SPEAKER VERIFICATION ON THE POLYCOST DATABASE USING FREQUENCY FILTERED SPECTRAL ENERGIES

Javier Hernando and Climent Nadeu

Polytechnical University of Catalonia
Barcelona, Spain
javier@tsc.upc.es

ABSTRACT*

The spectral parameters that result from filtering the frequency sequence of log mel-scaled filter-bank energies with a first or second order FIR filter have proved to be competitive for speech recognition. Recently, the authors have shown that this frequency filtering can approximately equalize the cepstrum variance enhancing the oscillations of the spectral envelope curve that are most effective for discrimination between speakers. Even better speaker identification results than using mel-cepstrum were observed on the TIMIT database, especially when white noise was added. In this paper, the hybridization of both linear prediction and filter-bank spectral analysis using either cepstral transformation or the alternative frequency filtering is explored for speaker verification. This combination, that had shown to be able to outperform the conventional techniques in clean and noisy word recognition, has yield good text-dependent speaker verification results on the new speaker-oriented telephone-line POLYCOST database.

1. INTRODUCTION

In current speaker recognition systems, the short-time spectral envelope of every speech frame is usually represented by a set of the Fourier series coefficients of its logarithm, i.e. the cepstral coefficients C(m), $1 \le m \le M$. These parameters are by far the most prevalent representations of speech signal [1] and contain a high degree of speaker specificity [2]. They are usually come either from a linear prediction (LP) analysis -LP-cepstrum-, or from a set of mel-scaled log filter-bank (FB) energies -mel-cepstrum [3]. Unfortunately, there are few comparative studies about the relative robustness to noise and distortions of mel-cepstrum with respect to LP-cepstrum.

Recently, the authors have considered a unified parameterization scheme for speech recognition that combines both LP and FB analysis [4]. It has been shown that an appropriate hybridization of both LP and FB approaches is capable of improving recognition results for both noisy and clean speech in continuous observation Gaussian density HMM digit recognition.

On the other hand, we may wonder if the cepstral coefficients are the best way of representing the speech spectral envelope, at least for some usual recognition systems. The sequence of cepstral coefficients C(m) is a quasi-uncorrelated and compact representation of speech spectra. Actually, the quefrency sequence is always windowed to eliminate the cepstral coefficients beyond M, and -for some type of recognizers- it also appropriately weights the remaining coefficients [2] [5] [6] [7]. However, cepstral parameters have at least three disadvantages: 1) they do not possess a clear and useful physical meaning as FB energies have; 2) they require a linear transformation from either log FB energies or the LP coefficients; and 3) in continuous observation Gaussian density HMM with diagonal covariance matrices, the shape of the cepstral window has no effect so that only its length, i.e. the number of parameters M, is a control variable.

In recent papers, the authors have shown that the spectral parameters that result from filtering the frequency sequence of log mel-scaled FB energies with a simple FIR filter of order 1 or 2 can be competitive with respect to mel-cepstrum for both speech recognition [8] [9] and text-independent speaker identification [10].

In the last work [10], it is shown that frequency filtering of log mel-scaled FB energies can approximately equalize the cepstrum variance enhancing the oscillations of the spectral envelope curve that are most effective for discrimination between speakers. Even better speaker identification results than using conventional mel-cepstrum were observed on the TIMIT database, especially when white Gaussian noise was added.

The aim of this paper is to gain some perspective of the merit of the hybridization of both LP and FB spectral analysis (section 2) using either cepstral transformation or the alternative frequency filtering (section 3) in speaker verification. The combination of these techniques has shown to be able to outperform the conventional LP and mel-cepstrum in clean and noisy word recognition [11]. In order to complement the last works [10] [11], text-dependent speaker verification experiments on the new speaker-oriented telephone-line POLYCOST database have been performed. In this way, better results than using conventional LP and mel-cepstrum have been observed in continuous observation Gaussian density HMM (section 4).

