論文 / 著書情報
Article / Book Information

Title	Speech Recognition and Synthesis Research at NTT Laboratories
Author	Sadaoki Furui, Kiyohiro Shikano, Hirokazu Sato
Journal/Book name	Proc. Korea-Japan Joint Symposium on Acoustics, "The Vision of Acoustics and Speech Communication Toward the 21th Century", , , pp. 37-42
発行日 / Issue date	1991,

In Commemoration of the 10th Anniversary of ASK
KOREA-JAPAN JOINT SYMPOSIUM ON ACOUSTICS
"The Vision of Acoustics and Speech Communication toward the 21st Century"

SPEECH RECOGNITION AND SYNTHESIS RESEARCH AT NTT LABORATORIES

Sadaoki Furui, Kiyohiro Shikano and Hirokazu Sato
古井 貞熙 鹿野 清宏 佐藤 大和

NTT Human Interface Laboratories
3-9-11 Midoricho, Musashino-shi, Tokyo, 180 Japan

I. OVERVIEW

This paper outlines major recent research activities on speech recognition and synthesis at NTT (Nippon Telegraph and Telephone Corporation) Laboratories.

The most important target of current speech recognition research is the creation of speaker-independent, large-vocabulary, continuous speech recognition systems. Various kinds of problems must be solved before such systems can be built. Figure 1 shows the basic structure of the continuous speech recognition systems, as well as major problems under investigation.

The flow of the recognition process is as follows. First, endpoints of speech periods are detected, typically by using the short-term energy level of input speech. Spectral analysis is then used to convert the speech wave into a time sequence of feature parameters. Cepstral and

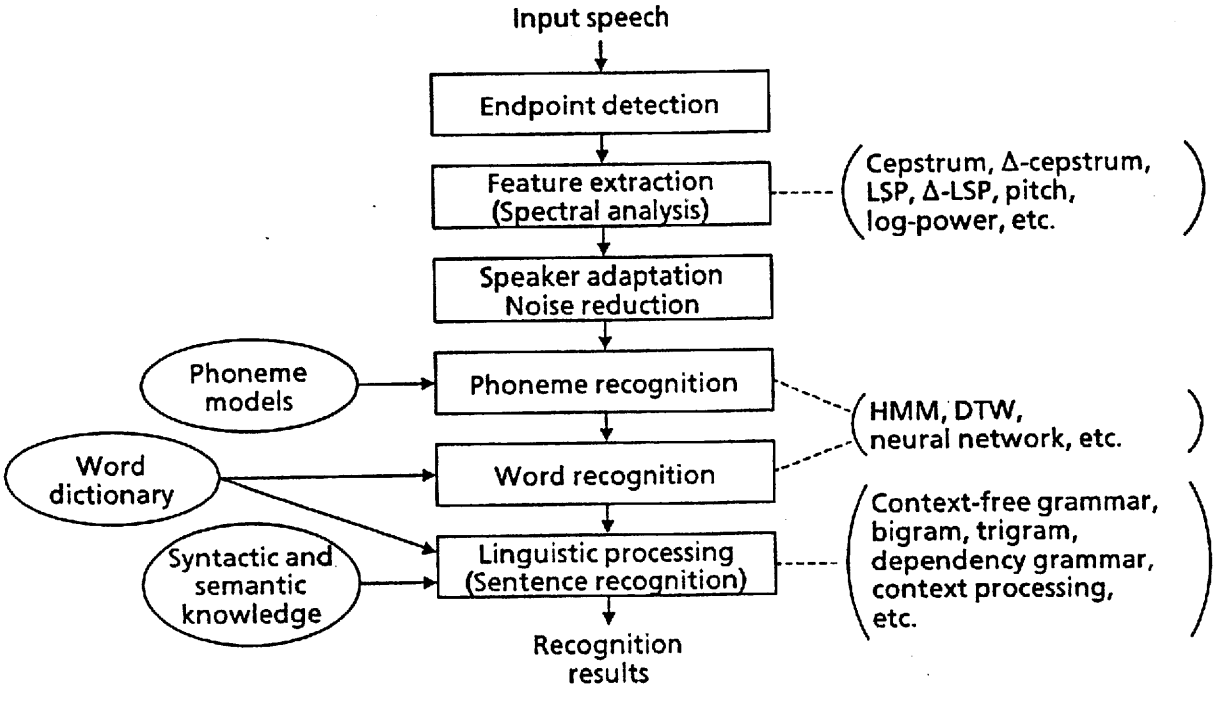

Fig. 1. Principal structure of continuous-speech recognition systems.

