EFFECT OF GSM SYSTEM ON TEXT-INDEPENDENT SPEAKER RECOGNITION PERFORMANCE

Nemat. S. Abdel Kader

Cairo University, Faculty of Engineering, Communication Dept.
E-mail nemat2000@hotmail.com

ABSTRACT

This paper introduces the influence of GSM encoder/decoder on text independent speaker recognition performance based on Vector Quantization (VQ) classifiers. A database consisting of 136 speakers is used to investigate the system performance. The speaker recognition is evaluated using the original database. The speech files are then passed through the GSM coder/decoder. The coded speech is finally used for training and testing the speaker recognition system. The effect of the additive white Gaussian noise (AWGN) and Rayleigh fading channels on the system performance is also investigated.

Keywords: Speech Coding, GSM System, Speaker Recognition, Vector Quantization.

1. INTRODUCTION

The GSM (Global System for Mobile Communication) system compresses the speech signal before its transmission to reduce the bits needed to represent the speech digitally while keeping an acceptable quality of the decoded speech. In recent years, due to the widespread use of wireless and mobile communication and computing terminal devices, there has been increasing interest in the performance of automatic speaker recognition (ASR) from coded speech. Mobile ASR capability can be applied both as a user interface to the terminal device as well as a data input/output modality between the user and remote applications. The constraints of the bit rate of the transmitted signal compound with potential exposure of the user to more intense and challenging acoustic environments make the problem of ASR in mobile environments more susceptible to performance degradation than fixed network speaker recognition applications.

Few papers address the effect of GSM speech coding on the speaker recognition performance [1-4]. The speaker models are performed using the
decoded speech or working directly without decoding. Three speech coders are standardized for use in the GSM wireless communication network. They are referred as the full rate (FR), half rate (HR) and enhanced full rate (EFR) [5].

This paper investigates the influence of the full rate GSM system on text independent speaker recognition performance taking into account the effect of channel nature and the background noise. The transcoded data was obtained by passing speakers speech files through the GSM coder/decoder. The general steps of ASR are: speech data acquisition, feature extraction, and pattern matching. Today, cepstral coefficients are the dominant features used for speaker recognition [6-7]. We use the VQ model to represent the statistical distribution of the features of each speaker.

The paper is organized as follows. The system model and GSM full rate is explained in section 2. The speaker recognition system is presented in Section 3. Speaker recognition experiments conducted on original and GSM FR coded speech are given in section 4. Finally conclusions and future work are drawn in section 5.

2. GSM SPEECH CODERS AND TRANSCODED DATABASE

Three speech coders are standardized for use in the GSM wireless communication network. They are referred as the full rate, half rate and enhanced full rate GSM coders. Their corresponding European telecommunications standards [5] are the GSM 06.10, GSM 06.20 and GSM 06.60, respectively. These coders work on a 13 bits uniform PCM speech input signal, sampled at 8 kHz. The input is processed on a frame-by-frame basis, with a frame size of 20 ms (160 samples).

2.1 Full Rate (FR) Speech Coder

The FR coder was standardized in 1987 [5]. This coder belongs to the class of Regular Pulse Excitation – Long Term Prediction - linear predictive (RPE-LTP) coders. In the encoder part, a frame of 160 speech samples is encoded as a block of 260 bits, leading to a bit rate of 13 kbps. The decoder maps the encoded blocks of 260 bits to output blocks of 160 reconstructed speech samples. The GSM full rate channel supports 22.8 kbps. Thus, the remaining 9.8 kbps are used for error protection. The FR coder is described in GSM 06.10 down to the bit level, enabling its verification by means of a set of digital test sequences which are also given in GSM 06.10. The whole path of data transfer through the GSM system, starting from speech production, GSM-FR source encoder, GSM transmitter data path, different communication channels, and GSM receiver data path is simulated. A conceptual block diagram of the GSM transmitter and receiver system is shown in figure 1.

2.2 Corpora

In this paper an Arabic based database is used to evaluate the proposed system. The database is collected from 136 speakers (69 men, 52 women, 10 boys, 5 girls) and every speaker is speaking five sentences for nearly 140 seconds, these sentences were recorded with 11kHz.

