INTERSPEECH
2009 BRIGHTON

# Selected topics from 40 years of research on speech and speaker recognition

*Sadaoki Furui* [1]

[1] Department of Computer Science, Tokyo Institute of Technology, Japan
`furui@cs.titech.ac.jp`

## Abstract

This paper summarizes my 40 years of research on speech and speaker recognition, focusing on selected topics that I have investigated at NTT Laboratories, Bell Laboratories and Tokyo Institute of Technology with my colleagues and students. These topics include: the importance of spectral dynamics in speech perception; speaker recognition methods using statistical features, cepstral features, and HMM/GMM; text-prompted speaker recognition; speech recognition using dynamic features; Japanese LVCSR; robust speech recognition; spontaneous speech corpus construction and analysis; spontaneous speech recognition; automatic speech summarization; and WFST-based decoder development and its applications.

**Index Terms**: speech recognition, speaker recognition, speech perception

## 1. Introduction

For almost four decades starting in 1970, I have been working on speech perception, speech analysis, speech synthesis, speech recognition, speaker recognition, multi-modal speech recognition, and question-answering systems, at NTT (Nippon Telegraph and Telephone) Labs and Tokyo Institute of Technology. I also worked at AT&T Bell Laboratories as a visiting researcher from 1978 to 1979, and visited several other research laboratories and universities in the world for short stays. I was very fortunate to have so many excellent colleagues to work with at all of these places. I have enjoyed all these environments and conducted various research with my colleagues and students. Speech processing is a very interesting research area and a wide variety of significant technical progress has been made, but it still has various difficult problems. Figure 1 shows generations of speech and speaker recognition research since 1952 [1].

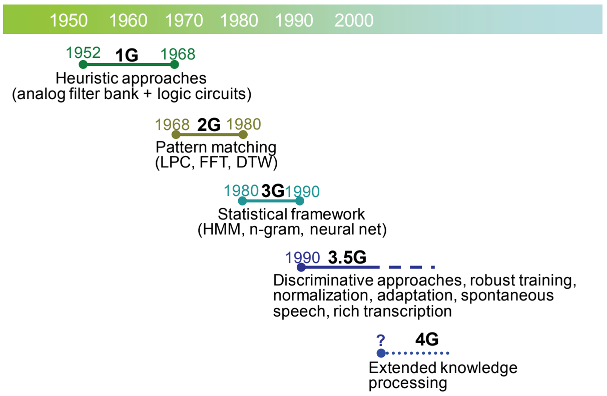

1950 1960 1970 1980 1990 2000
1952 1G 1968
Heuristic approaches
(analog filter bank + logic circuits)
1968 2G 1980
Pattern matching
(LPC, FFT, DTW)
1980 3G 1990
Statistical framework
(HMM, n-gram, neural net)
1990 3.5G
Discriminative approaches, robust training, normalization, adaptation, spontaneous speech, rich transcription
? 4G
Extended knowledge processing

Figure 1: *Four generations of speech and speaker recognition research.*

This paper focuses on speech and speaker recognition, which have been my main research areas, and summarizes several major results that we achieved. The paper concludes with future perspectives.

## 2. 1970s

### 2.1. Speaker recognition by statistical features

We conducted speaker recognition experiments, as one of the pioneers in this domain, using statistical features of speech spectra extracted from sentence and word utterances. The statistical features were extracted from long-term averaged spectrum of a sentence-long utterance [2] and time-averaged characteristics (a mean value, a standard deviation and a correlation matrix) of log-area-ratios and fundamental frequency derived from the voiced portion of spoken words [3, 4]. It was found that one of the most difficult problems in speaker recognition is that inter-session variability (variability over time) for a given speaker has a significant effect on recognition accuracy. Figure 2 shows the amount of spectral variation as a function of time interval between speech samples, measured using either log-area-ratios or long-term averaged spectra.

