TR－A－0131

Auditory front-end in DTW word recognition under noisy, reverberant and multi-speaker conditions.

Kazuaki Obara and Tatsuya Hirahara

1992. 1.22

ATR 視聴覚機構研究所

〒619-02 京都府相楽郡精華町光台 2-2　☎07749-5-1411

ATR Auditory and Visual Perception Research Laboratories

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

Telephone: +81-7749-5-1411
Facsimile: +81-7749-5-1408
Telex: 5452-516 ATR J

Auditory front-end in DTW word recognition under noisy, reverberant and multi-speaker conditions.

Kazuaki Obara and Tatsuya Hirahara
ATR Auditory and Visual Perception Research Laboratories,
Seika-cho, Soraku-gun, Kyoto 619-02, Japan

Abstract:

In this report three front-ends, a fixed Q cochlear filter (FQF), an adaptive Q cochlear filter (AQF), and a Bark DFT (DFT), are compared for use as the front-end of a DTW system. The FQF is a conventional cascade/parallel type cochlear filter which simulates the asymmetrical filtering characteristics of a basilar membrane system. The AQF is a nonlinear cochlear filter which simulates three level-dependent characteristics of a basilar membrane system [T. Hirahara *et al.*, Proc. ICASSP, 496-499 (1989)]. The DFT front-end generates 64-channel Bark scale coefficients based on a 512-point DFT magnitude spectrum. These three front-ends have 64 channels covering the frequency range from 1.5 to 19.5 Bark. Recognition performance for clean speech, speech degraded by adding noise and/or reverberation, and under multi speaker conditions, are compared. Four signal-to-noise ratios, S/N=∞ (clean), 40, 20 and 10 dB, are set by adding different levels of pink noise to speech data. For reverberant speech, the impulse responses obtained in the ATR reverberation room, RT=0.2 and 1.1 seconds, were convolved with speech data. Speech data used in the experiments were 216 phoneme-balanced Japanese words uttered by 2 male and 2 female speakers. A standard dynamic time warping (DTW) system was used as a back-end. The experiments results are as follows: (1) For clean speech, AQF performance is equal to that of DFT. (2) For noisy speech, AQF performance is equal to that of FQF but more robust than that of DFT. (3) For reverberant speech, AQF is affected more than DFT but the performance is better than that of FQF. (4) For speaker variation, AQF gives better performance than do FQF or DFT. While the advantage of the AQF front-end is small with an HMM back-end [T. Hirahara *et al.* Proc. ICSLP, 381-384 (1990)], these results show that the AQF can be a better front-end for a DTW recognition system.

1. Introduction

There have been many attempts to build an auditory model that simulates the signal processing which occurs in the auditory periphery, and to use the model as a recognition front-end. The underlying hypothesis of these studies is that if the model is designed properly, spectrum representation can be superior to that of a traditional spectrum. From this viewpoint, some recognition experiments have been made. M. Hunt *et al.*[1986,1988] and J. Cohen[1989] showed that their auditory model can outperform traditional front-ends. However other studies did not always show the superiority of the auditory front-ends (E. Zwicker *et al.*,[1979], M. Blomberg *et al.*[1982,1984], H. Hamada *et al.*[1989], R. Patterson *et al.*[1989], T. Hirahara[1990], S. Kajita *et al.*[1991]) Thus, the use of an auditory front-end has not been accepted widely in the automatic speech recognition field. We have built an adaptive Q cochlear filter which functionally simulates level dependent filtering characteristics of the basilar membrane (T.Hirahara *et al.*[1989]). In a previous study, recognition performance of the adaptive Q cochlear filter front-end was examined using an HMM back-end (T.Hirahara *et al.*[1990]), and LVQ2 back-end (T.Hirahara *et al.*[1991]) but the results were not satisfactory. One possible reason is that the modern stochastic pattern classifiers, such as HMM or LVQ2, are so powerful in classifying patterns that the feature extraction of the adaptive Q cochlear filter front-end might not have advantage, and another possible reason is that, owing to the HMM or LVQ2 back-end constriction, the original 55 channel feature vectors produced by the the adaptive Q cochlear filter front-end were merged into 16 channels, so feature vectors were not fully utilized.

In this report a DTW back-end was used to evaluate the performance of the adaptive Q cochlear filter front-end without merging the feature vectors. Recognition performance of the adaptive Q cochlear filter front-end (AQF) in a DTW word recognition system was compared with fixed Q cochlear filter front-end (FQF) , and bark scale DFT front-end (DFT) under noisy, reverberant and multi-speaker conditions.

2 Speech data

2.1 speech database

In the word recognition experiments, 216 phoneme-balanced words from the ATR speech database were used. These words were uttered two times by 2 male and 2 female speakers, and sampled at 20 kHz with 16-bit accuracy. The first utterance was used as a template and the second utterance was recognized.

2.2 Noisy speech

Noisy speech was made by adding pink noise to clean speech. The signal-to-noise ratio (S/N) was defined by global S/N, i.e.

$$S/N = 10 \cdot \log \left(\frac{\text{Total Energy of the Word}}{\text{Total Energy of the Noise}} \right)$$

Different noise was added to each word. S/N was set to 40, 20 and 10 dB.

2.3 Reverberant speech

Reverberant speech was generated by making a convolution of clean speech and a reverberation impulse response obtained from a variable reverberation room as shown in Fig.1. The reverberation impulse responses are shown in Fig.2. The reverberation time were 200 and 1070 ms respectively. The length of reverberation speech was set equal to that of clean speech.

3. Front-ends

In this study 3 front-ends were used, i.e. adaptive Q cochlear filter (AQF), fixed Q cochlear filter (FQF) and Bark scale DFT (DFT).

For a long time, basilar membrane was considered to be passive linear filter, but recently, adaptive filtering of the basilar membrane has been confirmed. (Johnstone, *et al.*[1986]) The filtering Q of the basilar membrane becomes high when the sound pressure level of input speech is low, and low when the sound pressure level of input speech is high.

We have developed an adaptive Q cochlear filter which simulates these level dependent filtering characteristics of the basilar membrane. To evaluate the performance of the AQF in DTW word recognition, we compared its performance with the other two: FQF and DFT.

3.1 Fixed Q cochlear filter

The block diagram of the fixed Q cochlear filter (FQF) is shown in Fig.3. The FQF is composed of a NOTCH-BPF (Band Pass Filter) combination, which simulates asymmetrical filtering characteristics of the basilar membrane: A steep high cut-off and a gradual tail at lower frequency. In this study the Q of the BPF was set to 5.0.

3.2 Adaptive Q cochlear filter

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.3. The adaptive Q circuit consists of a second order low pass filter (LPF) whose Q is determined by a Q decision circuit (Hirahara[1989]). The Q decision circuit determines the Q using the output power of the BPFs, that is, the Q of LPF becomes high when the output power of BPF is low, on the other hand 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 nonlinear reduction of the relative gain.

