TR－H－007

21

Speaker-Independent Speech Recognition Using an Auditory Model Front End that incorporates the Spectro-Temporal Masking Effect

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

Speaker-Independent Speech Recognition Using an Auditory Model Front End that incorporates the Spectro-Temporal Masking Effect

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:

Speaker-independent speech recognition experiments using an auditory model front end with a spectro-temporal masking model demonstrated the improvement in recognition performance and outperformed both auditory front ends without the masking model and traditional LPC-based front ends. An auditory model front end composed of an adaptive Q cochlear filter-bank incorporating spectro-temporal masking has been proposed [J. Acoust. Soc. Am., Vol. 92, No. 4, Pt. 2, pp.2476, 5pSP8, Oct. 1992]. The spectro-temporal masking model can enhance essential phonetic features by eliminating the speaker-dependent spectral tilt that reflects individual source variation. It can also enhance the spectral dynamics that convey phonological information in speech signals. These advantages result in an effective new spectral parameter to represent speech models for speaker-independent speech recognition. Speaker-independent word and phoneme recognition experiments were carried out for Japanese word and phrase databases. The masked spectrum was calculated by subtracting the masking level from logarithmic power spectra extracted using a 64-channel adaptive Q cochlear filter-bank. The masking levels were calculated as the weighted sum of the smoothed preceding spectra. To cover the variability of the time sequences of the spectrum, multi-template DTW and Hidden Markov Model were used as the back-end recognition mechanism.

1. Introduction

This paper proposes enhancement of 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 motivation for this is to represent speech spectra more precisely. We proposed an adaptive Q cochlear filter model. The adaptive Q cochlear filter is a non-linear filter that simulates the asymmetrical and power level dependent filtering of the 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., et. al. 1991a, b].

Recent auditory perception research has shown that the forward masking pattern becomes more widespread 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 with the aim of improving the performance of speaker-independent speech recognition. In this study, only the forward masking effect was taken into account because it might be more prominent than backward masking at the auditory peripheral level.

2. Adaptive Q cochlear filter

It is known that the filtering characteristics of the basilar membrane(BM) change adaptively according to the incoming sound intensity. In other words, the Q of the BM filtering becomes high when the sound pressure level of the input speech is low,and the Q becomes low when the sound pressure level of input speech is high. An adaptive Q cochlear filter(AQF) that simulates these level-dependent filtering characteristics of the BM was developed[Hirahara, 1989, 1990]. The adaptive Q

cochlear filter is composed of a NOTCH-BPF combination and adaptive Q circuits connected to each BPF output as shown in Fig.1.

Fig.1

The Adaptive Q circuit consists of a second-order low pass filter(LPF)whose Q is determined by a Q decision circuit. The Q decision circuit determines the Q using the output power of the BPFs, that is, the Q of the LPF becomes high when the output power of the BPF is low, and the filtering Q of the LPF becomes low when the output power of the BPF is high.

This AQF has the following features:

1) Level-dependent frequency selectivity.

2) Level-dependent automatic gain control.

3) Level-dependent resonance frequency shift.

The advantage of the third feature is not yet clear, the first two features seem to be useful for speech feature extraction because the signal-to-noise ratio of weak components is improved by increasing both the gain and the Q of the filter channel. Thus weak consonants and higher formants are enhanced and spectrograms obtained by AQF are much more distinct than those of the fixed Q cochlear filter or traditional DFT(Fig.2). In addition, abrupt spectral changes are also enhanced because of the lag in Q Adaptation. These advantages of the AQF seem to be effective for the front-end of a speech recognition system.

Fig.2

---

3. Forward Masking Model

3.1 Forward masking model and its formalization

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. Fig. 3 illustrates how the masking characteristics are modeled. In this figure, a spectral peak moves toward a lower frequency(the solid curve shows the current spectrum). The masking effect caused by the older spectrum gradually decays and becomes more frequency-smoothed. The smoothed spectrum is integrated into the masking pattern. The hatched area corresponds to the perceived effective spectrum.

Fig.3

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,

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

$$M(u,v) = A(u,v) + D(v) \tag{2}$$

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,

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

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

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 of the masking effect. N is the initial width of the Hamming window. α is the initial weight of the power normalized spectrum and β is the decay rate of the weight.

D(v) is calculated as,

$$D(v) = \sum_{k=1}^{K} \left(\gamma \cdot \delta^{(k-1)} \cdot S_{AV}(v-k)\right) \qquad (5)$$

$S_{AV}(v\text{-}k)$ represents the average power of the frame (v-k). γ is the initial weight of the average power and δ is the decay rate of the weight.

In this report, the masking model parameters are determined to be K=3, N=11, α=0.25, β=0.5, γ =0.5and γ=0.5, by preliminary experiment.

3.2 Characteristics of the masking model

Fig.4 schematically illustrates the effect of the masking model. In Fig.4(a), the spectral peak moves toward a lower frequency with a steady spectral tilt, and in Fig.4(b), the spectral peak moves toward a higher frequency at a different speed.