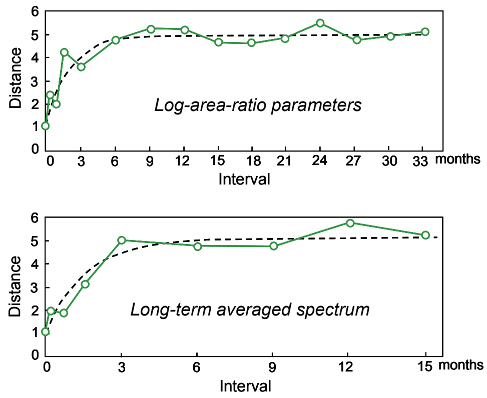
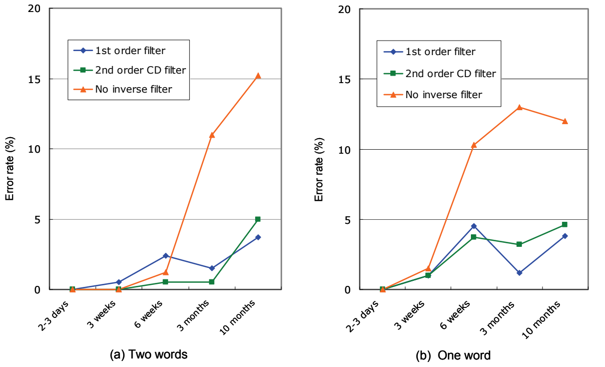


Figure 2: *The amount of spectral variation as a function of time interval.*

To solve this problem, various inverse filtering (spectral equalization) methods were investigated using a database consisting of speech utterances recorded over seven years. It was found that, when spectral equalization was applied to input utterances, high recognition accuracy could be obtained even if the training utterances for each speaker were recorded over a short period and the time difference between training and input utterances was as long as one year. Figure 3 shows the effectiveness of the inverse filtering on speaker recognition using statistical features of LPC parameters extracted from one or two words. From the viewpoint of the speech production mechanism, the effectiveness of inverse filtering means that vocal tract characteristics are much more stable than overall patterns of the vocal cord spectrum.



(a) Two words (b) One word

Figure 3: *Effectiveness of inverse filtering in reducing speaker recognition error rates. The horizontal axis indicates the time difference between training and testing data.*

### 2.2. Speaker recognition by cepstral dynamic features

New techniques for speaker verification using telephone speech were investigated [5]. The operation of the system was based on a set of functions of time obtained from acoustic analysis of a fixed, sentence-long utterance, "We were away a year ago". Figure 4 shows the overall structure of the system. Cepstrum coefficients were extracted by means of LPC analysis successively throughout an utterance to form time functions, and frequency response distortions introduced by transmission systems were removed (cepstral mean subtraction). The time functions were expanded by orthogonal polynomial representations (delta-cepstrum) and, after a feature selection procedure, brought into time registration with stored reference functions to calculate the overall distance. Reference functions and the decision threshold were updated for each customer (registered speaker). Several sets of experimental utterances were recorded over a conventional telephone connection. Utterances processed by ADPCM and LPC coding systems were also used. Verification error rate of one percent or less could be obtained even if the reference and test utterances were subjected to different transmission conditions.

## 3. 1980s

### 3.1. Spectral dynamics in speech perception

The relationship between dynamic spectral features and the identification of Japanese syllables modified by initial and/or final truncation was examined as shown in Figure 5 [6]. The experiments confirmed several main points as shown in Figures 6 and 7: (a) "Perceptual critical points," where the percent correct identification of the truncated syllable as a function of the truncation position changes abruptly ($T_i$, $T_f$), are related to maximum spectral transition positions ($T_m$); (b) A speech wave of approximately 10ms in duration that includes the maximum spectral transition position bears the most important information for consonant and syllable perception; (c) Consonant and vowel identification scores simultaneously change as a function of the truncation position in the short period, including the 10ms period for final truncation, which suggests that crucial information for both vowel and consonant identification is contained across the same initial part of each syllable; and (d) The spectral transition is more crucial than unvoiced and buzz bar periods for consonant (syllable) perception, although the latter features are of some perception importance. It was also found that vowel nuclei are not necessary for either vowel or syllable perception.

