論文 / 著書情報
Article / Book Information

Title	Speech Processing Technologies and Telecommunications Applications at NTT
Author	Noboru Sugamura, Tomohisa Hirokawa, Shigeki Sagayama, Sadaoki Furui
Journal/Book name	Interactive Voice Technology for Telecommunications Applications, , , pp. 37-42
発行日 / Issue date	1994, 9
権利情報 / Copyright	(c)1994 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA94) Kyoto Research Park, Kyoto, Japan
September 26 - 27, 1994

Speech Processing Technologies and Telecommunications Applications at NTT

Noboru Sugamura*, Tomohisa Hirokawa*, Shigeki Sagayama* and Sadaoki Furui**

NTT Human Interface Laboratories
*1-2356 Take, Yokosuka-shi, Kanagawa, 238-03 JAPAN
**3-9-11 Midoricho, Musashino-shi, Tokyo 180 JAPAN

Abstract

This paper describes major research and development in speech recognition and synthesis technologies at NTT from the telecommunications applications viewpoint. Technologies include speaker-dependent, speaker-independent word recognition based on DP matching, speaker-independent word spotting based on HMM, large vocabulary, speaker-independent continuous speech recognition based on HMM-LR and high-quality Japanese Text-to-Speech synthesis. A commercial ANSER system that uses speech recognition and synthesis technologies is also introduced.

1.Introduction

The "Multimedia Era" will come soon based on the advent of B-ISDN and FTTH. Under these circumstances, the various new services will utilize video, speech, text, data and other multimedia information. In these new services, speech will still play an essential role, because speech is the most natural and easiest communication media for human beings. Speech recognition and speech synthesis technologies have important roles in constructing better human/machine communication systems with the ability to communicate through speech over the network. In Japan, the number of subscriber telephone service contracts reached some 57 million at the end of March 1993. Therefore, it is better to develop new services around the telephone network for both users and venders; only ordinary telephone sets should be needed. Speaker-independent speech recognition has advanced dramatically over the last decade. Moreover the quality of synthesized speech has also been improved. These technologies have been realized by the use of high-performance CPUs and DSPs. As a leader in speech research and development, Nippon Telegraph and Telephone Corporation (NTT) has developed several speech technologies[1],[2],[3] and various types of equipment that offer a great deal of application potential. Speech recognition and speech synthesis equipments developed at NTT are introduced in the following sections from the viewpoint of telecommunications applications.

2. System Overview in the Real World

2.1 ANSER System

In Japan, NTT has combined speaker-independent speech (isolated word) recognition and speech synthesis technologies to form the telephone information system called ANSER(Automatic Answer Network System for Electrical Request)[4] . Since its introduction in 1981, the system has provided information services for the banking industry. Most Japanese banks (601 banks in 1990) now use ANSER, serving customers at the rate of several hundred thousand calls per day. Later, the system was also introduced into the securities industry.

ANSER's voice response and speech recognition capabilities let customers make inquiries and obtain information through dialogue with a computer. When ANSER was first developed in 1981, the system had only voice response capability and could accept input only from touch-tone telephones through DTMF signals. Speech recognition was added by the end of that year, permitting system access through ordinary dial telephones. Later, facsimile and modem access capabilities were added. Figure 1 shows a typical ANSER system configuration for a banking application. ANSER systems are in place in more than 75 cities across Japan, with all ANSER centers interconnected by a data communications network. Customers can access an ANSER center and obtain banking services for a small fee wherever they live.

Speaker-independent speech recognition is particularly difficult through telephone lines because, in addition to variations among speakers, telephone sets and lines cause varying amounts of distortion. The system's 16-word lexicon consists of the 10 digits and six control words in Japanese. A huge amount of telephone speech with a wide range of telephone-set and line variations and speaker characteristics was collected to form a speech database. The samples came from three regions of Japan and were generated by 1,564 male and female speakers ranging in age from 20 to 60 years. The basic idea for vocabulary-independent word recognition based on DP matching was introduced. Namely, each word is expressed as a sequence of phoneme templates.

