TR-I-0061

Speech Research at ATR
Interpreting Telephony Research Laboratories

Speech Processing Department,

ATR Interpreting Telephony Research Laboratories

December , 1988

Abstract

Speech research activities at the speech processing department of the ATR Interpreting Telephony Research Laboratories are introduced.

First, speech recognition research activities are summarized as follows :

(1) Hidden Markov phoneme models have been improved and successfully applied to Japanese phrase utterance recognition combined 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.

Second, the research activities on speech synthesis, voice conversion and noise reduction are summarized as follows :

(1) The speech synthesis system proposed is a synthesis-by-rule based on an optimal selection of non-uniform synthesis units which aims at producing natural, high quality speech sounds.

(2) Voice conversion is a method to change voice individuality. The conversion method proposed here is to make use of conventional vector quantization technique. The essential part of this technique is to make mapping codebooks between two different speakers for such acoustic parameters as spectrum, pitch frequency, and power level. The conversion experiments reveal that this method is effective and promissing for the conversion of voice individuality.

(3) Noise reduction is another technique which the interpreting telephony system should incorporate. This is done by a four-layered neural network with the back-propagation learning algorithm. The result reveals that the network can indeed learn to perform noise reduction even for speech and noise signals that were not part of the training data.

Speech Research at ATR Interpreting Telephony Research Laboratories

Speech Processing Department

Contents

I. Speech Recognition Research

I-1. Introduction

I-2. Continuous Speech Recognition by Hidden Markov Modeling

I-2.1. Improvement of HMM Phoneme Models

I-2.2. HMM Continuous Speech Recognition Using LR Parser

I-3. Phoneme Segmentation Using Spectrogram Reading Knowledge

I-4. Phoneme Recognition by Neural Networks

I-5. Speaker Adaptation by Fuzzy VQ and Spectrum Mapping

I-6. Summary

I-References

II. Speech Synthesis, Voice Conversion, and Noise Reduction Research

II-1. Introduction

II-2. Speech Database

II-3. Speech Synthesis

II-3.1. Outline of Speech Synthesis Scheme

II-3.2. The Synthesis Unit Entry Dictionary

II-3.3. Optimal Selection of the Unit Template Sequence

II-3.4. Control of Prosody

II-4. Voice Conversion

II-4.1. Voice Conversion through Vector Quantization

II-4.2. Conversion Experiments

II-4.3. Evaluation by Hearing Test

II-5. Noise Reduction

II-5.1. Network Architecture

II-5.2. Network Learning

II-5.3. Noise Reduction Experiments

II-6. Summary

II-References

Speech Research at ATR Interpreting Telephony Research Laboratories

Speech Processing Department

I. Speech Recognition Research

Abstract

Speech recognition research activities at the ATR Interpreting Telephony Research Laboratories are briefly described and summarized as follows:

(1) Hidden Markov phoneme models have been improved and successfully applied to Japanese phrase utterance recognition combined 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.

I-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, the integrated research of these technologies is also very important. We propose the interpreting telephony model shown in Figure I-1-1. In this model, language processing is split into a language source model stage and a language analysis stage. Our main research targets are fundamental research into speech and language processing and integration 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 the ATR Interpreting Telephony Research Laboratories. Efforts aimed at speaker-dependent phoneme recognition and speaker-independent phoneme segmentation have resulted in dramatically improved phoneme recognition performance. We are now pursuing three approaches: (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. Research progress is summarized in Sections 2, 3, and 4, respectively. For speaker-independent speech recognition, a speaker adaptation approach has been undertaken using the Vector Quantization and Spectrum Mapping concept, whose research

progress is summarized in Section 5. All research has been carried out using a Japanese, large-scale speech database with phoneme transcription developed at ATR.

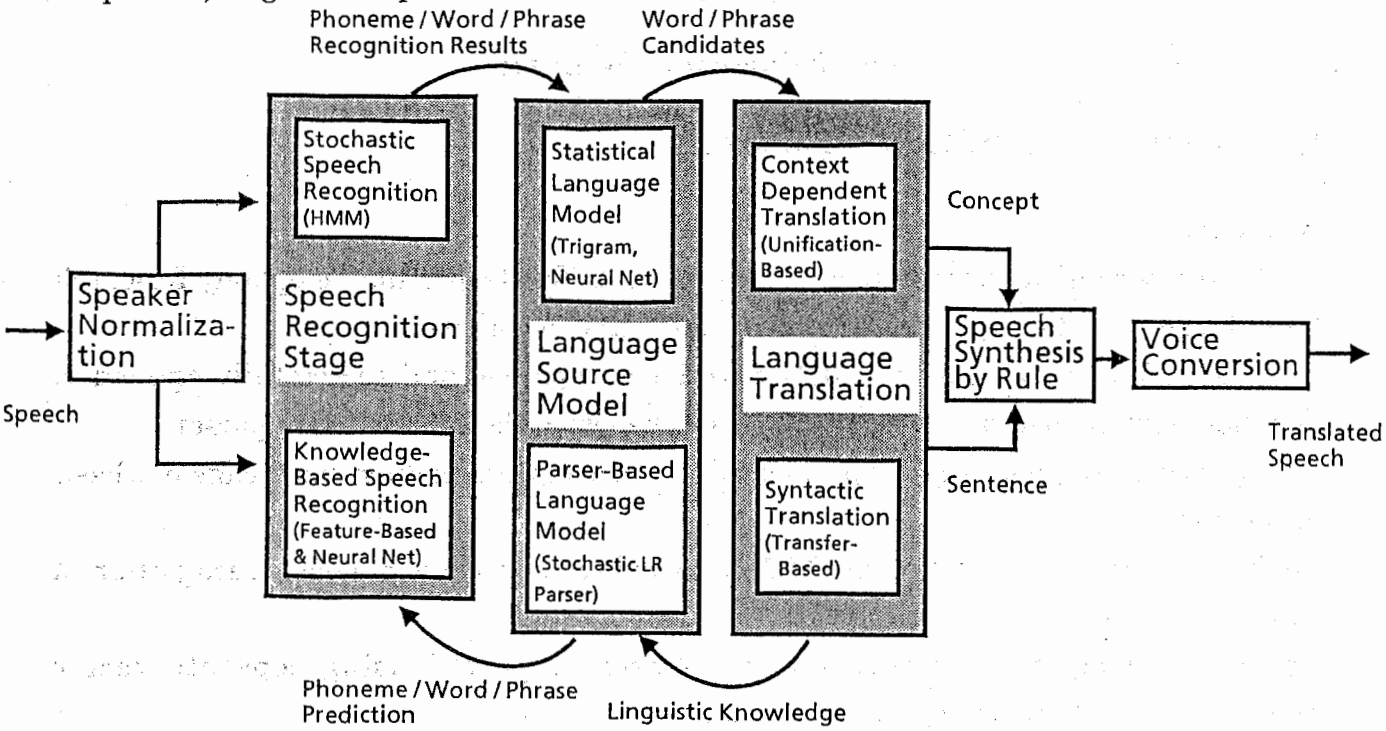


Figure I-1-1. Proposed Interpreting Telephony Experimental System.

