

From Read Speech Recognition to Spontaneous Speech Understanding

Sadaoki Furui

Tokyo Institute of Technology, Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
furui@cs.titech.ac.jp

Keywords: speech recognition, understanding, spontaneous speech, corpus, speech science

Abstract

This paper predicts future directions in speech recognition technologies. It describes the most important research problems, and tries to forecast where progress will be made in the near future. One of the most critical issues is how to build language models for spontaneous speech. It is indispensable to build a large corpus of spontaneous speech, since our knowledge of the structure of spontaneous speech is very limited. Recognition of spontaneous speech will require a paradigm shift from speech recognition to understanding where underlying messages of the speaker are extracted, instead of transcribing all the spoken words. It is crucial to analyze the function within the human brain and the function must then be realized using engineering models. Research should investigate how meanings conveyed by speech are understood.

1. Introduction

Speech recognition systems are expected to play important roles in an advanced multi-media society with user-friendly human-machine interfaces [1]. The field of automatic speech recognition has witnessed a number of significant advances in the past 5-10 years, spurred on by advances in signal processing, algorithms, computational architectures, and hardware. These advances include the widespread adoption of a statistical pattern recognition paradigm, a data-driven approach which makes use of a rich set of speech utterances from a large population of speakers, the use of stochastic acoustic and language modeling, and the use of dynamic programming-based search methods [2][3].

Read speech and similar types of speech, e.g. that from reading newspapers or from news

broadcast, can be recognized with accuracy higher than 90% using the state-of-the-art speech recognition technology. However, recognition accuracy drastically decreases for spontaneous speech. This decrease is due to the fact that the acoustic and linguistic models used have generally been built using written language or speech from written language. Unfortunately spontaneous speech and speech from written language are very different both acoustically and linguistically. In order to increase the recognition performance for spontaneous speech, it is crucial to build language models for spontaneous speech. Methods applying statistical language modeling such as bigrams and trigrams of words or morphemes to spontaneous speech corpus may not be adequate. Our knowledge of the structure of spontaneous speech is currently inadequate to achieve the necessary breakthroughs. Recognition of spontaneous speech will require a paradigm shift from speech recognition to understanding where underlying messages of the speaker are extracted, instead of transcribing all the spoken words.

Up to the present, the fields of speech perception and automatic speech recognition have been widely separate. However, in order to build spontaneous speech understanding systems, it is crucial to analyze the function within the human brain. The function must then be realized using engineering models. For these purposes, human speech perception research needs to shift from targeting short fragments such as phonemes and syllables to larger units such as words, phrases, sentences and paragraphs. Research should investigate how meanings conveyed by speech are understood. It is indispensable to build a large corpus of spontaneous speech and conduct corpus-based research on both the mechanisms of human speech perception and the engineering speech understanding systems with close connection and cooperation.

Ultimately speech 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.

2. Spontaneous Speech Recognition and Understanding

2.1 State-of-the-art speech recognition

Figure 1 shows the mechanism of state-of-the-art speech recognizers [4]. 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. 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, 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/spontaneous speech uttered by any speaker in any environment.

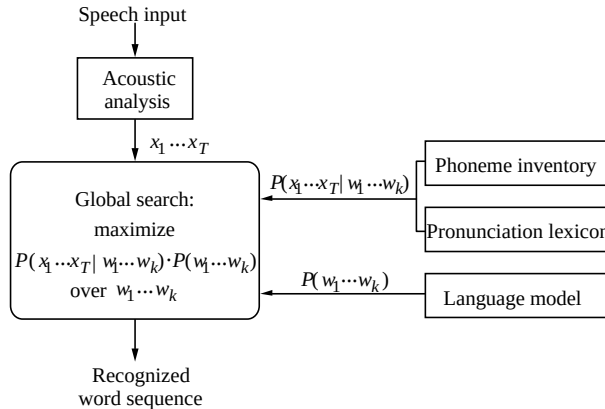


Fig. 1 - Mechanism of state-of-the-art speech recognizers.

2.2 Language modeling for spontaneous speech recognition

One of the most important present 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. 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. It is crucial to develop robust and flexible decoding 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 the spontaneous-speech specific problems [5]. 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 sentences rather than trying to accomplish a real task. Users actually trying to accomplish a task, however, use a different linguistic style. Adaptation of linguistic models according to tasks and topics is also a very important issue, since collecting a large linguistic database for every new task is difficult and costly.

