

# The Formant-emphasized Feature Vector for Speech Recognition in Noisy Condition

Young-Woo Son

Dept. of Electronic Engineering, Kyungpook  
National University  
1370 Sankyuk-dong, Buk-gu, Daegu 702-701,  
South Korea

Jae-Keun Hong

Dept. of Electronic Engineering, Kyungpook  
National University  
1370 Sankyuk-dong, Buk-gu, Daegu 702-701,  
South Korea

*Abstract* - Mel-frequency cepstral coefficients are widely used as the feature for speech recognition. In MFCC extraction process, the spectrum, obtained by Fourier transform of input speech signal is divided by mel-frequency bands, and each band energy is extracted for the each frequency band. The coefficients are extracted by the discrete cosine transform of the obtained band energy. In this paper, we calculate the output energy for each bandpass filter by taking the weighting function when applying mel-frequency scaled bandpass filter. The weighting function is Gaussian distributed function whose center is at the formant frequency. In the experiments, we can see the comparative performance with the standard MFCC in clean condition, and the better performance in worse condition using method we proposed.

**Keywords:** MFCC, Filter bank, Formant, Gaussian Distribution, Speech Recognition

## 1. Introduction

The importance of the noise robust recognition system is increasing, and a lot of research has been done. In the presence of noise, the accuracy and robustness of speech representation deteriorates dramatically, which makes serious spectral mismatch between the training and testing data. To alleviate this problem, many robust speech recognition techniques have been developed by many researchers. These techniques for the noise robust recognition are generally classified into three categories. The three categories are noise robust speech feature[1], speech enhancement[2], and model compensation[3]. Among the categories, the spectral compression is used for noise robust feature, and expressed as (1).

$$\hat{P}(k) = P(k)^\alpha, \quad 0 \leq \alpha \leq 1 \quad (1)$$

where  $\hat{P}(k)$  is the compressed speech power spectrum,  $P(k)$  is the original speech power spectrum, and  $\alpha$  is the compression factor.  $k$  is filter band index. As  $\alpha$  is smaller, the mismatch or variation caused by noise is much reduced and at the same time considerable amount of information is lost. Thus the spectral compression technique is a trade-off between information and pattern mismatch. The research[4][5] using a constant root  $\alpha$  has been carried and the research[6] for obtaining the function  $\alpha(k)$  according to SNR has been carried using spectral compression, and this is expressed as below (2).

$$\hat{P}(k) = P(k)^{\alpha(k)}, \quad 0 \leq \alpha(k) \leq 1 \quad (2)$$

where  $\alpha(k) = A \exp(-\lambda k) + A_0$ .  $A$  and  $A_0$  are used for restricting the dynamic range of the compression factor. The degree of spectral compression is determined by  $\alpha(k)$ . As  $\alpha(k)$  decreases, the effect of noise decreases, but the loss of speech information increases. Therefore the decision of  $\alpha(k)$  is very important process in the spectral compression. As the result, the part that has important information is needed to be emphasized without the loss of information for noisy speech recognition.

This paper also focuses on the robust speech feature extraction approach. In speech recognition, the vowel of speech has more information than consonant. Especially, consonants could be ignored in low SNR environments. Therefore information of vowel is an important thing in noisy environments. Among the properties of vowel, the effect of formants that have appreciable tendency even in noisy speech is crucial. We tried to emphasize the formants. Therefore, we use the weighting function of Gaussian distribution for increasing the effect of formants and decreasing the effect of high frequency that has no formants. The weighting function is Gaussian distributed with the mean which is the frequency band having formants, differently from the standard MFCC filter bank that has all same frequency response filters.

## 2. Formant-emphasized Filterbanks

### 2.1 Energy calculation of each band

In standard MFCC extraction process, each band energy is calculated with bandpass filtering power spectrum, and expressed as (3).

$$E(k) = \sum_i W_k(i)P(i) \quad (3)$$

where  $E(k)$  is the output energy for  $k$  th filter and  $W_k(i)$  is  $i$  th weighting value for  $k$  th filter, and  $P(i)$  is the speech power spectrum.  $W_k(i)$ , used for weighting in (3) is the triangular filter that has mel-frequency property. Figure 1 shows the triangular filter.

