TR－A－0118 8

Auditory Spectrograms in HMM Phoneme Recognition

Tatsuya Hirahara and Hitoshi Iwamida

1991. 6.27

ATR視聴覚機構研究所

〒619-02 京都府相楽郡精華町乾谷 ☎07749-5-1411

ATR Auditory and Visual Perception Research Laboratories

Inuidani,Sanpeidani, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

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

Auditory spectrograms in HMM phoneme recognition

Tatsuya HIRAHARA and Hitoshi IWAMIDA

ATR Auditory and Visual Perception Research Laboratories

Seika-cho, Soraku-gun, Kyoto, 619-02 JAPAN

This paper is submitted to *Computer Speech and Language* (Academic press).

Abstract

Several auditory spectrograms based on the adaptive Q cochlear filter and its relatives are compared in speaker dependent HMM phoneme recognition tests using clean speech, as well as speech degraded by adding pink noise. These spectrograms are created using a filter bank, an inner hair cell (IHC) model and a lateral inhibition (LINH) circuit, in different combinations. Eight different filter banks with three different types of filters are prepared: (1) a simple band pass filter with Q=4.5 and 30, (2) a conventional fixed Q cochlear filter with Q=4.5 and 30, and (3) an adaptive Q cochlear filter with feedback /feedforward control with a short/long adaptation time constant. Each filter bank is composed of 55 channel filters spaced in 1/3 Bark increments and spanning the frequency range from 1 to 18.7 Bark. The IHC model involves a saturated half wave rectifier and a short term adaptation circuit. The recognition task is to classify input tokens into 18 phoneme categories using 5,788 training tokens and 5,773 testing tokens. Results are as follows; (1) The adaptive Q cochlear filter with LINH gives better recognition performance than the other types of filter banks in all training/testing conditions. (2) The LINH effectively improves recognition performance. (3) The IHC model produces no benefit even for the noisy data set.

1. Introduction

There have been many attempts to build an auditory model that simulates the signal processing which occurs in the auditory periphery. One purpose of such modeling is to obtain an internal speech spectrum representation or its equivalent as a model output, then use the internal representation in speech science or in the speech engineering fields. In particular, applications of an auditory model for speech recognition front-ends have been attracting a great deal of interest. It has been believed that speech recognition performance can be improved by replacing a traditional front-end with an auditory peripheral model. The underlying assumption is that if a model could be designed properly, it would generate a more useful and efficient representation of the speech spectrum compared to traditional physical spectrum representations.

From this viewpoint, several speech recognition experiments using auditory front-ends have been reported. Hunt *et al.* (1986, 1988), Cohen (1989) and Meng *et al.* (190) showed that their auditory front-end outperformed a traditional front-end. However, other studies do not always show an auditory front-end to be superior to a traditional front-end. Some auditory front-ends are superior only for processing speech degraded by noise (Ghitza, 1988; Hunt *et al.*, 1986), but many show little, if any, superiority in processing clean speech (Zwicker *et al.*, 1979; Blomberg *et al.*, 1982, 1984; Hamada *et al.*, 1989; Patterson *et al.*, 1989; Hirahara, 1990; Kajita *et al.* 1991). Thus, this assumption has not yet been widely accepted in the field of automatic speech recognition.

We have also been developing an appropriate auditory model not only for a speech recognition system front-end but also for a general purpose sound spectrum analyzer in a speech perception study. As the first step, we developed an adaptive Q cochlear filter bank, which functionally simulates the level-dependent filtering characteristics of the basilar membrane system (Hirahara *et al.*, 1989, 1991). The output spectrogram of this adaptive Q filter bank seems to be useful as a characteristic vector for a speech recognition system. This is because speech cues are very well represented on the output spectrogram, even when their physical energy is low.

In this paper, several auditory spectrograms generated by this adaptive Q cochlear filter and its relatives are compared in speaker dependent HMM phoneme recognition tests using clean speech, as well as speech degraded by adding pink noise. The main purpose of the work is to evaluate the capacity of the adaptive Q cochlear filter bank as a front-end for an HMM phoneme recognition system. Further, based on the experiment results, we discuss whether an auditory front-end will pay off in automatic speech recognition or not.

2. An Adaptive Q Cochlear Filter

An adaptive Q cochlear filter (AQF) is a computational nonlinear filter which functionally simulates three level-dependent filtering characteristics of the basilar membrane vibrating system in the cochlea: level-dependent frequency selectivity, the level-dependent nonlinear reduction of the relative gain at the resonance frequency and the level-dependent resonance frequency shift (Johnstone *et al.*, 1986).

Figure 1

As shown in Fig.1, the adaptive Q cochlear filter consists of three parts: (1) cascaded second order notch filters (NOTCH), (2). second-order band pass filters (BPF) connected to each NOTCH output and (3) adaptive Q circuits (AQ) connected to each BPF output. In order to simulate the nonlinearity of the basilar membrane system mentioned above, the adaptive Q circuit in Fig.2 is introduced. The adaptive Q circuit consists of a second-order low-pass filter (LPF) of which Q is determined by a Q-decision circuit.

Figure 2

First, when the gain at DC is set at unity, the transfer function of the second order low-pass function LPF(s) is given by

$$\mathrm{LPF}(s) = \frac{\omega_l^2}{s^2 + (\ \omega_l/Q_l\ )s + \omega_l^2} \qquad [1]$$

where ω_l and Q_l are the pole frequency in radian and the Q (quality factor) at the pole of the LPF. Its magnitude frequency response

$$|\mathrm{LPF}(j\omega)| = \frac{Q_l}{\sqrt{Q_l^2\{1-(\omega/\omega_l)^2\}^2+(\omega/\omega_l)^2}} \qquad [2]$$

is shown in Fig.2 for four Q_l values. The maximum value of $|LPF(j\omega)|$ is

$$G_{max} = |LPF(j\omega_{max})| = \frac{Q_l}{\sqrt{1-1/(4Q_l^2)}} \qquad [3]$$

where

$$\omega_{max} = \{1-1/(2Q_l^2)\}\omega_l \qquad [4]$$

When Q_l is large, the formulae [3] and [4] show that the second order low-pass function reaches a maximum gain Q_l at $\omega=\omega_l$. At this time, its transfer characteristics are those of a low-pass filter with a single resonance at ω_l and with Q of the resonance nearly equal to Q_l. Then, the decrease in Q_l brings about the reduction of G_{max}, the lower shift of ω_{max} and the Q decrease simultaneously. When Q_l is below $1/\sqrt{2}$, the transfer characteristics become those of a simple low-pass filter, where ω_l and Q_l are not significant. This means that we can control the maximum gain, the resonance frequency and the Q, simultaneously by choosing adequate values of Q_l.

Next, let us consider a Q-decision circuit which calculates $Q_l(t)$, the Q of the second order low-pass function at time frame t, from controlling signal p(t) in every time frame using the following formulae.

$$\begin{array}{lll} & Q_{max} & p(t) \leq p_{min} \\ Q_l(t) = & (Q_{max}-Q_{min})\{1-\overline{p}(t)\}+Q_{min} & p_{min} \leq p(t) \leq p_{max} \qquad [5] \\ & Q_{min} & p_{max} \leq p(t) \end{array}$$

where

$$\bar{p}(t) = (p(t) - p_{min})/(p_{max} - p_{min}) \qquad [6]$$

While Q_{max}, Q_{min}, p_{max} and p_{min} are constants, p(t) is the logarithmic power of the controlling signal at time frame t. p(t) is given by

$$p(t) = \log(\int_{t-\tau}^{t} |y(t)|dt/\tau) \qquad [7]$$

where τ is a constant which determines the renewal period of $Q_l(t)$. The input-output function of the Q-decision circuit is shown in Fig.2. This Q-decision circuit generates the largest Q_l when p(t) is smaller than ρ_{min} and the smallest Q_l when p(t) is greater than p_{max}. When p(t) is between p_{min} and p_{max}, an intermediate Q_l value is generated inversely proportional to p(t). Two choices are available for the control signal of this Q-decision circuit. For feedforward control, the LPF input, *i.e.* the BPF output, is used as the control signal. For the feedback control, the LPF output itself is fed back to the Q-decision circuit input.

A conventional fixed Q cochlear filter is composed of a cascaded NOTCH-BPF combination (Lyon, 1982), which has asymmetrical filter characteristics: A steep high cutoff and a gradual tail at lower frequencies. Adaptive Q circuits in addition to this fixed Q cochlear filter realize the three level-dependent filtering characteristics: level-dependent frequency selectivity, the level-dependent nonlinear reduction of the relative gain at the resonance frequency and level-dependent resonance frequency shift. Although the advantage of the level-dependent resonance frequency shift for an auditory spectral analyzer is not yet clear, the other two level-dependent characteristics, the Q and gain adaptation, cause the system to act as a rational spectral analyzer.

