論文 / 著書情報
Article / Book Information

Title	Overview of Japanese Research and Development in Speech Information Processing Technology
Author	Sadaoki Furui
Journal/Book name	Speech Technology, , , pp. 18-21
Issue date	1992,

Overview of Japanese Research and Development in Speech Information Processing Technology

Speech research in Japan has had a long and active history throughout which new ideas and products have been created through cooperation among the various sectors of the speech community.

Sadaoki Furui
Furui Research Lab.
NTT Human Interface Labs.
Tokyo, Japan

Speech Analysis and Modeling

Speech research in Japan has a long history. In 1942, for example, Chiba and Kajiyama [1] reported the results of using X-ray pictures of the vocal tract to reveal the transmission characteristics of vowel utterances. However, most of the speech information processing technologies that have contributed to the progress of speech recognition, speech synthesis, and speech coding have been proposed within the last 25 years [2].

Two of the most important papers of this period were presented at the 6th International Congress on Acoustics held in Tokyo in 1968. The paper NTT Labs. presented described a speech analysis-synthesis system based on the maximum likelihood method [3], and the paper Bell Laboratories presented described predictive coding [4]. These papers focused on progress in speech information processing technology and introduced linear predictive coding (LPC), which has led to the creation of a new academic field.

Various research activities in the fields of speech production and perception—the most basic areas of speech signal processing—have been conducted in Japan. One of the greatest successes in these areas, achieved in cooperation between Japanese and American researchers, was the proposal of the two-mass model in which a vocal cord is modeled as two connected parts [5]. This model accurately expresses the actual vibration of human vocal cords.

Speech and Speaker Recognition

In acoustic processing, various spectral distance measures (mainly LPC-based measures) have been investigated [6]. The advantages of the cepstral distance measure have been recognized since 1970. This measure has been used in both speech recognition and speaker recognition [7]. Later, the delta-cepstral coefficient (defined as regression coefficients for the time series of cepstral coefficients) was combined with the instantaneous cepstral coefficient [8][9]. The combination halved recognition error rates in most recognition tasks. This success reflects the importance of dynamic (transitional) features in speech and speaker recognition.

One of the most important advances in speech recognition came with the proposal of the DTW or DP matching approach [10]. The DTW process nonlinearly expands or contracts the time axis to match phoneme positions between the input speech and reference templates. The idea of this method has greatly influenced many of the recognition algorithms, including the hidden Markov model method. Recently, neural network based recognition methods have also been widely investigated.

Speaker-dependent spectral variation is one of the most important problems in speech recognition, and various kinds of speaker adaptation algorithms (such as the VQ codebook mapping method) have been proposed to cope with this variation [11]. In these methods, codebook mapping rules are used to estimate speaker adaptive parameters from speaker-independent parameters. These methods can be classified either as supervised (text-dependent) methods, in which training words or sentences are known [12], or as unsupervised (text-independent) methods, in which arbitrary utterances can be used [13, 14].

Various techniques have been used in linguistic processing. These techniques include those originally proposed for Western languages and those specially designed for Japanese. A system at NTT Labs. is based on the combination of phonetic

hidden Markov models and a two-level grammar composed of an intra-phrase transition network grammar and an inter-phrase dependency grammar [15]. On the other hand, the HMM-LR system at ATR [16] uses generalized LR parsing, which quickly predicts phones according to a context-free grammar.

Speaker recognition algorithms for text-dependent [8] and text-independent [17] cases have been investigated for many years. In the former case, the time function of the spectral parameters for input speech is brought into time registration with the stored reference functions, whereas in the latter case, a speaker specific VQ codebook is made for each reference speaker. These codebooks vector quantize each short period of input speech, and quantization distortion is accumulated over the length of the input. The reference speaker whose codebook produces the smallest accumulated distortion is selected as the speaker. A method using hidden Markov models has also been investigated.

Speech Synthesis

In Japanese text-to-speech methods, special features exist in the text analysis and acoustic processing stages. In the former stage, to cope with difficulties such as the lack of spacing between words and the wide variety of symbols used (Chinese characters and two Japanese syllabaries), various techniques were tested [18]. In the present stage, synthesis methods based on the concatenation of phoneme or syllable units [19] are more frequently used for Japanese than for Western languages.