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 FQF or

DFT. In addition, abrupt spectral changes are also enhanced because of the lag of Q Adaptation. These advantages of the AQF seem to be effective for the front-end of a speech recognition system.

To determine the control parameter of the adaptive Q circuit, preliminary experiments were conducted under a noisy environment. The relationship between adaptive Q control parameter and word recognition performance are shown in Fig.4.

In these experiments, 56 words were selected from the speech data mentioned before, and performance was measured using the DTW (Dynamic Time Warping) word recognizer. Changing the Q control from [a] to [e] as shown in Fig.4, recognition performance of noisy speech was improved, while performance of clean speech was not changed. According to this preliminary experiment, the parameter of the adaptive Q circuits was determined.

3.3 DFT front-end

A Bark scale DFT front-end (DFT) was also used to evaluate the performance of the two cochlear filter front-ends. A feature extraction from the DFT is summarized in Fig.5.

Input speech was 20ms Hamming windowed and a 1024-point FFT computed every 10ms. Then a 512-channel DFT power spectrum was obtained and transformed into 64-channel bark scale coefficients. This transformation was done by summing up the DFT power spectrum components in each bark scale energy band. Finally, all coefficients were transformed into logarithmic values. The center frequency of each Bark channel , CH[n] , was set using the following equation:

$$CH[\,n\,] = 1.5+ (19.5-1.5)/63*n \quad ; n = 0 \text{ to } 63$$

For each channel, energy band was set to (19.5-1.5)/63

To convert Hz frequency to Bark frequency, the following equations were used (Seneff [1986]) :

$$\mathrm{Bark}\,(f) = \begin{array}{ll} 0.01*f & 0 <= f < 500 \\ 0.007*f+1.5 & 500.0<= f < 1220 \\ 6.0*\ln\,(f)-32.6 & 1220<= f \end{array}$$

3.4 Output of the three front-ends

Output of the three front-ends are shown in Fig.6. The utterance is [ikioi] in Japanese by male speaker (MST). Comparing cochlear filter (FQF and AQF) and DFT, the cochlear filter gives a relatively smooth spectrogram. This is because the frequency selectivity of the cochlear filter is lower than that of DFT at low frequency. Comparing AQF and FQF, AQF gives clearer spectrograms than does FQF at low energy level. This is because of the level dependent filtering characteristics of AQF, i.e. level dependent adaptive band width control and level dependent relative gain control. For example, the spectrum of weak consonant, /k /, is much clearer.

4. Word recognition experiments

The experimental diagram is shown in Fig.7. Frame rate and frame length are set to 10 and 20 ms respectively. All the 3 front-ends have 64 channels and each channel has the same center frequency and the same energy band width. In this study, dynamic time warping (DTW) was used as the back-end. The adjustment window was set to 10(Shikano, *et al.*[1982]).The dynamic programing algorithm used in this study is summarized in Appendix. A

4.1 Performance under a noisy environment.

Templates were clean speech, words to be recognized were noisy. For each front-end, Average performance of the four speakers is shown in Fig.8. Filled circles represent the performance of AQF, open circles represent FQF and open squares represent DFT.

When the S/N was high (S/N=40dB), AQF showed almost the same performance as DFT. When S/N became low (S/N=10dB), the performance of AQF and FQF were better than that of DFT. One possible explanation is that the feature extraction of the cochlear filter is smoother than that of DFT, so the spectrum change caused by noise is smaller in the cochlear filter than in DFT.

4.2 Performance under reverberant environment.

Templates were clean speech, words to be recognized were reverberated. For each front-end, average performance of the four speaker is shown in Fig. 9. Filled circles represent the performance of AQF, open circles represent FQF and open squares represent DFT.

When the reverberation time was not overly long (RT=200 ms), the performance differences of the three front-ends were small. But when reverberation time became long (RT=1070 ms), the order of performance was

DFT, AQF and FQF. This means that a front-end with a high Q, i.e. high frequency selectivity, is affected less than a front-end with a low Q.

4.3 Performance under multi-speaker conditions.

One speaker was used as a template and the other three speakers were recognized. Changing the template speaker, the experiments were repeated four times. The results for three front-ends are shown in Fig. 10. For each template speaker, the left bar shows the performance of AQF, the middle bar shows FQF and the right bar shows DFT. Each bar shows the average performance of three speakers.

In this experiment, AQF showed better performance than did FQF and DFT. A possible reason for this result is that the level dependent filtering characteristics of AQF play an important role in multi-speaker conditions. Furthermore, the performance of FQF is better than that of DFT. A possible reason for this result is that smooth feature extraction by cochlear filters gives less spectrum change for different speakers than does DFT.

4.4 $\chi 2$ test of the performance

A $\chi 2$ test was used to confirm the statistical significance of the experiment results. Calculation procedures of the $\chi 2$ test are summarized in Appendix B. The test results are summarized in Fig.11. The level of significance was chosen at $P <= 0.05$.

To summarize the $\chi 2$ test results,

Under noisy environment,

1) The performance of AQF is superior to that of DFT.

2) The performance of FQF is superior to that of DFT.

Under reverberant environment,

3) The performance of DFT is superior to that of AQF.

4) The performance of DFT is superior to that of FQF.

5) The performance of AQF is superior to that of FQF.

Under multi speaker conditions,

6) The performance of AQF is superior to that of DFT.

7) The performance of AQF is superior to that of FQF.

5 Summary and conclusion.

In this report the Adaptive Q cochlear filter in DTW word recognition was evaluated under noisy, reverberant and multi speaker conditions. Results are summarized as follows,

(1) Under noisy environment, Performance of AQF was as good as that of FQF but better than that of DFT. The possible reason is that, as frequency selectivity of cochlear filter is lower than that of DFT at the low frequency, feature extraction of cochlear filter (AQF and FQF) is smoother than that of DFT. For this smooth feature extraction, spectrum change caused by noise can be less than that of DFT.

(2) Under reverberant environment, performance of AQF was affected more than that of DFT but was better than that of FQF. These results show that a front-end with a high Q is less affected under reverberant environment.

(3) Under multi speaker conditions, AQF showed better performance than did FQF and DFT. The possible explanation is that the level dependent filtering characteristics of AQF play an important role in the multi-speaker conditions.

(4) Above results were checked using $\chi 2$ test.

To conclude the study, the AQF front-end can be a better front-end for DTW word recognition system.

Acknowledgement:

The authors would like to thank Mr. S.Tenpaku for his useful advice on programing and Drs. Y. Tohkura and A. Cheveigne for their helpful comment on writing this manuscript.