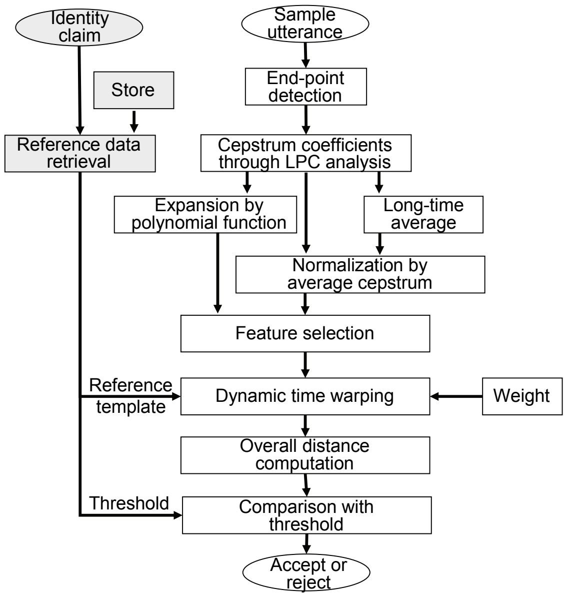


Figure 4: *Overall structure of the speaker verification system using dynamic features of cepstrum coefficients.*

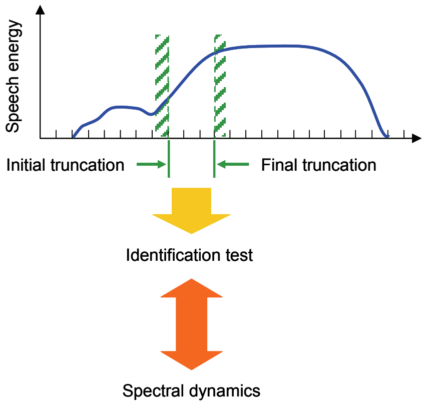


Figure 5: *Analysis of relationships between spectral dynamics and syllable perception using initial and final truncation syllables.*

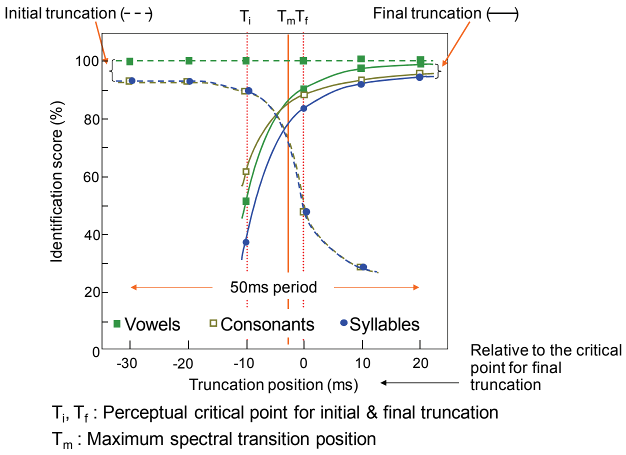


Figure 6: *Relationships between truncation position and identification scores for the truncated syllables.*

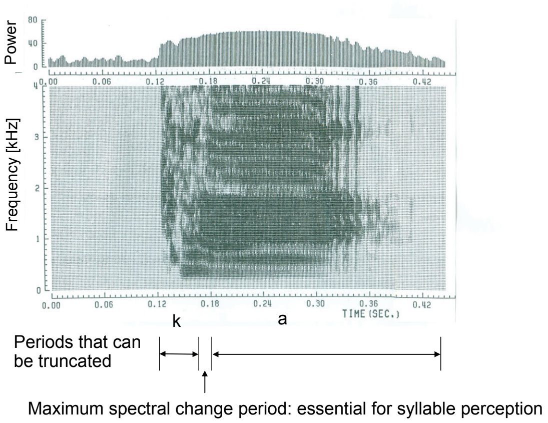

Power
Frequency [kHz]
TIME (SEC.)
k
a
Periods that can be truncated
Maximum spectral change period: essential for syllable perception

Figure 7: *Role of spectral transition for speech perception.*

### 3.2. Spectral dynamics in speech recognition

The speaker recognition technique based on a combination of instantaneous and dynamic features of the speech spectrum was slightly modified and applied to speech recognition [7]. Spoken utterances were represented by time sequences of cepstrum coefficients and energy, and the first order orthogonal polynomial coefficients (regression coefficients, delta features) for these time functions were extracted for every frame over an approximately 50ms period. The delta features were combined with time functions of the original cepstrum coefficients, and used with a DTW algorithm to compare multiple templates and input speech. Speaker-independent isolated word recognition experiments using a vocabulary of 100 Japanese city names indicated that a recognition error rate of 2.4% could be obtained with this method. Using only the original cepstrum coefficients the error rate was 6.2%. After 20 years, this method is still widely used in HMM-based speech recognition.

