TR－H－006

Word Recognition Using Auditory Model Front-End Incorporating Spectro-Temporal Masking

Kazuaki OBARA　Kiyoaki AIKAWA
Hideki KAWAHARA

1993. 3. 31

ATR人間情報通信研究所

〒619-02 京都府相楽郡精華町光台2-2　☎07749-5-1011

ATR Human Information Processing Research Laboratories

2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

Telephone: +81-7749-5-1011
Facsimile: +81-7749-5-1008

Word Recognition Using Auditory Model Front-End Incorporating Spectro-Temporal Masking

Kazuaki OBARA, Kiyoaki AIKAWA , and Hideki KAWAHARA
ATR Human Information Processing Research Laboratories
ATR Auditory and Visual Perception Research Laboratories
2-2 Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
email: obara@atr-hr.atr.co.jp

Abstract:

An auditory model front-end that reflects spectro-temporal masking characteristics is proposed. The model gives an excellent performance in the multi-speaker word recognition system. Recent auditory perception research shows that the forward masking pattern becomes more wide-spread over the frequency axis as the masker-signal interval increases. This spectro-temporal masking characteristics is modeled and implemented into the cochlear filter front-end for speech recognition. The current masking level is calculated as the weighted sum of the smoothed preceding spectra. The weight values become smaller and the smoothing window size becomes wider on the frequency axis as the masker-signal interval increases. The current masked spectrum is obtained by subtracting the masking levels from the current spectrum. Word recognition experiments demonstrated that the recognition performance is improved by incorporating the masking effect into the cochlear filter front-end. The performance was better than that with traditional LPC-based word recognizers.

1. Introduction

This paper improves the cochlear filter for speech recognition by implementing a spectro-temporal masking effect. There have been many attempts to implement auditory models into speech recognition. The major reason for this is to represent a speech spectral more precisely. We proposed an adaptive Q cochlear filter models. The adaptive Q cochlear filter is a non-linear filter that simulates the asymmetrical and power level dependent filtering of the basilar membrane. The fixed Q cochlear filter, on the other hand, is a reduced version of the adaptive Q cochlear filter. It only simulates asymmetrical filtering of basilar membrane. We showed that the adaptive Q cochlear filter combined with a lateral inhibition performs well in both noisy and reverberant environments. However the system performance was poor for unknown speakers [Obara, K., 1991a, b].

Recent auditory perception research has shown that the forward masking pattern becomes more wide-spread over the frequency axis as the masker-signal interval increases [Miyasaka, E., 1983]. This spectro-temporal masking characteristic is considered to be effective for eliminating the speaker-dependent spectral tilt that reflects individual source variations, and for enhancing the spectral dynamics that convey phonological information in speech signals. We implement this spectro-temporal masking effect into the cochlear filter. In this study, only the forward masking effect was taken into account because it might be more prominent than backward masking.

2. Forward Masking Model

The spectro-temporal masking is modeled so as to simulate two essential characteristics of the forward masking effects with increasing masker-signal interval: The exponential decay of the masking level and the smoothing of the masking pattern.

The masked spectrum, P(u, v), which corresponds to the perceived

spectrum, is modeled to be the current spectrum, S(u, v), and the current masking level, M(u, v), as,

$$P(u,v) = \mathrm{Max}\{\ \{S(u,v) - M(u,v)\}\ ,\ 0.0\ \} \quad (1)$$

$$M(u,v) = A(u,v) + D(v) \quad (2)$$

Here, u and v represent the channel number and frame number of the spectrum. The current masking level, M(u, v), is composed of two components, A(u, v) and D(v). A(u, v) represents the weighted sum of the power-normalized preceding spectra $S_N(u, v)$, and D(v) represents the weighted sum of the average power $S_{AV}(v)$ of the preceding spectra.

A(u, v) is calculated as,

$$A(u,v) = \sum_{k=1}^{K}\left[\sum_{j=-N-k}^{+N+k}\left(W(j,k)\cdot S_N(u+j, v-k)\right)\right] \quad (3)$$

$$W(j,k) = \left\{0.54 + 0.46\cdot\cos\left(\frac{p\cdot j}{(N+k)}\right)\right\}\cdot\alpha\cdot\beta^{(k-1)} \quad (4)$$

W(j, k) is the Hamming window for smoothing a spectrum, where the window width gets wider and the weight decreases exponentially as a function of the interval between the current and the preceding spectra. K limits the duration where the masking effect goes on. N is a initial width of the Hamming window. α is a initial weight for power normalized spectrum and β is a decay rate of the weight.