That is, the signal-to-noise ratio for weak components is improved by increasing not only the gain but also the resonance Q of the channel. Thus, weak consonants and higher formants are enhanced and the spectrograms obtained by the AQ filter are much more distinct than spectrograms obtained by the conventional fixed Q cochlear filters, or DFT spectrograms. In addition, the point where the spectral level changes abruptly is also enhanced by the AQ filter because of the time lag for the Q adaptation. These advantages of the adaptive Q type cochlear filter seem to be promising for the front-end of a speech recognition system.

3. Phoneme Recognition Experiments

3.1 Speech Data

The phoneme tokens used in the experiments are drawn from a large ATR database of 5,240 common Japanese words, which were uttered in isolation by a male professional announcer (MAU) (Kurematsu *et al.*, 1990). All utterances were recorded in a soundproof room and digitized at a 12kHz sampling rate with 16bit accuracy. The database was split into a training set and a testing set of 2,620 utterances each, from which phoneme tokens of 170ms duration were then extracted using manually applied acoustic-phonetic labels.

Each token includes one of the eighteen Japanese consonants /b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, /N/, /s/, /sh/, /h/, /z/, /ch/,./ts/, /r/, /w/ or /y/. The condition of each token extraction differs among consonant categories. With regard to the voiced stops /b, d, g/, the beginning point of the token extraction was

80ms before the succeeding vowel onset. With regard to the voiceless stops /p, t, k/, the beginning point of the token extraction was 40ms before the stop release. For other consonants, the beginning point of the token extraction was 120ms before the consonant-vowel boundary.

Finally, 5,788 tokens were prepared for the training set and 5,773 tokens were prepared for the test set. The number of tokens for each phoneme category is shown in Table 1.

Table 1

With regard to noisy data sets, 170ms pink noise (20Hz to 20kHz) data sampled at 12kHz with 16 bit accuracy was added to each clean token. This pink noise was generated by a signal generator (B&K 1049). The average signal-to-noise ratio for each token, i.e. total energy of a token over the total energy of the noise, was approximately 6dB to 3dB.

3.2 Front-ends

Figure 3 shows the block diagram of the front-ends used in the experiments. The front-end consists of four stages: a filter bank, an inner hair cell model (IHC), a temporal and channel integrator and a lateral inhibition circuit (LINH). Different combinations of filter banks, an inner hair cell model and a lateral inhibition circuit were examined.

Figure 3

3.2.1 Filter banks

Eight different filter banks of three different types were prepared. The first two are simple second-order band pass filter banks whose Q (Qb) is 4.5 or 30 (BPF4.5 and BPF30). The next two are cascade/parallel type fixed Q cochlear filter banks whose Qb is 4.5 or 30 (FQF4.5 and FQF30). The last four are adaptive Q cochlear filter banks with feedback/feedforward control with a short (τ=2ms) and a long (τ=10ms) adaptation time constant (AQFB2ms, AQFB10ms, AQFF2ms and AQFF10ms). Each filter bank was composed of 55 channel filters spaced at 1/3 Bark intervals and spanning the frequency range from 1 to 18.67 Bark (100Hz to 5,114Hz). Each filters' frequency responses in 1 Bark intervals are shown in Fig. 4.

Figure 4

It should be noted that the BPF4.5 and BPF30 are subsystems of the FQF4.5 and FQF30, respectively. Thus, the effect of the asymmetrical frequency response of the fixed Q cochlear filter, which reflects the frequency masking characteristics, is revealed by comparing the BPFs and FQFs. On the other hand, all AQFs are based on the FQF4.5, and the Q of the adaptive Q circuits was determined to vary from 5.0 to 30.0. Then, the actual Q of the AQFs varies from 7.0 to 25. Thus, the advantage of adaptive Q filtering can be seen by comparing the FQFs and AQFs.

3.2.2 Inner Hair Cell Model (IHC)

The inner hair cell model (IHC) that follows the filter bank stage consists of the saturated half-wave rectifier and the short term adaptation circuit proposed by Seneff (1988).

With regard to the saturated half-wave rectifier, it is defined mathematically as follows:

$$\begin{aligned} Y_i(t) &= 1 + A\cdot\tan^{-1}(B\cdot y_i(t)) && y_i(t) > 0 \\ &= \exp(A\cdot B\cdot y_i(t)) && y_i(t) \leq 0 \end{aligned} \qquad [8]$$

where, $Y_i(t)$ is the rectifier output of the i-th channel, $y_i(t)$ is the filter output of the i-th channel, A and B are constants of gain factors. While Seneff chose A=10 and B=65, we set A=10 and B=3 because of the level matching requirement between the filter bank and the rectifier stage. With regard to the short term adaptation circuit, the original parameter values were used.

The output of this IHC model is regarded as the equivalent of the firing rate at the primary auditory nerve. Fig.5 shows the block diagram of the IHC model and the input/output signals of the model.

Figure 5

3.2.3 Temporal Integrator and Channel Integrator

Given the limitation of the HMM recognition system we used, temporal and frequency resolution of either the filter bank output or the IHC model output had to be reduced. First, each channel output was averaged along the time axis by using a 10ms non-overlapped rectangular window. Next, the number of channels was reduced to 16 by combining four (below 10 Bark) or three (above 10 Bark) adjacent channels. Finally, all components were transformed into logarithmic values. This process is defined mathematically as follows:

$$Y_i(T) = \log_{10} \sum_i \frac{1}{N} \sum_t |Y_i(t)| \qquad [9]$$

where N=120 for 12kHz sampling rate.

Thus, 170ms tokens were converted into spectrograms containing 16 channels of logarithmic energies and 17 time frames of 10ms each.

In order to adjust the HMM system input vector size to that of DFT front-end, we used only 240 (16 channels by 15 time frames) dimensional vectors for each token by discarding the first and the last time frame data.

3.2.4 Lateral Inhibition Circuit (LINH)

The lateral inhibition process was performed on the 16 channel by 15 time frame vectors. The *j*-th spectrum at time frame *T* transformed by the lateral inhibition process $Y'_j(T)$ is obtained by a simple convolution of the input *j*-th component $Y_j(T)$ and the lateral inhibition coefficients λ_k.

$$Y'_j(T) = \log_{10} \sum_{k=-n}^{n} \lambda_k . Y_{j+k}(T) \qquad [10]$$

where n=3, λ_0=1.0, $\lambda_{\pm 1}$=0.6 and $\lambda_{\pm 2}=\lambda_{\pm 3}$=-0.3. This lateral inhibition enhances spectral contrast along the frequency axis.

3.2.5 DFT Front-end

Conventional DFT based mel scale spectrograms were prepared for comparison. The input token was 20ms-Hamming windowed and a 256 point FFT computed every 5ms. Then, a 128 channel by 31 time frame DFT spectrum was obtained from a 170ms token. This DFT spectrum was then transformed into 16 mel scale coefficients. This transformation was accomplished by adding the DFT power spectrum components in each mel scale energy band, where adjacent coefficients in frequency overlap by one spectral sample and are smoothed by reducing the shared sample by 50 percent (Waibel *et al.*, 1989). Adjacent coefficients in time were collapsed for further data reduction resulting in an overall 10ms frame rate. Discarding the first frame of the raw DFT spectrum, we then obtained 240 dimensional spectrum vectors of 16 channels by 15 time frames. Finally, all coefficients were transformed into logarithmic values.

Figure 6(a) shows 16 channel by 15 frame feature vectors of each front-end for a token /b/ (/aku<u>bi</u>/; yawing). The formant structures of /bi/ are better represented on the vectors of the IHC or AQF than those of other front-ends. Figure 6(b) shows feature vectors when the LINH was applied. It is clear that the LINH enhances spectral contrast.

Figure 6

3.3 HMM Phoneme Recognition System

Figure 7 shows the HMM phoneme recognition system used in the experiment. In the system, K-means clustering was used to make a codebook. The input vectors for the clustering procedure were a 16 channel by 7 frame partial vector. When this 7-frame vector was used, nine partial vectors were obtained from one token (16 channels by 15 time frames). In the experiments, 20 codebook vectors were assigned for each category.

Figure 7

A phoneme model with four states and six transitions was used in the system. The transition probabilities of the HMMs a_{ij} are all initialized so as to have equal values. The initial values b_{ik} are set, for each code k, at the number of observations of code k, divided by the number of observations of all codes. The Baum-Welch algorithm, based upon maximum likelihood estimation, is used to train the HMMs. The number of iterations was set at seven. A floor value of 10^{-6} was set on the output probabilities to avoid errors caused by zero probabilities.

The recognition task was to classify input tokens into 18 phoneme categories regardless of the following vowel.

4. Results

The results for the experiments are shown in Fig.8(a)-(d). In the figures, the abscissa represents front-end type and the ordinate is the recognition performance expressed in percent. The gray bars represent performance without LINH. The white bars represent improved performance due to LINH. The black lines in the gray bars represent degraded performance due to LINH.