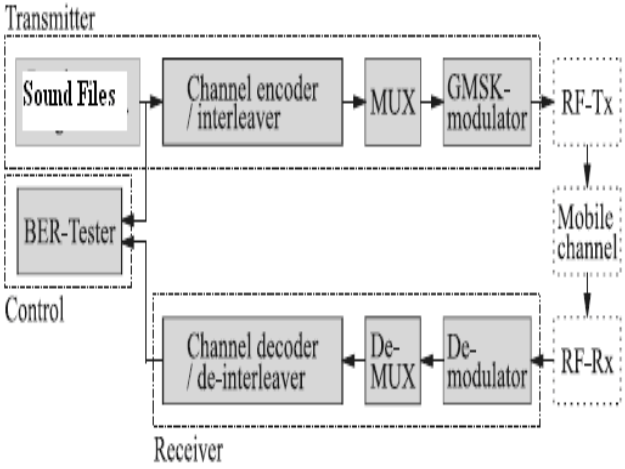

Figure 1: block diagram for a GSM transmitter/receiver system

In order for this database to be applicable in the GSM-FR encoder, all speech files are converted to a raw PCM file with 16-bits mono channel and enabling dithering with depth of 1 bit, also we down-sampled the speech database from 11kHz to 8Khz to be applicable for GSM-FR codec. The C-code implementations of the GSM coders assume 16-bit input format with 3 bits don't care. Thus, the first operation at the input of the coding program is a down-scaling by three bits (the three least significant bits are discharged). So in our work we compacted the information in the 16 bits in every sample into 13 bits only, and we don't care about the last 3 bits to comply with the GSM codec specifications.

2.3 GSM Transmitter and Receiver Simulation

The full-rate GSM speech codec [3] is a lossy speech coding- decoding algorithm based on a regular pulse excited long term prediction scheme. To develop software platform capable of generating a series of appropriate GSM data blocks, complex base-band representation is chosen to reduce the required simulation sample rate and also the overall simulation time and memory consumption. Moreover, the simulation is performed using a MATLAB toolbox.

The input speech frame is first pre-processed to produce an offset-free signal, which is then subjected to a first order pre-emphasis filter. The obtained samples are then analyzed to

determine the coefficients for the short term analysis filter (LPC analysis) to produce the short term residual signal. The filter parameters, termed reflection coefficients, are transformed to log area ratios, LARs, before transmission [6].

The speech frame is divided into 4 sub-frames with 40 samples of the short term residual signal in each. Each sub-frame is processed block-wise by the subsequent functional elements. The parameters of the long term analysis filter, (LTP lag and LTP gain), are estimated and updated, on the basis of the current sub-block of the present and a stored sequence of the previous reconstructed short term residual samples.

The resulting long term residual samples are fed to the Regular Pulse Excitation analysis (RPE) which performs the basic compression function of the algorithm. The RPE parameters are also fed to a local RPE decoding and reconstruction module which produces samples of the quantized version of the long term residual signal. By adding the quantized samples of the long term residual to the previous block of short term residual signal estimates, a reconstructed version of the current short term residual signal is obtained. The reconstructed short term residual signal samples are then fed to the long term analysis filter to produce the new short term residual signal estimates. During decoding, the residual signal is first reconstructed from the RPE-LTP information, and then filtered by the short-term synthesis filter, whose parameters are derived from the received LARs. Figure 2 shows a schematic representation of a general analysis by-synthesis coder.

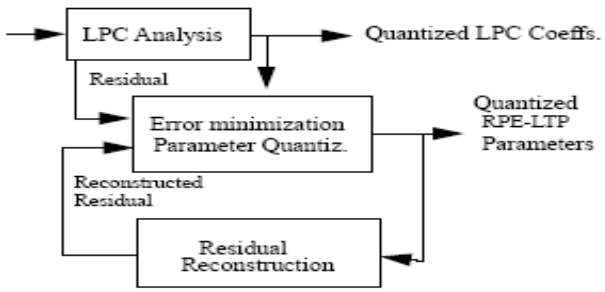

Figure 2: A simplified block diagram of a typical analysis by synthesis coder

3. SPEAKER RECOGNITION SYSTEM

We consider application of closed set text-independent speaker identification. In general, this means that we are looking at a system where mathematical models of voices of N speakers are created and stored in a speaker database. During the recognition a speech sample is compared to the models in the database. The result is the best identified speaker, or a list of several best matched speakers. A speaker recognition system consists of at least three functional modules (stages): data acquisition, feature extraction, and recognition decision, as in any biometric-based recognition system.

