論文 / 著書情報
Article / Book Information

Title	Recent Advances in Speech Processing Technology
Author	SADAOKI FURUI
Journal/Book name	International Symposium on Multi-Technology Information Processing (ISMIP'96), Vol. , No. , pp. 9-16
Issue date	1996, 12

ISMIP'96, Dec. 16 - 18, Hsinchu, Taiwan, pp. 9 - 16, 1996

Recent Advances in Speech Processing Technology

Sadaoki Furui
NTT Human Interface Laboratories
3-9-11, Midori-cho, Musashino-shi, Tokyo, 180 Japan
Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, 152 Japan
furui@splab.hil.ntt.co.jp, furui@cs.titech.ac.jp

ABSTRACT

This paper reviews recent speech processing technologies, including speech recognition, synthesis and coding, and discusses their current capabilities and appropriate applications. It also describes the most important research problems, and tries to forecast where progress will be made in the near future and what applications will become commonplace as a result of the increased capabilities. Robust performance in speech recognition and more flexibility in synthesizing speech will continue to be major problems that must be solved expeditiously.

1. INTRODUCTION

Speech recognition, synthesis, and coding systems are expected to play important roles in an advanced multi-media society with user-friendly human-machine interfaces [7]. Speech recognition systems include not only those that recognize messages but also those that recognize the identity of the speaker. Services using these systems will include voice dialing, database access and management, guidance and transactions, automated reservations, various order-made services, dictation and editing, electronic secretarial assistance, robots, automated interpreting (translating) telephony, security control, digital cellular communications, and aids for the handicapped (e.g., reading aids for the blind and speaking aids for the vocally handicapped).

Figure 1 shows a typical structure for task-specific voice control and dialog systems [5]. Although the speech recognizer, which converts spoken input into text, and the language analyzer, which extracts meaning from text, are separated into two boxes in the figure, they should work closely together, since the recognizer must use semantic information efficiently in order to obtain the correct text. How to combine these two functions is a most important issue, especially in conversational speech recognition (understanding). The meanings extracted by the language analyzer are used to drive an expert system to select the desired action, issue commands to various systems, and receive data from these systems. Replies from the expert system are sent to a text generator which constructs reply texts. These are then converted into speech by a text-to-speech synthesizer. "Synthesis from concepts" is performed by the combination of the text generator and the text-to-speech synthesizer.

Figure 2 shows hierarchical relationships among the various types of speech recognition, synthesis, and coding technology [2]. The higher the level is, the more abstract the information is. This figure is closely related to Fig. 1; speech recognition/understanding is the process extending upward from the bottom to one of the higher levels of Fig. 2, and speech synthesis is the process progressing downward from one of the higher levels to the bottom. Historically, speech technology originated from the bottom, and has developed toward the extraction and handling of higher-level information. Some of the technologies indicated in the figure remain to be investigated.