It should be noted that these results were obtained with exactly the same back-end conditions. That is, the very same tokens were chosen as an initial set for K-means clustering in each front-end. When different initial token set were chosen to make codebooks for training tokens, the performance was changed slightly. As for the DFT front-end, the averaged performance was 91.5% and it's standard deviation was 0.85% for ten tests with ten different initial token sets. As for the AQFF10ms front-end, averaged performance was 90.1% and its standard deviation was 0.44%.

Figure 8

Let us look first at the relative recognition performance when the HMM system was trained and tested on clean data (C_C). Among the BPFs and the FQFs, the Qb=30 system gave better performance than the Qb=4.5 system. When Qb was the same, the FQF outperformed the BPF. The AQFs gave better performance than the FQF regardless of either the Q control method or the adaptation time τ. The IHC model used with AQFF10ms degraded the performance. On the other hand, the lateral inhibition process effectively increased performance. The best performance was 92.4% achieved by the AQFF10ms with LINH.

Second, when the HMM system was trained on clean data and tested with noisy data (C_N), performance deteriorated from 20.9 to 24.9% compared to when the HMM system was trained and tested on clean data. The ranking of the nine front-ends is essentially the same as when clean data was used for the test set. The IHC model did not contribute to raising recognition performance. In contrast, the LINH effectively increased recognition performance in most front-ends, but did not work for AQFB, IHC or DFT. The best performance was 70.7% achieved by the AQFF10ms with LINH.

Third, when the HMM system was trained and tested on noisy data (N_N), performance deteriorated from 12.3 to 16.6% compared to when the HMM system was trained and tested on clean data. Among the BPFs and FQFs, a larger Qb gave higher recognition performance. The AQFFs outperformed the BPFs and FQFs. However, the AQFBs were inferior to the FQFs and the BPF30. The performance of the IHC model was the worst. The effect of the LINH was not particularly large compared to other training/testing conditions. The best performance was 76.3% achieved by the AQFF10ms with LINH.

Fourth, when the HMM system was trained on noisy data and tested on clean data (N_C), performance deteriorated from 8.6% to 13.4% compared to when the HMM system was trained and tested on clean data. When Qb was 4.5, the FQF gave performed better than the BPF. However, the result was the opposite when Qb was set at 30. Among the AQFs, shorter adaptation time τ and feedforward type Q control improved performance. The IHC model degraded the performance. The LINH process improved the performance in most front-ends. The best performance was 82.5% achieved by the DFT without LINH.

5 Discussions

5.1 Effect of the filter shape

Filter shape influences recognition performance. It is obvious that a larger Qb gives better recognition performance among the BPFs or FQFs regardless of the training and testing conditions. The FQF generally outperforms the BPF when Qb is the same. Since the FQF has sharper cutoff characteristics at higher frequencies than the BPF, the -3dB bandwidth of the BPF and FQF is nearly the same, but the -20dB bandwidth, for example, of the FQF is narrower than that of the BPF. Thus, these two results imply that the sharper filter gives better results.

5.2 Effect of the adaptive Q filter

Since the actual Q of the adaptive Q filters varies from 7.0 to 25.0 according to the signal level, the advantage of the AQFs is clarified by comparing the performance of the AQFs with that of the FQF4.5 and FQF30. If it were only the filter shape which contributed to the recognition performance, the AQFs' performance should fall between those of the FQF4.5 and the FQF30. Nevertheless, the results show that the adaptive Q cochlear filters outperform not only the FQF4.5 but also the FQF30. Thus, the advantage of the AQFs is not only based on the filter shape but comes from the level-dependent characteristics of the adaptive Q filtering.

With regard to the Q control method, the feedforward control always gives better results than the feedback control. In particular, the difference between the two control methods is obvious when dealing with noisy data. With regard to the adaptation time constant τ, a shorter τ gives better results except when the HMM system was trained on clean data and tested on noisy data. However, when the LINH is applied, a longer τ gives better performance.

5.3 Effect of the LINH

The lateral inhibition process effectively improves the recognition performance in most cases. When the HMM system is trained and tested on clean data, the LINH improves performance from 0.6% to 2.9%. When the HMM system is trained and tested on noisy data, the performance ranges from -1.8% worse to 1.6% better. The LINH degrades the performance of the AQFBs and the DFT. When the training and testing conditions are asymmetrical, performance improvement reaches 4.2%. In particular, the LINH improves the AQFFs' and the FQFs' performance considerably. By way of contrast, it does not work for the DFT.

5.4 Effect of the IHC model

The use of the IHC model is disappointing. Results show that the IHC model used with the AQFF2ms degrades performance not only for the clean data but also for the noisy data. We expected the use of the IHC model to bring about better performance, because the acoustic events of speech are well represented visually on the IHC model output. However, the result was the opposite.

One possible reason is that the IHC model we used is not properly designed. Another possible reason is that the IHC model outputs are adequate for our spectrogram reading knowledge but inadequate for the HMM to classify input tokens. It should be noted that this temporal contrast enhancement is obtained as a result of the level shift threshold mechanism of the short term adaptation circuit. That is, the spectrogram enhancement is accomplished either by emphasizing certain components or eliminating certain components. Thus, this elimination might lose some information required by the HMM. It is interesting that the DFT front-end gives good performance where no component is emphasized but all power spectrum information is preserved.

5.5 Comparison with the DFT front-end

The DFT front-end works surprisingly well for any HMM training/testing conditions. Without the LINH, the DFT front-end always gives the best performance. Furthermore, even when the AQ filter with LINH outperformed the DFT front-end, performance difference between the the two front-ends is less than 1.1%.

The frequency resolution of the DFT front-end original output (128 channels) is about double that of other front-ends (55 channels). Furthermore, the actual analysis length is 165ms for the DFT front-end while it is 150ms for the other front-end. These differences might enhance the performance of the DFT front-end. Hence, in a strict sense, it is not a fair comparison. However, we should say that the benefit of using the auditory front-ends is small under the experiment conditions we used.

5.6 Whether an auditory front-end?

Since the human auditory system is an excellent speech recognizer, it is worthwhile to consider how the human auditory system deals with speech. Nevertheless, it should be noted that the actual auditory periphery is designed for the actual higher level processes of the auditory system. Thus, the output of a properly designed auditory model will be suitable for a human pattern classifier or its equivalent model. However, it is obvious that the modern stochastic pattern classifiers such as HMM are not the model of how human beings classify speech patterns. Therefore, in order to judge whether a properly designed auditory model could generate a useful representation of speech for a speech recognition system or not, it is necessary to repeat the recognition experiment using an auditory front-end and a feature-based pattern classifier.

Figure 9

In addition, it is worthwhile to test a composite phoneme recognition system as shown in Fig. 9. The composite system consists of a time varying front-end, a time varying feature transformer and a pattern classifier. The front-end and feature transformer characteristics are controlled by the recognition result so as to generate a more efficient feature vector for the pattern classifier. The feature transformer plays the role of an "impedance matching section" between a front-end and the back-end used in the system. An auditory model such as the adaptive Q cochlear filter is suitable for the front-end and an artificial neural network such as a spatio-temporal lateral inhibition circuit is suitable for the feature transformer, because their characteristics are easily controlled. These will be the subject of future studies.

6 Summary and Conclusion

In this paper, we have examined several auditory spectrograms in speaker dependent HMM phoneme recognition tests. Results are summarized as follows:

(1) Among the simple second-order band-pass filter banks or the fixed Q cochlear filter banks, the Qb=30 system gives higher performance than the Qb=4.5 system.

(2) When Qb is the same, the fixed Q cochlear filter banks give higher performance than the second order band-pass filter banks.

(3) The adaptive Q cochlear filter bank with feedforward control outperforms traditional filter banks whose filtering characteristics are fixed.

(4) A lateral inhibition process applied on the logarithmic power spectrum improves recognition performance. In particular, the combination of the feedforward type adaptive Q filter and the lateral inhibition process provides the highest performance in most training/testing conditions.

(5) In contrast, the inner hair cell model used in the experiment degrades performance.

(6) Without the lateral inhibition, the DFT front-end always gives the best performance. Even when the adaptive Q cochlear filter with lateral inhibition outperformed the DFT front-end, performance difference between the two is less than 1.1%.

From these results, we conclude that the adaptive Q cochlear filter followed by the lateral inhibition process works well for an HMM phoneme recognition system. However, we should say that the benefit of using this auditory front-end is small under the experiment conditions we used. More systematic application research, such as recognition experiments testing distance measure and recognition methods, are needed to judge whether an auditory front-end pays off or not. In particular, it is necessary to test the combination of an auditory front-end and a feature-based phoneme classifier.

References

Blomberg, M., Carlson, R., Elenius, K. and Granström, B. (1982). Experiments with auditory models in speech recognition. In *The Representation of Speech in the Peripheral Auditory System* (Carlson, R. and Granström, B. Eds.), pp.197-201, Elsevier Biomedical Press, Amsterdam.

