SPEAKER VERIFICATION WITH GSM CODED TELEPHONE SPEECH

M. Kuitert *L. Boves*

Department of Language and Speech, University of Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, the Netherlands
e-mail: kuitert,boves@let.kun.nl

ABSTRACT

In this paper we investigate the impact on the performance of Speaker Verification (SV) systems of the signal and channel coding in GSM cellular telephone networks. In this study only the effects of the codec are investigated. This is done by transcoding the signals in an existing speech corpus, recorded in the fixed network, to GSM. We compared text dependent SV performance of systems trained with A-law speech and tested with A-law and GSM speech, as well as systems trained with GSM speech and tested with GSM speech. All SV systems compared were based on continuous density Gaussian mixtures HMM models, differing in acoustic resolution.
We have compared several parameter representations derived from FFT and LPC based spectral estimates. It is shown that (and why) LPC based estimates are to be preferred. It is also shown that it pays to extend the analysis bandwidth to the full 4 kHz offered by the digital telephone network.
The major conclusion of our research is that the impact of GSM coding on the parameter representations is marginal and can effectively be ignored.

1. INTRODUCTION

Now that cellular GSM networks are growing very rapidly, there is an urgent need to acquire knowledge on the impact of the coding and bit rate reduction in these networks on the performance of automatic speech recognition and speaker verification systems. In this paper we focus on the impact of GSM coding on Speaker Verification (SV) performance.
Designing and performing SV experiments are major tasks of their own. For the research presented here we could re-use the design and the tools developed in the CAVE project (Caller Verification in Banking and Telecommunication, LE1-1930) [1]. CAVE has provided us with a realistic corpus (consisting of connected digits recorded in the landline telephone network), with proven SV technology and with efficient experimentation protocols. All experimental procedures developed in CAVE are based on the Hidden Markov Toolkit (HTK) [4]. Thanks to this framework, we were able to run large numbers of experiments in relatively short time.
Due to the lack of realistic corpora of GSM recordings, the investigation of the impact of GSM on the performance of SV systems is limited to the effects of the coding algorithm proper. We have taken the A-law coded corpus used in CAVE and transcoded the speech into GSM. Although this is not representative of what happens in real applications, it is the best one can do until real GSM corpora are available. We have investigated the effects of the coder by comparing the performance of the CAVE reference system, based on the original A-law data, with the performance of the system with GSM transcoded data. We have studied SV systems trained with GSM data and tested with GSM data (matched condition) as well as SV systems where the enrolment was done with A-law signals and the testing with GSM coded signals.
In CAVE it has been shown that left-to-right HMM models with continuous mixture densities (separate models for each digit with a number of states depending on the number of phonemes in the canonical phonemic representation of the speech) give very good performance. CAVE has also shown that in most cases Linear Predictive Cepstral Coefficients (LPCC's) outperform FFT based Mel Frequency Cepstral Coefficients (MFCC's). Therefore, one aim of the present paper is to investigate whether the same difference between parameter representations is also found with GSM coded speech.
In the original approach, the use of MFCC's implied a band limiting operation: log-energy spectral values are computed in the frequency range between 300 Hz and 3.4 kHz only. However, in practice it appears that many telephone channels actually transmit spectral energy in a broader band, below 300 Hz as well as above 3.4 kHz. The default approach in HTK to compute LPCC's on 8 kHz signals takes account of the full frequency band from DC to 4 kHz. As a corollary to our study of the difference between MFCC's and LPCC's we included a study into the effects of the bandwidth reduction in default MFCC's.

2. DATABASE AND PARAMETERISATIONS

In this section the SESP database used in the experiments is described. Furthermore, the GSM transcoding is explained and the two types of parameterisations used are presented.

