論文 / 著書情報
Article / Book Information

Title	On the Use of Hierarchical Spectral Dynamics in Speech Recognition
Author	SADAOKI FURUI
Journal/Book name	IEEE ICASSP1990, Vol. , No. , pp. 789-792
発行日 / Issue date	1990, 4
権利情報 / Copyright	(c)1990 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Reprinted from PROCEEDINGS OF THE ICASSP 90 — 1990 INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, Albuquerque, New Mexico, April 3-6, 1990

S15a.10

ON THE USE OF HIERARCHICAL SPECTRAL DYNAMICS IN SPEECH RECOGNITION

Sadaoki Furui

NTT Human Interface Laboratories
Musashino-shi, Tokyo, 180 Japan

ABSTRACT

This paper proposes a recognition method which uses hierarchical spectral dynamic features extracted over multiple time lengths and it shows the effectiveness of these features in phoneme recognition and isolated word recognition. Speaker-independent isolated word recognition experiments are performed using a vocabulary of 100 Japanese words. Input speech is quantized by word-specific codebooks created as subsets of a universal codebook. When VQ distortion is used for word identification, a high recognition accuracy of 96 % is achieved, and when VQ distortion is used for preprocessing, the number of word candidates for each input utterance is reduced to 1 % of the vocabulary without increasing the error rate. Phoneme recognition experiments are performed for the /b/, /d/ and /g/ consonants in a large vocabulary of isolated words uttered by one male speaker. Using the proposed recognition method, a high recognition accuracy of 99 % is obtained. This paper also compares multiple codebook and single codebook methods.

I. INTRODUCTION

Dynamic (transitional) spectral features, as represented by Δcepstrum coefficients (Δcepstra) and Δpower, are widely used in speech recognition and their effectiveness for improving recognition accuracy has been verified in a large number of recognition systems [1][2]. Δcepstra and Δpower are linear regression coefficients of the time sequences of cepstrum coefficients (cepstra) and logarithmic energy. We have also previously reported that a VQ-based preprocessor, which uses word-specific codebooks created for the time functions of feature vectors consisting of cepstra, Δcepstra and Δpower, can effectively select word candidates in large vocabulary recognition [3]. In speech perception experiments, it has been confirmed that dynamic spectral features play important roles in the syllable and phoneme perception [4].

In previous methods, the number of speech frames or the time length for calculating regression coefficients has usually been set empirically at an optimum value of between 40 and 100 ms. Important dynamic features in speech spectra are, however, considered to exist not only over a single time length, but also over various time lengths corresponding to the duration of the phonemes and syllables, and their concatenations. This paper proposes a new VQ-based method for phoneme and word recognition, in which spectral dynamics over multiple time lengths are hierarchically extracted and used.

In VQ-based recognition systems, multiple features are, in many cases, represented by combining multiple codebooks, instead of creating a single codebook with all the features, to maintain small VQ distortion without increasing codebook size. However, the multiple codebook method inefficiently represents dependencies between different sets of features. This paper compares the performance of single and multiple codebooks under various conditions.

II. WORD RECOGNITION

2.1 Method

The principal structure of the word recognition system investigated in this section is shown in Fig. 1. A speech wave is analyzed by the time functions of the instantaneous cepstra, Δcepstra and Δpower. A universal VQ codebook for these time functions is then constructed based on a multi-speaker, multi-word database. Next, a separate codebook is designed for each word in the vocabulary as a subset of the universal codebook. These word-specific codebooks are used for front-end processing to eliminate word candidates whose distance or distortion scores are large. A dynamic time-warping (DTW) processor based on a word dictionary, in which each word is represented as a time-sequence of the universal codebook elements (SPLIT method), then resolves the choice among the remaining word candidates.

A vocabulary of 100 Japanese city names uttered by 20 male speakers (2000 samples) are used for the test utterances, and the same vocabulary uttered by a different set of four male speakers is used for the training utterances. These four speakers, considered to represent the entire range of male voices, were selected from among 30 male speakers. The speech waves are passed through a low-pass filter having a cut-off frequency of 4 kHz, and are digitized at a rate of 8-kHz. Word endpoints (beginning and ending positions) are detected, and linear predictive coding (LPC) analysis is performed on all frames within the word. The LPC analysis is a 10th-order analysis of 32-ms frames, spaced every 8 ms along the word. Each overlapping 32-ms section of speech is windowed using a Hamming window.