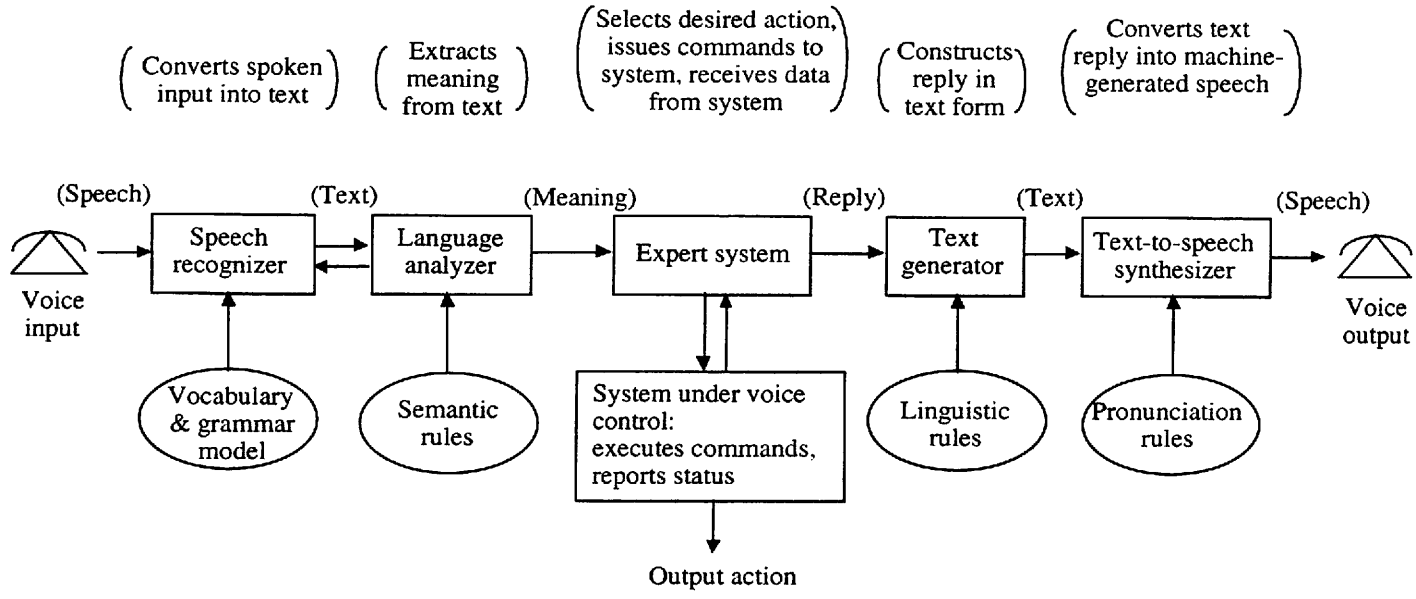

Figure 1 - Typical structure for task-specific voice control and dialog systems.

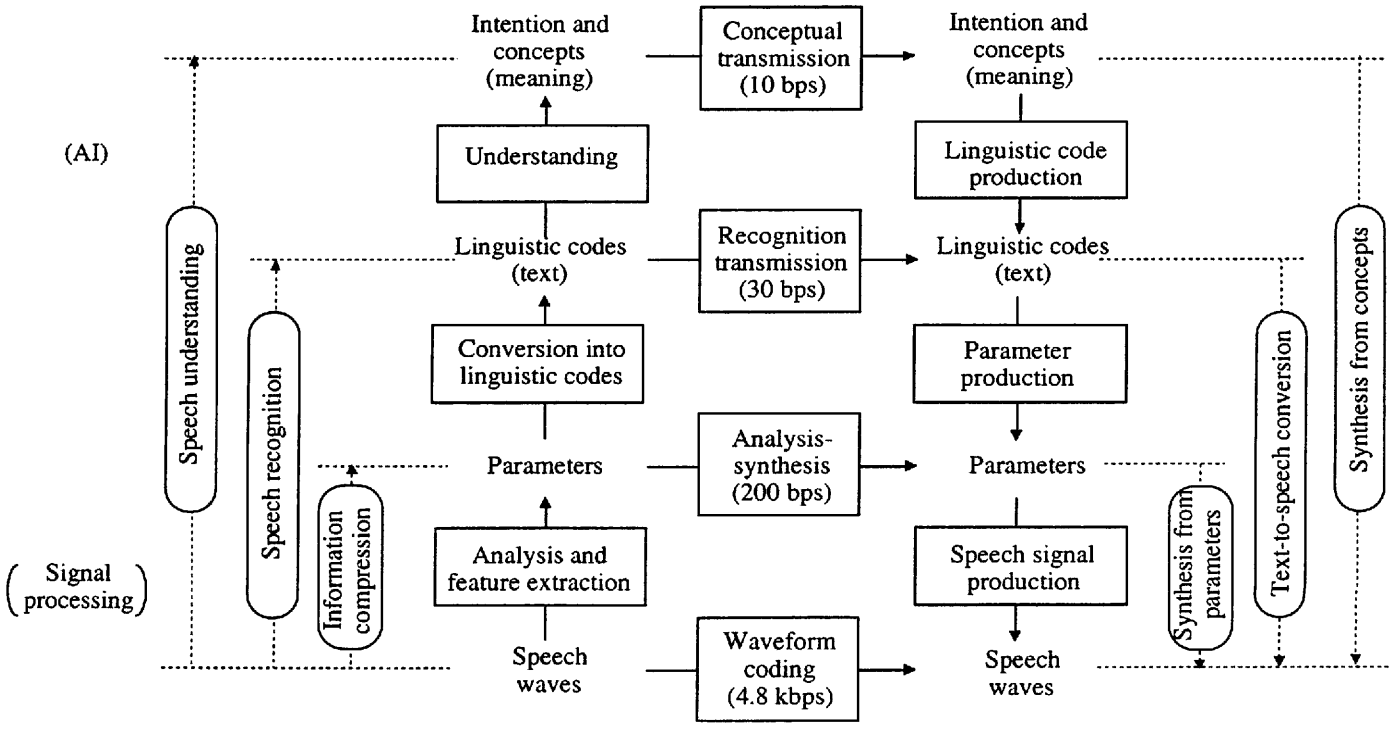

Figure 2 - Principal speech information proceesing technologies and their relationships.