### 3.3. Speaker recognition by HMM/GMM

A VQ (vector quantization)-distortion-based speaker recognition method and discrete/continuous ergodic HMM–based methods were investigated especially from the viewpoint of robustness against utterance variations [8]. It was shown that a continuous ergodic HMM was as robust as a VQ-distortion method when enough data was available and that a continuous ergodic HMM was far superior to a discrete ergodic HMM. It was also shown that the information on transitions between different states was ineffective for text-independent speaker recognition. Therefore, the speaker identification rates using a continuous ergodic HMM were strongly correlated with the total number of mixtures irrespective of the number of states as shown in Figure 8. This indicated that GMMs, which have only one state, and ergodic HMMs having multiple states are equally effective assuming that they have the same total number of mixtures.

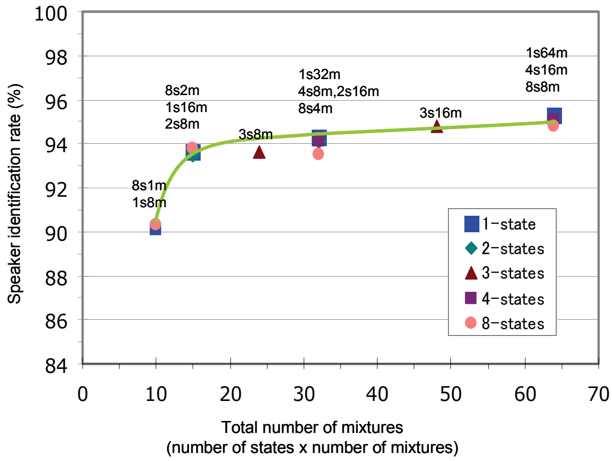


Figure 8: *Speaker identification rates as a function of the number of states and mixtures in each state in ergodic HMMs.*

## 4. 1990s

### 4.1. Japanese large-vocabulary continuous-speech recognition using a newspaper corpus and broadcast news

The first Japanese LVCSR experiments were conducted using newspaper reading speech and broadcast-news speech [9]. Since Japanese sentences are written without spaces between words, sentences were segmented into morphemes with a morphological analyzer and then n-gram language models were trained using those units. A morpheme is smaller than a "word"; words with inflection were divided into their stems and inflections and compound words were divided into elements. Although the Japanese language has many compound words and many types of inflections, much like the German language, morpheme-based units prevent the vocabulary size from becoming excessively large and can keep the lexical coverage high. It was also found that Japanese and French have more homophones than other European languages. The results for recognition of read newspaper speech with a 7k vocabulary and broadcast news speech with a 20k vocabulary were comparable to those for European languages. The research on broadcast news speech transcription was further extended by NHK (Japan Broadcast Company) Laboratories and an automatic closed captioning system was developed to be used for real broadcast news.

### 4.2. Robust speech recognition

Various techniques for increasing robustness in automatic speech recognition were investigated, mainly focusing on unsupervised or supervised acoustic model adaptation. They include a hierarchical spectral clustering-based method [10], an MCE (minimum classification error) training-based method [11], and an N-best-based method [12].

The hierarchical spectral clustering-based method was designed for unsupervised adaptation for VQ-based speech recognition. In this method, a set of spectra in training frames and the codebook entries are clustered hierarchically. Based on the deviation vectors between centroids of the training frame clusters and the corresponding codebook clusters, adaptation is performed hierarchically from small to large numbers of clusters. The spectral resolution of the adaptation process is therefore improved accordingly.

In the MCE training-based method, HMM parameters are first adapted to a new speaker by maximum a posteriori (MAP) estimation and then modified using MCE estimation during the supervised adaptation. Since the MCE estimation directly aims at minimizing the recognition error, the combination is effective in improving the recognition accuracy.

The N-best-based method was proposed to cope with the recognition errors in the supervision hypothesis used in unsupervised adaptation. Smoothed estimation and utterance verification were also introduced. Experimental results show that this method is effective in improving the recognition accuracy especially for difficult speakers.

