On the Decorrelation of Filter-Bank Energies in Speech Recognition

Climent NADEU, Javier HERNANDO and Mònica GORRICHO

Universitat Politècnica de Catalunya
Barcelona, Spain
E-mail: nadeu@tsc.upc.es

ABSTRACT

Cepstral coefficients are widely used in speech recognition. In this paper, we claim that they are not the best way of representing the spectral envelope, at least for some usual speech recognition systems. In fact, cepstrum has several disadvantages: poor physical meaning, need of transformation, and low capacity of adaptation to some recognition systems. In this paper, we propose a new representation that significantly outperforms both mel-cepstrum and LPC-cepstrum techniques in both recognition rate and computational cost. It consists of filtering the frequency sequence of filter-bank energies with an extremely simple filter that equalizes the variance of the cepstral coefficients. Excellent results of the new technique using a continuous observation density HMM recognition system and two very different recognition tasks, connected digits and phone recognition, are presented.

1. Introduction

In speech recognition, the short-time spectral envelope of every speech frame is often represented by a set of *M* cepstral coefficients *C(m), 1=m=M*, which are the Fourier series coefficients of its logarithm. These cepstral coefficients usually come either from a set of *Q* mel-scale log filter-bank energies (FBE) *S(k), k=1,...,Q* –*mel-cepstrum coefficients* (MCC) representation– or from a linear prediction analysis –*LPC-cepstrum* representation [1].

The sequence of cepstral coefficients *C(m)* is a quasi-uncorrelated and compact representation of speech spectra. In fact, in the MCC representation, the discrete cosine transform is an approximation of the optimal Karhunen-Loéve transform, and therefore it approximately decorrelates the frequency sequence *S(k), k=1,...,Q*. Thus, the lowest quefrency (index *m*) terms are those with the highest variance, a fact that provides a compact representation.

Actually, 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 speech recognition systems and for LPC-cepstrum, it also weights the remaining coefficients in order to approximately equalize their variance or, inseparably, deemphasize the low quefrency coefficients [2-4].

However, we may wonder if the cepstral coefficients are the best way of representing the speech spectral envelope, at least for some usual speech recognition systems. In fact, cepstral coefficients have at least three disadvantages: 1) they do not possess a clear and useful physical meaning as FBE have; 2) they require a linear transformation from either the log FBE 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 this work, in order to try to overcome those disadvantages, we present an alternative to the use of cepstrum that consists of a simple linear processing on the log FBE domain. Our approach is able to improve the speech recognition performance of the mel-cepstrum representation and the LPC-cepstrum representation by filtering the frequency sequence of log FBE to equalize the variance of the cepstral coefficients. Actually, a simple high-pass first order FIR filter suffices to obtain a significant improvement of the recognition rate. Inseparably from the equalization effect, the filter also produces a certain decorrelation of the log FBE. Moreover, the output of such a derivative-type filter actually is a spectral slope measure and, according to Klatt [5], the spectral slope is a perceptually important characteristic for phonetic distance.

2. Equalization of the Variance of the Cepstral Coefficients

Unless otherwise indicated, the sequence of *Q* mel-scale DFT-based log spectral energies [1] is used as the baseline speech spectral representation in this work. In continous observation Gaussian density HMM (CDHMM), using one mixture with diagonal covariance matrix per state, the log probability that a given observation vector **S**, whose components are the *Q* log FBE *S(k), k=1,...,Q*, has been generated by a given state *q* is

$$\log p(\mathbf{S}/q) = -\frac{1}{2}\sum_{k=1}^{Q}\log 2\pi s^2(k) - \frac{1}{2}\sum_{k=1}^{Q}\left|\frac{S(k)-m(k)}{s(k)}\right|^2 \quad (1)$$

where $m(k)$ and $s^2(k)$ are, respectively, the mean and variance of the *k*-th spectral parameter in the state *q*.

This work has been funded by the Spanish Government project TIC92-1026-C02-02

Note that, given the state q, the first term in (1) is constant, and thereby we will only consider the last term, which depends on the frequency sequence $S(k)$. To facilitate the reasoning, we will assume the same variance for all states (grand variance). In this way, the variance sequence is estimated over all the data, and we will consider it as constant, i.e. $s^2(k)=s^2$, which makes sense due to the fact that a constant value can always be obtained by a proper signal preemphasis.