2. SPEECH RECOGNITION

2.1 Overview

Typical large-vocabulary continuous speech recognition systems currently under investigation consist of parts for signal processing, feature extraction, phoneme recognition, word recognition, and sentence recognition [19]. A speech wave is first converted into digital form in the signal processing part, and then converted into a time series of feature parameters, such as cepstra and delta-cepstra, in the feature extraction part. Although various methods for extracting and using prosodic features, that is, time functions of voice pitch (height) and amplitude, have been investigated, no satisfactory method has yet been invented.

The system predicts a sentence (hypothesis) that is likely to be spoken by the user, based on the current topic, the meaning of words, and language grammar, and represents the sentence as a sequence of words. This sequence is then converted into a sequence of phoneme models which were created beforehand in a training stage. Each phoneme model is typically represented by an HMM. The likelihood (probability) of producing the time series of feature parameters from the sequence of the phoneme models is calculated, and combined with the linguistic likelihood (appropriateness) of the hypothesized sentence to calculate the overall likelihood (probability) that the sentence was uttered by the speaker. The (overall) likelihood is calculated for other sentence hypotheses, and the sentence with the highest likelihood score is chosen as the recognition result. The recent incorporation of stochastic language modeling has greatly improved the accuracy of speaker-independent, large-vocabulary, continuous speech recognition.

A very large vocabulary continuous speech recognition system with an 80,000-word vocabulary was built as a key element of a multi-modal dialogue system for telephone directory assistance at NTT Labs [16]. It is based on the HMM-LR algorithm using HMMs as phoneme models and a generalized LR parser as a language model.

2.2 Dynamic spectral features

Psychological and physiological research into human speech perception mechanisms shows that the human hearing organs are highly sensitive to changes in sounds, that is, to transitional (dynamic) sounds, and that the transitional features of the speech spectrum and the speech wave play crucial roles in phoneme perception [1].

The representation of the dynamic characteristics of speech waves and spectra has been studied, and several useful methods have been proposed. However, the performance of these methods is not yet satisfactory, and most of the successful speech analysis methods developed thus far assume a stationary signal. It is still very difficult to relate time functions of pitch and energy to perceptual prosodic information. If good methods for representing the dynamics of speech associated with various time lengths are discovered, they should have a substantial impact on the course of speech research.

2.3 Robust speech recognition

It is crucial to establish methods that are robust against voice variation due to individuality, the physical and psychological condition of the speaker, telephone sets, microphones, network characteristics, additive background noise, speaking styles, and so on [4][8][11]. It is also important for the systems to impose few restrictions on tasks and vocabulary. To solve these problems, it is essential to develop automatic adaptation techniques. To cope with the background noise problem, an HMM composition technique has been successfully used in many recognition systems [16].

Extraction and normalization of (adaptation to) voice individuality is one of the most important issues [3]. A small percentage of people occasionally cause systems to produce exceptionally low recognition rates. This is an example of the "sheep and goats" phenomenon. Speaker normalization (adaptation) methods can usually be classified into supervised (text-dependent) and unsupervised (text-independent) methods. Experiments have shown that people can adapt to a new speaker's voice after hearing just a few syllables, irrespective of the phonetic content of the syllables [13].

2.4 Language modeling

Stochastic language modeling, such as bigrams and trigrams, has been a very powerful tool, so it would be very effective to extend its utility by incorporating semantic knowledge. It would also be useful to integrate unification grammars and context-free grammars for efficient word prediction. Adaptation of linguistic models according to tasks and topics [14] is also a very important issue, since collecting a large linguistic database for every new task is difficult and costly.

2.5 Spontaneous speech recognition

One of the most important issues for speech recognition is how to create language models (rules) for spoken language [6]. When recognizing spontaneous speech in dialogs, it is necessary to deal with variations that are not encountered when recognizing speech that is read from texts. These variations include extraneous words, out-of-vocabulary words, ungrammatical sentences, botched utterances, restarts, repetitions, and style shifts. It is crucial to develop robust and flexible parsing algorithms that match the characteristics of spontaneous speech. How to extract contextual information, predict users' responses, and focus on key words are very difficult and important issues.

2.6 Speaker recognition

Speaker recognition accuracy has been improved by using HMMs, likelihood normalization, and the text-prompted method [9][21]. The text-prompted method was recently proposed at NTT Labs to cope with the problem that conventional systems are easily defeated by a recorded voice [15]. In this method, a new text is prompted by the system every time the system is used. The system accepts the input utterance only when it determines that the registered speaker uttered the prompted sentence. That is, the system not only recognizes speakers, but also rejects utterances whose text differs from the prompted text, even if it is uttered by a registered speaker. Because the vocabulary is unlimited, prospective impostors cannot know in advance the sentence they will be prompted to say. A recorded and played-back voice can thus be correctly rejected.

