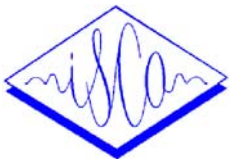

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

CDHMM SPEAKER RECOGNITION BY MEANS OF FREQUENCY FILTERING OF FILTER-BANK ENERGIES

J. Hernando and C. Nadeu
Universitat Polit cnica de Catalunya
Barcelona, Spain
javier@gps.tsc.upc.es

ABSTRACT*

Recently, the set of spectral parameters of every speech frame that result from filtering the frequency sequence of mel-scaled filter-bank energies with a simple first-order high-pass FIR filter have proved to be an efficient speech representation in terms of both speech recognition rate and computational load. In this paper, we apply the same technique to speaker recognition. Frequency filtering approximately equalizes the cepstrum variance, enhancing the oscillations of the spectral envelope curve that are most effective for discriminating between speakers. In this way, even better speaker identification results than using conventional mel-cepstrum were observed in continuous observation Gaussian density HMM, especially in noisy conditions.

1. INTRODUCTION

Cepstral coefficients C(m), $1 \le m \le M$, are the usual way of representing the short-time spectral envelope of a speech frame in current speaker recognition systems. These parameters are by far the most prevalent representations of speech signal [1] [2] and contain a high degree of speaker specificity [3]. The conventional mel-cepstrum coefficients come from a set of Q mel-scaled log filter-bank energies (LFBE) S(k), k=1,...,Q.

The sequence of cepstral coefficients C(m) is a quasi-uncorrelated and compact representation of speech spectra. In fact, in the mel-cepstrum representation, the discrete cosine transform is an approximation of the optimal Karhunen-Lo ve transform. The quefrency sequence C(m) is always windowed before entering a distance or probability computation in the pattern matching stage of the recognition process. That window eliminates the cepstral coefficients beyond a quefrency M. And, for some type of recognition systems, it also appropriately weights the remaining coefficients [3] [4] [5] [6].

However, we may wonder if the cepstral coefficients are the best way of representing the speech spectral envelope, at least for some usual recognition systems. In fact, cepstral coefficients have at least three disadvantages: 1) they do not possess a clear and useful physical meaning as LFBE have; 2) they require a linear transformation from either LFBE or the LPC 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 [7] [8], in order to try to overcome those disadvantages, the authors present an alternative to the use of cepstrum in speech recognition that consists of a simple linear processing on the LFBE domain. The transformation of the sequence S(k) to cepstral coefficients is avoided by filtering that sequence. We hereafter will call this operation frequency filtering to denote that the convolution is performed in the frequency domain.

As shown in [7], frequency filtering produces both effects, decorrelation and weighting, in only one step using a simple high-pass 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.

The aim of this paper is twofold: 1) to find an appropriate quefrency weighting for speaker recognition (section 2); and 2) to show that this discriminative quefrency weighting can be performed by filtering the sequence of LFBE (section 3). In this way, even better speaker identification results than using conventional mel-cepstrum were observed in continuous observation Gaussian density HMM, especially in noisy conditions (section 4).

2.DECORRELATION AND DISCRIMINATION

The filter-bank-based spectral estimate, implemented with the DFT (or more efficiently with the FFT), is a way to obtain a small set of parameters, the so-called filter-bank energies, that represent the speech spectrum envelope in a given frame. It actually removes pitch information and reduces estimation variance (error) by integrating the periodogram (the square value of the DFT samples) in frequency bands. And it offers the possibility of easily distributing the position of the bands in the frequency axis and defining their width in any desired way. For this purpose, a mel or a Bark scale are traditionally employed.

HMM are mostly employed with diagonal covariance matrices. In that case, they implicitly assume uncorrelated spectral parameters. That is true for the Gaussian pdf of continuous density HMM (CDHMM) and semicontinuous density HMM (SCHMM), and also

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

for the Mahalanobis distance of discrete HMM. Conversely, the frequency sequence of log filter-bank energies LFBE S(k) is strongly correlated. The usual mel-cepstrum are a way of obtaining from S(k) an almost uncorrelated set of parameters. Actually, by approximating the random process S(k) with a first-order Markov model, it follows that the discrete cosine transform is almost equivalent to the Karhunen-Lo ve transform.