2. HYBRID SPECTRAL ANALYSIS

The strength of LP method arises from its close relationship to the digital model of speech production, so an appropriate deconvolution between vocal tract response and glottal excitation can be expected from it.

*This work has been supported by the grants TIC 95-1022-C05-03 and TIC 95-0884-C04-02

LP is a full-band approach to spectrum modeling. Conversely, the filter-bank (FB) approach removes pitch information and reduces estimation variance by integrating the periodogram in frequency bands. The FB approach separately models the spectral power for each band, and it offers the possibility of easily distributing the position of the bands in the frequency axis (a mel scale is employed in the so-called mel-cesptrum) and defining their width and shape in any desired way, to take advantage of the perception properties of the human auditory system. This sub-band working mode also has several advantages derived from the frequency localization of the parameters. For example, if the SNR of each band is known, it can be used in straightforward ways: noise subtraction, noise masking,...

The combination of LP and FB analysis may yield improved spectral parameters. One possible approach is to apply FB analysis on the signal prior to LP analysis [12] [13]. It will be referred to as FB-LP and it is computed similarly to the PLP coefficients [12], but using a higher order LP analysis without perceptual weighting and amplitude compression. An alternative approach is to use LP analysis followed by FB analysis (it will be referred to as LP-FB).

Both conventional LP and mel-cepstrum parameterizations and the cepstrum representations corresponding to the two new hybrid FB-LP and LP-FB can be encompassed in a unified parameterization scheme [4], that can lead to other novel speech parameterization techniques.

3. FREQUENCY FILTERING

HMM are mostly employed with diagonal covariance matrices. In that case, they implicitly assume uncorrelated spectral parameters. Conversely, the frequency sequence of log band energies is strongly correlated. The usual mel-cepstrum is a way of obtaining from log mel-scaled FB energies an almost uncorrelated set of parameters.

Decorrelation is thus a desired property for the sets of spectral parameters due to the particular way they are used in our current recognition systems. Nevertheless, what is really relevant to the own classification process is the discrimination capacity of those parameters

It is a known fact that the variance of C(m) decreases along the axis m [6]. Thus, the low quefrencies m will generally dominate the probability or distance computations in the classifier. We may ask whether this is the best we can do or a proper global variance equalization of C(m) could help to increase recognition performance, much like it occurs in speech recognition. Let us note that there exists a close relationship between equalization of the variance of C(m) at low quefrencies and decorrelation of log band energies.

However, a flat variance may not be the most adequate goal for recognition purposes. For example, when the frequency interval between bands is not large enough, that equalization gives too much weight to the estimation noise carried out by C(m) for high quefrencies. Another reason for not flattening it completely can be the presence of the acoustic channel characteristics or broad-band additive noise, which may require a stronger attenuation of the lowest quefrencies.

In [10] the ratio between inter-speaker and global variances of each cepstral coefficient was considered as a measure of its discrimination capacity. In the case of the clean TIMIT database, it was observed that the dynamic range of this ratio sequence was clearly smaller than that of the global variance. This fact suggested that an approximate equalization of the variance could help to increase the discrimination capability of the cepstral sequence, at least for clean speech.

On the other hand, this ratio sequence showed a slight increasing tilt along the quefrency index m. This fact led to think that the most discriminative information is located in the higher quefrencies, i.e. in the fast alternation of peaks and valleys of the spectral curve; and it is not in the lowest quefrencies, i.e. in the spectral tilt. Actually, most speaker recognition systems do use a higher number of cepstral parameters than speech recognizers. Even it could be convenient to slightly overemphasize the higher quefrencies.

Cepstral liftering (weighting on m) has been the usual way to compensate for the excessive weight of the lowest m terms in both speech [5] [6] [7] and speaker [2] recognition systems. In this case, two steps are needed for obtaining the final parameters from the log FB energies or the LP coefficients: 1) a linear transformation, that significantly decorrelates the sequence of parameters, and 2) a discriminative weighting (liftering) of the cepstral coefficients. Furthermore, in continuous observation Gaussian density HMM with diagonal covariance matrices, the shape of the cepstral window has no effect due to the intrinsic variance normalization of the Gaussian pdf.