In previous research [3], the number of frames for calculating Δcepstra and Δpower was set at 7. To investigate the effectiveness

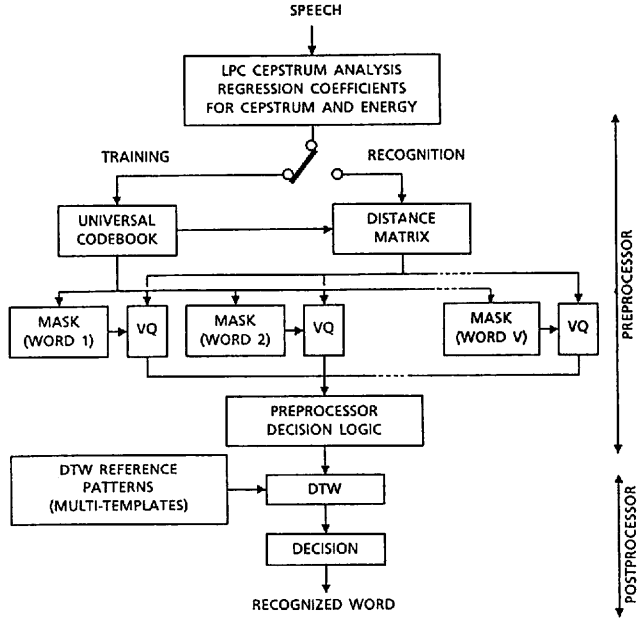

Fig.1 - Block diagram of word recognizer incorporating a VQ preprocessor and a SPLIT postprocessor.

of the hierarchical dynamic feature method, Δcepstra and Δpower extracted from 21 frames (168 ms) are combined with those extracted from 7 frames (56 ms) and instantaneous cepstra at each frame, as shown in Fig. 2. Δc_m and Δp_m in the figure indicate regression coefficients calculated over m frames for cepstra and log-energy, respectively. The codebook size for each word is set at 64, and the universal codebook size is varied among three values: 256, 512, and 1024.

The weighting factor for each feature parameter in the distortion (distance) calculation is set as follows. The variance values of all the coefficients are averaged for cepstra and Δcepstra sets separately, and all the coefficients of each set are uniformly weighted by inverse values of the mean variances. The weighting factor for Δpower is calculated by multiplying the inverse variance by a fixed value, which is empirically set based on preliminary experiments.

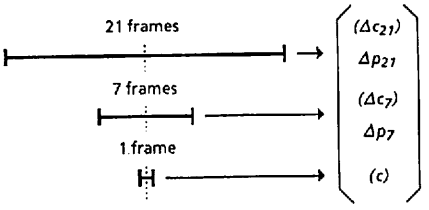

Fig.2 - Structure of feature vectors for the hierarchical dynamic feature method.

2.2 Recognition by VQ Distortion

Table 1 shows recognition error rates for the condition that only the top candidate is selected for each input speech based on VQ distortion. The table also shows the results of the DTW-based method (SPLIT method) for comparison.

In the case of the VQ-distortion-based method, four kinds of conditions indicated in the table have been tried. First, regression coefficients are calculated over 7 frames. Since instantaneous cepstra are always combined with regression coefficients to create a feature vector, the number of elements in the feature vector is 21: 10 cepstra, 10 Δcepstra, and one Δpower $(c, \Delta c_7, \Delta p_7)$. Second, the regression coefficients are calculated over 21 frames. Third, the weighted sum of the distortion values, D, with two codebooks, 7-frame and 21-frame, are used for the recognition decision;

$$D_{sum} = \{D(c, \Delta c_7, \Delta p_7) + \alpha D(c, \Delta c_{21}, \Delta p_{21})\}/(1+\alpha), \qquad (1)$$

where α is the weighting factor. Fourth, the 7-frame and 21-frame regression coefficients and cepstra are combined to build a long feature vector having 32 elements $(c, \Delta c_7, \Delta p_7, \Delta c_{21}, \Delta p_{21})$, and a set of these vectors are used to create a single codebook. We refer to the third method as a multiple codebook method, and the other methods as single codebook methods.

