Published in Proceedings of the International Conference on
Acoustics, Speech, and Signal Processing (ICASSP2000) 2, 1085-1088, 2000
which should be used for any reference to this work

GSM SPEECH CODING AND SPEAKER RECOGNITION

L. Besacier[1,2] *, S. Grassi*[1] *, A. Dufaux*[1] *, M. Ansorge*[1] *, F. Pellandini*[1]

(1) Institute of Microtechnology, University of Neuchâtel, A.L. Breguet, 2 – 2000 Neuchâtel (Switzerland)
(2) now with CLIPS/IMAG, GEOD team, University Joseph Fourier, BP 53 – 38041 Grenoble (France)
laurent.besacier@imag.fr , sara.grassi@imt.unine.ch

ABSTRACT

This paper investigates the influence of GSM speech coding on text independent speaker recognition performance. The three existing GSM speech coder standards were considered. The whole TIMIT database was passed through these coders, obtaining three transcoded databases. In a first experiment, it was found that the use of GSM coding degrades significantly the identification and verification performance (performance in correspondence with the perceptual speech quality of each coder). In a second experiment, the features for the speaker recognition system were calculated directly from the information available in the encoded bit stream. It was found that a low LPC order in GSM coding is responsible for most performance degradations. By extracting the features directly from the encoded bit-stream, we also managed to obtain a speaker recognition system equivalent in performance to the original one which decodes and reanalyzes speech before performing recognition.

1. INTRODUCTION

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 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. As GSM transcoding (the process of coding and decoding) modifies the speech signal, it is likely to have an influence on speaker recognition performance, together with other perturbations introduced by the mobile cellular network (channel errors, background noise). Furthermore, as the demand for mobile communications is continuously increasing, it is expected that an increasing number of transactions using speaker recognition will take place through the mobile cellular network. Thus, this paper proposes an in-depth look at the influence of GSM speech coding on text independent speaker recognition performance. To our knowledge, few contributions [1], [2], [3] were made on this subject, whereas the effect of perturbations in a mobile framework has been more extensively studied for automatic speech recognition, where we can cite among others [4] and [5].

The three existing GSM speech coders are briefly described in *Section 2*. The whole TIMIT database was passed through these coders, obtaining three transcoded databases, as explained in *Section 3*. Two different experiments were carried out. In the first experiment (see *Section 4*) the speaker identification and verification performance degradation due to the utilization of the three GSM speech coders was assessed. In the second experiment (see *Section 5*) the features for the speaker recognition system were calculated from the information available in the encoded bit stream (only for the GSM FR coder). This experiment allows a measurement of the degradation introduced by different aspects of the coder, and gives some guidelines for a better use of the information available in the bit stream, for speaker recognition purposes. Finally, in *Section 6* the results obtained are discussed and possible future work is described.

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 coders. Their corresponding European telecommunications standards [6] are the GSM 06.10, GSM 06.20 and GSM 06.60, respectively. 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 [6] 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 [7].

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 [8]. 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 [9], except for tandem and background noise conditions. The normative GSM 06.06 [6] gives the bit-exact ANSI-C code for this algorithm, while GSM 06.07 gives a set of digital test sequences for compliance verification.

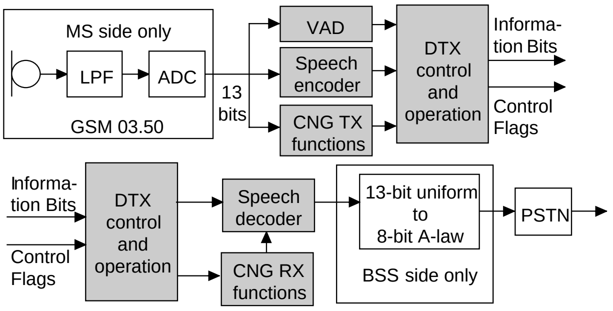

Figure 1. Speech path when accessing services requiring speaker verification through the mobile phone (BSS = Base Station System, VAD = Voice Activity Detection, CNG = Comfort Noise Generation, DTX = Discontinuous Transmission, TX = transmitter, RX = Receiver).

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 [10].

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 [6] 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 [6]: 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 recognition systems.

2.5 Speech Path

Figure 1 shows the typical speech path when a user is accessing services that require speaker recognition 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 (e.g. at the centralized bank service). The audio part of the Mobile Station [11] 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.

3. GSM TRANSCODED DATABASES

3.1 TIMIT database

The TIMIT database [12] is used during the various experiments. Even if this database is mono session, it offers the advantage of being largely used in the literature for comparison, being suited to text independent tasks, and proposing a large number of speakers (438 male and 192 female speakers).

