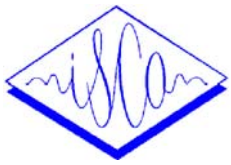


5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998

ON THE INTERACTION BETWEEN TIME AND FREQUENCY FILTERING OF SPEECH PARAMETERS FOR ROBUST SPEECH RECOGNITION

Dušan Macho and Climent Nadeu***

*Dept. of Telecommunications, Slovak Technical University and Dept. of Speech Analysis and Synthesis, Slovak Academy of Sciences, Bratislava, Slovakia
**Dept. of Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona, Spain

ABSTRACT

One of the great today's challenges in speech recognition is to ensure the robustness of the used speech representation. Usually, the recognition rate is strongly reduced when the speech is corrupted, e.g. by convolutional or additive noise, and the speech features are not designed to be robust. In this paper we study the effect of additive noise on the logarithmic filter-bank energy representation. We use time and frequency filtering techniques to emphasize the discriminative information and to reduce the mismatch between noisy and clean speech representation. A 2-D spectral representation is introduced to see the regions most affected by noise in the 2-D quefrency-modulation frequency domain and to help to design the frequency and time filter shapes. Experiments with one and two dynamic feature sets show the usefulness of the combination of time and frequency filtering for both, white and low-pass noise speech recognition. At the end the power time and frequency filtering technique is presented.

1. INTRODUCTION

Only a part of the information contained in the speech signal is used for speech recognition. Moreover, a speech signal can be distorted by non-speech components (e.g. channel or microphone distortion, additive noise, reverberation…). It is necessary to extract phonetically important features with good discriminative properties and robustness when used in adverse environments.

For recognition purposes, speech is often converted to a time sequence of log filter-bank energies (log FBE). In this way, the considered speech unit is represented as a two-dimensional (2-D) time-frequency sequence. This sequence is further processed in order to obtain more robust and discriminative features (e.g. transformed to mel-cepstrum, RASTA filtered [1]…). Recently, the authors in [2] showed that a simple filtering performed on the frequency dimension of every frame (Frequency Filtering – FF) gives better recognition results for clean speech than cepstral coefficients. The FF can be seen as a liftering operation performed in the spectral domain. The frequency filters in [2] were designed to equalize the variance of cepstral coefficients and a simple, database independent, second-order filter $z\text{-}z^{-1}$ (here denoted as FF2) was found as a good compromise. In [3], the FF features appeared more robust than cepstral coefficients when speech is distorted by additive white noise and a first-order filter $1\text{-}z^{-1}$ (FF1) gave good recognition results.

The components of the speech feature vector vary in time, according to the changes of the speech signal, describing time trajectories. The spectrum of the time trajectory is called modulation spectrum. The typical speech modulation spectrum decreases along the modulation frequency axis [4]. Thus, the low modulation frequencies generally dominate the distance computation in the classifier (similarly, as do the low quefrency components) but they do not carry the most discriminative information [5]. Moreover, when the speech is corrupted by stationary convolutional noise, the 0^{th} modulation frequency is the most affected in the log FBE representation. Thus, filtering on the time dimension (Time Filtering – TF) can remove undesirable parts of the modulation spectrum.

In [5], both time and frequency filtering were presented jointly, but considering that there is not interaction between them. However, we recently observed some facts that led us to consider that the interaction exists. Firstly, the noticeable better clean speech performance of FF with respect to cepstrum that is obtained when only one static feature set (without TF) is used, may be reduced to a slight difference if dynamic features are included in the representation. Second, FF looses its good performance for noisy speech when the noise is colored.

In this work, we gain more insight into that interaction problem by using the 2-D modulation spectrum representation obtained from log FBE sequence. We observed, for example, that the mean value of that 2-D function for noisy speech shows higher values at low indices than the corresponding function for clean speech. Thus, TF and FF can improve the recognition rate by attenuating the most distorted regions. Moreover, in the same way that it can be convenient to use slightly different time filters in two different frequency bands [6], the use of different frequency filters in different modulation frequency regions can also increase the recognition performance for speech distorted by additive noise. For designing the filters, we can take advantage of that 2-D spectral representation.