3.1 Feature Extraction

The typical set of features derived from the speech signal is the cepstral vector. Cepstral frame-based systems typically generate features at a rate of 100 frames per second, spanning a segment of speech with duration of approximately 25 msecs. This means that the features will have a certain overlap and thus cannot be completely conditionally independent. In addition to the cepstral features, typical feature vectors include first and second order time derivatives of the cepstral information. These cepstral time derivatives carry information related to the time varying properties of the signal, and have been found to enhance recognition considerably [7]. In addition, they help to make the observations more conditionally independent. A typical cepstral front-end also performs a human perception based scaling of the frequency axis based on perceptually derived warping functions (*e.g.*, mel scale, Bark scale). When cepstral features are derived from spectra warped according to a mel scale, we refer to them as mel scale cepstral coefficients, or MFCCs, which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

A 14-dimensional mel-cepstral vector (which consists of 0th order and 12 normal MFCC and log energy MFCC) is extracted from the speech signal every 10 ms using a 20 ms Hamming window and a pre-emphasis factor of 0.97. The

mel-cepstral vector is computed using a simulated triangular filterbank on the DFT spectrum.

3.2 Data modeling

Data modeling is a crucial stage in speaker recognition. A model for each person enrolled in the speaker recognition system, referred to as a client model, is required to describe the distinct distributional characteristics of data from this client. There are two types of data models, template models and stochastic models, both of which may be used for speaker recognition. In template modeling, a template is chosen in the recognition process based on the minimal distance between a given sequence of input samples and the template's frames. As it is based on this distance measure instead of probability, the template-based matching approach is deterministic.
The VQ model is used to represent the statistical distribution of the features of each speaker. The feature vectors extracted from the speech signal contain all available information and can be used to represent the speaker, but this is not practical where there are a large number of feature vectors. The extracted feature vectors are processed by a VQ algorithm for locating the clusters (centriods) in the feature space and then reducing the amount of data (feature vectors) we need to store for each speaker as a part of database [8-9].

LBG is an iterative algorithm which models speaker's data by a set of vectors, referred to as speaker's codebook. The codebook is obtained by the splitting method. An initial code vector is first set as the average of the entire training vectors. This code vector is then split into two. The iterative algorithm is run with these two vectors as the initial codebook. The final two code vectors are split into four and the process is repeated until the desired number of code vectors is obtained. We use Euclidean distance as the distortion criterion for training the codebooks.

The process of speaker identification is divided into two main phases. The first phase, speaker enrollment, speech samples are collected from the speakers, and they are used to train their models. In the identification phase, a test sample from an unknown speaker is compared against the speaker database. Based on these comparisons the final decision about speaker identity is made. This process is represented in figure 3.

4. EXPERIMENTAL RESULTS

In this section we describe the results of a series of speaker recognition experiments using cepstral features derived from the reconstructed waveforms. The database of the 136 speakers is used. Each speaker is speaking five sentences for nearly 140 seconds. For the training of the speaker models we use 6000 frames for training and 4000 frames for testing, the frame length is 20 ms and the frame rate is 10 ms, therefore the experiments can be considered as totally text-independent. By applying the Euclidian distance algorithm we can get the claimant speaker.

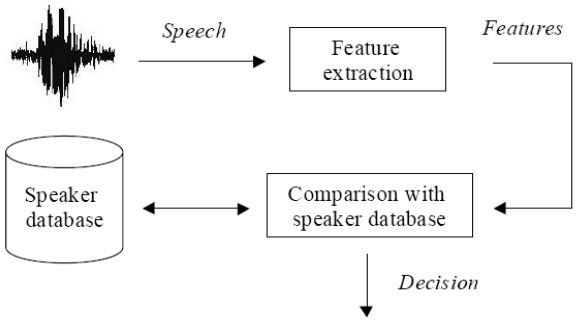

Figure3 Identification Phase

4.1 Recognition Accuracy using Original Speech Waveforms

The first set of experiments is applied to clean speech independent of GSM system with different testing frames table 1 and fig. 4). The performance of the clean speech system reaches 99% with 4 seconds testing time.

