into source code. The main component providing versatility is variable number of modules, with which is control module able to communicate. After starting synthesis module is module able to register on control module and then is created client.

Compatibility of communication between modules is ensured through standardized XML formats. The output of one module must be compatible with further input etc. Because the project brings together a larger number of modules, which have been created in the Institute of Telecommunications, it was necessary to solve the mutual compatibility between different modules. These formats can be easily, if necessary, modified by adjust or exchange by classes for creating and processing XML documents in individual modules.

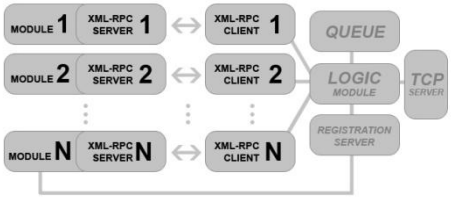


Fig. 6 Detail block diagram of modular architecture of speech synthesizer.

4.1 XML -RPC communication

Modules were programmed in many languages so it was necessary to ensure compatible way of communication. For this purpose we choose XML RPC protocol, whichis easy to implement, extensible and allows ensure mutually communication between used modules.

4.2 Communication with control module

As was mentioned the communication between the control module and synthesis module runs through XML RPC connection. XML RPC is Remote Procedure Call to call remote procedure. In term remote procedure we understand procedure (function alternatively method in OOP) which is called in other code. The remote procedure is called at the hand of http request in server, which implements this protocol. Through http request client server sends procedure parameters in xml format. As soon as server, which implements XML RPC protocol, obtains request, procedure will be called with accepted parameters and the result of operation will be send in http answer back to client. (see Fig. 8)

4.3 Principle of synthesis

User trying to synthesize text can connect through TCP client on server and sends text for synthesis in adequate XML form. TCP server receives the message, determines which type of message is it (synthesis or resynthesis) and includes into queue of requests. In queue of requests waits until synthesizer resources allows handling the request (synthesizer can be occupied with other request). [7] In case that synthesizer is free (isn't occupied with any request) from queue (FIFO stack) picks up another request. (see Fig. 9)

– Synthesizer send request to first module and subsequently receives answer
– Answer from first module is then send to second module and next receives an answer
– Module goes on this way until the request is processed with all available modules
– After request processing module sends through TCP server an answer

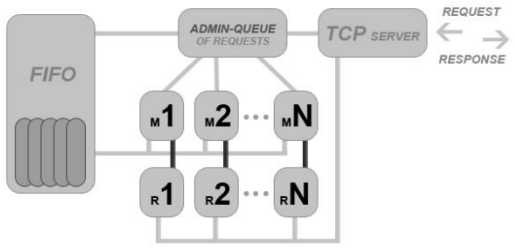


Fig. 7 Block diagram of report management, Mn -n-th thread, Rn - operation of n-th requirement.

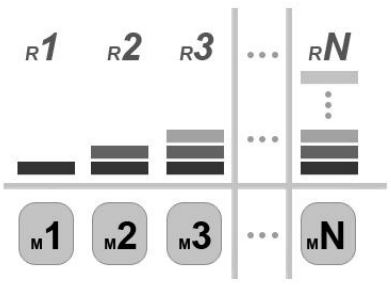


Fig. 8 The processing of requirements in modules in time=t, Mn - n-th module, Rn - n-tn requriment.

On picture (see Fig. 10) is represented way of request processing by individual modules. Received request is firstly processed with first module and progressively with all other modules. Processing is designed to be occupied every module in given moment (after releasing of module is again occupied with other request). Every module enriches outgoing XML file. This added information into XML file is subsequently used by other modules. [6]

5. Conclusion

Our goal was to create a functional modular synthesizer from modules which were created on the Department of telecommunications. We had to find solutions for problems with communication using http servers, writing data into xml file, compatibility and extensibility of the whole system. We solved all these tasks and our system is able to synthesize sentence or statement input into http interface. Final quality of the synthesized speech is equal to quality of used modules. Improvement of modules wasn't part of our team project but other projects on department of telecommunications. This project has different options of extensions and improvements, e.g. in communication between used modules, adding new modules, in part of resynthesizes and lot of others.

Acknowledgements

This work has been supported by the projects VEGA 1/0718/09 and FP7-ICT-2011-7 HBB-Next.