For the recognition tests presented in this paper, we used the following conditions: single digits from the adult portion of the TI database, decimated from 20 kHz to 8 kHz sampling rate; no preemphasis; 30 ms long Hamming windowed frames with 10 ms shift; 13-order log FBE basic parameterization scheme; continuous density HMMs with 8 states per digit and 3 states for the silence model; for noisy speech, either stationary white additive noise or low-pass additive noise with cut-off frequency 1100 Hz were added to the clean speech to obtain SNR equal to 20 dB and 10 dB. Training was performed always with clean speech and testing with noisy speech.

This work was carried out during the stay of D. Macho at UPC Barcelona and sponsored by Spanish government and partially by Slovak Academy of Sciences.

2. THE 2-D MODULATION SPECTRUM

For better analysis purposes, we spread modulation spectrum representation [4] in two dimensions, where the modulation spectrum of every cepstral coefficient is present. Let *log S(k,n)* be the short-time log FBE estimate of the speech signal with *k* denoting the filter-bank output and *n* the frame index. The 2-D modulation spectrum (in [7], the modulation spectrogram has been introduced which displays the evolution of low modulation frequencies in time and frequency) is then estimated by computing and averaging function $|C(m,\theta)|^2$ over a speech database. $|C(m,\theta)|^2$ is obtained by inverse discrete-Fourier transforming from the frequency domain *k* to the quefrency *m* and by the Fourier transforming from the time domain *n* to the modulation frequency domain θ ,

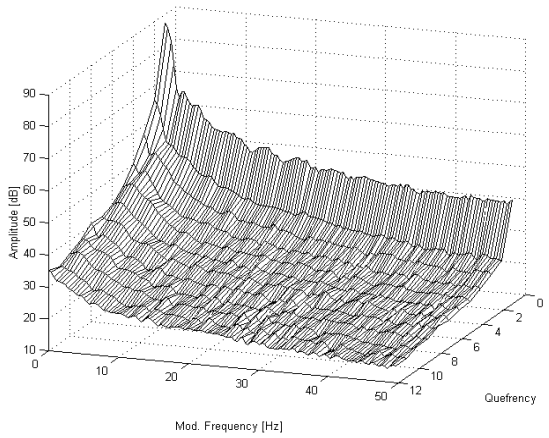
$$\log S(k,n) \xrightarrow{IDFT_k} c(m,n) \xrightarrow{FT_n} C(m,\theta) \xrightarrow{|\cdot|^2} |C(m,\theta)|^2 . \quad (2)$$

The 2-D modulation spectrum estimated from clean isolated digits database is shown on Figure 1(a). The decreasing tilt in both dimensions can be observed. Figure 1(b) shows the 2-D modulation spectrum of speech distorted by additive white noise. The low indices in quefrency and modulation frequency seem to be the most affected by noise.

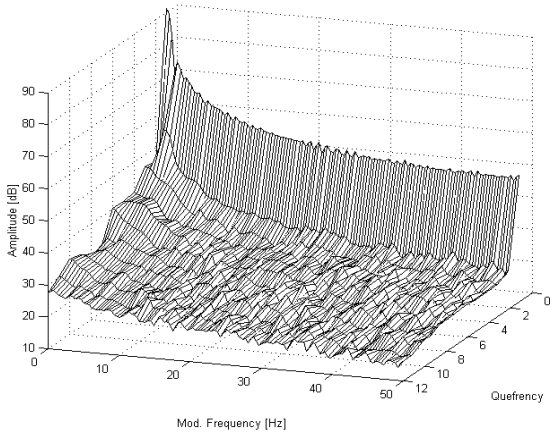
The mismatch between training and testing log FBE representation is the main reason of the poor recognition results obtained when the speech corrupted by additive noise is used for testing. We computed the mismatch between the clean and noisy speech representation as

$$\left|C_{noisy}(m,\theta) - C_{clean}(m,\theta)\right|^2 \quad \text{for all } m \text{ and } \theta , \quad (3)$$

and averaging it over many speakers and utterances we estimated the 2-D modulation spectrum of mismatch. Figure 2 shows the 2-D modulation spectra of mismatch for speech corrupted by additive white noise (a) and additive low-pass noise (b), both for *SNR=10dB.* The largest mismatch is situated