Fig.4

When the speed of the spectral peak is 0, the masking model gives the output shown in thin lines. When the spectral peak moves faster, the masking mode gives the output shown in thick lines. This means that steady spectral features, such as spectral tilt, are reduced. On the other hand, the spectral dynamics, i.e. spectral peak movement, is enhanced by the masking model. These characteristics of the masking model seem to be effective for feature extraction in speech recognition. This is because the spectro-temporal masking model can enhance common phonetic

features by eliminating the speaker-dependent spectral tilt that reflects individual source variation and also enhance the spectral dynamics that convey phonological information in speech signals.

Fig. 5 shows the spectrogram of the utterance "iyoiyo" by a male speaker. Fig.5(a) is the spectrogram of the AQF output, Fig.5(b) is the masked spectrogram by the masking model and Fig.5(c) is the masking pattern calculated by the masking model.

Fig.5

The masking pattern(Fig.5(c)) reflects the spectral tilt of the AQF spectrum. As a the consequence, the masking model reduces the spectral tilt of the masked spectrum. Comparing Fig.5(a) and Fig.5(b), spectral dynamics of the frequency and time axes are enhanced by the masking model. The masked spectrum gave a much more distinct spectral representation than did the AQF output.

The effectiveness of the masking model in speaker-independent speech recognition was investigated in multi-template DTW word recognition and HMM phomene recognition.

4. Experiments

4.1 Experimental conditions

Speech recognition experiments were performed for 216 phonetically balanced Japanese words uttered by 10 male and 10 female speakers. The sampling rate of the speech was 12 kHz. Two types of cochlear filters were tested: the adaptive Q cochlear filter and the fixed Q cochlear filter. The fixed Q cochlear filter is a simplified version of the adaptive Q cochlear filter, which simulates only the asymmetrical filtering of the basilar membrane. The cochlear filters were compared with a Bark scale bandpass filter based on DFT, LPC cepstrum parameters and dynamic cepstrum parameters which incorporates the spectro-

temporal masking effect in cepstral representation. A 16th-order LPC cepstrum was extracted using linear predictive coding. The dynamic cepstrum enhances spectral dynamics and gives better speech recognition performance than dose the conventional LPC cepstrum parameter.[Aikawa, K. et. al., 1992].

In the DFT frontend, 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. The three filters(AQF, FQF, DFT) 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.

Two types of LPC-based spectral parameters, cepstrum(CEP16) and dynamic cepstrum(DyCEP16), were compared with the filter-based parameters.

4.2 Speaker-independent word recognition experiments using multi-template DTW

Multi-template DTW Word recognition experiments were conducted using the masking model parameters from preliminary experiments. The experiments were repeated 10 times rotating the test speaker. In each experiment, 9 speakers' utterances served as the template, while the utterances of the other speaker served as the test speech. Fig. 6 shows the experimental results.

Fig.6

In the figure, each bar represents the average recognition rate of 10 experiments. By introducing the spectro-temporal masking model, the recognition rate of AQF, FQF, DFT were improved. The adaptive Q cochlear filter with the masking model gives the best performance, improving the average recognition rate to 98.3% from 96.9%, 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. These experiments demonstrated that the adaptive Q cochlear filter with the spectro-temporal model gives feature parameters that are less affected on speaker variations than those of traditional feature parameters. Although the parameter sizes are not the same, the cochlear-filter-based spectra exhibit better performance than do the LPC-based spectra.

4.3 Speaker independent phoneme recognition experiments using continuous HMM.

Speaker-independent phoneme recognition experiments using continuous HMM were performed to examine the effectiveness of the masking model. A tied-output-probability, 3-state model were used for the HMMs. The number of Gaussian mixtures was 8. The masking model was tested in 5-vowel, 8-consonant, and 23-phoneme recognition. The phoneme HMMs were trained on a Japanese database of 216 phonetically balanced word with nine of ten speakers. For testing, one speaker's utterances, which were not used in training, were recognized. Experiments were repeated 10 times rotating the testing speaker. Fig. 7 shows the results of 5-vowel recognition experiments.

Fig.7

The introduction of the masking model to AQF, FQF, and DFT improved the recognition rate, with the AQF giving the best performance. Fig. 8 shows the results of 8-consonant phoneme recognition experiments.

Fig.8

The introduction of the masking model to AQF, FQF, and DFT also improved the recognition rate, and the improvements being greater than those of 5-vowel recognition. This implies that the masking model enhance the dynamics of the consonant features. Among frontends, the AQF again gives the best performance.

Fig. 9 shows the results of 23-phoneme recognition experiments . Fig. 9(a) shows the results with male speakers, and Fig. 9(b) shows the results with female speakers.

Fig.9

Regardless of the speakers gender, the introduction of the masking model to AQF, FQF, and DFT improved the recognition rate. Among the frontends, the AQF gives the best performance. Table 1 compares the performance improvement among AQF, FQF, DFT by introducing the masking model

---

Table 1

In 5-vowel recognition, the difference was not significant.In 6-consonant and 23-phoneme recognition, the improvement of the AQF and FQF were significantly greater than that of the DFT. This shows that the characteristics of the cochlear filter are more suitable for the masking model to give speaker-independent speech feature parameters. These experiments have shown that cochlear filters with the masking model give better feature representations in speaker-independent phoneme recognition than traditional LPC-based representations.

