COMPARISON OF TIME & FREQUENCY FILTERING AND CEPSTRAL-TIME MATRIX APPROACHES IN ASR

Dušan Macho[1,2], *Climent Nadeu*[1], *Peter Jančovič*[2,*], *Gregor Rozinaj*[2], *Javier Hernando*[1]

[1]TALP Research Center, TSC Dept., UPC, Barcelona, Spain
[2]Dept. of Telecommunications, STU, Bratislava, Slovakia
e-mail: dusan@gps.tsc.upc.es

ABSTRACT

In current speech recognition systems, speech is represented by a 2-D sequence of parameters that model the temporal evolution of the spectral envelope of speech. Linear transformation or filtering along both time and frequency axes of that 2-D sequence are used to enhance the discriminative ability and robustness of speech parameters in the HMM pattern-matching formalism. In this paper, we compared two recently reported approaches which operate on the sequence of logarithmically compressed mel-scaled filter-bank energies: the first approach - TIFFING (TIme and Frequency FilterING) - applies FIR filters to that 2-D sequence along both axes, while the second one - CTM (Cepstral Time Matrix) - uses the DCT to compute a set of parameters in the 2-D transformed domain. They are compared in several ways: (1) analytically, using Fourier transformation, (2) statistically and (3) performing recognition tests with clean and noisy speech.

1. INTRODUCTION

The HMMs used in current speech recognition systems assume that the observation vectors can be modeled by Gaussian distributions with diagonal covariance matrices, i.e., they assume the elements of each vector are uncorrelated. However, the logarithmically compressed filter-bank energy (log FBE) parameters are highly correlated (e.g., the correlation coefficient for two adjacent bands is 0.92 for the TI database) and, thus, a linear transformation is needed to decorrelate them. Due to its closeness to the optimal KL transform, the discrete cosine transform (DCT) not only nearly decorrelates the log FBE vector but also sorts the transformed parameters – cepstral coefficients – in variance order. After DCT, the rectangular lifter window is usually applied to retain the highest-energy cepstral coefficients and to smooth in this way the spectral envelope of speech. The use of lifters different from the rectangular one has no effect on the recognition rates in the context of HMMs with diagonal covariance matrices [1], so the cepstral representation can not benefit from discriminative weighting of cepstral coefficients.

Recently, it was suggested [1] that the transformation to the cepstral domain can be avoided by performing frequency filtering (i.e., filtering along the frequency axis) of the log FBE vector. Frequency filtering (FF) decorrelates the vector coefficients and, as it is equivalent to weighting in cepstral domain, it can emphasize the most discriminative cepstral coefficients. In this way, FF allows to use a wide spectrum of lifter window shapes by means of different frequency filters that can be tuned to the database, task, or noise conditions [2]. Among them, the database-independent second-order frequency filter $z\text{-}z^{-1}$ has shown a rather good performance for a wide range of conditions.

Another assumption present in HMM modeling is that each observation vector is uncorrelated with its temporal neighbors. Therefore, similarly as in the case of frequency dimension, a linear transformation or filtering of parameters along the time axis may decorrelate the time sequence of parameters. Moreover, from the point of view of modulation spectrum analysis, the discriminative ability of the sequence of parameters can be improved by enhancing of modulation frequencies around 3-4 Hz [3].

Summarizing the previous paragraphs Figure 1 shows four alternatives how to diminish the correlation and to enhance the discrimination of the 2-D time sequence of log FBE parameters. The most common way is that from cell 3 where the parameters are cepstral coefficients and their time derivatives. In this paper, we put in relation two recently reported approaches: tiffing (cell 1) [4,5] and cepstral-time matrices (CTM, cell 4) [6]. In the context of 2-D modulation spectrum, the objective of both approaches is the same: to emphasize the most discriminative and robust region of the 2-D modulation spectrum. So far, both approaches have been tested in tasks with different conditions (different database, noise conditions...). In order to compare them and to stress their advantages/disadvantages we tested both approaches using the same recognition task consisting of clean and real-noise single digit recognition.

filtering in frequency and **filtering** in time (tiffing) 1	**filtering** in frequency and **transformation** in time 2
transformation in frequency and **filtering** in time (delta-cepstrum) 3	**transformation** in frequency and **transformation** in time (CTM) 4

Figure 1 Four alternatives of further processing of the log FBE sequence.