### 4.3. Text-prompted speaker recognition

Conventional text-dependent speaker verification systems could easily be circumvented, because someone could play back the recorded voice of a registered speaker uttering key words or sentences into the microphone and be accepted as the registered speaker. Another problem was that people often do not like text-dependent systems because they do not like to utter their identification number, such as their social security number, within hearing of other people. To cope with these problems, a text-prompted speaker recognition method was proposed in which key sentences are completely changed every time [13]. The system accepted the input utterance only when it determined that the registered speaker uttered the prompted sentence. Because the vocabulary was unlimited, prospective impostors could not know in advance the sentence they would be prompted to say. This method not only accurately recognized speakers, but could also reject an utterance whose text differed from the prompted text, even if it was uttered by a registered speaker. Thus, a pre-recorded utterance could be correctly rejected.

This method used speaker-specific phone models as basic acoustic units as shown in Figure 9. One of the major issues in this method was how to properly create these speaker-specific phone models when using training utterances of a limited size. The phone models were represented by Gaussian-mixture continuous HMMs or tied-mixture HMMs, and they were made by adapting speaker-independent phone models to each speaker's voice. In the recognition stage, the system concatenated the phone models of each registered speaker to create a sentence HMM, according to the prompted text. Then the likelihood of the input speech given the sentence model was calculated and used for speaker verification. Robustness of this method was confirmed through various experiments.

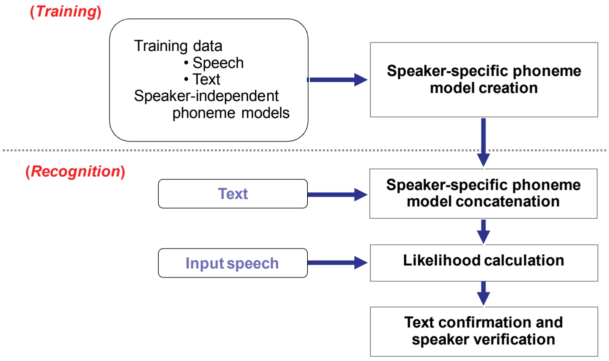

(Training)
Training data
• Speech
• Text
Speaker-independent phoneme models
Speaker-specific phoneme model creation
(Recognition)
Text
Speaker-specific phoneme model concatenation
Input speech
Likelihood calculation
Text confirmation and speaker verification

Figure 9: *Text-prompted speaker recognition method.*

## 5. 2000s

### 5.1. Spontaneous speech project and the CSJ corpus

A five-year large-scale national project entitled "Spontaneous Speech: Corpus and Processing Technology" was conducted in Japan from 1999 to 2004, spending approximately $10M, with the following objectives [14, 15]:

(a) Constructing a large-scale spontaneous speech corpus, the Corpus of Spontaneous Japanese (CSJ), consisting of roughly 7M words (morphemes) with a total speech length of 650 hours.
(b) Acoustic and language modeling for spontaneous speech recognition and understanding using linguistic as well as para-linguistic information in speech.
(c) Developing spontaneous speech recognition and summarization technology.

Mainly recorded in the CSJ were monologues such as academic presentations (AP) and extemporaneous presentations (EP) as shown in Table 1. The CSJ also included smaller databases of dialogue and read speech for the purpose of comparison. Figure 10 shows the process of the corpus construction. The recordings were manually given orthographic and phonetic transcriptions. Spontaneous speech-specific phenomena, such as filled pauses, word fragments, reduced articulation or mispronunciation, and non-speech events like laughter and coughing were also carefully tagged. One-tenth of the utterances, named Core, were tagged manually and used for training a morphological analysis and part-of-speech (POS) tagging program for automatically analyzing all of the 650-hours of utterances. The Core was also tagged with para-lingusitic/intonation information, dependency-structure, discourse structure, and summarization information. Various analyses, modeling and recognition experiments were conducted using the CSJ. Figures 11 and 12 show speech recognition results as a function of the size of training data for language and acoustic models.