The D(v) is calculated as,

$$D(v) = \sum_{k=1}^{K}\left(g\cdot d^{(k-1)}\cdot S_{AV}(v-k)\right) \quad (5)$$

The $S_{AV}(v-k)$ represents the average power of the frame (v-k). g is a initial weight for the average power and d is a decay rate of the weight.

3. Experiments

3.1 Experimental conditions

Word recognition experiments were performed for 216 phonetically balanced Japanese words uttered by 10 male and 10 female speakers. The sampling rate was 12 kHz. The block diagram of the speech recognition system is shown in Fig. 1. We tested two types of cochlear filters. One was the adaptive Q cochlear filter and the other was the fixed Q cochlear filter. The cochlear filters were compared with a Bark scale band pass filter based on DFT. In the DFT front-end, spectra were obtained from 1024 point FFTs every 10 ms. Then, a 512-channel DFT power spectrum was transformed into 64-channel Bark scale coefficients. The speech spectrum was obtained every 10 ms as the logarithmic power of the filter outputs. All these three filters cover the frequency range from 1.5 to 18.5 Barks with 64 channels. The speech spectrum was obtained as the logarithmic power of the filter outputs. Then, the outputs of the one of the filters was sent to the masking model. The spectral sequence of produced by one of the filters was sent to a Dynamic Time Warping(DTW) word recognizer.

3.2 Effect of the masking model parameters

The effectiveness of the masking model was investigated using the utterances of 2 male speakers. One speaker's utterance served as the template and the other speaker's utterance served as the testing speech. Feature extractions were made by the Fixed Q cochlear filter with the masking model. Several combinations of masking weight, a and g were tested. The results are shown in Fig. 2. The original recognition performance without the masking model is labeled "baseline". The word recognition rate was improved to 97.2% from 93.1%, when α=0.25 and γ =0.5.

3.3 Multi-speaker word recognition

Word recognition experiments were conducted using the best model parameters from the preliminary experiments. The experiments were repeated rotating the template speaker. In each experiment, one speaker's utterance served as the template, while the utterances of the other 9 speakers served as the testing speech. Fig. 3 shows the experimental results. In the figure, each bar represents the average recognition rate of 9 speakers. The light bars represent recognition rates without masking and the dark bars represent recognition rates with the proposed masking model. With spectro-temporal masking, the recognition rate was improved to 87.7% from 82.7% , on average. A statistical $\chi 2$ test demonstrated that the improvement in the recognition performance by introducing the spectro-temporal masking model is statistically significant for all speakers. Fig. 4 shows the experimental results for the female speakers. The recognition rate improved about 5% from 75.5% to 80.7%.

3.4 Comparison with the LPC parameters

The cochlear-filter-based front-ends were compared with LPC-based front-ends for the recognition of an unknown speaker's voice. A 16th-order LPC cepstrum was extracted using linear predictive coding. Three types of LPC-based spectral parameters, cepstrum, weighted cepstrum and dynamic cepstrum, were compared with filter-based parameters. The dynamic cepstrum which enhances spectral dynamics was proposed by K. Aikawa [Aikawa, K., 1992]. The results are shown in Fig. 5. Although the parameter sizes are not the same, the cochlear-filter-based spectra exhibits better performance than that of the LPC-based spectra. The best recognition performance was obtained with the adaptive Q cochlear filter implemented with the spectro-temporal masking model.

3.5 Masking model in noisy environment

Fig. 6 shows the effectiveness of the masking model in noisy environments. The spectro-temporal masking model seems to be effective in noisy environments, because it enhances the spectral dynamics of the speech conveying phonological features. As a result, it can enhance the signal-to-noise ratio of the speech. In these experiments, the utterances of 4 speakers, 2 male and 2 female, were used. Each speaker uttered the same set of words twice. The first utterance served as the template and the second utterance distorted by additive pink noise was recognized. The average performance of the four speakers are shown. Three front-ends, the adaptive Q cochlear filter, the fixed Q cochlear filter and the Bark scale DFT, were compared with and without the masking model. The dashed lines represent recognition rates of the original three filters without the masking model, and the solid lines represent recognition rates of the three filters with the masking model. This figure shows that the recognition performance is significantly improved by implementing the spectro-temporal masking effect.