References

[1] VACHO, I., *ARCHITEKTÚRA SYSTÉMU PRE SYNTETIZÁTOR SLOVENČINY*, Diploma thesis, may 2010.

[2] HEROUT, P., *Java a XML*. ISBN: 80-7232-307-5, KOPP, 2007.

[3] TOTH, J., *Fonetická transkripcia skratiek pri syntéze reči.* Diploma thesis, may 2010.

[4] KONDELOVA, A., *Analýza prozodických vlastností slovenskej reči.* Diploma thesis, may 2010.

[5] HEROUT, P., *Java – grafické užívatelské prostředí a čeština.* ISBN-8072322370, KOPP, 2006.

[6] BLOCH, J., *Java efektivně.* ISBN- 8024704161, Grada, 2002.

[7] ROZINAJ, G., Towards More Intelligent Speech Interface. 17th International Conference on Systems, Signals and Image Processing: IWSSIP. ISBN: 978-85-228-0565-5, Rio de Janeiro (Brazil), 2010. p. 49-52.

The use of IRKR system for service resembling library

Matúš TICHÝ[1]

[1] Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
tichy.matus@gmail.com

Abstract. *The aim of this paper is to explain the functionality of IRKR system (intelligent speech communication interface) with emphasis on use for library application service. The IRKR, a modular system using speech synthesis and speech recognition software, is build to communicate with user in order to provide him certain services. In this paper we will at first introduce the system IRKR and its architecture. We will show how it works and what are the requirements and software dependencies necessary for proper execution of this software. Then we will explain the functionality of the dialog of the whole system and the types of files used for the dialog execution. We will introduce the library application and its connection to the IRKR system. We will look at modifications of some parts of IRKR system, especially modifications inside of dialog and grammar files, which were necessary for the implementation of the library service.*

Keywords

Speech analysis, spoken dialog systems, voice xml,

1. Introduction

The system IRKR – Intelligent Speech Communication Interface demonstrates one of many uses of speech analysis which is the use for spoken dialog systems. It is a dialog system delivered through voice. Call center applications, some entertainment or chatting applications, informational or transactional applications are well known examples of dialog systems where spoken dialog goes between an automated system and human user. IRKR system was created to offer information about weather forecast and transport schedule services. We tried to implement a new service into the system which should provide information on available books in library.

2. System IRKR

The IRKR is modular system which means that each module of the system holds responsibility for certain functionality. Modules communicate with each other and exchange information according to beforehand specified rules.

2.1 IRKR modules

There are two modules determined to speech analysis. One is used for speech synthesis which processes prescribed dialog parts to provide the user with spoken part of the dialog. Second module recognizes speech of the human user and generates text used by other modules. This text form of user spoken words is decisive in determining the following direction of the dialog. Module called Dialog Manager (DM) controls the flow of the dialog according to rules taken from certain configuration files. Gateway is the next module which can be viewed either as an interface between the system and client telephone used by user or as a manager of the session with user. The registration of modules and their communication by internal messages is managed by the module named Hub. The Fig. 1 shows schematic sketch of the IRKR modular architecture. The sketch shows existing communication flows between modules and protocols used for their realization.

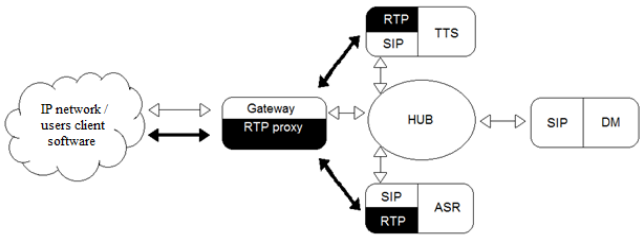


Fig. 1. Modular architecture of IRKR system

The system is configured using config.xml files for each module. In these files lies the specification of communication ports of each module, their SIP address and other necessary parameters. Exchange of control information is realized using SIP protocol messages and exchange of useful data is done with RTP protocol.

To run the system properly it is recommended to use Windows XP with installed service pack for MS Visual Studio 2005. Newer versions of Microsoft operating systems can be used but are not recommended because of possible incompatibility with Visual Studio version.

