RECOGNITION FROM GSM DIGITAL SPEECH

A. Gallardo-Antolín, F. Díaz-de-María and F. Valverde-Albacete

Departamento de Tecnologías de las Comunicaciones

Escuela Politécnica Superior. Universidad Carlos III de Madrid

C/Butarque, 15, 28911-Leganés (Madrid) SPAIN

e-mail: (gallardo, fdiaz, fva)@tsc.uc3m.es

ABSTRACT

This paper addresses the problem of speech recognition in the GSM environment. In this context, new sources of distortion, such as transmission errors or speech coding itself, significantly degrade the performance of speech recognizers. While conventional approaches deal with these types of distortion after decoding speech, we propose to recognize from the digital speech representation of GSM. In particular, our work focuses on the 13 kbit/s RPE-LTP GSM standard speech coder.

In order to test our recognizer we have compared it to a conventional recognizer in several simulated situations, which allow us to gain insight into more practical ones. Specifically, besides recognizing from clean digital speech and evaluating the influence of speech coding distortion, the proposed recognizer is faced with speech degraded by random errors, burst errors and frame substitutions. The results are very encouraging: the worse the transmission conditions are, the more recognizing from digital speech outperforms the conventional approach.

1. INTRODUCTION

Nowadays, the extensive use of digital mobile telephony is opening a wide range of opportunities for designing new Automatic Speech Recognition (ASR)-based applications, which benefit from the inherent mobility of these systems. However, mobility also challenges the ASR systems by introducing new sources of degradation. The main ones are the following [1, 2, 3]:

- noisy environment: many different situations (public places, running cars, etc), hands-free operation mode, etc.;
- speech coder distortion: low to medium bit rate speech coders (5.6 and 13 kbit/s are the standard half-rate and full-rate, respectively, in GSM) unavoidably introduce certain amount of distortion which has a significant effect on speech recognition performance [2, 3];
- transmission errors: due to the nature of the radio channel [4]; and
- typical "ad hoc" subsystems/characteristics of digital mobile telephony networks: such as voice activity detector (VAD), discontinuous transmission (DTX), or insertion of comfort noise.

The first of the above-mentioned sources of distortion has been addressed in many different ways in the context of GSM environment: speech enhancement, robust parameterizations, model compensation, etc., with promising results [1, 5, 6].

The remaining types of distortion, however, have not received, as far as we know, the proper attention. Several studies have been conducted to quantify the influence of speech coding algorithms on ASR [2, 3], concluding that the lower the bit rate is, the more significant the degradation of ASR performance.

In this paper, we investigate the influence of speech coding and transmission errors on conventional speech recognition systems, and propose a novel approach to cope with these types of distortion: recognizing from a parameterization directly derived from the digital speech representation used in GSM. In particular, our experiments focus on the full-rate coder [7].

The paper is organized as follows: section 2 briefly presents the baseline system and the data-base used for the experiments. In section 3, we describe the procedure to derive a suitable parameterization from the digitally encoded speech signal. In section 4, we present the experiment conditions and show the results. Finally, we draw conclusions and outline future work.

2. BASELINE SYSTEM AND DATA-BASE

For the speech recognition experiments, we use a data-base integrated by 72 speakers and 11 utterances per speaker for the ten Spanish digits. This data-base was recorded at 8 kHz and in clean conditions. In addition, we have digitally encoded this data-base using the full-rate GSM standard (software freely available at [8]), so that we have both the clean and the encoded data-bases at our disposal.

We have divided each data-base in two sets: a training set consisting of 7040 utterances from 64 speakers, and a test set formed by the 880 utterances from the remaining 8 speakers. None of the speakers in the testing set is used in the training process.

The baseline is an isolated-word, speaker independent HMM-based ASR system developed using the HTK package [9]. Left-to-right HMM with continuous observation densities are used. Each of the whole-digit models contains a different number of states (which depends on the number of allophones in the phonetic transcription of each digit) and three Gaussian mixtures per state.

3. TWO PARAMETERIZATIONS, TWO ASR SYTEMS

In all of the experiments, we have used a MFCC-based front-end as parametric representation of the speech signal. The feature vectors consist of 12 mel-cepstral, one log-energy, 12 delta-cepstral and one delta log-energy coefficients, for a total dimension of 26.