2.2 PC-based ANSER System

Since its original introduction, the system has been improved in terms of processing capacity, performance of speech recognition, and quality of synthesized speech. ANSER is one of the biggest commercial applications of speech recognition technology in the

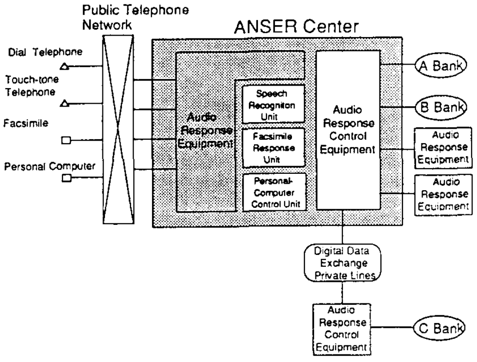


Fig. 1 ANSER System Configuration

0-7803-2074-3/94/$4.00 © 1994 IEEE

world. However, the market for voice information services could be expanded by reducing the size and cost of the speech recognition unit used in the ANSER system. Therefore, NTT has developed a single speech recognition processor LSI to realize a more compact and economical speech recognition board. The speech recognition LSI was created as a 20K gate array using 1.5-mm C-MOS design rules. Spectral analysis is executed on a general purpose DSP chip. The board has sufficient memory (one 128-KW RAM and four 128-KW Flash ROMs) for a maximum vocabulary size of 2,048 words for single-word template usage, however only 512 words among them can be recognized at the same time. The new speech recognition board is approximately 75% smaller than the original unit, making it possible to plug the recognizer into a conventional PC (Personal Computer).

Basically, this board can be used for both speaker-independent and speaker-dependent isolated word recognition. The average recognition accuracy of the ten numerals is 97% for speaker-independent recognition. A PC based ANSER system for personal usage can be easily configured by combining the recognition board with a PC and a text-to-speech board developed by NTT.

3. Speech Recognition Technologies

3.1 A Speaker-Independent Word Spotting Board and Trial for Voice Dialing

While using an isolated word recognition system, any extraneous speech or breathing noises prior to or following the target word can cause recognition errors. This problem can be solved by using a word spotting technique to recognize a set of key words embedded in conversational speech or in noisy speech. We have developed a speaker-independent word spotting board based on the HMM (hidden Markov model)[5]. Over the last ten years, HMM based speech recognition techniques have been successfully applied to various speech recognition systems, especially speaker-independent recognition and continuous speech recognition.

The block diagram of the HMM-based word spotting system is shown in Fig. 2. There are essentially four stages in the process: spectral (LPC) analysis, fuzzy vector quantization, Viterbi decoding, and postprocessing. In the spectral analysis stage, the input signal is analyzed by the LPC method, and spectral feature vectors are obtained. Fuzzy vector quantization represents each input vector as a weighted combination of the code vectors. In the Viterbi decoding stage, time series of fuzzy observation symbols are scored frame-synchronously by the Viterbi decoding algorithm with reference to the HMMs of the key words. After the decoding process, recognition candidates are selected in the postprocessing stage, which includes pruning by duration control and assignment of likelihood scores. After detecting the locally peaked likelihood score, spotting results are obtained.

The maximum length of a key word is 2.4 seconds, but utterance length is unlimited. Each key word is represented by an 11-state discrete hidden Markov model trained using speech data of several hundred male and female speakers. An experiment showed that 97% keyword recognition accuracy was obtained for the speaker-independent ten Japanese numeral spotting task. For this task, each HMM was trained using about 100 male speaker tokens collected through long distance telephone lines with SN ratios ranging from 20dB to 10dB[5].

The board uses two general purpose DSPs for spectral analysis and fuzzy vector quantization, and nine Transputers for Viterbi decoding and postprocessing. The nine Transputers are connected in a tree structure, with the root Transputer performing some postprocessing and system control operations. This architecture allows a set of 88 key words to be processed in real time. A maximum of 224 vocabulary words can be stored on about 10 Megabytes of on-board memory. This board has a VME-bus interface and can be easily plugged into a workstation as shown in Fig. 3.

