Published in Proceedings of the International COST 254 Workshop on
Intelligent Communication Technologies and Applications, with Emphasis on Mobile Communications, 307-311, 1999
which should be used for any reference to this work

INFLUENCE OF GSM SPEECH CODING ALGORITHMS ON THE PERFORMANCE OF TEXT-INDEPENDENT SPEAKER IDENTIFICATION

S. Grassi, L. Besacier, A. Dufaux, M. Ansorge, F. Pellandini

Institute of Microtechnology, University of Neuchâtel
Rue A.-L. Breguet 2, 2000 Neuchâtel, Switzerland
Phone: +41 32 7183432; Fax: +41 32 7183402
Email: sara.grassi@imt.unine.ch , laurent.besacier@imt.unine.ch

ABSTRACT

This paper investigates the influence, on the performance of a text-independent speaker identification system, of the three speech coding algorithms standardized for use in the GSM wireless communication network. The speaker identification system is based on Gaussian Mixture Models (GMM) classifiers. Only the influence of the speech coding algorithms was taken into account. This was done by passing the whole TIMIT database through each coding/decoding algorithm obtaining three transcoded databases. These databases were used for training and testing the speaker identification system. For comparison purposes, the speaker identification performance was also assessed using the original TIMIT and its 8 kHz downsampled version.

The results showed a significant performance degradation when using the GSM transcoded databases, compared to the normal and downsampled versions of TIMIT. These results are in correspondence with the subjective speech quality of each coder.

1. INTRODUCTION

Speaker recognition is the task of automatically recognizing the identity of a speaker using her / his voice. It has different applications such as banking over telephone network, telephone shopping, database access services and security control for confidential information. A big deal of these transactions will eventually take place through the telephone network. Moreover, as the demand for mobile communications is continuously increasing, it is expected that an increasing number of transactions using voice recognition will take place through the mobile cellular network.

GSM (Global System for Mobile Communications) is the pan-European cellular mobile standard. Three speech coding algorithms are part of this standard. The purpose of these speech coders is to compress the speech signal before its transmission, reducing the number of bits needed in its digital representation, while keeping an acceptable quality of the decoded output. We will use the term transcoding to indicate the process of coding and decoding the speech signal. As transcoding modifies the speech signal, it is likely to have an influence on voice recognition performance, together with other perturbations introduced by the mobile cellular network (channel errors, background noise). The effect of these perturbations have been studied for automatic speech recognition in [1] and [2].

This paper gives preliminary results about the influence of GSM speech coding on the performance of a speaker identification system. The three existing GSM speech coder standards were considered. These coders are briefly described in *Section 2*. *Section 3* shows the typical speech path when a user is accessing services that requires speaker identification. The whole TIMIT database was passed through these coders, obtaining three transcoded databases, as explained in *Section 4*. The speaker identification system used in these experiments is based on Gaussian Mixture Model (GMM) classifiers and is presented in *Section 5*. Speaker identification experiments conducted on normal and GSM-coded speech are given in *Section 6*. Finally, *Section 7* discusses the results obtained and draws some future work.

2. GSM SPEECH CODERS

There exist three different GSM speech coders, which are referred to as the full rate, half rate and enhanced full rate GSM coder. Their corresponding European telecommunications standards [3] are the GSM 06.10, GSM 06.20 and GSM 06.60. These coders work on a 13 bit 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). A brief description of these coders follows.

2.1 Full Rate (FR) Speech Coder

The FR coder was standardized in 1987. 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 [4], [3] 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.

A public domain bit exact C-code implementation of this coder is available [5], [6].

2.2 Half Rate (HR) Speech Coder

The HR coder standard was established to cope with the increasing number of subscribers. This coder is a 5.6 kbps VSELP (Vector Sum Excited Linear Prediction) coder from Motorola [7]. In order to double the capacity of the GSM cellular system, the half rate channel supports 11.4 kbps. Therefore, 5.8 kbps are used for error protection.