2.3 Message-driven speech recognition and understanding

As shown in Fig. 1, state-of-the-art automatic speech recognition systems employ the criterion of maximizing $P(W|X)$, where $W=w_1, \dots, w_k$ is a word sequence, and $X=x_1, \dots, x_T$ is an acoustic observation sequence. This criterion is reasonable for dictating read speech.

However, the ultimate goal of automatic speech recognition is to extract the underlying messages of the speaker from the speech signals. Hence we need to model the process of speech generation and recognition as shown in Fig. 2 [6], where M is the message (content) that a speaker intended to convey. There is also a possibility to give feedback from the "understanding module" to the speech recognition module such that decoding hypotheses can be properly adjusted and, hopefully, converge to the most correct word sequence as well as the most correct understanding of the utterance.

According to this model, the speech recognition process is represented as the maximization of the following a posteriori probability [6]:

$$\max_M P(M | X) = \max_M \sum_W P(M | X) P(W | X). \quad (1)$$

Using Bayes' rule, Eq. (1) can be expressed as

$$\max_M P(M | X) = \max_M \sum_W \frac{P(X | W)P(W | M)P(M)}{P(X)}. \quad (2)$$

For simplicity, we can approximate the equation as

$$\max_M P(M | X) = \max_{M,W} \sum_W \frac{P(X | W)P(W | M)P(M)}{P(X)}. \quad (3)$$

$P(X|W)$ is calculated using hidden Markov models in the same way as in usual recognition processes. This new formulation of speech recognition was applied to the Japanese broadcast news transcription, and it was found that word error rates were slightly reduced by this method.

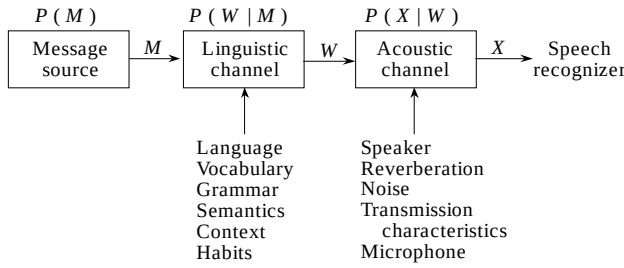


Fig. 2 - A communication-theoretic view of speech generation and recognition.

2.4 Speech summarization

Transcribed spontaneous speech usually includes various kinds of redundant information. Therefore, a summarization technique to compress the information is expected to be useful in many applications, such as for closed captioning, indexing speech data for automatic retrieval, making abstracts of presentations, minutes of meetings and voicemails, and presenting information in news-on-demand systems.

We have proposed a method of automatically summarizing speech, sentence by sentence, by extracting a limited number of relatively important words from an automatic transcription according to a target compression ratio of the number of characters [7][8]. A word set that maximizes a summarization score consisting of a significance (topic) score I and a confidence score C of each word, a linguistic score (likelihood) L of word concatenation, and a word modification (concatenation) score T_r is extracted.

Given a transcription result consisting of N words, $W=w_1, w_2, \dots, w_N$, the summarization is performed by extracting a set of M ($M < N$) words,

$V=v_1, v_2, \dots, v_M$, which maximizes the summarization score given by eq.(1).

$$S(V) = \sum_{m=1}^M \{L(v_m | \dots v_{m-1}) + \lambda_I I(v_m) + \lambda_C C(v_m) + \lambda_T T_r(v_{m-1}, v_m)\} \quad (4)$$

where λ_I , λ_C and λ_T are weighting factors for balancing among L , I , C and T_r .

The significance score is calculated for nouns based on the amount of information conveyed by each word, a flat score being given to words other than nouns. A posterior probability of each transcribed word, i.e. the ratio of the word hypothesis probability to that of all other hypotheses, is calculated using a word graph obtained by the decoder and used as a confidence measure. The confidence score is incorporated to give a penalty for acoustically as well as linguistically unreliable hypotheses. A trigram probability is used as a linguistic score. The word modification score is determined by a dependency structure obtained using Stochastic Dependency Context Free Grammar (SDCFG).

A set of words maximizing the summarization score is efficiently selected using a dynamic programming (DP) technique. In order to evaluate the summarization scores of various summarized sentences with different summarizing ratios obtained from the same original sentence, the summarization score is normalized according to the summarization ratio. Japanese broadcast news speech transcribed using our large vocabulary continuous speech recognition system was summarized. Experimental results show that all the scores constituting the summarization score are effective to make the summarized sentences readable and meaningful. The summarization method is being now applied to presentation speech and extended to summarize articles consisting of multiple sentences.