Blomberg, M., Carlson, R., Elenius, K. and Granström, B. (1984). Auditory models in isolated word recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 17.9.1-17.9.4

Cohen, J (1989). Application of an auditory model to speech recognition. *Journal of the Acoustical Society of America*, **85**, 2623-2329

Ghitza, O. (1986). Auditory nerve representation as a front-end for speech recognition in a noisy environment. *Computer Speech and Language*, **1**, 109-130

Ghitza, O. (1988). Auditory neural feedback as a basis for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 91-94

Ghitza, O. (1989). Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment. *Journal of Phonetics*, **16**, 109-123

Hamada, H., Hirahara, T., Imamura, A., Matsuoka, T. and Nakatsu, R. (1989). Auditory-based filter-bank analysis as a front-end processor for speech recognition. In *Proceedings of the. Eurospeech*, **2**, 396-399

Hirahara, T. and Komakine, T. (1989). A computational cochlear nonlinear preprocessing model with adaptive Q circuits. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 496-499

Hirahara, T. (1990). HMM speech recognition using DFT and auditory spectrograms (Part 2). *ATR Technical Report*, **TR-A0075**

Hirahara, T. (1991). A nonlinear cochlear filter with adaptive Q circuits. *Journal of the Acoustical Society of Japan*, **47**, 327-335

Hunt, M. and Lefèbvre, C. (1986). Speech recognition using a cochlear model. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 37.7.1-37.7.4

Hunt, M. and Lefèbvre, C. (1987). Speech recognition using an auditory model with pitch-synchronous analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 20.5.1-20.5.4

Hunt, M. and Lefèbvre, C. (1988). Speaker dependent and independent speech recognition experiments with an auditory model. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 215-218

Hunt, M. and Lefèbvre, C. (1989). A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 262-265

Kajita, S. and Itakura, F. (1991) Speech recognition using synchrony spectrum. Technical Report of the Institute of Electronics, Information and Communication Engineers, **EA91-4**, 1-8

Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H. and Shikano, K. (1990). ATR Japanese speech database as a tool of speech recognition and synthesis. Speech Communication, **9**, 357-363

Lyon, R. F. (1982). A Computational model of filtering, detection and compression in the cochlea. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1282-1285

Meng, H. M. and Zue, V. W. (1990) A comparative study of acoustic representations of speech for vowel classification using multi-layer perceptrons. In *Proceedings of the International Conference on Spoken Language Processing*. vol.**2**, 1053-1056

Patterson, R. and Hirahara, T. (1989). HMM Speech Recognition using DFT and Auditory Spectrograms. *ATR Technical Report*, **TR-A0063**

Robinson, T., Holdworth, J. Patterson, R. and Fallside, F. (1990). A comparison of preprocessors for the Cambridge recurrent error propagation network speech recognition system. In *Proceedings of the International Conference on Spoken Language Processing*. **2**, 1033-1036

Seneff, S. (1988). A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics*, **16**, 55-76

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. and Lang L. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transaction on· Acoustics, Speech and Signal Processing*,. **ASSP-37**, 328-339

Zwicker, E. and Terhardt, E. (1979). Automatic speech recognition using psychoacoustic models. *Journal of the Acoustical Society of America*, **65**, 487-498

Figure captions

Table 1 Number of tokens used in the experiment.

Figure 1 Block diagram of a cascade/parallel type adaptive Q cochlear filter bank. The adaptive Q cochlear filter consists of three parts: (1) cascaded second order notch filters (NOTCH), (2) second-order band pass filters (BPF) connected to each NOTCH output and (3) adaptive Q circuits (AQ) connected to each BPF output. This adaptive Q cochlear filter functionally simulates three level-dependent filtering characteristics of the basilar membrane vibrating system in the cochlea.

Figure 2 Block diagram of the adaptive Q circuit. Frequency responses of a second-order low-pass-filter (LPF) at four Q values (left) and input-output relationship of a Q-decision circuit, which calculates the second-order LPF's Q from control signal by formulae (5) (6) and (7) (right).

Figure 3 Block diagram of the front-ends used in the experiments.

Figure 4 Filter responses of three types of filter-banks. In each panel, filter responses are drawn in 1 Bark intervals, *i.e.* in 3 channel intervals. BPF4.5 and BPF30 show the responses of the second-order band-pass-filter whose Q (Q_b) is 4.5 and 30. FQF4.5 and FQF30 show the responses of the cascade/parallel type fixed Q cochlear filter bank whose Q_b is 4.5 and 30. AQF minimum Q and AQF maximum Q show the adaptive Q cochlear filter responses when Q of the adaptive Q circuit of all channel is set at minimum value

(Q_l=3.0) and maximum value (Q_l=45). Q_b of AQFs is set at 4.5.

Figure 5 Block diagram of the inner hair cell model (IHC) which consists of a half-wave rectifier (HWR) and a short term adaptation circuit (STA) proposed by Seneff (1988). Input for the HWR, *i.e.* filter output, HWR output and STA output of 6th, 33rd and 55th channel for one second of speech data are also depicted.

Figure 6 (a) 240 dimensional feature vectors (16 channels by 15 frames) for a token /b/ (/aku<u>bi</u>/; yawing) obtained by each front-end. (b) The same feature vectors obtained by each front-end' with lateral inhibition circuit (LINH).

Figure 7 **(a)** Block diagram of the HMM phoneme recognition system and **(b)** a phoneme model structure. K-means clustering was used to make a codebook, where input vectors for the clustering procedure were a 16 channel by 7 frame partial vector.

Figure 8 The results for the phoneme recognition experiments: The abscissa represents the front-end type and the ordinate represents the recognition performance expressed in percent. The gray bars represent performance without LINH. The white bars represent improved performance due to LINH. The heavy black lines in the gray bars represent degraded performance due to LINH. BPFs are the simple second-order band-pass filter, FQFs are the fixed Q cochlear filter, AQFFs are the adaptive Q cochlear filter with feed-forward Q control, AQFBs are adaptive Q cochlear filter with

feedback Q control, IHC is the AQFF2ms used with IHC model and DFT is the DFT based mel scale filter.

Figure 9 Concept of a composite phoneme recognition system which consists of a time varying front-end, a time varying feature transformer and a pattern classifier. The front-end and feature transformer characteristics are controlled by the recognition result so as to generate a more efficient feature vector for the pattern classifier. An auditory model such as the adaptive Q cochlear filter is suitable for the front-end and a spatio-temporal lateral inhibition circuit is suitable for the feature transformer. The feature transformer plays the role of an "impedance matching section" between the front-end and the back-end used in the system.

T. Hirahara & H. Iwamida

Token	Training	Test	Token	Training	Test
/ b /	218	227	/ s /	475	538
/ d /	202	179	/ sh /	186	177
/ g /	260	252	/ h /	214	207
/ p /	32	15	/ z /	115	115
/ t /	425	440	/ ch /	79	71
/ k /	1152	1164	/ ts /	212	177
/ m /	471	481	/ r /	754	722
/ n /	260	265	/w /	71	81
/ N /	503	488	/ y /	159	174