References:

Anderson, T. (1990): "Speech processing using an auditory model and neural networks," J. Accoust. Soc. Am. Suppl.1, Vol.87, RR20

Blomberg, M., Carlson, R., Elenius, E. and Granstorm, B. (1984): "Auditory models and isolated word recognition," Q Prog. Stat. Rep., Speech Transmiss. Lab. (Royal Institute of Technology, Stockhom), pp. 1-15

Blomberg, M., Carlson, R., Elenius, E.and Granstorm, B. (1982): "Experiment with auditory models in speech recognition," *The Representation of Speech in the Peripheral Auditory System,* R. Carlson, B. Granstorm Eds, Elsevier Biomedical Press, pp. 197-201

Cohen, J. (1989): "Application of an auditory model to speech recognition," J. Acoust. Soc. Am. 85 (6), pp. 2623-2629

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

Hamada, H., Hirahara, T., Imamura, A., Matsuoka, T.and Nakatu, R. (1989): "Auditory-based filter-bank analysis as a front-end Processor for speech recognition," Proc. Eurospeech 89, Vol.2 pp. 396-399

Hirahara, T. and Komakine, T. (1989): "A computational cochlear nonlinear processing model with adaptive Q circuits," ICASSP-88,pp. 496-499

Hirahara, T.(1989); "HMM speech recognition using DFT and auditory spectrograms," Part2 ATR Technical Report TR-A-0075

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): "An adaptive Q cochlear filter in phoneme recognition" Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, Session 1-4 (No page assignment)

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

Hunt, M., Lefebvre, C. (1986): "Speeker recognition using a cochlear model," Proc. ICASSP86, pp. 37.7.1-37.7.4

Hunt, M., Lefebvre, C. (1988): "Speeker dependent and independent speech recognition experiments with an auditory model," Proc. IEEE Int.Conf. Acoustics, Speech&Signal Processing, ICASSP-88, pp. 215-218 (1988)

Johnstone, B., Patuzzi, R. and Yates, G. K. (1986): "Basilar membrane measurements and the travelling wave," Hearing Research, 22, pp. 147-153

Kajita, S. and Itakura, F. (1991): "Speech recognition using synchrony spectrum," IEICE Technical Report, EA91-4

Koizumi, T. and Taniguchi, S.(1989): "Speech recognition based on a model of the auditory system," Tech. Report of IEICE, SP89-41, pp.15-22

Liu, W., Andreou, G., Goldstein, M. (1990): "Analog speech processor based on the auditory periphery," J. Accoust. Soc. Am. Suppl.1, Vol.87, 5SP16

Meng, H. and Zue, V. (1990): "A comparative study of accoustic representation of speech for vowel classification using multi layer perceptrons," Int. Conf. on Spoken Language Processing, ICSLP-90, pp. 1053-1056

Obara, K. and Hirahara, T. (1991) "Cochlear filters in DTW word recognition under noisy, reverberant and multi-speaker conditions," Proc. Fall meeting of the Acoust.Soc.Japan, vol.1, pp.15-16

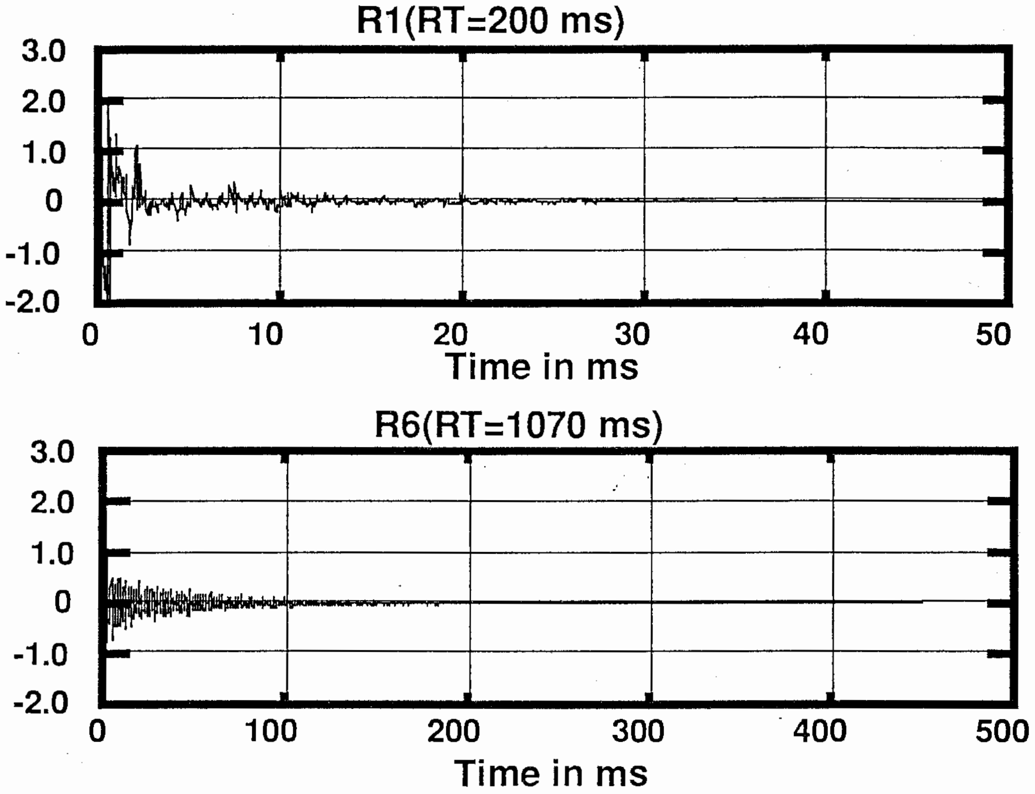
Patterson, R., Hirahara, T. (1989): "HMM speech recognition using DFT and auditory spectrogram," ATR Tech. Report TR-A-0063

Slany, M. and Lyon, F. (1990): "Visual representation of speech - A computer model based on correlation," J. Accoust. Soc. Am. Suppl.1, Vol.88, 2SP3

Seneff, S. (1986): "A computational model for the peripheral auditory system: application to speech recognition research," Proc. ICASSP, 37.8.1-37.8.4

Shikano, K. and Sugiyama, M. (1982): "Evaluation of LPC spectral matching measures for spoken word recognition," Trans. IECE, Vol. J65-D, No. 5, pp. 535-541

Zwicker, E. and Terhardt, E. (1979): "Automatic speech recognition using psychoacoustic models," J. Acoust. Soc. Am. 65, pp. 487-498



R1(RT=200 ms)

R6(RT=1070 ms)