4. Conclusion

In this study, we have proposed the cochlear filter front-end that incorporates the forward masking characteristics and evaluated its effectiveness in DTW word recognition system. The best recognition performance were obtained with the Adaptive Q cochlear filters implementing the masking model. The recognition rate in multi-speaker word recognition system improves about 5% from 83.4 % to 87.5 % by implementing the masking model. In the noisy conditions, the Adaptive Q cochlear filter that incorporates the masking model also improves the recognition performance to 96% from 83%(S/N=10dB).

Acknowledgement:

The authors would like to thank Drs. Y. Tohkura, president of ATR Human Information Processing Research Laboratories and T. Hirahara of NTT Basic Research Laboratories for their helpful comments.

References:

Aikawa, K., Kawahara, H. and Tohkura, Y.(1992): "Dymanic Cepstrum parameter Inporporating Time-Frequency Masking and It's Application to Speech Recognition," J. Acoust. Soc. Am., Vol. 92, No. 4, Pt. 2, pp.2476, 5pSP5.

Miyasaka, E. (1983): "Spatio-temporal characteristics of masking of brief test-tone pulses by a tone-burst with abrupt switching transients," Journal of the Acoustical Society of Japan, Vol. 39,No. 9, pp.614-623.

Ghiza, O. (1988): "Temporal non-place information in the auditory nerve firing patterns as a front-end for speech recognition in a noisy environment," Journal of Phonetics, Vol.16, No.1 pp.109-123.

Hirahara, T. and Iwamida, H. (1990): "Auditory spectrograms in HMM phoneme recognition," Int. Conf.on Spoken Language Processing, ICSLP-90, pp. 381-384.

Obara, K.and Hirahara, T (1991a): "Auditory Front-end in DTW Word Recognition under noisy, reverberant and multi-speaker conditions," J. Accoust. Soc. Am., Vol. 90, No. 4, Pt. 2, 3SP11, pp. 2274.

Obara, K.and Hirahara, T (1991b): "Evaluation of Auditory Front-end in DTW Word Recognition," IEICE Technical Report, SP91-81, pp. 1-8.

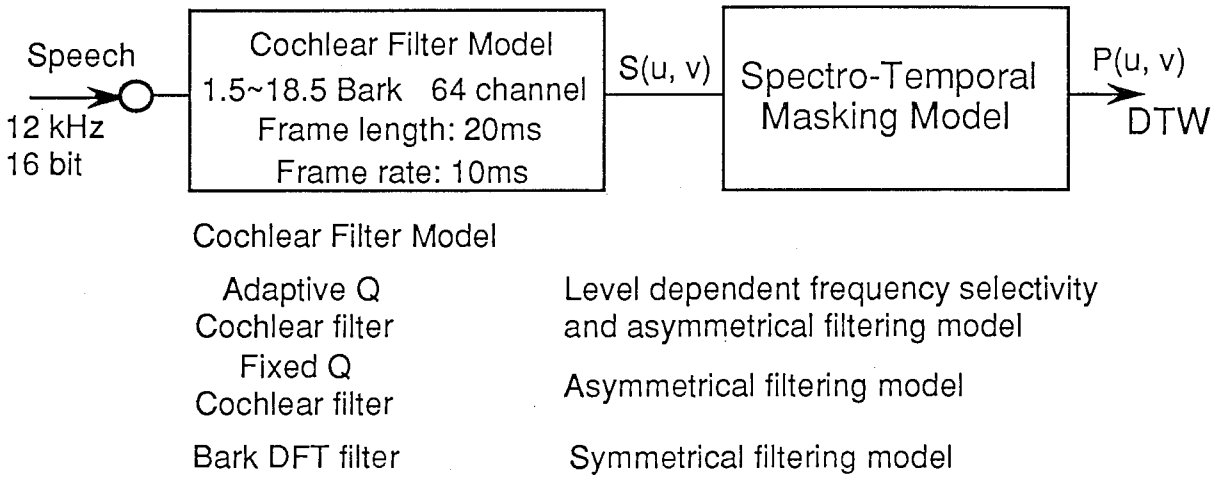

Fig. 1 Block diagram of the speech recognition system.