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. And also because decorrelation may provide a less redundant representation. 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 [5]. Figure 1 shows an estimation of this variance for the TIMIT database [9] using Q=20 mel-scaled frequency bands. Note the zero value corresponding to zero quefrency, which is caused by the subtraction of the average S(k) value.

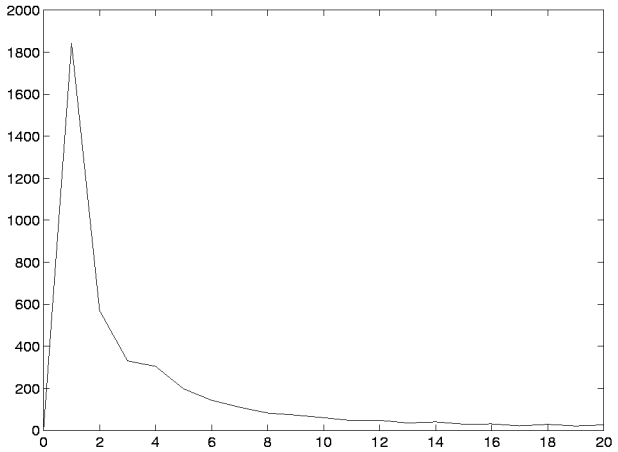

Fig. 1. Variance of the cepstral coefficients for the TIMIT database.

Thus, the low quefrencies m will generally dominate the probability or distance computations in the classifier [7]. 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 [7] [8]. Let us note that there exists a close relationship between equalization of the variance of C(m) at low quefrencies and decorrelation of S(k).

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.

A possible measure of the discrimination capacity of each cepstral coefficient can be the ratio between its inter-speaker and global variances. Figure 2 shows an estimation of this ratio for the TIMIT database using Q=20 mel-scaled frequency bands.

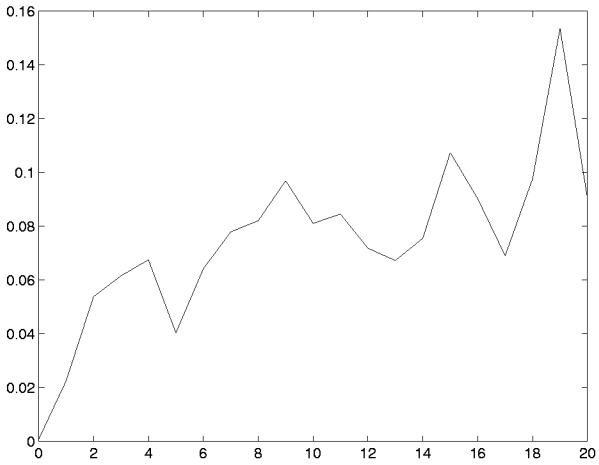

Fig. 2. Estimation of the ratio between inter-speaker and global cepstrum variances for the TIMIT database

As it can be seen in Figure 2, the dynamic range of this ratio sequence is smaller than that of the global variance shown in Figure 1. This fact suggests that an approximate equalization of the variance can help to increase the discrimination capability of the cepstral sequence, at least for clean speech.

On the other hand, Figure 2 shows a slight increasing tilt along the quefrency index m. This fact leads 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 and speaker recognition systems. In this case, two steps are needed for obtaining the final parameters from the log filter-bank energies: 1) a linear transformation (discrete cosine transform), that significantly decorrelates the sequence of parameters, and 2) a 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.

3. FREQUENCY FILTERING

We aim to perform an approximate equalization of the variance of the cepstral coefficients by filtering the frequency sequence of log filter-bank energies LFBE. Since this filtering is implemented as a circular

convolution with the sequence h(k), the cepstral coefficients are multiplied -weighted- by the DFT of h(k), here denoted by H(m).