4. Conclusion

In this study, a cochlear filter frontend that incorporates the forward masking characteristics was proposed and its effectiveness evaluated in a speaker-independent speech recognition system. The masking model can enhance essential phonetic features by eliminating the speaker-dependent spectral tilt that reflects individual source variation. It can also enhance the spectral dynamics that convey phonological information in speech signals. The speaker-independent speech recognition experiments showed that the incorporation of the forward masking model improves the recognition performance and that the adaptive Q cochlear filter incorporating the forward masking model gives the best recognition performance.

In conclusion, the adaptive Q cochlear filter incorporating the forward masking is very effective in extracting speaker-independent spectral representations and improves the recognition performance of a speaker independent-speech recognition system.

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.(1992a): "Dynamnic Cepstrum Parameter Incorporating Time-Frequency Masking and Its Application to Speech Recognition," J. Acoust. Soc. Am., Vol. 92, No. 4, Pt. 2, pp.2476, 5pSP5.

Aikawa, K., Kawahara, H. and Tohkura, Y.(1992b): "Speech Recognition Using Dynamnic Cepstrum with Continuous Mixture HMMs," IEICE Technical Report, SP92-103, pp. 1-8.

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 Komakine T,(1989): "A Computational Cochlear Nonlinear Processing Model with Adaptive Q circuits," ICASSP 88, pp496-499

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

Hirahara, T.(1991): "A nonlinear cochlear filter with adaptive Q circuits," J. Acoust. Soc. Jpn,47, 5, pp.327-335

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.

Obara, K., AIKAWA K., and KAWAHARA H.(1992a): "Word Recognition Using Auditory Model Front-End Incorporating Spectro-Temporal Masking," J. Accoust. Soc. Am., Vol. 92, No. 4, Pt. 2, pp.2476, 5pSP8.

Obara, K., AIKAWA K., and KAWAHARA H.(1992b): "Feature Extraction Using Auditory Model Front-ends Incorporating Spectro-Temporal Masking," IEICE Technical Report, H92-59, pp. 1-8.

Figure Captions:

Fig.1

(a) Block diagram of an adaptive Q cochlear filter model.
(b) Adaptive Q circuit of the i-th channel.

Fig.2

Spectro-temporal masking model.

Fig.3

Spectrum enhancement by the masking model for two different spectrum movement speeds. (a) spectrum movement toward a lower frequency, (b) spectrum movement toward a higher frequency.

Fig.4

Spectrum enhancement by the masking model. (a)Adaptive Q cochlear filter output. (b) Masking model output. (c)Masking pattern.(Utterance: "iyoiyo" by a male speaker)

Fig. 5

Effect of the masking parameters on word recognition.

Fig. 6

Multi-speaker word recognition results for male speakers. Each bar represents the average recognition rate of 9 speakers.

Fig. 7

Multi-speaker word recognition results for female speakers. Each bar represents the average recognition rate of 9 speakers.

Fig. 8

Performance comparison of various front-ends in multi-speaker word recognition. Each bar represents the average recognition rate for 10 experiments. In each experiment, the average recognition rate of 9 speakers is obtained.

Fig. 9

Word recognition results in a noisy environment. Each point represents the average recognition rate of 4 speakers.

Table 1

List of words corrected by a masking model.

