AVOIDING DISTORTIONS DUE TO SPEECH CODING AND TRANSMISSION ERRORS IN GSM ASR TASKS

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

Departamento de Tecnologías de las Comunicaciones
EPS-Universidad Carlos III de Madrid
C/ Butarque, 15, 28911-Leganés (Madrid), SPAIN

ABSTRACT

In this paper, we have extended our previous research on 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 compared the performance of a conventional system and the one we propose on a speaker independent, isolated- digit ASR task. For the half and full-rate GSM codecs, from our results, we conclude that the proposed approach is much more effective in coping with the coding distortion and transmission errors. Furthermore, in clean speech conditions, our approach does not impoverish the recognition performance, even recognizing from GSM digital speech, in comparison with a conventional system working on unencoded speech.

1. INTRODUCTION

In a very near future, many of the Automatic Speech Recognition (ASR) applications dealing with interactive remote access to information and services will have to deal with speech coming from mobile phones. But, before these ASR systems become a reality, effective procedures should be developed to tackle the new sources of degradation introduced by the digital mobile telephony systems. The main ones are listed below:

- *noisy environments:* (public places, car cockpit, etc), hands-free operation mode, etc. [1];
- *speech codec (encoder-decoder) distortion* [2,3,4];
- *transmission errors:* due to the nature of the radio channel; and
- *characteristic subsystems of digital mobile telephony networks:* such as discontinuous transmission (DTX), or insertion of comfort noise.

During the last years, several techniques have been proposed to tackle the ASR problem when speech comes from a digital mobile phone: speech enhancement [5], robust parameterizations [1], model compensation [4], etc. All of these approaches aim at coping with almost any source of noise without caring about its origin, since it can be very varied. But, distortions characteristic to these scenarios can be better faced by specific procedures which take advantage of our knowledge about their origin, from our point of view.

In this paper, we extend previous work [6] in which we proposed a novel approach to deal with distortions due to speech coding and transmission errors: recognizing from a parameterization directly derived from the digital speech representation used in the digital mobile telephony system. In [6] we presented a preliminary research on the GSM system, focusing on the full-rate standard [7]. Here, we continue this work to cover both half- [8] and full-rate standards. Thus, from now onwards, we will concentrate on distortions due to speech coding and transmission errors.

This paper is organized as follows. in Section 2, we describe the procedure to derive a suitable parameterization from the digitally encoded speech signal. In Section 3, we present the experimental protocol and results. Finally we draw conclusions and outline our future work.

2. RECOGNIZING FROM DIGITAL SPEECH

2.1 Influence of Speech Coding and Transmission Errors on Recognition Performance

The main problem for noisy speech recognition lies in reducing the mismatch between training and testing conditions. The conventional solution to tackle this problem when the noise comes from the source coding algorithm and transmission errors involves training the recognizer using decoded speech. However, the results will be always poorer than those achievable when recognizing unencoded speech.

Very limited work has been reported on the influence of coding distortion in speech recognition [1,2,3,4]. However, it seems clear that:

- recognition rates for coded speech significantly improve when the speech recognizer is trained using speech processed by the same coding algorithm [2];
- speech codecs working at or above 16 kb/s do not affect significantly the recognition performance, even when the speech goes through several tandeming stages [3];
- for bit rates below 16 kb/s, however, the loss of recognition accuracy is very significant (the lower the bit rate is, the poorer the accuracy), even with matched training and test conditions [1,2,3,4].

Since speech codecs for mobile telephony operate, in general, below 16 kb/s (in particular, the half- and full-rate GSM codecs work at 5.6 and 13 kb/s at rates, respectively), it can be concluded that the speech coding distortion will significantly affect the recognition performance.

0-7803-5041-3/99 $10.00 © 1999 IEEE

2.2 Avoiding Distortions Due to Speech Coding and Transmission Errors

Since all of the digital mobile telephony systems use standard codecs, we know exactly how the speech signal is coded. Therefore, we propose to benefit from the digital speech representation, obtained directly from clean speech, to make the recognition task easier.

