SPEECH RECOGNITION RESEARCH AT ATR

Kiyohiro Shikano, Takeshi Kawabata, Alex Waibel,
Kaichiro Hatazaki, Hidefumi Sawai, Satoshi Nakamura,
Toshiyuki Hanazawa, Kenji Kita, Akira Kurematsu
(ATR Interpreting Telephony Research Laboratories)

Abstract

Speech recognition research activities in ATR Interpreting Telephony Research Laboratories are briefly described. The activities are summarized as follows:
(1) Hidden Markov phoneme models have been improved and successfully applied to Japanese phrase utterance recognition combining with the LR predictive parser.
(2) A phoneme segmentation expert based on spectrogram reading knowledge has been developed.
(3) Time-Delay Neural Networks (TDNN) have been applied to phoneme recognition in word utterances.
(4) Speaker adaptation algorithms have been improved using separate vector quantization and fuzzy vector quantization.

1. Introduction

An automatic telephone interpretation system is a facility which enables a person speaking in one language to communicate readily by a telephone with someone speaking another language. At least three constituent technologies are necessary for such a system: speech recognition, machine translation and speech synthesis. Moreover, integration research of these technologies are also very important. We propose an interpreting telephony model shown in Figure 1-1. In this model, the language processing is split into a language source model stage and a language analysis stage. Main targets of our research laboratories are fundamental research of speech and language processing and integrations of speech and language processing technologies to show the feasibility of an automatic telephone interpretation system.

In this paper, we describe speech recognition research efforts in ATR Interpreting Telephony Research Laboratories. Efforts aimed at speaker-dependent phoneme recognition and speaker-independent phoneme segmentation have resulted in dramatically improved phoneme recognition performances. We are now pursuing three approaches. They are (1) Hidden Markov Model approach for continuous speech recognition, (2) Feature-Based approach especially for accurate phoneme segmentation, and (3) Neural Network approach for accurate phoneme recognition. These research progresses are summarized in Section 2, 3, and 4, respectively. For speaker-independent speech recognition, a speaker adaptation approach has been undertaken using a concept of Vector Quantization and Spectrum Mapping, whose research progress is summarized in Section 5. These researches have been carried out using ATR developed Japanese large scale speech database with phoneme transcription.

2. Continuous Speech Recognition by Hidden Markov Modeling

HMM phoneme models have been improved and successfully combined with the LR predictive parser to recognize Japanese phrase utterances.

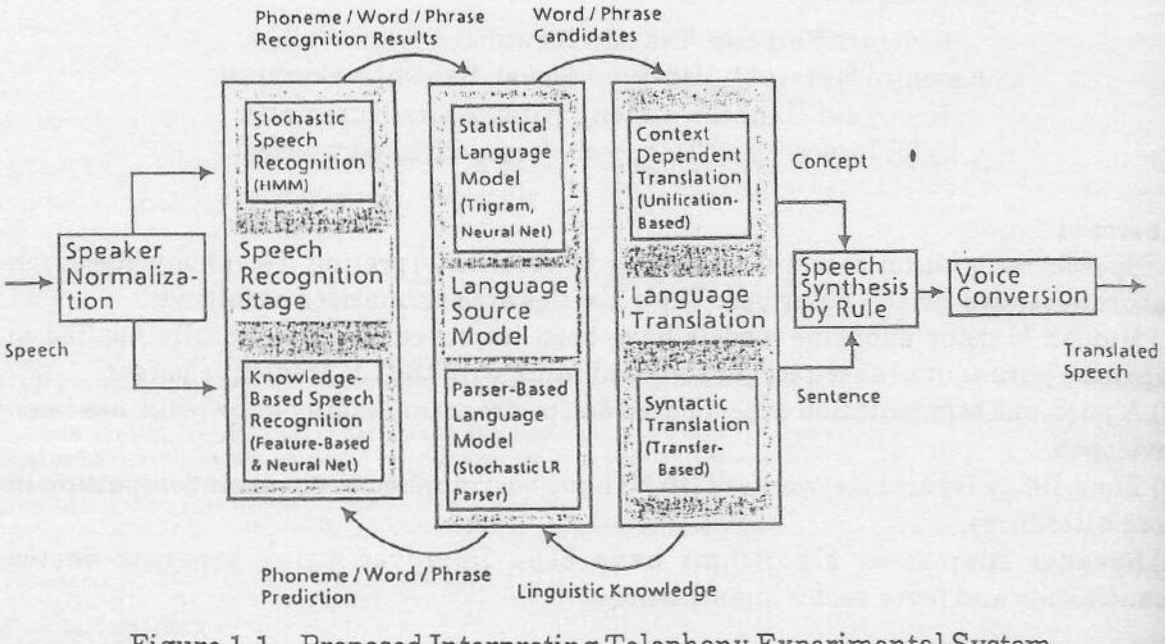
Phoneme / Word / Phrase Recognition Results
Word / Phrase Candidates
Stochastic Speech Recognition (HMM)
Statistical Language Model (Trigram, Neural Net)
Context Dependent Translation (Unification-Based)
Concept
Speaker Normalization
Speech
Speech Recognition Stage
Language Source Model
Language Translation
Speech Synthesis by Rule
Voice Conversion
Translated Speech
Sentence
Knowledge-Based Speech Recognition (Feature-Based & Neural Net)
Parser-Based Language Model (Stochastic LR Parser)
Syntactic Translation (Transfer-Based)
Phoneme / Word / Phrase Prediction
Linguistic Knowledge

