FREQUENCY AND TIME FILTERING OF FILTER-BANK ENERGIES FOR HMM SPEECH RECOGNITION

Climent Nadeu, José B. Mariño, Javier Hernando and Albino Nogueiras[*]

Universitat Politècnica de Catalunya
Barcelona, Spain
climent@gps.tsc.upc.es

ABSTRACT

In speech recognition, a discriminative quefrency weighting can be achieved by somewhat decorrelating the frequency sequence of log mel-scaled filter-bank energies with a computationally inexpensive filter. In this paper, we show how the spectral parameters that result from this kind of frequency filtering, both alone and combined with filtering of their time trajectories, are competitive with respect to the conventional cepstral representations of speech signals.

1. Introduction

An usual way to spectrally represent a speech frame in speech recognition consists of computing –through the DFT– the mel-scaled log filter-bank energies (logFBE) [1], i.e. the spectral sequence $S(k)$, where $k=1,...Q$ is the index of the frequency band. After that spectral estimation step, a discrete cosine transformation (DCT) to (real) cepstral coefficients is often performed to obtain a cepstral sequence $c(m)$, the mel-cepstrum, where m denotes the quefrency index.

On the other hand, by taking into account a particular frequency band, a time sequence of the logFBE of that band is obtained. Thus, we can consider a two-dimensional spectral-time sequence of log energies $S(k,n)$ in which n denotes the frame index.

Given a DFT-based spectral estimator that obtains the FBEs (or given the other conventional alternative, the LPC estimator), there is a way of taking advantage of the usual pattern matching approaches for speech recognition, e.g. the hidden Markov model (HMM). It consists of carrying out a cepstral weighting (liftering) and/or using some kind of differential (delta) parameters instead of, or along with, the basic spectral parameters. Both operations account for the convenience of de-emphasizing slow changes in the frequency and the time sequences.

In recent papers [2-4], we interpreted both operations as filtering $S(k,n)$ in the two dimensions in a separate way. The transformation of the sequence $S(k,n)$ to cepstral coefficients was avoided by performing, for every speech frame, a filtering of that frequency sequence [3], which we hereafter will *call frequency filtering* to distinguish it from the *time filtering* involved in the computation of the differential parameters.

If the conventional mel-cepstrum coefficient (MCC) and linear prediction cepstral coefficient (LPCC) parameterizations are liftered before entering the classifier, two steps are needed for obtaining the final parameters from the logFBEs: 1) a transformation, that significantly decorrelates the sequence of parameters, and 2) a weighting of the cepstral coeficients. As shown in [3], frequency filtering produces both effects in only one step and using an extremely simple filter. In this way, even better recognition results than using MCC and LPCC were observed in continous density HMM (CDHMM).

In this paper, we present both time and frequency filters in a parallel way, within a 2-D formalism, as a further attempt to better understand why they are useful in speech recognition. And some experimental tests that combine both filters in continous, semicontinous and discrete HMM are also reported.

2. Spectral-time representation of the speech signal

The filter-bank-based spectral estimate, implemented with the DFT (or more efficiently with the FFT), is an excellent way to obtain a small set of parameters, the so-called filter-bank energies (FBE), that represent the speech spectrum 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.

The two-dimensional spectral-time sequence of logFBEs $S(k,n)$ carries all the information about the current speech signal that is given to the classifier. That is also true for its 2-D Fourier counterpart $C(m,\theta)$. The transformation from the frequency domain k to the quefrency domain m is performed by the DFT, and the transformation from the time domain n to the modulation frequency domain θ is represented by a Fourier transform (FT). Hereafter, we will denote the pair quefrency index – modulation frequency variable by QMF.

$$S(k,n) \xrightarrow{\text{inverse DFT}} c(m,n) \xrightarrow{\text{FT}} C(m,\theta)$$

[*] This work has been funded by the Spanish government projects TIC95-1022-C05-03 and TIC95-0884-C04-02.

3. Decorrelation vs discrimination

HMMs assume i.i.d. signal observations, but the time sequences of spectral parameters are not so. Differential parameters are usually presented as a way of coping with that limitation of the models by including the rate of temporal change in each observation. Moreover, they actually are less correlated themselves between frames than the basic (non-differential) spectral parameters since they usually show a wider MF band [2].

HMMs are mostly employed with diagonal covariance matrices. In that case, they implicitly assume uncorrelated spectral parameters. That is true for the Gaussian pdf of continous density HMM (CDHMM) and semicontinous density HMM (SCHMM), and also for the Mahalanobis distance of discrete HMM. Conversely, the frequency sequence of logFBE $S(k)$ is strongly correlated. For instance, in the TI digits data base, the correlation between logFBE of adjacent spectral bands is 0.92 for 12 bands. The usual MCC 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 DCT is almost equivalent to the discrete Karhunen-Loéve transform, since the value of the real pole of the model is close to 1 (0.92 for the above mentioned data base) [5], so the cepstral coefficients are almost globally uncorrelated.