First of all, since in the usual mel-scaled filter-bank there are not any filters centered at frequencies $\omega=0$ and $\omega=\pi$, a zero is appended at both ends of the sequence, i.e. $S(0)=S(Q+1)=0$, to represent the low energy contained at those extreme bands. Then, according to the usual practice [2], in every frame, the average value of the even sequence S(k) over index k is subtracted.

After that, S(k) is circularly convoluted with h(k) to obtain a filtered sequence. Since only the values of the filtered sequence between k=1 and k=Q are used as observations in the recognition system, we can employ the shortest h(k), i.e. a length 2, with no interference of the symmetric S(k), k=-1,...,-Q, samples in the computation of the used segment of the filtered sequence. In this way, we can refer to the process as an actual linear filtering, with h(k) being the impulse response.

A first-order FIR filter that maximally equalizes the variance of the cepstral coefficients can be easily obtained by a least-squares modeling in the following way. Firstly, the variance is estimated by averaging over all the frames of a given database. Then, after performing an inverse DFT, the quotient r between the values of the resulting sequence –the variance of S(k)– at index 1 and index 0 is computed. Thus, the first-order FIR filter that maximally flattens the variance will be $H(z)=1-rz^{-1}$.

Figure 3 shows the product of the cepstrum global variance corresponding to the TIMIT database [9] using Q=20 mel-scaled frequency bands -shown in Figure 1- by the magnitude of the sampled filter response H(m), that was computed following the above procedure. The resulting value of r is 0.75. As it can be seen, the cepstrum variance tilt has been approximately equalized.

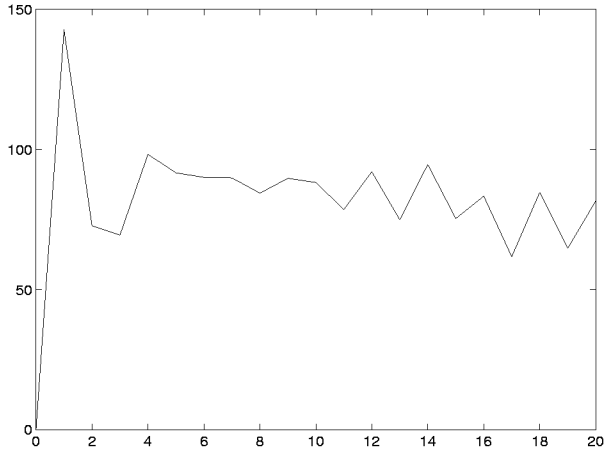

Fig. 3. Cepstrum variance equalized by the sampled filter response H(m).

However, the filtered LFBE representation (FLFBE) may improve its performance if this frequency filter is empirically optimized, perhaps taking into account the slight increasing tilt along the axis m of the ratio between inter-speaker and global cepstrum variances shown in Figure 2.

Finally, it is worth noting the computational simplicity of filtering with respect to the mel-cepstrum representation. A way of further reducing the computations is to use the filter $z-z^{-1}$, since it does not need products and avoids the average subtraction due to its zero at zero quefrency. It consists in subtracting the two LFBE of the bands adjacent to the current one. That simple filter does not depend on the database and it seems to yield speech recognition results close to those of the optimal filter, which is data base dependent [7].

4. RECOGNITION EXPERIMENTS

We carried out speaker recognition experiments in both clean and noisy conditions by filtering the average-subtracted frequency sequence of log filter-bank energies LFBE in several ways, and using the filtered sequence as the speech representation, with no addition of supplementary differential features. A speaker recognition system based on continuous-density HMM was used.

The TIMIT [9] database was used in our experiments. 200 speakers (100 male and 100 female) were selected. Clean speech was used for training in all the experiments. Noisy speech for testing was simulated by adding zero mean white Gaussian noise to the clean signal so that the SNR of the resulting signal becomes 20 dB.