Figure 1-1. Proposed Interpreting Telephony Experimental System.

2.1. Improvement of HMM Phoneme Models [4,11]

The following techniques are introduced and evaluated for discrete HMM phoneme recognition [5].

(a) Duration control techniques [3],
(b) Separate vector quantization techniques [1],
(c) Fuzzy VQ techniques [2].

These techniques are evaluated on phoneme recognition in word utterances using large (2,620 words) and small (216 words) size training data sets.

Effective duration control is realized by combining two duration control techniques. One is a phoneme duration control for each HMM phoneme model and the other is a state duration control for each HMM state. The phoneme duration control is carried out by weighting HMM output probabilities with phoneme duration histograms obtained from training sample statistics. State duration control is realized by state duration penalties calculated by modified forward-backward probabilities of training samples.

The separate vector quantization techniques for HMM phoneme recognition is useful for reducing VQ distortion. In our case, spectral features, spectral dynamic features [6] and energy are quantized separately. In the training stage, output vector probabilities of these three codebooks are estimated simultaneously and independently, and in the recognition stage the whole output probabilities are calculated as a product of output vector probabilities in these codebooks.

HMM training procedures are performed using the large training data (2,620 words) set uttered by one male speaker. Recognition experiments for male speakers are carried out using another 2,620 word set, which is composed of different words and is uttered by the same speaker. Phoneme boundaries are specified accurately by visual examination of spectrogram outputs. The phoneme boundary information is used in training procedures and recognition experiments to use the boundary information.

Improvements of the recognition rates using the large training data set are shown in Table 2-1, where (a) uses a single codebook for spectral features and energy, (b) uses duration control techniques with a single codebook, (c) uses three separate codebooks for spectral features, spectral dynamic features, and energy, and (d) uses duration control techniques with three separate codebooks for spectral features, spectral dynamic features, and energy. Duration control and separate codebook techniques are effective for HMM phoneme recognition. These recognition experiments result in 7.5% improved phoneme recognition rate from 86.5% to 94.0% on the average of three speakers using the separate codebook techniques and duration control techniques.

The fuzzy VQ technique is effective for parameter smoothing when the number of training samples is insufficient, so this technique is evaluated using the small training data (216 words) set from a male speaker. The phoneme recognition rate is improved by about 7% as shown in Table 2-2.