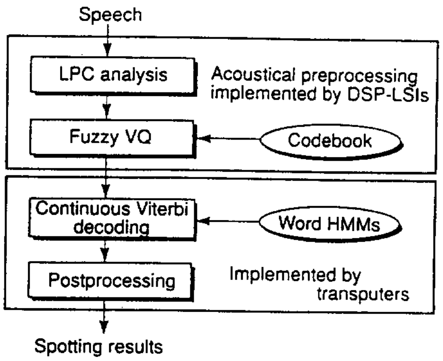
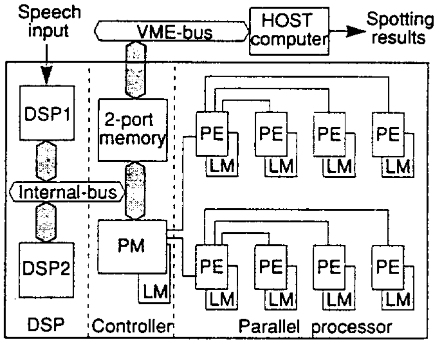

Speech
LPC analysis
Acoustical preprocessing implemented by DSP-LSIs
Fuzzy VQ
Codebook
Continuous Viterbi decoding
Word HMMs
Postprocessing
Implemented by transputers
Spotting results

Fig.2 Block Diagram of the Word Spotting System



PM: Processing Master, PE: Processing Element, LM: Local Memory

Fig.3 Hardware Architecture of the Word Spotting Board

Using this word spotting board, a voice dialing system called "Name Dial" was tested in Advanced Intelligent Networks (Advanced IN). Speech recognition experiments were performed using an IN experimental system that included the speech recognition function. A voice dialing service, which allowed one of a closed user group to call a member of the group by speaking his section and name, was tested and achieved a voice dial success rate of 77 %[6].

We have also tested the voice activated telephone intermediary system using this word spotting board[7].

3.2. Voice Dialing Telephone for Automobile

The need for a noise-robust and hands-free speech recognizer is rapidly increasing, especially one that can be used in automobiles. It is thought difficult, however, to improve speech recognition performance under noisy conditions, and speech uttered in a running automobile is always contaminated by nonstationary environment noises. Conventional noise subtraction techniques using a single microphone do not provide adequate speech recognition performance in an automobile, so we have designed a speech recognizer that is effective even in environments filled with automobile noise.

The proposed speech recognizer performs noise subtraction in the autocorrelation domain, not the spectral (frequency) domain[8]. Signals from primary and secondary microphones are used, and the autocorrelation function of each signal is calculated over a short time period. The primary microphone is placed so as to pick up the speaker's voice and maximize the signal-to-noise ratio, whereas the secondary microphone is positioned to pick up the environmental noise and minimize the SN ratio. The two autocorrelation functions can be used for adaptive noise subtraction. The subtraction ratio can be controlled according to the energy of the signal input through the primary microphone and the energy ratio of the two input signals during the noise measurement period. The subtraction result is then used to extract LPC cepstrum coefficients, and these parameters are used for spectral pattern matching. A dynamic programing algorithm was used for spectral pattern matching.

Tests were performed to determine the optimal secondary microphone location after placing the primary microphone on the driver-side visor. The best recognition accuracy was obtained by autocorrelation subtraction when the secondary microphone was located at the center of the ceiling.

We have implemented the speech recognition algorithm in a prototype speech recognizer. Components of the speech recognizer are shown in Fig. 4. A prototype of the voice dialing telephone, which is equipped with a speech recognizer, is shown in Fig. 5. While the car is parked, the vocabulary words are registered as required to permit speaker dependent usage. After registration, the recognizer can accept the vocabulary words, even if noise is present. In high noise environment, better speech detection accuracy can be obtained by using these two inputs rather than only one input(9). The recording block plays back guidance messages during vocabulary registration and speech recognition. During vocabulary registration, the speaker must utter words ("yes","no" and "cancel") to control the recognizer. Signals indicating the start and the end points are sent from the speech recognition block to the speech recording block as recording trigger signals. The recorded speech is played back through a loudspeaker for verbal confirmation of the recognition candidates.

When a recognized candidate word is uttered by the loudspeaker, the speaker may confirm by saying "yes", or he can say "cancel". To evoke the recognizer, there is a key in the operation block. This block also includes numerical keys and a few control keys.