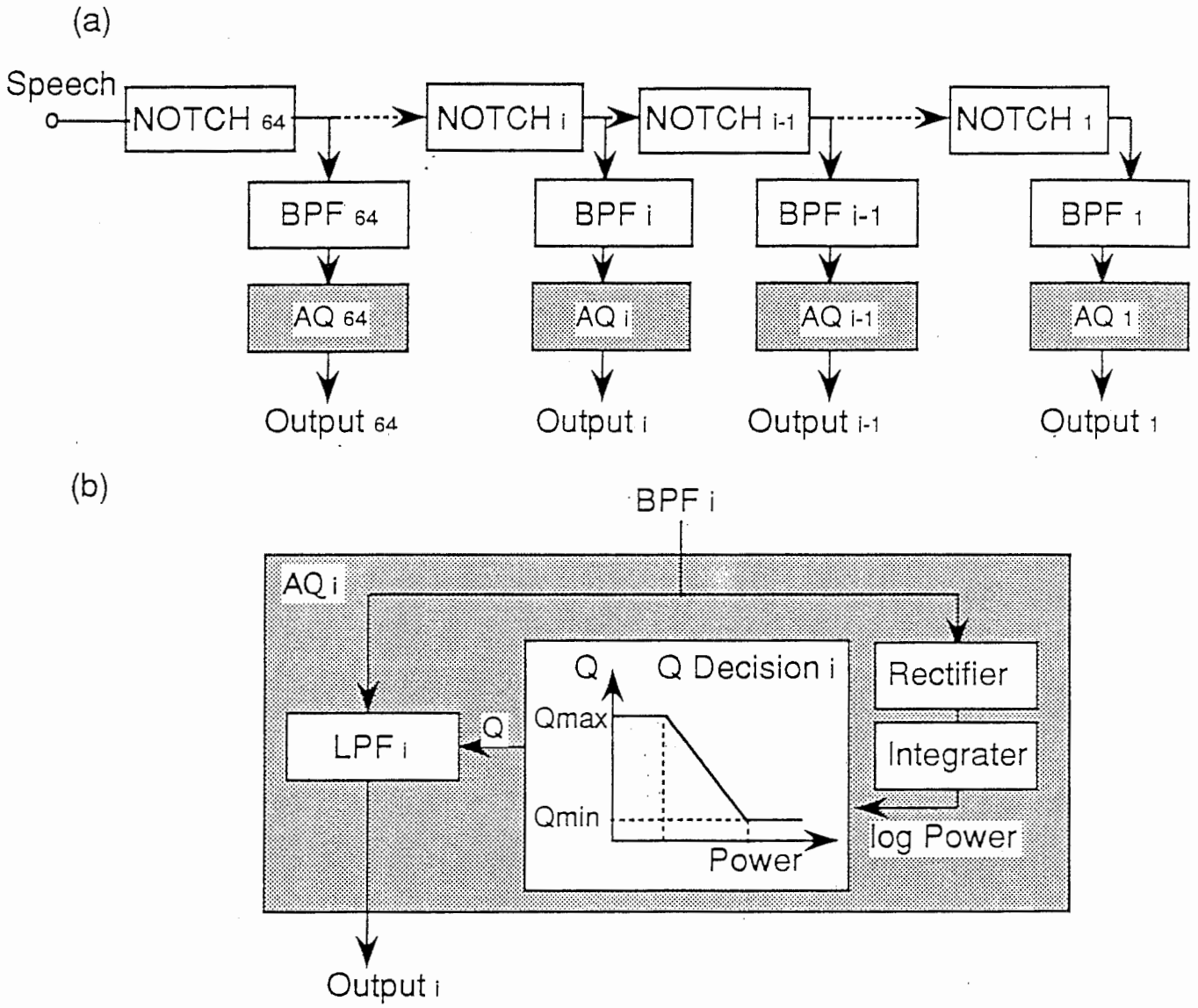

Fig.1 Block diagram of the adaptive Q cochlear FIlter bank.