Table 1: Speaker identification results of clean speech

Number of frames (Time)	ID. Performance
100 (0.5 sec)	60.54%
200 (1 sec)	77.06%
400 (2 sec)	92.13%
800 (4 sec)	99.41%

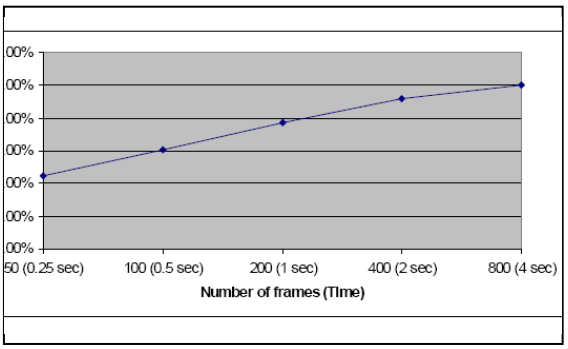

Figure 4 Speaker Identification using clean speech

4.2 Speaker Identification through GSM-FR codec under matched and mismatched conditions

Speaker recognition over wireless channels is problematic because of the additional noise and distortion introduced into voice signals during the coding, transmission (e.g., due to fading or packet loss), and decoding stages. Noise-degraded voice signals present in wireless environments are often substantially different from the original voice signal, leading to degradation in ASR performances when standard ASR techniques are applied.
An experiment is applied to evaluate the system performance where the encoded-decoded speech files are used in training and testing process (matched condition). The system is then trained using 6000 frames from the files before GSM-FR encoder and tested with the remaining 4000 frames applied to GSM-FR encoder-decoder (mismatched condition). Results are shown in table 2 and figure 5.

Number of frames (Time)	ID. Performance	
	(Matched Conditions)	(Mismatched Conditions)
100 (0.5 sec)	52.12%	41%
200 (1 sec)	69.19%	53%
400 (2 sec)	83.46%	63%
800 (4 sec)	91.62%	71%

Table 2: Speaker identification (ID) results through GSM-FR codec under matched and mismatched conditions

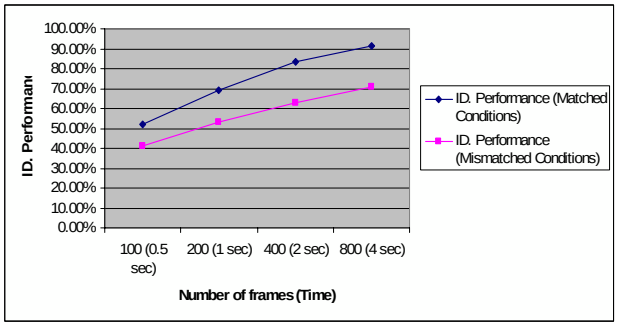

Fig.5 Speaker identification (ID) results through GSM-FR codec under matched and mismatched conditions

It was shown that the performance of automatic speaker recognition system is significantly degraded by acoustic mismatches between training and testing conditions. Such acoustic mismatches are commonly encountered in systems that operate on speech collected over GSM network. For the previous two cases, at 8 seconds test time, the identification rate reaches 95% under matched condition, while it decreases to 50% under mismatched conditions.

The GSM-FR codec causes degradation in speaker identification process by nearly 9% under matched condition and 53% under mismatched conditions after 4 seconds of testing time compared to the same period of clean speech. By comparing matched conditions with mismatched conditions, speaker identification process can be affected by enormous way which reaches 45% after 8 seconds.

Also we can see that under mismatched conditions, increasing the test time has no great effect on the performance of speaker identification process. We have improvement of 22% after extending the testing period by 7.5 seconds in mismatched conditions, while with matched conditions it reaches 43%, and in clean speech the improvement is 55% with 3.75 seconds test time.

4.3 Speaker Identification through GSM System with Additive White Gaussian Noise (AWGN) channel under matched and mismatched conditions

At this experiment the speaker identification process is applied through the whole GSM system including GSM-FR codec, GSM

transmitter and receiver. In addition to that we add the AWGN
communication channel. The AWGN Channel adds white Gaussian noise to a real or complex input signal. The AWGN channel is simulated with different SNR ratios (20dB, 10dB, 5dB). The results with different test times under matched and mismatched conditions are shown in table 3 and figures 6, 7.

4.4 Speaker Identification through GSM System with Rayleigh fading channel under matched and mismatched conditions

The transmission of a band-pass and narrowband signal to a mobile is modeled as a multipath channel structure. The fading simulation receives the input signal s (t), adds simulated fading effects

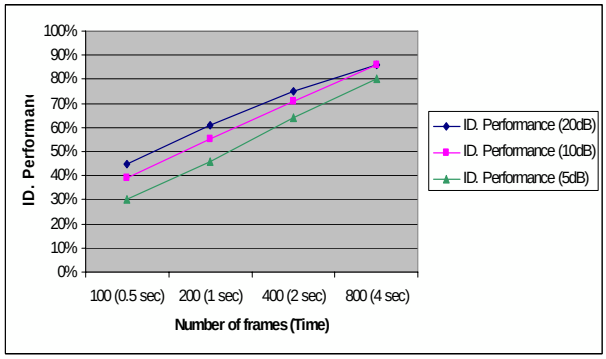

Figure 6 Speaker identification results through GSM System and AWGN channel with different SNR under matched conditions