2.2 Dialog of the system with user

The dialog of the whole system is driven by decisions of DM based on information taken from specific vxml files. VoiceXML (VXML) is the W3C's standard XML format for specifying interactive voice dialogues between a human and a computer. It uses simple tags similar to HTML tags but they are used to specify whether the text enclosed by the opening and closing tag is to be synthesized or if it is a text which takes part in the speech recognition. There are also tags used to specify grammar used during the recognition process or tags for redirecting the dialog flow to other part of the file or to other vxml file.

3. Dialog system with service library

Let us make an example for better understanding of the dialog system. User makes call to IRKR dialing on his softphone application the SIP address of Gateway module with correct port. After the connection being established the session starts between the user and IRKR. User is greeted by the system and is offered the option to listen to the help/manual. Then the system waits for users answer. If the user answers YES, the system tells him the list of actual services with their description. This part of the dialog is described by the file named start.vxml. After listening to the help/manual or if the users answer to the systems first question was NO the system is redirected to file option.vxml which determines the next flow of the dialog. This time is the user asked by the system to choose one of offered services. According to users answer which has to be recognized by the responsible module the system is redirected to another vxml file.

3.1 VXML files

All vxml files present in the system use already mentioned simple tags to ensure its functionality. Line containing the <prompt> tag is used for the question. The text enclosed by the opening and closing tag is sent to the synthesizer module and after being processed is received by user as audio (see Ex. 1).

```
<prompt count="1" bargein="false">
Zvo¾te si službu. POÈASIE CESTOVNÝ PORIADOK</prompt>
```
(1)

The tag <if> is used for the answer and redirection (see Ex. 2). The word between apostrophes in the opening tag is involved in comparison to the recognized answer given by user. If the words equal then the <goto> tag is taken into consideration and the system follows the direction given by the tag, which in this example is the file mixed_initiative.vxml.

```
<if cond="chosenService=='weather'">
    <goto next="applications/weather/mixed_initiative.vxml"/>
</if>
```
(2)

These vxml files can be considered as rules according to which the DM module manages the whole dialog. The amount of necessary vxml files depends on the programmers decision, whether he wants to divide the succession of the dialog into more files. The system must read and store in memory each file it needs for the dialog at the moment. It is wise to free the memory that was reserved for the file after the dialog moves to the next vxml file. From this point of view it is better to have more number of smaller vxml files than one large so that the whole system consumes the smallest possible amount of memory. It comes into consideration how the process of reading the vxml files and freeing the memory after their use slows down the whole run of the system. Since we do not know the answer now this should be next step of development of the system which we can classify as optimization phase.

3.2 Library service as part of IRKR

In order to implement a new service to the system we had to add lines with similar tags as those used in examples but with changes made to provide library service we named books (KNIHY).

```
<prompt bargein="false">IRKR portál ponúka služby POÈASIE CESTOVNÝ PORIADOK a KNIHY.</prompt>

<prompt bargein="false">Služba KNIHY vám umožòuje nájs□ □ knihu v zozname knižnice.</prompt>

<prompt count="1" bargein="false">Zvo¾te si službu. POÈASIE CESTOVNÝ PORIADOK alebo KNIHY. </prompt>
```
(3)

From this example (see Ex. 3) we can see how the letters with diacritics are represented in the language of vxml. According to these changes the system awaits one of three phrases POČASIE, CESTOVNÝ PORIADOK, KNIHY as users answer to the question of service choice. That means we had to make sure the recognition of word KNIHY worked. To do this it was necessary to add the word KNIHY to the grammar files. The file final.dic holds all phrases that the recognizer should await as users input with their transcription to something very similar to SAMPA (Speech Assessment Methods Phonetic Alphabet) which is computer-readable phonetic script used to differentiate all possible characters in certain language. We have added the needed line “knihy k nm i h i” to this file. This specifies a new word in the dictionary which is essential for selecting the service named KNIHY we implemented according to users requirement. Second file service.xml.bnf provides smaller set of phrases which are the possible user answers at the exact point of the dialog. This types of bnf files are processed by small program called HParse.exe with the output of slf type of files. These files are necessary for the right execution of the recognition. The last files which hold the grammar for the

certain point of the dialog are xml files (they contain the list of possible user answers that are linked to already described vxml files). Examples of such file could be train_stations.xml which lists all possible stations provided in the transport schedule service or date.xml file which collects all needed grammar for the correct date selection.