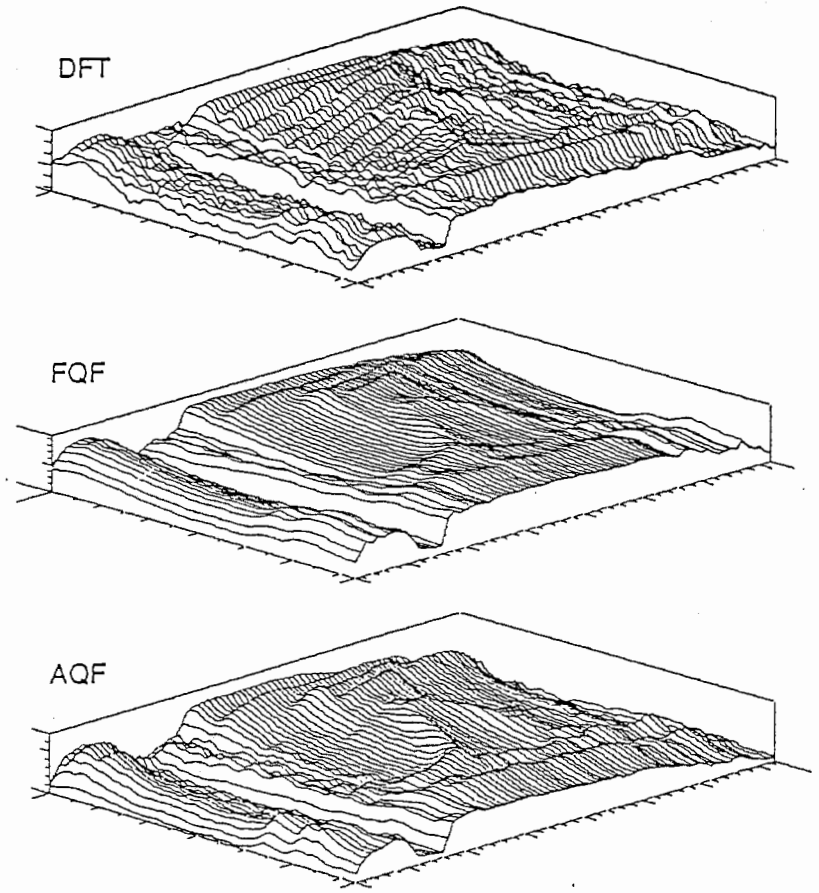

Fig. 2 Spectrogram of DFT,AQF, FQF
Utterance:"ikioi", Male Speaker

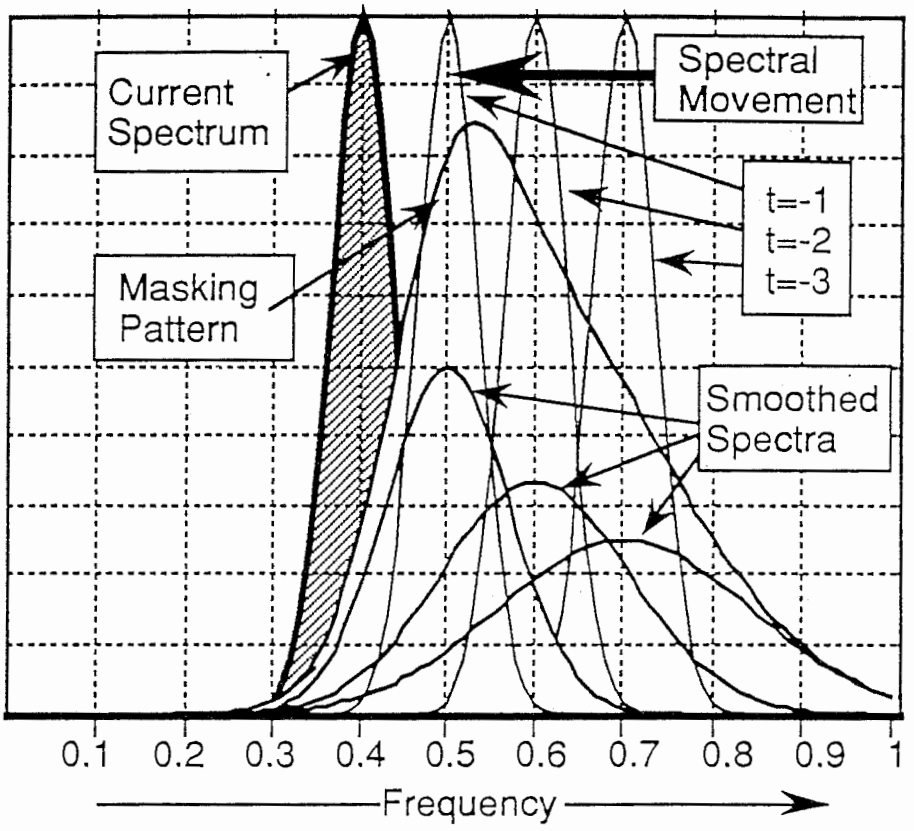

Fig.3 Schematic representation of the forward masking characteristics.

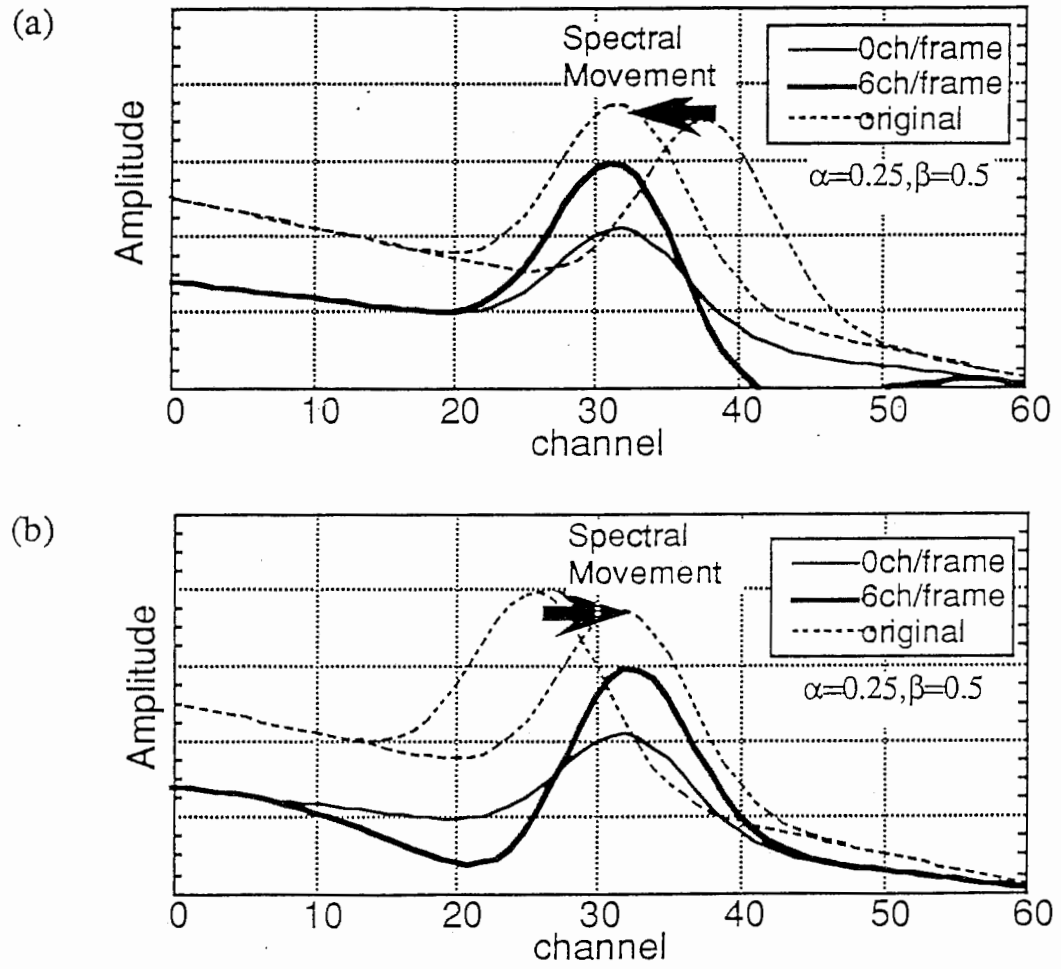

Fig.4 Schematic illustration of tthe masking model characteristics. In Fig.4(a), spectral peak moves toward a lower frequency with steady spectral tilt, and in Fig.4(b), speacral peak moves toward a higher frequency with different moving speed. For detail, See text.

