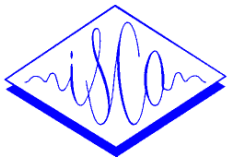

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

FILTERING THE SPECTRAL PARAMETERS TO MITIGATE THE INFLUENCE OF TRANSMISSION ERRORS ON ASR SYSTEMS

C. Peláez-Moreno, A. Gallardo-Antolín, J. Vicente-Peña and F. Díaz-de-María

Departamento de Teoría de la Señal y Comunicaciones
Universidad Carlos III de Madrid, Spain
fdiaz@tsc.uc3m.es

ABSTRACT

When voice-enabled remote access to information and services is considered, new sources of distortions -due to modern communication networks- should be taken into account. Focusing on mobile communication systems, speech coding distortion and transmission errors severely affect the performance of automatic speech recognition systems.

In this paper, we propose a filtering technique to mitigate the influence of transmission errors on recognition performance. In particular, considering the temporal evolution of each spectral parameter used for recognition as a temporal signal, we postulate that artificial high frequencies are generated in theses signals due to transmission errors. We suggest filtering them out. The proposed technique is assessed for two different parameterisations (type of spectral parameters): MFCC (*Mel Frequency Cepstral Coefficients*) and LSP parameters (*Line Spectral Pairs*). The results show that filtering is clearly effective.

1. INTRODUCTION

One of the factors that have contributed to the proliferation of digital multimedia contents is the possibility of accessing them remotely through communications networks. A friendly interface is one of the key subsystems for a successful access. A voice-enabled interface is perhaps the most friendly of the interfaces –in fact, speech is the most natural way of communication among human beings–.

Nevertheless, the modern communication systems pose new problems to voice-enabled services. Thus, the wide deployment of such interfaces will have to wait for effective solutions to these problems. In particular, focusing on mobile communication systems, speech coding distortion and transmission errors severely affect the performances of Automatic Speech Recognition (ASR) systems [13], [9], [4], [11].

During the last years, the ASR community has paid much attention to sources of distortion like variations of the acoustic environment, transducers and (phone) channel, or speaker variability. There is a lot of literature on this matter ([7], [2], [15], [14], [10] and [8] are good examples). Nevertheless, specific methods to deal with coding distortion and transmission errors are much more recent and scarce ([9], [4], [17] and [11]).

In this work, we focus on the distortion due to transmission errors and propose a specific technique to deal with it. We start from a simple idea: the transmission errors eventually produce abrupt changes in the temporal evolution of spectral parameters used for recognition purposes. In other words, they increase the bandwidth of these spectral parameters –individually considered as a time series. Consequently, we propose lowpass filtering these spectral parameters time series in order to mitigate the influence of transmission errors on ASR systems. Although the problem so stated is new –transmission errors have not been specifically tackled in ASR–, the filtering technique is not novel itself. In fact, Hermansky et al. [7] and Nadeu et al. [14][15] proposed similar ideas in the context of noisy speech recognition.

The paper is organized as follows. Section 2 describes a general application scenario for these techniques and the particular model we have used in our experiments. Section 3 explains in detail the proposed method. Experimental results are presented in Section 4. And finally, conclusions and further work are outlined in Section 5.

2. THE APPLICATION SCENARY

As mentioned in the Introduction, our work aims at coping with the influence of transmission errors on ASR systems. This problem arises when remote recognition is considered. Figure 1 illustrates the general scenario. The speech is encoded at the user terminal, transmitted over a mobile communication network –eventually affected by transmission errors–, and decoded just before entering the ASR system.

For our experiments, we have considered the European GSM mobile system and the Half-Rate

Standard Speech Codec [1]. We have taken into account both the source and the channel coding. Bursty transmission errors have been simulated using the model described in [3]. In order to assess our proposal in several conditions, we have considered four different Bit Error Rates (BERs): 0, 10^{-3}, 10^{-2} and $5x10^{-2}$.