In recent papers, in order to try to overcome those disadvantages and to have parameters that posses frequency meaning, an alternative to the use of cepstral parameters was introduced for speech recognition [8] [9] and speaker identification [10]. It consists in a simple linear processing in the log mel-scaled FB energy domain. The transformation of the sequence of log mel-scaled FB energies to cepstral coefficients is avoided by performing a filtering of that sequence, which we hereafter will call frequency filtering to denote that the convolution is performed on the frequency domain.

This frequency filtering produces both effects, decorrelation and discrimination, in only one step using a simple first or second order FIR filter. Moreover, frequency filtering is able to produce a cepstral weighting in an implicit way in continuous observation Gaussian density HMM with diagonal covariance matrices. Furthermore, it is worth noting the computational simplicity of frequency filtering with respect to the conventional cepstrum representations.

A first-order FIR filter that maximally equalizes the variance of the cepstral coefficients can be easily obtained by a least-squares modeling. Computation details can be found in [8]. However, even better results can be obtained if this frequency filter is empirically optimized. In the case of text-independent speaker identification on the TIMIT database, good results were obtained taking into account the slight increasing tilt along the

axis m of the ratio between inter-speaker and global cepstrum variances mentioned above [10].

The spectral parameters that result from filtering the frequency sequence of log mel-scaled FB energies have proved to be competitive with respect to mel-cepstrum for both speech recognition [8] [9] and text-independent speaker identification [10]. Even better speaker identification results than using conventional mel-cepstrum were observed on the clean TIMIT database, especially when white Gaussian noise was added.

However, frequency filtering not only can be performed on log FB energies, but it can also be applied when an LP analysis is performed, as it is described in [8], even in the case of the hybrid spectral analysis described in section 2. Text-dependent speaker verification experiments in all these cases will be reported in the next section on the new speaker-oriented telephone-line POLYCOST database.

4. VERIFICATION EXPERIMENTS

4.1. POLYCOST Database

The POLYCOST is a new speaker-oriented database that has been recorded as a common initiative within the European COST 250 action entitled 'Speaker Recognition in Telephony'. The database was collected through the European telephone network during January-March 1996. The recording has been performed with an 8 kHz sampling rate.

It contains around 10 sessions recorded by 134 subjects from 14 countries. The majority of non native English speakers gives the possibility to experiment intra-, inter-speaker, language and country variabilities. One session is set up of 15 prompts including one prompt for DTMF detection, 10 prompts with connected digits uttered in English, 2 prompts with sentences uttered in English and 2 prompts in mother tongue.

A set of baseline experiments has been defined. The text-dependent speaker verification experiment has been chosen in this work in order to complete the last work about text-independent speaker identification [10]. The task in this experiment is speaker verification on a fixed password phrase, which is common to all speakers, concretely, "*Joe took father's green shoe bench out*".

A client model per speaker has been built from the first 4 sessions. A world-model has been built from the first 5 sessions of 22 speakers that have been set aside as an off-line database.

True-identity tests are made on 5^{th} session and later sessions. With existing sessions for 110 client speakers chosen, this gives 666 true-identity tests. To simulate impostor attempts against speaker X, the 5^{th} session from all speakers in the database except speaker X is used. With 109 impostor tests per client, there are 11990 impostor tests in the experiment.

For comparison purposes, a common scoring software is used, which was developed within the CAVE project. This implementation is based on the methods described in the EAGLES handbook [14].

4.2. Recognizer Setup

The HTK recognition system, based on Continuous-Density Hidden Markov Models (CDHMM), was appropriately modified to perform speaker verification experiments with the novel speech representations.