We have already introduced the dialog capabilities with respect to the use of our new library service. We have given the service a working title books (KNIHY) since it should provide the user with information about requested book. At the moment the service works with manually created list of books. The realization of the list is done using xml files. The requirements on the service as part of the IRKR system are following:

- User is guided through the dialog of the service by the systems queries
- Users answer is properly recognized or he is instructed to use other phrase (to preserve the flow of the dialog)
- After proper recognition the user is given requested information and is guided to the next part of the dialog
- User should have the option to voluntarily finish the service or the whole session

The flow of the dialog we ensured by the vxml files. After the user chooses the service named KNIHY the system redirects to the file books_start.vxml. This file specifies the dialog of our library service. The file starts with vxml version tag. This is followed by the greeting text of the service and request for the name of the book. These are sent to synthesizer using the tag <prompt>. Then follow <if> tags which ensure the comparison of the recognized users answer with the names of books taken from the list stored in books_gram.xml file. This file is linked to the dialog by the use of <grammar> tag in the books_start.vxml file. If the comparison output is true the users request is confirmed and he is ensured that the book was found in the list. The file books_gram.xml is typical grammar file used in system. In this case the word grammar means list of words from which one of them is the awaited recognizable answer of the user. Grammar files provide the system with necessary set of words or phrases during the whole dialog. This means that for each part of the dialog there has to exist one grammar file. Such file can be associated with more than one part of the dialog. For example every time the system awaits the users answer to be YES or NO the grammar file yesno.xml is requested by the dialog. The specification of certain file connection to certain part of dialog is set in the vxml files. For better understanding of the .vxml files see the example in Fig. 2.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<vxml version="1.0" lang="Slovak">
 <form id="begin">
   <block>
     <prompt bargein="false">Vitajte v službe KNIHY.</prompt>
     <goto next="#inputData"/>
   </block>
 </form>
 <form id="inputData">
     <field name="bookname">
      <grammar src="grammars/books/books_gram.xml"
type="application/grammar+xml"/>
      <prompt count="1" bargein="false">Prosím, zadajte
názov knihy.</prompt>
       <filled>
        <if cond="bookname=='Meno Ruze'"/>
          <goto next="applications/option.vxml"/>
        <elseif cond="bookname=='koniec'"/>
          <goto next="applications/goodbye.vxml"/>
        <else/>
        </if>
       </filled>
     </field>
 </form>
</vxml>
```

Fig. 2. Simple example of the books_start.vxml file

We have already mentioned that to choose the service books (KNIHY) the word KNIHY needs to be contained in the dictionary files. In order to secure requested functionality of the recognizer during the part of the dialog when service KNIHY runs it is essential that all names of listed books need to be present in the dictionary files final.dic and books_gram.xml.bnf/slf.

From the previous text it is clear that vxml files used for the dialog have strictly prescribed succession. That means that certain files are accessible in the dialog just from files that allow it. This succession is set by <goto> tags present in the vxml files. Whether an action of redirection is done (according to these <goto> tags) usually depends on the users answer. Mostly the dialog is redirected between vxml files but in some cases the dialog is directed to other point inside the same file.

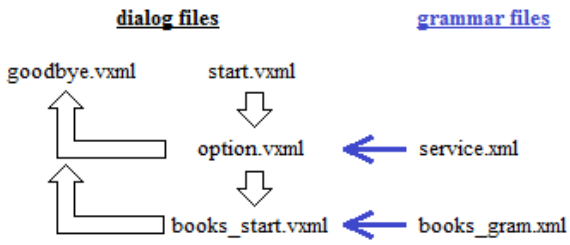