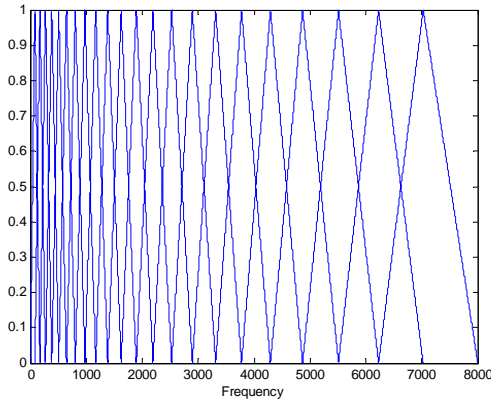


Figure 1. Triangular bandpass filters

In noisy condition that has higher frequency than speech has, the property of speech is decreasing respectively. Therefore, the property of speech is needed to be emphasized or de-emphasized in MFCC extraction process for speech recognition. The property of speech could be emphasized applying weighting function to frequency band filter in (3).

$$E(k) = \sum_i \alpha(k)W_k(i)P(i) \quad (4)$$

where  $\alpha(k)$  is the weighting function, and expressed in (5).

### 2.2 The bandpass filters with Gaussian distribution based on formant band

$$\alpha(k) = \sum_{n=1}^m \frac{\exp\left[-\frac{(x_{n,k} - \mu_n)^2}{2\sigma^2}\right]}{\sqrt{2\pi}\sigma} + \beta \quad (5)$$

The weighting function is the Gaussian distribution function and  $k$  is the position of bandpass filter, and  $\mu_n$  is the position of bandpass filter that has the  $n$  th formant.  $x_{n,k}$  is the position of  $k$  th bandpass filter centering the filter band that has  $n$  th formant.

$\sigma^2$  is the deviation of Gaussian distribution,  $\beta$  determines the dynamic range. Also  $m$  is the number of formants. The example of modified filterbank using equations (4) and (5) shows in figure 2. where the deviation,  $\sigma$  is set by 4 and the number of formants,  $m$  is set by 2. As figure 2, the weighting value of each band decreases as the each band becomes more distant from the band that has formant.

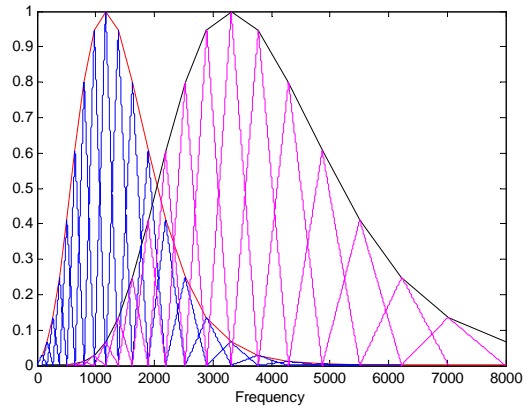


Figure 2. Example of bandpass filters with weighting function of Gaussian distribution

To calculate the weighting in speech region, the crisis filter band that has formants should be obtained. There are various methods[7] to find formants. However, because it's not necessary to search for the exact position of formants, but necessary to search for the position of filter bands that have formants, the peak-searching method from speech spectrum is used to search for formants instead of the exact formant searching method with LPC. The weighting of filter band beside to filter band that has formant is gained by (5). The deviation of the Gaussian distribution is important factor when calculating the weighting that has Gaussian distribution. Because if the deviation is too small, the weighting function has effect on only frequency band where formants exist, so the spectrum is too distorted in the frequency band where formants do not exist, and if the deviation is too large, the weighting function has effect on all the frequency bands, so the weighting function does not make any difference. These Gaussian distribution weighting function makes the weighting large in the low frequency band, and the weighting small in the high frequency band that does not have formants, so the effect of noise decreases.