Table 2-1. Phoneme Recognition Rates for Separate Codebooks and Duration Control. (2620 word training set)

speaker	(a) PWLR	(b) PWLR DUR	(c) WLR& DCEP& POW	(d) WLR& DCEP& POW DUR
MAU	84.8%	89.8%	93.2%	94.1%
MHT	90.1%	92.4%	95.2%	95.3%
MNM	84.5%	88.7%	91.9%	92.7%
average	86.5%	90.3%	93.4%	94.0%

Table 2-2. Phoneme Recognition Performances for Fuzzy VQ. (216 word training set, male speaker MAU)

	VQ	Fuzzy VQ
(a) PWLR	64.6%	72.1%
(c) WLR& DCEP&POW	70.9%	78.1%
(d) WLR& DCEP&POW DUR	-	80.9%

2.2. HMM Continuous Speech Recognition Using the LR Parser [7]

The HMM phoneme models are integrated with the generalized LR predictive parser as shown in Figure 2-1. The LR parser originally developed for compiler and extended to handle arbitrary context-free grammar [8]. An LR parser is guided by an LR table automatically created from context-free grammar rules, and proceeds left-to-right without backtracking. In the LR parsing mechanism, the next parser action (accept, error, shift, or reduce) is determined by looking up in the LR table with the current state of the parser and next input symbol. This parsing mechanism is valid only for symbolic data and cannot simply apply to continuous data such as speech.

In our approach, the LR table is used to predict the next phoneme in the speech input. For the phoneme prediction, the grammar terminal symbols are phonemes instead of the grammatical category names generally used in natural language processing. That is, a lexicon for the task is embedded in the grammar. The following describes the system operation. First, the parser picks up all phonemes which the initial state of the LR table is expecting, and invokes the HMM phoneme models to verify the existence of these expected phonemes. During this time, all possible parsing trees are constructed in parallel. The phoneme verifier (HMM phoneme model) receives a probability array, which includes end point candidates and their probabilities, and updates it using an HMM phoneme probability calculation process (trellis algorithm). This probability array is attached to each node of the partial parsing tree. When the highest probability in the array is below a threshold level, the parsing tree is pruned, and also pruned by a beam searching algorithm. The parsing process stops if the parser detects an accept action in the LR table and an end of an utterance.

This integration algorithm is applied to Japanese phrase recognition, whose task is the secretary service of the international conference. Utterances are uttered phrase by phrase. The syntax of phrases includes a general Japanese syntax structure of phrases, whose perplexity a phoneme is about five. Supposing that the average phoneme length of words is four, the perplexity of words is more than six hundreds.

The HMM phoneme models are trained using 5240 words. The duration control parameters are modified according to the ratio of utterance speed between word utterances and phrase utterances. The phrase recognition rate is 83% for 276 phrase inputs, as shown in Table 2-3.

The integration of the HMM and the LR parser is further developed to deal with continuous speech using a word spotting algorithm[21].

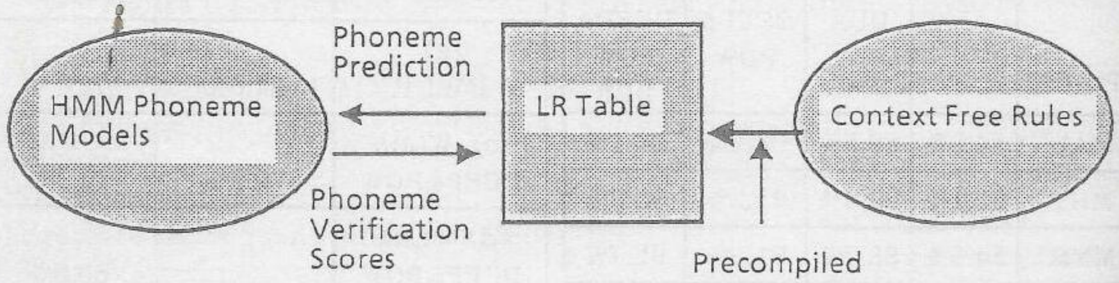

Figure 2-1. HMM-LR Continuous Speech Recognition System.

Table 2-3. Phrase Recognition Experiment Results

	Phrase Recognition Rate (%)
Phrase Recognition Rate	83.2 %
within Top TWO choices	94.3 %
within Top THREE choices	97.1 %

3. Phoneme Segmentation Using Spectrogram Reading Knowledge [9]

The phoneme segmentation approach by an expert system utilizing spectrogram reading strategy and knowledge used by human experts to read spectrograms is described. The expert system, into which the strategy and knowledge are incorporated, detects phonemes in continuous speech and determines their boundaries as well as their coarse categories. The system configuration is shown in Figure 3-1.