2.7 Minimum error discriminative training

Unlike conventional maximum likelihood training, which estimates a model based only on training utterances from the same category, a discriminative training approach takes into account models of other competing categories and formulates the optimization criterion such that category separation is enhanced and the classification/recognition error rate on the training data is directly minimized. The optimization solution is obtained with a generalized probabilistic descent algorithm. This method is, therefore, called MCE/GPD (minimum classification error with generalized probabilistic descent). Unlike the Bayesian framework, it does not require estimation of probability distributions, which usually cannot be reliably obtained. This method has been successfully applied in various experimental studies for both speech and speaker recognition [12].

3. SPEECH SYNTHESIS

3.1 Overview

A typical structure of a text-to-speech synthesizer consists of text-preprocessing, text-to-phoneme conversion (letter-to-sound), prosodic contour generation (pitch and duration), and speech signal reconstruction (synthesis). Although a large number of text-to-speech synthesis systems are commercially available and they can generally provide intelligible speech, they all still lack naturalness. The quality of sound attained has recently jumped up, as the result of using waveform concatenation methods for speech signal reconstruction [10][18], but it is still a long way from a naturally-articulated voice. The most serious problems are in the first two stages. First, the reading of a written text is often error-prone, such as incorrect pronunciation of proper names, foreign words, abbreviations, acronyms etc. Second we do not have accurate rules for prosody control. Prosody is the major cause of the unnaturalness of current synthesized speech. It tends to be perceived as monotonous or jerky, and synthesized speech fails to transmit "meaning" [22].

3.2 Text preprocessing and text-to-phoneme conversion

In general, the first two modules for converting letters to phonemes combine heuristic rules and the utilization of an orthographic-phonetic dictionary. Actually, the pronunciations of many words and symbols, especially abbreviations, are word-dependent and so ambiguous that they are only clarified through their context. Overcoming this problem will require dictionaries and methods that are much more complex and sophisticated than those currently used. They must include rules based on a large database and natural language understanding algorithms taking into account the syntactic, semantic, and pragmatic aspects. It is obvious that the accuracy of pronunciation depends on how well the message content or the surrounding context is understood.

3.3 Prosodic contour generation

Correct prediction of stress location for accented languages needs the removal of grammatical ambiguities in utterances (e.g., verb/noun ambiguities). Although an adequate prosodic counter for each utterance could ultimately be generated by its meaning, at least partial syntactic parsing of the utterance is indispensable. The lack of sufficient syntactical information leads in many cases to prosodic segmentation errors, which are unacceptable for listeners.

Prosodic contours are generally specified by rules, either directly elaborated from real contour analysis or derived from formal models. It could also be possible to achieve further improvement in the reconstruction of prosodic contours by optimal use of prosodic databases compiled from live recordings.

3.4 Controlling synthesized voice quality

The problems for speech synthesis include controlling synthesized voice quality and choosing an individual speaking style, multi-lingual and multi-dialectal synthesis, choice of application-oriented speaking styles, and the addition of emotion. In current commercial speech synthesizers, voice quality can be selected from male, female, and children's voices. No system has been constructed, however, that can precisely select or control the synthesized voice quality. Automatic creation of new voices from a given voice is an interesting and important issue especially for future automatic translation telephony. Research into the mechanism underlying voice quality, including voice individuality, is thus needed to ensure that a synthesized voice is capable of imitating a desired speaker's voice or to provide any desired voice quality such as harshness or softness.

4. SPEECH CODING

4.1 Overview

Various efforts have been devoted to speech coding, that is, reducing the transmission rate (or memory requirement) for speech without degrading its quality. In the last few years, the demand for speech coding has become very high, especially for digital cellular communication systems (mobile and portable telephones). Progress in reducing the transmission rate is shown in Fig. 3 [17]. The horizontal axis shows the year in which each coding method was developed or approved as a standard.

Most new low-bit rate coders belong to the class of CELP (Code-Excited Linear Prediction) coders. In both LPC vocoder and CELP, the speech waveform is generated by a linear prediction synthesis filter with an excitation source filter. The vocoder models the excitation source by either a pulse sequence of pitch period or a fixed noise sequence. In CELP, however, a codebook is used, and the best excitation vector is selected so that the perceptually weighted distortion between input and the synthesized speech is minimized. In addition, a periodic signal is generated by repeating the excitation signal in the previous frame, and a non-periodic signal is selected from a random excitation codebook. Naturalness and robustness are significantly improved by CELP.