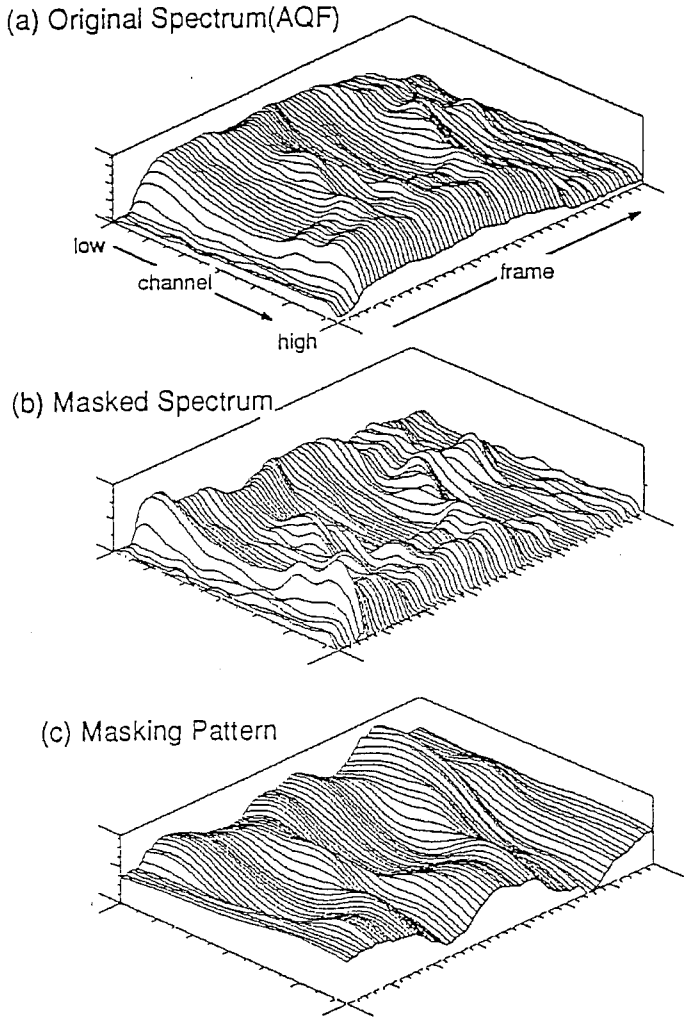

Fig. 5 Spectrogram of AQF(a), Masking model output(b) and masking pattern(c).

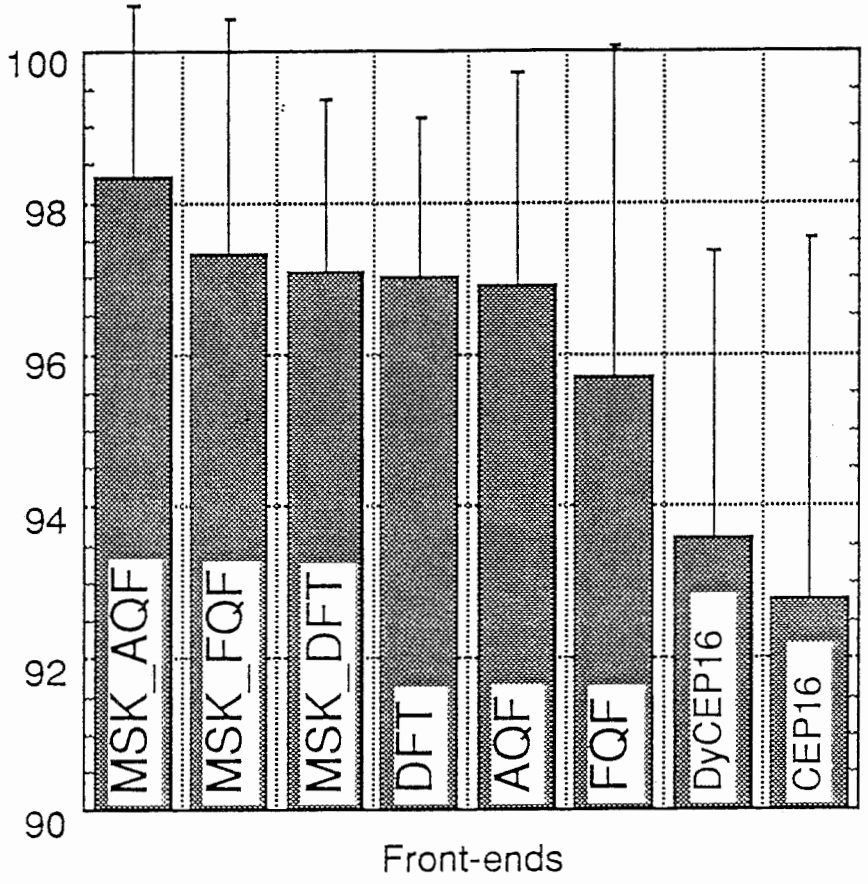

Fig. 6 Speaker independent word recognition using DTW multi-templete recognizers. Each bar represent average word recognition rate of ten speakers.