TOTAL **5788** tokens for training and **5773** tokens for testing

Table 1 Number of tokens used in the experiment

T. Hirahara & H. Iwamida

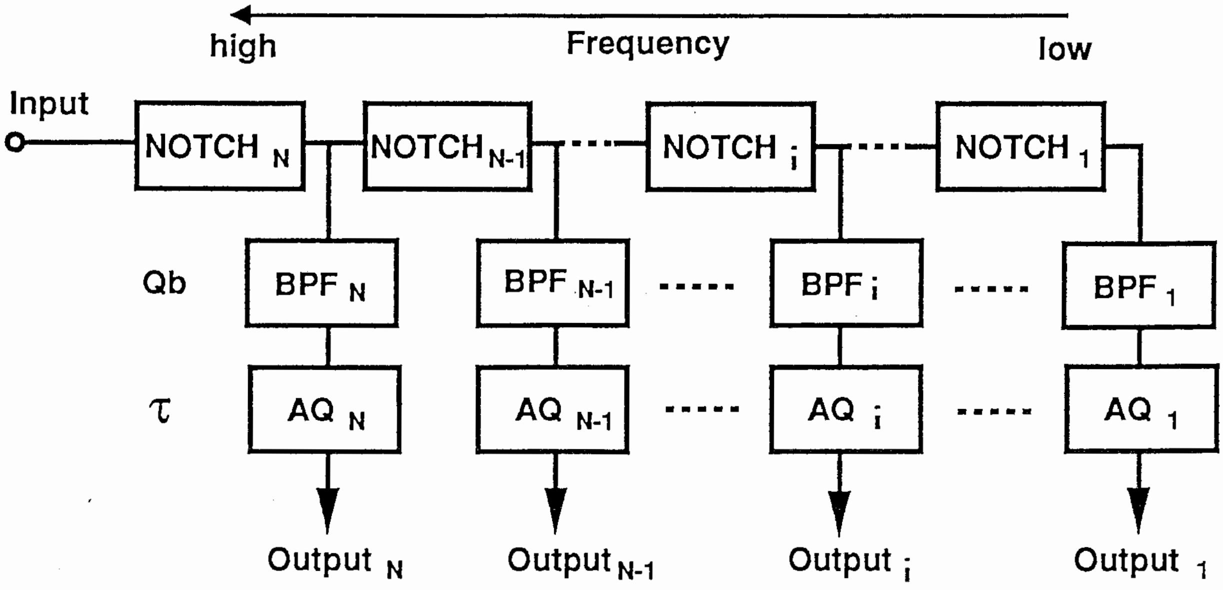


Figure 1

T. Hirahara & H. Iwamida

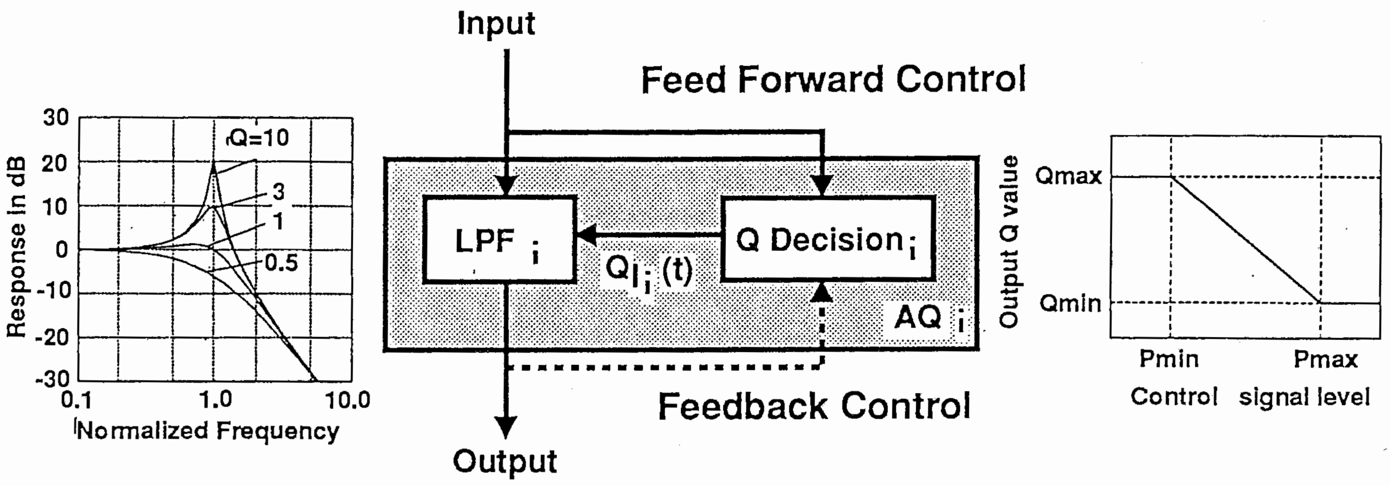


Figure 2

T. Hirahara & H. Iwamida

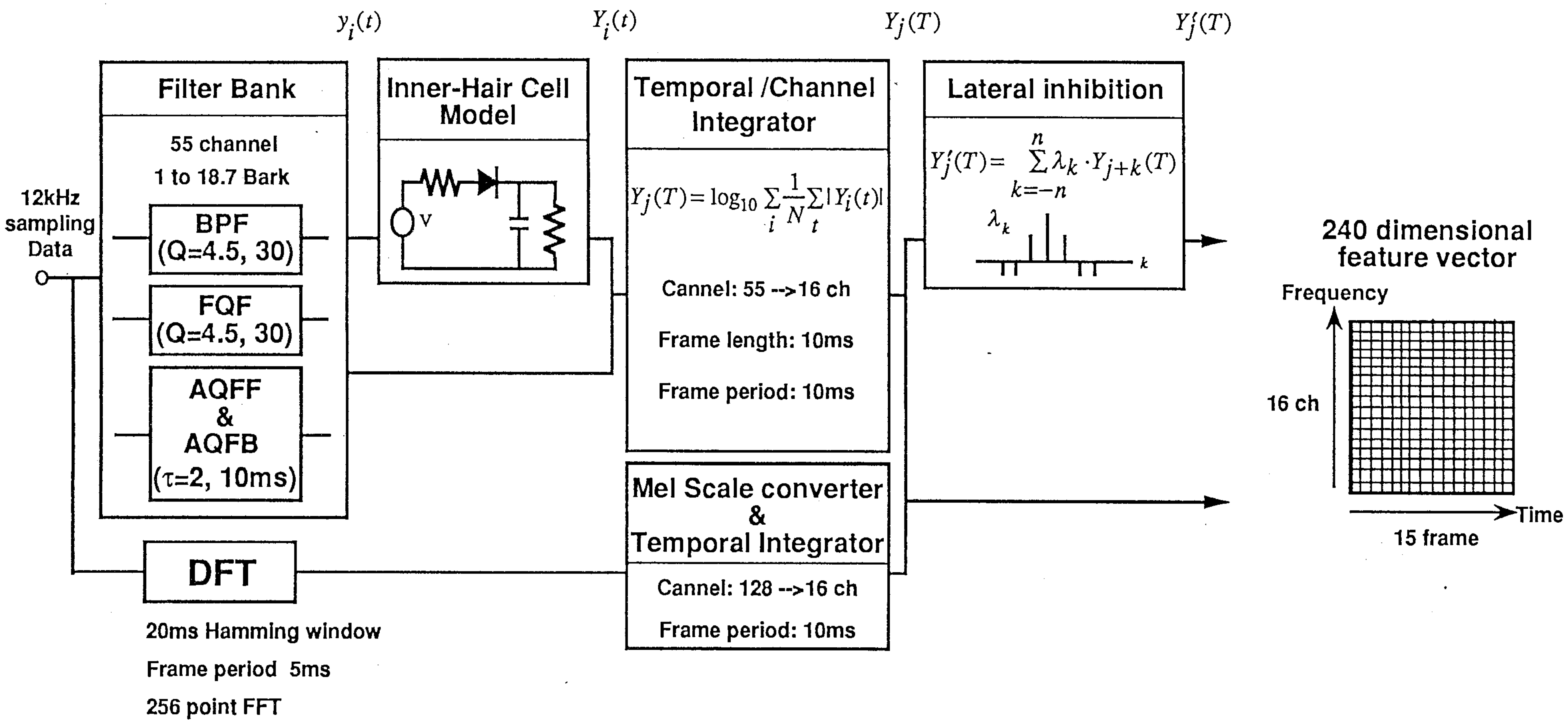
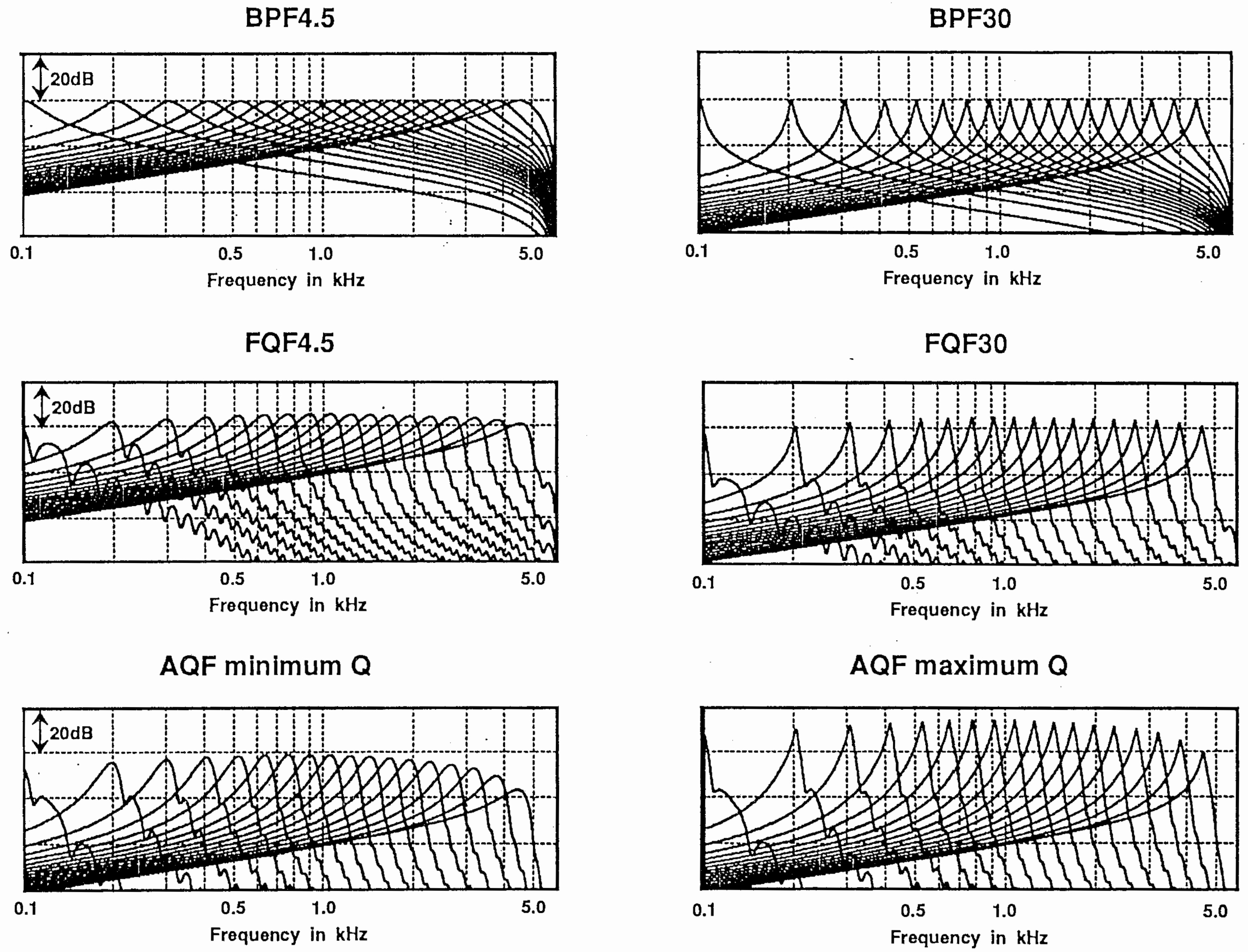