I-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.

I-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-size (2,620 words) and small-size (216 words) training data sets.

Effective duration control is realized by combining two duration control techniques. One is phoneme duration control for each HMM phoneme model and the other is a state duration control for each HMM state. 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.

Separate vector quantization (multiple codebook) techniques for HMM phoneme recognition are useful for reducing VQ distortion. In our case, spectral features,

spectral dynamic features [6] and energy are quantized separately. In the training stage, the output vector probabilities of these three codebooks are estimated simultaneously and independently, and in the recognition stage all the output probabilities are calculated as a product of the output vector probabilities in these codebooks.

HMM training procedures are performed using the large-size 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 also used in recognition experiments to evaluate phoneme recognition performance. Improvements in the recognition rates using the large training data set are shown in Table I-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 resulted in a 7.5% improvement from 86.5% to 94.0% in the phoneme recognition rate for the average of three speakers using separate codebooks 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-size training data (216 words) set uttered by a male speaker. The phoneme recognition rate is improved by about 7% as shown in Table I-2-2.

Table I-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 I-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%

I-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 I-2-1. The LR parser was originally developed for use as a 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 the current state of the parser and next input symbol in the LR table. This parsing mechanism is valid only for symbolic data and cannot have been applied to continuous data such as speech.

In our approach, the LR Table I-is used to predict the next phoneme in the speech input. For 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 the end of an utterance.

This integration algorithm is applied to Japanese phrase recognition, whose task is secretarial service for an international conference. Utterances are uttered phrase by phrase. The syntax of phrases includes a general Japanese syntax structure of phrases, whose perplexity per phoneme is about five. Supposing that the average phoneme length per word is three, the perplexity of words is more than one hundred.

The HMM phoneme models are trained using 5,240 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 279 phrase inputs, as shown in Table I-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].

I-3. Phoneme Segmentation Using Spectrogram Reading Knowledge [9]

The phoneme segmentation approach by an expert system utilizing the spectrogram reading strategy and knowledge used by human experts to read spectrograms is

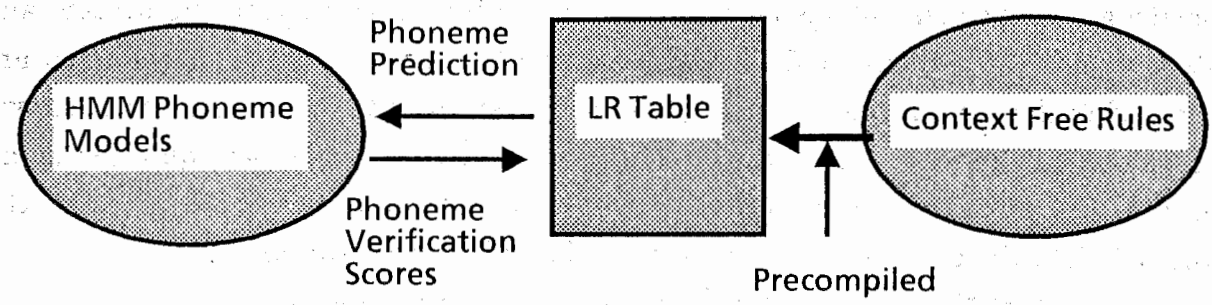


Figure I-2-1. HMM-LR Continuous Speech Recognition System.

Table I-2-3. Phrase Recognition Experiment Results

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

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 I-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 of 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 as well as their categories, with high accuracy. The method proposed here utilizes the 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 separately under their own phonetic context hypotheses. 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 I-3-1. The result shows that the system achieves performance equal to human experts'. In paticular, boundary alignment error is small, that is, most of the boundaries obtained are within 10 msec of the hand labeled boundaries.