2. TIFFING

In tiffing, the 2-D sequence of log FBE vectors is filtered first along the frequency axis and then along the time axis. In the frequency filtering part, FIR filters of lengths 2 or 3 are usually employed as a consequence of the relatively low number of bands (usually about 12 log FBEs are used). Note that only 2 or 3 adjacent bands of the log FBE parameter vector are involved in the computation of each FF coefficient, a fact that differs from the whole-band approach used in the CTM computation. Moreover, the FF coefficients still lie in the frequency domain.

In this work, we use the frequency filter $z\text{-}z^{-1}$, denoted as FF2, whose impulse response is $h(k)=\{1,0,-1\}$. As the Fourier transform of $h(k)$ is a sine wave, FF2 emphasizes the middle

[*] P. Jančovič is also with Slovak Academy of Sciences, Bratislava, Slovakia.

quefrency indexes that convey important information about the formant structure of speech spectrum.

The effect of time filtering is reflected in the modulation spectrum of filtered parameters. When speech is represented by various feature sets (static and time-filtered), the modulation frequency bands of time filters are distributed along the interval of modulation frequencies that both is phonetically relevant and does not carry an excessive spectral estimation noise [4]. As time filters we used a set of Slepian FIR filters that are orthogonal and show a maximum concentration of power in that interval of interest. In this work, Slepians are always cascaded with a first-order equalizer $1\text{-}0.97z^{-1}$ that approximately equalizes the modulation spectrum of speech [3]. As will be shown in Section 3, by using a set of DCT basis sequences as filters very similar effects to that of Slepians+equalizer on the modulation spectrum of speech can be obtained.

3. CEPSTRAL-TIME MATRIX (CTM)

A cepstral-time matrix, $C_{\mathrm{CTM}}(m,i)$, is computed by applying a 2-D DCT to the spectral-time matrix consisting of L stacked adjacent log FBE vectors. Since the 2-D DCT can be decomposed in two 1-D DCTs, the CTM computation can be performed in the following way:

$$\log S(k,n) \xrightarrow{DCT_k} c(m,n) \xrightarrow{DCT_n} C_{CTM}(m,i) \qquad (1)$$

Thus, by applying the first DCT to the frequency index k, a 2-D time sequence of cepstral coefficients $c(m,n)$ is obtained, where n is the time frame index. The second DCT transforms a time sequence of L stacked cepstral vectors to the modulation frequency domain. To each DCT basis sequence can be associated a band in the modulation frequency domain with a central frequency that depends upon both the number of cepstral vectors L and the frame rate. Consequently, to each column of CTM a modulation frequency can be assigned. On the other hand, each line of CTM corresponds to a quefrency index. Typically, only a sub-matrix of CTM is used for recognition and in this way, modulation frequencies and quefrency indexes are selected.

In [7], the authors pointed out that both cepstral derivatives and columns of CTM are produced by weighted summation of cepstral vectors involved in their computation. Actually, the transformation performed by the second DCT in (1) has an effect in the modulation frequency domain very similar to that of time filtering. To clarify this assertion, Figure 2 illustrates the modulation spectra of time-filtered speech when both DCT basis and Slepians+equalizer were employed as time filters. To compute the modulation spectra, the first three DCT basis sequences and Slepians (both DCT sequences and Slepians+equalizer of length 15) have been used together with the function $1/(1\text{-}0.97z^{-1})$ which was taken as an approximation of average speech modulation spectrum. The two high-energy bands in Figure 2 with the lowest-frequency content correspond to the first DCT basis sequence and the first Slepian+equalizer filter. The two high-energy bands in the middle correspond to the second and the last two bands to the third DCT basis and Slepian+equalizer filter. Remarkable similarity between the appropriate high-energy bands of DCT and Slepian-filtered parameters can be observed. Perhaps the largest difference lies in the first band, where the DCT filter removes the 0th modulation frequency component, while the Slepian filter not.