The essential difference between a conventional ASR system and the proposed approach is the source signal from which the parameterization is derived. Thus, we build two ASR systems, each one starting from a different signal to compute the feature vectors. The first starts form the decoded speech and proceeds as usually, while the second starts from the LPC spectrum previously computed by the full-rate GSM standard encoder. These two different ways of computing the feature vectors are described more in-depth in the next subsections.

3.1. Parameterization Derived from Decoded Speech

In this approach, feature extraction is carried out on the decoded speech signal, which is analyzed once every 10 ms employing a 20 ms analysis Hamming window, using the HTK package [9]. Twelve mel-spaced cepstral coefficients are obtained using a FFT-based filter bank with 40 channels. Then, the log-energy, and the 12 delta-cepstral and the delta-log energy coefficients are appended.

3.2. Parameterization Derived from GSM Digital Speech

Since the full-rate GSM coder estimates the LPC spectrum from clean speech, we propose to take advantage of this clean speech representation in the front-end of the ASR system.

Recognition from this clean speech parameterization has two important advantages: first, we are avoiding the speech coding distortion and second, and more important, we are also circumventing all the transmission errors that do not directly affect to the encoded LPC spectrum.

The scheme in Figure 1 illustrates the suggested parameterization procedure along with the conventional one.

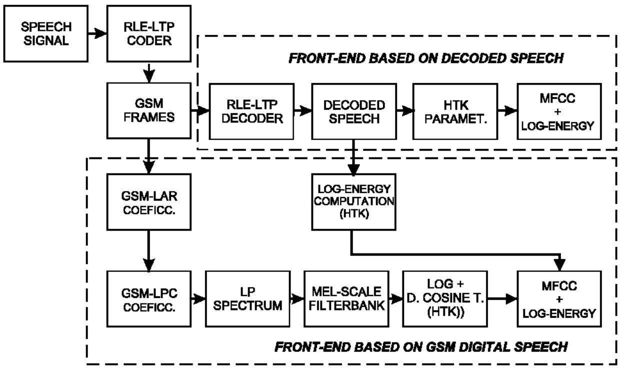

Figure 1: Parameterization procedures.

Our work, as detailed next, mixes our own procedures with some facilities of the HTK (HTK Toolkit) package. More precisely, the trans-parameterization (from LAR to MFCC) is described in detail through the following steps:

1. For each GSM frame (20 ms of speech for the full rate standard) its eight LAR ("Log-Area-Ratio") coefficients are extracted. After decoding them, they are converted to LP coefficients. Note that in this case, we do not overlap windows, so the number of frames in a certain utterance is half that in the previous case.
2. A 256-point spectrum of the speech frame is computed from the LP coefficients.
3. A filter bank composed of 40 mel-scale symmetrical triangular bands is applied to weight the LP-spectrum magnitude, yielding 40 coefficients.
4. The 40 coefficients obtained by the mel-band weighting are converted to 12 mel cepstrum coefficients (this is done by the HTK software).
5. A log-energy coefficient is appended. For simplicity, we have extracted this information from the decoded speech waveform using HTK, but similar coefficients could have been obtained from the digital GSM stream directly. The performances of both approaches have not been compared, although no significant differences are expected.
6. Dynamic parameters are computed (by HTK) for all the 12 MFCC and the log-energy, so that, a 26-parameter vector is used to represent each GSM-encoded speech frame.

4. EXPERIMENTAL RESULTS

In order to determine the impact of different types of GSM distortion in the performance of ASR systems and evaluate the effectiveness of our approach under these conditions, we have carried out three different sets of experiments:

- Training with clean data and testing with GSM decoded data (labeled as "clean-decoded"). Clean data refers to speech data without coding.
- Training and testing with parameterization derived from GSM decoded data (labeled as "decoded-decoded").
- Training and testing with parameterization derived from LAR coefficients obtained from GSM frames (labeled as "digital-digital").

GSM data used for training does not contain transmission errors.

In addition, a baseline experiment using clean data for training and testing (labeled as "clean-clean") has been performed for comparative purposes.

4.1. Influence of coding distortion

Table 1 shows the recognition results for different training and test configurations described above. It can be seen from this table below that full-rate GSM coding does not affect significantly the recognition rate when there are no transmission errors. In our opinion, because the RPE-LTP is a coder that was designed taking in account some perceptually related criterion. Similar results have been reported in [1]. However, we expect that using a half-rate GSM codec will markedly decrease the ASR performance as shown in [2-3].