The measured output speech quality for the HR coder is comparable to the quality of the FR coder in all tested conditions [8], except for tandem and background noise conditions.

The normative GSM 06.06 [3] gives the bit-exact ANSI-C code for this algorithm, while GSM 06.07 gives a set of digital test sequences for compliance verification.

2.3 Enhanced Full Rate (EFR) Speech Coder

The EFR coder was the latest to be standardized. This coder is intended for utilization in the full rate channel, and it provides a substantial improvement in quality compared to the FR coder [9].

The EFR coder uses 12.2 kbps for speech coding and 10.6 kbps for error protection. The speech coding scheme is based on Algebraic Code Excited Linear Prediction (ACELP).

The bit exact ANSI-C code for the EFR coder is given in GSM 06.53 [3] and the verification test sequences are given in GSM 06.54.

2.4 DTX / VAD / CNG

Spectrum efficiency can be increased through the use of Discontinuous Transmission (DTX), switching the transmitter on only during speech activity periods. Voice Activity Detection (VAD) is used to decide upon presence of active speech. To reduce the annoying modulation of the background noise at the receiver (noise contrast effects), Comfort Noise Generation (CNG) is used, inserting a coarse reconstruction of the background noise at the receiver.

The three GSM coders described above include the functions of DTX, VAD and CNG. Their corresponding normative references are [3]: GSM 06.31, GSM 06.32 and GSM 06.12 for the FR coder, GSM 06.41, GSM 06.42 and GSM 06.22 for the HR coder, and GSM 06.81, GSM 06.82 and GSM 06.62 for the EFR coder.

The use of DTX is associated with potential degradation of the speech quality due to speech clipping (speech detected as noise) and noise contrast effects. It is thus expected that the use of DTX has a negative impact on the performance of speaker identification systems.

3. SPEECH PATH

Figure 1 shows the typical speech path when a user is accessing services that requires speaker identification using his / her mobile phone. The speech path goes from the audio input in the Mobile Station (MS) to the digital interface of the Public Switched Telephone Network (PSTN). The speaker recognition task occurs after the PSTN (at the centralized site, e.g. bank service).

The audio part of the Mobile Station [10] includes the microphone and analog to digital conversion (ADC). This audio part gives a 13-bit uniform Pulse Code Modulated (PCM) signal to the encoder.

In the work reported in this paper, only the effects introduced by the shadowed blocks in Figure 1 (Encoder / Decoder and DTX) are studied.

4. GSM TRANSCODED DATABASES

4.1 TIMIT database

The TIMIT database [11] contains speech from 630 speakers (438 male and 192 female), each of them speaking 10 phonetically-rich sentences. The speech signal is recorded through a high quality microphone, in a very quiet environment, with a 0-8 kHz bandwidth. All recordings took place in a single session (contemporaneous speech).

The text material in the TIMIT prompts, consists of 2342 sentences, divided into 2 dialect sentences (SA sentences), 450 phonetically compact sentences (SX sentences) and 1890 phonetically diverse sentences (SI sentences). The SA sentences are meant to expose dialectal variants of the speakers and were read by all