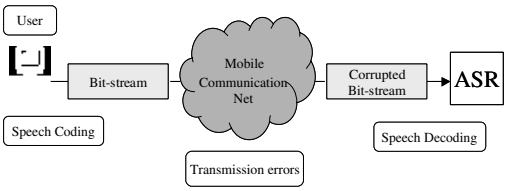

Figure 1.- A general scenario for remote automatic speech recognition

3. FILTERING THE TEMPORAL SEQUENCES OF SPECTRAL PARAMETERS

3.1. Motivation and related works

We postulate that transmission errors produce artificial high frequencies in the spectrum of the temporal sequences of spectral parameters used for recognition. Consequently, we suggest lowpass filtering these sequences to alleviate –to some extent– the devastating influence of transmission errors on ASR systems.

In our opinion, this is possible because the allocated bandwidth in GSM for the representation of these spectral features is higher than that needed. In particular, the GSM HR speech codec transmits the spectral envelope once every 20 ms; thus assuming a 25 Hz bandwidth. In the next subsection we will show how some representations of the spectral information requires lower bandwidths. This leaves room for the suppression of the high frequency components due to transmission errors.

As mentioned in the Introduction, similar filtering approaches have already been proposed [8][14][15]. We observe two fundamental differences between these contributions and ours: first, they emphasize removing low-frequencies (associated to channel distortion); and second, their work does not address the distortion due to transmission errors. Besides, as will be shown next, we apply this technique on different spectral parameters.

3.2. Effective bandwidth estimation

With the purpose of verifying the feasibility of our proposal, we have performed estimations of the aforementioned bandwidths for two alternative spectral parameters, namely: the MFCC (–*Mel Frequency Cepstral Coefficients*–), widely used in ASR, and LSP parameters (–*Line Spectral Pairs*–), mostly employed in speech coders due to their excellent transmission behaviour (see for example [12]). Our motivation for using LSPs is the following: they exhibit exceptional quantization characteristics which makes us expect lower bandwidths and, consequently, more chances for the proposed filtering technique.

The bandwidth estimation method we have employed is briefly outlined below:

- Firstly, we have extracted in parallel (always from clean speech) 10 LSP coefficients and 12 MFCC using a very small frame period: 0.25 ms – in other words, a sampling frequency of 4000 Hz. Using this frame rate, we are able to avoid any eventual spectral aliasing, which could corrupt posterior bandwidth estimations.
- Next, we have windowed the individual temporal sequences corresponding to every parameter using large Hamming windows. Specifically, we have used 2 s. windows with a 50 % overlap between neighbouring windows. On the one hand, such a large window duration obliges to us to work with a very poor temporal resolution; but, on the other hand, the frequency resolution is high, making possible to estimate effective bandwidths with an accuracy around 1 Hz –as it will be seen next, we are working with very low effective bandwidths.
- Finally, we have obtained the power spectral density of each window and averaged it over all of windows. We have computed what we have named *Effective Bandwidth* (Ebw) that corresponds to the frequency interval within which a specific fraction of the spectral power is concentrated. In particular we have employed a 90 % EBw.

The estimated EBw obtained following the previous steps is between 8 and 15 Hz for the different LSP parameters, being this bandwidth smaller for the upper order parameters. However, for the MFCC, the Ebw ranges from 10 Hz to 30 Hz and, contrary to the LSP behaviour, this bandwidth increases for the upper ones.

As can be observed these bandwidths are, with some exceptions, smaller than that allocated by GSM codec: 25 Hz. It is worth noting, as well, that the LSP parameters use, as expected, a considerably narrower bandwidth.

3.3. Filtering the parameterizations

While the MFCC parameterization is the most commonly employed in ASR, the previously described bandwidth measurements suggest that we can obtain more advantages by filtering the LSP coefficients

sequences. Therefore, we have applied the proposed method to both MFCC and LSP sequences.

For recognition purposes, the filtered LSP sequences were finally transformed into MFCC. For every speech frame, we have obtained a 256-point spectral envelope from the LSP coefficients (see [18]) and applied a filter bank composed of 40 mel-scale symmetrical triangular bands to weight the LP-spectrum magnitude. The outputs of the filterbank are subsequently converted into 12 mel cepstrum coefficients.

The frame rate employed for both parameterizations is 10 ms, thus providing a maximum bandwidth of 50 Hz, from which we have eliminated the high frequency band above the EBw estimated for each component of the feature vector.