In the following, we will express the last term in (1) in terms of the cepstral sequence $C(m)$ corresponding to $S(k)$. Since in the usual mel-scale filter-bank distribution there are not any filters centered at frequencies $w=0$ and $w=p$, 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. As the log spectrum is an even (and periodic) function, we can write (1) in terms of the even sequence $S(k)$, $k=-Q,...,0,...,Q+1$, where $S(-k)=S(k)$, $k=1,...,Q$. Thus, the last term in (1) is proportional to

$$\sum_{k=-Q}^{Q+1} |S(k) - \mu(k)|^2 \qquad (2)$$

Then, by applying the Parseval relation [6], it follows that that term is also proportional to

$$\sum_{m=-Q}^{Q+1} |C(m) - M(m)|^2 \qquad (3)$$

where the even cepstral sequences $C(m)$ and $M(m)$, for $m=-Q,...,0,...,Q+1$ are, respectively, the discrete Fourier transforms of the even frequency sequences $S(k)$ and $\mu(k)$.

Expression (3) shows that, although spectral-type observations are used in our HMM framework, the probability can be computed from the cepstral coefficients. Since $M(m)$ is also the mean of $C(m)$ in the state q, every cepstral coefficient $C(m)$ contributes in an additive way to the probability according to its square distance to the mean value in the state, exactly like $S(k)$ in (2). However, although the grand variance has been equalized in (2), it is not so in (3). Hence, the cepstral coefficients can be weighted in order to equalize their variance, with the purpose of obtaining an even contribution of them to the probability computation. Note the coincidence of that conclusion with the cepstral weighting studied in [2-4]. In fact, a Euclidean distance on the cepstral coefficients was assumed in those works, and expression (3) actually is a Euclidean distance.

On the other hand, the equalization of the cepstral variance produces a certain decorrelation of the log FBE. In fact, it can easily be shown that the cepstral variance of a non-symmetric uncorrelated log FBE sequence $S(k)$, $k=1,...,Q$, is flat.

The variance of the cepstral coefficients has a decreasing tilt along the quefrency axis [3]. For this reason, the number Q of frequency bands has to be accurately chosen, since too large a value would imply the existence of high quefrencies which would be strongly amplified by the inverse variance weighting. As those high cepstral indexes carry much spectral estimation error [2], the recognition performance would worsen. Hence, in the case of a large Q value, the equalization should take place only in the low quefrency region.

The conclusions drawn in the previous paragraphs for CDHMM are also valid for discrete HMM and for any other speech recognition system that uses a Euclidean-type metric to incorporate the observations into the recognition process.

3. Filtering of the Frequency Sequence of Filter-Bank Energies

We aim to perform the equalization of the variance of the cepstral coefficients by filtering the frequency sequence of log FBE. Since the filtering is implemented as a circular convolution with the sequence $h(k)$, the cepstral coefficients are multiplied by the DFT of $h(k)$, here denoted by $H(m)$, so that expression (3) turns out to be

$$\sum_{m=-Q}^{Q+1} |C(m) - M(m)|^2 |H(m)|^2 \qquad (4)$$

First of all, according to the usual practice [1], in every frame, the average value of the even sequence $S(k)$ over index k is subtracted, so the term in (3) corresponding to the zero quefrency is removed whereas the other terms remain unchanged. After that, $S(k)$ is circularly convolved 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. The same is true if length 3 is used and $h(k)$ is centered around $k=0$. 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 covariance of $S(k)$– at index 1 and index 0 is computed. Thus, the first-order FIR filter that maximally flattens the variance will be

$$H(z) = 1 - rz^{-1} \qquad (5)$$

Figure 1 shows the estimated variance corresponding to the TI digits database [7] using $Q=12$ mel-scale frequency bands, along with the inverse square magnitude of the sampled filter response $H(m)$, that was computed following the above procedure. The resulting value of r is 0.5. Analogously, the coefficients of the least-squares second-order FIR filter are -0.5 and -0.05, a fact that shows how a first-order filter already obtains an accurate modelling of the inverse variance. Note in Figure 1 the zero value of the zero quefrency variance, which is caused by the subtraction of the average $S(k)$ value.

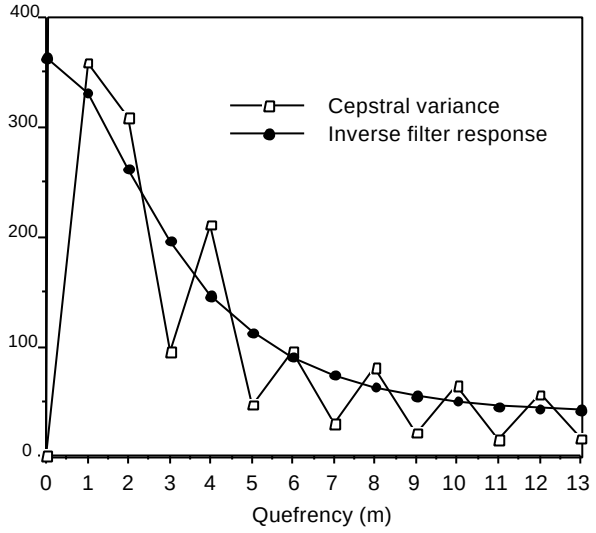

Fig. 1. Approximation of the TI digits estimated variance with the inverse square magnitude of the filter $1-0.5z^{-1}$