The HTK software, based on the Continuous-Density Hidden Markov Models (CDHMM), was modified to perform speaker recognition experiments with the novel speech representation. In the parameterization stage, after pre-emphasizing the signals with a zero at z=0.95, Hamming windowed frames of 25 ms were taken every 10 ms. Each frame was represented by M=20 parameters, derived from a bank of Q=20 mel-scaled filters. Each speaker was characterized by a Markov model of one state with 32 mixtures with diagonal covariance matrices. The silence was also characterized by a Markov model, but with 3 states and only one mixture. For each speaker, the model was trained with 5 TIMIT sentences. The other 5 TIMIT sentences were used separately as test signals.

Table 1 shows the speaker identification rates (ID) in clean and noisy conditions obtained with the conventional mel-cepstrum coefficients (MFCC) representation along with the ones obtained with the filtered log filter-bank energies (FLFBE) using several high-pass first order FIR filters: $1-0.75z^{-1}$, which equalizes the TIMIT database for M=Q=20, as it is used in this work; and $1-0.8z^{-1}$, $1-0.9z^{-1}$ and $1-z^{-1}$, that are inspired by the increasing tilt of the curve of the ratio between inter-speaker and global cepstrum variances in Figure 2.

Parameters / ID	clean	20 dB
MFCC	98.1	32.4
FLFBE ($1-0.75z^{-1}$)	98.3	46.1
FLFBE ($1-0.8z^{-1}$)	98.5	52.8
FLFBE ($1-0.9z^{-1}$)	98.4	61.8
FLFBE ($1-z^{-1}$)	98.3	64.4

Table 1. Speaker identification rates

It can be seen in Table 1 that the new FLFBE parameterization is competitive with conventional mel-cepstrum representation in clean conditions. When the optimal equalizer for the TIMIT database $1-0.75z^{-1}$ is used, FLFBE outperforms conventional mel-cepstrum. However, the best results are obtained with the filter $1-0.8z^{-1}$, which slightly overemphasizes higher quefrencies with respect to the equalized cepstrum.

The simple database-independent filter $z-z^{-1}$, that yielded results close to those of the optimum filter in clean speech recognition [7], has not provided so good results in speaker recognition, 97.8 % identification rate. It is due to the band-pass characteristics of this filter. Actually, a high-pass filter, like the ones considered in the Table 1, is more convenient in order to properly emphasize higher quefrencies.

Regarding to noisy conditions, excellent results have been obtained by using the new FLFBE approach. Using the optimum equalizer $1-0.75z^{-1}$, there is an identification error rate reduction of almost 30 % respect to conventional mel-cepstrum. The results are even better by using filters that put more emphasis on higher quefrencies. Setting the zero at $z=1$, there is an identification error rate reduction of almost 50 %. Filters with zero close to 1 are more convenient in the presence of broad-band noise due to the fact that cepstral parameters of lower index are globally more affected by this type of noise than higher order ones.

5. CONCLUSION

We have explored a new speech representation in speaker recognition that consists in filtering the frequency sequence of mel-scaled filter-bank energies with a simple high-pass first-order FIR filter. For clean speech the empirically optimum zero of the filter is very close to the one resulting from flattening the cepstrum variance by linear prediction. In this case, the set of parameters of a given frame is the sequence obtained at the otput of the optimum first-order prediction error filter driven by the filter-bank energies sequence. For noisy speech, it is preferable to use a zero closer to 1. Actually, the best results in our experiments correspond to a zero of value 1 and offer almost 50 % error rate reduction with respect to mel-cepstrum.

6. ACKNOWLEDGMENTS

The authors would like to thank Jordi Mu oz and Javier Mart n for their help in software development.

7. REFERENCES

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

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

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

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

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

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

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

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

[9] W. Fisher, V. Zue, J. Bernstein, D. Pallet, "An Acoustic-Phonetic Data Base", J. Acoust. Soc. Amer. Suppl. (A), 81, S92, 1986.