Fig. 1 Reverberation Impulse used in this Experiments

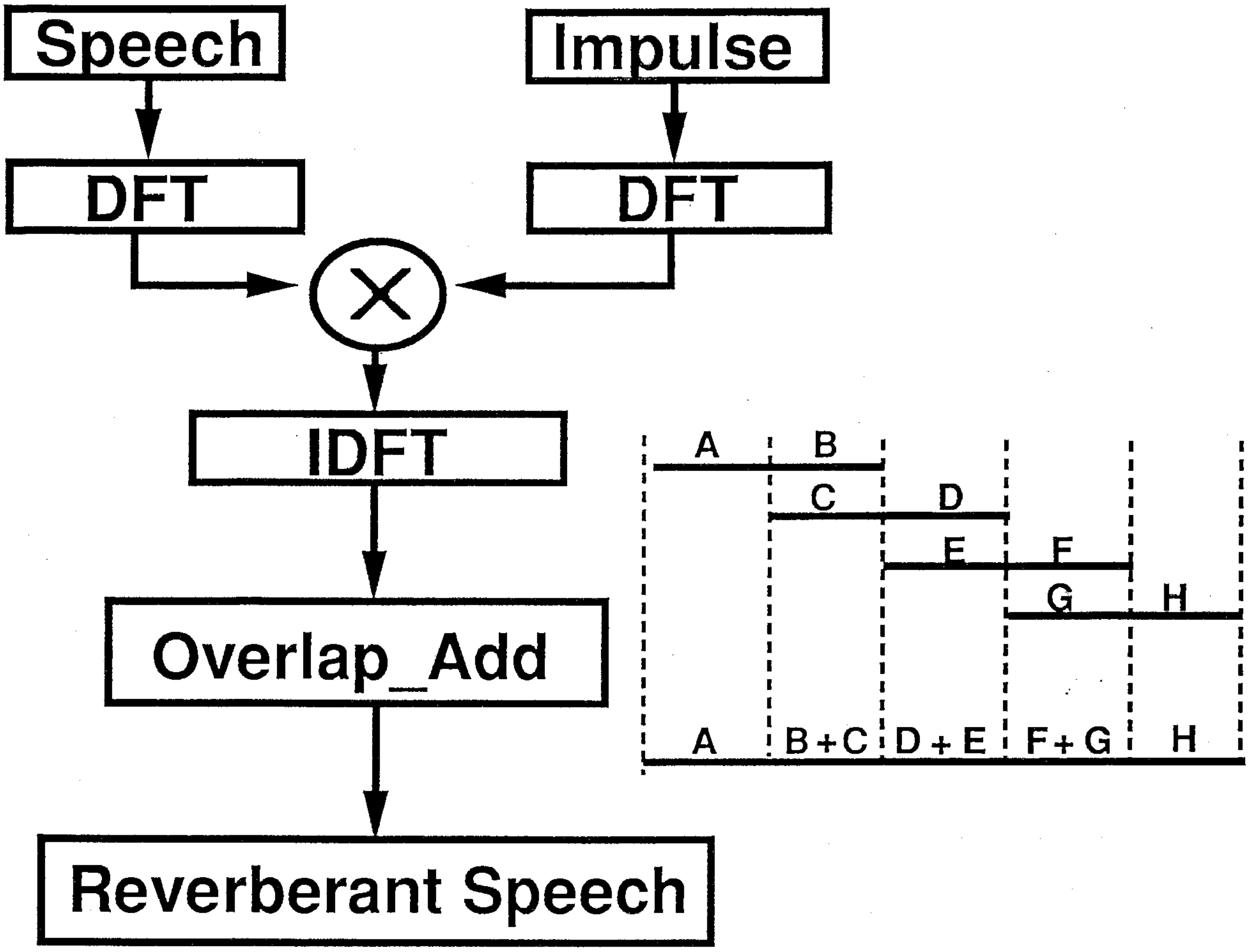


Fig. 2 Signal processing of making reverberation speech.

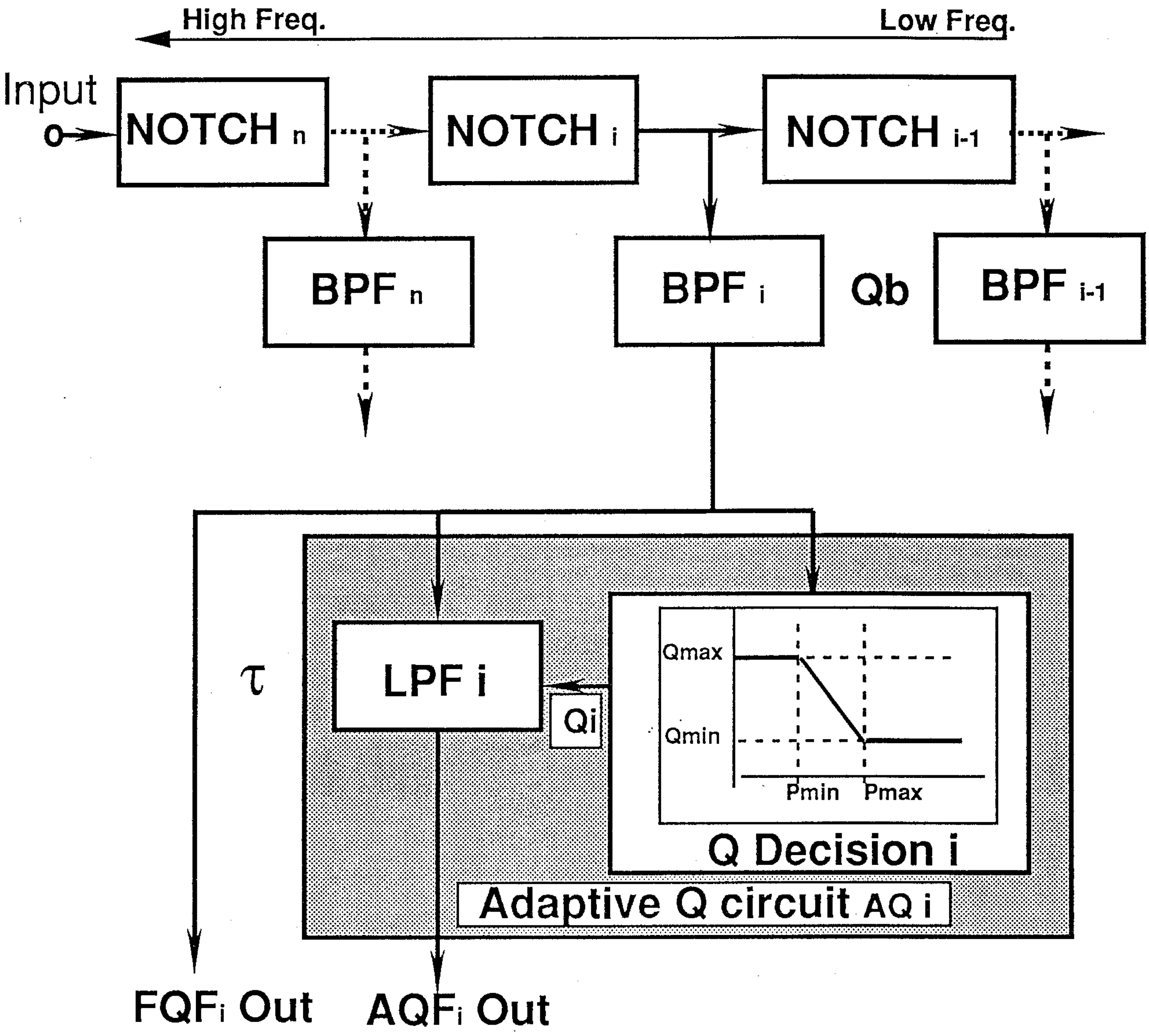


Fig. 3 Block diagram of the fixed Q Cochlear Filter bank and adaptive Q Cochlear Filter bank