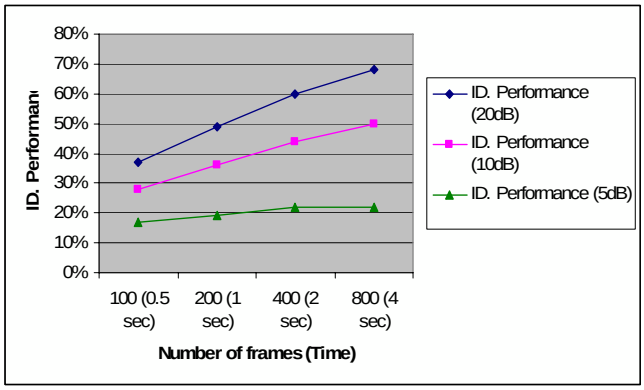

Figure 7 under mismatched conditions

NUMBER OF FRAMES(TIME)	ID PERFORMANCE					
	20 dB		10 dB		5 dB	
	m.	mis.	m.	mis.	m.	mis.
100(0.5 sec)	45%	37%	39%	28%	30%	17%
200(1 sec.)	61%	49%	55%	36%	46%	19%
400(2 sec.)	75%	60%	71%	44%	64%	22%
800(4 s3c.)	86%	68%	86%	50%	80%	22%

Table 3: Speaker identification results through GSM System and AWGN channel with different SNR under matched(m) and mismatched(mis) conditions

based on the fading parameters and outputs the signal y (t). The input signal s (t) is originated in a Transmitter, while the output signal y(t) eventually reaches a Receiver (figure 8).

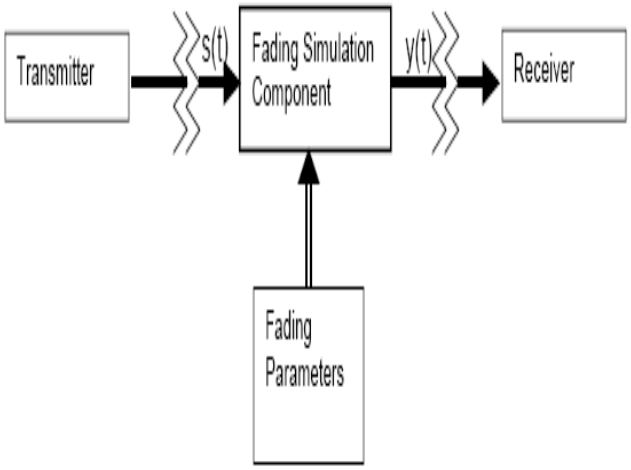

Figure 8: Black-box view of the Simulation Component

The fading simulation component consists of a Tapped Delay Line (TDL) and a number of Complex Gain Generators [10].

In our simulation we used autoregressive model of order 100, Doppler frequency equals 100 Hz at maximum and symbol frequency equals 271Kbps which equals the symbol frequency of transmitted bits.

The whole blocks of GSM transmitter and receiver path in addition to Rayleigh fading communication channel is applied to the system with additive noise equals 20 dB. Results with different test times under matched and mismatched conditions are shown in table 4 and figure 9.

A significant performance degradation is observed when using GSM system with fading channel. In matched conditions, the degradation in performance reaches 15% while it reaches 38% after 4 seconds of testing time under mismatched conditions compared to the same period of clean speech. From the above two experiments we can see that the performance in case of fading channel is almost the same as that of AWGN with SNR 20dB.

Table 4: Speaker identification results through GSM System with Rayleigh fading channel (with SNR 20dB) under matched and mismatched conditions

Number of frames (Time)	ID. Performance		
	(20dB)	(10dB)	(5dB)
100 (0.5 sec)	37%	28%	17%
200 (1 sec)	49%	36%	19%
400 (2 sec)	60%	44%	22%
800 (4 sec)	68%	50%	22%