The Automatic Answer Network System for Electrical Request (ANSER) [20, 21], developed in 1981 by NTT Labs., is now widely used for on-line banking services via telephone. Customers can use this system to make inquiries and receive answers through dialogue with a computer. ANSER, used at 95% of the financial institutions in Japan (about 500 banks), is the world's largest system using speech recognizers and synthesizers. The average monthly traffic is about 16 million calls.

Text-to-speech systems in supporting systems for newspaper proofreading and revision have also been successfully used since 1987. Text-to-speech systems are also used in services providing information over telephone networks, in verbal collation for word processors, and for assisting the visually impaired.

Speech Coding

In speech coding, several new algorithms have been proposed, including the APC-AB method for speech coding at approximately 16 kbps [22]. "Full-rate" standardization for digital cellular radio at 11.2 kbps including error correction codes has already been settled, and research on "half-rate" coding (5.6 kbps and below) is under way at many laboratories [23]. To cope with the impairment of the transmission system by coding delay, several laboratories are also investigating low-delay speech coding at 8 kbps.

Sites and Cooperation

Speech research in Japan is being pursued at a very large number of universities, private companies, and public laboratories. Major universities include Kyoto Univ., Kyoto Institute of Technology, Kyushu Univ., Nagoya Univ., Nagoya Institute of Technology, Osaka Univ., Shinshu Univ., Shizuoka Univ., Tohoku Univ., Tokyo Univ., Tokyo Institute of Technology, Toyohashi Univ. of Technology, Waseda Univ., Yamagata Univ., and Yamanashi Univ. Major private laboratories include ATR, Cannon, Fujitsu, Hitachi, KDD, Matsushita, Mitsubishi, NEC, NHK, NTT (Nippon Telegraph and Telephone), NTT-Data, Rikoh, Sanyo, Secom, Sharp, and Toshiba. Public laboratories include ETL and CRL. The largest of these speech research groups are NTT and ATR Labs.

A nationwide project involving cooperation between academic groups was supported by the Ministry of Education, Science, and Culture by a Grant-in-Aid for Scientific Research on Priority Areas. From 1987 to 1990, Professor Fujisaki organized and led the project [24], entitled "Advanced Man-Machine Interface through Spoken Language." This cooperative effort improved research facilities at universities and created opportunities for the mutual exchange of ideas, not only within academic groups, but also between these groups and the people from company laboratories.

Presentation and Publication

Results of speech research and development in Japan are presented at domestic and international conferences. The major domestic conference is the Meeting of the Acoustical Society of Japan (ASJ), which is held in spring and fall each year. Roughly 200 speech-related papers are presented at every meeting, and a two-page summary of each presentation appears in the proceedings of the meeting. The Institute of Electronics, Information, and Communication Engineers (IEICE) also has spring and fall meetings where fewer speech papers are presented.

Full papers are presented at the Meeting of the Speech Technology Committee, organized by the ASJ and IEICE and held almost every month. Roughly 150 papers per year are presented at these meetings. Workshops on special topics such as "HMM and neural network technologies" and "dialogue analysis and processing" are also held periodically. Full papers are published in the journals of the IEICE and ASJ, and both these journals have Japanese and English transactions. Papers are also published in American and European journals like

IEEE Transactions on Signal Processing, the Journal of the Acoustical Society of America, and Speech Communication.

Several international conferences on speech research have been held in Japan. These include the International Congress on Acoustics in 1968; the IEEE International Conference on Acoustics, Speech, and Signal Processing in 1986; and the International Conference on Spoken Language Processing in 1990. ASJ and ASA Joint Meetings were held in Hawaii in both 1978 and 1988. The ASJ Acoustical Society of Korea Joint Workshop on Speech Recognition was held in Seoul, Korea in 1991.

REFERENCES

1. Chiba, T. and Kajiyama, M., "The vowel its nature and structure," Tokyo-Kaiseikan (1942).

2. Furui, S., "Digital speech processing, synthesis, and recognition," Marcel Dekker (1989).

3. Itakura, F. and Saito, S., "Analysis synthesis telephony based on the maximum likelihood method," Proc. 6th Int. Cong. Acoust., C-5-5 (1968).

4. Atal, B. S. and Schroeder, M. R., "Predictive coding of speech signals," Proc. 6th Int. Cong. Acoust., C-5-4 (1968).

5. Ishizaka, K. and Flanagan, J. L., "Synthesis of voiced sounds from a two-mass model of the vocal cords," Bell Syst. Tech. J., 51, 6, pp. 1233-1268 (1972).