The speech recognizer mainly consists of two μ-law CODECs for analog-to-digital conversion, two DSPs for speech recognition, a 256 kbyte RAM for speech patterns and some working area, a single ADM CODEC LSI and one Mbyte of memory for recording. An 8-bit CPU is used to control the entire recognizer. The maximum number of vocabulary words to be registered is 48 when LPC cepstrum coefficients and their regression coefficients are used for each 1.6-second word pattern.

Speech recognition experiments in a moving passenger car were carried out using the developed speech recognizer. The speaker sat in the passenger's seat. The primary and the secondary microphone were mounted on the passenger's visor and at the center of the ceiling, respectively. At first, the names of 48 Japanese cities were uttered for registration when the car was parked (engine idling). The windows were completely closed and the ventilation fan was set at its lowest level. The experiments were continued under various (several) driving conditions. The results obtained by one male speaker are shown in Table 1. The recognizer worked well under all driving conditions when windows were closed. The main reason for the recognition error observed is considered to be the error in detecting the speech period and spectral distortion when the background noise level was extremely high.

3.3 Speech Recognition Technologies for Future Telecommunications Applications(10)

The telephone directory service is an important service for users in telecommunications. To realize this service using speech recognition, we have to establish large-vocabulary, continuous, speaker-independent speech recognition. Algorithms for recognizing large-vocabulary continuous speech require (1)accurate scoring for phoneme sequences, (2)reduction of trellis calculation, and (3) efficient pruning of phoneme sequence candidates. To meet these requirements, we have proposed the following four methods.

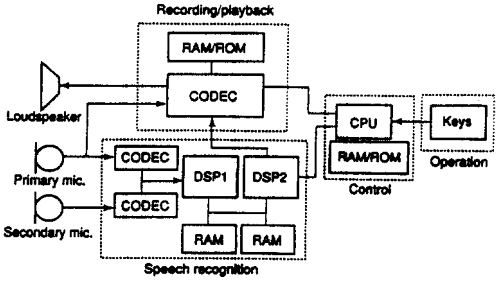


Fig.4 Block Diagram of Speech Recognizer in Voice Dialing Telephone

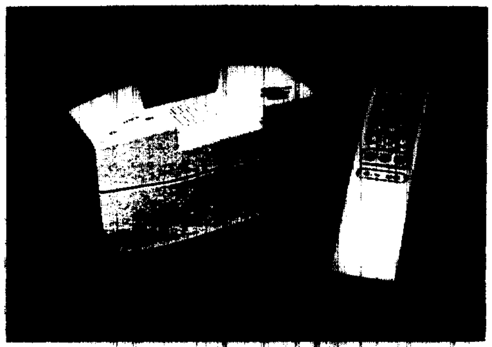


Fig.5 Prototype of Voice Dialing Telephone for Automobiles

Table 1 Speech Recognition Performance in an Automobile

Driving Condition	Recognition Accuracy (%)		
	1st Place	Up to 2nd	Up to 3rd
Engine Idling (Window Closed)	95.8	100	—
Light Driving (Window Closed)	89.6	97.9	97.9
Light Driving (Window Down)	68.8	85.4	85.4
Highway Driving (Window Closed)	85.4	93.8	93.8
Highway Driving (Window Partly Down)	77.1	85.4	89.6

(a)Two-Stage LR Parser

The two-stage LR parser uses two classes of LR tables: a main grammar table and sub-grammar tables as shown in Figure 6. These grammar tables are separately compiled from a context-free grammar. The sub-grammar tables deal with semantically classified items, such as city names, area names, block numbers, and subscriber names. The main grammar table controls the relationships between these semantic items. Dividing the grammar into two classes has two advantages; since each grammar can be compiled separately, the time needed for compiling the LR table is reduced, and the system can easily be adapted to many types of utterances by simply changing the main grammar rules.

The flow of the speech recognition process is as follows:

(1) The two stage LR parser predicts the following phonemes using the LR tables. It transfers them to an HMM phoneme verifier.

(2) The phoneme verifier calculates the phoneme likelihood using the trellis algorithm and returns the likelihood to the two stage LR parser.