Fig. 3. Dialog files succession for library service

Figure 3. shows which files can be accessed from which and what grammar files are connected to them. For example during the execution of the dialog from the file option.vxml the file books_start.vxml can take the responsibility for the dialog and can redirect the dialog to be executed from the goodbye.vxml file (this will lead to end of the IRKR session with user). We also see that it uses

the grammar file books_gram.xml. This would happen when the user would ask for the service KNIHY and then would request to end the session by saying word KONIEC.

4. Conclusion and possibilities of next development

The functionality of the system is at the moment incomplete. So far the system offers three services: weather forecast, transport schedule and library. Our goal was to implement the library service which should provide the user with information about requested book from the library list. The service just lets the user know if requested book is in the library list at the moment. Next step in development of the service should be implementation of more complex functionality. The service should be able to provide more information about listed books such as the author, year of release or maybe short description of the book. It would be interesting also to improve the whole systems behavior for example the speech analysis part. To do this it would be necessary to analyze and document in detail the action and work of speech processing modules.

Acknowledgment

This work has been supported by the projects VEGA 1/0718/09 a FP7-ICT-2011-7 HBB-Next.

References

[1] ADAM H., DAVID C. *Definitive VoiceXML*. Prentice Hall Professional, 2003.

[2] Voice Extensible Markup Language (VoiceXML) 2.1. W3C Recommendation. http://www.w3.org/TR/2007/REC-voicexml21-20070619/ , 2007

[3] VoiceXML Development Guide. http://www.vxml.org/ , 2007

[4] VLASÁK, J. Využitie protokolu SIP pre komunikáciu IRKR. Diploma project. FEI STUBA, 2008.

[5] JANÍK M. Integrácia protokolu SIP do modulov IRKR. Diploma project. FEI STUBA, 2008.

[6] JUHAR, J. - CIZMAR, A. - RUSKO, M. - TRNKA, M. - ROZINAJ, G. – JARINA, R.: Voice Operated Information System in Slovak In: Computing and Informatics, Volume 26, 2007, Number 6, ISSN 1335-9150

Concept Design of configurable GUI for Speaker Verification Software VeriSp

Juraj VOJTKO[1], *Patrik PIDA*[1]

[1] Institute of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
juraj.vojtko@stuba.sk, patrik.pida@gmail.com

Abstract. *This paper describes software for speaker verification VeriSp. Graphic user interface of this software provides insufficient options mainly of configuration, goal of this paper is point at GUI disadvantages and propose improvements making VeriSp more configurable and confortable for users.*

Keywords

Speaker identification, speaker verification, VeriSp, GUI, Matlab

1.Introduction

Speaker identification and speaker verification is one type of speech signal analysis where machine should ask question "who is the speaker?" or "is it speaker XY?". Human speech is one kind of biometrical data. Using only speech as biometrical data has some disadvantages, like changes speech characteristics when user is sick or changes according to psychic tendency. Therefore it is suggest using combination of speaker identification and some other identification technic.

2.Speech Processing

Generally, we can divide speech processing into two major groups – speech analysis and speech synthesis. Low level analysis or simply called analysis aims extract relevant information from signal to next processing. This process of feature extraction is called parametrization. Higher level of speech analysis process includes speech recognition and speaker recognition (identification and verification). [4]

Typical application using this approach is Voice Operated Information System in Slovak [10].

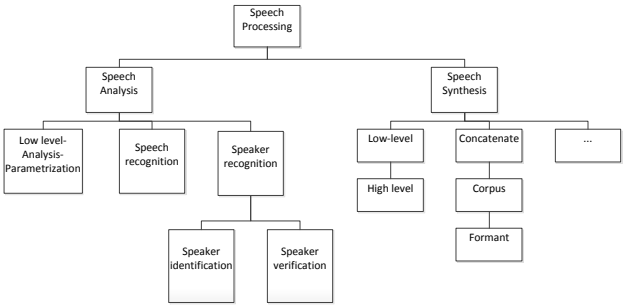


Fig. 1. Schema of speech processing scope