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.

If $S(k,n)$ is considered a stationary process in n, its mean only depends on k. In the following, we will assume that this mean equals zero in order to avoid distinction between variance and mean square magnitude. Actually, if the mean of $S(k,n)$ is zero for all k and n, the mean of $C(m,\theta)$ is also zero for all m and θ. Thus, the variance of $C(m,\theta)$ is

$$T(m,\theta) = E\left\{|C(m,\theta)|^2\right\} \qquad (1)$$

It is a known fact that $T(m,\theta)$ decreases along the axis m [6] and also along the axis θ [2]. Thus, the low quefrencies m and modulation frequencies θ will generally dominate the probability or distance computations in the classifier. Therefore, we may ask whether this is the best we can do or a proper global weighting of $C(m,\theta)$ could help to increase discrimination and therefore recognition performance.

Let us note that there exists a close relationship between compensation of the relatively high values of $T(m,\theta)$ at low QMF and decorrelation of $S(k,n)$. In fact, if the 2-D sequence $S(k,n)$ were driven to the classifier after it had been approximately decorrelated, the corresponding QMF function $T(m,\theta)$ would be rather flat in both dimensions.

However, a flat variance $T(m,\theta)$ may not be the most adequate goal for recognition purposes. For example, when the frequency interval between bands or the time interval between frames are not large enough, that equalization gives too much weight to the estimation noise carried out by $C(m,\theta)$ for high QMF. Another reason for not flattening it completely can be the presence of the acoustic channel and the speaker characteristics in the lowest QMF.

4. Weighting vs filtering

As the lowest QMF indices carry the slowest oscillations of the 2-D sequence $S(k,n)$, they represent the spectral tilt (in k) and the long-term changes (in n). However, the most discriminative information is located in the alternation of peaks and valleys in the spectral sequence and in the alternation of stationary and transitional portions of the time sequence. And both are represented by higher QMF.

Cepstral liftering (weighting on m) and differential parameters (weighting on θ) have been the usual ways to compensate for the excessive weight of the lowest m and θ terms in the probability computation. In fact, both relatively emphasize medium QMF.

Liftering, as it is employed in the LPCC representation, performs a sort of inverse variance weighting of low quefrencies which mainly enhances the quefrencies between m=6 and m=8 [7]. Notice that these values correspond to oscillations of the spectral sequence that show from 3 to 4 peaks up to 4 KHz. Thus, these quefrencies account for the fluctuations in the $S(k)$ curve produced by the formant peaks and the between-formant valleys.

Differential parameters result from filtering the 2-D sequence along the time dimension. As explained in [2], when several of these parameters are used in speech recognition supplementing the basic spectral parameter, their modulation frequency bands are distributed along with that of the basic spectral parameter in the first interval of the θ axis, in such a way that the composite spectrum is rather flat in that interval. Additionally, when only a filtered parameter is used, it has been shown for clean speech that a proper enhancement of the modulation frequency band around 3 Hz, the typical syllabe rate, may noticeably improve recognition performance [4]. This result suggests that the most relevant oscillation has one peak per syllable, which is quite sensible.

Thus, filtering in time n means weighting in its Fourier transform domain θ. And weighting in m by $H(m)$ means filtering in its DFT domain k, where $h(k)$, the DFT of $H(m)$ is the impulse response of the filter. By choosing $H(m)$ to somewhat equalize the variance of $c(m,n)$, two beneficial effects are obtained: 1) the filtered frequency sequence is decorrelated to some extent, and 2) the low quefrencies are de-weighted as in the common liftering.

That frequency filtering is the approach considered in [3], where the new parameter sequence obtained by using a simple filter of order one or two is shown even to outperform both the current mel-cepstrum and LPC-cepstrum coefficients for CDHMM and for two different data bases. Actually, cepstral weighting has no effect on the recognition results of CDHMM with diagonal covariance matrix due to the intrinsic variance normalization of the Gaussian pdf, whereas

frequency filtering is able to produce a cepstral weighting in an implicit way.

As shown in Fig.1, there are four ways of implementing 2-D filtering in the original domain and 2-D weighting in the QMF domain. The most usual way consists of filtering in time *n* and weighting in quefrency *m*, i.e. differential or dynamic features and liftering (the shaded cell in Fig.1). Due to the advantages of frequency filtering, we propose to avoid cepstrum computation and to perform filtering in both *n* and *k*. An opposite alternative has been reported in the literature [8,9]. It uses as spectral representation the 2-D sequence *C(m,i)*, where *i* is an index that results from discretizing the variable θ. The two reported approaches sample the modulation frequency variable, since the authors either work with the 2-D sequence of the utterance as a whole [8] or consider 2-D blocs with a fixed number of frames [9]. In [8], *C(m,i)* does not come from the FBE sequence but directly from the DFT samples, without integration. However, it is also weighted along both *m* and i.