Recognizing from encoded (digital representation) speech will be possible as long as some of the parameters extracted by the speech encoder are suitable to feed the recognizer, i.e., keep up the relevant information from the recognition point of view. Fortunately, CELP-type (Code Excited Linear Predictive) coders always used at the typical rates of cellular systems, extract and code the appropriate spectral information, from which recognition can be successfully carried out. Therefore, we suggest to perform speech recognition from a parameterization directly derived from the one provided by the encoding algorithm, instead of reconstructing the (decoded) speech and then, obtaining the parameterization.

The advantages of the proposed approach are very significant:

- First of all, to avoid the speech reconstruction (decoding) distortion by recognizing from a parameterization obtained from clean speech. As a consequence, recognition rates close to those achieved for unencoded speech are expected.
- The time requirements are reduced, since a complete decoding is not necessary.
- Furthermore, under noisy channel conditions, the proposed procedure prevents the recognition rate from being affected by most of the transmission errors: only the errors which affect the spectral envelope coding will produce a loss of recognition accuracy.

2.3 Deriving the Parameterization from GSM Digital Speech

The only difference between a conventional ASR system and the proposed approach is the source signal from which the parameterization is derived. The first one starts from the decoded speech and proceeds as usually, while the second one starts from the previously computed LPC spectrum, output by the GSM standard encoder. In any case, we have used a MFCC-based front-end as parametric representation of the speech signal. More precisely, 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 block diagram in Figure 1 illustrates the suggested parameterization procedure as compared to the conventional one. Our implementation mixes our own procedures with some facilities of the HTK (HTK Toolkit) package [9]. More precisely, the trans-parameterization (from the quantized LP parameters used by the encoder to MFCC) is described below step by step:

1. For each GSM frame (20 ms of speech for both the half- and the full-rate standards), the quantized LP parameters (ten reflection coefficients for the half-rate codec, and eight LAR -Log Area Ratio- for the full-rate one) are extracted. After decoding them, they are converted to LP coefficients.
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, which are converted to 12 mel cepstrum coefficients using the HTK software.
4. A log-energy coefficient is appended.
5. Dynamic parameters are computed (by HTK) for all the 12 MFCC and the log-energy.

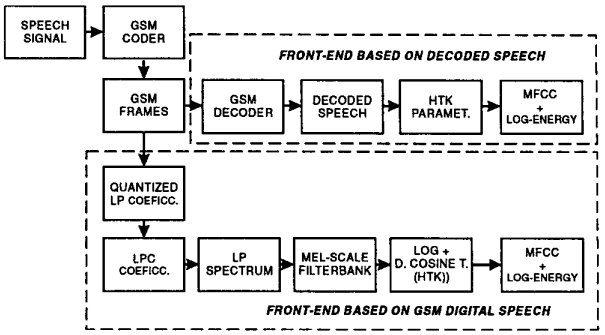

Figure 1: Parameterization procedures.

On the other hand, in the conventional approach which serves as reference, 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. Note that the number of feature vectors per utterance is twice that in the proposed procedure because of the frame rate.

3. EXPERIMENTAL RESULTS

3.1 Baseline System and Data-Bases

For the speech recognition experiments, we use a data-base consisting of 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 both the half- and the full-rate GSM standards [8, 7], so that we have three different data-bases at our disposal.

Since the data-bases are quite limited to achieve reliable speaker-independent results, we have used a 9-fold cross validation to artificially extend them. Specifically, we have split each data-base into 9 balanced groups; 8 of them for training and the remaining one for testing, averaging the results afterwards.

The baseline is an isolated-word, speaker independent HMM-based ASR system developed using the HTK package. 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.

3.2 Experiments and results

In this section we present the experiments carried out in order to analyze the influence of half- and full-rate GSM standard coding on the performance of an ASR system in several mismatch conditions. For both coders we have used the same experimental protocol; three sets of experiments have been conducted for each standard:

- Models trained with unencoded data and testing with GSM decoded data (labeled as "unencoded-decoded").
- Training and testing with GSM decoded data (labeled as "decoded-decoded").
- Training and testing with the parameterization derived from the quantized LP coefficients from GSM frames (labeled as "digital-digital").