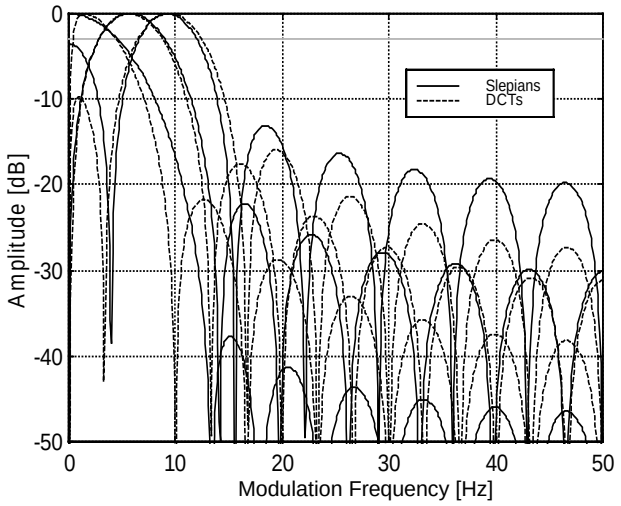


Figure 2 Theoretical modulation spectra of speech time-filtered by DCT basis and Slepian+equalizer filters.

4. DISCRIMINATIVE REGION IN 2-D MODULATION SPECTRUM

The 2-D modulation spectrum (2-D MS), introduced in [2], is estimated from the 2-D sequence of log FBEs by computing and averaging function $|C(m,\theta)|^2$ over a speech database. The function $C(m,\theta)$ is obtained by inverse discrete-Fourier transforming $\log S(k,n)$ from the frequency domain k to the quefrency m and by the Fourier transforming the resulting sequence $c(m,n)$ from the time domain n to the modulation frequency domain θ,

$$\log S(k,n) \xrightarrow{IDFT_k} c(m,n) \xrightarrow{FT_n} C(m,\theta) \xrightarrow{|\cdot|^2} |C(m,\theta)|^2 \qquad (2)$$

Note the similarity of the first two steps in (2) with the computation of CTM as expressed by (1). Since the spectrum is an even function, the first step coincides for both transformations. In the second step, the cepstral vectors involved in the CTM computation can be selected by applying a rectangular window $w(n)$ of length L on the time sequence of each cepstral coefficient, i.e.,

$$c_{\mathrm{w}}(m,n) = c(m,n)w(n) \qquad \text{for each } m. \qquad (3)$$

Then, the relationship between the function $C(m,\theta)$ and the cepstral-time matrix $C_{CTM}(m,i)$ can be written as

$$C_{CTM}(m,i) = \left[\mathrm{Re}\{C(m,\theta) * W(\theta)\}\right]\Big|_{\theta=\frac{2\pi i}{L}} \qquad (4)$$

where $W(\theta)$ is the Fourier transform of the window $w(n)$. The equation (4) justifies the use of 2-D MS in the context of CTMs since the cepstral-time matrix is a sampled version of the real part of the function $C(m,\theta)$ convolved with $W(\theta)$.

Both CTM and tiffing can benefit from the conclusions that could be drawn from the analysis performed on the 2-D MS. Actually, both tiffing and CTM approaches are able to modify the 2-D MS and enhance its most discriminative and robust components.

A way to find out a region with good discrimination capability is to compute the quotient between the between-class and the within-class variances of the 2-D modulation spectrum. We used 13 log FBE parameterization and 55 male and 57 female speakers from the adult portion of the TI database to compute the variances for the TI single digit task. Assigning a class to each digit, the within-class variance is obtained by computing the variance in each class and averaging them over all 11 classes. The between-class variances are also computed and averaged over all speakers.