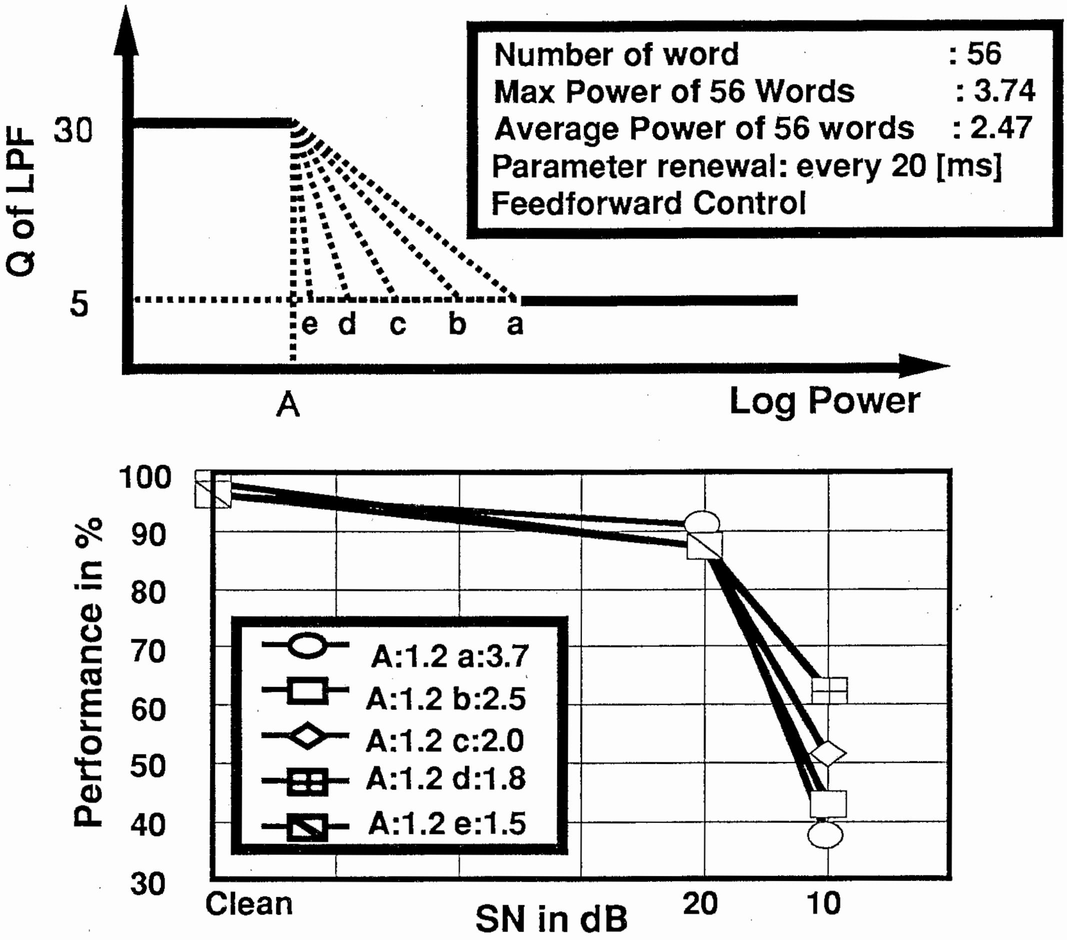


Fig. 4 Adaptive Q parameter vs recognition performance

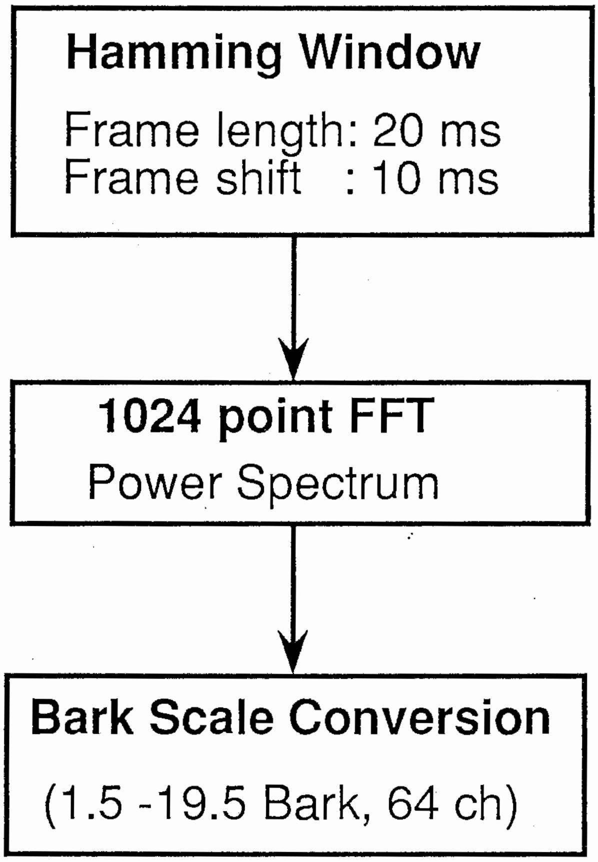
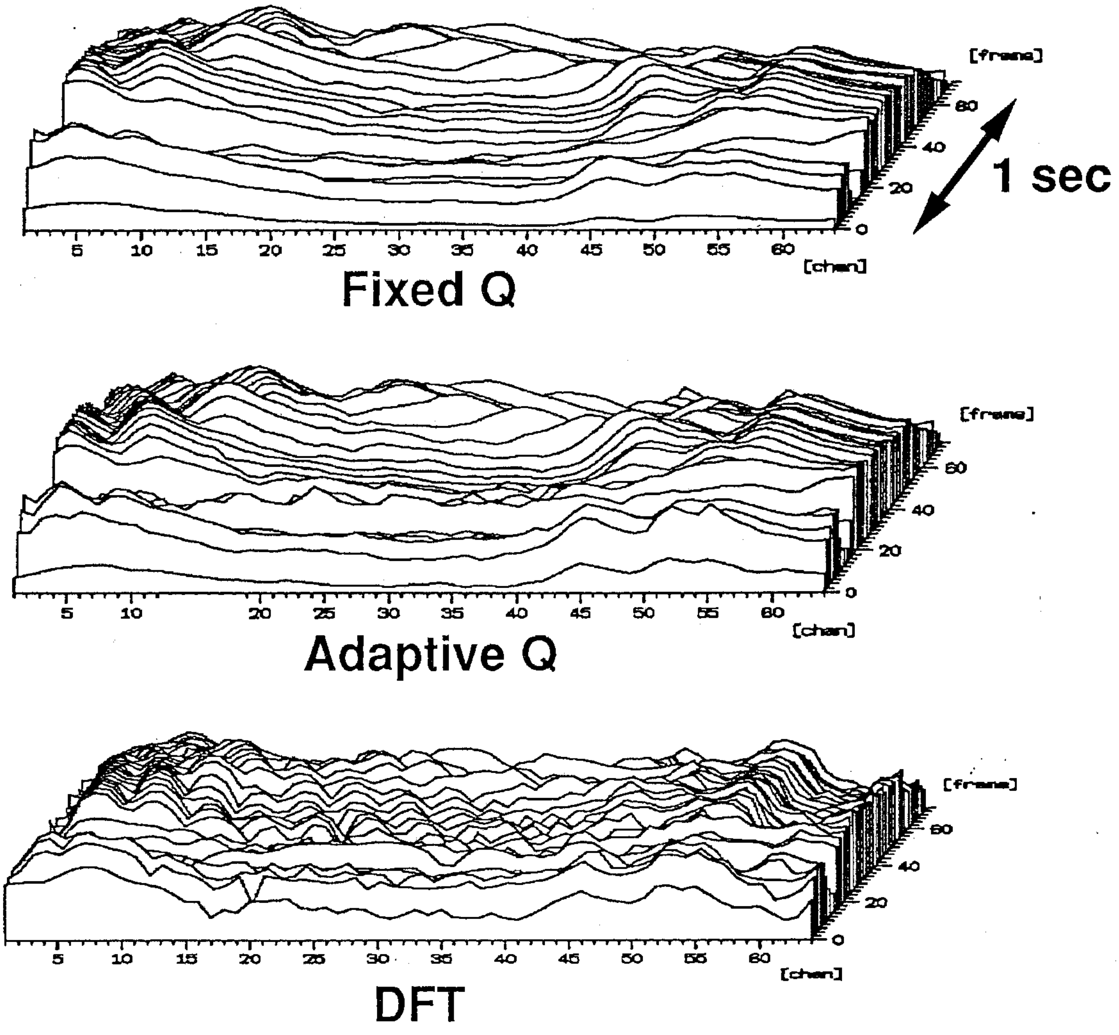


Fig. 5 Bark Scale DFT Frontend

Utterance: IKIOI(Male Speaker)