Since Zue and his colleagues [10] showed that a trained spectrogram reader is able to identify phonetic segments in an unknown speech spectrogram with high accuracy, several speech recognition systems based on spectrogram reading knowledge have been developed. The previous research proved the effectiveness of the experts' knowledge for phoneme identification rather than phoneme segmentation. However, human experts perform phoneme segmentation and identification simultaneously and, as the result, are able to determine the phoneme boundaries with high accuracy, as well as their categories. The method proposed here utilizes this experts' strategy and knowledge for phoneme segmentation in continuous speech. Phoneme boundaries obtained by this system are so accurate that the phonemes can be identified using a stochastic or neural network phoneme recognition method [4,12].

The expert system is constructed based on the experts' strategy and knowledge which can be expressed easily and naturally, as follows:

(a) The system adopts assumption-based inference, which makes it easy to describe segmentation rules depending on phonetic context. These rules are applied under their own phonetic context hypotheses separately. Hypotheses which are assigned large certainty factors survive.

(b) Acoustic features are extracted from the spectrogram when they are referred to by rules under certain hypotheses. This makes it possible to extract various kinds of global and local features.

(c) Some acoustic features are assigned certainty factors, which makes it possible to describe human experts' fuzzy knowledge. Distinct thresholds can be avoided.

Knowledge of Japanese phoneme segmentation is incorporated into the system and tested using continuously spoken Japanese words. The phoneme boundaries are compared to the boundaries labeled by a spectrogram reader, whose results are shown in Table 3-1. The result shows that the system achieves performance equal to human experts'. Especially, the boundary alignment error is small, that is, most of boundaries obtained are within 10 msec of the hand labeled boundaries.

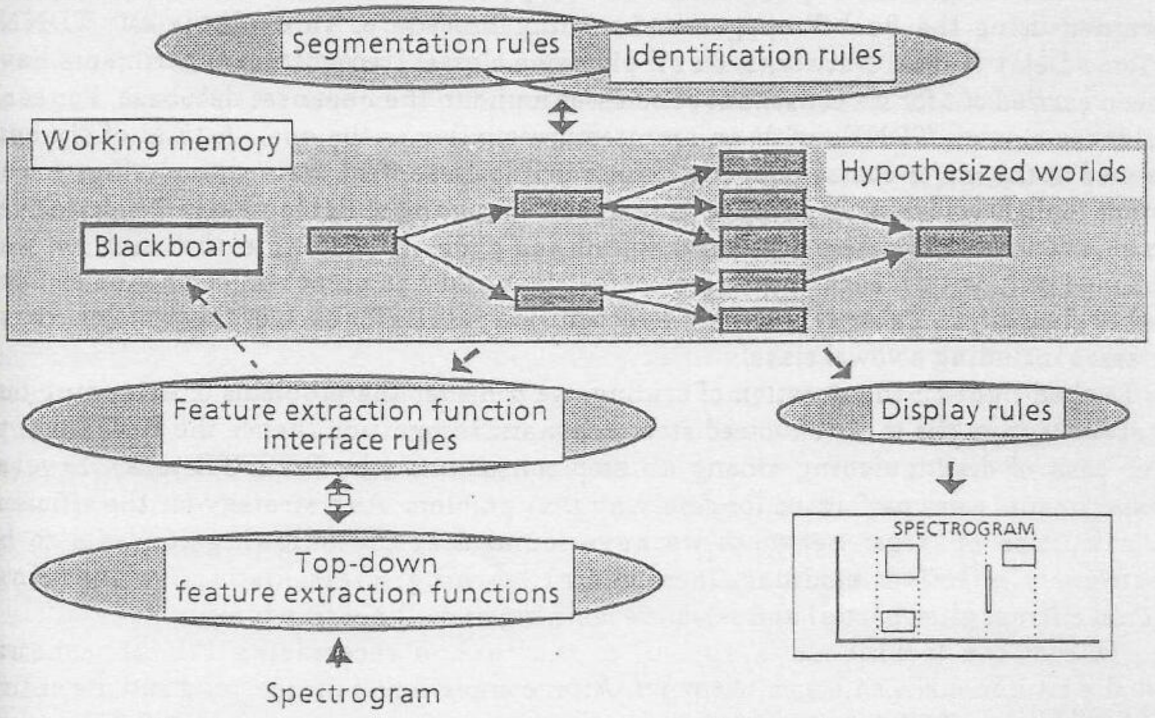

Figure 3-1. Phoneme Recognition Expert System Architecture.

