論文 / 著書情報
Article / Book Information

Title	Future Directions in Speech Information Processing
Author	Sadaoki Furui
Journal/Book name	16th ICA and 135th Meeting ASA, , , pp. 1-4
発行日 / Issue date	1998, 6

Proc. 16th ICA and 135th Meeting ASA, Seattle, pp. 1-4 (1998)

Future Directions in Speech Information Processing

Sadaoki Furui

Dept. Computer Science, Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku, Tokyo, 152 Japan
furui@cs.titech.ac.jp

Abstract: This paper predicts future directions in speech processing technologies, including speech recognition, synthesis and coding. It 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. The latter half of the paper focuses on speech recognition problems. Unified approach over speech recognition, synthesis and coding, and research about the human brain are crucial for the future development.

INTRODUCTION

Speech recognition, synthesis, and coding systems are expected to play important roles in an advanced multimedia society with user-friendly human-machine interfaces [3]. 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).

The most promising application area for speech technology is telecommunications [9]. 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 will become popular, and everybody will have their own portable telephone. Several technologies will play major roles in this communications revolution, but speech processing will be one of the key technologies. By using advancing speech synthesis / recognition technology, telephone sets will become useful personal terminals for communicating with computer systems. 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 and so forth.

GENERAL DIRECTIONS

Ultimate Speech Recognition and Synthesis Systems

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.

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 and so forth.

Unified Approach

Individual speech information processing technologies, such as speech recognition, speaker recognition, speech synthesis, and speech coding, have so far been investigated independently by separate approaches. For example, research on voice individuality has been examined differently in speaker adaptation for speech recognition, automatic

speaker verification, voice conversion in speech synthesis, and problems of very-low-bit-rate speech coding quality variation from speaker to speaker. However, handling all these issues deals with extracting common phenomena in voice individuality, essentially solving the same problem from different aspects. These approaches should therefore be based on a common mathematical model. They are now, however, based on completely different models, most of them nor reflecting the nature of speech. From this point of view, we should seek to pursue a unified approach which can be universally applied to different speech information processing areas by approaching real nature of speech and essential solutions of the problems.

Statistical Approaches and Speech Science

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 mismatch between the training database and the testing data. Therefore, even the statistical approaches must be based on reasonable models which can only be created by observing actual phenomena with our knowledge of speech science.

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 the problems 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.

Articulatory and Perceptual Constraints

Knowledge and technology from a wide range of areas, including the use of articulatory and perceptual constraints, will be necessary to develop speech technology. For example, when several phonemes or syllables are continuously spoken, as in the case of usual sentence speech, the tongue, jaw, lips, etc. move asynchronously in parallel, and yet with coupled relationships. Current speech analysis techniques, however, represent speech as a simple time series of spectra. It will become necessary to analyze speech by decomposing it into several hidden factors based on speech production mechanisms. This approach seems to be essential for solving the coarticulation problem, one of the most important problems in both speech synthesis and recognition.

The human hearing system is far more robust than machine systems - more robust not only against the direct influence of additive noise but also against speech variations (that is, the indirect influence of noise), even if the noise is very inconsistent. Speech recognizers are therefore expected to become more robust when the front end uses models of human hearing. This can be done by imitating the physiological organs or by reproducing psychoacoustic characteristics.

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 [3].

Research on the Human Brain

Although observation and modeling of the movement of vocal systems along with the physiological modeling of auditory peripheral systems have recently made great progresses, the mechanism of speech information processing in our human brain has hardly been investigated. Psychological experiments on human memory clearly showed that speech plays a far more important and essential role than vision in the human memory and thinking processes. Whereas models of separating acoustic sources have been researched in "auditory scene analysis", the mechanisms of how meanings of speech are understood and how speech is produced have not yet been made clear.

It will be necessary to clarify the process by which human beings understand and produce spoken language, in order to obtain hints for constructing language models for spoken language, which is very different from written language. It is necessary to be able to analyze context and accept ungrammatical sentences. It is about time to start active research on clarifying the mechanism of speech information processing in the human brain so that epoch-making technological progress can be made based on the human model.

SPEECH AND SPEAKER RECOGNITION

Overview

A series of (D)ARPA projects have been a major driving force of the recent progress in research on large-vocabulary, continuous-speech recognition. Specifically, dictation of speech reading newspapers such as north America business newspapers including Wall Street Journal, conversational speech recognition using an ATIS task, and, recently, broadcast news dictation have been actively investigated [10]. Common features of these systems exist in using cepstral parameters and their regression coefficients as speech features, triphone HMMs as acoustic models, vocabularies of several thousand or several ten thousand entries, and statistical language models such as bigrams and trigrams [8]. Such methods have been applied not only to English but also to French, German, Italian, and Japanese, and, although there are several language-specific characteristics, similar recognition results have been obtained. Recently, Switchboard and Call Home tasks using natural conversational speech have been actively investigated. In spite of the remarkable recent progress, we are still far behind our ultimate goal of understanding free conversational speech uttered by any speaker under any environment.