It is worth noting the computational simplicity of the filtering with respect to the DCT computation of the mel-cepstrum representation. In fact, average subtraction plus first-order filtering requires *3Q* additions and *Q+1* multiplications, whereas, assuming *M* cepstral coefficients are used, DCT requires *MQ* additions and multiplications, which, for typical values of *Q* and *M* , can be an order of magnitude higher. A way of further reducing the computations is to use the filter $z-z^{-1}$ which only requires *Q-2* subtractions, since it does not need multiplications and avoids the average subtraction due to its zero at zero quefrency.

Note that the cepstral coefficients can also be replaced by filtered spectral energies in the case of LPC spectra. Recognition results will be reported in the next section based on both filter-bank and LPC spectral estimates.

4. Recognition Experiments

We carried out speech recognition experiments by filtering the average-subtracted frequency sequence of log FBE in several ways, and using the filtered sequence as the speech representation, with no addition of supplementary differential features. A speech recognition system based on continous observation density hidden Markov models (CDHMM) was used (HTK software). The experiments correspond to two different databases: the TI connected digits database [7] and the EUROM1 [8] phonetic database.

4.1 Connected Digits Database

Firstly, training and testing were carried out with the single and connected digit utterances of the adult portion of the TI database. After decimating the signals from 20 KHz to 8 KHz sampling rate and pre-emphasizing them with a zero at z=0.95, Hamming windowed frames of 30 ms were taken every 10 ms. Each of the 11 digit left-to-right hidden Markov models consisted of 8 states, and the silence model had 3 states. Only one diagonal covariance mixture was employed per state.

DFT-based mel-scale FBE

First of all, the new filtered log FBE sequence has been compared with the MCC sequence. After trying several values, 20 frequency bands (*Q*=20) and 8 cepstral coefficients (*M*=8) were chosen as the empirically optimal parameters for MCC. Table 1 shows the MCC recognition results along with the ones obtained with the first-order and the second-order equalization filters proposed in the previous section. The energies corresponding to 12 mel-scaled frequency bands (*Q*=12) were used for every filter.

	String	Word	Del	Subs	Ins
MCC	22.59	8.09	3.63	4.46	1.07
order 1	18.02	5.79	1.98	3.81	1.27
order 2	18.08	5.81	2.01	3.80	1.30

Table1 Percentage of recognition errors for DFT-based mel-scale FBE.

Note, in Table 1, the significant improvement achieved by the filtered log FBE with respect to conventional MCC: 20% in string error rate, and 28% in word error rate. The second order equalization yields almost exactly the same rates as those of the first order filter, since the second coefficient is very small.

In the reported experiments, the new spectral representation technique requires more features per frame than MCC (12 instead of 8). In order to check, in these preliminary results, the need to have a larger number of features, we performed an experiment with the first-order equalization filter and *Q*=10. The resulting string and word error rates were, respectively, 18.07% 6.03%, only slightly larger than the ones obtained for *Q*=12.

Applying the Karhunen-Loéve transform to the average-subtracted FBE in order to globally decorrelate them, 20.60% string error rate and 7.49% word error rate were obtained, scores worse than those of the filtered log FBE. Consequently, although our HMMs assume uncorrelated features, the important fact appears to be the particular type of probability measure (1-3) that arises from this assumption. Also experiments using full covariance matrices in the Gaussian densities were carried out. The resulting string error rate is 13.30% and the word error rate is 4.42%, results better than those of the filtered log FBE with diagonal matrices. Hence, the state decorrelation achieved by using full matrices appears preferable to a global one. However, the computational load is substantially enlarged.

Our reasoning in Section 2 assumes only one Gaussian mixture per state. For this reason, we performed an experiment using 8 mixtures. The results are given in Table 2. Note that there is also a significant relative improvement like for one mixture, and therefore we can hope than the conclusions drawn from that reasoning are also valid for multiple mixtures.

	String	Word	Del	Subs	Ins
MCC	15.57	5.26	1.75	3.51	0.89
order 1	13.08	3.94	1.29	2.65	1.02

Table2. Recognition error rates using 8 mixtures.