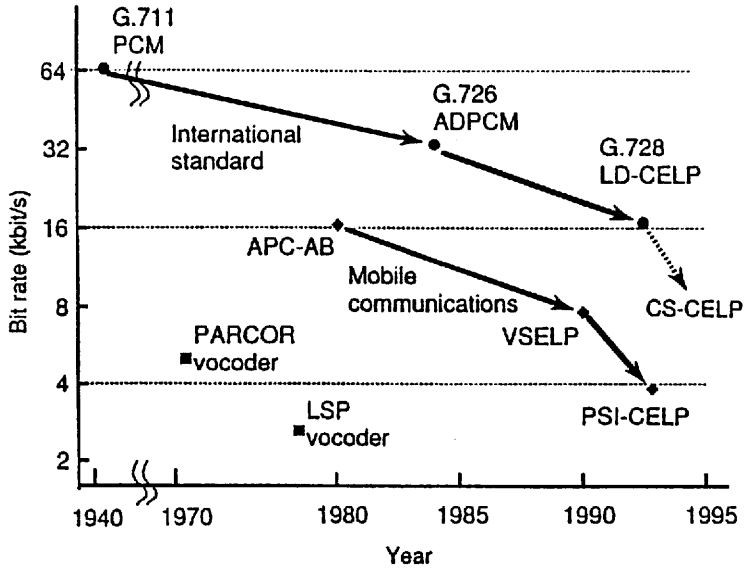

Figure 3 - Progress in transmission rate reduction.

Several years ago, extensive efforts to develop new coding algorithms and error correction techniques as well as the progress of LSI technology enabled the 8kb/s codec (coder and decoder) to be used for digital mobile telephones. In 1990, the VSELP (Vector Sum Excitation Linear Prediction) codec was selected as a full-rate speech codec standard for digital cellular systems both in North America and Japan. VSELP is a CELP coding method invented in Motorola, in which the excitation source is generated by summing several fixed basis vectors.

However, the digital cellular system is expected to run out of radio channels in Japan if the demand continues to grow at the current rate. Therefore, as a result of extensive research, NTT group has developed the PSI-CELP (Pitch Synchronous Innovation CELP) speech coding method and has achieved quality comparable to the current standard (VSELP), while using half the bit rate. PSI-CELP has been chosen as the half-rate codec standard for digital cellular system in Japan [17].

4.2 PSI-CELP speech coder for the digital cellular system

One of the most important features of the PSI-CELP speech coder is that random excitation vectors are given pitch periodicity for voiced speech by a pitch synchronization process. This feature is the origin of the name PSI-CELP. It significantly reduces the distortion and improves the quality for voiced speech, especially for female voices. In addition to the pitch synchronization, a two-channel conjugate structure and fixed codebook for transient speech signals were newly developed. These two features reduce the memory requirement, enhance robustness against channel errors, and improve the quality for the transient part.

5. FUTURE PROSPECTS OF SPEECH TECHNOLOGY

5.1 Ultimate speech recognition and synthesis systems

Ultimate speech recognition systems should be capable of robust, speaker-independent or speaker-adaptive, continuous speech recognition [5]. Speaker recognition techniques are expected to be widely used in the future as methods of verifying the claimed identity in telephone banking and shopping services, information retrieval services, remote access to computers, credit-card calls, etc.

Ultimate speech synthesis and recognition systems that are really useful and comfortable for users should match or exceed human capabilities. That is, they should be faster, more accurate, more intelligent, more knowledgeable, less expensive, and easier to communicate with than human staff. For this purpose, the ultimate systems must be able to handle conceptual

information, the highest level of information in Fig. 2.

It is, however, neither necessary nor useful to try to use speech for every kind of input and output in computerized systems. Although speech is the fastest and easiest means of input and output for a simple exchange of information with computers, it is inferior to other means in conveying complex information. There needs to be an optimal division of roles and cooperation in a multimedia environment that includes images, text, tactile signals, handwriting, etc.

5.2 Articulatory and perceptual constraints

To solve various problems, it is necessary to promote sure and steady research and development by grasping the essence of speech phenomena, instead of developing methods by simply looking at them superficially. Speech technology is related to many scientific and engineering fields, such as physiology and psychology of speech production and perception, acoustics (physics), signal processing, communication and information theory, computer science, pattern recognition, and linguistics; it has an inter-disciplinary nature. It can also be said that speech research exists at the boundary between natural science and engineering. Knowledge and technology from a wide range of areas, including the use of articulatory and perceptual constraints, will be necessary to develop speech technology.