The recognition results indicate that error rates can be greatly reduced by summing the two distortion values for the 7-frame and 21-frame codebooks. For example, when universal codebook size is 1024, the error rate for this method is 6.65 %, whereas the error rates for the single codebook methods using either 7 or 21-frame codebooks are 10.90 % and 10.40 %, respectively. It is also indicated that a single codebook that includes hierarchical dynamic features, that is, 7-frame and 21-frame regression coefficients, produces much better results than the distortion summation method using multiple codebooks. The error rate for this method is 3.60 %, which is almost half the error rate of the SPLIT method without using regression coefficients (6.50 %). These results indicate the effectiveness of the hierarchical dynamic features.

As for the universal codebook size, it is indicated that the results for 1024 are better than those for 512. An additional experiment indicates that 1024 is large enough, and that no further improvement can be obtained, even when there is no binding using a universal codebook.

Table 1 - Word recognition error rates

Method	Codebook	Codebook Size 256	512	1024
VQ Distortion*	Single $(c,\Delta c_7,\Delta p_7)$	—%	13.45%	10.90%
	Single $(c, \Delta c_{21}, \Delta p_{21})$	—	16.45	10.40
	Multiple $D(c,\Delta c_7,\Delta p_7) + \alpha D(c,\Delta c_{21},\Delta p_{21})$	—	9.05	6.65
	Single $(c,\Delta c_7, \Delta p_7, \Delta c_{21},\Delta p_{21})$	—	4.45	3.60
SPLIT	Single (c)	6.50	6.20	6.50
	Multiple $D(c) + \alpha D(\Delta c_7,\Delta p_7)$	2.15	2.35	2.25
	Single $(c,\Delta c_7,\Delta p_7)$	3.00	2.15	1.95

* Codebook size for each word is 64.

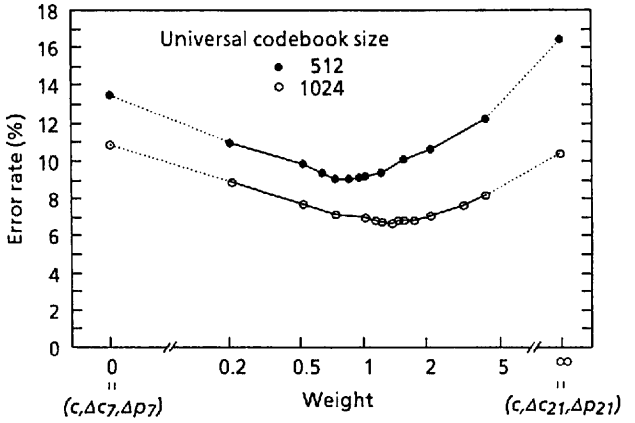

Fig.3 - Error rates for word recognition based on the weighted sum of the two VQ-distortions obtained using the 7-frame and 21-frame regression coefficient codebooks, respectively.

The results for the VQ distortion summation methods shown in the table are those obtained with the optimum weighting factor for summing the two distortion values associated with the different codebooks. Figure 3 shows the error rate as a function of the weighting factor α. It can be observed that the error rate does not fluctuate greatly, and that optimum results can be easily obtained when the weighting factor is set at around 1.

2.3 Recognition by SPLIT Method

In the experiments using the SPLIT method, three conditions have been tried. First, a single codebook consisting of only cepstra is used. Second, multiple codebooks for cepstra and 7-frame regression coefficients are used. Third, a single codebook including cepstra and regression coefficients is created.

As for the universal codebook size, the results in the Table 1 indicate that 256 is large enough in the cases of single cepstra codebook and multiple codebook methods. When two kinds of features consisting of cepstra and regression coefficients are used and the codebook size is set at 256, the multiple codebook method produces much better results than the single codebook method. This result is similar to the results reported by Lee [2]. However, the error rate with a single codebook including these two kinds of features decreases as the codebook size increases, up to 1024. When the codebook size is 1024, the single codebook method achieves better results (1.95 %) than the multiple codebook method

(2.25 %). This means that instantaneous and dynamic features have word-specific correlation characteristics. A supplementary experiment using this method indicates that 1024 is large enough and the error rate under this condition (1.95 %) is almost the same as the results without vector quantization.