2.1. SESP database

The SESP database for SV research comprises 24 males and 22 females, all adult native speakers of Dutch. All speakers made between 14 and 32 calls to a Rhetorex ISDN platform, which stores 64 kbit/s A-law coded signals. Recordings started in March 1994, and were completed in May 1994. Among the items recorded in each session were three tokens of a 14 digit card number (*scope* numbers, the calling card programme of PTT Telecom). The subject's own *scope* card number was pronounced twice in each session. The third token was the *scope* card number of another subject of the same sex. The latter tokens were recorded as impostor material. The subjects

called from a wide variety of locations (a substantial proportion of the calls came from foreign countries) and environments (restaurants, public phones on airports, etc., in addition to home, office and quiet hotel rooms). Due to the sampling protocol most calls used a unique handset; only a small proportion of the handsets was cordless; no call originated from a cellular network.
In order to avoid the complexities caused by an integration of Automatic Speech Recognition and Speaker Verification research, the corpus material was limited to those utterances that contain exactly 14 digits.

2.2. GSM codec simulation

The A-law coded signals were transformed to GSM with software that is publicly available (GSM 06.10) [2]. In the GSM transcoding only the effects of the LPC coding and data compression were implemented. No attempt was made to simulate the effects of bit errors due to distortions in the radio transmission channel. However, a substantial part of the original A-law signals already contained relatively high levels of background noise.

The GSM codec simulation transforms 20 ms input frames to short-time predictable parts, long-term predictable parts and a remaining residual signal. During the first stage, 8 reflection coefficients are calculated (LPC). The long-term prediction (LTP) stage divides the LPC residual into subframes of 5 ms. Subsequently, each subframe is correlated with 15 ms of (reconstructed) LPC residuals preceding it. This LTP stage yields the time offset of the subframe with the maximum correlation and a scaling factor. In the residual pulse excitation (RPE) stage the scaled subframe is subtracted from the LPC residual. Next, this reduced residual is divided into 4 evenly spaced subsequences. The sequence with the maximum energy is coded using APCM and transmitted together with the corresponding index.

The decoder starts by expanding the sequence into a residual pulse of 5 ms, zero-padding the gaps in the downsampled signal. The resulting residual is fed through the LTP synthesis stage, which adds the reconstructed subframe indicated by the offset and scaling factor parameters. In the last stage of decoding 4 of these reconstructed short-time residual frames are fed through the short-time synthesis filter whose reflection coefficients were calculated during the LPC stage of coding.

2.3. Parameterisations

One aim of our research was to compare the effectiveness of MFCC's and LPCC's as parameter representations in SV with GSM coded signals (and implicitly also for A-law signals). In addition, we wanted to investigate the impact of widening the frequency band represented by the cepstrum coefficients to the full band from 0 to 4 kHz. Experience shows that a large proportion of the recordings contains substantial spectral energy below and above these cut-off frequencies. Especially the high part of the spectrum is assumed to contain useful information on the speaker's identity [5]; therefore, it makes sense to try to use that information in automatic SV systems.

In all experiments we applied a 97% pre-emphasis and used a Hamming window (width 25.6 ms, shift 10 ms) as a first step towards the calculation of 12 cepstrum parameters. Cepstral mean subtraction over the total utterance duration was applied in all cases. A sine-lifter was applied to the cepstrum coefficients, since liftering was shown to improve performance in the mainstream CAVE research [1]. The exact same parameterisations have been applied to both the A-law and the GSM-transcoded SESP-data. We have used three different parameterisations (the [g] prefix refers to GSM coded signals):

- ([g]MFCC):
  liftered Mel-frequency cepstral coefficients obtained from an FFT analysis of the speech frames, after which the spectral coefficients were combined to log-energies in equally wide filter bands on a Mel frequency scale. [g]MFCC's have been used in two versions, viz. band limited to $300 - 3400$ Hz and full band ($0 - 4000$ Hz).
- ([g]LPCC):
  liftered linear prediction cepstral coefficients, obtained by a cepstral transformation of the LPC coefficients on a linear frequency axis. Due to the way in which HTK implements the computation, [g]LPCC coefficients cover the full 4 kHz frequency band.
- ([g]MLPCC):
  liftered Mel-frequency cepstral coefficients obtained by computing the spectrum from the LPC parameters, and subsequently combining the spectrum coefficients to log-energies in filter bands on a Mel frequency scale in exactly the same way as with the MFCC's. [g]MLPCC's have been used in two versions, viz. band limited to $300 - 3400$ Hz and full band ($0 - 4000$ Hz).