5.3 Speech science vs. statistical approaches

There is no doubt that most recent progress in speech and speaker recognition, even in speech synthesis, came from statistical approaches, such as HMM and stochastic language modeling. These approaches were made possible by recent remarkable progress in computer power. Statistical approaches are usually more reliable and, in many cases, more powerful than knowledge-based approaches, provided that we can obtain a large enough database. However, there is always some limit to the size of the database and we always encounter some discrepancy between the training database and the testing data. Therefore, even in the statistical approaches we need reasonable models.

Although it is not always necessary or efficient for speech synthesis/recognition systems to directly imitate human speech production and perception mechanisms, it will become more important in the near future to build mathematical models based on these mechanisms in order to improve performance [2].

5.4 Telecommunications applications

The most promising application area for speech technology is telecommunications [20]. The fields of communications, computing, and networking are now converging in the form of personal information/communication terminals. In the near future, personal communications services (known as PCS) will become popular, and everybody will have their own portable telephone. Several technologies will play major roles in this communications revolution, but one of the key ones will be speech processing technology. By using advancing speech synthesis/recognition technology, telephone sets will become useful personal terminals for communicating with computer systems.

6. CONCLUSION

Speech recognition, synthesis and coding systems are expected to play important roles in an advanced multi-media society with user-friendly human-machine interfaces. This paper reviewed the most important research problems to be solved in order to achieve ultimate recognition, synthesis, and coding systems. The problems include dynamic spectral features, robustness against various voice variation, adaptation/normalization techniques, language modeling, and MCE/GPD approach in speech/speaker recognition, text processing, signal processing and speaking style control in synthesis, synthesis from concepts, and use of articulatory and perceptual constraints.
Although speech recognition, synthesis and coding research has thus far been done independently for the most part, there will be increasing interaction between these aspects until common problems are being investigated and solved simultaneously.

REFERENCES

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

[2] S. Furui: "Digital speech processing, synthesis, and recognition", Marcel Dekker, Inc., New York (1989)

[3] S. Furui: "Speaker-independent and speaker-adaptive recognition techniques", in *Advances in Speech Signal Processing*, ed. by S. Furui and M. M. Sondhi, Marcel Dekker, Inc., New York, pp.597-622 (1992)

[4] S. Furui: "Toward robust speech recognition under adverse conditions", ESCA Workshop on Speech Processing in Adverse Conditions, pp.31-42 (1992)

[5] S. Furui: "Towards the ultimate synthesis/recognition system", *Voice Communication between Humans and Machines*, ed by D. B. Roe & J. G. Wilpon, National Academy Press, Washington D. C., pp.450-466 (1994)

[6] S. Furui: "Prospects for spoken dialogue systems in a multimedia environment", Proc. ESCA Workshop on Spoken Dialogue Systems, pp.9-16 (1995)

[7] S. Furui: "Recent advances in speech processing technology", Proc. 15th Int. Cong. on Acoustics, pp.III-497-504 (1995)

[8] S. Furui: "Flexible speech recognition", Proc. EUROSPEECH '95, pp.1595-1603 (1995)

[9] S. Furui: "An overview of speaker recognition technology", *Automatic Speech and Speaker Recognition*, ed by C.-H. Lee, F. K. Soong & K. K. Paliwal, Kluwer Academic Publishers, Boston, pp.31-56 (1996)

[10] T. Hirokawa, K. Itoh and H. Sato: "High quality speech synthesis based on wavelet compilation of phoneme segments", Proc. Int. Conf. on Spoken Language Processing, Th.sAM.2.4, pp.567-570 (1992)

[11] B. H. Juang: "Speech recognition in adverse environments", Computer Speech and Language, 5, pp.275-294 (1991)

[12] S. Katagiri, C.-H. Lee and B.-H. Juang: "New discriminative algorithm based on the generalized probabilistic descent method", Proc. IEEE Workshop on Neural Networks for Signal Processing, pp.299-309 (1992)

[13] K. Kato and S. Furui: "Listener adaptability for individual voices in speech perception", IEICE Technical Report, H85-5 (1985)

[14] R. Kuhn and R. De Mori: "A cache-based natural language model for speech recognition", IEEE Trans. Pattern Anal., Machine Intelligence, PAMI-12, 6, pp.570-583 (1990)

[15] T. Matsui and S. Furui: "Concatenated phoneme models for text-variable speaker recognition", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.II-391-394 (1993)

[16] Y. Minami, K. Shikano, S. Takahashi, T. Yamada, O. Yoshioka and S. Furui: "Large-vocabulary continuous speech recognition algorithm applied to a multi-modal telephone directory assistance system", Speech Communication, 15, pp.301-310 (1995)

[17] T. Moriya, T. Kaneko, and N. Kitawaki: "PSI-CELP speech coder for mobile communications", NTT Review, 6, 1, pp.97-100 (1994)