Figure 3

T. Hirahara & H. Iwamida



BPF4.5

BPF30

FQF4.5

FQF30

AQF minimum Q

AQF maximum Q

Figure 4

T. Hirahara & H. Iwamida

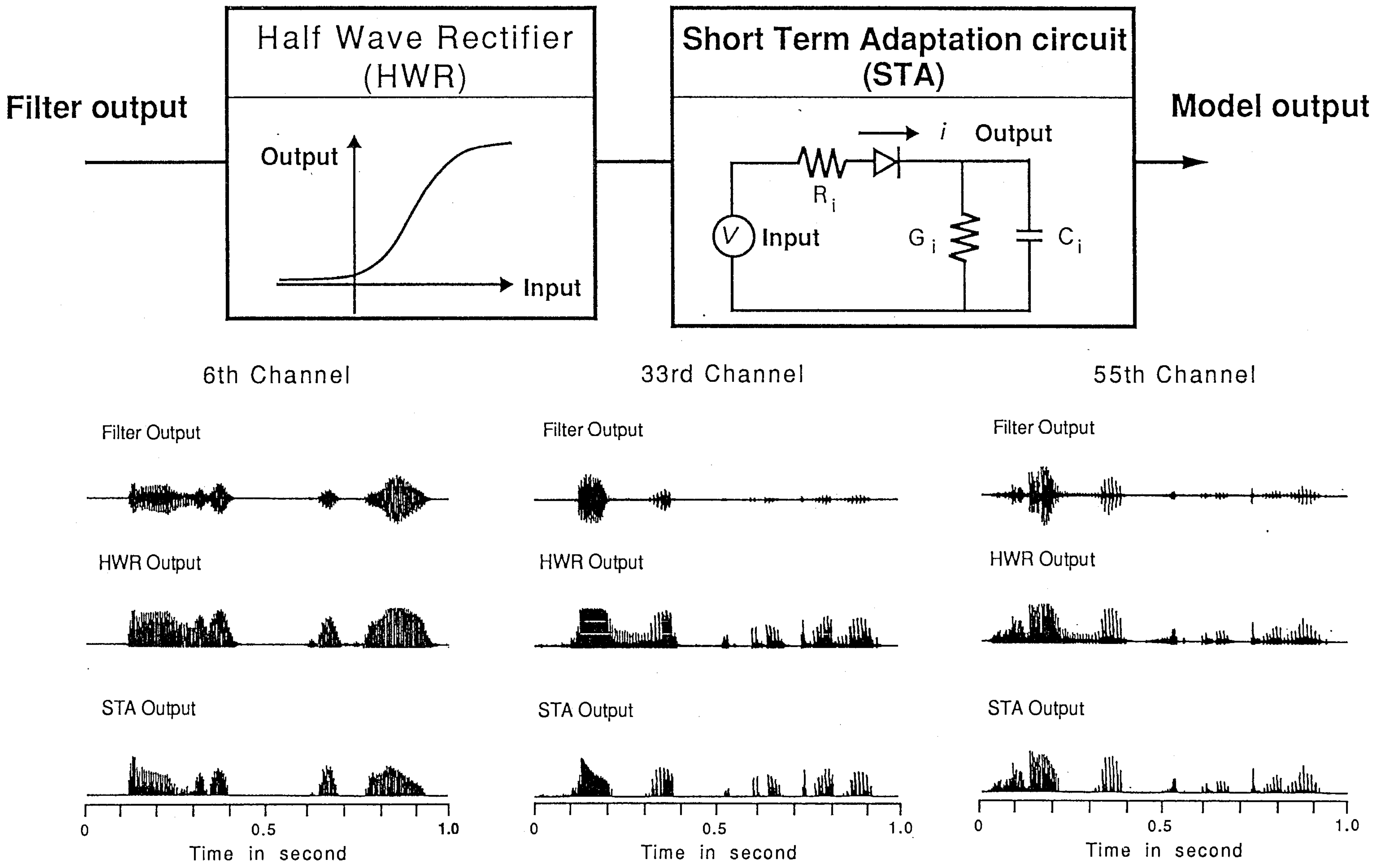


Figure 5

T. Hirahara & H. Iwamida

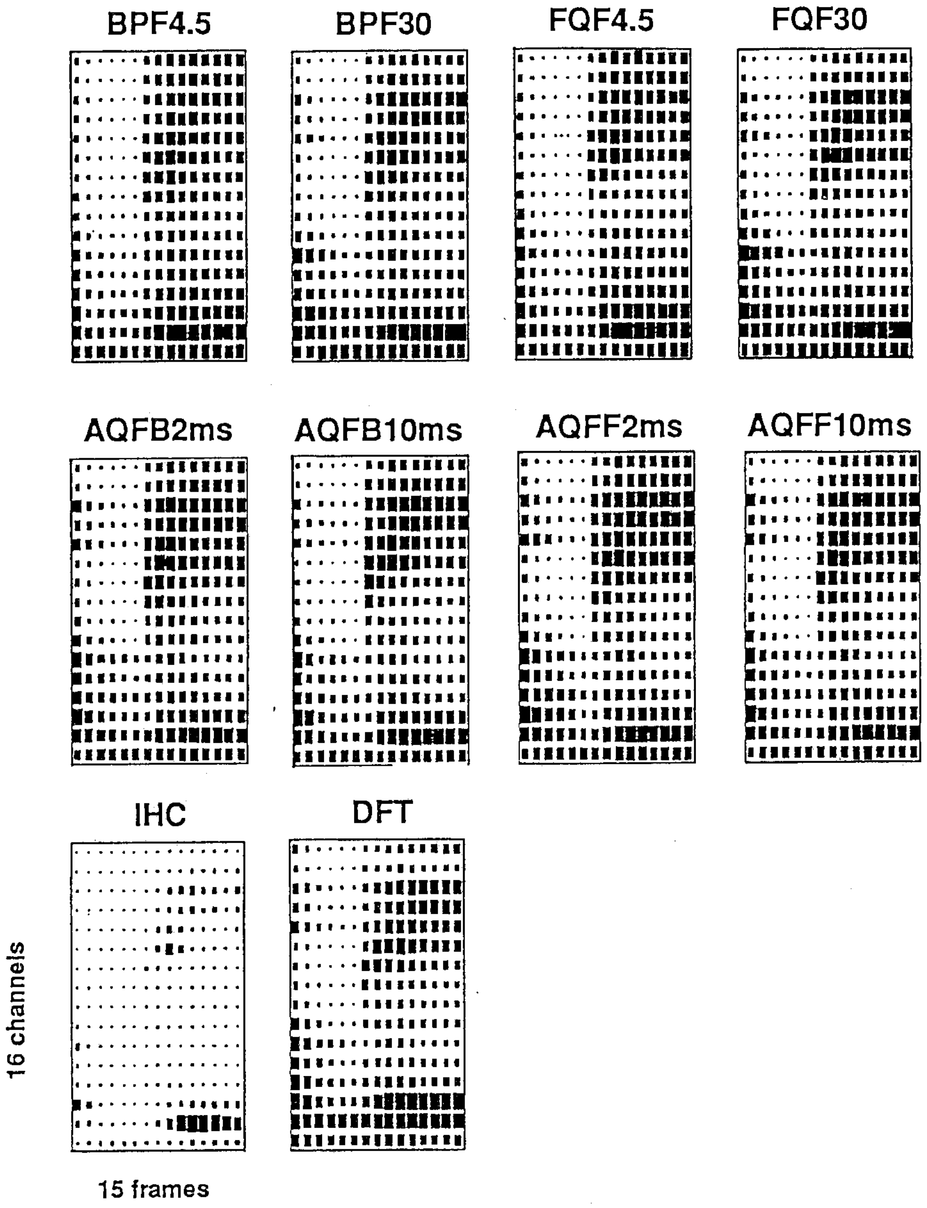


Figure 6 (a)