4. EXPERIMENTS AND RESULTS

4.1. Databases and Baseline Systems

The database in which we have focused our experiments for the CSR task is the well-known Resource Management RM1 Database [16], which has a vocabulary of 991 words. The training corpus consists of 3990 sentences and the test set contains 1200 sentences, which corresponds to a compilation of the first four official test sets. Originally, RM1 was recorded at 16 kHz and in clean conditions; however, our experiments have been performed using a (downsampled) version at 8 kHz. We have employed context-dependent acoustic models (three-mixture cross-word triphones) and a simple language model (a word-pair grammar).

Both transmission errors and coding distortion are considered, as described in section 2, by using the GSM Half-Rate Standard Speech Codec (taking into account both the source and the channel coding) and a bursty channel model.

For the computation of the final parametric representation of the speech signal from which the recognition is performed two different parameterizations are employed:

- MFCC: 12 mel-cepstral and a log-energy coefficients are extracted every 10 ms using a hamming window of 25 ms from the decoded speech. Each individual coefficient is filtered using a tuned –the cutoff frequency is given by the estimated Ebw estimation– 8^{th} order lowpass FIR filter. Finally, 12 delta-cepstra and a delta log-energy coefficients are appended.
- LSP: 10 LPC (*–Linear Prediction Coefficients–*) and an energy coefficient are firstly computed from the decoded speech at the same rate than the previous parameterization. They are subsequently transformed into 10 LSP coefficients plus the energy. These parameters are filtered using the same procedure described for the previous parameters and later on transformed into 12 MFCC and energy using the procedure described in section 3.3. Finally, as in the previous case, 12 delta-cepstral and a delta log-energy coefficients are appended.

4.2. Confidence measures

In order to state the statistical significance of the experimental results shown in the next subsections, we have calculated the confidence intervals (for a confidence of 95%) using the following formula [19], (pp. 407-408):

$$\frac{band}{2} = 1.96\sqrt{\frac{p(100-p)}{n}} \qquad (1)$$

where p is the word accuracy and n is the number of examples to be recognized (10,288 words). Thus, any recognition rate in the tables below is presented as belonging to the band

$$\left[p - \frac{band}{2}, p + \frac{band}{2}\right]$$

with a confidence of 95%.

4.3. Results

Table 1 shows the recognition rates obtained for different channel conditions. Namely, for BERs of 10^{-3}, 10^{-2} and $5x10^{-2}$ (the first row shows the performance for a clean channel in which only the coding distortion is present[1]). Both, MFCC and LSP parameterizations are compared and the results are shown for each of them with and without the use of the proposed filtering.

As it can be seen, filtering is always effective for both parameterizations. Furthermore, the improvements become bigger while the channel worsens: from 2.01% to 6.21% for MFCCs and from 0.37% to 3.74% for LSPs.

Furthermore, it is worth noting a couple of things: first, even when no transmission errors are present, the filtering provides a significant improvement in the case of MFCCs; and second, though the improvement of the filtering in the case of MFCC is larger, the LSP parameterization always performs better. These observations lead us to conclude that recognizing from some type of smoothed spectrum seems to be more robust.

[1] As a reference the recognition rate of the ASR system without coding distortion and transmission errors is of 90.83%.

This conclusion agrees with well-established previous results on noisy speech recognition [5].

BER	MFCC		LSP	
	No filtering	Lowpass filtering	No filtering	Lowpass filtering
0	85.35 (84.5 ,86.1)	87.07 (86.3 ,87.8)	86.26 (85.5 ,87.0)	86.58 (85.8 ,87.3)
10^{-3}	85.30 (84.5 ,86.1)	87.11 (86.4 ,87.8)	86.14 (85.4 ,86.9)	86.62 (85.9 ,87.4)
10^{-2}	83.13 (82.3 ,83.9)	85.73 (85.0 ,86.5)	85.11 (84.3 ,85.9)	85.51 (84.7 ,86.3)
5.10^{-2}	55.37 (54.3 ,56.5)	58.81 (57.7 ,59.9)	58.29 (57.2 ,59.4)	60.47 (59.4 ,61.5)

Table 1: Recognition Rates illustrating the improvements achieved by filtering both MFCC and LSP parameterizations. Confidence intervals are provided in brackets.

5. CONCLUSIONS AND FURTHER WORK

From our results a couple of conclusions can be drawn: first, lowpass filtering of the spectral parameters have shown to be effective to deal with transmission errors; and second, recognizing from some type of smoothed spectral representation –either a lowpass filtered one or a LPC-based one– seems to be more robust against transmission errors than MFCC.