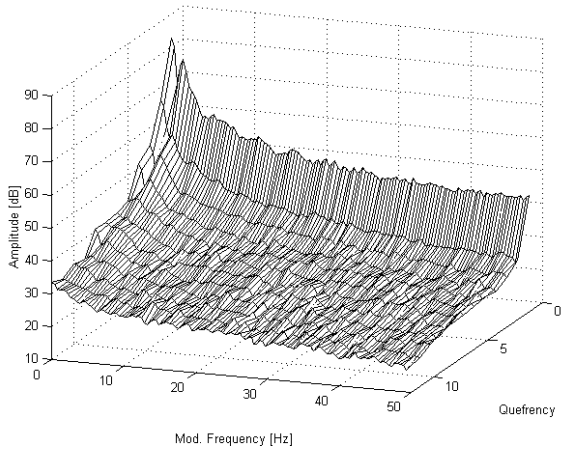


(a)

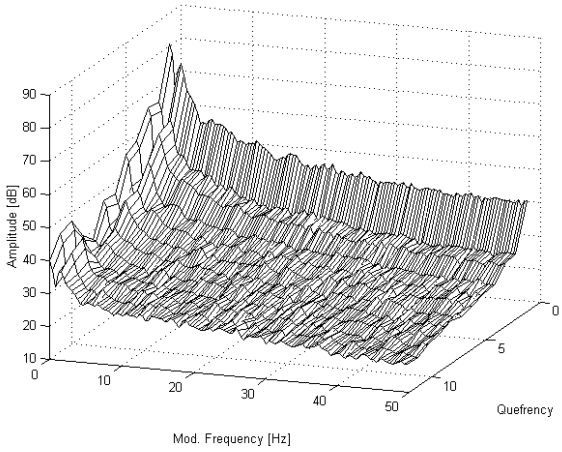


(b)

Figure 1: 2-D modulation spectra of (a) clean and (b) white noise speech with *SNR=10dB*



(a)



(b)

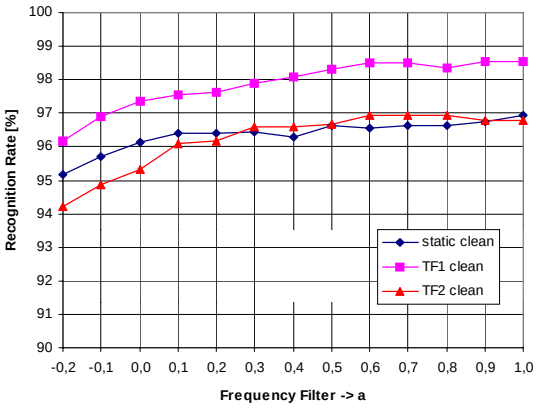
Figure 2: 2-D modulation spectra of mismatch for (a) additive white noise and (b) additive low-pass noise, both with *SNR=10dB*

in low quefrencies and modulation frequencies with its maximum at the (0,0) point. Note the difference along quefrency between both figures (especially at low modulation frequencies), while along modulation frequency they are similar. In both figures, when the modulation frequency increases, low and middle quefrencies are less affected and can be used for recognition. Using FF and TF, we can remove the distorted part of the 2-D modulation spectrum and even better discriminative properties of features can be obtained. If we emphasize two different regions in the modulation frequency dimension by using two different time filters, one for each of two feature sets, we can use a different frequency filter for every region in order to weight differently in the quefrency dimension. In the following sections, the effect of frequency and time filtering on the recognition performance is shown.

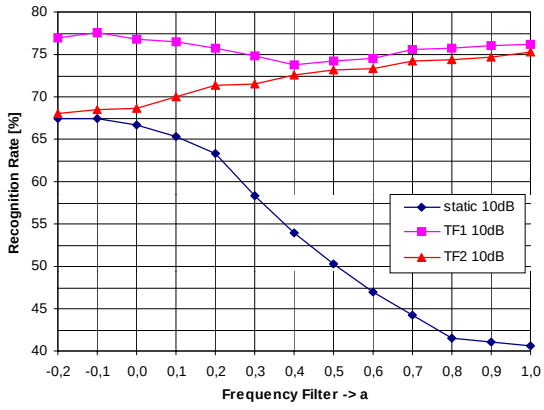
3. RECOGNITION TESTS

3.1 Static Feature Set

If no TF is used, we refer to the speech representation as a static feature set. In the following, only the effect of the frequency filtering is presented. For this purpose, we used 13 different frequency filters of length 3 with system function $(z-1)(z+a)$, where a changes from $-0,2$ to $1,0$ with step $0,1$ (note, that the filter with $a=0$ is FF1 and $a=1$ is FF2). The transform response of the filter is a quefrency function (a lifter). Changing the parameter a of the filters in the interval $<-0,2; 0,2>$, the shape of the lifters in low and middle quefrencies changes, while does not change much in high quefrencies. When the parameter a changes in the interval $<0,6; 1,0>$, the lifter shape in the high quefrencies changes while in low and middle quefrencies it does not. Since the first and the last filtered log FBE of each frame contain absolute energy [2] they can carry much noise, so that they were not used in this feature set.