Δcepstral coefficients are widely used to represent both instantaneous and dynamic features of speech waveforms [Furui, 1986]. We have recently evaluated two new methods: one using hierarchical dynamic features [Furui, 1990a], the other using LSP (line spectrum pair) and ΔLSP parameters [Gurgen et al., 1990]. We have also proposed a new way of using prosodic features [Takahashi et al., 1990], which have previously been used with little success in speech recognition. Prosodic features are useful for recognizing words that are phonetically similar but have different accent patterns.

Although a high speaker-independent recognition accuracy can be obtained by training the recognition system with utterances from a large number of speakers, this method has a performance limitation [Furui, 1990b]. It is therefore necessary to introduce techniques for automatically adapting the system to a new speaker [Furui, 1989b, Imamura, 1991].

Word templates or word models are convenient for recognizing isolated words, but recognizing large-vocabulary continuous speech requires units that are smaller than words. It is also important for so-called co-articulation effects to be considered when creating these units. We have proposed an efficient method using a clustering technique to create context-dependent phoneme units [Sagayama, 1989]. We have compared several representations of HMM (hidden Markov model) phoneme models from the viewpoint of their robustness against variation of speaking styles, such as the differences between word-by-word utterances, phrase-by-phrase utterances, and continuous speech [Matsuoka et al., 1991]. Although neural-network phoneme modeling has various new possibilities, it has not yet surpassed the performances of models based on HMM. New training methods have been investigated for stabilizing the performance of neural networks [Gurgen et al., 1991].

The last and the most important stage of the continuous speech recognition is linguistic processing, where syntactic and semantic knowledge are used. Local language modeling, such as a transitional network grammar or bigram-/trigram-based grammar, has recently been used in many recognition systems. How to combine these methods with a global syntax like case grammar is one of the important research subjects [Matsunaga et al., 1990].

Applications of continuous speech recognition can be classified into two categories: voice-input word processors (dictation systems); and information services like the guidance, reservation, and order receiving services that use natural conversational speech. We have recently implemented a prototype dictation system [Tsuboi et al., 1990] and an experimental dictation system [Yamada et al., 1991]. We have also implemented word-spotting hardware for a dialog system [Imamura et al., 1990].

Speaker recognition using speaker-specific information in speech waves is also an important research subject; it is expected to be used in various applications such as security control. We recently proposed an efficient method of text-independent speaker recognition [Matsui et al., 1991].

Speech synthesis techniques that convert written text to natural speech are important for meeting the demands of voice response systems, text-reading systems/services, and various guidance services. Our text-to-speech conversion system, which is implemented on a single personal-computer board, creates natural voice from arbitrary Japanese text by using phonetic units based on the COC (Context-Oriented Clustering) technique [Mizuno et al., 1991].

The following sections will describe some of these techniques in more detail.

II. FEATURE EXTRACTION IN SPEECH RECOGNITION

2.1 Prosodic Features

Similarity of pitch pattern, which is a part of the prosodic information in speech waves, has been combined with spectral similarity for word recognition [Takahashi et al., 1990]. Experimental results show that this combination reduces word recognition error rates to 2/3 of what they are when only spectral similarity is used.

We have also proposed a method for automatically detecting phrase-boundaries according to the alignment between speech and concatenated accent-phrase models built using HMMs [Takahashi et al., 1991]. With duration-controlled pitch-pattern HMMs, the automatic phrase-boundary detector achieves almost the same boundary detection accuracy as manual detection by

a human. Phrase-boundary detection is expected to reduce the number of calculation in continuous speech recognition.