LPC-based FBE

LPC-based experiments were also carried out. After computing 13 cepstral coefficients *C(m), m=0,...,12*, from a 10th order LPC analysis, they were transformed to the spectral domain using a 24 point DFT. The obtained 13 values were considered as log FBE, and thereby their average value was subtracted from them and they were filtered as it was done with the mel-scale DFT-based FBE. This procedure does not have a practical interest but the results can give more support to the basic principle of our technique. The value of the zero of the first-order filter is, in this case, 0.53, quite similar to that of the previous DFT-based case.

LPC	String	Word	Del	Subs	Ins
LPC-ceps	24.03	7.76	2.47	5.29	2.10
order 1	19.71	6.92	2.31	4.61	0.99
($z-z^{-1}$)	19.67	6.90	2.21	4.60	1.08

Table3. LPC-based recognition error rates.

Table 3 shows the recognition results for the conventional LPC-cepstrum representation and two filtered LPC-based log FBE. Even though these results are not so good as those of Table 1, we observe that filtering improves again the recognition performance with respect to cepstrum. Additionally, the second-order filter $z-z^{-1}$ achieves almost the same recognition performance as the first-order equalizing filter. It can be shown that that extremely simple second-order filter is equivalent in the quefrency domain to the weighting proposed in [2].

4.2 Phonetic Database

Tests were also carried out with the EUROM1 phonetic Spanish database [8]. 842 utterances from 186 different phonetically balanced sentences and using 42 speakers were used for training, and 225 utterances from 61 sentences and 17 speakers for testing. Speakers are balanced by gender. The sampling rate is 16 KHz. Hamming windowed frames of 25 ms were taken every 10 ms. A CDHMM speech recognition system like that of the last section was used. However, three diagonal covariance Gaussian mixtures were employed per state in this case, and each of the 33 left-to-right phone models consisted of 3 states.

First of all, the new filtered log FBE representation was compared with the MCC one. After trying several values, 20 frequency bands (Q=20) and 12 cepstral coefficients (M=12) were chosen as the empirically optimal parameters for MCC.

	Accur	Phone	Del	Subs	Ins
MCC	16.77	51.98	13.43	34.59	35.20
order 1	43.90	56.89	13.47	29.64	12.99
order 2	44.14	57.08	13.19	29.73	12.94
($z-z^{-1}$)	41.30	55.33	13.13	31.54	14.03

Table 4 Percentage of recognition rates and error rates for the EUROM1 database.

Table 1 shows the MCC recognition results in terms of phone recognition accuracy, correct phone recognition and error rates, along with the ones obtained with three different filters: 1) the first-order equalization filter, whose zero is, in this case, z=0.26; 2) the second order equalization filter whose coefficients are -0.23 and -0.13; and 3) the second-order filter $z-z^{-1}$. The energies corresponding to 16 frequency bands (Q=16) were used for every filter as a logical extension to 16 KHz of the mel scale from Q=12 for 8 KHz.

The filtered FBE improve over MCC for all the performance rates, except for the deletion rate which is similar. The low accuracy of MCC in the EUROM1 database is due to its high number of insertions. The second-order filter slightly improves the first-order one, and the filter $z-z^{-1}$ produces again a remarkable improvement, with results close to those of the equalization filters.

5. Concluding Remarks

We have proposed a new parameterization techni-que that outperforms the almost universally employed cepstrum representations in both recognition rate and computational cost by filtering the frequency sequence of filter-bank energies. A first-order FIR filter suffices to equalize the variance of the cepstral coefficients, and it is able to obtain noticeable better recognition results than the mel-cepstrum and the LPC-cepstrum speech representations. Even the computationally inexpensive $z-z^{-1}$ filter achieves remarkable recognition results. Note that the coefficients of this filter are not computed from the cepstral variance, so the filter has not to be adapted to the current database.

A second-order filter may become necessary if the number of bands is excessively large since, in that case, the high quefrency coefficients carry a large amount of estimation error, which can be attenuated with a second zero close to z=-1. On the other hand, if the best performance did not correspond to a completely flat variance, a second-order FIR filter could also be employed. One real zero would be used to deemphasize the lower quefrency components, whereas the other zero would deemphasize the equalized higher quefrency ones.

References

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

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

[3] Y. Tohkura, “A weighted cepstral distance measure for speech recognition”, *Proc. ICASSP'86*, pp.761-4.

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

[5] D.H. Klatt, “Prediction of perceived phonetic distance from critical band spectra: A first step”, *Proc. ICASSP'82*, pp.1278-81.

[6] A.V. Oppenheim, R.W. Schafer, *Discrete-Time Signal Processing*, Prentice-Hall, 1989.

[7] R.G. Leonard, "A database for speaker-independent digit recognition", *Proc. ICASSP'84*, pp. 42.11.1-4.

[8] ESPRIT project: Speech Technology Assessment in Multilingual Applications (SAM-A). Document SAM-A/6002, 1993.