Fixed Q

Adaptive Q

DFT

Fig. 6 Power spectrum of three front-ends

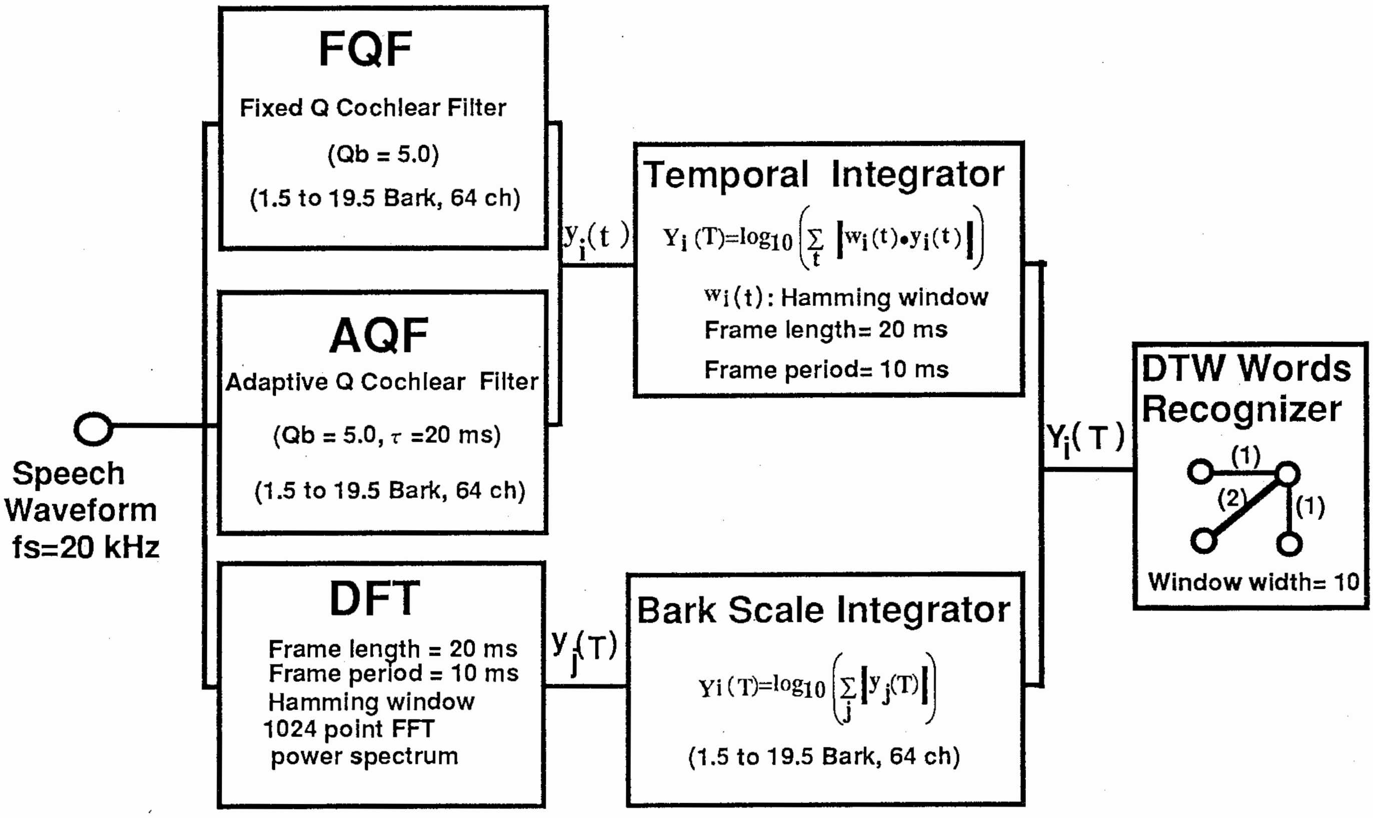


Fig. 7 System diagram of the experiments

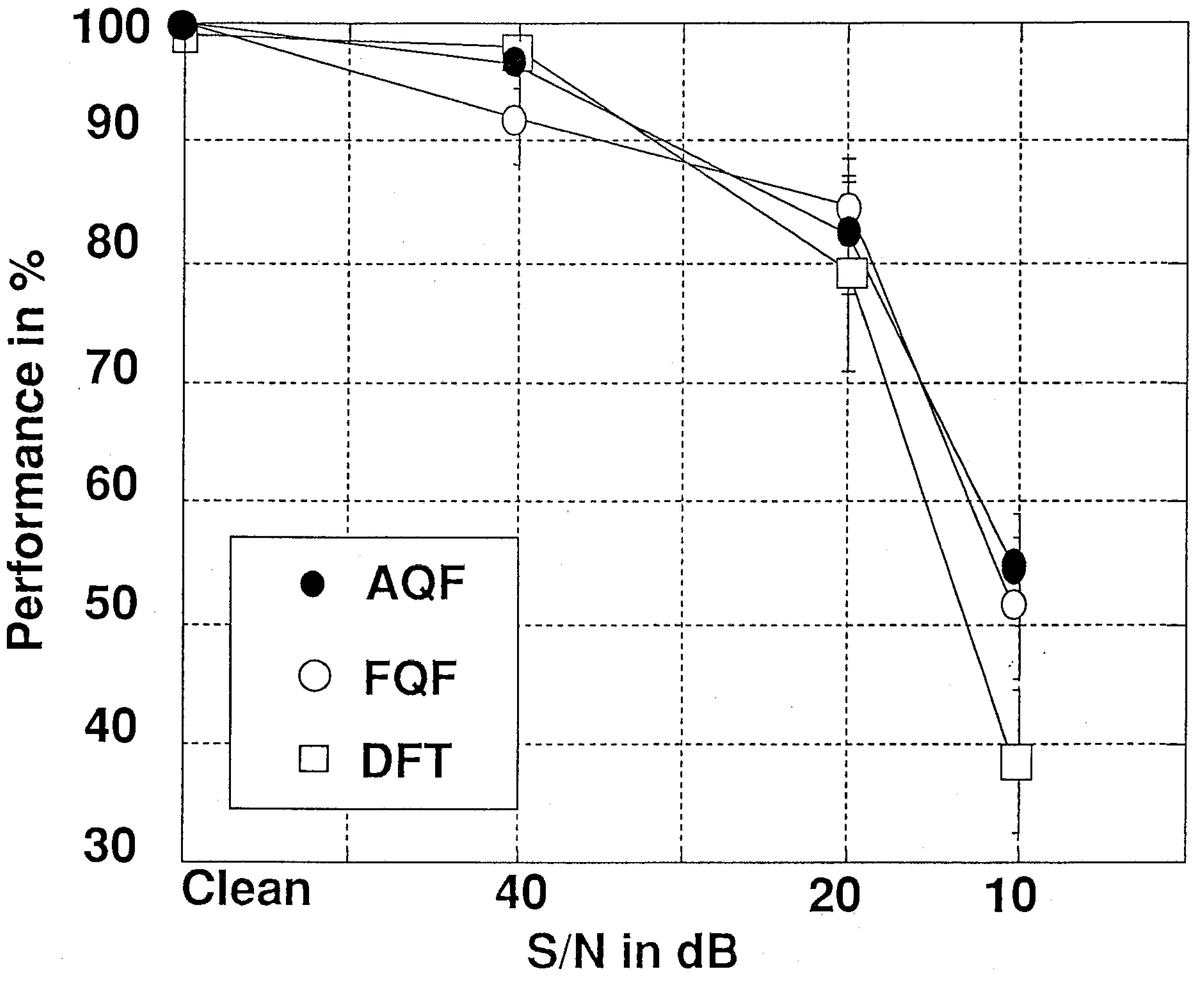