2.4 VQ-Based Preprocessing

As the next step, VQ-based preprocessing for selecting candidate words has been investigated. In this method, words whose VQ distortion scores are smaller than the designated threshold are selected as the candidate words. Threshold θ is determined by

$$\theta = \theta_0 + Min\{D_n\}, \qquad (2)$$

where D_n is the distortion using the codebook for the nth word, and θ_0 is a fixed bias which is set at a predetermined value. Therefore, the threshold varies according to the minimum distortion for each input utterance [3].

Figure 4 shows the results for the multiple codebook method, in which 7-frame and 21-frame codebooks are separately built and the distortion values with these codebooks are summed using an optimum weighting factor. On the other hand, Fig. 5 shows the results for the single codebook method incorporating hierarchical dynamic features. These figures indicate the preprocessing error rate, that is, the probability that the true word is not selected, and the candidate/vocabulary ratio, that is, the ratio of the number of selected words to vocabulary size, as a function of bias θ_0.

By comparing Figs. 4 and 5, it can be concluded that the single codebook method reduces the number of candidates more effectively than the multiple codebook method. When $\theta_0 = 0.2$, the probability that the true word is included in the candidates is 99.95 % for both methods. However, the averaged numbers of selected words are 1.67 and 6.61 out of 100 for the single codebook and multiple codebook methods, respectively.

Figure 6 shows the probability of obtaining a single candidate for each input utterance. When $\theta_0 = 0.2$, a single candidate is selected for 71.8 % of the input utterances with the single codebook method, and for 28.0 % of the input utterances with the multiple codebook method. When only one candidate is selected, it is unnecessary to perform DTW calculation. Consequently, when preprocessing using the single codebook is applied, DTW calculation is necessary for only 0.95 % of the input utterances.

III. PHONEME RECOGNITION

3.1 Method

In addition to isolated word recognition, phoneme recognition has also been tried using codebooks including spectral dynamic features. In this experiment, /b/, /d/ and /g/ utterances are employed. They are contained in a 5240 word database, which was spoken by a professional announcer (Speaker MAU) and labeled at the ATR Lab. Previous experimental results of speech perception research using Japanese syllables whose initial and/or final parts were truncated indicated that the most important information for the syllable and phoneme perception exists in the speech waves of the short periods where the amount of spectral transition from consonant to vowel achieves maximum [4]. Based on this knowledge, fixed length periods, having boundaries between consonants and following vowels at their centers, are extracted from the database and used for recognition experiments.

The numbers of /b/, /d/ and /g/ periods included in the database are 445, 382 and 512, respectively. From them, odd-numbered periods are used for training and even-numbered periods are used for testing. Although analysis conditions are almost the same as those for word recognition experiments, the frame length and frame period are shortened slightly to 25 ms and 7.5 ms, respectively, since phoneme information is considered to be more precise than word information. The number of frames for extracting Δcepstra and Δpower is varied between 3 and 11 in the experiments.

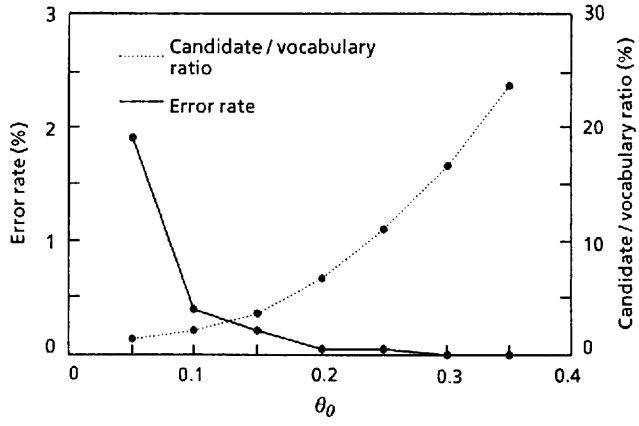

Fig.4 - Preprocessor performance for the multiple codebook method as a function of the threshold bias θ_0.