In all experiments the 12-cepstrum coefficients were augmented with energy (dB) and their first and second time derivatives, making for a 39-element feature vector.

3. EXPERIMENTAL PROTOCOL

In order to determine the impact of the GSM transcoding we used HMM_LR word-based models with diagonal covariance matrices [1]. This specific SV method was chosen because it appeared to yield excellent results on A-law data in the CAVE mainstream experiments.

The complexity of a model can be described by the number of states p for each phoneme in the canonical phonemic transcription of the digits and the number of Gaussian mixtures q in each state. For this research the complexities $p * q = 2 * 1$, $2 * 2$, $3 * 2$ and $4 * 2$ were used to build word models.

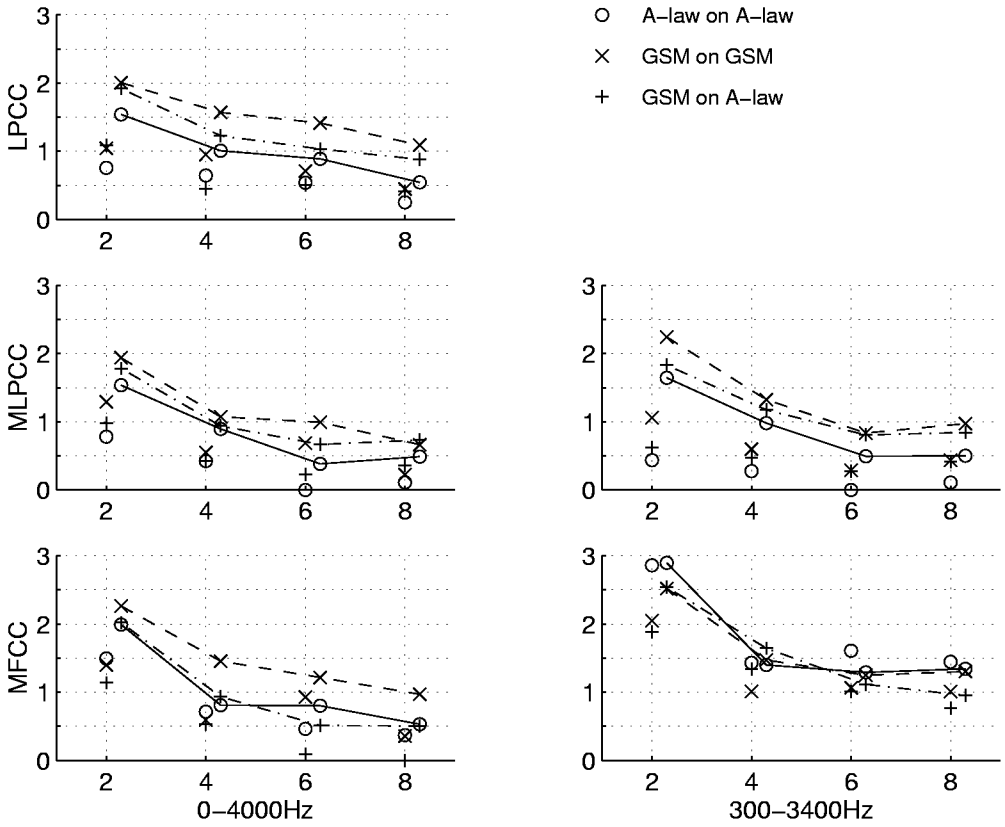
3.1. Enrolment

The enrolment procedure consists of learning the parameters for each speaker and each word in the vocabulary, viz. the 10 digits, based on training material that was pre-segmented automatically.

The enrolment set is composed of 8 *scope* numbers per client (21 males and 20 females), selected from 4 sessions originating from at least 2 different quiet recording environments (probably corresponding with different handsets).