Table 3-1. Segmentation Results for Unvoiced Fricatives.

Word set	Phoneme	Number of phonemes	Number of missed boundaries	
			Left	Right
(a) 216 words	/s/	32	1 (3%)	1 (3%)
	/sh/	25	1 (4%)	1 (4%)
	total	57	2 (3.5%)	2 (3.5%)
(b) 5240 words	/s/	1086	36 (3.3%)	38 (3.5%)
	/sh/	783	21 (2.7%)	25 (3.2%)
	total	1869	57 (3.0%)	63 (3.4%)

4. Phoneme Recognition by Neural Networks [12]

A number of studies have recently demonstrated that connectionist architectures capable of capturing some critical aspects of the dynamic nature of speech can achieve superior recognition performance for small but difficult phoneme discrimination tasks [13]. A problem that emerges, however, as we attempt to apply neural network models to the full speech recognition problem is the problem of scaling. In this section we demonstrate based on a set of experiments aimed at phoneme recognition that is indeed possible to construct large neural networks by exploiting the hidden structure of smaller trained subcomponent networks. A set of successful techniques is developed that bring the design of practical large scale connectionist recognition systems within the reach of today's technology.

For the recognition of phonemes, a four layer net is constructed. The network is trained using the Back-Propagation Learning Procedure. To evaluate our TDNNs (Time-Delay Neural Networks) on all phoneme classes, recognition experiments have been carried out for six consonant subclasses found in the Japanese database. For each of these classes, TDNNs with an architecture similar to the one. A total of six nets aimed at the major coarse phonetic classes in Japanese were trained, including voiced stops /b,d,g/, voiceless stops /p,t,k/, the nasals /m,n/ and syllabic nasals /N/, fricatives /s,sh,h/ and /z/, affricates /ch,ts/, and liquids and glides /r,w,y/ . Note, that each net was trained only within each respective coarse class and has no notion of phonemes from other classes yet. Table 4-1 shows the recognition results for each of these major coarse classes including a vowel class.

To shed light on the question of scaling, we consider the problems of extending our networks from the tasks of voiced stop consonant recognition (hence the BDG task) to the task of distinguishing among all stop consonants (the BDGPTK-task). Several experiments were performed for resolving that problem. As a strategy for the efficient construction of larger networks we have found that the following concepts to be extremely effective: modular, incremental learning, class distinctive learning, connectionist glue, partial and selective learning and all-net fine tuning.

One of the techniques is applied to the task of recognizing all consonants (/b,d,g,p,t,k,m,n,N,s,sh,h,z,ch,ts,r,w,y/). After completion of the learning run the entire net achieves a 95.0% recognition accuracy. All net fine tuning yields 96.0% correct consonant recognition over testing data. The TDNN consonant recognition rate of 96.0% is superior to the HMM rate of 93.8%.

Table 4-1. TDNN Phoneme Recognition Rates within coarse classes.

task	phoneme rec. rate
b,d,g	98.6 %
p,t,k	98.7 %
m,n,N	96.6 %
s,sh,h,z	99.3 %
ch,ts	100 %
r,w,y	99.9 %
coarse classes	96.7 %
a,i,u,e,o(vowels)	98.6 %

Table 4-2. TDNN All Consonant Recognition Rate after All-Net Fine Tuning.

task	Phoneme recognition rate
18 consonants	96.0 %
HMM	93.8 %

5. Speaker Adaptation by Fuzzy VQ and Spectrum Mapping [14,15,17]

This section describes an approach to speaker adaptation which is achieved by spectral mapping from one speaker to another. This algorithm realizes general speaker adaptation which does not depend on speech recognition systems as post-processing. Evaluation experiments on HMM and voice conversion [16] have already clarified the performance and general applicability.