Fig. 8 Performance of three front-ends under noisy environment.

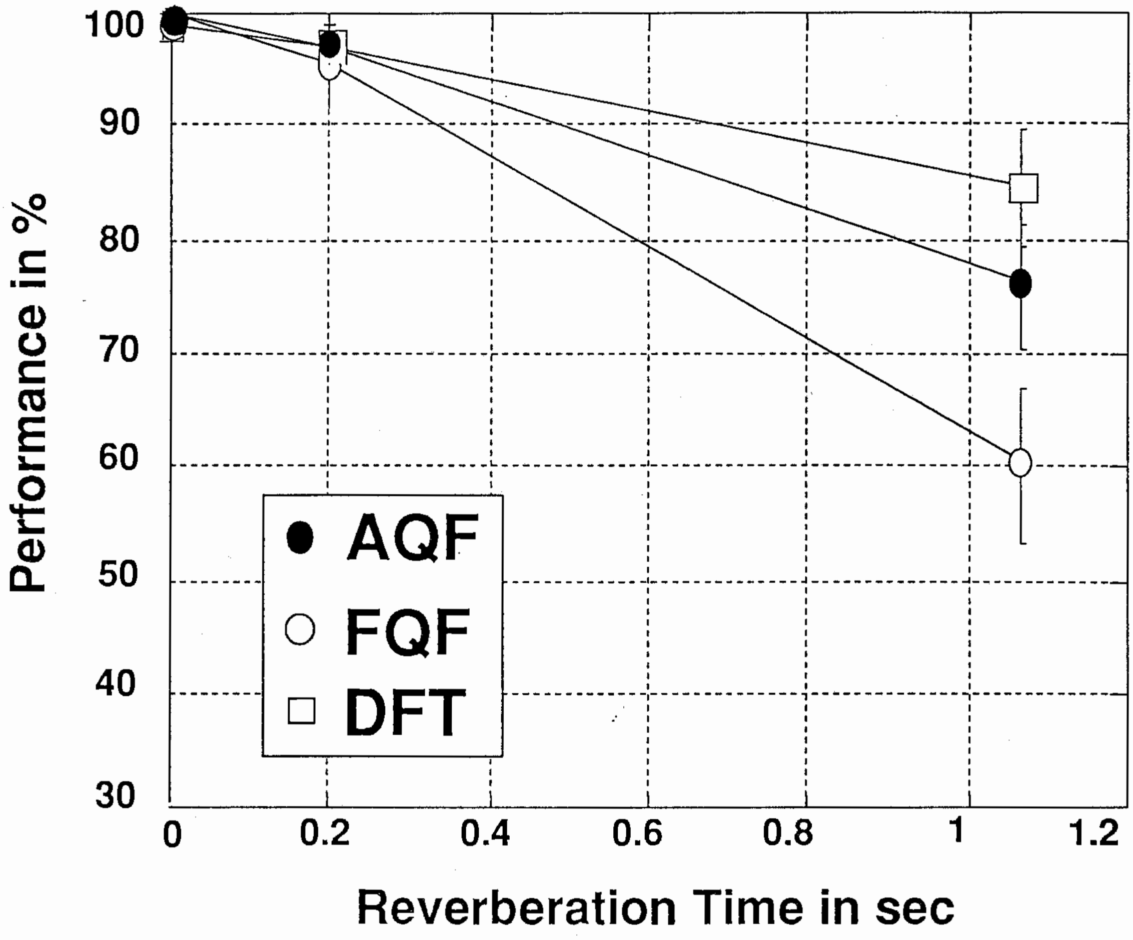


Fig. 9 Performance of three front-ends under reverberant environment.