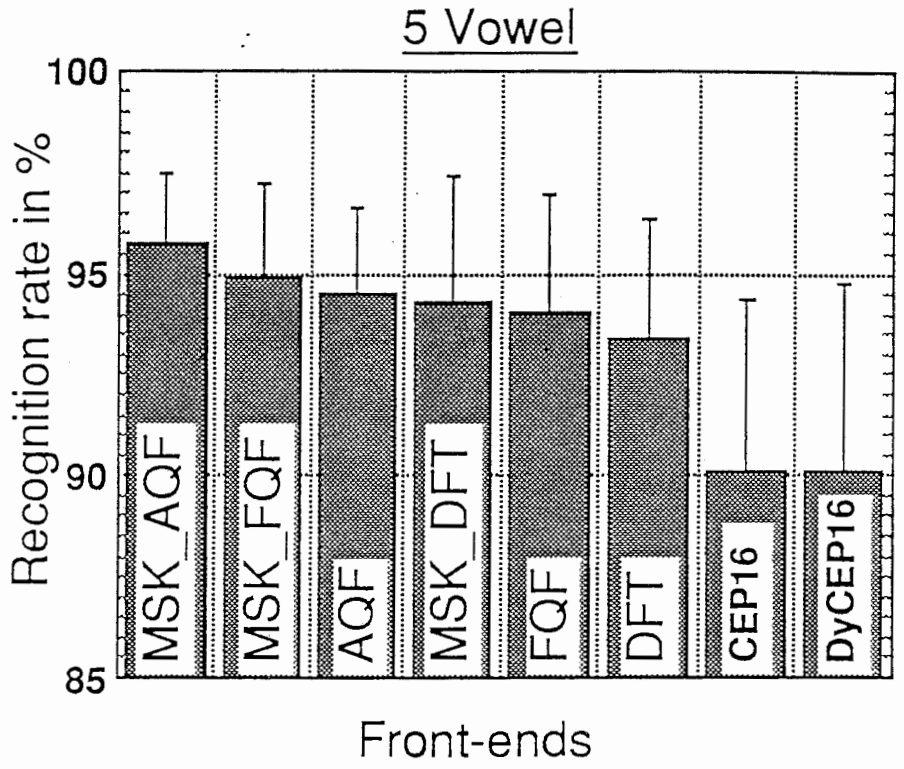

Fig. 7 The results of 5-vowel recognition experiments using continuos HMM recognizer. Each bar represent average word recognition rate of ten speakers.

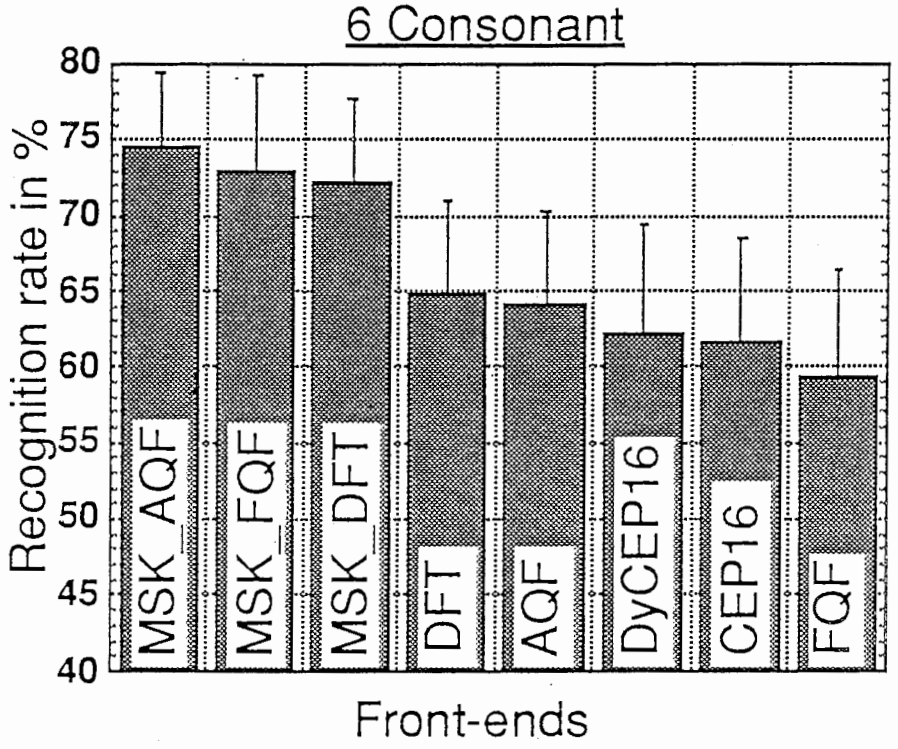

Fig. 8 Results of 6-consonant recognition experiments using continuous HMM recognizer. Each bar represent average recognition rate of ten speakers.