3.2.1 Influence of coding distortion

Our aim now is to measure the loss of recognition accuracy due to the distortions introduced by the full and half-rate coding. We can observe from the Table 1 that higher recognition rates are obtained when an acoustical match between training and test conditions exists ("decoded-decoded") for both standards and are very similar to the baseline experiment ("unencoded-unencoded"). However, applying unencoded models to half-rate decoded speech makes the performance of the recognizer decrease as compared to the baseline and full-rate, due to the loss of useful information produced by the lower bit rate. Similar conclusions have been extracted in [2, 3].

The last row of the Table 1 ("digital-digital") shows the good performance of the system based in our proposal, which indicates that the GSM digital representation is suitable for recognition purposes.

		Recognition Rate (%)	
TRAINING	**TEST**	**Full-Rate**	**Half-Rate**
Unencoded	Unencoded	99.66%	99.66%
Unencoded	Decoded	99.43%	97.39%
Decoded	Decoded	99.53%	99.37%
Digital	Digital	99.25%	99.33%

Table 1: Recognition results for full and half-rate coders.

3.2.2 Influence of transmission errors

As transmission errors are very frequent in wireless communications, GSM contains several subsystems for protecting the streams of speech data.

On the other hand, the GSM channel coding is not capable of detecting and correcting all the errors, and some of them may be present in frames labeled as correct. In particular, if various consecutive frames are seriously damaged, they are replaced by an increasingly attenuated version of the last correctly-received frame. When the number of replaced frames is more than 5 (100 ms), the decoder mutes the output; in other words, a sudden disappearance of signal occurs ("GSM holes" [10]). GSM holes affect drastically the performance of the ASR.

The aim of this set of experiments is to evaluate separately the influence on recognition performance of both disturbing conditions: frame substitution and remaining transmission errors.

Frame substitutions

We simulate frame substitution by randomly setting the flag BFI ("Bad Frame Indicator") of each frame to 1 (unusable frame) at different rates. In theses cases, the decoder acts according to the procedure described in [11]. Recognition rates for the half-rate coder are summarized in Figure 2. It can be seen that performance decreases significantly when the number of replaced frames is high because the risk of GSM holes increases. This loss of accuracy is more significant in the case of using unencoded models ("unencoded-decoded").

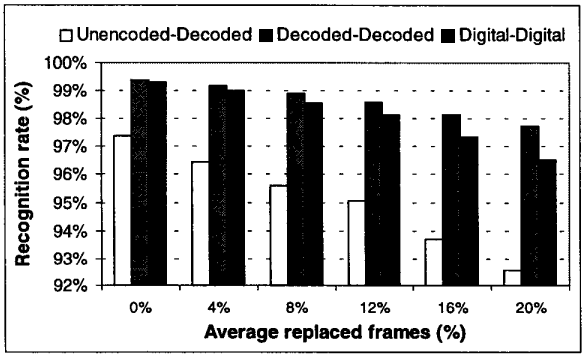

Figure 2: Performance for different frame substitutions rates (half-rate standard).

The performance of our approach ("digital-digital") is slightly worse compared to the "decoded-decoded". Normally, speech is resynthesized in the decoder, using the information of the actual frame and the state of the decoder, which depends on previous frames, whereas, in our approach, identical static MFCC parameters are obtained from identical GSM frames. This decrease of useful information affects the recognition rate.

Random and burst errors

We have also artificially degraded the GSM-encoded-speech with both random and burst errors at different BERs ("Bit Error Rates") with the purpose of evaluating the influence of transmission errors on the ASR system without frame substitutions.

Insertion of random errors is performed by adding random (unrealistic) errors and burst errors (more proper) to GSM-coded frames. The latter are inserted with a simple model composed by two states described in [6].