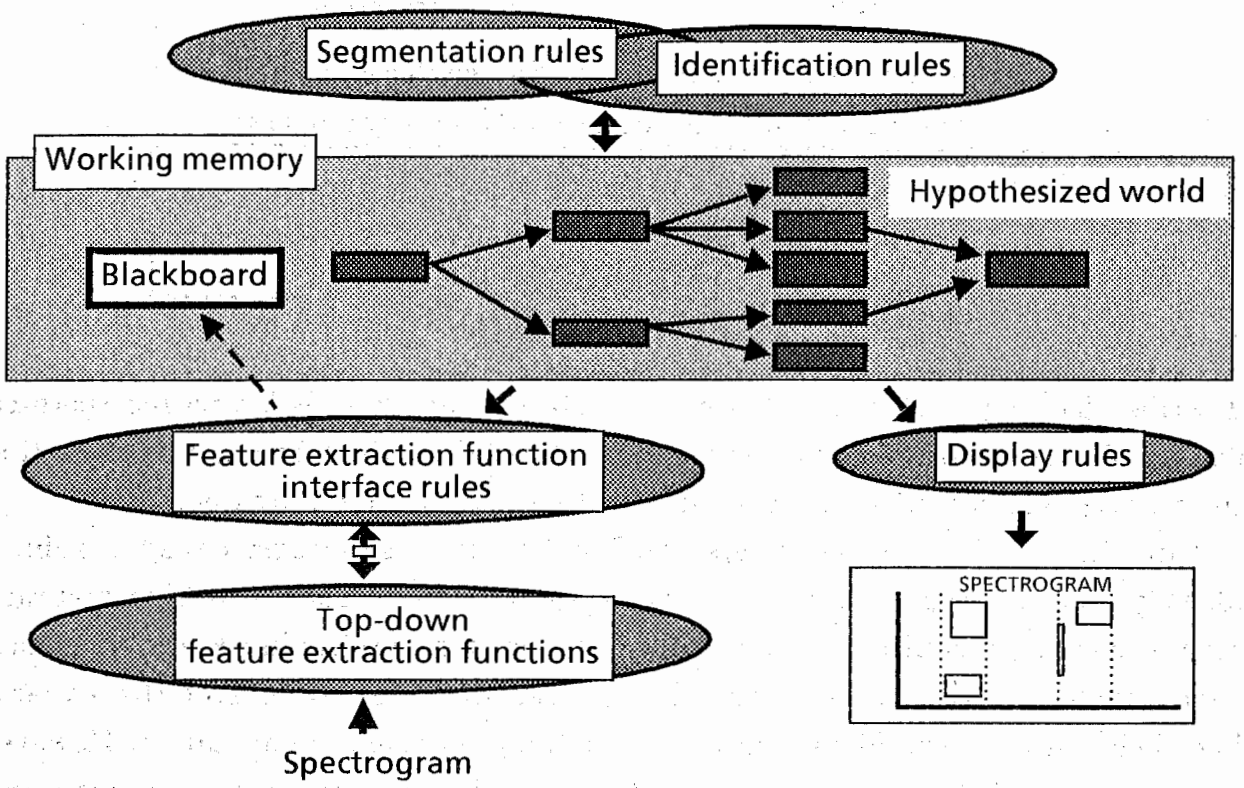


Figure I-3-1. Phoneme Recognition Expert System Architecture.

Table I-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) 5,240 words	/s/	1,086	36 (3.3%)	38 (3.5%)
	/sh/	783	21 (2.7%)	25 (3.2%)
	total	**1,869**	**57 (3.0%)**	**63 (3.4%)**

I-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]. One 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 that, based on a set of experiments aimed at phoneme recognition, it is indeed possible to construct large neural networks by exploiting the hidden structure of smaller trained subcomponent networks. A set of successful techniques that bring the design of practical large scale connectionist recognition systems within the reach of today's technology is developed.

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 a similar architecture is used. 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 had no notion of phonemes from other classes. Table I-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 considered 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 found 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%.

I-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 for post-processing. Evaluation experiments on HMM and voice conversion [16] have already clarified the performance and general applicability.

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

task	Phoneme recognition rate
18 consonants	96.0 %
HMM	93.8 %

The spectrum mapping method is based on the following three ideas. The first is accurate representation of input vectors by separate VQ and fuzzy VQ. The second is accurate establishment of spectral correspondence based on the fuzzy relationship 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 the target speaker's space. These algorithms are evaluated from the viewpoint of spectral distortion. The evaluation results are summarized in Figure I-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 the fuzzy relationship 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 I-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.

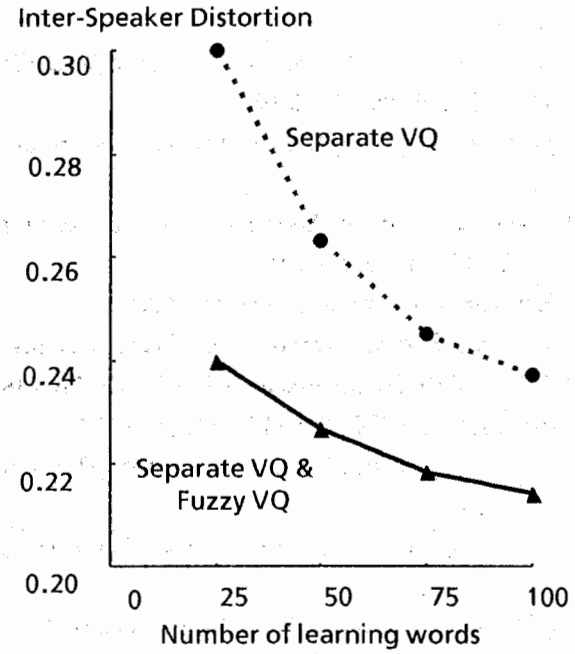


Figure I-5-1. Speaker Adaptation Algorithm Evaluation by Spectral Distortion.

Table I-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

I-6. Summary

Speech recognition research activities at ATR were summarized. In addition to the above research activities, the following research activities have also been 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 needed to develop automatic telephone interpretation technologies.

I-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)

II. Speech Synthesis, Voice Conversion, and Noise Reduction Research

Abstract

The outline of studies on speech synthesis, voice conversion and noise reduction which are now being conducted at ATR are described. The speech synthesis system proposed is a synthesis-by-rule based on an optimal selection of non-uniform synthesis units which aims at producing natural, high quality speech sounds. Voice conversion is a method to change voice individuality. The conversion method proposed here is to make use of conventional vector quantization technique. The essential part of this technique is to make mapping codebooks between two different speakers for such acoustic parameters as spectrum, pitch frequency, and power level. The conversion experiments reveal that this method is effective and promissing for the conversion of voice individuality. Noise reduction is another technique which the interpreting telephony system should incorporate. This is done by a four-layered neural network with the back-propagation learning algorithm. The result reveals that the network can indeed learn to perform noise reduction even for speech and noise signals that were not part of the training data.

II-1. Introduction

Since the foundation of ATR in 1986, we have been conducting basic research in speech recognition and speech synthesis which are parts of an automatic interpreting telephony system we are trying to develop. The final goal of the system is to enable people speaking different languages to communicate smoothly by telephone. To achieve this goal, three basic research should be taken into consideration: (1) speech recognition, (2) machine translation, and (3) speech synthesis. Since telephone conversation must be dealt with by this system, speech recognition must recognize continuous speech speaker-independently. After the speech recognition is completed, machine translation will follow. Difficulties of translation lie in the fact that we have to translate spoken dialogue not written language. The output of machine translation which is a sequence of symbols of another language is the input to the text-to-speech system to produce speech sounds at the other telephone. This paper outlines the speech synthesis, voice conversion, and noise reduction research that is now being conducted at ATR.