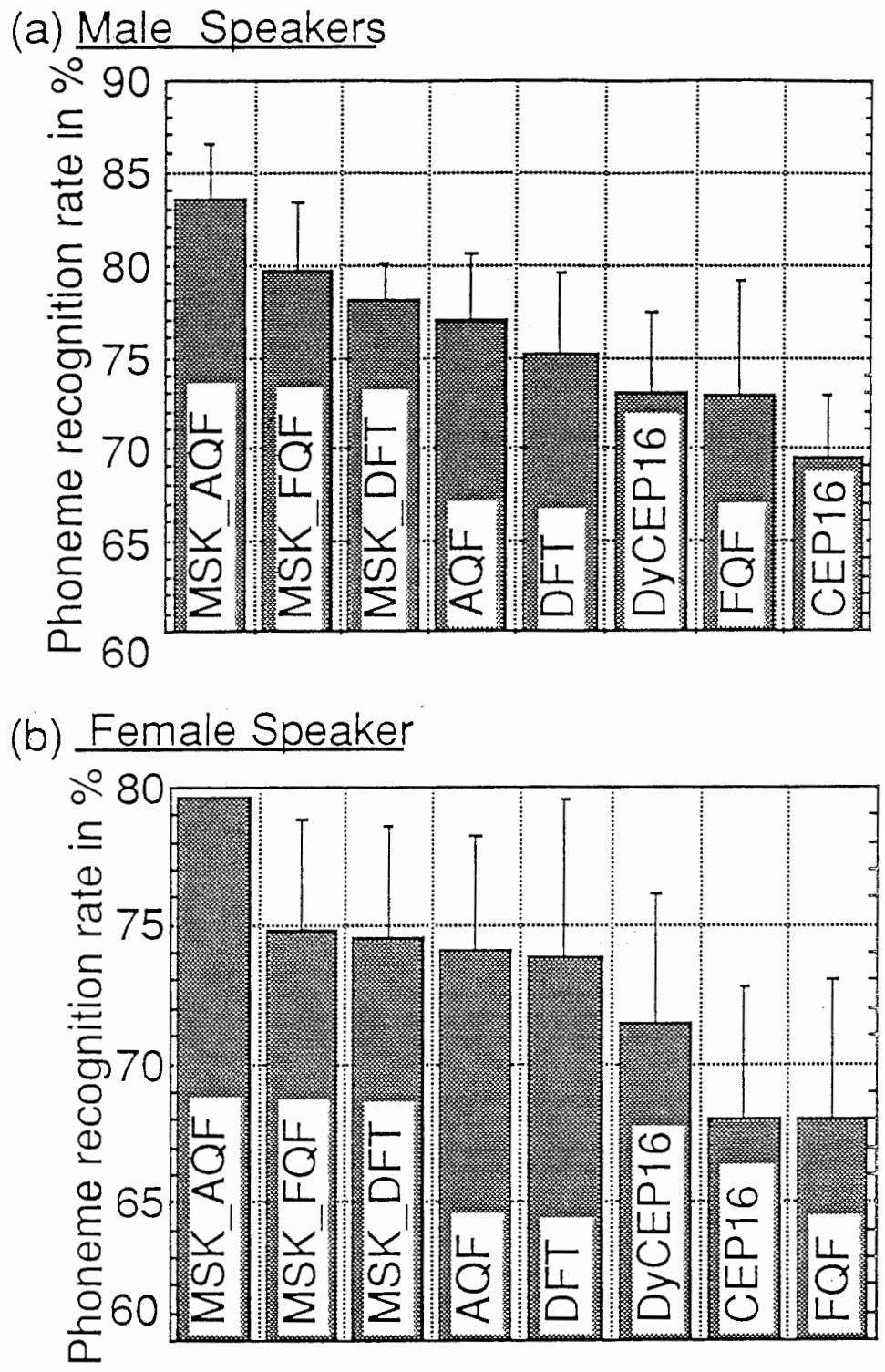

Fig.9 Results of 23-phoneme recognition experiments. Each bar represent average recognition rate of ten speakers. Fig. 9(a) :Average recognition rate of 10 Male speakers. Fig. 9(b) :Average recognition rate of 10 Female speakers.

5-Vowel	Masking	Original	Δ
AQF	95.79	94.50	1.29
FQF	94.93	94.05	0.88
DFT	94.30	93.38	0.92
6-Consonant			
AQF	74.54	64.04	10.50
FQF	72.93	59.38	13.55
DFT	72.14	64.74	7.40
23-CV Phoenme			
AQF	83.35	76.86	6.49
FQF	79.58	72.73	6.85
DFT	77.97	74.99	2.98

Δ= Masking - Original

Table 1 Performance improvement comparison among AQF, FQF, DFT by introducing the masking model.