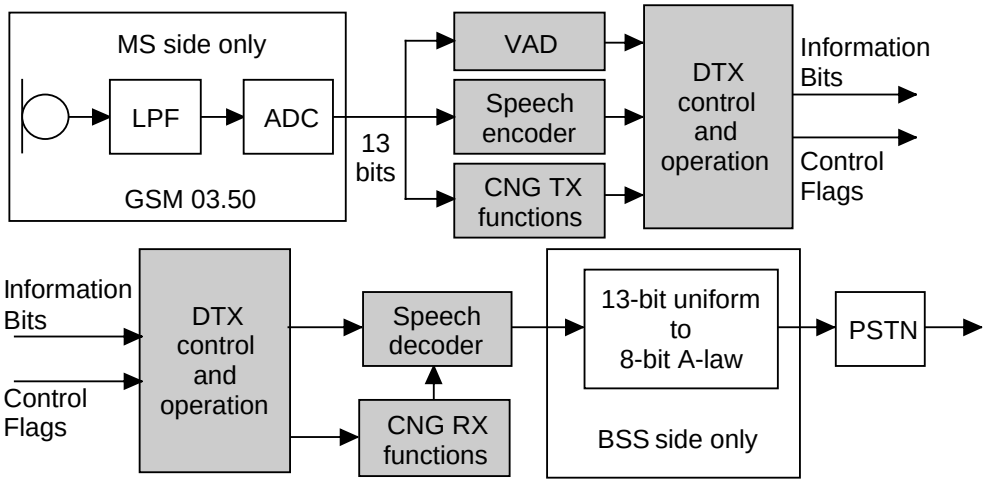

Figure 1: Speech path when accessing services requiring speaker verification through the mobile phone (MS = Mobile Station, BSS = Base Station System, PSTN = Public Switched Telephone Network, VAD = Voice Activity Detection, CNG = Comfort Noise Generation, DTX = Discontinuous Transmission, TX = transmitter, RX = Receiver, LPF = Low Pass Filter, ADC = Analog to Digital Converter).

630 speakers. The SX sentences are comprehensive and compact. Each speaker read 5 of these sentences and each text was spoken by 7 different speakers. The SI sentences are intended to add diversity in sentence types and phonetic contexts. Each speaker read 3 of these sentences, with each sentence being read by only a single speaker.

4.2 GSM transcoding

The whole TIMIT database was downsampled from 16 kHz to 8 kHz, using a 158th-order linear-phase FIR half-band filter, with a very steep transition band (150 Hz of transition band), a very flat passband (passband ripple < 0.1 dB), and more than 97 dB of attenuation in the stop band. Thus, the downsampled speech files contain basically all the frequencies of the original TIMIT in the 0-4 kHz range. Hereafter, the downsampled database will be referred to as TIMIT 8k, while the original will be referred to as TIMIT 16k. We are aware that the actual anti-aliasing low pass filter of a mobile phone (see Section 3) may not have such ideal characteristics. However, to the extend of our knowledge, this filter is not specified in the GSM standards [10].

TIMIT 8k was transcoded using the three GSM speech coders. The public domain C-code implementation of the FR coder was used (see Section 2.1), as well as the ANSI-C code for the HR and the EFR provided by ETSI (see Section 2.2 and 2.3). These C-code implementations were compiled and verified using the test vectors provided by ETSI [3], before their utilization.

To investigate the use of DTX, two more transcoded databases were built, using the HR and EFR programs, with the DTX option activated. This option is not available in the existing FR program.

4.3 Note on the Scaling of the Input Speech

In building the transcoded databases, no scaling was applied to the TIMIT 8k before transcoding.

The C-code implementations of the GSM coders assume the following input format (16-bit fixed point 2's complement) after the ADC (see Figure 1):

S.v.v.v.v.v.v.v.v.v.v.v.v.x.x.x

where S is the sign bit, v a valid bit, and x a "don't care" bit. Thus, the first operation at the input of the three coding programs is a down-scaling by three bits (the three least significant bits are discharged). If the input speech file range is well adjusted to a 16-bit range, there will not be a great loss in precision. On the other hand, if the input speech file has a range corresponding, e.g., to 13 bits, the loss in precision is greater. The maximum amplitude of the TIMIT 8k speech files was measured, and it was found that 45% of the files have a range corresponding to 13 bits or less.

The loss in precision at the input could decrease the performance of the coding, affecting also the identification performance. A new set of transcoded databases in which the input is scaled to its maximum range, is being built to investigate this effect.

5. SPEAKER IDENTIFICATION SYSTEM

5.1 Acoustic Analysis

The speech analysis module extracts 16 cepstral coefficients. The frame length is 30 ms and the frame rate is 10 ms.