The goal of speech recognition is provide information what was uttered. There are a few approaches to reach that, mainly used are statistical methods, using in [5] [6] [7], and methods based on pattern compare.

As it was written above in Introduction, speaker recognition solves two problems: get information who is speaking or confirm identity of speaker. The main difference between identification and verification is that verification know declared identity of speaker which have to confirm opposite to identification, where system have to designate identity. It implies that speaker identification application can be used to verification, too.

3. VeriSp

VeriSp is speaker verification software working in mode of identification. Its name originates from the first characters of words "VERIfication of SPeaker". It has been developed during two diploma thesis on Dept. of Telecommunication FEI STU in years 2008-2010. In the first phase, neural networks approach was implemented [2], in the second phase Support Vector Machine approach extended the VeriSp [3].

VeriSp was developed in mathematical-simulation environment MATLAB [8] with utilization of toolboxes for neural networks and Voicebox [9].

3.1 Current state

Actual VeriSp version supports 3 types of parametrization:

- Linear Predictive Coding (LPC)
- Cepstral Coefficients (CC)
- Cepstral Coefficients in Mel frequency scale (MFCC)

There are two "classification" methods:

- Neural networks (perceptron)
- Support Vector Machine (SVM)

Speech input options:

- online - from microphone
- offline - read from wav file

Graphic user interface (GUI) is divided into 3 parts:

- Speaker verification
- Add speaker to using network
- Batch verification

Figure 2 shows current version of VeriSp's GUI.

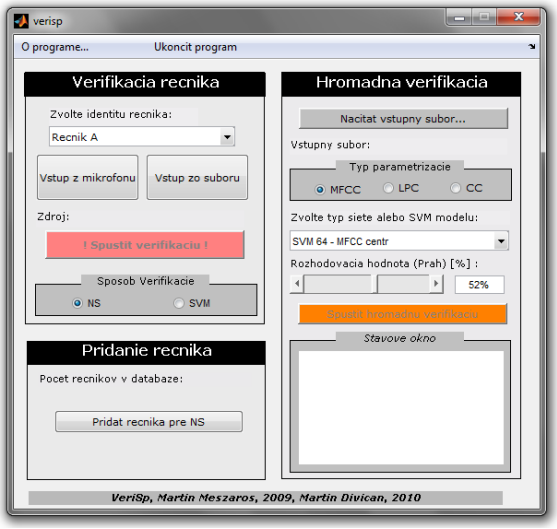


Fig. 2. Current VeriSp GUI

3.1.1 VeriSp lifecycle

Application loads existing trained network (neural or SVM) during startup. User using verification block can choose declared identity of speaker and way of input speech signal to application. Start of verification can be perform by button clicking after insert the record. This is real-time (interactive) verification. There are 4 output possibilities:

- successful verification
- high probability of successful verification
- verification failed – other speaker from database was identified
- verification failed

Second block permit user to add new speaker to database of speakers and train new extended network. Analogous to verification block, user can choose if new speaker record is inserted as a file or through microphone. If user choose microphone, speech is recorded and saved into wav file, because every new training of network need all speaker records to train.

Last part of GUI serve to general verification and test of system – combination of choose network, parametrization, criteria and threshold. Input of this part is description file which obtains records filename, declared identity and expected result of verification. State window show state of batch process and final result of batch verification is percentage of success result.

3.2 Disadvantages of current GUI

Current GUI has some disadvantages which complicate user work and limit him:

- user cannot interactively choose network type, name of network is hardcoded, to change it is necessary make changes in application source code
- user cannot change type of network during application run
- definition of new network is possible only as run Matlab function, no GUI support is available (only add speaker and retrain is possibly)
- information about loading network are not displayed to user
- there are several decision criteria implemented, but user cannot choose it from GUI, only by changing code comments
- count of parametrizations coefficient is not modifiable

These enumerated disadvantages are not decrease value of current application. There was not limiting in time of VeriSp creation, but nowadays we suppose user appreciative more option in configuration and easily graphic interface.

