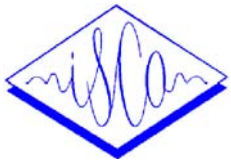

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

LIFTERED FORWARD MASKING PROCEDURE FOR ROBUST DIGITS RECOGNITION

YAO Kaisheng ⋆ †, *Bertram SHI* †, *Pascale FUNG* †, *CAO Zhigang*⋆

† Human Language Technology Center,
Department of Electrical and Electronic Engineering,
University of Science and Technology, HKUST, Clear Water Bay, Hong Kong
⋆ Department of Electronic Engineering,
Tsinghua University, 100084, Beijing, P.R.China

ABSTRACT

Using TI digits recognition experiments, we show that a combination of two dynamic speech features, Liftered Forward Masked (LFM) MFCC and 2-D cepstrum, can improve system robustness to additive Volvo noise while maintaining system performance comparable to standard MFCC features in clean conditions. Through experiments, we show that the information extracted by forward masking and by the 2D cepstrum are in some sense orthogonal. By combining the LFM MFCC and the 2-D cepstrum plus Δ 2-D cepstrum, we achieve a recognition rate above 90% on the TI connected digits task, even in additive Volvo noise condition with SNR as low as 0dB. This corresponds to a SNR gain over 30dB compared with standard MFCC plus dynamic and acceleration coefficients.

1. INTRODUCTION

In real environments, a speech recognizer can encounter distortions which degrade recognizer accuracy, such as additive background noise. The use of dynamic features (e.g. delta and acceleration coefficients) in addition has been proposed to improve robustness of speech recognition system in mismatched conditions [2, 3, 6, 8, 9]. However, Hanson[4] showed that while dynamic features such as delta and acceleration MFCC's alone are more robust to mismatch due to additive noise, the addition of static features improves performance in clean conditions. Thus, there appears to be a tradeoff between performance in clean matched conditions and robustness to additive noise. People commonly use a combination of static features and dynamic features.

In this paper, we show that a proper combination of two dynamic speech features, liftered forward masked (LFM) MFCC and the 2-D cepstrum, can result in speech recognizers which are robust to additive Volvo noise, while maintaining performance in clean environments comparable to that of standard MFCC features. A speech recognizer trained using the proposed features in clean speech achieved a clean speech word recognition rate of 99.0% on the TI connected digit task and a word recognition rate of 91.0% in 0dB additive Volvo noise. In contrast, the word recognition rate of a recognizer trained on standard MFCC plus dynamic and acceleration coefficients are 98.4% and 51.6%. Assuming a minimum acceptable recognition rate of 90%, there is a 30dB SNR gain using the proposed features over the standard MFCC based features. We emphasize that models in our experiments were trained only with clean speech, but tested in both clean and additive Volvo noise conditions without adaptation.

One of the possible reasons that the combination of these two features is so effective is that it seems that the information extracted using the forward masking technique is in some sense orthogonal to the information extracted by the 2-D cepstrum. A speech recognizer using the 2-D cepstrum calculated without forward masking performs better than a system trained using the 2-D cepstrum calculated with LFM procedure.

In contrast to ordinary forward masking procedures [7] which operated at the full sampling rate, the LFM procedure here operates at frame rate, similar to that proposed by Strope[2]. However, the forward masking parameters chosen by Strope were obtained based on physiological experiments, whereas forward masking parameters here by optimizing recognition accuracy on the digits task.

This paper is organised as follows. Section 2 describes the Liftered Forward Masking (LFM) procedure and the 2-D cepstrum. Section 3 compares the performance of speech recognizers trained using different parameters and feature combinations. Finally, Section 4 summarizes the conclusions of our experiments.