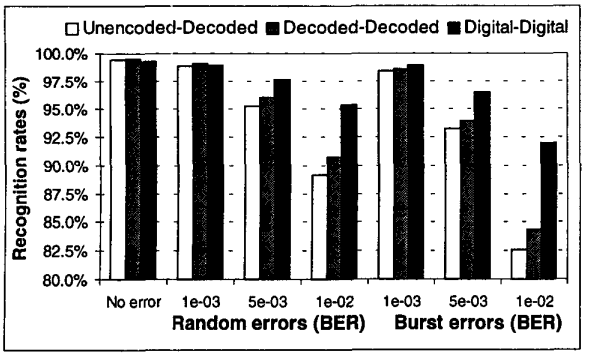

Figure 3: Recognition rates for both random and burst errors at different channel conditions (full-rate standard).

Figure 3 and Figure 4 summarize the recognition rates for both random and burst errors at different BERs for the full- and half-rate codecs, respectively. As expected, results show that recognition rate decreases dramatically when error rate increases in both cases. In most BER situations, little improvements have been achieved in matched conditions ("decoded-decoded"), probably because the effect of the distortion of the codec itself is avoided, but not that produced by errors. For all different BER and for the full and half-rate standards, our proposed approach ("digital-digital") exhibits higher robustness than using the decoded speech. The reason is that errors in bits other than those of the LP coefficients are ignored with this procedure.

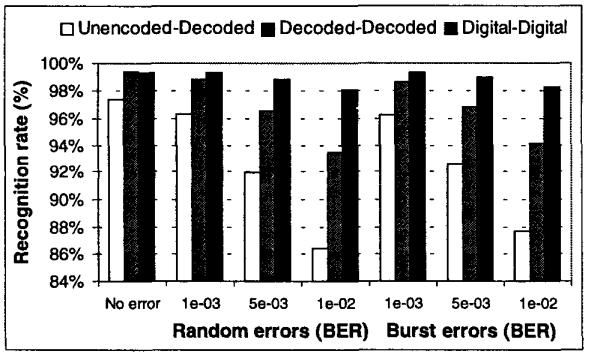

Figure 4: Recognition rates for both random and burst errors at different channel conditions (half-rate standard).

4. CONCLUSIONS AND FURTHER WORK

In this paper, we have extended our previous research [6] on 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 compared the performance of a conventional system and the one we propose on a speaker independent, isolated- digit ASR task. For the half and full-rate GSM codecs, from our results, we conclude that the proposed approach is much more effective in coping with the coding distortion and transmission errors.

Furthermore, in clean speech conditions, our approach does not impoverish the recognition performance, even recognizing from GSM digital speech, in comparison with a conventional system working on unencoded speech. Thus, conventional techniques to tackle noisy speech recognition can use the proposed front-end to avoid these types of distortion, and focus on other sources of noise or distortion.

We plan to continue this research by completing our simulation system to include all the GSM subsystems involved.

On the other hand, more elaborate solutions to frame substitutions should be developed taking into account all the information of previous frames.

5. ACKNOWLEDGMENTS

This work is partly supported by ERICSSON, S.A.

6. 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,* Atlanta, USA, Vol. 2, pp. 77-80, 1996.

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

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

[4] Salonidis, T. and Digalakis, V. "Robust Speech Recognition for Multiple Topological Scenarios of the GSM Mobile Phone System". *ICASSP-98*, Seattle, USA, Vol. 1, pp. 101-104, 1998.

[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,* 23:141-159, 1997.

[6] Gallardo-Antolín A., Díaz-de-María F. and Valverde-Albacete F. "Recognition from GSM Digital Speech". *ICSLP-98,* Sidney, Australia, 1998 (to be published).

[7] ETSI recommendation 6.10, "Full Rate Speech Transcoding".

[8] ETSI recommendation 6.20, "Half Rate Speech Transcoding".

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

[10] Karray, L., Jelloun, A.B. and Mokbel, C. "Solutions for Robust Recognition over the GSM Cellular Network". *ICASSP-98,* Munich, Germany, Vol. 1, pp. 261-246, 1998.

[11] ETSI recommendation 6.21, "Substitution and Muting of Lost Frames for Half Rate Speech Traffic Channels".