[18] M. Moulines and F. Charpentier: "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", Speech Communication, 9, pp.453-467 (1990)

[19] L. R. Rabiner and B. H. Juang: "Fundamentals of speech recognition" Prentice-Hall, Inc., New Jersey (1993)

[20] L. R. Rabiner: "The role of voice processing in telecommunications", Proc. 2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, pp.1-8 (1994)

[21] A. E. Rosenberg and F. K. Soong: "Recent research in automatic speaker recognition", in *Advances in Speech Signal Processing*, ed. by S. Furui and M. M. Sondhi, Marcel Dekker, Inc., New York, pp.701-737 (1992)

[22] C. Sorin: "Speech synthesis applications and research: Present and future", Proc. 13th Int. Cong. on Acoustics (1989)

AUTHOR

Sadaoki Furui, The Research Fellow and the Director of Furui Research Laboratory at NTT Human Interface Laboratories, and a Visiting Professor at the Graduate School of Information Science and Engineering, Tokyo Institute of Technology. His specialty is in the areas of speech analysis, speech recognition, speaker recognition and speech perception.

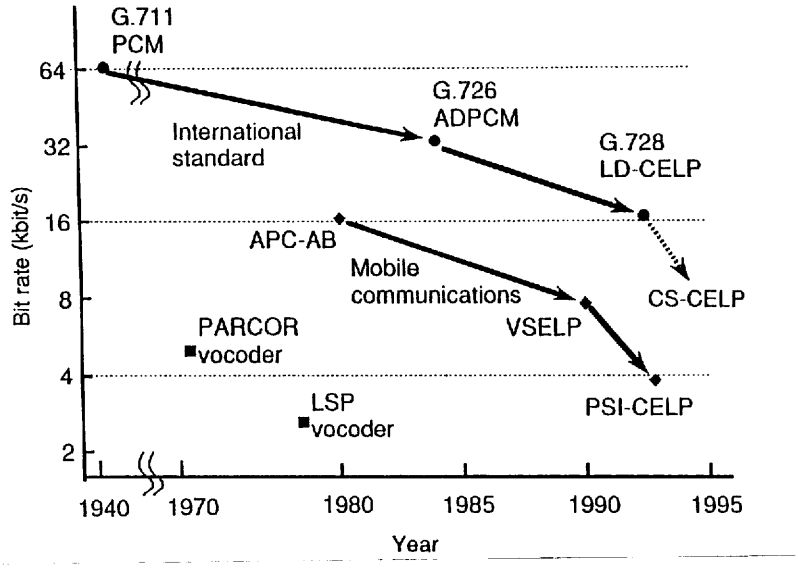

Figure 3 - Progress in transmission rate reduction.

Several years ago, extensive efforts to develop new coding algorithms and error correction techniques as well as the progress of LSI technology enabled the 8kb/s codec (coder and decoder) to be used for digital mobile telephones. In 1990, the VSELP (Vector Sum Excitation Linear Prediction) codec was selected as a full-rate speech codec standard for digital cellular systems both in North America and Japan. VSELP is a CELP coding method invented in Motorola, in which the excitation source is generated by summing several fixed basis vectors.

However, the digital cellular system is expected to run out of radio channels in Japan if the demand continues to grow at the current rate. Therefore, as a result of extensive research, NTT group has developed the PSI-CELP (Pitch Synchronous Innovation CELP) speech coding method and has achieved quality comparable to the current standard (VSELP), while using half the bit rate. PSI-CELP has been chosen as the half-rate codec standard for digital cellular system in Japan [17].

4.2 PSI-CELP speech coder for the digital cellular system

One of the most important features of the PSI-CELP speech coder is that random excitation vectors are given pitch periodicity for voiced speech by a pitch synchronization process. This feature is the origin of the name PSI-CELP. It significantly reduces the distortion and improves the quality for voiced speech, especially for female voices. In addition to the pitch synchronization, a two-channel conjugate structure and fixed codebook for transient speech signals were newly developed. These two features reduce the memory requirement, enhance robustness against channel errors, and improve the quality for the transient part.

5. FUTURE PROSPECTS OF SPEECH TECHNOLOGY

5.1 Ultimate speech recognition and synthesis systems

Ultimate speech recognition systems should be capable of robust, speaker-independent or speaker-adaptive, continuous speech recognition [5]. Speaker recognition techniques are expected to be widely used in the future as methods of verifying the claimed identity in telephone banking and shopping services, information retrieval services, remote access to computers, credit-card calls, etc.