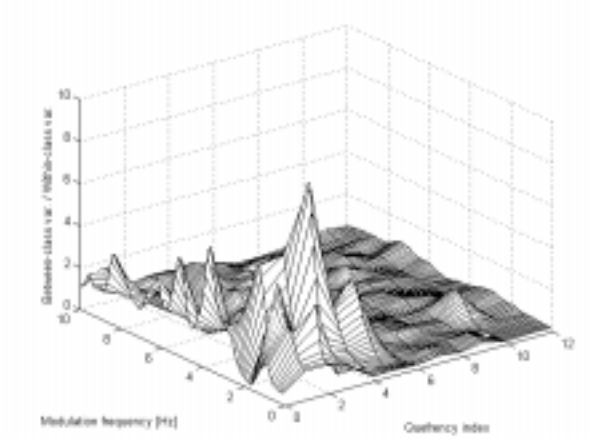
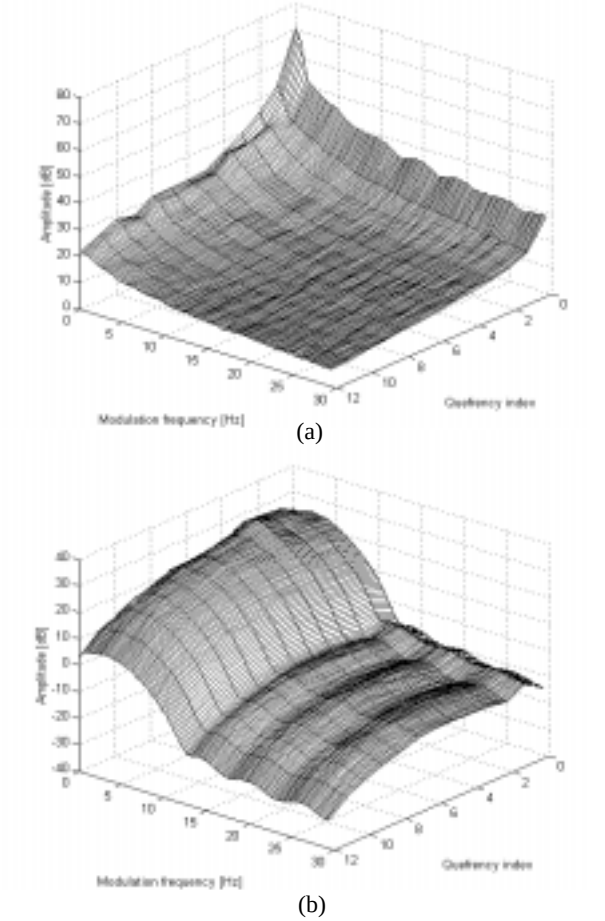


Figure 3 The quotient $Variance_{between}/Variance_{within}$ obtained from the TI isolated digit database.

Figure 3 shows the resulting quotient for modulation frequencies up to 10 Hz. It can be observed that in the region roughly around θ = 3-4Hz and m = 3-4 the between-class variance is significantly higher than the within-class variance. As will be shown in Subsection 5.2, the best recognition rates are obtained when either CTMs or tiffing enhances that region.

The effect of tiffing on the 2-D MS is illustrated in Figure 4. It can be seen in Figure 4(a) that the 2-D MS corresponding to the sequence of log FBE vectors without tiffing shows a high content of very low modulation frequencies and quefrencies. On the other hand, as depicted in Figure 4(b), the quefrencies around 3-4 and a modulation frequency band approximately up to 4Hz dominate in the log FBE sequence tiffed by TF1 and FF2 (see Subsection 5.1 for filter notation). Thus, the domain with potentially high discrimination ability is emphasized by filtering in comparison to non-filtered parameters. In the CTM case, since a non-zero weighting applied on CTMs has no effect on the recognition rate for HMMs with diagonal covariance matrix, weighting reduces to selecting of elements to be kept for recognition in each CTM.



(a)

(b)

Figure 4 2-D MS of clean speech computed over a part of TI database before (a) and after (b) tiffing has been applied. For tiffing, FF2 and TF1 have been used.