2.5 Japanese national project on spontaneous speech corpus and processing technology

For building language models for spontaneous speech, we need to have large spontaneous speech corpora. In this context, we started a Science and Technology Agency Priority Program (Organized Research Combination System) entitled "Spontaneous Speech: Corpus and Processing Technology" in 1999 supervised by S. Furui [9]. The principal organizations working together to conduct this project are the National Language Research Institute under the Ministry of Education, the Communication Research Laboratory under the

Ministry of Posts and Telecommunications, and Tokyo Institute of Technology.

The project will be conducted over a 5-year period under the following three major themes as shown in Fig. 3.

- 1) Building a large-scale spontaneous speech corpus consisting of roughly 7M words with the total speech length of 800 hours. Mainly recorded will be monologues such as lectures, presentations, and news commentaries. They will be manually given orthographic and phonetic transcription. One-tenth of the utterances ("Core") will be tagged manually and used for making a morphological analysis program for automatically analyzing all of the 800-hour utterances. The Core will also be tagged with para-linguistic information including intonation.
- 2) Acoustic and linguistic modeling for spontaneous speech understanding using linguistic as well as para-linguistic information in speech.
- 3) Investigating spontaneous speech summarization technology.

The technology created in this project is expected to be applicable to wide areas such as indexing of speech data (broadcast news, etc.) for information extraction and retrieval, transcription of lectures, preparing minutes of meetings, closed captioning, and aids for the handicapped.

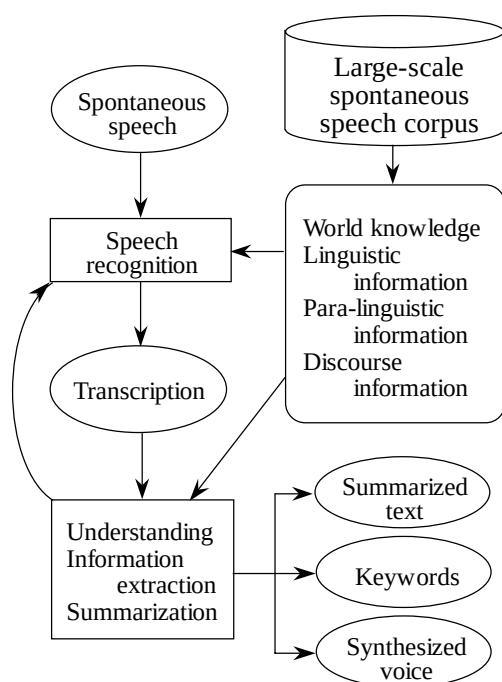


Fig. 3 - Overview of the Japanese national project on spontaneous speech corpus and processing technology.

3. Speech Science and Engineering

3.1 Statistical approaches and speech science

There is no doubt that most recent progress in speech and speaker recognition comes from statistical approaches, such as HMMs 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, especially for spontaneous speech. 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, including 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.

3.2 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 speech 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 utilizes 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 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 [10].

3.3 Research on the human brain

Figure 4 shows a diagram of human speech generation and perception process. 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 generated 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 and utilize contextual information to handle anaphora and ellipsis frequently used in human dialogues. It is 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.

3.4 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 [11][12]. 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 [13]. 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 at least for each basic

short period. 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.

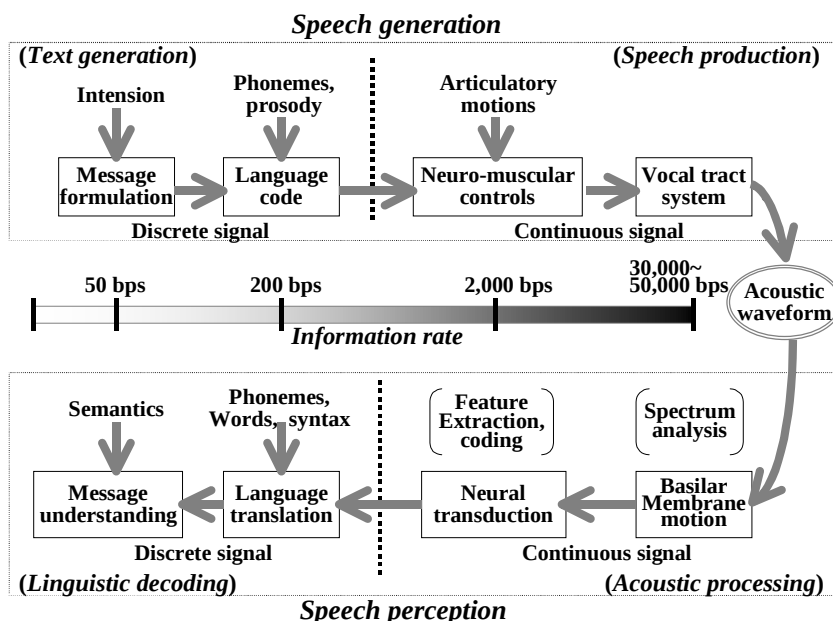


Fig. 4- Alternative view of speech-generation/speech-perception process.