(a)



(b)

Figure 3: Recognition rate in terms of the FF and TF used in the parameterization for (a) clean and (b) white noise speech with *SNR=10dB*

Figure 3 shows the recognition rates for all filters. The clean speech recognition rate (Figure 3(a)) increases when a increases and FF2 gives the best results. However, when the speech is corrupted by additive white noise (Figure 3(b)), filters that attenuate low and middle quefrencies are preferable. This is due to fact that, although the low and middle quefrencies are useful for clean speech recognition, they are severally affected by noise [8].

3.2 One Time-Filtered Feature Set

In this case, time filtering is applied to the sequence of features. As time filters we used the two different Slepian filters (the same as those in [4] with parameters $K=1$, $W=12$, $L=14$, denoted as TF1 and $K=2$, $W=12$, $L=14$ denoted as TF2) joint with equalization $1-0,97z^{-1}$. TF1 preserves the modulation frequencies of speech roughly from 0 Hz to 3 Hz and TF2 from 2 Hz to 9 Hz.

The first test we performed was without frequency filtering. From the first two lines of Table 1 it seems that the features from the TF1 region yield more discriminative information (97,71% recognition rate for clean speech) than those from the TF2 region (95,13%). However, when noisy speech is recognized, the TF2 features give better results and are more robust than the features from the TF1 region.

Technique	Clean	SNR=20dB	SNR=10dB
White noise			
TF1, no FF	97,71	50,70	16,10
TF2, no FF	95,13	64,10	41,01
TF1, FF1 13/12	97,79	94,37	82,98
TF1, FF2 13/12	99,16	95,90	81,17
TF2, FF1 13/12	97,26	92,84	77,02
TF2, FF2 13/12	98,39	95,13	79,60
Low-pass noise			
TF1, FF1 13/12	97,79	90,38	78,11
TF1, FF2 13/12	99,16	90,30	77,14
TF2, FF1 13/12	97,26	92,23	77,99
TF2, FF2 13/12	98,39	93,32	78,63

Table 1: Recognition rates in % using TF1, TF2 without frequency filtering and with FF1 and FF2

The situation changes when FF is used in conjunction with TF. Figure 3 shows the behavior of both, TF1 and TF2 feature sets in terms of different FFs. For clean speech, the TF1 features give better results for every FF than the TF2 features (see Figure 3(a)). In the noisy case, the frequency filtering partially reduces the high content of noise in TF1 region and the recognition rates even outperform the TF2 results (Figure 3(b)). Moreover, a different behavior of the feature sets from two mentioned modulation frequency regions can be observed. For the TF1 region, the frequency filters which attenuate more the low and

middle quefrencies (those with *a=<-0,2; 0,2>*) give slightly better results than the others for noisy speech. Attenuating the high quefrencies in this region seems to improve the recognition too (filters with *a=<0,6; 1,0>*). For the TF2 region, an increasing tendency in the recognition rate can be observed when the coefficient *a* increases and FF2 is the optimal filter.

We tried to include the first and the last filtered log FBE to the feature vector. Only including of the first one improved the noisy speech recognition. In general, the first log FBE contains more speech energy and is not affected by additive noise so much as the last one, which contains less speech energy. Table 1 shows the recognition rates with the first log FBE included in the feature vector.

In the additive low-pass noise case, the results from experiments when only FF is used are very low (near 17% for FF1 or FF2 and *SNR=10dB*). This is due to fact, that the filtered log FBEs include the step of the transition band of the noise spectrum. Since the step effect is constant in the time, it can be almost canceled by the time filter. Results with different time and frequency filters are in the Table 1.