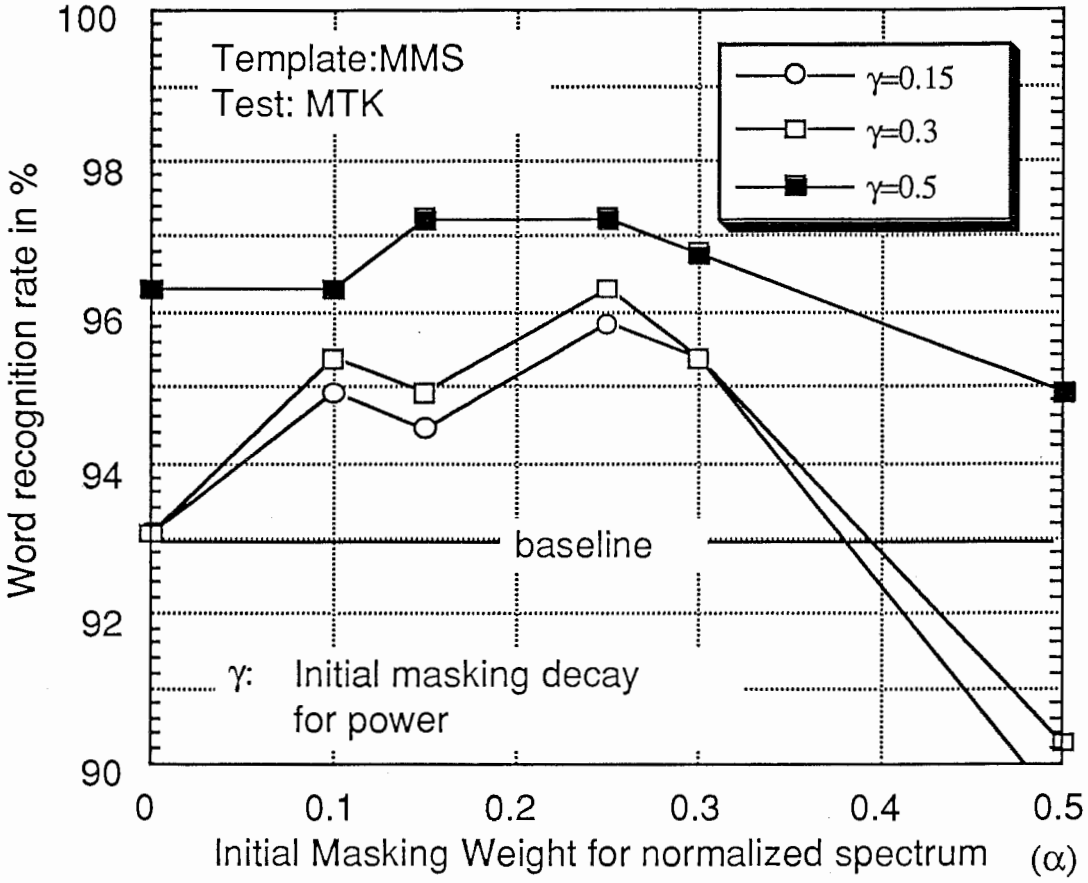

Fig. 2 Effect of the masking parameters.