In the parameterization stage, the speech signal (non-preemphasized) was divided into frames of 20 ms at a rate of 10 ms, and each frame was characterized by M=20 parameters obtained by any of the spectral analysis techniques considered above -LP, FB, LP-FB, FB-LP-, and using either cepstral transformation or frequency filtering. When an LP analysis was performed, the prediction order was fixed to 20. Also when a FB was used the number of filters was fixed to 20.

Only static parameters were used, neither energy nor delta-parameters. Each speaker was characterized by a Markov model of one state with 32 mixtures of diagonal covariance matrix. The silence was also characterized by a Markov model, but with 3 states and only one mixture.

4.3. Experimental Results

Table 1 shows the speaker verification results in terms of equal error rate (ERR) obtained with the conventional spectral analysis FB and LP techniques and also the two hybrid approaches, FB-LP and LP-FB, both using cepstral transformation and frequency filtering of log band energies.

Several high-pass first order FIR filters have been considered: $1-0.5z^{-1}$, which equalizes the variance of mel-cepstrum in the isolated digit utterances of the adult portion of the TI database for M=Q=12, as it is used in [8]; $1-0.75z^{-1}$, which equalizes the variance of mel-cepstrum in the TIMIT database for M=Q=20, as it is used in [10]; and $1-z^{-1}$, that is inspired by the increasing tilt of the curve of the ratio between inter-speaker and global cepstrum variances mentioned in section 3.

This last filter, that is equivalent to a slope lifter [7], further reduces the computational load, since it does not need products and avoids the usual practice of subtracting the average value of the frequency sequence of log band energies [3] [8] due to its zero at zero quefrency. It just consists in subtracting the last log band energy to the current one. Furthermore, this simple filter is independent on the database.

Also has been tested the band-pass second order filter $z-z^{-1}$, that is equivalent to the so-called band-pass liftering [5]. It consists in subtracting the two log band energies adjacent to the current one and it has the same advantages than the filter $1-z^{-1}$. It seems to yield speech recognition results close to those of the optimal equalizer filter, which is data base dependent [8].

The speaker verification results in terms of EER of Table 1 have also represented graphically in Figure 1. As it can be seen, when the conventional cepstral transformation is used, the best results are obtained by using linear prediction analysis. Standard LP-cepstrum otains a 3.396 % EER, meanwhile conventional mel-cepstrum gives a 3.748 % EER. Hybrid spectral analysis yield intermediate results: FB-LP-cepstrum gives 3.476 % EER and LP-FB cepstrum gives 3.405 % EER.

Analysis	Cepstrum	$1-0.5z^{-1}$	$1-0.75z^{-1}$	$1-z^{-1}$	$z-z^{-1}$
FB	3.748	4.223	4.674	4.883	2.546
LP	3.396	3.095	2.921	2.648	2.961
FB-LP	3.476	3.044	2.684	3.045	2.706
LP-FB	3.405	4.444	4.142	4.629	3.463

Table 1: Speaker verification results in terms of equal error rate

% EER

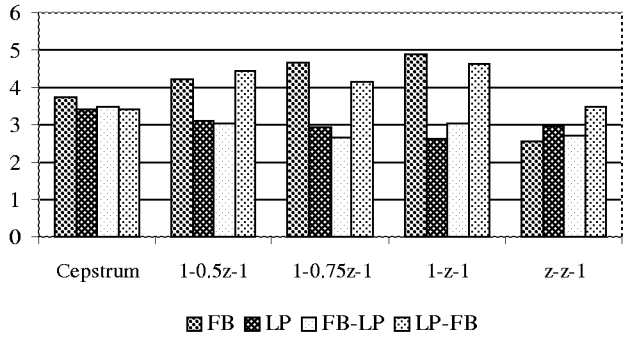

Figure 1: Graphic representation of speaker verification results.