5.2 GMM-based Classifier

Gaussian Mixture Models (GMMs) can be used to represent speaker's voices. GMM classifiers were introduced for the task of speaker identification / verification by Reynolds [12]. The GMMs can be viewed as hybrids between unimodal gaussian classifiers and vector quantizer codebooks. They are able to take into account very small clusters of sounds and have shown to give good results for the speaker identification task.

More precisely, the distribution of feature vectors extracted from the signals of a particular speaker is modeled by a gaussian mixture density. For a feature vector, denoted as x_t, the mixture density is defined as

$$p(x_t \mid \lambda_s) = \sum_{i=1}^{N} p_i^s b_i^s(x_t) . \qquad (1)$$

The density is a weighted linear combination of *N* components unimodal gaussian densities $b_i^s(x_t)$, each parameterized by a mean vector μ_i^s and a covariance matrix Σ_i^s; p_i^s are the mixture weights. The parameters of a speaker model *s* are then denoted as $\lambda_s = (p_i^s, \mu_i^s, \Sigma_i^s)_{1 \le i \le N}$. Given a sequence $(x_t)_{1 \le t \le T}$ of feature vectors from a speaker signal, maximum likelihood estimates of the model parameters are obtained using the Expectation-Maximization (EM) algorithm. Given an unknown sequence of signal $(y_t)_{1 \le t \le T'}$, the recognized speaker $\hat{s}$ is then obtained with the maximum likelihood (ML) decision rule :

$$\hat{s} = \arg\max_{s} \frac{1}{T'} \sum_{t=1}^{T'} \log p(y_t \mid \lambda_s) . \qquad (2)$$

6. EXPERIMENTS AND RESULTS

6.1 Protocols

A well-known protocol is used on TIMIT (see Section 4.1) for speaker identification. It is called the *"long training / short test protocol"* [13]. For the training of the speaker models, we use all 5 SX sentences concatenated as a single reference pattern for each speaker. The average total duration is 14.4 seconds. For the test of the speaker identification system, each of the SA and SI sentences is tested separately. The whole test set thus consists of 630x5=3150 test patterns of 3.2 seconds each, in average. Even though the SA sentences are the same for each speaker, these sentences are used in the test set. Therefore, the experiments can be considered as totally text independent.

6.2 Experiments and Results

A GMM classifier of N=16 mixtures was tested. We choose 16 gaussians which is a good trade-off between complexity and performance. Diagonal covariance matrices were used for gaussian densities, since there are no strong correlations between cepstral coefficients. These experiments were conducted using *h2m*, a set of *Matlab* functions designed by O. Cappe [14]. Table 1 shows the identification results obtained with this speaker identification system on TIMIT 16k, TIMIT 8k, and the GSM transcoded TIMIT (FR, HR and EFR). For the HR and EFR coders, the effect of DTX / NO DTX was also investigated. The use of only 10 cepstral coefficients was also investigated, but the results are not reported since the performance was always lower than with 16 coefficients.

For each of the five experiments, training and testing are both made on the same database.

TIMIT 16 kHz	TIMIT 8 kHz	DTX	TIMIT GSM/FR	TIMIT GSM/HR	TIMIT GSM/EFR
97.4%	84.9%	no	65.0%	57.2%	66.7%
		yes		55.8%	62.3%

Table 1: Speaker identification results (%) for normal and GSM transcoded speech – GMM of 16 mixtures – 16 Cepstral Coeff. – Long training / Short test protocol – 630 speakers - 3150 tests.

7. DISCUSSION

We have investigated the influence of the three GSM speech coders on a text-independent speaker identification system, based on GMM classifiers. Only the effects introduced by the speech coding were taken into account.

The results showed a significant performance degradation when using GSM transcoded databases, compared to the normal and downsampled versions of TIMIT. The performance achieved using GSM transcoded speech is not sufficient in a practical context. Improvement in the results could be achieved by finding which block of each encoder introduces the most important performance degradations, and then try possible corrections, such as finding new features which are better preserved by the coder.