5. RECOGNITION TESTS

5.1 Test conditions

Isolated digits from the adult portion of TI database were used in our tests. For training, clean speech was always used. For noisy tests, speech was contamined by pub, railway station and white noises with SNR=10dB. Both real noises were driven from SUNROM-1 noisy database [8]. In the front-end, no preemphasis has been used and the sequence of log FBEs was computed from 30 ms long Hamming windowed segments obtained every 10 ms from the signal. CDHMMs with 8 emitting states, one Gaussian per state and diagonal covariance matrix have been used to model the digit units. For silence, HMM with 3 emitting states was used.

For the time part of tiffing, we used Slepian filters along with the first-order equalizer $1\text{-}0.97z^{-1}$. Three different filters were employed (see [3] for notation and design procedure): TF1: k=1, W=12, L=14; TF2: k=2, W=12, L=14 and TF3: k=3, W=12, L=14. The length of each TF was 15. The FF2 filter was employed for frequency filtering. In the recognition tests with DCT basis functions as time filters, the first three DCT basis functions (excluding the d.c. one) of length 15 were employed.

5.2 Preliminary tests

In our preliminary tests, we intended to select the discriminative region of 2-D MS found in Section 4 by using both tiffing and CTM approaches. For that purpose we used 13 log FBEs tiffed by TF1 and FF2 whose corresponding 2-D MS is depicted in Figure 4(b). Only 6 CTM coefficients were chosen, according to Figure 3, from the CTM computed from a 16x16 spectral-time matrix. Recognition rates 99,4% and 99,44% were obtained for tiffing and CTM, respectively.

Next, we applied both approaches to the case of additive white noise. For tiffing, the combination of TF1&FF2 and TF2&FF1 (FF1 is the first-order filter $1\text{-}z^{-1}$) was employed and the high-frequency endpoint of the frequency-filtered vector was discarded due to its high contamination by white noise (in total, 24 tiffed parameters were used to represent each speech frame). From the original 16x16 CTM, the 33 most robust elements were chosen after several experiments. In this way, 88,01% and 87,73% recognition rates were yielded by tiffing and CTM, respectively (the clean speech recognition rate is 99,12% for both approaches in this case).

The previous experiments showed that very similar results can be achieved by both approaches. The CTM approach works well with a small number of coefficients (only 6) in the clean digit recognition task. For noisy speech, however, tiffing needed less parameters (24) than CTM (33), and the way the robust parameters were chosen was more straightforward for tiffing than for CTM (the final matrix possessed a rather sparse form).

5.3 One time-filtered feature set

In the experiments presented in this and the next subsection, either DCT basis sequences or Slepians+equalizer have been used to filter (along the time axis) both FF and MFCC parameter sequences. In the case of FF parameters, 13 log FBEs have been frequency filtered with FF2. In the MFCC case, 20 log FBEs have been transformed by using DCT and the first 12 cepstral coefficients have been retained (without $c(0)$). As the frame energy improved recognition rates in all conditions for MFCC, we added it to each MFCC vector (note that this differs from the usual way the CTMs are employed). Recognition results are in Table 1, where the first two lines correspond to original tiffing, the second two lines can be interpreted as a variant of tiffing, the third two lines are time-filtered cepstra and each of the last two lines represents a column of CTM with the frame energy added.

	Clean	Pub	Station	White
TF1, FF2	99,40	73,76	92,72	62,74
TF2, FF2	99,03	76,94	89,98	74,41
1st DCT, FF2	99,40	74,33	89,26	78,15
2nd DCT, FF2	98,67	74,77	88,17	69,22
TF1, MFCC_E	98,47	69,76	90,74	69,34
TF2, MFCC_E	98,11	74,73	89,38	77,55
1st DCT, MFCC_E	98,75	72,35	87,79	75,37
2nd DCT, MFCC_E	98,18	70,99	89,38	69,66

Table 1 Percentage recognition rates for one time-filtered feature set.