Table 1. *Contents of the CSJ.*

| Type of Speech | # Speakers | # Files | Monologue/ Dialogue | Spontaneous/ Read | Hours |
|---|---|---|---|---|---|
| Academic presentations (AP) | 838 | 1006 | Monolog | Spont. | 299.5 |
| Extemporaneous presentations (EP) | 580 | 1715 | Monolog | Spont. | 327.5 |
| Interview on AP | * (10) | 10 | Dialog | Spont. | 2.1 |
| Interview on EP | * (16) | 16 | Dialog | Spont. | 3.4 |
| Task oriented dialogue | * (16) | 16 | Dialog | Spont. | 3.1 |
| Free dialogue | * (16) | 16 | Dialog | Spont. | 3.6 |
| Reading text | *(244) | 491 | Dialog | Read | 14.1 |
| Reading transcriptions | * (16) | 16 | Monolog | Read | 5.5 |
| *Counted as the speakers of AP or EP | | | | Total hours | 658.8 |

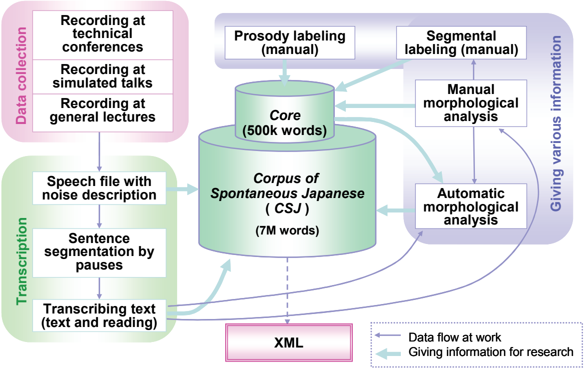

Data collection
Recording at technical conferences
Recording at simulated talks
Recording at general lectures
Transcription
Speech file with noise description
Sentence segmentation by pauses
Transcribing text (text and reading)
Prosody labeling (manual)
Segmental labeling (manual)
Core (500k words)
Corpus of Spontaneous Japanese ( CSJ ) (7M words)
Manual morphological analysis
Automatic morphological analysis
Giving various information
XML
Data flow at work
Giving information for research

Figure 10: *CSJ corpus construction process.*

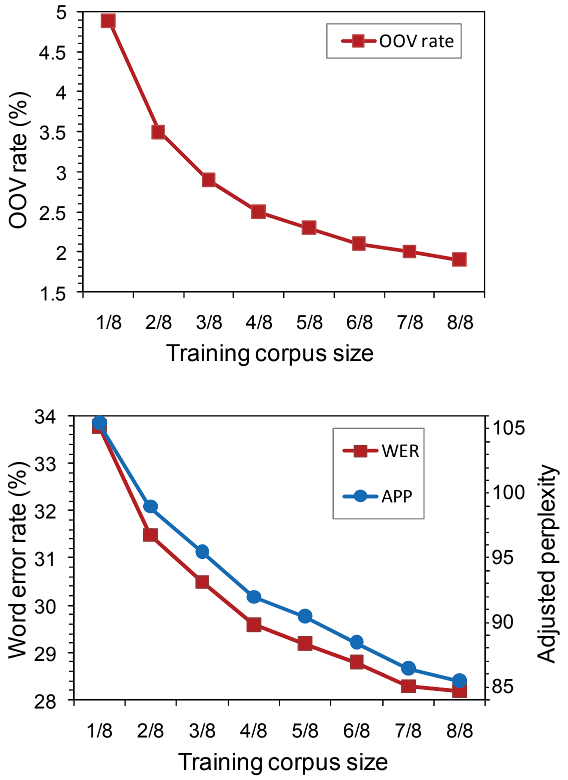


Figure 11: *Out-of-vocabulary (OOV) rate, word error rate (WER) and adjusted test-set perplexity (APP) as a function of the size of language model training data (8/8 = 6.84M words).*

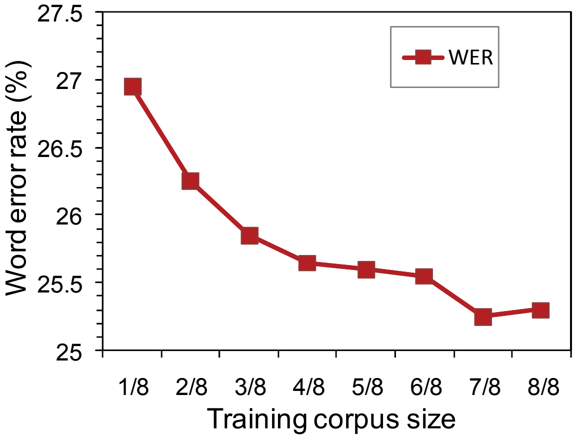


Figure 12: *Word error rate (WER) as a function of the size of acoustic model training data (8/8 = 510 hours).*