The spectrum mapping method is based on the following three idears. The first is accurate representation of input vectors by separate VQ and fuzzy VQ. The second is accurate establishment of spectral correspondence based on fuzzy relation of membership function obtained from supervised training procedure by DTW. The third is continuous spectral mapping from one speaker to another by fuzzy mapping. In this algorithm, the input vector represented by fuzzy membership function is mapped onto the target speaker's space by fuzzy mapping theory. This fuzzy mapping allows continuous mapping of the input vector onto target speaker's space. These algorithms are evaluated from the viewpoint of spectral distortion. The evaluation results are summarized in Figure 5-1.

In the application to HMM, the input vector is represented as the weighted combination of fuzzy membership function U_{ai} and codevector. The mapping function calculated from the correspondence histogram h_{ij} is fuzzy relation between codevector i and codevector j of each speaker, therefore the output probability of HMM is calculated as a product of U_{ai} and h_{ij}. At the same time, separate vector quantization with spectrum, difference spectrum and power term is adopted as the product of each output probability. The /b,d,g/ recognition results are shown in Table 5-1.

Evaluation experiments are carried out and the results are as follows:

(a) Average intra-speaker VQ distortion is reduced by about 28% using fuzzy VQ techniques and k-nearest neighbor rule.

(b) Inter-speaker mapping distortion is reduced 10% using the fuzzy VQ and fuzzy continuous mapping technique rather than the conventional technique.

(c) The number of training words required for finding correspondence is reduced from 100 to 30.

(d) Phoneme recognition experiments on the /b,d,g/ task by HMM were carried out. The recognition rate for the /b,d,g/ task is 78% on average. Improvement of about 27% in the recognition rate is accomplished.

Phoneme recognition experiments on the TDNN neural networks through the speaker adaptation algorithms are being carried out.

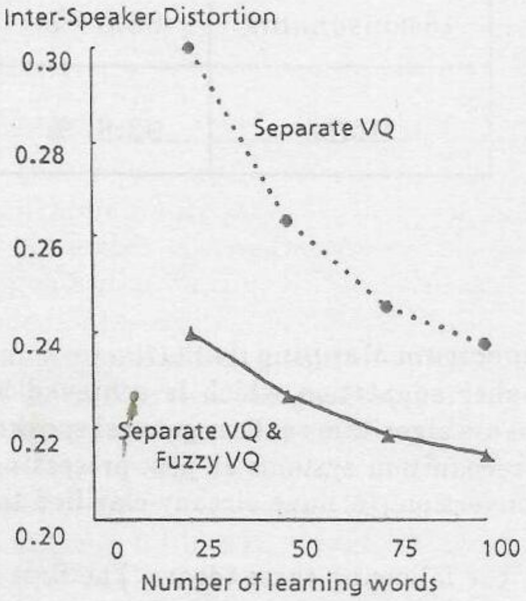

Figure 5-1. Speaker Adaptation Algorithm Evaluation by Spectral Distortion.

Table 5-1. /b,d,g/ Recognition Rates by HMM Speaker Adaptation, which is the average of male to male and male to female.

Method	Recognition Rate (%)
without adaptation	51.7
Mapped Codebook [22]	66.4
Fuzzy Mapping [23]	72.1
Fuzzy Mapping + SPVQ	73.2
Fuzzy Mapping + SPVQ + FZVQ	75.7
Fuzzy Mapping + SPVQD + FZVQ	78.1

SPVQ: Separate vector quantization with spectrum and power term

SPVQD: Separate vector quantization with spectrum and power and difference spectrum term

FZVQ: Fuzzy vector quantization

6. Summary

Speech recognition research activities in ATR were summarized. Besides of the above research activities, the following research activities have been also carried out.

(1) Word category prediction by N-gram neural networks [18].

(2) English word recognition by HMM phoneme models.

(3) Phoneme spotting by TDNN neural networks [20].

(4) Fast back-propagation algorithm for neural networks in speech [19].

(5) Continuous speech recognition using HMM word spotting and LR parser [21].

We are focusing our speech research on the speech recognition research itself and the integration with language processing to show the possibility of an automatic telephone interpretation system. Moreover, international research collaboration to handle many languages is highly needed to develop automatic telephone interpretation technologies.