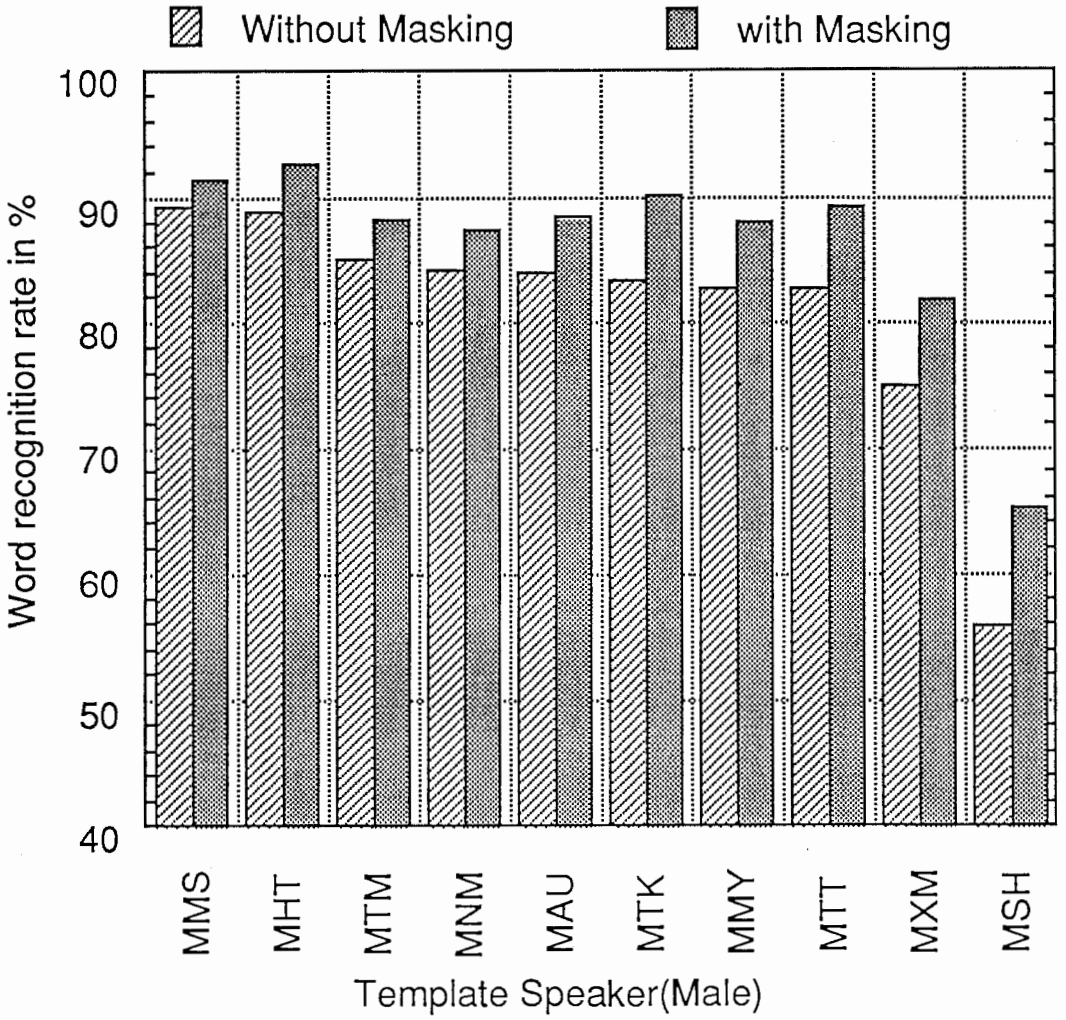

Fig. 3 Multi-speaker word recognition rate of male speakers. Each bar represents average recognition rate of 9 speakers.