Speech synthesis is done by a synthesis-by-rule using non-uniform synthesis units which aim at producing high-quality, natural sounding speech. Based on a large-scale speech database, an entry dictionary for non-uniform speech units is compiled. According to the input phoneme sequences, speech units are selected optimally from

the entry dictionary based on some measure of appropriateness; Any speech segments that correspond to phoneme substrings of the given set can be used as unit templates from the dictionary. Statistical properties of phoneme sequences are also analyzed using a word dictionary and a text database to estimate the amount of speech data that covers all likely phoneme sequences necessary to produce Japanese sentences.

Voice conversion is a technique that can give the output speech from a computer speech individuality. With this method, we aim to produce speech sounds that mimic the original speaker's voice. The basic idea of this technique is to make use of speaker adaptation by vector quantization to make a speaker-dependent speech recognition system to be a speaker-independent. Codebooks for both spectrum and pitch frequency parameters are made separately for different speakers using a set of learning words. Then, based on the histogram, mapping codebooks which represent the correspondence between the codebooks of different speakers are made. Use of a listening test to evaluate this technique is also discussed.

Noise reduction study has been conducted here using a method different from the conventional one, namely, using a neural network. A four-layered feed-forward network is trained with the back propagation algorithm using a set of training samples to realize a mapping from noisy signals to noise-free signals. Giving the noisy speech waveforms directly to the input layer and the noise-free signals to the output layer for training, the network has been shown to learn to perform noise reduction. Analysis has been made to examine how the noisy signals are processed at each layer. The performance of noise reduction has also been discussed.

II-2. Speech Database

For the multiple purposes of speech research, a large-scale speech database is now under construction in ATR.[1]-[9] It consists of (1) a word database, (2) a continuous speech database. The word database contains 5,240 common Japanese words selected from a dictionary and the continuous speech database is a set of 503 phonetically balanced short sentences. Fine acoustic-phonetic transcriptions are made in five different layers. Professional announcers' speech sounds were collected first and databases for fifteen speakers, eight males and seven females, have been completed so far. A database management system has also been developed to handle the database easily and efficiently. Efforts are now being made to extend the database to include non-professional speakers.

II-3. Speech Synthesis

In this paper, we propose a synthesis scheme that provides an efficient selection of speech segments in a given speech data set according to an input phoneme string(2).

Any speech segments corresponding to phoneme substrings of the given set can be used as unit templates with their contextual information. The statistical properties of phoneme sequences are analyzed using a Japanese word dictionary and a text database to estimate the desired amount of speech data that might provide sufficient coverage of all likely phoneme sequences in Japanese.

II-3.1 Outline of Speech Synthesis Scheme

Figure II-3.1 shows a flow diagram of the speech synthesis scheme focusing on its unit template selection part. Speech synthesis is carried out as follows:

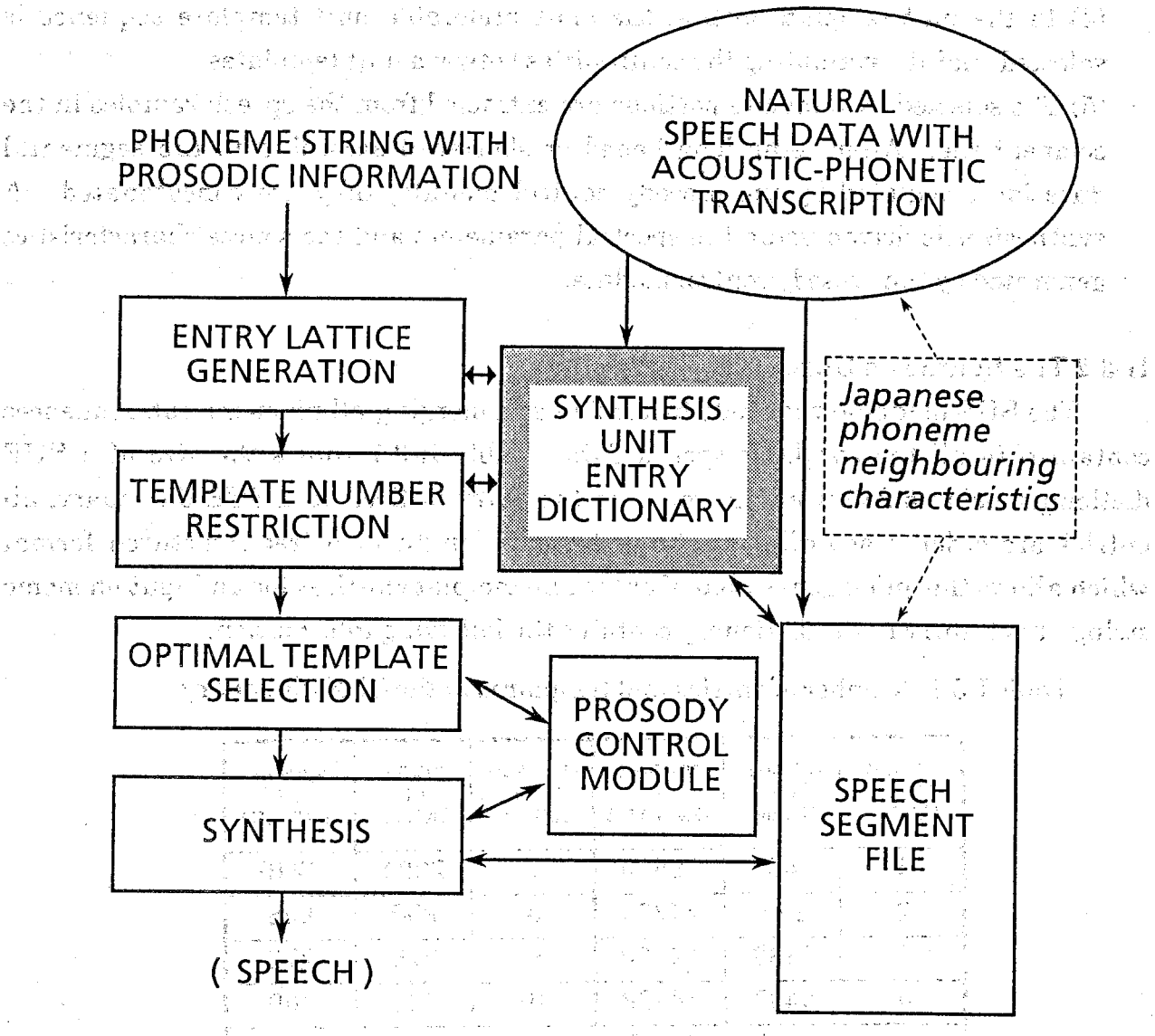


Figure II-3.1 Outline of the synthesis scheme using non-uniform units

(1) For an arbitrary input phoneme string, all phoneme substrings in each breath group are listed.
(2) After the search of the synthesis unit entry (SUE) dictionary, phoneme substrings found in the dictionary are collected in the form of a unit entry lattice.
(3) As each synthesis unit entry in the lattice can be extracted from multiple speech samples, only a few templates are selected in each unit entry for an efficient search for the most preferable unit template sequence. Comparison of the contextual similarities between speech samples having this entry and the corresponding portion in the input string reduces the number of all possible templates to a small set of suitable candidates.
(4) In the unit template lattice, the most preferable unit template sequence is selected mainly evaluating the continuities between unit templates.
(5) The selected by template portions are extracted from the speech samples in the segment file. After being lengthened or shortened according to the segmental duration calculated by the prosody control module, they are concatenated. A synthesizer is driven using the spectral parameters and the source characteristics generated by the prosody control module.

II-3.2 The Synthesis Unit Entry Dictionary

The SUE dictionary is made by sorting and merging all phoneme sub-sequences contained in all the available speech data. Table II-3.1 shows the size of a SUE dictionary using 5,240 words to generate the entry lattice. In a SUE dictionary, all entries are ordered according to their phoneme length in a tree-structured format which allows the quick generation of entry and template lattices for an input phoneme string. Each entry in the dictionary contains the following information.

Table I-3.1 Number of entries and templares in the SUE dictionary

PHONEME STRING LENGTH	NUMBER OF ENTRIES	NUMBER OF TEMPLATES	PHONEME STRING LENGTH	NUMBER OF ENTRIES	NUMBER OF TEMPLATES
1	43	29974	7	2084	2210
2	327	24734	8	890	908
3	2138	19498	9	275	278
4	5497	14328	10	100	100
5	6513	9396	11	19	19
6	4539	5226	12	2	2

(1) The total number of branch entries.
(2) The address of branch entries.
(3) The total number of speech templates that contain the entry.
(4) Addresses of speech templates that contain the entry.

Using the SUE dictionary, an entry lattice is automatically made by searching for the longest right-matching entry for each phoneme position in the input phoneme string.

II-3.3 Optimal Selection of the Unit Template Sequence

In the concatenation type synthesis system, discontinuities between units introduce speech quality degradation. To decrease such degradation, a unit template sequence is selected using the following criteria:

(1) Conservation of consonant-vowel transitions. The concatenation at C-V boundaries, especially if it is not continuant, is highly penalized.

(2) Conservation of vocalic sound succession. Concatenation at inter-vocalic boundaries is also penalized.

(3) Long unit preference. Since any concatenation degrades the resulting speech quality, units should be as long as possible.

(4) Template overlap. The greater the overlap between selected units that make up the synthesized utterance, the smaller the resulting discontinuities. To minimize the discontinuities, we try to maximize the overlap by evaluating the fit to the left and right of a candidate unit.

According to these criteria, there is a tendency for preferential selection of long unit templates followed by voiceless consonants if their combination exists in the input phoneme string. For example, the string /**hanagasakidashita**/ could be realized by the following template sequence: /**hana**banashii/+/na**ga**sa/+/mura**sa**ki/+ /hi**kida**shi/+/a**shita**/.

II-3.4 Control of Prosody

We are trying to find rules for prosodic control to improve "naturalness". Prosodic characteristics have been investigated for words uttered in seven different ways: fast, normal, slow, loud, weak, high, and low.[10] Their prosodic parameters, pitch frequency (F_0), power, and segmental duration, were compared. The results reveal that there is a strong correlation between F_0 and power, and a difference in F_0 patterns between speaking styles. F_0 pattern shows systematical changes as the speaking style changes. Estimation of the F_0 control parameter has been made to realize these tendencies in speech synthesis by rule system.

II-4. Voice Conversion

Speech individuality generally consists of two major factors: acoustic features and prosodic features. As the first step in this research, we are trying to control acoustic features. In this paper, we propose a new voice conversion technique without separating these acoustic parameters.[11]-[13] The basic idea of this technique is to make use of codebooks for several acoustic parameters. These codebooks carry all information about the speech individuality in terms of the varying acoustic features. A conversion of acoustic features from one speaker to another is, therefore, reduced to the problem of mapping the codebooks of the two speakers.

II-4.1 Voice Conversion through Vector Quantization

1) Learning Step

The mapping codebooks are codebooks that describe a mapping function between the vector spaces of two speakers. The block diagram in Figure II-4.1 illustrates how a mapping codebook for spectrum parameters is generated.

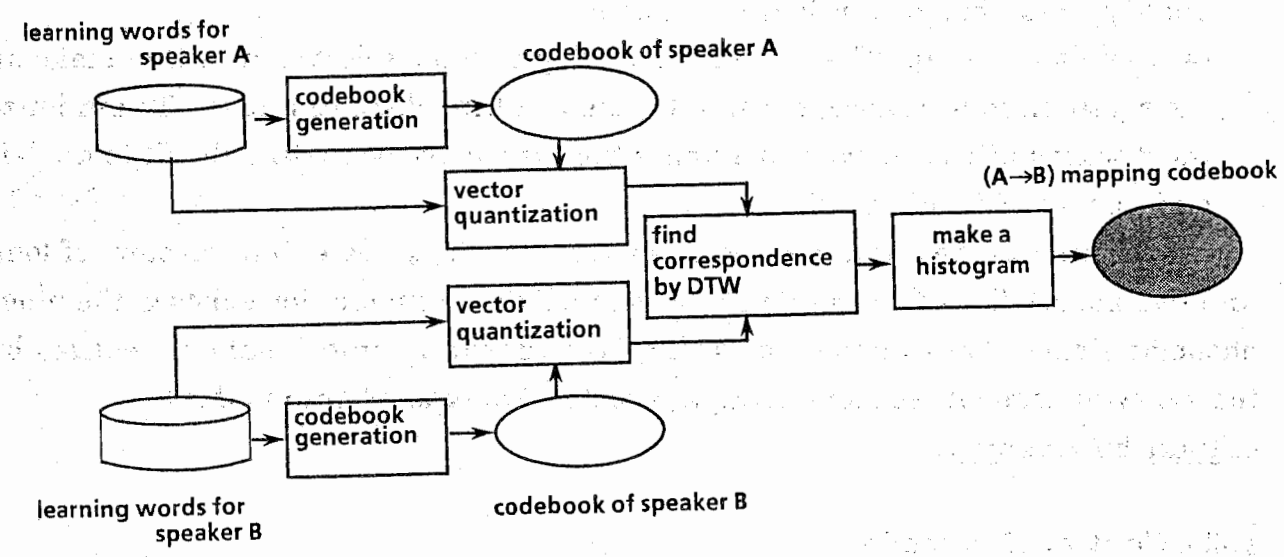


Figure II-4.1 Method of generating a mapping codebook

(1) Two speakers, A and B, pronounce learning words. Then, all words are vector-quantized frame by frame.