(3) The state of the candidate then moves to the next state in the LR table via shift and reduce operations.

(4) During steps (1)-(3), all possible phoneme sequence candidates are constructed in parallel. These phoneme sequence candidates are pruned by using their HMM likelihoods.

(b) Accurate Scoring

The algorithm uses a backward trellis as well as a forward trellis so as to accurately calculate the score of a phoneme sequence candidate. The backward trellis likelihood is calculated without any grammatical constraints on the phoneme sequences; it is used as a likelihood estimate for potential succeeding phoneme sequences.

(c) Adjusting Window

An algorithm for determining an adjusting window that restricts calculation to a probable part of the trellis for each predicted phoneme has been proposed. The adjusting window has a length of 50 frames (400 ms). The score within the adjusting window is calculated by convoluting the forward and backward trellises. In this procedure, the likelihood in the backward trellises is multiplied by (1-ε), where ε is a small value.

(d) Merging Candidates

The LR tables need multiple pronunciation rules to cover allophonic phonemes, such as the devoicing and long vowels noted in Japanese pronunciation. These multiple rules cause an explosion of the search space. To reduce the search space, phoneme sequence candidates as well as grammatical states are merged when they are phonetically and semantically the same. Candidate word sequences having the same meaning are further merged, ignoring the differences in non-keywords.

This algorithm was applied to a telephone directory assistance system that recognizes spontaneous speech that contains addresses and names of more than 70,000 subscribers. Table 2 shows the size of the vocabulary for each semantic item. The grammar used in this system has various rules for interjections, verb phrases, post-positional particles, etc. It was made by analyzing 300 sentences in simulated telephone directory assistance dialogs. The word perplexity was about 70,000. In this task, no constraints by the directory database were placed on the combination of addresses and subscriber names.

Two types of speaker-independent HMM were prepared to evaluate the algorithm: 56 context-independent phoneme HMMs, and 358 context-dependent phoneme HMMs. The proposed algorithm was evaluated on the basis of 51 sentences that included 184 keywords. These utterances contained various interjections and verb phrases. They were spontaneously uttered by eight different speakers. Experimental results confirmed the effectiveness of merging at the meaning level and the context-dependent HMMs. These techniques achieved an average sentence understanding rate of 65% and an average keyword recognition rate of 89%. The results show that the proposed algorithm performs well in spite of the large perplexity.

3.4 Direction of Speech Recognition Technologies

There are many kinds of services that need using speech recognition in the telecommunications world. However, the words used in each service are usually completely different. To expand services using speech recognition, speech recognition technologies should have the following capabilities for corresponding customers' requests.

(1) Flexible Vocabulary (Vocabulary-independent) Recognition

With the whole-word-model based speech recognition, it is time and cost consuming to construct word model or word templates. It would be better if new vocabularies could be created without having to collect training data.

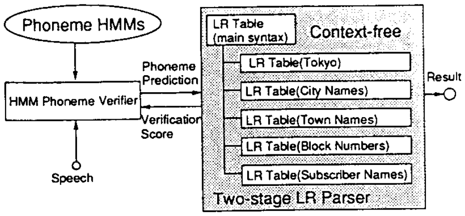


Fig.6 Continuous Speech Recognition System for Telephone Directory Assistance

Table 2 Vocabulary Sizes for Telephone Directory Assistance Task

Semantic Item	City Names	Town Names	Block Numbers	Subscriber Names
Vocabulary Size	2	27	620	71,251

In our new recognition system named "ECLAIR", it is very easy to change recognized words and to construct a new system. We define a vocabulary by providing only a phonemic transcription for each word.

(2) Easy Implementation and Low-cost Realization

The speech recognition algorithm itself improves day by day. Thus it is very efficient to implement the revised (newest) algorithm easily, even after system introduction. To realize this function, the recognition function should be realized using only software without special hardware or firmware. We have realized these requirement by using high speed computers. Such as speech recognition server can be shared by the plural clients over the network. Speech recognition algorithms can be used easily and economically both for users and system developers without worrying about the details of speech recognition.