	Time *n*	Mod. freq. θ
Freq. band *k*	*2-D filtering*	*Filtering (k)* *Weighting (q)*
Quefrency *m*	*Weighting (m)* *Filtering (n)*	*2-D weighting*

Fig.1 Four alternative ways of implementing filtering and weighting.

5. Frequency and time filtering

Filtering is separately performed in the dimensions *k* and *n*. However, the effect of filtering in both dimensions is remarkedly similar. In fact, we define both the frequency filter and the time filter as the combination of three operations: 1) removal (or high attenuation) of the average value; 2) approximate variance equalization in the transform domain (quefrency for *k*, or modulation frequency for *n*), with a first-order high-pass FIR filter; and 3) smoothing with a low-pass filter that shapes the (equalized) band.

In the time domain, the three operations are achieved with the classical regression or Lagrange filter and also with the Slepian-equalizer filter proposed in [2]. The conventional cepstral (or logFBE) mean subtraction (CMS) also is a sort of filtering [4], but it only carries out the first operation in order to cancel linear distortion.

In the frequency domain, we will use the most simple filter proposed in [3], i.e.

$$H(z)=z-z^{-1} \qquad (2)$$

It consists of a subtraction of the two FBE of the bands adjacent to the current one. That extremely simple filter does not depend on the data base and it seems to yield results close to the optimum which is data base dependent. We will not take a large *Q* because otherwise the high quefrencies should be de-emphasized and more parameters would be used. Conversely, in the θ domain there exists a large interval at high modulation frequencies that does not contain speech information.

6. Recognition experiments

The 2-D filtering approach has been tested with CDHMM (using HTK software), SCHMM and DHMM, and using various data bases.

CDHMM and clean digit data base

Firstly, training and testing were carried out with the TI connected digit database. The experimental setup was already reported in [3]. As there, the new parameterization, that we will denote by FLFBE (filtered logFBE), is compared with the MCC representation but here time filtering is also included. 12 basic spectral parameters and no energy were employed. As mentioned before, the frequency filter was $z-z^{-1}$, and the time filter consisted of the same Slepian-equalizer filter used in [4] for LPCC representation. Fig.2 shows the string recognition percentages. Substitution means that only the filtered parameter is used (one feature), and supplementation means that both the non-filtered parameter and the filtered one are used together (two features). The good performance of FLFBE already observed in [3] is kept when time filtering is considered.

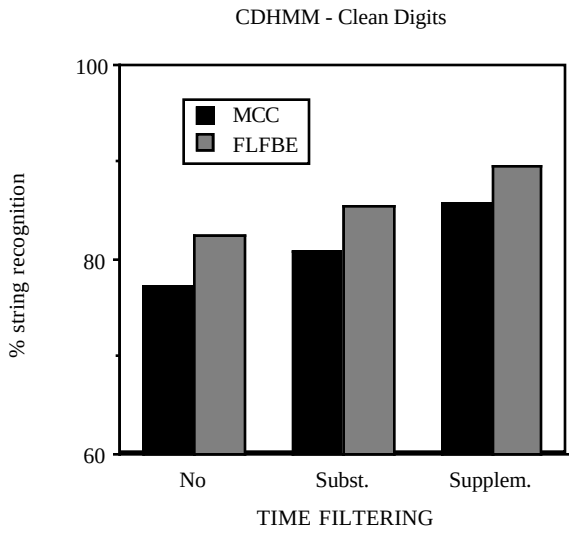

Fig.2 String recognition percentages.

When the (non-normalized) energy is included as an additional parameter and no time filtering is employed, MCC increases string recognition 1.8% whereas FLFBE decreases the same amount. For the supplementation case, MCC increases 3.75% and FLFBE increases less than 0.1%. In both cases, FLFBE still performs better than MCC, but the differences are much smaller than when the energy parameter is not considered. These results suggest that the FLFBE parameterization somehow includes the energy information through both ending points of the filtered frequency sequence, since their values actually are the log energies of the second and the next to last (with a minus sign) bands.

It is worth noting that the FLFBE representation can improve its performance if the frequency filter is empirically optimized. Recognition percentage increased 1% using the

filter $(1-0.7z^{-1})(1+0.3z)$. That filter attenuates low and high quefrencies less than the above filter $z-z^{-1}=(1-z^{-1})(1+z)$ so that it has a response closer to that of the optimal equalizer obtained in [3] with linear prediction, i.e. $1-0.5z^{-1}$.