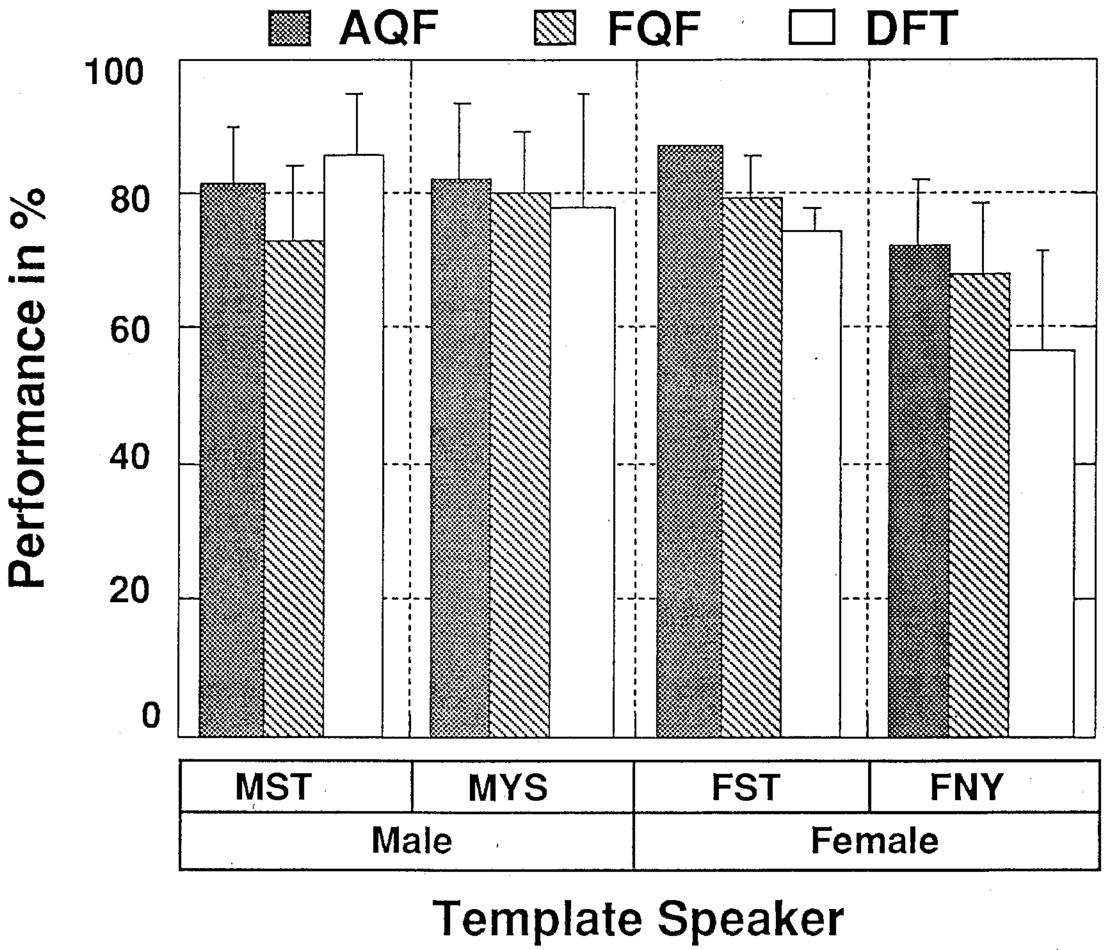


Fig. 10 Performance of three front-ends
under multi-speaker conditions

χ2 test results (Condition: S/N variable)

	AQF-FQF	FQF-DFT	DFT-AQF
SN40	AQF>FQF	DFT>FQF	-
SN20	-	FQF>DFT	-
SN10	-	FQF>DFT	AQF>DFT

χ2 test results (Condition: Reverberation variable)

	AQF-FQF	FQF-DFT	DFT-AQF
R1	AQF>FQF	DFT>FQF	-
R6	AQF>FQF	DFT>FQF	DFT>AQF

χ2 test results (Condition: Multi Speaker variable)

AQF-FQF	FQF-DFT	DFT-AQF
AQF>FQF	-	AQF>DFT

(5 % Significant level)

Fig. 11 chi-square test result of the experiment results

Dynamic Time Warping Algorithm

Initialize:

$$G(1,1) = 2.0 \bullet D(1,1)$$

$$G(1,j) = G(1,j-1) + D(1,j) \quad ;2 \le j < r/2$$

$$G(i,1) = G(i-1,1) + D(i,1) \quad ;2 \le i < r/2$$

$$G(1,j) = \infty \quad ;r/2 < j \le r$$

$$G(i,1) = \infty \quad ;r/2 < j \le r$$

Iteration:

$$G(i,j) = \min \begin{cases} G(i,j-1) + D(i,j) \\ G(i-1,j-1) + 2 \bullet D(i,j) \\ G(i-1,j,) + D(i,j) \end{cases}$$

$$(i,j); \qquad 2 \le i \le Iend, 2 \le j \le Jend$$

MatchingScore:

$$S(A;B) = \min_{(i,j)\ :\text{end region}} \{G(i,j)/(i+j)\}$$

Appendix A Dynamic programing algorithm used in this experiments.

	Correct	Error	Samples
Experiment A	Ac	Ae	At = Ac + Ae
Experiment B	Bc	Be	Bt = Bc + Be
	Ac + Bc	Ae + Be	T = At + Bt

$$\chi^2 = \frac{(Bc \cdot Ae - Ac \cdot Be)^2 \cdot T}{At \cdot Bt \cdot (Ac + Bc) \cdot (Ae + Be)}$$

At:Total Sample of Words in experiment A
Ac:Correct Sample Number in experiment A
Ae:Error Sample Number of experiment A

Bt:Total Sample of Words in experiment B
Bc:Correct Sample Number in experiment B
Be:Error Sample Number in experiment B

T: Total sample number(At+Bt)

if $\chi^2 >= 3.841$ Performance difference is Significant

else Performance difference is NOT Significant

Appendix B χ2 test of the recognition performance of the two frontend.