(2) The correspondence between vectors of the same words from the two speakers is determined using dynamic time warping (DTW).

(3) The vector correspondences between two speakers are calculated as histograms.

(4) Using each histogram as a weighting function, the mapping codebook is defined as a linear combination of speaker B's vectors.

(5) Steps 2, 3 and 4 are repeated to refine the mapping codebook.

Mapping codebooks for pitch frequencies and power values are also generated because these parameters contribute a great deal to speech individuality. These mapping codebooks are generated at the same time using almost the same procedure mentioned above. The differences are: 1) pitch frequencies and power values are scalar-quantized, 2) the mapping codebook for pitch frequency is defined based on the maximum occurrence in the histogram.

2) Conversion-Synthesis Step

As shown in Figure II-4.2, after the LPC anaysis of speaker A's speech, the spectrum/pitch parameters are vector/scalar quantized using his/her codebook. Then, synthesis is carried out by decoding using mapping codebooks between speakers A and B. The output speech will have the voice individuality of speaker B.

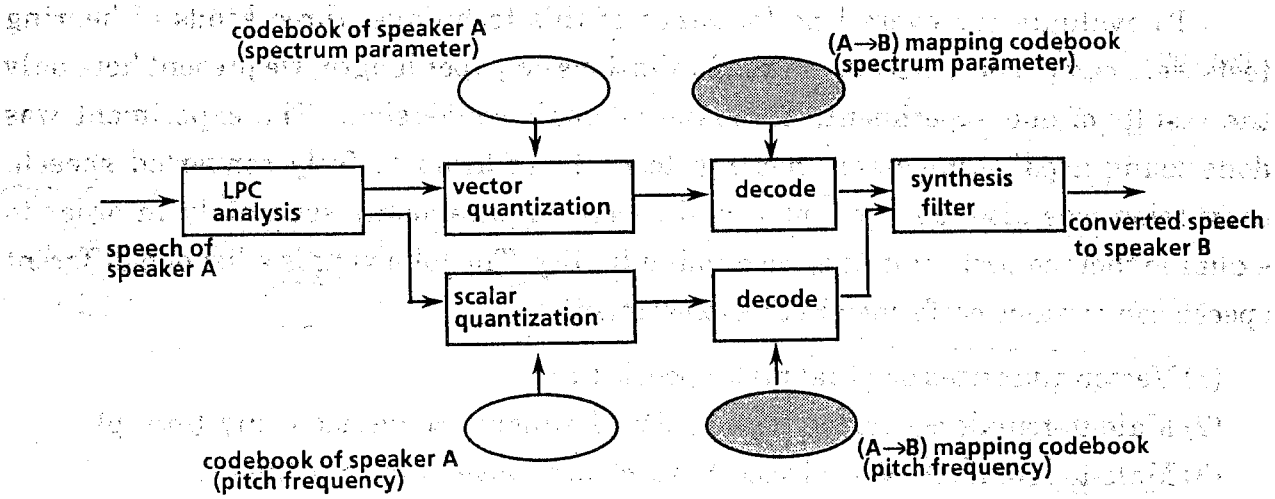


Figure II-4.2 Block diagram of voice conversion from speaker A to speaker B

II-4.2 Conversion Experiments

To evaluate the performance of the conversion technique, distortion measurements were carried out on the spectrum parameters and pitch frequencies. Table II-4.1 shows the results of the open test. After vector-quantization, two kinds of spectrum distortions between two speech samples were calculated; between the input and target speaker's, and between the converted speech and the target speaker's speech. In the female-to-female conversion, the distortion decreased by 27% compared

Table II-4.1 Spectrum distortion before and after conversion

speaker combination	before conversion	after conversion
female 1 → female 2	0.2759	0.2109
female 1 → female 3	0.2070	0.1489
male 1 → male 2	0.3364	0.1717
male 1 → male 3	0.2851	0.1550
male 1 → female 1	0.6084	0.2193

to the non-conversion, in the male-to-male conversion by 49%, and in the male-to-female conversion by 66%.

Pitch frequency conversion was also carried out through the same process. According to the result, 60 words are considered sufficient for making a mapping codebook for pitch frequency regardless of speaker combinations, and the average pitch frequency are less than 15Hz.

II-4.3 Evaluation by Hearing Test

To evaluate the overall performance of this technique, three kinds of hearing tests were carried out. Because of the limitations on paper length, we present here only the results of one experiment, the male-to-female conversion. The experiment was done using a pair-comparison hearing test. In addition to fully converted speech, conversion was also done for pitch and spectrum parameters separately in order to examine their contribution to speech individuality. The following is a list of 5 different speech conversions performed in this experiment.

(1) Vector-quantized original male speech (m)
(2) Male-to-female converted speech; pitch frequency conversion only (mp-fp)
(3) Male-to-female converted speech; spectrum conversion only (ms-fs)
(4) Male-to-female converted speech on all parameters (m-f)
(5) Vector-quantized original female speech, the target for the conversions (f)

Stimulus pairs were presented to listeners through a loud-speaker in a sound-proof room. Twelve participants in the experiment were asked to rate the similarity for each pair on five categories: "similar","slightly similar","difficult to decide","slightly dissimilar","dissimilar." Hayashi's forth method of quantification was applied to the experiment data of the hearing test. This method places stimuli on a two-dimensional space according to the similarities between every two stimuli. The projection onto a two-dimensional space is shown in Figure II-4.3 which represents the relative similarity-distance between stimuli. In this figure, converted speech "m→f" is most closely placed to the speech "f". This indicates that the male speech was properly

converted to the target female speech by this technique. Judging from the positions of "mp→fp" and "ms→fs", it is observed that the first and second axes roughly correspond to pitch frequency and spectrum differences, respectively. The result indicates that neither pitch frequency nor spectrum carries enough information about speech individuality, and that both are necessary.