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 length of the time windows in which sound transitions are perceived have a hierarchical structure and range from the order of several milliseconds to several seconds. The hierarchical layers correspond to various speech features, such as phonemes, syllables, and prosodic features. It has also been reported that the human hearing mechanism perceives a target value estimated from the transitional information extracted using dynamic spectral features.

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.

Robust Speech Recognition

Ultimate speech recognition systems should be capable of robust, speaker-independent or speaker-adaptive, continuous 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 [5]. 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.

Extraction and normalization of (adaptation to) voice individuality is one of the most important issues [2]. 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 adaptation (normalization) methods can usually be classified into supervised (text-dependent) and unsupervised (text-independent) methods. Unsupervised, on-line, incremental adaptation is ideal, since the system works as if it were a speaker-independent system, and it performs increasingly better as it is used.

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 [6] is also a very important issue, since collecting a large linguistic database for every new task is difficult and costly.

Detection-Based Approach for Spontaneous Speech Recognition

One of the most important issues for speech recognition is how to create language models (rules) for spontaneous speech. 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, disfluency, partial words, repairs, hesitations, and repetitions. It is crucial to develop robust and flexible parsing algorithms that match the characteristics of spontaneous speech. A paradigm shift from the present transcription-based approach to a detection-based approach will be important to solve such problems. How to extract contextual information, predict users' responses, and focus on key words are very important issues.

Style shifting is also an important problem in spontaneous speech recognition. In typical laboratory experiments, speakers are reading lists of words rather than trying to accomplish a real task. Users actually trying to accomplish a task, however, use a different linguistic style.

Speaker Recognition

Recently, various topics of research interest in speaker recognition have led to new approaches and techniques. They include VQ- and ergodic-HMM-based text-independent recognition methods, a text-prompted recognition method [7], parameter / distance normalization techniques, model adaptation techniques, and methods of updating models as well as *a priori* thresholds in speaker verification. However, there are still many problems for which good solutions remain to be found. The open questions include [4] (a) How can human beings correctly recognize speakers? (b) What feature parameters are appropriate for speaker recognition? (c) How can we fully exploit the clearly evident encoding of identity in prosody and other suprasegmental features of speech? (d) Is the "sheep and goats" problem (a small percentage of speakers account for the majority of errors) universal? (e) Can we ever reliably cluster speakers on the basis of similarity / dissimilarity? (f) How do we deal with long-term variability in people's voices?

CONCLUSION

This paper discussed the most important research problems to be solved in order to achieve ultimate speech recognition, synthesis, and coding systems. Speech recognition problems include dynamic spectral features, robustness against voice variations, adaptation / normalization techniques, language modeling, use of articulatory and perceptual constraints, and detection-based approach for spontaneous speech recognition.

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. The necessity for research about the human brain will also increase in order to solve various fundamental problems in speech recognition and synthesis.

REFERENCES

1. Furui, S., "On the role of spectral transition for speech perception", *J. Acoust. Soc. Am.*, **80**, 4, 1016-1025 (1986)
2. Furui, S., "Speaker-independent and speaker-adaptive recognition techniques", in *Advances in Speech Signal Processing*, ed. by Furui, S. and Sondhi, M. M., New York, Marcel Dekker, Inc., 1992, pp. 597-622
3. Furui, S., "Towards the Ultimate Synthesis/Recognition System" in *Voice Communication between Humans and Machines*, ed by Roe, D. B. & Wilpon, J. G., Washington D. C., National Academy Press, 1994, pp. 450-466
4. Furui, S., "Recent advances in speaker recognition", *Proc. First International Conference on Audio- and Video-based Biometric Person Authentication*, Crans-Montana, Switzerland, 237-252 (1997)
5. Furui, S., "Recent advances in robust speech recognition", *Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, 11-20 (1997)
6. Kuhn, R. and DeMori, R., "A cache-based natural language model for speech recognition", *IEEE Trans. Pattern Anal., Machine Intelligence*, **PAMI-12**, 6, 570-583 (1990)
7. Matsui, T. and Furui, S., "Concatenated phoneme models for text-variable speaker recognition", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Minneapolis, II-391-394 (1993)
8. Rabiner, L. R. and Juang, B. H., *Fundamentals of Speech Recognition*, New Jersey, Prentice-Hall, Inc., 1993
9. Rabiner, L. R., "The role of voice processing in telecommunications", *Proc. 2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, Kyoto, Japan, 1-8 (1994)
10. *Proceedings of the DARPA Speech Recognition Workshop*, San Francisco, Morgan Kaufmann Publishers, 1997