(3) Multi-modal System

Personal computers or workstations will be introduced in the home as they become cheaper in the near future. In these circumstances, users can use any input/output device they want to use. Speech recognition could be used when the number of choices is large and it is difficult to find the desired item. In the telephone directory assistance system mentioned above, multi-modal dialogue functions were implemented and tested. A multi-modal speech dialog system for telephone directory assistance with three input devices (microphone, keyboard, and mouse) and two output devices (speaker and display) has been constructed as shown in Fig. 7. The system has been evaluated from the human-machine-interface point of view[11].

4. Speech Synthesis Technologies

4.1 Overview of Speech Synthesis Technologies

Speech synthesis is the key technology in the transfer to users of computer processing results and/or stored information. With the recent increase in the number of computer systems, the demand to communicate with the systems via speech is increasing, and speech synthesis technologies are expected to satisfy these demands[12],[13].

Speech synthesis technologies are classified into two main types: speech digitizing or analyzing to develop prerecorded speech data and speech production by rules. The former method creates prerecorded words or phrases, selects the appropriate speech units from a speech file, and simply concatenates them to yield continuous voice messages that match the input data. This method usually offers high speech

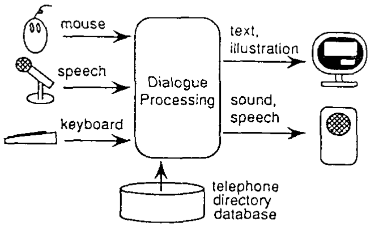


Fig.7 Functions of Multi-modal Dialogue System for Telephone Directory Assistance

quality because the voice units are relatively long and meaningful. Almost all existing systems in the real world, such as arrival and departure announcements at railroad stations, booking systems over the telephone network and telephone number guidance, have adopted this method. While it is preferable where there is a finite and limited vocabulary, it's not suitable where the vocabularies are very large and are continually being changed and updated.

On the other hand, rule based speech synthesis creates voice messages from phoneme string and prosody information using various kinds of rules and tables. Among these, the speech synthesis method that accepts ASCII text input is called Text-To-Speech (TTS) synthesis. While TTS is capable of generating an arbitrary vocabulary, user acceptance of TTS output is still not as high as that of digitized voice. TTS applications remain sporadic in the public domain. Recent research efforts have tried to advance TTS performance and its speech quality has steadily improved. It is expected that TTS will be one of the critical medium transformation technologies in the coming multimedia age(14).

4.2 Current Applications of Speech Synthesis Technologies over the Telephone Network

Various voice services using the telephone network have been developed and are well accepted by the public. There are several reasons for this. First, only a regular telephone set is needed by the user. Next, the user can input information using the keypad of the telephone set. Advances in speech technology make it easier to process speech data. The initial services were large scale services providing several hundreds of telephone ports such as banking systems. Lately, small scale unique services such as information offering systems for a limited number of members, similar to the 900 services in the U.S., have been developed and adopted.

In the ANSER System, prerecorded speech is basically used for voice response messages, however account names and the amount of money are generated for transfer acknowledgment and balance inquiry by rule based speech synthesis. In the information inquiry domain, a recent trend is for public departments such as city office, tax office and police stations, to offer voice guidance for administrative tasks via the telephone network. From the view point of equality and simplicity, these services will spread to private companies.

4.3 Problems with the Prerecorded Method

The most significant problem with the prerecorded method is the excessive time needed to prepare a speech file. It normally takes one or two months to create a one minute message, from voice recording to loading the speech data into the target memories. Recorded samples must be evaluated, and if the voice quality is insufficient, the sample must be rerecorded under modified conditions. There are other difficulties. The same announcer must be retained until the service is closed and additions to the vocabulary must be uttered so as to achieve smooth integration when inserted into the carrier sentences in the service situations.

TTS technology has, on the other hand, advanced in recent years and if the correct prosody and phoneme data is provided, the synthetic voice is well accepted by most users. Therefore, using TTS technology, especially that proposed by NTT, can create a truly effective voice message creation system(15). In the system, the target text is initially converted into phoneme strings and prosody data. The operator then revises the text analysis results such as reading for Kanji characters, pause locations and accentual positions by hand. The TTS function is then executed to synthesize the speech. The operator is able to finely modify the synthesized speech by altering speech parameters if necessary. The total time required to produce a voice message is much less than the digitizing method. Accordingly, this system significantly reduces the cost of preparing the speech file.