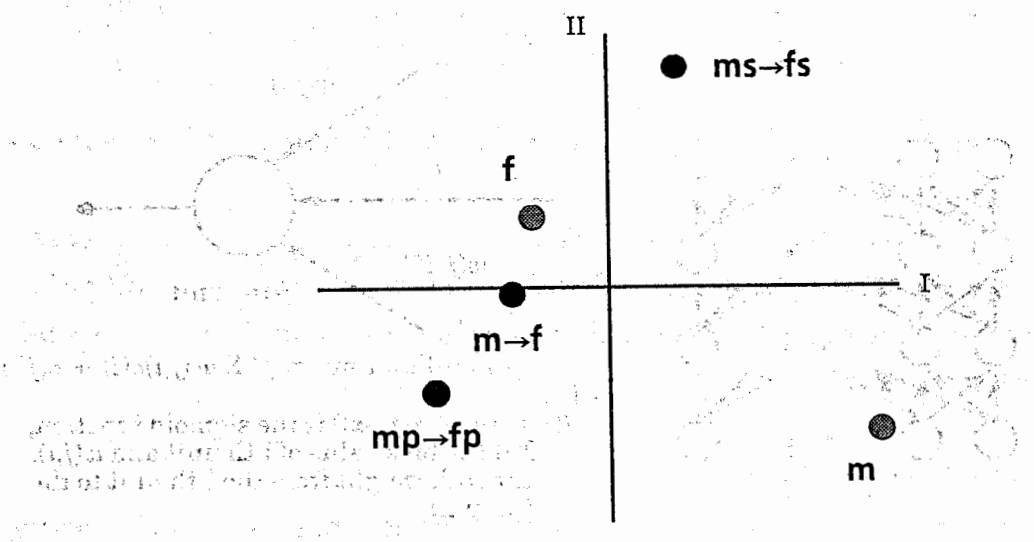


Figure II-4.3 Distribution of psychological distances in the male-to-female voice conversion

II-5. Noise Reduction

We propose a new noise reduction method using connectionist models.[14]-[15] Noise reduction can be viewed as a mapping from a set of noisy signals to a set of noise-free signals. Connectionist models are attractive for such mapping for the following reasons.

(1) An arbitrary decision surface can be formed in a multi-layered connectionist network. Any complex mapping from the set of noisy speech signals to the set of noise-free speech signals can, in principle, be realized .

(2) Simple learning algorithms exist to construct a suitable mapping function.

(3) Connectionist networks have attractive generalizing properties.

In the following, we first describe a connectionist model for mapping noisy to noise-free speech signals and then show the effectiveness of this approach by computer experiments.

II-5.1 Network Architecture

A four-layer feed-forward network was chosen for the architecture because it can, in principle, realize any mapping function. Each layer has 60 units and is fully

interconnected with its next higher layer as shown in Figure II-5.1. The network's state, by way of each unit's output, is updated synchronously on each layer and signals flow upward from the input layer to the output layer. For the network to use as much information about speech and noise as possible, the input and output of the network is given by the waveform itself, the units on the output and input layers are all linear units, i.e., are not passed though a non-linear output function. A unit

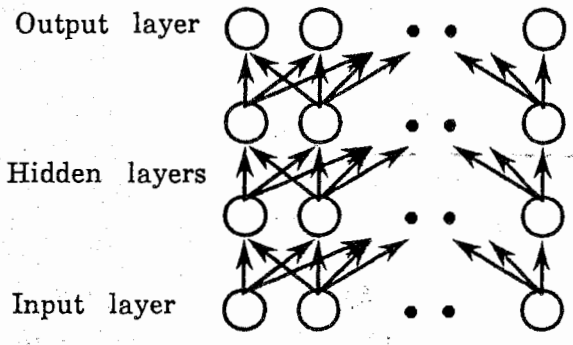
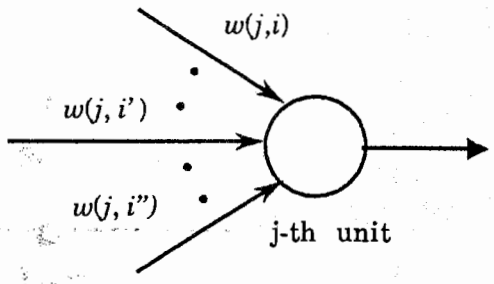


Figure II-5.1 Network for noise reduction



J-th unit's output = $f(\Sigma\, w(j,i)o(i) + \theta(j))$,
where
$f(x) = 1/(1+exp(-x))$ is the sigmoid function,
$\theta(j)$, the bias value of j-th unit and $w(j,i)$, the link weight from the i-th unit to the j-th unit

Figure II-5.2 Property of an element

element is one of many simple processors that make up the network. It first computes the weighted sum of all its inputs (including a bias input) and then deforms this sum by passing it through a nonlinear function, in our case the sigmoid function as shown in Figure II-5.2.

II-5.2 Network Learning

A subset of 216 phoneme balanced words from this database was used for our experiments. Computer room noise was chosen as non-stationary noise. Noisy speech data was generated artificially by adding the computer room noise to the speech data. The resulting S/N ratio was about -20db. During this phase, the back-propagation learning procedure repeatedly adjusts the network's internal link weights in an attempt to find an optimal mapping between noisy and noise-free signals. Using the waveforms of the 216 phoneme-balanced words as the target output and their noise added versions as the training input, the network scans each training utterance from beginning to end at a rate of 60 data points per input frame. When the network reaches the end of the training data, it returns to the beginning for additional learning passes. This procedure is repeated until the network's squared error rate converges to a sufficiently small value.

II-5.3 Noise Reduction Experiments

Figures II-5.3 and II-5.4 show the results of testing the network's ability to find a generalized noise reduction mapping based on the observations in the training data. In Figure II-5.3, we show as input to the network the Japanese word "kakuritsu" fully.