On the other hand, using the parameterization derived from GSM digital speech ("digital-digital") does not decrease the performance significantly. This approach is thus suitable for recognition purposes.

TRAINING	TEST	Recognition Rate (%)
Clean	Clean	99.77%
Clean	Decoded	99.89%
Decoded	Decoded	99.89%
Digital	Digital	99.66%

Table 1: Recognition results for different combination of training and testing conditions.

4.2. Influence of transmission errors

The GSM coding scheme contains several mechanisms to protect encoded speech from transmission errors due to the radio interface [4]. When heavy errors are detected, the damaged frame is discarded and replaced by a previous correctly-received frame. Nevertheless, non-replaced frames may contain errors, since some bits of the stream are not protected.

We have decided to simulate separately both disturbing conditions, replaced frames and transmission errors, in order to evaluate their influence on recognition performance.

Frame substitutions

We simulate frame substitution by randomly repeating frames at different rates. In these experiments, non-replaced frames do not contain errors. Recognition rates are summarized in Figure 2. It can be seen that performance does not degrade significantly when the number of replaced frames is not very high (less than 20 %).

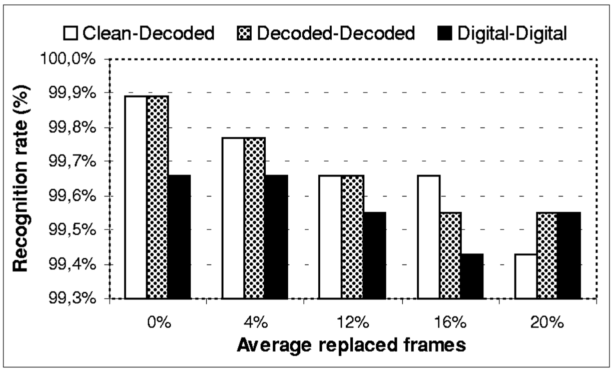

Figure 2: Performance for different frame substitutions rate.

Random and burst errors

In order to measure the influence of transmission errors on the ASR system, we have artificially degraded the GSM-encoded speech with both random and burst errors at different BERs ("Bit Error Rates"). Nevertheless, there are no frame substitutions, even when a frame is heavily affected.

Simulating random errors is performed by adding errors to the GSM coded frames at the bit level. Since such kind of errors does not describe in a realistic way the transmission conditions (multipath, fading, etc.) in mobile communications, we have added burst errors to the GSM-coded frames.

Burst errors are inserted using a simple model [10] composed by two states, the first one with low bit error rates (P_1) and the second one in which transmission errors are highly probable ($P_2 >> P_1$). In normal conditions, the system is in the first state, but it shifts to the second state when the effect of fading is randomly simulated. In real GSM communications systems, the transition probability from state one to state two (P_s) is rather low. Figure 3 shows the structure of this model.

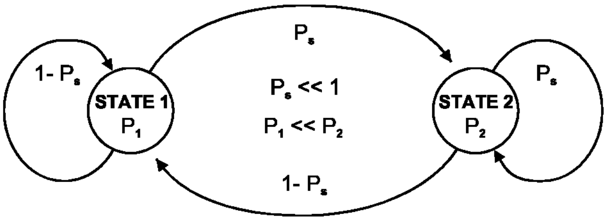

Figure 3: Model of burst-errors simulation.

The bit error rate (BER) is computed as follows:

$$BER = (1 - P_s)P_1 + P_s P_2$$

In our implementation, we contaminate each frame at the bit level according to the error probability of the current state. Skips from one state to another are only allowed on a frame-by-frame basis.

Figure 4 summarizes the recognition rates for both random and burst errors at different BERs, that we suppose feasible in mobile transmissions. In Tables 2 and 3 we compare the recognition performance in both cases.

		BER for Random-Errors		
TRAIN.	TEST	1e-3	5e-3	1e-2
Clean	Decoded	99.55%	96.59 %	90.80%
Decoded	Decoded	99.32%	95.91%	90.34%
Digital	Digital	99.20%	97.73%	95.23%

Table 2: Recognition rates for different bit-error probabilities (BER): Random errors.

		BER for Burst-Errors		
TRAIN.	**TEST**	1e-3	5e-3	1e-2
Clean	Decoded	99.09%	93.64%	81.59%
Decoded	Decoded	98.86%	92.84%	80.80%
Digital	Digital	99.20%	96.02%	90.91%

Table 3: Recognition rates for different bit-error probabilities (BER): Burst errors.