4.4 TTS Technology and Trend

Most Japanese TTS systems have three processes as shown in Figure 8. The first process is text analysis; morphological analysis is used to separate the input text into individual words by referring to a word dictionary. Kanji (Chinese) characters are then converted into Kana characters which correspond to syllables. At the same time, accentual phrases and pause locations are determined by word compound rules. Accentual position in a compounded word is shifted according to accent moving rules. Second, prosody patterns such as fundamental frequency pattern, segmental power pattern and phoneme duration pattern are calculated using pattern setting models and parameter values prestored

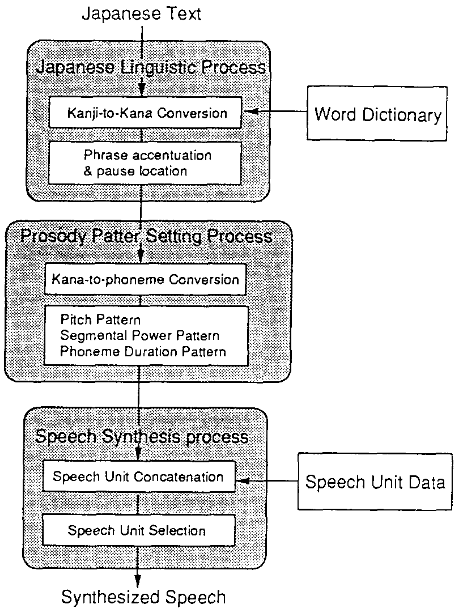


Fig.8 Japanese Text-to-Speech Processes

in tables. Last, each speech unit, which corresponds to a phoneme segment, is chosen one by one from a speech unit file and concatenated to yield continuous speech.

In recent years, several speech synthesis methods based on waveform concatenation have been proposed that offer high speech quality. The intelligibility of their output is higher than that of conventional parametric synthesis methods. The hardware architecture is rather simple because proprietary DSPs are not needed.

One interesting trend is the rapid increase in CPU performance of Personal Computers (PCs) and Work Stations (WSs); high grade PCs offer several tens of MIPS. Thus it is now feasible to execute TTS synthesis in software alone.

4.5 Service Segment of TTS Synthesizers

The NTT group developed the first TTS product in Japan in 1986, and released a TTS card with PC interface in 1989. It is called "Syaberinbo" (literally chattering boy). The first version of "Syaberinbo" employed the LSP (Line Spectrum Pair)[16] speech synthesis method, a kind of LPC, and the speech units were CV(Consonant-Vowel) and VC dyphone type segments. The second version of the TTS card, called "Syaberinbo HG (High Grade)", was developed in 1991. This TTS card adopts a COC (Context Oriented Clustering) method[17] for speech unit generation and so offers a more fluent voice.

Several voice-based information services named "Dial Q^2" (corresponds to AT&T's 900 service called "Multi-Quest"), are very popular in Japan. These services use the starting numbers of "0990". The pronunciation of Japanese "9" is the same of that of "Q" in English, so this is called "Dial Q^2" service. In these services, the TTS systems work well in services whose vocabularies frequently change. The application described below is interesting and confirms the advantages of TTS.

A newspaper company provides daily sports information such as professional baseball games, horse races and Sumo wrestling winners to its customers via the telephone network. One example is the progress of a baseball game; the score is sent from a PC in the stadium to the newspaper company via the telephone network and the user can hear the latest score through the TTS service. A very new service which transmits CD recordings over the telephone network has just started from July 1994. In the system, TTS technology developed by NTT is used to introduce the musicians and music titles. These applications suggests other similar inquiry services such as stock market reports, traffic jam information, and news. A reservation system that uses a TTS card for medical examinations is being used in several hospitals in Japan.