3.3 New concept of VeriSp GUI

New concept of GUI tries eliminating limit described above. It contains six blocks incorporated into 3 logical blocks. Figure 3 shows logical structure of blocks:

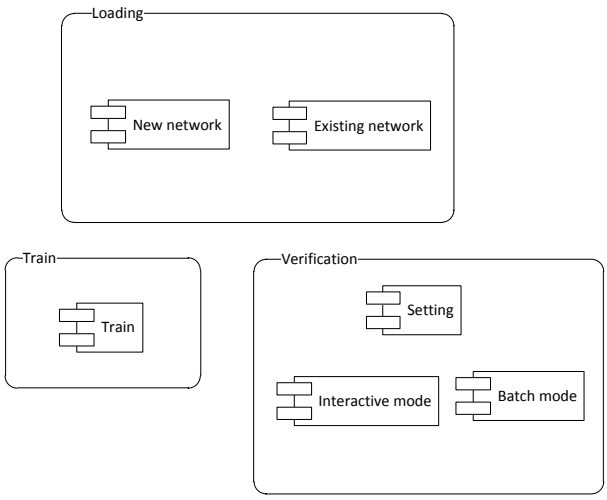


Fig. 3. Logical view of GUI concept design

First logical part "Loading" allows user create new network by choosing all necessary parameters. Next there is option create new network from template. Application configure network according to template, user can change some parameter and create network. Last option of loading network is load existing saved network. Parameter of using network are displayed in train block after loading or creating network (user cannot change them). Type of network is defined by its type (neural or SVM), internal structure and parametrization.

Network information panel and button allowing saving network extended second logical part "Train".

Last logical part "Verification" obtains three blocks: interactive verification, batch verification and setting, common block for both. Setting of classification allows user to define which classification method will be use and which value of threshold will be applied. This setting is common for both verification blocks. Nowadays there are implemented following classification methods:

- criteria 50% - winner have to have least probability 50%
- criteria 2x – winner have to have probability least 2 multiple than next in order
- criteria 1,5x – winner have to have probability least 1,5 multiple than next in order

Figure 4 shows graphic representation of new configurable concept

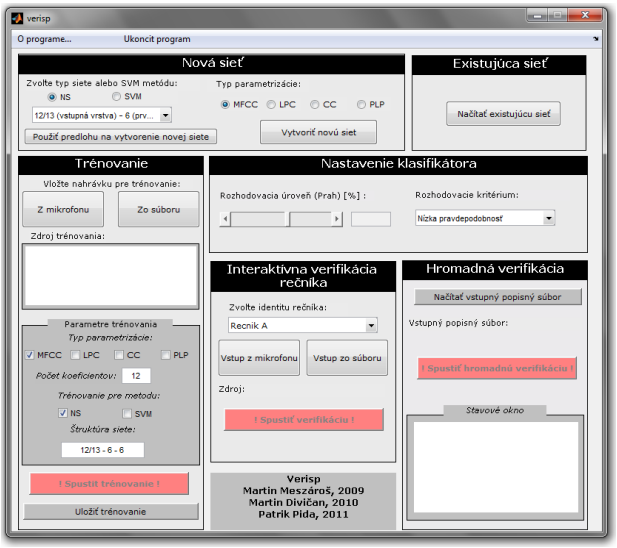


Fig. 3. Design of new GUI

Last design change is extension of possibilities of parametrization. New parametrization will be added – Perceptive Linear Prediction.

4. Conclusion

We analyzed current version of software and we identified some problematic issues in VeriSp. We regard as main problem the impossibility to save network and easily create a new network. The next problem is hidden configuration parameters.

We made a proposal of extended GUI, which should resolve listed disadvantages. In the future we plan to implement designed changes of GUI and a new parametrization.

Acknowledgements

Authors hereby declare that this work was created with support of projects VEGA 1/0718/09, FP7-ICT-2011-7 HBB-Next and FP7 Newton.

References

[1] Meszároš, Martin - Vojtko, Juraj:Speaker Verification Based on Speech Characteristics – Impact of Neural Network Configuration In: 3rd International Workshop on Speech and Signal processing Redžúr 2009, September 24, 2009, Bratislava, Slovak Republic, pp. 39-42, ISBN 978-80-227-3137-9

[2] Meszároš, Martin: Speaker verification based on voice charakteristics. Graduation theses. FEI STU Bratislava, 2009.

[3] Divičan, Martin: Speaker verification based on voice charakteristics using SVM, STU Bratislava, 2010.

[4] Psutka, Josef – Müller, Luděk – Matoušek, Jindřich – Radová, Vlasta: Mluvíme s počítačem česky. Praha: Academia, 2006, ISBN 80-200-1309-1

[5] Young, Steve. 2004. ATK. An Application Toolkit for HTK. Version 1.4.1. http://mi.eng.cam.ac.uk/~sjy/ATK_Manual.pdf

[6] Young, Steve. 2005. The HTK Book (for HTK Version 3.3). http://htk.eng.cam.ac.uk/docs/docs.shtml

[7] Sphinx-4: A speech recognizer written entirely in the JavaTM programming language. http://cmusphinx.sourceforge.net/sphinx4/

[8] Mathworks: MATLAB http://www.mathworks.com/products/matlab/

[9] "Voicebox"toolbox documentation: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[10] Juhar, Jozef - Cizmar, Anton - Rusko, Milan - Trnka, Marian - Rozinaj, Gregor – Jarina, Roman: Voice Operated Information System in Slovak, In: Computing and Informatics, Volume 26, 2007, Number 6, ISSN 1335-9150

[11] Kačur, Juraj; Rozinaj, Gregor: Building accurate and robust HMM models for practical ASR systems. Telecommunication Systems. Berlin : Springer Verlag, 2011. (in print) ISSN: 1018-4864.