2. THE PROCEDURE OVERVIEW

2.1. Liftered Forward Masked MFCC (LFM MFCC)

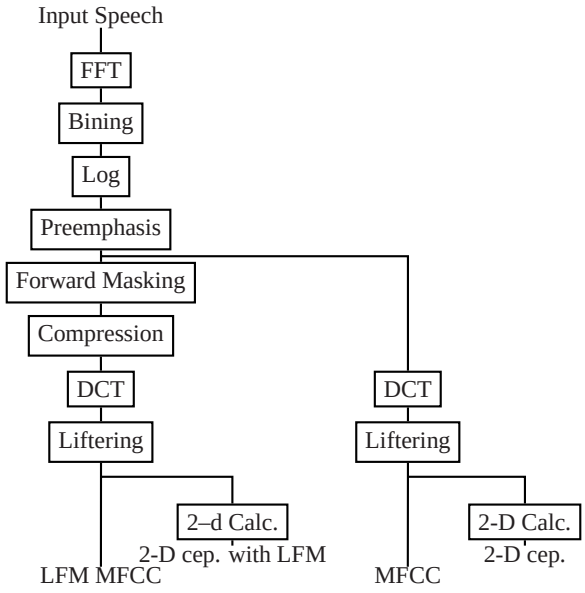

Figure 1: Procedure of LFM MFCC

A block diagram of the procedure stages of LFM MFCC is shown in the left branch of Figure 1. Similar to the first stages of the

computation of MFCC's, the power spectrum of each frame of speech is estimated by the FFT and then binned by triangular filters whose center frequency are equally spaced on the mel frequency scale. Delta energy is also extracted.

Preemphasis is applied using the physiologically based equal loudness curve used in [5]. In the log domain, this corresponds to an additive bias. At this stage, standard MFCC processing as shown on the right hand branch would apply the DCT and raised sine liftering to yield liftered MFCC coefficients. To obtain LFM MFCC, we apply a forward masking procedure to each critical band log power spectrum and cubic root compression [5] independently before the DCT operation. In forward masking, the log power spectrum at each critical band and the delta log energy value are processed independently according to the equation

$$c(n) = \begin{cases} T_s/\mu_a \cdot (x(n) - c(n-1)) \\ +(1 - T_s/\mu_b) \cdot c(n-1) & : \quad c(n-1) \leq x(n) \\ (1 - T_s/\mu_b) \cdot c(n-1) & : \quad otherwise \end{cases} \tag{1}$$

where $x(n)$ is the input log power spectrum and $c(n)$ is the output. The constants μ_a and μ_b are onset and offset time constants, respectively. The T_s is the step size.

From this equation, we can see that the forward masking procedure at each critical band acts as a non-linear filter with different onset and offset time constants. It extracts comparatively lower frequency components of modulation spectrum.

The output of the LFM procedure is scaled by a factor of 0.33 and exponentiated. Although this performs a similar function to the logarithm of the standard MFCC procedure, our experiments (not described here) show a slight performance increase when the 0.33 compression is used.

2.2. 2-D Cepstrum

The 2-D cepstrum is the Fourier transform of the time trajectory of cepstrum. Kandedera[8] indicated that some components of the 2-D cepstrum, especially around 4Hz, are more important for recognition of speech than others. There are two kinds of 2-D cepstrum. The FFT based 2-D cepstrum keeps the real and imaginary parts of the transform, whereas the DCT based only uses resulting absolute amplitudes. In our experiments, we used the frequency component at 4.88Hz. The Δ 2-D cepstrum was calculated by obtaining the delta values of components of 2-D cepstrum in two adjacent frames.

3. EXPERIMENTAL RESULTS

In section 3.1, optimal onset and offset time constant for our forward masking procedure were obtained through TI isolated digits recognition experiments. In section 3.2, we show that FM procedure and the calculation procedure of 2-D cepstrum extract orthogonal information in some sense. Section 3.3 describes the performance of LFM MFCC combined with 2-D cepstrum and Δ 2-D cepstrum. In section 3.4, we show that the clean speech performance of the proposed features is comparable with that achieved by MFCC based features.

In these experiments, the TI-46 connected digits database was used. Digits models and background noise model were trained on clean speech utterances using HMM toolkit (HTK) adopting a flat start approach. Contaminated speech for test was generated by artificially adding different levels of Volvo noise [1] to the clean speech.

For the connected digits recognition experiments, 500 connected digits utterances from 15 speakers were used for training, and 100 connected digits utterances from 4 speakers were used for testing. For the isolated digits recognition experiments , 341 isolated digits utterances from 31 speakers were used for training, and 110 isolated digits utterances from 10 speakers were used for testing. In both cases, the testing and training speakers were disjoint.