Also, speaker independent world models were built for the ten digits, using a pool of speakers from the Dutch POLYPHONE corpus, in order to allow us to use a log-likelihood *ratio* evaluation during the verification phase.

Figure 1. Impact of GSM-coding on SV results: In each of the panels the results of the three conditions tested are shown. Horizontal axis: models of increasing complexity. Vertical axis: EER. The reference test results are indicated by 'o's, 'x's denote the results of the GSM data on GSM models, while the '+'s give the results for GSM data on the A-law models. Connected symbols denote the SS EER, the isolated symbols denote the FF EER. The figures in the left column show the results of the experiments on the full-bandwith data,the right column shows the results on the band-limited data.



3.2. Verification

The verification phase does not require any segmentation, since it uses the linguistic content of the known verification utterance to constrain the word sequence (forced recognition). In computing the optimal alignment of client and world models with the speech utterance, optional speaker-independent silence model may be inserted between successive words. The verification set contains 2674 utterances, 1658 of which are true attempts.

The log-likelihood score for the claimant is obtained by summing of the log-likelihood scores of his/her individual word models for the digits in the test utterance. The log-likelihood obtained with the optimal alignment of the world models is then subtracted from the claimant's log-likelihood. This is equivalent to likelihood ratio scoring. The performance evaluation is based on a posteriori EERs (e.g. gender-balanced same-sex Equal Error Rate [SS EER]) [1].

4. RESULTS AND DISCUSSION

To investigate the impact of GSM coding on SV performance we carried out three series of tests. In the first series, the speaker models were built from A-law data and the verifications tests also used A-law data. The results of these experiments are used as a reference for the results obtained with the GSM data. In the second series the speaker models were built with GSM transcoded data and tested with GSM data. In the third series speaker models built from A-law data were tested with GSM coded speech.

The experiments were conducted with several parameter representations, i.e., MFCC's and [M]LPCC's. Moreover, we have compared a band-limited condition with results obtained for full band signals. The results are summarised in Figure 1.

4.1. Reference system

The best configuration for the A-law speech data is obtained with $p * q = 3 * 2$, with a SS EER of 0.382% and MLPCC's for feature vectors. Arguments for the superior performance of MLPCC's compared to MFCC's and linear LPCC's are given below. The slightly better performance of models with 6 Gaussians per phoneme (compared to 4 and 8 Gaussians) is most probably due to insufficient enrolment material to obtain stable estimates for more than 6 mixtures, while it is still true that more mixtures help improve performance. In other words, of the complexities tested 6 Gaussians per phoneme is the maximum acoustic resolution that allows for stable parameter estimates.

4.2. GSM results

For most conditions, tests with GSM coded speech yield worse performance than tests with A-law speech. It is

interesting, however, that the matched condition *GSM training and GSM test* yields lower performance than the mismatch condition, where the training used A-law signals and the test is on GSM. This finding remains to be further investigated and explained.
On average, the performance degradation due to GSM coding is rather small. Apparently, the effects of the signal and channel coding on the spectral properties of the reconstructed signals are marginal; they do not seem to warrant the development of special purpose parameter representations for GSM coded signals.
Contrary to our expectation, models based on LPC derived cepstral coefficients are not more robust against GSM coding than models based on FFT parameters. Apparently, the fact that GSM is a form of LPC coding does not translate into an advantage here. This may be due to the fact that Mel scale transformation of FFT derived spectra removes spectral zeros in about the same manner as does LPC spectral estimation.

It should be emphasised that the results in our simulation experiment do not imply that it is not necessary to train separate models for A-law and GSM coded speech in real applications. We have only investigated the impact of signal and channel coding, and we have found these to be marginal. In real applications other channel effects (echo, bit errors in radio transmission) are very likely to overshadow the effects of the coding algorithm.