3.2 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 TIMIT8k, while the original will be referred to as TIMIT16k. We are aware of the fact that the actual anti-aliasing low pass filter of a mobile phone may not have such ideal characteristics. However, to the extent of our knowledge, this filter is not specified in the GSM standards [11].

TIMIT8k 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 [6], before their utilization.

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

3.3 Note on the Scaling of the Input Speech

In building the transcoded databases, no scaling was applied to the TIMIT8k 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 TIMIT8k 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, and affect the recognition performance. As part of future work we would build a new set of transcoded databases with the input scaled to its maximum range, to investigate this effect.

Original		*GSM Transcoded*			
TIMIT16k	*TIMIT8k*	*DTX*	*FR*	*HR*	*EFR*
2.2%	13.1%	no	31.5%	38.5%	28.2%
		yes	-	39.8%	34.6%

Table 1. Speaker identification results (% errors identification) for original and GSM transcoded speech – 430 speakers - 2150 tests.

Original		*GSM Transcoded*			
TIMIT16k	*TIMIT8k*	*DTX*	*FR*	*HR*	*EFR*
1.1%	5.1%	no	7.3%	7.8%	6.6%
		yes	-	8.7%	6.2%

Table 2. Speaker verification results (%EER) for original and GSM transcoded speech – 430 speakers - 2150 client accesses and 2150 impostor accesses.

4. FIRST EXPERIMENT

4.1 Protocols

A well-known protocol is used on TIMIT for speaker identification and verification. It is called the *"long training / short test protocol"* [13]. For the training of the speaker models, we use all the 5 SX sentences concatenated as a single reference pattern for each speaker. The average total duration is 14.4 seconds. For the testing of the speaker identification system, each of the SA and SI sentences is tested separately.

430 speakers (147 women and 283 men) of the database are used and the whole test set thus consists of 430x5=2150 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.

The remaining 200 speakers of the database are used to train the background model needed for the speaker verification experiments. 2150 client accesses and 2150 impostor accesses are made (for each client access, an impostor speaker is randomly chosen among the 429 remaining speakers).

All the experiments were carried out under matching conditions (i.e. training and testing are both made using the same database).

4.2 Speaker recognition system

The speech analysis module extracts 16 cepstral coefficients. The frame length is 30 ms and the frame rate is 10 ms. A GMM classifier [14] of N=16 mixtures was tested. 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 [15]. During recognition, the *verification* score for an utterance is the log-likelihood ratio computed by taking the difference between the log-likelihoods of the claimant model and the background model; whereas the *identification* score is the log-likelihood of the speaker models.

4.3 Results

Table 1 and Table 2 show the identification and verification results respectively obtained on TIMIT16k, TIMIT8k, 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 studied, but the results are not reported since the performance was always lower than with 16 coefficients.

4.4 Comments

The results show a significant performance degradation when using GSM transcoded databases, compared to the normal and downsampled versions of TIMIT even if training and testing were both performed with transcoded speech. The results obtained are in correspondence with the perceptual speech quality of each coder. That is, the higher the speech quality is, the higher the measured recognition 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. We see that the degradation of the performance is less important for speaker verification than for speaker identification, but is still significant. These results are equivalent to those obtained in [3], whereas [1] and [2] suggest that the GSM coding does not introduce major degradations. From our point of view, the performance achieved using GSM transcoded speech is not sufficient in a practical context. Thus, the following section is devoted to studying the source of the degradation observed with GSM transcoded speech (only the FR coder is studied). The possibility of performing recognition using directly codec parameters rather than parameters extracted from the resynthesized speech is also investigated.

5. SECOND EXPERIMENT

The purpose of this experience is twofold, namely to find out which portions of the encoder are responsible for major degradations and to improve the performance with respect to the results obtained by extracting the features from resynthesized speech.

Line (1) in Table 3, corresponding to the baseline, lists the values reported from the TIMIT FR experiment in Table 1 and 2.

Training and testing were made for matching conditions. All the experiences (lines (2) to (8)) were carried out using TIMIT8k, but the feature extraction was made compatible with the FR coder characteristics: 20 ms segmentation, calculation of 8-th order LPC (LPC8), calculation of cepstral coefficients c1-c15 from the LPC using the well known recursion for minimum phase signals, and calculation of c0 using log (E), where E is the energy of the LPC residual. The results obtained with this feature extraction are given in line (2) of Table 3. For lines (2) to (4) the feature extraction is done with a C-program, using double-precision floating point arithmetic:

(3) Uses only cepstral coefficients c1-c15 (no energy term c0),
(4) Uses an LPC model order of 12 instead of 8.