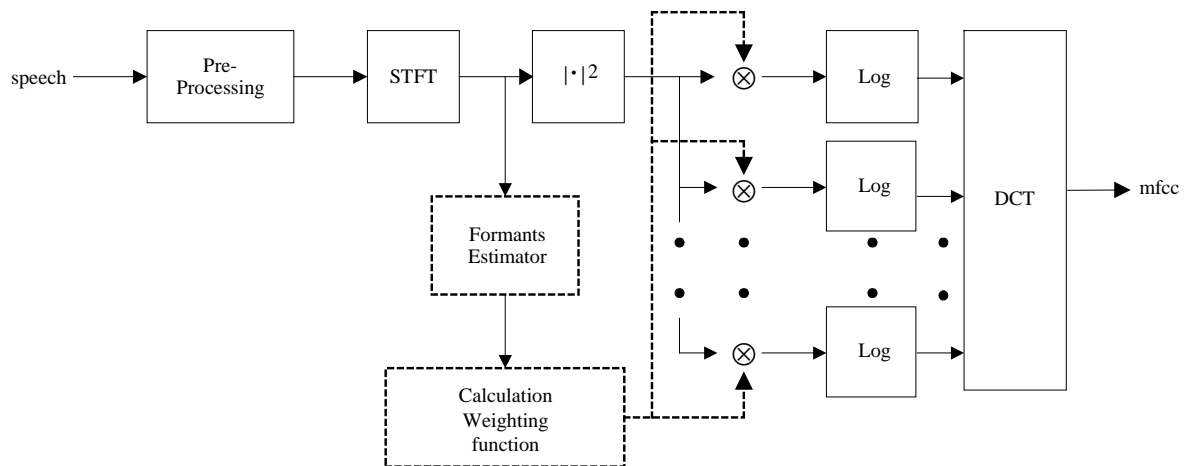


Figure 3. Feature extraction using weighting function

### 3. Experiment and Results

Figure 3 shows the total block diagram of extraction process using the weighting function. Bold dashed line shows the proposed method. Formants are searched after the preprocessing and FFT transform of input speech signal. Figure 4 shows the magnitude spectrums and formants of clean speech and noisy speech with about 10dB white Gaussian noise. The vertical line shows the position of formant frequencies by using the spectrum.

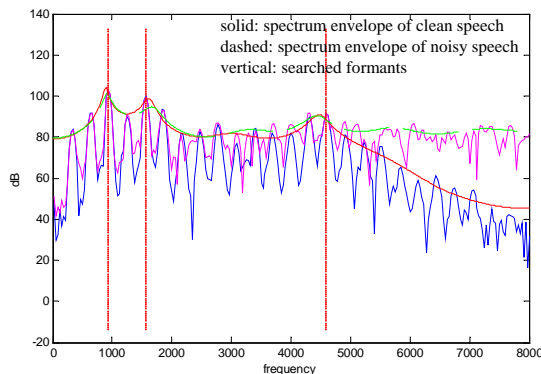


Figure 4. Formants and frequency responses of speech according to SNR

As you see figure 4, the loss of information could happen due to wrong formants because it is very difficult to search for the formants of noisy speech. Therefore, the number of formants need to be restricted. In the experiments, we used below 3 formants. The weighting factor is calculated using the weighting function of Gaussian distribution that is expressed in (5). Where  $\sigma$  is set 4 and  $\beta$  is set 0.2.

To restrict the dynamic range of the weighting function we excluded the denominator of Gaussian distribution.  $\beta$  is the lowest value of the weighting function. If  $\beta$  approaches to the maximum of the dynamic range, the better recognition performance could not be expected because of no decline of the high frequency components that have much noise, as the filter property in the frequency band that is not affected by the Gaussian distribution function is becoming almost same to the filter property in the standard MFCC filter bank. If  $\beta$  approaches to the minimum of the dynamic range, the information of formant band and its neighboring bands just remain and the rest almost disappear. Therefore, it is crucial to determine the reasonable  $\beta$ . The experiments are done with MFCC modified by the weighting from Gaussian distribution weighting function for each band, after formants are searched with the method above, under the experimental circumstances below

- 1) Database: 445DB provided by ETRI(Electronics and Telecommunications Research Institute) in Korea
- 2) Sampling rate: 16kHz
- 3) Pre-processing factor: 0.97
- 4) Window function: hamming window
- 5) Frame rate: 20ms frame, 10ms shift
- 6) Recognizer: HTK ver. 3.2.1[8]
- 7) Noise: white Gaussian noise

Table 1. Comparison of recognition rates of conventional MFCC and the proposed method