4.3. Bandwidth effects

Comparing columns (especially the bottom row with MFCC data) in Fig. 1 shows that enlarging the analysis bandwidth to the full $0 - 4000$ Hz band does improve the performance. The relative contributions to SV performance of the low and high frequency additions remain to be investigated. However, we assume that the major improvement comes from the addition of the high frequency band. From the acoustic phonetic literature it appears that *speech* information is concentrated in the frequency range up to approximately 2500 Hz, whereas the higher frequency bands contain more *speaker* information [5].
In the MFCC case, the advantage of the full bandwidth only holds for the models trained with A-law signals (the '+' and 'o' symbols in the figure), but not for the models trained on GSM coded speech. Awaiting further experiments we assume that this is due to the less precise coding of high frequency information in GSM signals.

4.4. Spectral estimation effects

From Figure 1 it can be seen that the LP-based parameterisations outperform the FFT based parameterisation for both the band limited and full bandwidth data. Moreover, the Mel-scaled LPCC's perform slightly better than the linear LPCC's. The latter effect can be explained by the high frequency smoothing inherent in Mel scale filters: random variation in the high frequency region is reduced by the increasing bandwidth of the Mel scale filters.
Especially in the band limited condition (the right column in Fig. 1) it is clear that spectral estimates obtained from an LPC analysis outperform the estimates based on an FFT analysis. Although this remains to be proven, we assume that the advantage of LPC over FFT analysis is connected to the inherent smoothness of the LPC spectral estimates. Since LPC provides estimates of the spectral envelope, while FFT analysis may yield a mix of harmonic and envelope information, we surmise that the advantage of LPC is completely due to its superior suppression of the harmonic information. It is well known that, especially for high pitched speech, the lowest FFT based Mel scale filter bands may or may not happen to contain a single harmonic [3]. The unpredictable presence or absence of such a harmonic in spectrum has a large effect on the cepstrum coefficients. Therefore, smoothing FFT based Mel spectra in the low frequency range should perhaps be made standard part of any MFCC representation. However, the default HTK procedures do not perform this smoothing, and as yet we did not have the time to integrate it into the software.
Our explanation of the disadvantage of FFT based analysis relative to LPC spectral estimates is corroborated by the SS EERs obtained for females (FF EERs are represented by the unconnected symbols). In most conditions the results for the females are better than average. The most obvious exception is the band limited MFCC condition, where the problem with interaction between harmonics and spectral envelope is worst for high pitched female speech.

5. CONCLUSIONS

In this paper it is shown that the speech and channel coding in the GSM cellular network has only marginal effects on the performance of HMM models for speaker verification. The effects are so small that the development of new parameter representations is not warranted.
We have also shown that LPC derived spectral estimates are to be preferred over FFT based estimates, mainly because of the superior performance of LPC in estimating the spectral envelope. Especially in the low frequency bands FFT derived Mel scale filter energy estimates may suffer from interaction between individual harmonics and envelope information. This effect is especially detrimental with high-pitched female voices.
Despite the formal limitation of telephone channels to the frequency band of $300 - 3400$ Hz it is shown that using information on the full band from DC to 4 kHz in A-law coded speech improves speaker verification performance.
Last but not least, the experimental protocols developed in the CAVE project have been invaluable in allowing us to design and perform the experiments.

REFERENCES

[1] Bimbot, Hutter, Jaboulet, Koolwaaij, Lindberg, Pierrot (1997) Speaker verification in the telephone network: An overview of the technical development activities in the CAVE project. *Proceedings EUROSPEECH-97.*

[2] J. Degener (1994) Digital speech compression: Putting the GSM 06.10 RPE-LTE algorithm to work. *Dr. Dobb's Journal*, December 1994

[3] van Alphen (1992) HMM-based continuous-speech recognition: systematic evaluation of various system components, *PhD thesis*, UvA Amsterdam.

[4] Young, Jansen, Odell, Ollason, Woodland (1995) *The HTK Book*, HTK 2.0 Manual.

[5] Furui, S. (1986) Research on individuality features in speech waves and automatic speaker recognition techniques, *Speech Communication, Vol. 5*, pp. 183-197.