Some questions still remain open and deserve further investigation. Some examples follows: 1) Could the cutoff frequency of the filters be reduced for improving the recognition performance? or in other words, could better results be expected considering a 80 % Ebw or even lower? 2) We use a specific cutoff frequency for each parameter, is it worthwhile? 3) How would the approaches proposed by Hermansky et al. [7] and Nadeu et al. [14][15] perform if transmission errors are considered? …

ACKNOWLEDGEMENTS

This work has been partially funded by the Spanish grant CAM-07T-0018-2000.

6. REFERENCES

[1] ETSI Recommendation GSM 6.20, "Digital cellular telecommunications systems; Half Rate speech; Part 2: Half Rate Speech Transcoding", 1999.

[2] Gales, M. J. F., ""Nice" Model-Based Compensation Schemes for Robust Speech Recognition", *Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, Francia, pp. 55-66, 1997.

[3] Gallardo-Antolín, A., Vázquez-Castro, M., Díaz-de-María, F., Valverde-Albacete, F. and Pérez-Fontán, F., "BER Performance Assessment of the Land Mobile GSM Channel with Application to Automatic Speech Recognition Tasks", *Proc. 5th Bayona Workshop on Emerging Technologies in Telecommunications*. Vol. 1, pp. 212-216, 1999.

[4] Gallardo-Antolín, A., Peláez-Moreno, C., Díaz-de-María, F., "A robust front-end for ASR over IP and GSM networks: an integrated scenario", *Proc. of European Conference on Speech Communication and Technology (Eurospeech)*, vol.2, pp. 1103-1106, Aalborg, Dinamarca, Sep. 2001.

[5] Gold, B., Morgan, N., "Speech and Audio Signal Processing", New York, NY: John Wiley & Sons, 2000.

[6] Hanson, B., Applebaum, T., Junqua, J.-C, "Spectral dynamics for speech recognition under adverse conditions", in *Automatic speech and speaker recognition: advanced topics*, Editors: Lee, C.-H, Soong, F. K. y Paliwal, K.K, Ed. Kluwer Academic Publishers, pp. 331-356, 1996.

[7] Hermansky, H., Morgan, N., "RASTA processing of speech", *IEEE Trans. On Speech and Audio Processing*, vol. 2, no. 4, pp. 587-589, Oct. 1994

[8] Hircsh, H. G., "HMM Adaptation for Applications in Telecommunication", *Speech Communication*, vol. 34, issue 1-2, pp. 127-139, abril 2001.

[9] Huerta, J. M., "Speech Recognition in Mobile Environments", *PhD Thesis*, April 2000.

[10] Junqua, J. C., "Robust speech recognition in embedded systems and PC applications", Kluwer Academic Publishers, 2000.

[11] Kim, H. K., Cox, V., "A bitstream-based front-end for wireless speech recognition on IS-136 communications system", *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, Jul. 2001.

[12] Kondoz, A. M., "Digital speech : coding for low bit rate communication systems", Ed. John Wiley & Sons, 1996.

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

[14] Nadeu, C., Pachès-Leal, P., Juang, B.-H, "Filtering the time sequences of spectral parameters for speech recognition", *Speech Communication*. vol. 22, no. 4, pp. 315-32, Sep. 1997.

[15] Nadeu, C., Macho, D. Hernando, J., "Time and frequency filtering of filter-bank energies for robust HMM speech recognition", *Speech Communication*, vol 34, pp. 93-114, 2001.

[16] NIST, The Resource Management Corpus (RM1). Distributed by NIST, 1992.

[17] Peláez-Moreno, C., Gallardo-Antolín, A. and Díaz-de.-María, F., "Recognizing Voice Over IP. A Robust Front-End for Speech Recognition on the WWW", *IEEE Trans. on Multimedia*, vol 3, no. 2, jun. 2001.

[18] Sugamura, N., Itakura, F., "Speech analysis and synthesis methods developed at ECL in NTT –from LPC to LSP-", *Speech Communications*, vol. 5, pp. 199-215, 1986.

[19] Weiss, N.A., Hasset, M.J., "Introductory statistics", Third Edition. Reading, MA: Addison-Wesley, 1993.