T. Hirahara & H. Iwamida

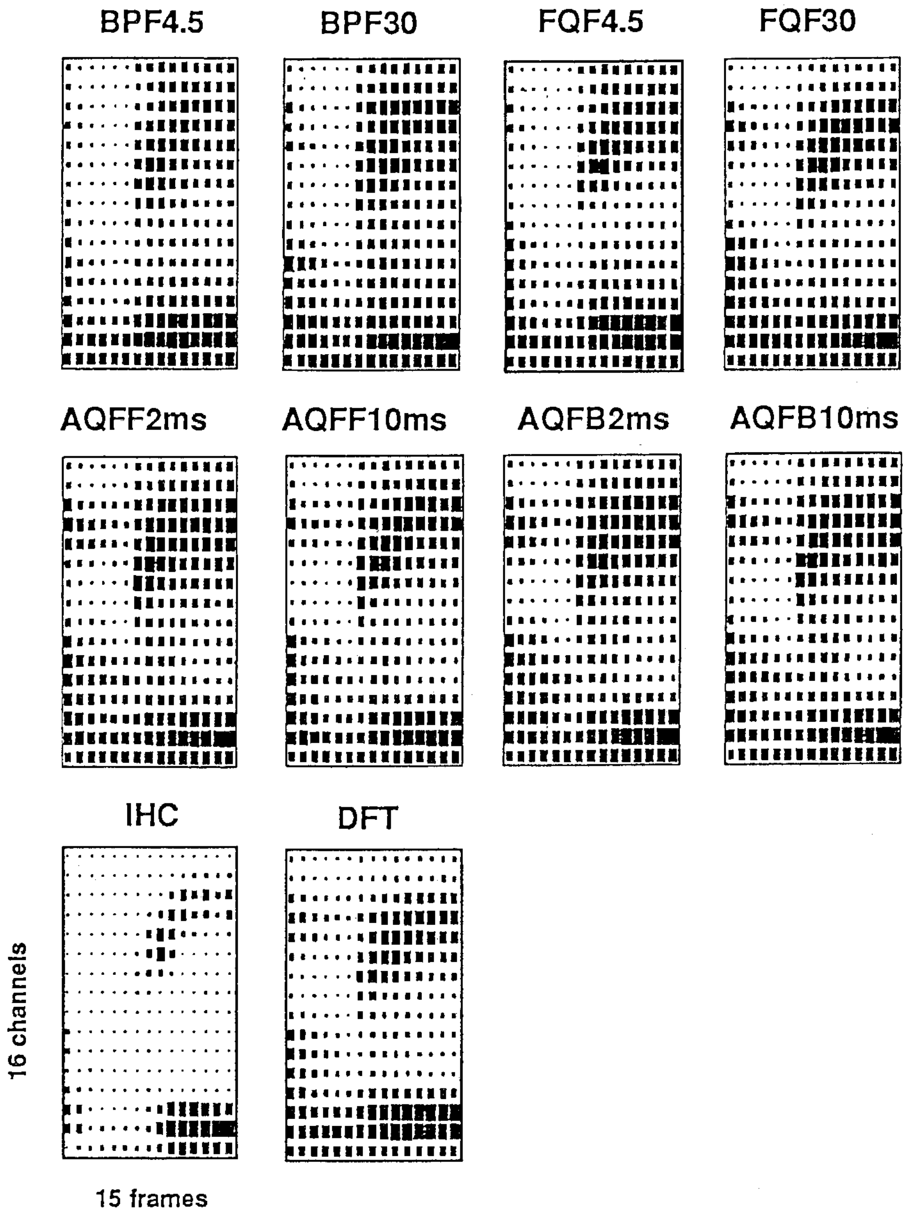


Figure 6 (b)

T. Hirahara & H. Iwamid

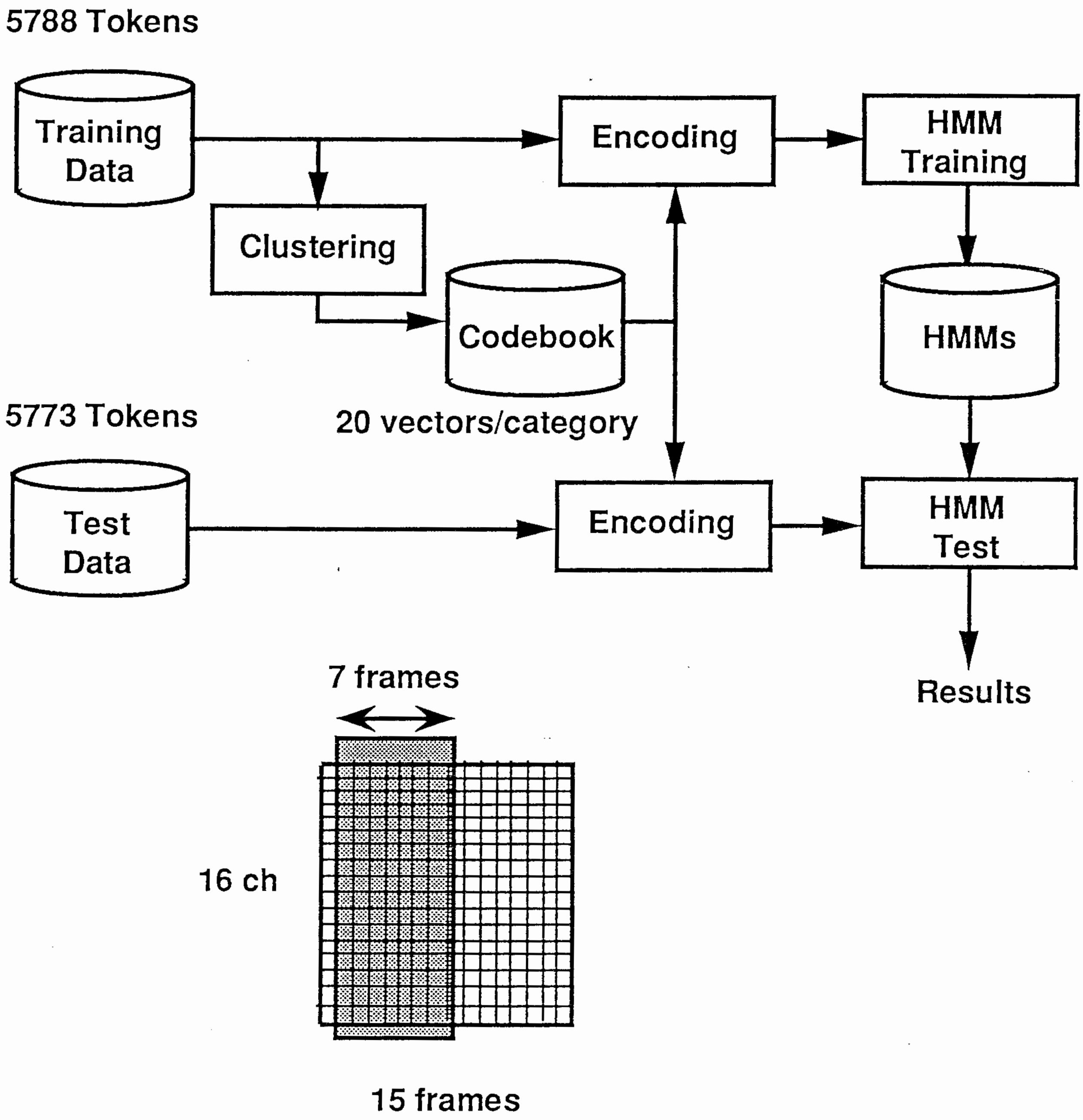
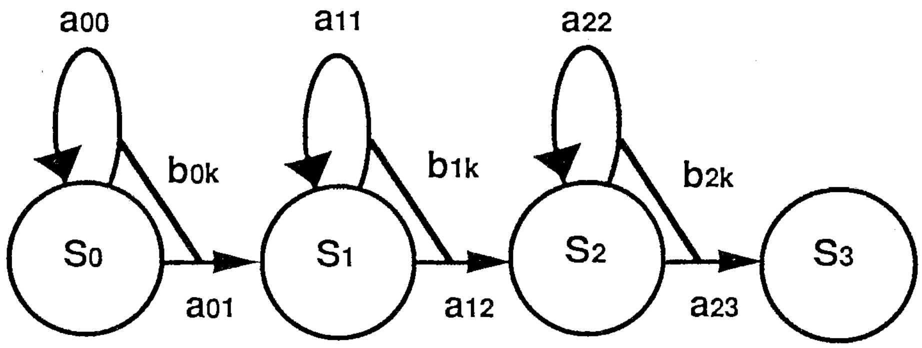