3.5 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, etc. [14]. 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 [15]. A small percentage of people occasionally cause systems to produce exceptionally low recognition rates. This is an example of the “sheep and goats” phenomenon [3]. Speaker adaptation (normalization) methods can usually be classified into supervised (text-dependent) and unsupervised (text-independent) methods. Unsupervised, on-line, instantaneous/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. However, since we have to adapt many phonemes using a limited size of utterances including only a limited number of phonemes, it is crucial to use reasonable modeling of speaker-to-speaker variability or constraints. Modeling of the mechanism of speech production (e.g. the eigenvoice approach [16]) is expected to provide a useful representation of speaker-to-speaker variability. Humans have been found to have a mechanism of unsupervised, on-line speaker adaptation [17].

4. Conclusion

This paper discussed the most important research problems to be solved in order to achieve ultimate speech recognition systems. The problems include language modeling for spontaneous speech, message-driven speech recognition and understanding, speech summarization, spontaneous speech corpus building, use of articulatory and perceptual constraints, dynamic spectral features, and robustness against voice variations. A paradigm shift from speech recognition to understanding, where underlying messages of the speaker, that is, meaning/content that the speaker intended to convey, are extracted, instead of transcribing all the spoken words, will be indispensable. The necessity for research about the human brain will increase in order to solve various fundamental problems.

References

- [1] B.-H. Juang and S. Furui: “Automatic recognition and understanding of spoken language – A first step towards natural human-machine communication”, *Proc. IEEE*, 88, 8, pp. 1142-1165, 2000
- [2] Rabiner, L. R. and Juang, B. H., *Fundamentals of Speech Recognition*, New Jersey, Prentice-Hall, Inc., 1993
- [3] Furui, S., *Digital Speech Processing, Synthesis, and Recognition*, 2nd Edition, New York, Marcel Dekker, 2000
- [4] Ney, H., “Corpus-based statistical methods in speech and language processing”, in *Corpus-based Methods in Language and Speech Processing*, Young, S. and Bloothoof, G. Ed., pp. 1-26, 1997
- [5] Juang, B.-H., “From speech recognition to understanding: Shifting paradigm to achieve natural human-machine communication,” *Proc. 16th ICA and 135th Meeting ASA*, pp. 617-618, 1998
- [6] Ohtsuki, K., Furui, S., Iwasaki, A. and Sakurai, N., “Message-driven speech recognition and topic-word extraction”, *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Phoenix, pp. 625-628, 1999
- [7] Hori, C. and Furui, S., “Automatic speech summarization based on word significance and linguistic likelihood”, *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Istanbul, pp. 1579-1582, 2000
- [8] Hori, C. and Furui, S., “Improvements in automatic speech summarization and evaluation methods”, *Proc. Int. Conf. Spoken Language Processing*, Beijing, pp. IV-326-329, 2000
- [9] Furui, S., Maekawa, K., Isahara, H., Shinozaki, T. and Ohdaira, T., “Toward the realization of spontaneous speech recognition – Introduction of a Japanese priority program and preliminary results –”, *Proc. Int. Conf. Spoken Language Processing*, Beijing, pp. III-518-521, 2000
- [10] 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
- [11] Furui, S., “On the role of spectral transition for speech perception,” *J. Acoust. Soc. Am.*, **80**, 4, pp. 1016-1025, 1986
- [12] Furui, S. and Akagi, M., “On the role of spectral transition in phoneme perception and its

- modeling,” Proc. 12th ICA, Toronto, Canada, A2-6, 1986
- [13] Furui, S., “Speaker-independent isolated word recognition using dynamic features of speech spectrum,” *IEEE Trans. Acoust., Speech, Signal Process.*, **ASSP-34**, 1, pp. 52-59, 1986
 - [14] 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
 - [15] 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
 - [16] Kuhn, R., et al., “Eigenvoices for speaker adaptation,” *Proc. ICSLP*, Sydney, pp. 1771-1774, 1998
 - [17] Kato, K. and Furui, S., “Listener adaptability for individual voice in speech perception,” *Trans. Committee, Hearing Res.*, Acoust. Soc. Japan, H-85-5 (in Japanese), 1985