6. Shikano, K. and Itakura, F., "Spectrum distance measures for speech recognition," in Furui, S. and Sondhi, M. M. (Ed.) "Advances in speech signal processing", Marcel Dekker, pp. 419-452 (1991).

7. Furui, S., "Talker recognition by the longtime averaged speech spectrum," Trans. IECE, 55-A, 10, pp. 549-556 (1972).

8. Furui, S., "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust., Speech, Signal Processing, ASSP-29, 2, pp. 254-272 (1981).

9. Furui, S., "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust., Speech, Signal Processing, ASSP-34, 1, pp. 52-59 (1986).

10. Sakoe, H. and Chiba, S., "Recognition of continuously spoken words based on time-normalization by dynamic programming," J. Acoust. Soc. Jap., 27, 9, pp. 483-500 (1971).

11. Furui, S., "Speaker-independent and speaker-adaptive recognition techniques," in Furui, S. and Sondhi, M. M. (Ed.) "Advances in speech signal processing," Marcel Dekker, pp. 597-622 (1991).

12. Shikano, K., Lee, K.F., and Reddy, R., "Speaker adaptation through vector quantization," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo, Japan, 49.5, pp. 2643-2646 (1986).

13. Shiraki, Y. and Honda, M., "Speaker adaptation algorithms based on piecewise woving adaptive segment quantization method," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Albuquerque, S12.5 (1990).

14. Matsumoto, H. and Yamashita, Y., "Unsupervised speaker adaptation of spectra based on minimum fuzzy vector quantization error criterion," J. Acoust. Soc. Amer., Suppl. 1, 84, p. S15 (1988).

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

16. Hanazawa, T., Kita, K., Nakamura, S., Kawabata, T., and Shikano, K., "ATR HMM-LR continuous speech recognition system," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Albuquerque, S2.4 (1990).

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

18. Sagisaka, Y., "Speech synthesis from text," IEEE Comm. Magazine, January, pp. 35-41 (1990).

19. Sato, H., "Speech synthesis for text-to-speech systems," in Furui, S. and Sondhi, M.M. (Ed.) "Advances in speech signal processing," Marcel Dekker, pp. 833-853 (1991).

20. Nakatsu, R., "ANSER: An application of speech technology to the Japanese banking industry," IEEE Computer, 23, 8, pp. 43-48 (1990).

21. Sato, H., "Japanese text-to-speech equipment: Current applications and trends," Proc. 1990 Int. Conf. Spoken Language Processing, Kobe, Japan, 20.4 (1990).

22. Honda, M. and Itakura, F., "Bit allocation in time and frequency domains for predictive coding of speech," IEEE Trans. Acoust., Speech, Signal Processing, ASSP-32, 3, pp. 465-473 (1984).

23. Honda, M. and Shiraki, Y., "Very low-bit-rate speech coding", in Furui, S. and Sondhi, M. M. (Ed.) "Advances in speech signal processing," Marcel Dekker, pp. 209-230 (1991).

24. "The second symposium on advanced man-machine interface through spoken language," Makaha, Oahu, Hawaii, (November 1988).

FOR MORE INFORMATION

Contact Sadaoki Furui, Furui Research Lab., NTT Human Interface Labs., 3-9-11, Midiori-cho, Musashino-shi, Tokyo, 180 Japan. (81) 422 59 3910; Fax: (81) 422 60 7808.

SADAOKI FURUI *received his B.S., M.S., and Ph.D in Mathematical Engineering and Instrumentation Physics from Tokyo University, Tokyo, Japan in 1968, 1970, and 1978 respectively. After joining the Electrical Communications Labs. at NTT Corp. in 1970, he studied the analysis of speaker characterization information in the speech wave, its application to speaker recognition, the vector quantization based speech recognition algorithm, spectral dynamic features for speech recognition, and the analysis of the speech perception mechanism. He currently is the Research Fellow and heads the Furui Research Lab. at NTT Human Interface Labs. From 1978 to 1979, he was with the Staff of the Acoustics Research Dept. at AT&T Bell Labs., New Jersey, as a visiting researcher working on speaker verification. He is the author of "Digital Signal Processing, Synthesis, and Recognition" (Marcel Dekker, 1989).*