References

[1] K.F.Lee, H.W.Hon,"Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM", ICASSP88, pp123-126, (1988-04)

[2] H.P.Tseng, M.J.Sabin, E.A.Lee,"Fuzzy Vector Quantization Applied to Hidden Markov Modeling", ICASSP87, (1987-04)

[3] S.E.Levinson,"Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition". Computer Speech and Language, 1, pp29-45, (1986)

[4] T.Hanazawa, T.Kawabata, K.Shikano, "Study of Separate Vector quantization for HMM Phoneme Recognition", ASJ Fall Meeting, 2-P-21, (1988-10) (in Japanese)

[5] R.Schwartz, Y.Chow, O.Kimball, S.Roucos,M.Kransner, J.Makhoul, "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech", ICASSP85, (1985-03)

[6] S.Furui,"Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", IEEE tr.ASSP, Vol.ASSP-34, No.1, (1986-02)

[7] K.Kita, T.Kawabata, H.Saito, "HMM Continuous Speech Recognition Using Predictive Parsing", Trans. Tech. Group Speech Acoust. Soc. Japan, S88- , (1988-10) (in Japanese)

[8] M.Tomita,"Efficient Parsing for Natural Language", Kluwer Academic Publishers, (1986)

[9] K.Hatazaki, S.Tamura, T.Kawabata, K.Shikano, "Phoneme Segmentation by an Expert System Based on Spectrogram Reading Knowledge", Speech88, 7th FASE Symposium, pp927-934, (1988-08)

[10]V.W.Zue, R.A. Cole, "Experiments on Spectrogram Reading,", ICASSP79, pp116-119, (1979-04)

[11]T.Hanazawa,T.Kawabata, K.Shikano,"Output Probability Smoothing for HMM Phoneme Recognition", ASJ Spring Meeting, 3-P-1, (1987-03) (in Japanese)

[12]A.Waibel, H.Sawai, K.Shikano,"Phoneme Recognition by Modular Constraction of Time-Delay Neural Networks", ASJ Fall Meeting, 2-P-12, (1988-10)

[13]A.Waibel, T.Hanazawa, G.Hinton,K.Shikano, K.Lang,"Phoneme Recognition: Neural Networks vs. Hidden Markov Models", ICASSP88, pp107-110, (1988-03)

[14]S.Nakamura,K.Shikano, "Spectrogram Normalization Using Separate Vector Quantization", Speech88, 7th FASE Symposium, pp31-38, (1988-08)

[15]S.Nakamura,K.Shikano, "Spectrogram Normalization Using Fuzzy Vector Quantization",Trans. Tech. Group Speech Acoust. Soc. Japan, S87-123 , (1988-02) (in Japanese)

[16]M.Abe, S.Nakamura,K.Shikano, H.Kuwabara, "Voice Conversion Through Vector Quantization", ICASSP88, pp655-658, (1988-04)

[17]S.Nakamura, T.Hanazawa, K.Shikano, "Phoneme Recognition Evaluation of HMM Speaker Adaptation Using Fuzzy Vector Quantization", ASJ Fall Meeting, 2-P-20, (1988-10) (in Japanese)

[18]M.Nakamura, K.Shikano, "A Study of N-Gram Word Category Prediction Based on Neural Networks", ASJ Fall Meeting, 2-P-2, (1988-10) (in Japanese)

[19]P.Haffner, A.Waibel, K.Shikano, "Fast Back-Propagation Learning Methods for Neural Networks in Speech", ASJ Fall Meeting, 2-P-1, (1988-10)

[20]H.Sawai, A.Waibel, K.Shikano, "A Preliminary Study on Spotting Japanese CV-Syllables by Time-Delay Neural Networks", ASJ Fall Meeting, 2-P-11, (1988-10) (in Japanese)

[21]T.Kawabata, K.Shikano,"Japanese Phrase Recognition Based on HMM Phone Units",ASJ Fall Meeting, 2-P-11, (1988-10) (in Japanese)

[22]K.Shikano, Kai-Fu Lee, Raj Reddy,"Speaker Adaptation through Vector Quantization", ICASSP86, pp2643-2646, (1986-04)

[23]M.Feng, F.Kubala,R.Schwarzt, J.Makhoul, "Improved Speaker Adaptation Using Text Dependent Spectral Mapping", ICASSP87, pp131-134, (1987-04)