2.2 Phoneme Models and Speaking-Style Adaptation

Context-dependent phoneme HMMs have proved effective for phoneme recognition, word recognition, and word spotting [Sagayama, 1989, Matsuoka, 1990]. In real situations, only a few samples are available for creating some of the phoneme models, and the number of samples is sometimes too small to make a triphone model. We proposed a new clustering method for context-dependent phoneme HMMs [Matsuoka, 1990]. This method automatically selects triphone, biphone, or uniphone models according to the number of available training samples.

Recognition systems used in real situations must be able to cope with speech variations due to the speaker and due to the surrounding environment. The former variations include individuality, dialect, fluency, stress, speaking rate, level, and pitch. Variations and transitions between phonetic features can be efficiently and flexibly represented in a probabilistic manner by HMMs trained using utterances by many speakers.

The robustness against speaking-style variation was compared for six types of phoneme HMMs [Matsuoka et al., 1991]. These HMMs were trained with isolated-word utterances and tested with phrase-by-phrase utterances and continuous utterances. Robustness was highest with fuzzy vector quantization (VQ) HMMs, diagonal mixture Gaussian HMMs, and full-covariance Gaussian HMMs.

A new model-adaptation method using deleted interpolation has also been proposed [Matsuoka et al., 1991]. This method produces several HMMs in advance for different speaking styles, and the interpolation coefficients of there models are optimized for the test speech. The adapted models recognizes phonemes as well as or better than models trained with many speaking styles together. The proposed model synthesis method requires much less calculation than full training with all the samples.

III. SPEAKER CHARACTERIZATION

3.1 Speaker Adaptation Using Stochastic Speaker Characteristics

We recently proposed a new speaker-adaptive recognition method based on a statistical speech recognition algorithm [Imamura, 1991]. The problem is formulated as finding the best word sequence W' maximizing the *a posteriori* probability $P(W|Y,S)$ for input acoustic string Y and speaker individuality S. Using Bayes' rule, this probability is computed as

$$P(W|Y,S)=P(W)P(Y,S|W)/P(Y,S), \tag{1}$$

where $P(Y,S)$ is the *a priori* joint probability of acoustic string Y and speaker individuality S, $P(W)$ is the *a priori* probability of word sequence W given by a language model, and $P(Y,S|W)$ is the category-conditional joint probability given by a speaker-constrained acoustic HMM.

A stochastic speaker classifier is used as the feature extractor for speaker individuality information. This speaker classifier includes several speaker classes represented by speaker Markov models that are estimated by clustering the training speech uttered by many speakers. For each input speech token Y, the classifier computes the category-conditional probability $P(Y|S)$ for each speaker class S. This probability is used to select the speaker class (feature sub-space) that is most suitable for the input speech. Speaker individuality is quantized by these category-specific probabilities $P(Y|S)$. In the subsequent word-decoding phase, speaker-constrained acoustic HMMs use the outputs of the speaker classifier and the acoustic pre-processor to compute $P(Y,S|W)$.

Experiments using a telephone-speech database of Japanese digits recognized words with an accuracy of 98.1%. The corresponding error rate is roughly 1/2 that of the conventional speaker-independent (pooled training) method.

3.2 Text-Independent Speaker Recognition

To increase robustness against inter-session variations of parameters, a VQ-based text-

independent speaker recognition method using vocal tract and pitch information has been investigated [Matsui et al., 1991]. This method uses a combination of cepstrum, Δcepstrum, pitch, and Δpitch information. A new normalization method, Talker Variability Normalization (TVN), takes the inter- and intra-speaker variability into consideration. A new distance measure between input speech and speaker-specific codebooks, the Distortion-Intersection Measure (DIM), includes a factor reflecting the size of the intersection between distributions of input speech frames and codebook elements. The combination of these methods achieved high-performance speaker recognition.