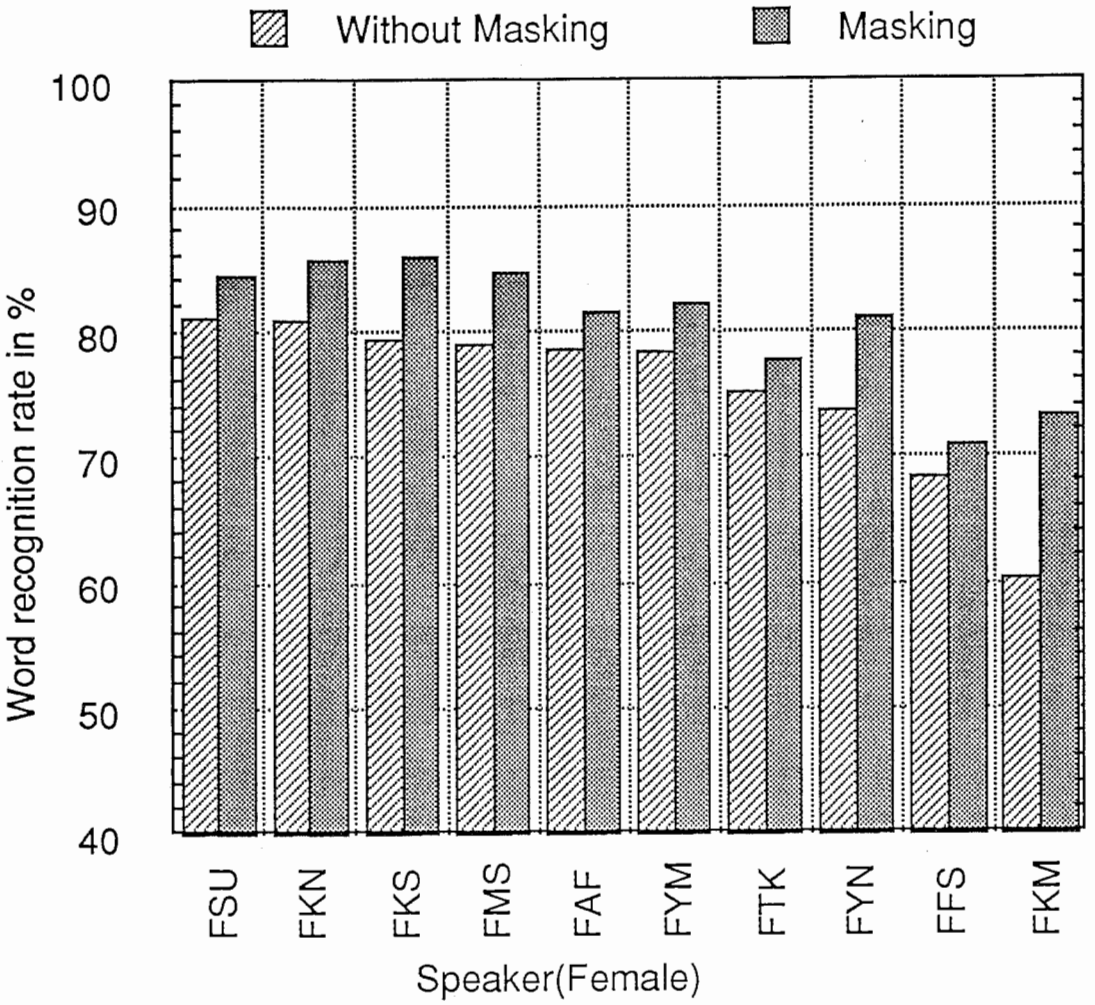

Fig. 4 Multi-speaker word recognition rate of female speakers. Each bar represents average recognition rate of 9 speakers.