## 5.2. Spectral reduction in spontaneous speech

We have analyzed spectral differences between spontaneous and read speech using the CSJ [16]. In order to avoid the effect of individual differences, utterances in various different styles by the same five male and five female speakers were compared. Utterances by a larger set of speakers were used in speech recognition experiments. It was clarified that spectral distribution/difference of phonemes in spontaneous speech is significantly reduced compared to that of read speech as shown in Figure 13. It has also been found that the more spontaneous the speech, the smaller the distances between phonemes become, and that there is a high correlation between the mean phoneme distance and the phoneme recognition accuracy as shown in Figure 14. This indicates that spectral reduction is one major reason for the decrease of recognition accuracy of spontaneous speech. It was also confirmed that the word recognition accuracy is highly correlated with test-set perplexity. Therefore, word recognition accuracy can be approximated by combining mean Mahalanobis distance between phonemes and test-set perplexity.

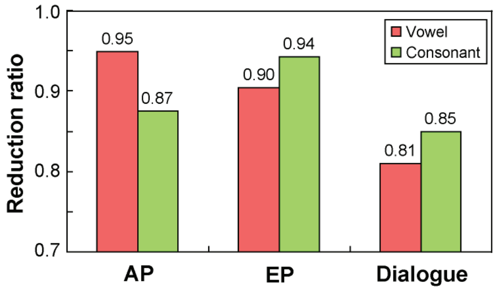
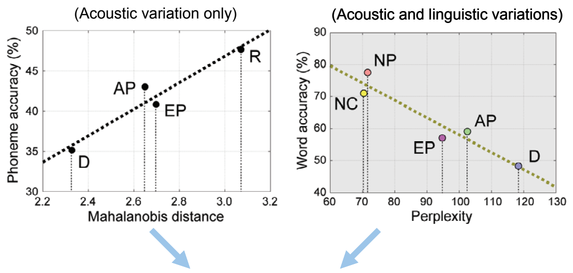


Figure 13: *Mean reduction ratios of vowels and consonants for each speaking style (AP: academic presentations, EP: extemporaneous presentations).*



$$Accuracy \simeq ax + by + c$$

$x$ : Mean Mahalanobis distance between phonemes

$y$ : Test – set perplexity

$a,b,c$ : Constant

Figure 14: *Approximation of word recognition accuracy by combining mean Mahalanobis distance between phonemes and test-set perplexity (R: read speech, D: dialogue, NP: newspaper articles, NC: news commentary).*

### 5.3. Automatic speech summarization and evaluation

We proposed a two-stage summarization method consisting of sentence extraction and word-based sentence compaction for summarizing broadcast news speech and presentation speech as shown in Figure 15 [17, 18]. After automatic sentence segmentation and removing all the fillers based on speech recognition results, a set of relatively important sentences was extracted, and sentence compaction was applied to the set of extracted sentences. The ratio of sentence extraction and compaction was controlled according to a summarization ratio determined by the user. The relatively important sentences were extracted using a significance score or dimension reduction based on SVD, combined with a sentence location-based score. Word units were extracted and concatenated to maximize the weighted sum of linguistic likelihood, amount of information, confidence measure, and grammatical likelihood of concatenated units [19]. Several evaluation metrics, including "Summarization accuracy" and its variations, were proposed and evaluated in comparison with other metrics, such as ROUGE. Although the correlation between the subjective and objective scores averaged over presentations was high, the correlation for each individual presentation was not very high due to the large variation of characteristics across presentations.