Ultimate speech synthesis and recognition systems that are really useful and comfortable for users should match or exceed human capabilities. That is, they should be faster, more accurate, more intelligent, more knowledgeable, less expensive, and easier to communicate with than human staff. For this purpose, the ultimate systems must be able to handle conceptual

RECENT ADVANCES IN SPEECH PROCESSING TECHNOLOGY

Sadaoki Furui

NTT Human Interface Laboratories
3-9-11, Midori-cho, Musashino-shi, Tokyo, 180 Japan
and
Tokyo Institute of Technology
2-12-1, Ookayama, Meguro-ku, Tokyo, 152 Japan
furui@splab. hil.ntt.jp, furui@cs.titech.ac.jp

SUMMARY

This talk reviews recent speech processing technologies, including speech recognition, synthesis and coding, and discusses current capabilities and the associated applications. For the majority of mankind, speech production and understanding are quite natural and unconsciously acquired processes performed quickly and effectively throughout our daily lives. Speech recognition, synthesis, and coding systems are expected to play important roles in advanced multi-media society with user-friendly human-machine interfaces. Speech recognition systems include not only those that recognize messages but also those that recognize the identity of the speaker. Services using these systems will include voice dialing, database access and management, guidance and transactions, automated reservations, various order-made services, dictation and editing, electronic secretarial assistance, robots, automated interpreting (translating) telephony, security control, digital cellular communication and aids for the handicapped (e.g., reading aids for the blind and speaking aids for the vocally handicapped). This talk also describes the most important research problems that should be solved for accomplishing ultimate recognition, synthesis, and coding systems. The problems include dynamic spectral features, robust speech/speaker recognition, adaptation/normalization, language modeling, discriminative approach, text processing, signal processing and speaking style control in synthesis, synthesis from concepts, use of articulatory and perceptual constraints, and evaluation methods. Although speech recognition, synthesis and coding research have thus far been done independently for the most part, they will encounter increased interaction until commonly shared problems are being investigated and solved simultaneously. Finally the talk tries to forecast where we see progress being achieved in the near future and what applications will become commonplace as a result of the increased capabilities.

OUTLINE

BIOGRAPHY

Sadaoki Furui received B.S., M.S., and Ph.D. degrees in mathematical engineering and instrumentation physics from Tokyo University, Tokyo, Japan in 1968, 1970, and 1978, respectively.

After joining the Electrical Communications Laboratories of Nippon Telegraph and Telephone Corporation in 1970, he studied the analysis of speaker characterizing information in speech waves and its application to speaker recognition as well as interspeaker normalization and adaptation in speech recognition. He also studied vector-quantization-based speech recognition algorithms, spectral dynamic features for speech recognition, speech recognition algorithms that are robust against noise and distortion, and analysis of the speech perception mechanism. He has authored or coauthored over 280 published articles.

He is currently a Research Fellow and the Director of the Furui Research Laboratory at NTT Human Interface Laboratories. He is also a Visiting Professor at the Graduate School of Information Science and Engineering, Tokyo Institute of Technology, and an Instructor at Tokyo University and Chuo University.

From December 1978 to December 1979 he was with the Staff of the Acoustics Research Department at Bell Laboratories, Murray Hill, New Jersey, as a visiting researcher working on speaker verification.

Dr. Furui is a Fellow of the IEEE. He is currently a Vice President of the Acoustical Society of Japan. He served on the IEEE Technical Committee on Speech, and the Technical Program Committees of ICASSP86 in Tokyo as well as ICSLP90 in Kobe. He served on ICSLP94 in Yokohama as Vice Chairman of the Conference Committee. He also serves on several international advisory boards in the US and Europe. He is an Advisory Committee member of the Journal of Speech Communication and an Editorial Board member of the journal Computer Speech and Language.

He received the Yonezawa Prize and the Paper Award from the Institute of Electronics, Information and Communication Engineers of Japan (IEICE) in 1975, 1988 and 1993, and the Sato Paper Awards from the Acoustical Society of Japan (ASJ) in 1985 and 1987. He received the Senior Award from the IEEE ASSP Society and the Achievement Award from the Minister of Science and Technology, both in 1989. He also received the Book Award from the IEICE in 1990.

He is the author of "Digital Speech Processing, Synthesis, and Recognition" (Marcel Dekker, 1989) in English, "Digital Speech Processing" (Tokai University Press, 1985) in Japanese, and "Acoustics and Speech Processing" (Kindai-Kagaku-Sha, 1992) in Japanese. He edited "Advances in Speech Signal Processing" (Marcel Dekker, 1992) jointly with Dr. M.M. Sondhi. He translated "Fundamentals of Speech Recognition" authored by Drs. L.R. Rabiner and B.-H. Juang into Japanese (NTT Advanced Technology, 1995).