IV. CONTINUOUS SPEECH RECOGNITION

4.1 Language Processing

Language processing of our continuous speech recognition system is based on "two-level grammar," which consists of syntactic grammar for intra-phrase structures and a semantic dependency grammar for inter-phrase sentence structures [Matsunaga et al., 1990]. The phrase-dependency grammar is effective for pruning wrong phrase sequences. A joint likelihood - combining acoustic, syntactic, and semantic likelihoods derived from acoustic processing and language processing - is maximized to obtain the optimum solution. This procedure takes into account the redundancy of speech and the large freedom of phrase order that is characteristic of the Japanese language. Experimental results showed that the dependency parser decreased the average phrase recognition error rate to roughly half of that achieved without dependency analysis.

4.2 Dictation System Using Phoneme Source Modeling

We are investigating a phonetic typewriter that uses the underlying syntactic and statistical structure of Japanese phoneme and character sequences [Yamada et al., 1991]. A schematic diagram of this system, which consists of an HMM-based acoustic processing part and phoneme source modeling, is shown in Fig. 2. Because Kana (Japanese syllabary alphabets) roughly correspond to consonant-vowel (CV) syllables, a syllable trigram approach to language source modeling is effective for Japanese.

For our phonetic typewriter, a general Japanese syllable sequence structure is written using context-free rewriting rules. This structure is precompiled into an LR table with syllable trigram probabilities calculated from a large text database. The predictive LR parser predicts possible

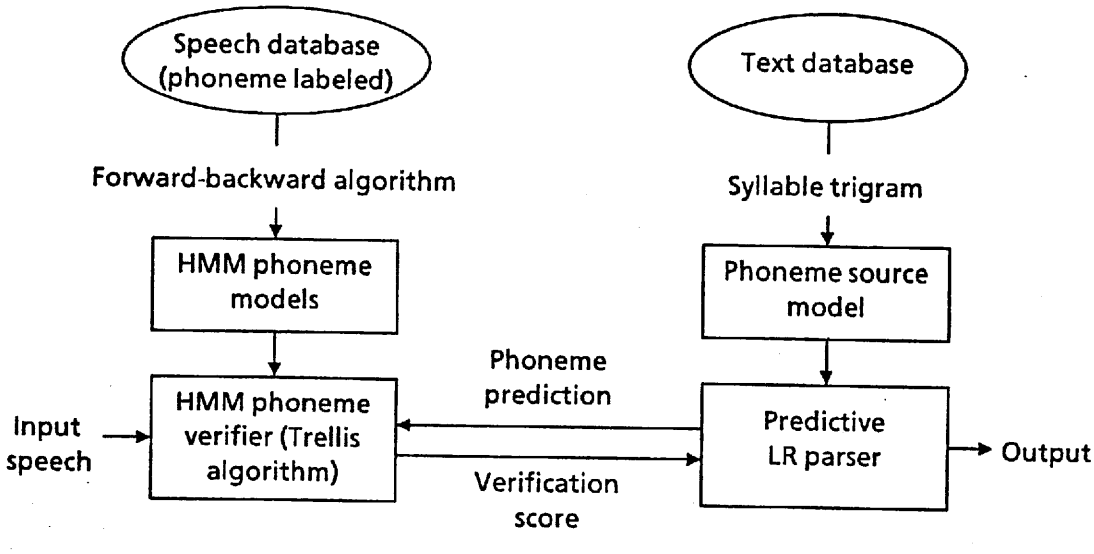

Fig. 2. Schematic diagram of a phonetic typewriter based on the HMM-LR method.

phoneme sequences from left to right according to the general Japanese phoneme sequence syntax. The parser then calculates phoneme-sequence probabilities based on syllable trigram and HMM probabilities; speech is recognized as the phoneme sequence with the highest probability.

This phonetic typewriter has been tested using 279 phrases uttered by one male speaker. For a test-set phoneme perplexity of 3.9, the syllable source model achieved a phoneme recognition rate of 94.9%, whereas the phoneme recognition rate was only 73.2% without the syllable trigram.

A trigram model based on character sequences (Kana and Chinese character) in usual Japanese sentences has also been studied. Compared with the syllable trigram, this character trigram model significantly reduces phoneme perplexity.