Table 1: Feature used in the isolated and connected TI digits recognition

No.	Feature Name	Vector Size	
		1st	2nd
1	Liftered MFCC	11	
2	LFM MFCC	11	
3	2-D cep	22	
4	MFCC + Δ MFCC	11	11
5	MFCC + ΔMFCC + $\Delta\Delta$MFCC	11	11 + 11
6	MFCC + LFM MFCC	11	11
7	MFCC + 2-D cep	11	22
8	LFM MFCC + 2-D cep	11	22
9	LFM MFCC + 2-D cep + Δ 2-D cep	11	11 + 11

There were 12 whole word models for 11 digits (zero is pronounced as oh or zero) and one silence model. Each digit was modeled by a 10 states (including a nonemitting initial and final state) left to right HMM without skip states. The silence model had the same model structure as the digit models. For the experiments in 3.1 and 3.2, single stream single mixture HMM models were used. In experiments of 3.3.1, 3.3.2, and 3.4, we used two stream models. The first stream of these models had a two Gaussian mixtures with stream weight of 1.0, and the second stream had four Gaussian mixtures with stream weight of 0.8. Diagonal covariance matrices were used in all experiments.

Speech signals were sampled at a rate of 20kHz. Window size was 25.6ms with a step size of 12.8ms. Fifty six (56) filters were used in the binning stage for both the Liftered Forward Masking procedure and liftered MFCC.

For both MFCC and LFM MFCC, 10 cepstral coefficients (C1 through C10) were computed. The 11th feature was the delta log energy or forward masked delta log energy. Since we return both the real and imaginary parts of the 2-D cepstrum, these 11 coefficients result in 22 2-D cepstrum coefficients. In order to keep the size of the parameter vector constant, when computing the 2-D cepstrum plus Δ 2-D cepstrum, we keep the real and imaginary parts of the FFT of the first 5 cepstral coefficients. The 11th coefficient of the Δ 2-D cepstrum is the amplitude of the FFT of the 6th cepstral coefficient.

Features, their combinations and their feature vector size in different streams are shown in table 1. For convenience, in some of the latter figures representing experimental results, feature index instead of feature name is used.

3.1. Performance of the LFM MFCC

Different sets of onset and offset time constants have different effects on the performance of LFM MFCC. In our experiments, the onset and offset coefficients for each critical band were the same. In this paper, apart from our own forward masking parameters, we also used data from [2]. Parameters from [2] are based on some physiological experiments to get release and attack time constants for an auditory procedure. We used the release and attack time constants [2] at the frequency of 1,000Hz, which are 16.0ms for μ_a and 49.0ms for μ_b, respectively. We also set our own onset and offset time constants to 54.5ms and 17.5ms, respectively, by comparative recognition experiments.

Word error rates of these two sets of parameters are shown in Figure 2 along with the word error rate of using MFCC. Due

to the recursive nature of the forward masking procedure, the initial value of $c(t)$ also influences the system performance. In this paper, we set the initial value of $c(t)$ to zero.

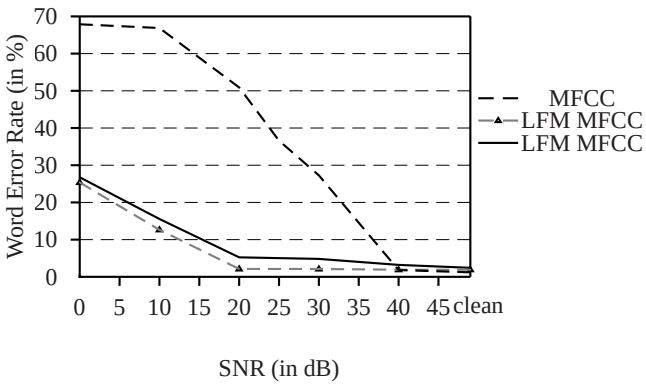

Figure 2: Performance of LFM MFCC using different parameters. The black line shows the performance of the LFM MFCC with μ_a = 16.0ms and μ_b = 49.0ms; The gray triangle line shows performance with μ_a = 54.5ms and μ_b = 17.5ms.

One observation from Figure 2 is that LFM MFCC using our parameters perform better than the LFM MFCC using parameters from [2], in both clean and noisy conditions. As further seen in this figure, in clean environments, the error rate of LFM MFCC using either set of parameters is higher than that of MFCC. We also note that performance of MFCC starts to degrade below a SNR of 35 dB.

3.2. Effects of LFM Procedure on 2-D Cepstrum

In these isolated digits recognition experiments, we compared the performance of the 2D cepstrum computed from MFCC's with and without forward masking. Figure 3 shows, although adding forward masking improves performance of liftered MFCC coefficients, the adding forward masking procedure when computing the 2-D cepstrum actually decreases performance. It appears that some of the information extracted by the 2-D cepstrum that is important for recognition is actually suppressed by the forward masking procedure. Thus, in some sense the information extracted in the LFM MFCC are orthogonal to that extracted by the 2-D cepstrum.