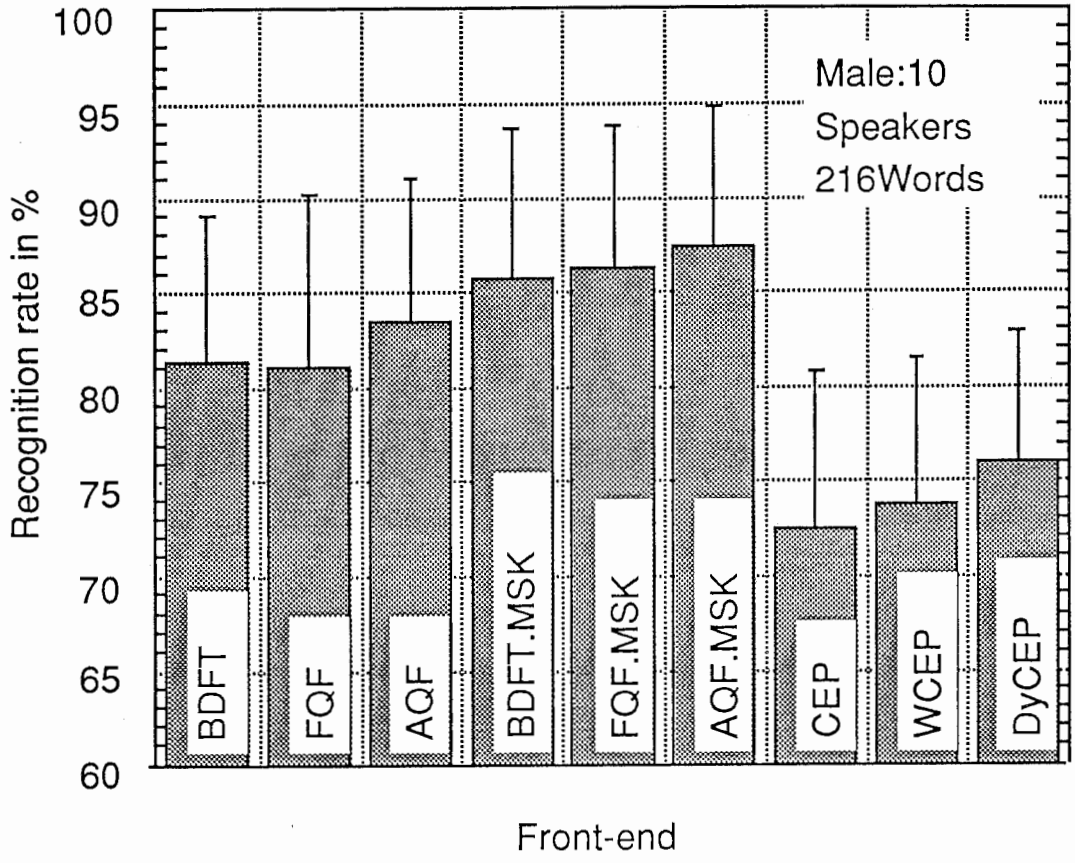

Fig. 5 Performance comparison of various front-ends in multi-speaker word recognition. Each bar represents average recognition rate of 10 speakers.

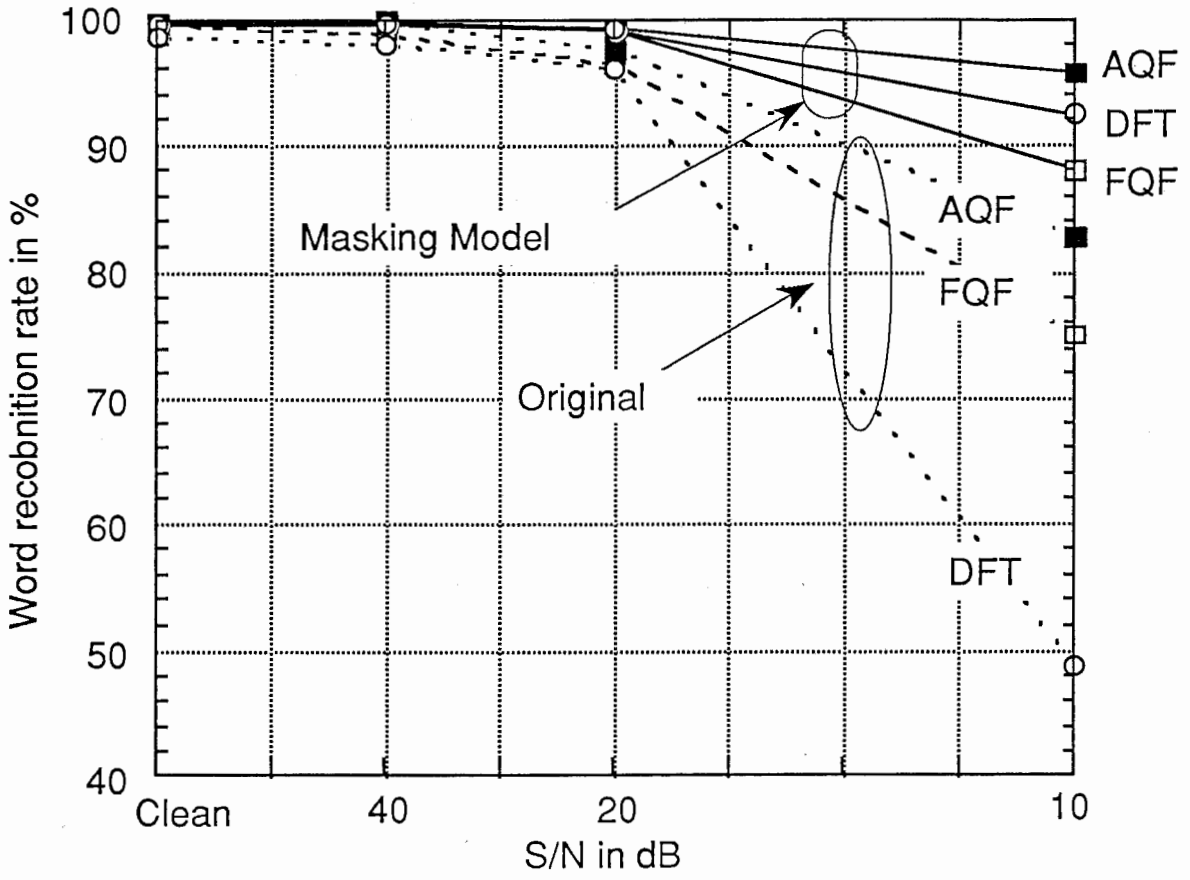

Fig. 6 Word recognition result in noisy environment. Each point represents average recognition rate of 4 speakers.