In general, the results shown in Table 1 do not differ much, but some interesting observations can be picked out from them. Observing the clean speech column, it is obvious that the FF2 parameters yield the best recognition rates (mainly due to the discriminative weighting capability of FF). The best recognition rates for real noises were obtained by tiffing. On the other hand, in the white noise case, the 1st DCT basis function performs considerably better than the TF1 filter (note, that TF1 does not entirely remove the d.c. modulation frequency component). This problem of tiffing can be solved by discarding the high-frequency endpoint of FF vector; in fact, in that case, the result improves from 62,74% to 81,17%. Observe that, except the case where TF1 is used for white noise, the recognition rates yielded by tiffing are always better than those obtained by columns of CTM (the last two lines of Table 1).

5.4 Two and three time-filtered feature sets

Indeed, when two or three time-filtered feature sets are used to represent each speech frame, both clean and noisy speech recognition rates improve. Recognition rates are in Table 2 whose lines have the same interpretation as those from Table 1.

	Clean	Pub	Station	White
TF1, TF2; FF2	99,76	82,01	95,09	69,54
TF1, TF2, TF3; FF2	99,68	84,43	94,85	76,22
1st, 2nd DCT; FF2	99,72	81,69	93,96	79,28
1st, 2nd, 3rd DCT; FF2	99,64	83,50	93,08	78,35
TF1, TF2; MFCC_E	99,52	81,05	95,05	80,48
TF1, TF2, TF3; MFCC_E	99,36	81,17	94,29	81,89
1st, 2nd DCT; MFCC_E	99,32	79,44	93,88	80,89
1st, 2nd, 3rd DCT; MFCC_E	99,28	78,71	93,08	80,56

Table 2 Percentage recognition rates using two and three time-filtered feature sets.

Clean speech recognition rates become closer for all parameter sets. Tiffed features perform better than CTM features, except the white noise case (discarding the high-frequency endpoint of FF vectors yields 84,59% recognition rate in the case of two feature sets). The advantage of tiffing is especially noticeable in the pub (bubble) noise case and when the third feature set is added. However, the inclusion of the third feature set does not imply an improvement in most cases.

6. CONCLUSIONS

In this paper, two recently reported approaches operating on the 2-D sequence of log FBE parameter vectors – tiffing and CTM – have been compared. We showed that both approaches could take advantage of the analysis performed in the 2-D modulation spectrum of speech. CTM works particularly well with a small number of parameters in clean isolated digit recognition. Although the differences in recognition rate of both approaches have not been large, however, tiffed features have shown a consistently better performance for both clean speech and speech contamined with real noises.

7. ACKNOWLEDGMENTS

This work has been supported by Spanish Government Agency CYCIT, projects TIC98-0683 and TIC98-0423-C06-01, and the work of one of the authors, P. Jančovič, has been founded by the Tempus-Telecomnet project S_JEP 09326-95 and by Slovak Grant Agency, number of grant 2/6120/99.

8. REFERENCES

[1] C. Nadeu et al., “On the Decorrelation of Filter-Bank Energies in Speech Recognition”, Proc. Eurospeech, 1995, pp. 1381-84.

[2] D. Macho and C. Nadeu, “On the Interaction between Time and Frequency Filtering of Speech Parameters for Robust Speech Recognition”, Proc. ICSLP, 1998, pp. 1487-90.

[3] C. Nadeu et al., “Filtering the Time Sequence of Spectral Parameters for Speech Recognition”, Speech Communication 22, 1997, pp. 315-322.

[4] C. Nadeu, “On the Filter-bank-based Parameterization Front-end for Robust HMM Speech Recognition”, Accepted for Workshop on Robust Methods in Tampere, 1999, Finland.

[5] D. Macho et al., “Time and Frequency Filtering for Speech Recognition in Real Noise Conditions”, Accepted for Workshop on Robust Methods in Tampere, 1999, Finland.

[6] B.P. Milner and S.V. Vaseghi, “Speech Modelling using Cepstral-Time Feature Matrices and Hidden Markov Models”, Proc. ICASSP, 1994, pp. 601-604.

[7] B.P. Milner and S.V. Vaseghi, “An Analysis of Cepstral-Time Matrices for Noise and Channel Robust Speech Recognition”, Proc. Eurospeech, 1995, pp. 519-522.

[8] SUNROM-1: CD-ROM with noises from the ESPRIT 2094 Project.