5. Conclusion

This paper has briefly introduced several systems, devices and technologies developed at NTT, associated with speech recognition and speech synthesis. These technologies include isolated word recognition applicable to both speaker-dependent and speaker-independent recognition, speaker-independent word spotting, speaker-independent, large vocabulary, continuous speech recognition and text-to-speech synthesis. Some of them are being used in the real word for providing several services over the telephone network .

Speech technologies will play an increasingly important role in the coming "Multimedia Era". As speech technologies move from the laboratory into the real world, there is a need for systems that respond to a wide variety of real world problems. In spite of the great progress made in recent years, there remain a large number of problems to be solved.

Acknowledgment

The technologies and applications described in this paper are results of the research and development efforts of a large number of individuals within NTT Laboratories over a long period. Several individuals must be acknowledged for their personal contributions to the material presented here. These include Dr. R. Nakatsu, Dr. K. Shikano, Dr. Y. Minami, Mr. O. Yoshioka, Mr. Y. Suzuki, Mr. M. Tobita and Mr. A. Imamura for their work in speech recognition; Dr. H. Sato, Mr. K. Hakoda, Mr. S. Nakajima and Mr. H. Mizuno for their work in speech synthesis.

References

(1) S. Furui, K. Shikano, S. Matsunaga, T. Matsuoka, S. Takahashi and T. Yamada, "Recent Topics in Speech Recognition Research at NTT Laboratories", Proceedings of DARPA Speech and Natural Language Workshop, pp.162-167 (1992).

(2) S. Furui, "Recent Advances in Speech Recognition Technology at NTT Laboratories", Speech Communication, pp.195-204 (1992).

(3) T. Hirokawa, "Applications of Japanese text-to-speech synthesizer", Speech Tech89, pp30-32 (1989).

(4) R. Nakatsu, "Anser An Application of Speech Technology to the Japanese Banking Industry", Computer, pp. 43-48 (1990).

(5) A. Imamura and Y. Suzuki, "Speaker-independent word spotting and a Transputer-based implementation", Proceedings of ICSLP 90, pp.537-540 (1990).

(6) Y. Suwa, Y Noda and M. Midorikawa, "A Study on the Application of Speech Recognition to Advanced IN", Technical Report of IEICE, SSE93-88, IN93-95, CS93-111, pp.147-1152 (1993.10).

(7) M. Kitai and H. Nishi, "The Evaluation of Trial Results for a Voice Activated Telephone Intermediary System", in the Proceedings of IVTTA94 (1994).

(8) Y. Suzuki and M. Tobita, "Designing a Speech Recognizer for Use in Automobile", Proceedings of AVIOS 92, pp.273-279.

(9) Y. Nakadai and N. Sugamura, "A Speech Recognition Method for Noise Environments Using Dual Inputs", Proceedings of ICSLP 90, pp.1141-1144 (1990).

(10) Y. Minami, K. Shikano, S. Takahashi and T. Yamada, "Search Algorithm that Merges Candidates in Meaning Level for Very Large Vocabulary Spontaneous Speech Recognition", Proceedings of ICASSP 94, II-141-144 (1994).

(11) O. Yoshioka, Y. Minami and K. Shikano, "Multi-modal Dialogue System for Telephone Directory Assistance", Proceedings of ICSLP94 (1994).

(12) R. Nakatsu, "Market trends of speech recognition and synthesis technologies", JASJ, Vol.48, No.1, pp.60-65 (1992).

(13) Y. Mitome, "Applications and a future prospect of speech synthesis", JASJ, Vol.49, No.12, pp.875-880 (1993).

(14) T. Hirokawa, K. Itoh and H. Sato, "High quality speech synthesis system based on waveform concatenation of phoneme segments", Trans., IEICE, Vol.E76-A, No.11, pp.1964-1970 (1993).

(15) T. Hirokawa, K. Itoh and K. Hakoda, "Speech editor based on enhanced user-system interaction" , Proceedings of AVIOS, pp.39-45 (1993).

(16) N. Sugamura, F. Itakura, "Speech analysis and synthesis methods developed at ECL in NTT (From LPC to LSP), " Speech Communication No.5, pp.199-215 (1986).

(17) S. Nakajima and H. Hamada, "Automatic generation of synthesis units based on context oriented clustering", Proceedings of ICASSP88, pp.133-136 (1988).