Results show that, as expected, recognition rate decreases dramatically when error rate increases. The effect is even worse for burst-errors, because in this condition, several contiguous frames can be seriously damaged.

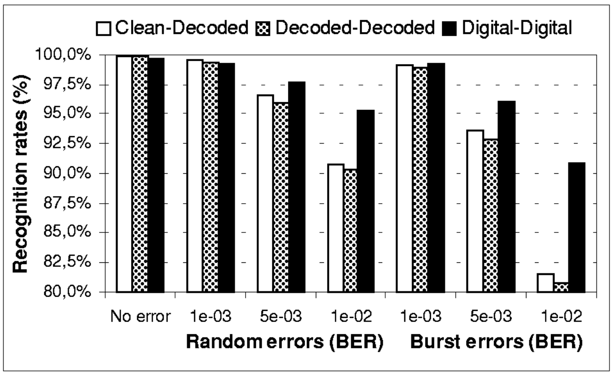

Figure 4: Recognition rates for both random and burst errors at different channel conditions.

Improvements are not achieved using GSM decoded speech for training and testing in the case of transmission errors. Probably, because the distortion produced by errors is more important than the distortion of the codec itself.

Our proposed approach ("digital-digital") exhibits higher robustness to all sorts of BER situations than using the decoded speech. The reason is that errors in bits non-corresponding to the LAR coefficients are avoided in this procedure.

5. CONCLUSIONS AND FURTHER WORK

In this paper, we have presented a new approach to ASR in the GSM environment. Instead of recognizing from the decoded speech signal, our system works from the digital speech representation used by the GSM encoder (we have focused on the full-rate standard).

We have studied the influence of coding distortion and transmission errors on the performances of the proposed speaker independent, isolated-digit ASR system in comparison to a conventional one. And, even though simulations concerning transmission errors do not represent accurately the GSM environment conditions, the achieved results allow us to conclude that the proposed approach is much more effective in coping with these problems than conventional approaches.

We plan to continue this research by completing our simulation system to include all the GSM subsystems involved and extending our results to the half-rate GSM standard.

On the other hand, evaluating the effects of tandemings of ADPCM and LD-CELP to GSM standards on the recognition performance will be worthwhile (although only small degradations can be expected), to take into account phone calls coming from the fixed telephone network.

6. ACKNOWLEDGMENTS

This work is partly supported by Ericcson. The authors would like to thank Rodrigo Bravo for his help in the early stage of this work.

7. REFERENCES

1. Dufour S., Glorion, C. and Lockwood, P. "Evaluation of the Root-Normalised Front-End (RN_LFCC) for Speech Recognition in Wireless GSM Network Environments". *ICASSP-96, Vol. 2, pp. 77-80.* 1996.

2. Euler, S. and Zinke, J. “The Influence of Speech Coding Algorithms on Automatic Speech Recognition”. *ICASSP-94, Vol. 1, pp. 621-624.* 1994.

3. Lilly, B. T. and Paliwal, K. K. "Effect of Speech Coders on Speech Recognition Performance". *ICSLP-96, Vol. 4, pp. 2344-2347.* 1996.

4. Mouly M. and Pautet M. B. "The GSM System for Mobile Communications". *Published by Moulin Consultant,* 1992.

5. Mokbel, C., Mauuary, L., Karray, L., Jouvet, D., Monné, J., Simonin, J. and Bartkova K. “Towards improving ASR robustness for PSN and GSM telephone applications”. *Speech Communication, Vol. 23, pp. 141-159.* 1997.

6. Salonidis, T. and Digalakis, V., “Robust Speech Recognition for Multiple Topological Scenarios of the GSM Mobile Phone System”. *ICASSP-98, Vol. 1, pp. 101-104.* 1998.

7. Vary, P. Hoffman, R. Hellwig, K. and Sluyter, R. "A Regular-Pulse Excited Linear Predictive Coder". *Speech Communication, Vol. 7, no. 2, pp. 209-215.* 1988.

8. Degener, J. and Bormann, C. "GSM 06.10 13 kbit/s RPE/LTP speech compression software". *Technische Universitaet Berlin,* 1992.

9. Young S. et al. "HTK-Hidden Markov Model Toolkit (ver. 2.1)". *Cambridge University.* 1995

10. Lin, S. and Costello, D. J. "Error Control Coding: Fundamentals and Applications". *Prentice-Hall.* 1983.