3.3 Two Time-Filtered Feature Sets

The recognition rate can be improved using features from both time-filtered regions in two different feature sets. We have found the static feature set is a source of errors when used together with time-filtered features and we do not use it. In the Table 2, the recognition tests are presented for three combinations of time and frequency filters. For clean speech, the best recognition result is obtained when FF2 filter is used for both time-filtered feature sets. When FF1 is used, the recognition for clean speech decreases, but increases for noisy speech. From Figure 3(b) it can be observed (here the feature sets were used separately), that for the TF1 feature set the frequency filters which attenuate low and middle quefrencies are preferable and for TF2 features, the FF2 is the best filter. Using this observation, an additional improvement for noisy speech recognition was obtained. At the end of Table 2 the best recognition rates for low-pass noise speech are mentioned.

Technique	**Clean**	**SNR=20dB**	**SNR=10dB**
White noise			
FF1 13/12, TF1 & FF1 13/12, TF2	98,15	96,18	86,68
FF2 13/12, TF1 & FF2 13/12, TF2	99,48	97,06	84,59
FF1 13/12, TF1 & FF2 13/12, TF2	99,12	96,74	88,01
Low-pass noise			
FF1 13/12, TF1 & FF2 13/12, TF2	99,12	94,16	81,41

Table 2: Recognition rates in % for two feature sets

3.4 Power Frequency and Time Filtering

In this technique we assumed, that in the log FBE representation of noisy speech the high-energy coefficients are less affected by noise than the coefficients with low energy content. Thus, we use simple power operation on the log FBEs before they enter to the FF in order to emphasize the high-energy coefficients. In general, the power operation can be expressed as $\left|\log S(k,n)\right|^{\gamma}$. Table 3 shows the results from the same experiments as Table 2 but using square-power frequency and time filtering ($\gamma = 2$) for two feature sets. A clear improvement for noisy speech can be obtained while the recognition rates for clean speech do not decrease.

Technique	**Clean**	**SNR=20dB**	**SNR=10dB**
White noise			
FF1 13/12, TF1 & FF1 13/12, TF2	98,43	97,22	91,51
FF2 13/12, TF1 & FF2 13/12, TF2	99,28	98,03	90,95
FF1 13/12, TF1 & FF2 13/12, TF2	99,16	97,63	92,31
Low-pass noise			
FF1 13/12, TF1 & FF2 13/12, TF2	99,16	96,62	89,66

Table 3: Recognition rates in % for two feature sets using square-power frequency and time filtering

4. CONCLUSIONS

So far, time and frequency filtering have been studied separately. In this paper, we offer an introduction to their joint investigation. We showed TF-FF features are robust against stationary, additive white and low-pass noises for isolated digit recognition. A great advantage of this technique is that it does not decrease clean speech recognition results. In the further work, a 2-D filter can be designed, which will include different FF in different TF regions in one feature set. Moreover, the power coefficient can be optimized. Also, we want to extend the mentioned techniques to more difficult tasks.

5. REFERENCES

1. Hermansky, H., Morgan, N. "RASTA Processing of Speech", IEEE Trans. on Speech and Audio Processing, Vol.2, No. 4, 1-12, October 1994.
2. Nadeu, C., Hernando, J., Gorricho, M., "On the Decorrelation of Filter-Bank Energies in Speech Recognition", Proc. Eurospeech, 1381-84, 1995.
3. Hernando, J., Nadeu, C. "Robust Speech Parameters Located in the Frequency Domain", Proc. Eurospeech, 417-20, 1997.
4. Nadeu, C., Paches-Leal, P., Juang, B. H. "Filtering the Time Sequence of Spectral Parameters for Speech Recognition", Speech Communication 22, 315-322, 1997.
5. Nadeu, C., Marino, J. B., Hernando, J., Nogueiras, A. "Frequency and Time Filtering of Filter-Bank Energies for HMM Speech Recognition", Proc. ICSLP, 430-33, 1996.
6. Avendaño, C., Vuuren, S., Hermansky, H., "Data Based Filter Design for RASTA-like Channel Normalization in ASR", Proc. ICSLP, 2087-90, 1996.
7. Greenberg, S., Kingsbury, B. E. D., "The Modulation Spectrogram: in Pursuit of an Invariant Representation of Speech", Proc. ICASSP, 1647-50, 1997.
8. Hernando, J., Nadeu, C., "Linear Prediction of the One-Sided Autocorrelation Sequence for Noisy Speech Recognition", IEEE Trans. on SAP, Vol. 5, N 1, 80-84, 1997.