The results obtained are in correspondence with the subjective speech quality of each coder. That is, the higher the speech quality is, the higher the measured identification performance.

It was observed that the DTX has a negative impact on the performance, due to speech clipping (speech detected as noise). Nevertheless the degradation was very small, probably due to the short duration of the silence periods in the TIMIT database.

The effect of the scaling of the input signal before transcoding is also being investigated. A small increase in the performance is expected when all the input signals are scaled up to use all the dynamic range available at the input of the GSM coders.

Due to the lack of realistic corpora of GSM recordings [15], the investigation was limited to the influence of the coding algorithms. While waiting for real GSM corpora, a possible direction of future work would be the study of the effects of simulated channel errors on the performance of the speaker identification system.

8. ACKNOWLEDGEMENTS

This work was supported by the Swiss Federal Office for Education and Science under Grant OFES C97.0050 (COST 254 project).

9. REFERENCES

[1] S. Dufour, C. Glorion, P. Lockwood, “Evaluation of root-normalised front-end (RN LFCC) for speech recognition in wireless GSM network environments”, *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, ICASSP'96, Atlanta, USA, Vol. 1, pp. 77-80, May 1996.

[2] L. Karray, A. B. Jelloun, C. Mokbel, “Solutions for robust recognition over the GSM cellular network”, *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, ICASSP'98, Seattle, USA, Vol. 1, pp. 261-264, May 1998.

[3] http://www.etsi.org/.

[4] ETS 300 961 : Digital cellular telecommunications system (Phase 2+); Full rate speech; Transcoding (GSM 06.10 version 5.1.1), second edition, May 1998.

[5] J. Degener, "Digital Speech Compression : Putting the GSM 06.10 RPE-LTP algorithm to Work", Dr. Dobb's Journal, Dec. 1994.

[6] http://kbs.cs.tu-berlin.de/~jutta/toast.html

[7] I. Gerson and M. Jasiuk, “A 5600 bps VSELP speech coder candidate for half rate GSM”, *Proc. European Conference on Speech Communication and Technology*, EUROSPEECH' 93, Berlin, Germany, Vol. 1, pp. 253-256, Sep. 1993.

[8] TR 101 641 : Digital cellular telecommunications system (Phase 2+); Half rate speech; Performance characterization of the GSM half rate speech codec (GSM 06.08 version 6.0.0 Release 1997).

[9] K. Järvinen et al. “GSM Enhanced Full Rate Codec”, *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, ICASSP'97, Munich, Germany, Vol. 2, pp. 771-774, April 1997.

[10] EN 300 903: Digital cellular telecommunications system (Phase 2+); Transmission planning aspects of the speech service in the GSM Public Land Mobile Network (PLMN) system (GSM 03.50 version 6.1.0), 1997.

[11] W. Fisher, V. Zue, J. Bernstein, D. Pallet, "An acoustic-phonetic database", JASA, suppl. A, Vol. 81(S92), 1986.

[12] D. Reynolds, “Speaker Identification and Verification using Gaussian Mixture Speaker Models”, in Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, April 5-7, 1994, pp. 27-30.

[13] F. Bimbot, I. Magrin-Chagnolleau, L. Mathan, "Second-order Statistical Methods for Text-Independent Speaker Identification", Speech Communication, n° 17 (1-2), Aug. 1995, pp. 177-192.

[14] O. Cappe, “h2m : A set of MATLAB functions for the EM estimation of hidden Markov models with Gaussian state-conditional distributions”. ENST/Paris http://sig.enst.fr/~cappe/h2m/html/.

[15] M. Kuitert and L. Boves, "Speaker verification with GSM coded telephone speech", *Proc. European Conference on Speech Communication and Technology*, EUROSPEECH'97, Rhodes, Greece, Vol.2, pp. 975-978, Sep. 1997.