As further seen in Figure 2 and 3, the 2-D cepstrum alone is more robust than the LFM MFCC alone in additive Volvo noise conditions. However its performance in clean conditions is not as good as that of LFM MFCC.

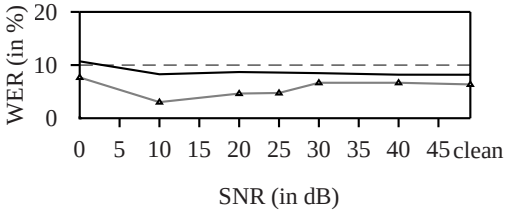

Figure 3: Word Error Rate of 2-D cepstrum calculated from liftered MFCC or from LFM MFCC. Black line represents the 2-D cepstrum calculated from LFM MFCC; Gray triangle line represents 2-D cepstrum calculated from liftered MFCC.

3.3. Experiments on Combination of LFM MFCC with 2-D Cepstrum

Because the LFM MFCC and 2-D cepstrum contain different information about the speech signal, but both yield improved accuracy when compared with static MFCC coefficients, it seems likely that combining these two features may result in further improved performance. In this section, we describe experiments which support this belief. The two features were combined by assigning LFM MFCC to one stream and the 2-D cepstrum or 2-D cepstrum plus Δ 2-D cepstrum to another stream. For the aim of comparison, several other combinations were also tested.

In section 3.3.1, performance of different combinations of speech features were compared through isolated digit recognition experiments. The combinations with the two highest recognition rates in the isolated digits recognition experiments were chosen for the connected digits recognition experiments in section 3.3.2.

3.3.1. Isolated Digits Recognition Results

The isolated digits recognition experiments show that the combination of LFM MFCC with 2-D cepstrum does improve robustness when compared with standard MFCC combined with its dynamic and acceleration coefficients. From Figure 4, we can see that the SNR gain of the combination of LFM MFCC with 2-D cepstrum (Feature 8) is above 30dB for a word error rate below 10%, compared with MFCC plus Δ MFCC plus $\Delta\Delta$ MFCC (Feature 5). One observation from Figure 4 is that the acceleration parameter, Δ 2-D cepstrum, could slightly further improve system performance in noisy conditions. A further observation from this isolated digits recognition experiment is that the combination of MFCC with LFM MFCC (Feature 6) is more robust than the combination of MFCC with 2-D cepstrum (Feature 7) in additive Volvo noise conditions.

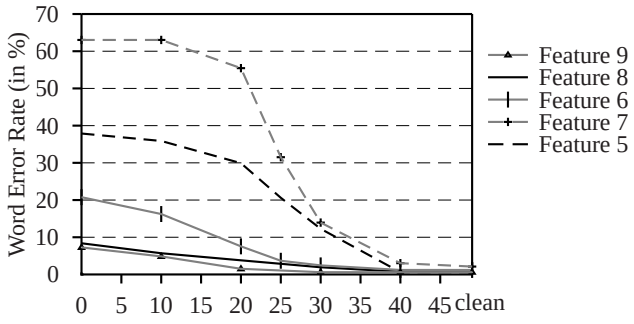

Figure 4: Results on Isolated Digits Recognition Experiments

3.3.2. Connected Digits Recognition Results

Three kinds of speech features were tested for connected digits recognition. They were MFCC plus Δ MFCC plus $\Delta\Delta$ MFCC, LFM MFCC plus 2-D cepstrum, and LFM MFCC plus 2-D cepstrum plus Δ 2-D cepstrum. Word error rates are shown in Figure 5.

From Figure 5, we can see that the SNR gain of LFM MFCC plus 2-D cepstrum is above 30dB for a word error rate below 20%, compared with MFCC plus Δ MFCC plus $\Delta\Delta$ MFCC. As further seen in this figure, the acceleration parameter, Δ 2-D cepstrum, could improve system performance to above 90% in 0dB additive Volvo noise condition.