Figure 7 (a)

T. Hirahara &
H. Iwamid



S: State
a: Transition probability
b: Output Probability

Figure 7 (b)

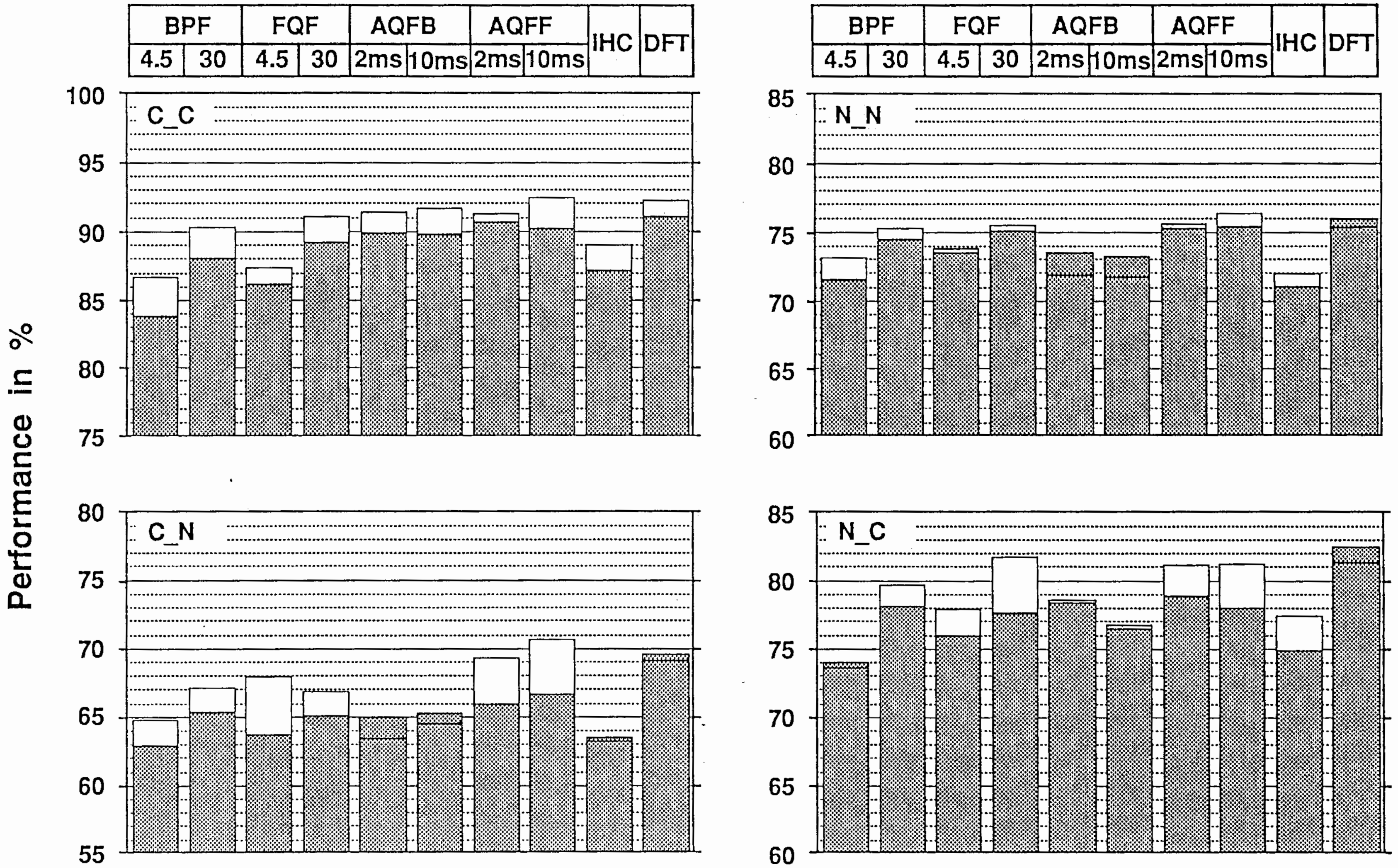

BPF
FQF
AQFB
AQFF
IHC
DFT
4.5
30
4.5
30
2ms
10ms
2ms
10ms
Performance in %
C_C
100
95
90
85
80
75
N_N
85
80
75
70
65
60
C_N
80
75
70
65
60
55
N_C
85
80
75
70
65
60

Figure 8

T. Hirahara & H. Iwamida

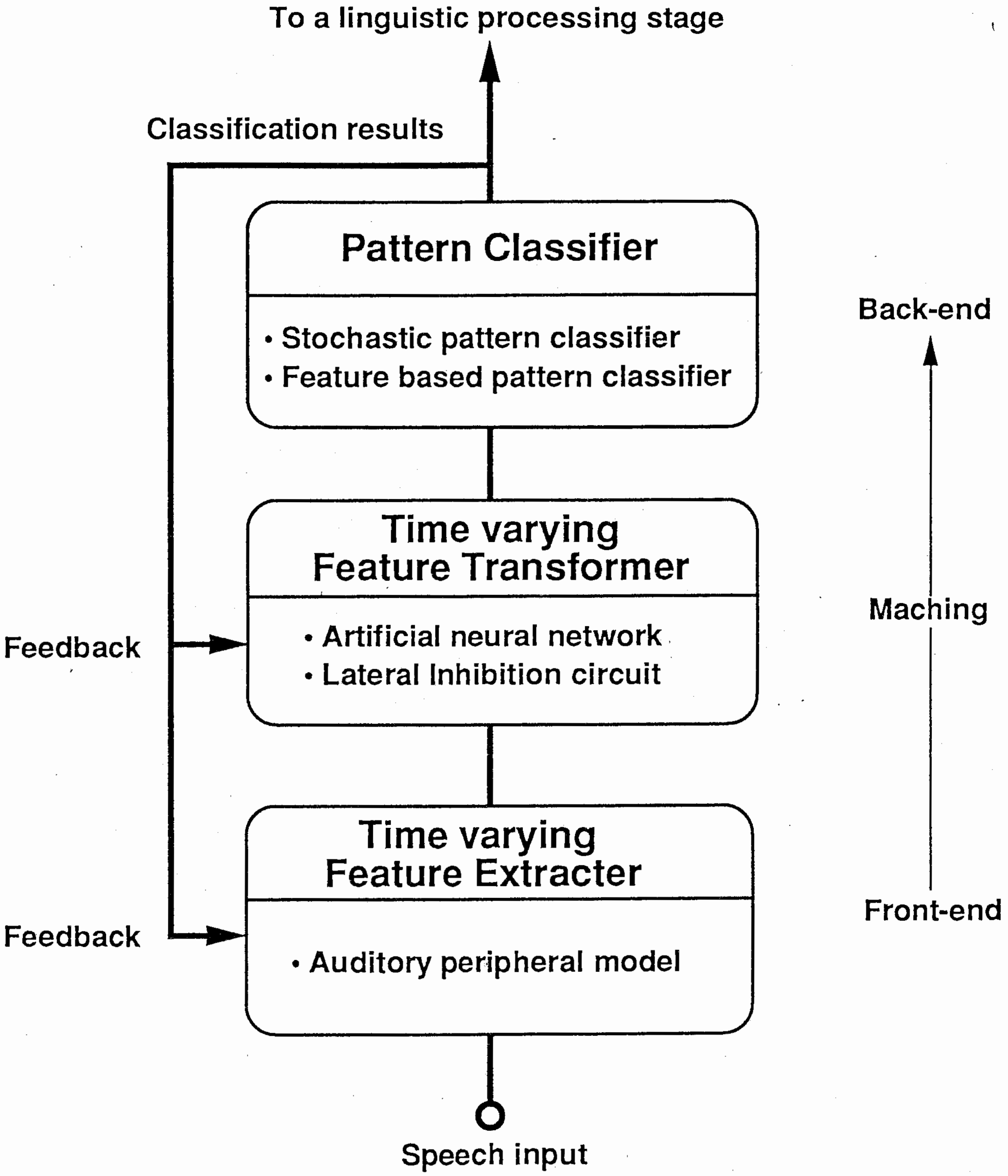

To a linguistic processing stage
Classification results
Pattern Classifier
• Stochastic pattern classifier
• Feature based pattern classifier
Time varying Feature Transformer
• Artificial neural network
• Lateral Inhibition circuit
Time varying Feature Extracter
• Auditory peripheral model
Feedback
Feedback
Speech input
Back-end
Maching
Front-end

Figure 9