V. SPEECH SYNTHESIS

A Japanese text-to-speech synthesizer that produces speech that is more natural and intelligible than the speech generated by existing synthesizers was also developed. This synthesizer improves the quality of output speech by using the COC method and by expanding the frequency band of synthesized speech. The COC method uses a statistical technique for automatically generating phonetic units from a natural speech database.

To permit easy installation in personal computers, the text-to-speech synthesizer was implemented on a single PC board with a general purpose DSP. The quality of the synthesized speech was confirmed to be better than that produced using the conventional dyad method. The method and performance of the synthesizer are described in detail in another paper [Mizuno et al., 1991].

VI. CONCLUSION

Speech is one of the most natural and easiest communication methods for human beings. Speech will therefore play important roles in the various new communication services that will be provided in the future. To create these new services, it is important to investigate various problems from both the technological and human viewpoints.

From the technological point of view, future problems will be related to speech individuality, robust and proper statistical modeling, and new technologies like sophisticated neural networks [Furui, 1989a]. By solving these problems, we will be able to create new communication and information services based on speech processing.

REFERENCES

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

Furui, S. (1989a): "Digital Speech Processing, Synthesis, and Recognition," Marcel Dekker, New York

Furui, S. (1989b): "Unsupervised speaker adaptation method based on hierarchical spectral clustering," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Glasgow, Scotland, S6.9

Furui, S. (1990a): "On the use of hierarchical spectral dynamics in speech recognition," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Albuquerque, S15a.10

Furui, S. (1990b): "Speaker-dependent-feature extraction, recognition and processing techniques," Proceedings of the ESCA Tutorial and Research Workshop on Speaker Characterization in Speech Technology, pp. 10-27

Gurgen, F., Sagayama, S. and Furui, S. (1990): "Line spectrum pair frequency-based distance measures for speech recognition," Proc. Int. Conf. Spoken Language Processing, Kobe, 13.1

Gurgen, F., Aikawa, K. and Shikano, K. (1991): "The improvement of phoneme recognition performance of a neural network using fuzzy training," Proc. SYNAPSE'91, Osaka

Imamura, A. and Suzuki, Y. (1990): "Speaker-independent word spotting and a transputer-based implementation," Proc. Int. Conf. Spoken Language Processing, Kobe, 13.5

Imamura, A. (1991): "Speaker-adaptive HMM-based speech recognition with a stochastic speaker classifier," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto, 58.S13.3

Matsui, T. and Furui, S. (1991): "A text-independent speaker recognition method robust against utterance variations" Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto, 57.S6.3

Matsunaga, S., Sagayama, S., Homma, S. and Furui, S (1990): "A continuous speech recognition system based on a two-level grammar approach," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Albuquerque, S11.7

Matsuoka, T. (1990): "Word spotting using context-dependent phoneme-based HMMs," Proc. Int. Conf. Spoken Language Processing, Kobe, 13.7

Matsuoka, T. and Shikano, K. (1991): "Robust HMM phoneme modeling for different speaking styles," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto, 10.S5.4

Mizuno, H., Nakajima, S., Hirokawa, T. and Hakoda, K. (1991): "A new one-board text-to-speech synthesizer," Korea-Japan Joint Workshop on Advanced Technology of Speech Recognition and Synthesis

Sagayama, S. (1989): "Phoneme environment clustering for speech recognition," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Glasgow, Scotland, S8.3

Takahashi, S., Matsunaga, S. and Sagayama, S. (1990): "Isolated word recognition using pitch pattern information," Proc. Int. Conf. Spoken Language Processing, Kobe, 13.9

Takahashi, S. and Matsunaga, S. and Shikano, K. (1991): "Accent phrase boundary detection in continuous speech using hidden Markov models," Proc. Spring Meeting of Acoust. Soc. Jap., 2-5-13 (in Japanese)

Tsuboi, T. and Sugamura, N. (1990): "A prototype for a speech-to-text transcription system," Proc. Int. Conf. Spoken Language Processing, Kobe, 20.8

Yamada, T., Hanazawa, T., Kawabata, T., Matsunaga, S. and Shikano, K. (1991): "Phonetic typewriter based on phoneme source modeling," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto, 56.S3.4