Regarding to the use of the new frequency filtering technique, the results depend drastically on the type of spectral analysis. In the case of FB analysis, the only filter that outperform cepstral transformation is the second-order band-pass filter $z-z^{-1}$. However, using this filter a 2.546 % EER is achieved, the best results among all those obtained in this work. The relative improvement with respect mel-cepstrum is about 32 %.

In the case of LP spectral analysis, all of the first-order high-pass filters outperform cepstral transformation. The best result, a 2.648 % EER, is obtained by using $1-z^{-1}$, which is equivalent to the slope lifter and overemphasizes higher quefrencies with respect to the equalized cepstrum. In this case, the relative improvement with respect LP-cepstrum is about 22 %.

With respect to the hybrid spectral analysis, the performance of frequency filtering is quite different. In the case of LP-FB analysis, the use of frequency filtering does not outperform cepstral transformation. Regarding to the FB-LP analysis, a 2.706 % EER is obtained by using the band-pass filter $1-z^{-1}$ and a 2.684 % EER by using the high-pass filter $1-0.75z^{-1}$. In this case, the relative improvement with respect FB-LP-cepstrum is about 23 %.

5. CONCLUSION

In this paper, two ways of obtaining more robust parameters have been explored for speaker verification: the hybridization of both linear prediction (LP) and filter-bank (FB) analysis, and the frequency filtering of log band energies as an alternative to cepstrum. This combination, that had shown to be able to outperform conventional techniques in clean and noisy word recognition, has yield good text-dependent speaker verification scores on the new speaker-oriented telephone-line POLYCOST database. The best results have been obtained by using a second-order band-pass filter for FB spectral analysis and a first-order high-pass filter for LP and FB-LP (FB prior to LP) analysis.

6. ACKNOWLEDGMENTS

The authors would like to thank Javier Martín for his help in software development.

7. REFERENCES

1. Naik, J.M., "Speaker verification: A tutorial", IEEE Communications Magazine, January 1990, pp. 42-

2. Thompson, J., Mason, J.S., "Within class optimization of cepstra for speaker recogniton", Proc. EUROSPEECH'95, PP. 165-168.

3. Picone, J.W., "Signal modeling techniques in speech recognition", Proc. IEEE, Vol.81, No.9, Sept.1993, pp. 1215-47.

4. Hernando, J., Nadeu, C., "A unififed parameterization scheme for noisy speech recognition", Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-à-Mousson, France, April 1997, pp. 115-8.

5. Juang, B.H., Rabiner, L.R., Wilpon, J.G.,"On the use of bandpass liftering in speech recognition", Proc. ICASSP'86, pp. 765-8.

6. Tohkura, Y., "A weighted cepstral distance measure for speech recognition", Proc. ICASSP'86, pp. 761-4.

7. Hanson, B.A., Wakita, H., "Spectral slope based distorsion measures for all pole models of speech", Proc. ICASSP'86, pp. 757-60.

8. Nadeu, C., Hernando, J., Gorricho, M., "On the decorrelation of filter-bank energies in speech recognition", Proc. EUROSPEECH'95, pp. 1381-1384.

9. Nadeu, C., Marino, J.B., Hernando, J., Nogueiras, A., "Frequency and time-filtering of filter-bank energies for HMM speech recognition", Proc. ICSLP'96, pp. 430-433.

10. Hernando, J., Nadeu, C., "CDHMM speaker recognition by means of frequency filtering of filter-bank energies", Proc. EUROSPEECH'97, pp. 2363-2366.

11. Hernando, J., Nadeu, C., "Robust speech parameters located in the frequency domain", Proc. EUROSPEECH'97, pp. 417-420.

12. Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech", JASA, Vol. 87, No. 4, , pp. 1738-52, 1990.

13. Rahim, M.G., Juang, B.H., "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition", IEEE Trans. SAP, Vol. 4, No. 1, pp. 19-30, 1996.

14. Bimbot, F., Chollet, G., "Assesment of speaker verification systems", In.: *Spoken Resources and Assesssment, EAGLES Handbook.*