SCHMM and subword modeling

TELEMACO [10] is a system for automatic voice dialling based on the recognition of commands in fluent speech. The set of commands is composed by the digits and fifteen dialling words (call, answer, transfer, number, etc.). The system decodes each utterance in commands and fillers. An order can require from one to nine commands. The command models are built with context dependent phone models and the fillers are models of group of syllables. Both types of models are trained with the same database, which is independent from the application.

43 speakers and 842 utterances (EUROM1 data base [3]) were used for training and 17 speakers and 512 utterances for testing. CMS was used since the training and testing data bases were recorded differently. First and second differential parameters plus the differential energy were employed. The size of the codebooks was 128 (32 for the differential energy). Gaussian SCHMM with quantization to the 6 (2 for the energy) closest codewords were used. The units were modeled with 4 states, allowing transitions to the two next states. Syllabic group models had 8 states with no skips. Table 1 shows the results for 12 basic parameters, both for MCC and FLFBE. Again, FLFBE performs noticeably better than MCC.

	Sentence recognition	Word Accuracy	Word Correct
MCC	75.0	89.7	92.85
FLFBE	84.6	93.6	95.0

Table 1. Recognition percentages by using frequency and time filtering.

In the FLFBE test reported in Table 1, the differential energy was not included. Using it, sentence recognition goes down to 81.1%, but it still is much better than the MCC score. Moreover, the performance also is much better than using liftered LPCC (76%).

DHMM and telephone digits

Other few tests were carried out with a small Spanish connected digit data base collected through the public switching telephone network. 1022 digit strings were used for training and 339 for testing. Again, *Q=12* for FLFBE. MCC used 12 liftered coefficients computed from 20 bands. The energy of each frame was also employed for both MCC and FLFBE, and each set of parameters was quantized with 64 codewords. The DHMM results using Euclidan distance are depicted in Fig.3. In this case, there are small differences in performance between both representations, and a consistent behaviour that favours one of them is not observed.

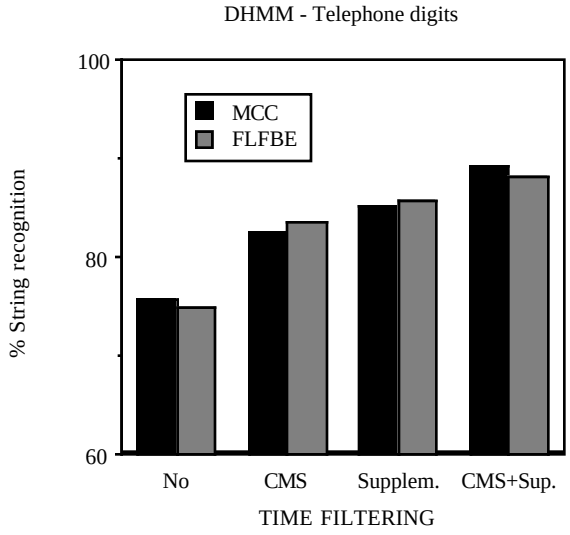

Fig.3 String recognition percentages.

7. Concluding remarks

In our experiments, the FLFBE technique has shown the capacity to achieve equal or better results than conventional cepstrum-based alternatives as MCC and LPCC in a wide variety of conditions, both alone and combined with time filtering. Actually, when diagonal covariance matrices are used, it allows to include in the HMMs a quefrency weighting in order to enhance the most discriminative oscillations of the spectral envelope curve.

Acknowledgments

The authors wish to thank the students Joan Marí, Montse Pla and Raul Gracia for their valuable help to obtain the experimental results.

References

1. L.R. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.
2. C. Nadeu, B.H. Juang, *Proc.ICSLP'94*, pp.1927-30.
3. C. Nadeu, J. Hernando, M. Gorricho, *Proc. EUROSPEECH'95*, pp. 1381-4.
4. C. Nadeu, P. Paches-Leal, B.H. Juang, *Proc. EUROSPEECH'95*, pp. 923-26.
5. A.N. Akansu, R.A. Haddad, *Multirresolution signal decomposition*, Acad. Press, 1992.
6. Y. Tohkura, *IEEE Trans. ASSP*, Vol. 35, No.10., Oct. 1987.
7. B.H. Juang, L.R. Rabiner, J.G. Wilpon, *IEEE Trans. ASSP*, Vol. 35, No.7., July 1987, pp. 947-53.
8. H.C. Wang, H.-F. Pai, *Proc. EUROSPEECH'93*, pp. 341-4.
9. S.V. Vaseghi, P.N. Conner, B.P. Milner, *IEE Proc. I*, Vol. 140, No. 5, Oct. 1993, pp. 317-20.
10. E. Lleida, J. B. Mariño, A. Moreno, *Proc. EUROSPEECH'93*, pp. 1801-4.