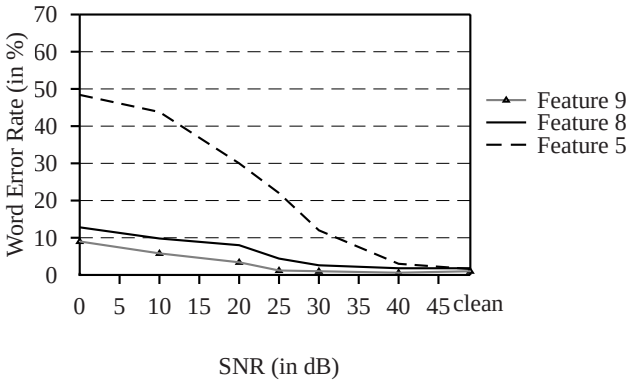

Figure 5: Results on Connected Digits Recognition Experiments

3.4. Performance in Clean Environments

Performance of different combinations of features in clean environments are shown in table 2. They were tested in isolated digits recognition experiments. We can see that, in clean conditions, the combination of MFCC with Δ MFCC performs better than that of combination of MFCC with either 2-D cepstrum or LFM MFCC. However, as further seen in this table, the combination of the two dynamic features, LFM MFCC and 2-D cepstrum, could make system performance comparable to that of MFCC plus Δ MFCC in clean conditions. We also note that the combination of two dynamic features with one acceleration feature, LFM MFCC with 2-D cepstrum plus Δ 2-D cepstrum, could maintain a system performance comparable to that of MFCC combined with its dynamic and acceleration, MFCC plus Δ MFCC plus $\Delta\Delta$ MFCC, in clean conditions. The forward masking parameters in this experiment were μ_a = 54.5ms and μ_b = 17.5ms.

Table 2: Performance (Word Error Rate) of combined features in clean conditions. S stands for static features, D stands for dynamic features, and A stands for acceleration features. WER is in %.

Comb.	Feature Name	WER
S+D	MFCC + Δ MFCC	0.61
	MFCC + LFM MFCC	1.21
	MFCC + 2-D Cep	2.12
D+D	LFM MFCC + 2-D Cep	0.65
S+D+A	MFCC + Δ MFCC + $\Delta\Delta$ MFCC	0.61
D+D+A	LFM MFCC + 2-D Cep + Δ 2-D Cep	0.61

4. CONCLUSION

The combination of two dynamic features, LFM MFCC and the 2-D cepstrum, leads to speech recognizers which are more robust to additive Volvo noise than standard MFCC features, while maintaining recognition accuracies nearly identical to those achieved by standard MFCC features in clean matched conditions. Through experiments, we found that the information extracted by forward masking is in some sense orthogonal to that extracted by the 2-D cepstrum, which may explain why system performance improves when they are combined. For a word recognition rate of 90%, the SNR gains on isolated and connected digits recognition experiments in additive Volvo noise for the LFM MFCC plus 2-D cepstrum plus Δ 2-D cepstrum are over 30dB compared with MFCC plus Δ MFCC plus $\Delta\Delta$ MFCC.

5. ACKNOWLEDGMENTS

This work was supported by the Hong Kong Research Grants Council under grant number HKUST CA97/98.EG02.

6. REFERENCES

[1] Signal Processing Information Base. Volvo noise. http://spib.rice.edu/cgi-bin/spib-bin/prog1.

[2] B.Strope and A.Alwan. A model of dynamic auditory perception and its application to robust word recognition. *IEEE Trans. Speech and Audio Proc.*, 5(5):451–464, 1997.

[3] B.Strope and A.Alwan. Robust word recognition using threaded spectral peaks. In *ICASSP*, pages 625–628, 1998.

[4] Brian A. Hanson and Ted H. Applebaum. Robust speaker–independent word recognition using static, dynamic and acceleration features: Experiments with lombard and noisy speech. In *ICASSP*, pages 857–860, 1990.

[5] H. Hermansky. Perceptual linear predictive (plp) analysis for speech. *J.Acoust.Soc.Amer.*, 87(4):1738–1752, 1990.

[6] H.Hermansky and N.Morgan. Rasta processing of speech. *IEEE Trans. Speech and Audio Process*, 2(4):578–589, 1994.

[7] Jean-Claude Junqua and Jean-Paul Haton. *Robustness in Automatic Speech Recognition*. Kluwer Academic Publisher, 1996.

[8] Noboru Kanedera, Hynek Hermansky, and Takayuki Arai. On properties of modulation spectrum for robust automatic speech recognition. In *ICASSP*, pages 613–616, 1998.

[9] S.Furui. Speaker–independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. Acoust. Speech Signal Process*, pages 52–59, 1986.