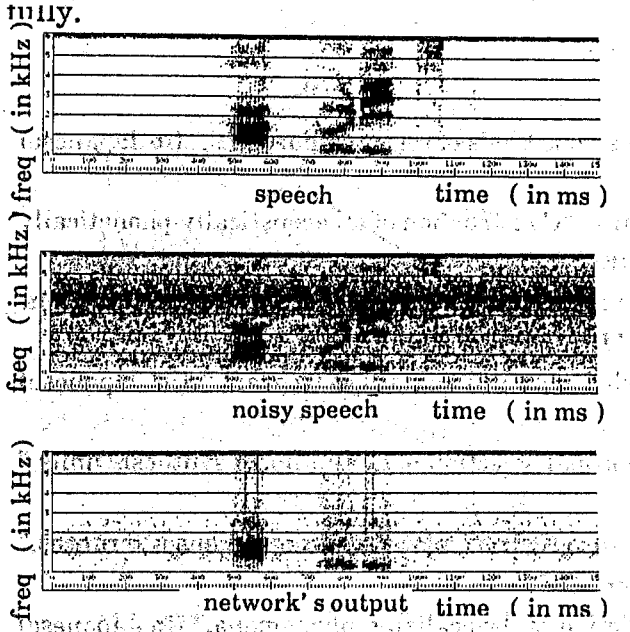


Figure II-5.3 Spectrograms of original speech (upper), noisy speech(middle), and noise reduction (lower) for a non-training word

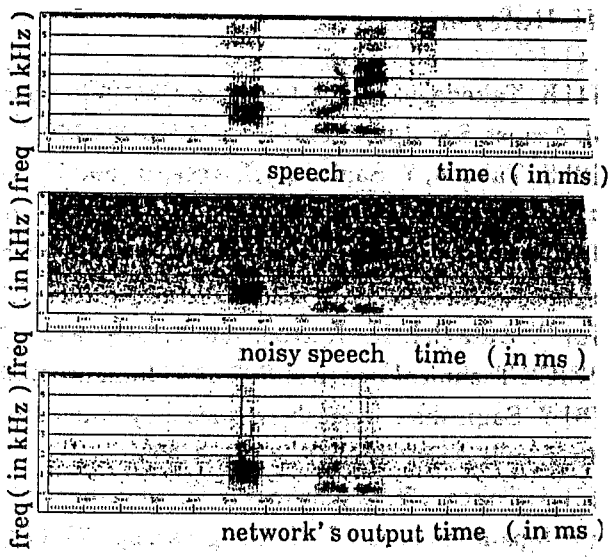


Figure II-5.4 Spectrograms of original speech (upper), noisy speech(middle), and noise reduction (lower) for a non-training word and noise

corrupted by noise. This utterance was not part of the training data. Again, we observe that the network's mapping suppresses the input noise successfully. In Figure II-5.4, we show the result of a more difficult problem. Here, the same word, "kakuritsu" has been corrupted by computer-generated white noise. Despite the fact, that the network was trained on a different kind of noise (non-stationary computer room noise), it produces a substantially cleaner output signal, without adversely affecting the speech signal.

II-6. Summary

Basic studies on speech synthesis, voice conversion, noise reduction and speech database in ATR have been reviewed. The results of these studies will be parts of the automatic interpreting telephony system which we are trying to establish within a limited domain. The speech synthesis system is a synthesis-by-rule system based on

non-uniform synthesis units which will be optimally selected from the entry dictionary compiled from a large-scale Japanese speech database. Voice conversion is another technique to give the synthetic speech voice individuality. Using mapping codebooks through vector quantization, it is possible to convert the speech of one speaker to that of another quite well. Noise reduction is conducted by a four-layered neural network giving speech waveforms directly to the network for learning. The network can learn to perform noise reduction even for unknown speech signals.

II-References

[1] K. Takeda, et al : "A Japanese speech database for various kinds of research purposes," (in Japanese) J. Acoust. Soc. Jpn., Vol. 44, pp.747-754 (1988)

[2] K. Takeda, Y. Sagisaka, S. Katagiri and H. Kuwabara : "Construction of an acoustically-phonetically transcribed Japanese speech database," (in Japanese) IEICE Technical Report, SP87-19 (June 1987)

[3] K. Takeda, H. Kuwabara and S. Morikawa : "Construction of a speech database management system," (in Japanese) IEICE Technical Report, SP87-116 (January 1988)

[4] S. Katagiri, K. Takeda and Y. Sagisaka : "Speech labeling using a spectrogram," (in Japanese) IEICE Technical Report, SP87-115 (January 1988)

[5] Y. Sagisaka : "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," Proc. of ICASSP'88, pp.679-682 (1988)

[6] K. Takeda, Y. Sagisaka and H. Kuwabara : "Vowel duration in isolated and continuous utterances," (in Japanese) Proc. of ASJ Congress, pp.177-178, (March 1988)

[7] K. Takeda and H. Kuwabara : "Analysis and prediction of devocalizing phenomena," (in Japanese) Proc. of ASJ Congress, pp.105-106, (October 1987)

[8] K. Abe and Y. sagisaka : "A synthesis unit selection method adapting to an input phoneme," (in Japanese) Proc. of ASJ Congress, pp.161-162, (March 1988)

[9] K. Abe and Y. Sagisaka : "Spectrum analysis by syllable environments," (in Japanese) Proc. of ASJ Congress, pp.199-200, (October 1987)

[10] M. Miyatake and Y. Sagisaka : "On the prosodic characteristics in various utterances with different speaking styles," (in Japanese) IEICE Technical Report, SP87-62 (October 1987)

[11] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara : "Voice conversion through vector quantization," Proc. of ICASSP'88, pp.655-658 (1988)

[12] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara : "Voice conversion through vector quantization," (in Japanese) IEICE Technical Report, SP87-124 (February 1988)

[13] M. Abe, S. Tamura and H. Kuwabara : "A speech modification method by signal reconstruction using short-time Fourier Transform," (in Japanese) IEICE Technical Report, SP88-48 (September 1988)

[14] S. Tamura and A. Waibel : "Noise reduction using connectionist models," Proc. of ICASSP'88, pp.553-556 (1988)

[15] S. Tamura : "An analysis of a noise reduction neural network which takes waveforms as input and output," (in Japanese) Proc. of ASJ Congress, pp.237-238, (October 1988)