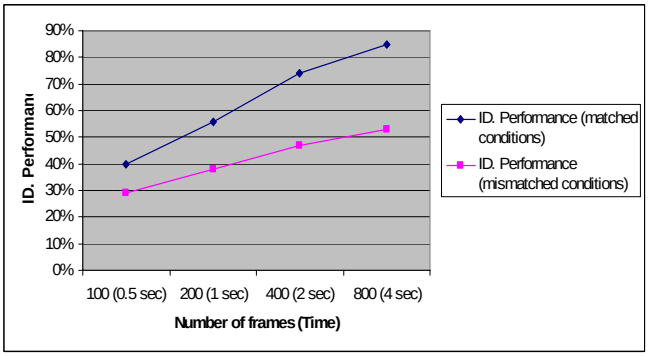

Figure 9 Speaker identification through GSM System with Rayleigh fading channel (with SNR 20dB) under matched and mismatched conditions

5. CONCLUSIONS AND FUTURE WORK

In this paper we have investigated the influence of GSM speech coding on a text-independent speaker recognition system based on VQ classifier. The effect of channel nature and the background noise produce degradation in the recognition performance. One reason for this is that the converted voice signal is distorted in the decoding process. Even without channel effects, the segmental-signal-to-quantization-noise-ratio (SSNR) of the decoded voice is often only approximately 20 dB in typical wireless environments. This latter signal condition arises because in seeking to meet low-bit-rate constraints, many features, e.g., excitation signals are not adequately coded and represented. In

addition, channel fading and interference cause many speech coder parameters to be unreliable for high-quality conversion back to voice signals. Also, the converted voice signal has a heavy dependency on the particular coding scheme used in the speech coder. Most importantly, the synthesized speech from the speech coder is usually very different from the human speech used to establish the speech recognition model; typically many characteristics of a speaker's voice are altered or lost in the synthesis process.

It was shown that the performance of the speaker recognition system is significantly degraded by acoustic mismatches between training and testing conditions. Using GSM-FR codec parameters, at 4 seconds test time, the identification rate reaches 92% under matched condition, while it decreases to 47.5% under mismatched conditions.

Adding the AWGN or Rayleigh fading channels to the encoded parameters, the degradation in performance reaches 15% and 38% after 4 seconds of testing time under matched and mismatched conditions compared to the same period of clean speech.

6. REFRENCES

[1]. M. Kuitert and L. Boves, "Speaker Verification with GSM Coded Telephone Speech", Proc. Eurospeech' 97, Vol. 2, pp. 975-978, 1997.

[2]. [2] T.F. Quatieri, E. Singer, R.B. Dunn, D.A. Reynolds, J.P. Campell, "Speaker and Language Recognition Using Speech Codec Parameters", Proc. Eurospeech'99, Vol. 2, pp. 787-790, 1999.

[3]. [3] S. Grassi, L. Besacier, A. Dufaux, M. Ansorge, F. Pellandini, "Influence of GSM Speech Coding on the Performance of Text-Independent Speaker Recognition", Proc. of EUSIPCO 2000, European Signal Processing Conference 2000, Tampere, Finland, September 4-8, 2000, pp. 437-440.

[4] L. Besacier, S. Grassi, A. Dufaux, M. Ansorge, F. Pellandini, "GSM Speech Coding and Speaker recognition", ICASSP 2000, International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, June 5-9, 2000, Vol.II, pp. 1085-1088.

[5] European Telecommunication Standards Institute, "European digital telecommunications system (Phase 2); Full rate speech processing functions (GSM 06.01)", ETSI 1994.

[4]. [6] Kroon, P., Deprettere, E. F., Sluyter, R. F. "Regular-Pulse Excitation - A Novel Approach to Effective and Efficient Multipulse Coding of Speech", *IEEE Transactions on Acoustics*, October 1986. *Speech and Signal Processing*, Vol. ASSP-34 No. 5:1054-1063,

[5]. [7] T. Kinnunen, "Spectral Features for automatic Text-Independent Speaker Recognition", Ph.Lic, Thesis, Univ. of Joensuu, Dept. of Computer Science, Feb. 2004.

[6]. [8] Linde, Y., Buzo, A., and Gray, R., M., "An algorithm for vector quantizer design", *IEEE Trans. Communication*, vol.28, Jan. 1980, pp. 84-94.

[7]. [9] Kinnunen, T., Kilpeläinen, T., and Fränti, P., "Comparison of clustering algorithms in speaker identification", *Proc. IASTED Int. Conf. Signal Processing and Communications*,Marbella, Spain, Sept. 2000, pp. 222-227.

[8]. [10]V.Erceg, K.V. S. Hari, M. S. Smith and D. S.Baum, " Channels Models for Fixed Wireless Applications", IEEE Task Group Contribution 802-16-3, Feb. 2001.