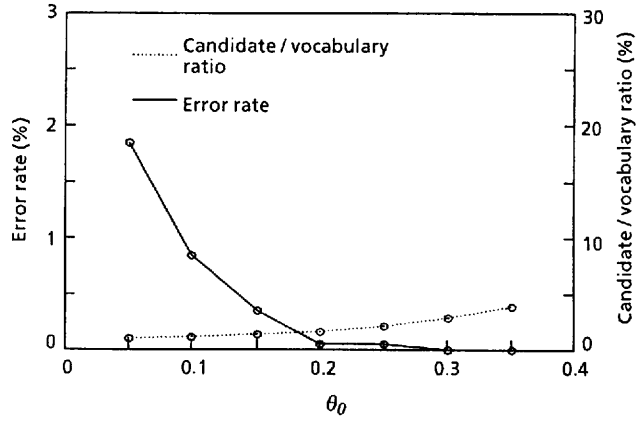

Fig.5 - Preprocessor performance for the single codebook method incorporating hierarchical dynamic features.

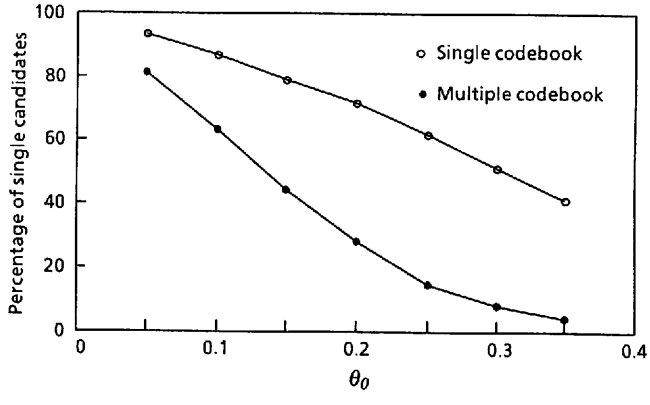

Fig.6 - Probability of obtaining a single candidate for each input utterance as a function of the threshold bias θ_0.

3.2 Effectiveness of Dynamic Features

In the first experiment, the length of the speech periods for training and recognition is set at 26 frames (195 ms), and the number of frames for regression analysis is set at 5 frames. Two

kinds of feature vectors are tested; instantaneous cepstra only *(c)*, and combination of instantaneous and dynamic features, that is, the combination of cepstra, Δcepstra and Δpower *(c, Δc_5, Δp_5)*. Table 2 shows recognition error rates obtained when the size of the codebook characterizing each phoneme is varied between 32 and 256. These results indicate that 64 and 128 are critical sizes of the codebooks in the instantaneous feature case and the instantaneous and dynamic features combination case, respectively. It also indicates that, in both cases, the error rates still decrease as the codebook size increases beyond these critical numbers. When the regression coefficients are combined with instantaneous cepstra, and the codebook size is set at 256, an error rate of 1.35 % is obtained, which is less than half of the error rate obtained using only cepstra. Based on these results, the regression coefficients are combined with instantaneous cepstra and the codebook size is set at 256 in the following experiments.

Experimental results show that the optimum number of frames for regression analysis is 5 or 7, and that variations in the error rate as a function of the speech length used for training and recognition are small within the range of between 8 and 22 frames. When these numbers of frames are set at 5 (37.5 ms) and 12 (90 ms) frames, respectively, the minimum error rate of 1.20% is obtained.

Table 2 - /b/,/d/,/g/ recognition error rates

		Codebook size			
		32	64	128	256
Codebook (Feature vector)	*(c)*	7.32 %	3.89 %	3.74 %	2.99 %
	(c, Δc_5, Δp_5)	—	3.44	1.49	1.35

3.3 Single Hierarchical vs. Multiple Dynamic Features

The performance of the single hierarchical dynamic feature codebook method and the multiple codebook method are compared, using regression coefficients extracted from 5 frame periods and 11 frame periods. Similar to the word recognition experiments, two distortion values associated with separate codebooks, *D(c, Δc_5, Δp_5)* and *D(c, Δc_{11}, Δp_{11})*, are summed in the multiple codebook method. On the other hand, hierarchical regression coefficients obtained for the two kinds of the speech length are combined with instantaneous cepstra and used for creating a single codebook *(c, Δc_5, Δp_5, Δc_{11}, Δp_{11})* in the single codebook method.