	Speaker													
	LYI		CJD		LKS		HKW		BJS		SUN		Average	
	MFCC	Proposed method	MFCC	Proposed method	MFCC	Proposed method	MFCC	Proposed method	MFCC	Proposed method	MFCC	Proposed method	MFCC	Proposed method
Clean	97.42	96.74	93.48	93.37	96.63	96.07	95.84	95.84	96.39	95.83	96.61	96.50	96.06	95.73
30dB	92.47	95.17	88.99	89.44	94.61	95.06	93.60	93.93	92.45	94.70	93.12	95.37	92.54	93.95
25dB	86.63	92.02	72.58	83.03	89.00	92.25	85.73	90.34	84.89	90.76	87.92	94.24	84.46	90.44
20dB	66.18	86.18	36.40	65.84	68.43	84.64	62.25	86.18	65.84	86.25	73.70	92.21	62.13	83.55
15dB	30.45	67.30	10.67	36.74	26.52	69.10	21.80	70.79	36.87	70.24	46.39	85.44	28.78	66.60
10dB	10.34	39.44	2.81	14.72	5.39	28.43	4.72	33.15	12.74	48.48	15.80	65.91	8.63	38.36
Average	63.92	79.48	50.82	63.86	63.43	77.59	60.66	78.37	64.86	81.04	68.92	88.28	62.10	<b>78.10</b>

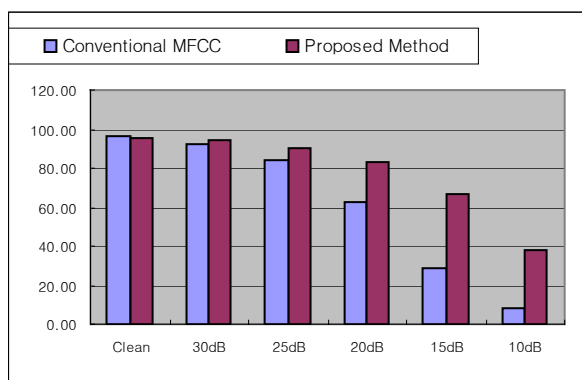


Figure 5. Comparison of recognition rates of conventional MFCC and the proposed method

Table 1 and figure 5 represents MFCC filter bank modified by the Gaussian weighting function whose mean is the formant frequency band according to SNR shows generally better recognition performance than the standard MFCC under HTK recognition system.

#### 4. Conclusion

The techniques for noise robust recognition have been developed in the part of MFCC extraction as a speech feature vector. In this, we proposed the new MFCC to obtain the noise robust speech feature as emphasizing the formant components that is less affected by noise. To emphasize formant components, we apply the weighting to the each triangular filter that has formant in the standard MFCC filter bank. The weighting for each band is calculated by the crisis value according to SNR and the Gaussian distribution function whose mean is at the formant frequency band. New MFCC with the band filters modified by weighting is obtained. The proposed method generally shows better recognition performance. If the formants are found more correctly in noisy condition, much

better recognition rate would be shown. Also the experiments with various speech database and noise are needed.

#### References

- [1] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am* 87, pp. 1738-1752, April 1990.
- [2] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection for robust speech recognition in cars", *Speech Communication*, vol.11, pp. 215-228, June 1992.
- [3] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition in noise", *Speech Communication*, vol.12, pp. 231-239, 1993.
- [4] K. K. Chu, S. H. Leung and C. S. Yip, "Perceptually non-uniform spectral compression for noisy speech recognition", *Proc. ICASSP 2003*, pp. 404-407, 2003.
- [5] K. K. Chu, S. H. Leung, "Feature extraction based on perceptually non-uniform spectral compression for speech recognition", *Proc. ISCAP 2003*, pp. 726-729, 2003.
- [6] K. K. Chu and S. H. Leung, "SNR-dependent non-uniform spectral compression for noisy speech recognition", *Proc. ICASSP 2004*, pp. 973-976, 2004.
- [7] L. Welling and H. Ney, "Formant estimation for speech recognition", *IEEE Trans. On Speech and Audio Processing*, vol. 6, no. 1, Jan. 1998.
- [8] S. Young, D. Kershaw, J. Odell, D. Ollason, and P. Woodland, *The HTK Book version 3.2.1*, 2002.