Feature extraction for lines (5) to (8), is done from the FR C-program, which uses a simulated 16-bit fixed-point arithmetic:

(5) Uses c1-c15, from LPC before quantization (LPC coding-decoding),
(6) Uses c1-c16, from LPC before quantization,
(7) Uses c1-c15, from LPC after quantization.

Coefficients	*id. error*	*EER*
(1) Baseline: resynthesized speech FR	**31.5%**	**7.3%**
(2) LPC8 → c0-c15	31.8%	7.0%
(3) LPC8 → c1-c15	38.0%	7.8%
(4) LPC12 → c0-c15	24.0%	5.5%
(5) FR (no q) → c1-c15	43.7%	7.5%
(6) FR (no q) → c1-c16	43.6%	7.5%
(7) FR (with q) → c1-c15	40.8%	8.4%
(8) Codec param. FR (with q) → c0-c15	**35.7%**	**7.0%**

Table 3: Speaker identification and verification results for the second experiment.

(8) Uses c1-c15, from LPC after quantization, and c0, which is calculated using $\log(\hat{E})$, where $\hat{E}$ is the energy of the reconstructed LPC residual.

Comments from pair-wise comparison on Table 3:

(1)-(2): The use of the new feature extraction (more compatible with the FR characteristics), does not introduce significant distortion.

(2)-(3): The use of c0 (more laborious to calculate from the bit-stream) is crucial for good performance.

(2)-(4): A low LPC order in GSM FR coding (LPC8) is responsible for most performance degradations. Better results are likely to be obtained in experiences using the EFR, which has a 10-th order LPC. Working on the decoded speech allows possible recover of higher order LPC information that has “leaked” in other encoded parameters (LTP lags and gains, and RPE pulses [6]). A possible direction of future work is to obtain this higher order information from the decoded speech.

(5)-(6): No performance improvement is expected by retaining cepstral coefficients beyond c15 without increasing the LPC order.

(5)-(7): LPC quantization in the FR coder decreases the performance in the verification and improves in the identification. Not conclusive.

(7)-(8): The c0 calculated from the reconstructed residual improves the performance.

(1)-(8): By extracting the features directly from the information in the encoded bit-stream, we have managed to obtain a speaker recognition system that is quasi equivalent to the baseline.

6. DISCUSSION AND FUTURE WORK

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

Two experiments were done. In the first experiment, it was found that usage of GSM coding degrades significantly the identification and verification performance. The second experiment provides a measurement of the different performance degradation sources within the FR coder. Moreover, it enlightens the perspective to directly exploit the coder output parameters instead of decode and reanalyze speech.

Future work would consist in performing the second experiment using the EFR coder. The possibility of obtaining a real GSM database from a national telephone operator is under negotiation.

7. ACKNOWLEDGEMENTS

This work was partially supported by the Swiss National Science Foundation under Grant FN 20-53’843, and by the Swiss Federal Office for Education and Science under Grant OFES C97.0050 (COST 254 project).

8. REFERENCES

[1] M. Kuitert and L. Boves, "Speaker verification with GSM coded telephone speech", Proc. Eurospeech’97, Vol.2, pp. 975-978, 1997.

[2] M. El-Maliki, P. Renevey and A. Drygajlo, "Speaker verification for noisy GSM quality speech", COST 254 Workshop, Neuchâtel, Switzerland, (in print), May 5-7, 1999.

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

[4] S. Dufour, C. Glorion, P. Lockwood, “Evaluation of root-normalised front-end (RN LFCC) for speech recognition in wireless GSM network environments”, Proc. ICASSP’96, Vol. 1, pp. 77-80, 1996.

[5] L. Karray, A. B. Jelloun, C. Mokbel, “Solutions for robust recognition over the GSM cellular network”, Proc. ICASSP'98, Vol. 1, pp. 261-264, 1998.

[6] http://www.etsi.fr

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

[8] I. Gerson and M. Jasiuk, “A 5600 bps VSELP speech coder candidate for half rate GSM”, Proc. Eurospeech'93, Vol. 1, pp. 253-256, 1993.

[9] 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).

[10] K. Järvinen et al. “GSM Enhanced Full Rate Codec”, Proc. ICASSP'97, Vol. 2, pp. 771-774, 1997.

[11] 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.

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

[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] 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.

[15] 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/.