Figure 7 compares recognition error rates for the multiple and single codebook methods. The former results are shown as a function of the weighting factor for *D(c, Δc_{11}, Δp_{11})*, whereas the latter are shown as a function of the weighting factor for the 11 frame regression coefficients. Although the effectiveness of combining the regression coefficients obtained from speech periods of different durations is small in both the multiple and single codebook methods, the effectiveness in the latter method is relatively larger than the former method. When the weighting factor is set at 0.4, minimum error rates can be obtained in both methods. The error rate with the single codebook method under this condition is 0.90 %. This result is almost the same or better than that obtained with the HMM or Neural Network method [5].

3.4 Phoneme Segment Spotting

The experiments so far have used speech periods extracted from word utterances based on manual labeling. In order to evaluate the effectiveness of the new method in more general situations, spotting-based /b/, /d/ and /g/ recognition has been tried. In this experiment, a single hierarchical dynamic feature codebook for each phoneme is constructed using 12-frame speech segments around consonant-to-vowel transitions. The length of the input speech for recognition is expanded to 300 ms containing the phonetic transitions. An 18 frame (135 ms) window is shifted at

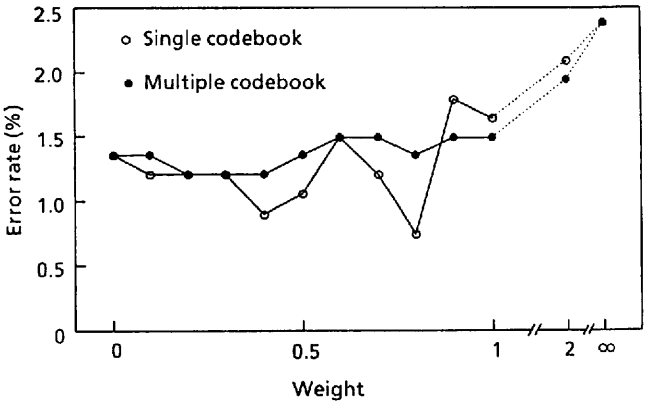

Fig.7 - Comparison of /b/, /d/, /g/ recognition error rates using the multiple codebook and single codebook methods as a function of the weighting factor combining two distortions.

every frame throughout the input speech, and feature vectors within the window are vector-quantized by the phoneme-specific codebooks. VQ distortion is averaged within the window, and the position of the window where the averaged distortion is the minimum is selected. The input speech is recognized to be the phoneme whose codebook produces the minimum distortion. A recognition error rate of 1.05 %, which is almost equal to the previous method, has been obtained. It confirms the robustness of the new method.

IV. CONCLUSION

A new VQ-based recognition method which uses feature vector codebooks containing hierarchical spectral dynamics has been proposed. This method is highly effective for reducing the number of candidates in word recognition and achieving a high recognition accuracy in /b/, /d/ and /g/ recognition. Since this method does not need time alignment, it has the advantage of a small amount of computation and the ease of parallel processing. Since it was suggested that recognition methods using transitional features are robust against variation in speaking rate [6], the recognition method proposed in this paper is expected to have an advantage over conventional methods.

Experimental results comparing the performance of the multiple codebook and single codebook methods indicate that, when the codebook size is restricted to being small, the multiple codebook method is better than the single codebook method. However, if the codebook size is reasonably large, the single codebook method displays better performance than the multiple codebook method.

REFERENCES

[1] S.Furui,"Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans. Acoust., Speech, Signal Processing, ASSP-34, 1, pp.52-59 (1986)

[2] K.-F. Lee,"Large-vocabulary speaker-independent continuous speech recognition using HMM", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, New York, S3.7 (1988)

[3] S.Furui,"A VQ-based preprocessor using cepstral dynamic features for speaker-independent large vocabulary word recognition" IEEE Trans. Acoust., Speech, Signal Processing, ASSP-36, 7, pp.980-987 (1988)

[4] S.Furui,"On the role of spectral transition for speech perception", J. Acoust. Soc. Amer., 80(4), pp.1016-1025 (1986)

[5] A.Waibel et al., "Phoneme recognition using time-delay neural networks", IEEE Trans. Acoust., Speech, Signal Processing, ASSP-37, 3, pp.328-339 (1989)

[6] M.Akagi and Y. Tohkura.,"On the application of spectral target prediction model to speech recognition", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, New York, S3.11 (1988)