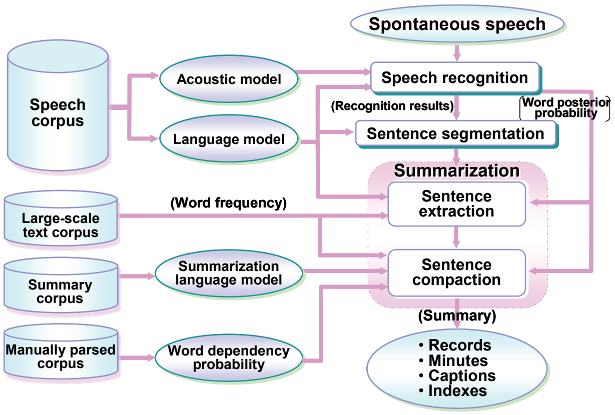


Figure 15: *Speech summarization by sentence extraction and compaction.*

### 5.4. Development of WFST-based decoder and application

We are currently developing the Tokyo Tech Transducer-based decoder, $T^3$ *decoder*, based on the Weighted Finite State Transducer (WFST) framework with the aim of achieving high performance and flexibility [20, 21, 22]. In the WFST framework, models used for speech recognition are all expressed as WFSTs. WFST operations can then be used to compose and optimize the models together, as shown in Figure 16, and this can give a final, highly efficient network that can achieve high recognition performance.

Drawbacks with the approach are that large models require large amounts of memory during optimization and decoding, and access to the original knowledge sources is lost so that making on-line changes is very difficult. To reduce memory usage and increase flexibility during decoding, we have proposed extended/generalized on-the-fly dynamic composition techniques. We have also proposed a fast method for computing acoustic likelihoods that makes use of a Graphics Processing Unit (GPU). After enabling the GPU acceleration the main processor runtime dedicated to acoustic scoring tasks was reduced from the largest consumer to just a few percent even when using mixture models with a large number of Gaussian components. Experimental results showed a large reduction in decoding time with no change in accuracy.

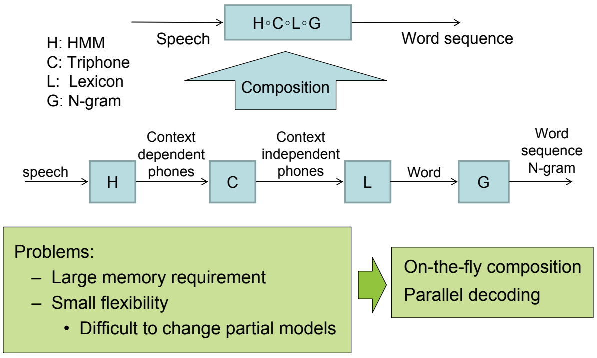
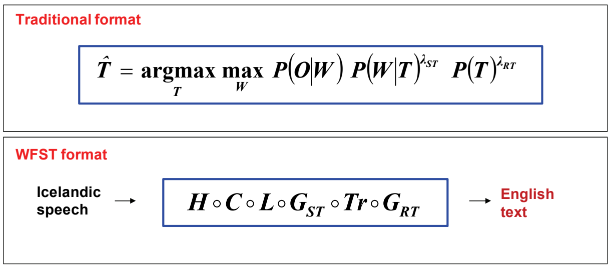


Figure 16: *WFST-based decoder.*

The decoder has been successfully applied to various LVCSR tasks. It has also been applied to speech recognition of a resource-deficient language (Icelandic) using a language model for a resource-rich language (English) and a translation model as shown in Figure 17. In a weather information domain with a large English corpus, good recognition results were obtained for an experimental system which accepted Icelandic speech as input and produced English text as output [23].


Traditional format
$\hat{T} = \underset{T}{\mathrm{argmax}} \max_{W} P(O|W)\, P(W|T)^{\lambda_{ST}}\, P(T)^{\lambda_{RT}}$
WFST format
Icelandic speech
$H \circ C \circ L \circ G_{ST} \circ Tr \circ G_{RT}$
English text

- $P(O|W)$: Icelandic acoustic model
- $P(W|T)$: English to Icelandic translation model
- $P(T)$: English language model

Figure 17: *Icelandic speech recognition using English language model, translation model, and the WFST-based decoder.*

### 5.5. Unsupervised cross-validation and aggregated adaptation methods

In order to reduce the over-training problem and suppress the negative effects of recognition errors in the unsupervised batch-mode acoustic model adaptation, we have proposed cross-validation (CV) and aggregated adaptation algorithms [24]. The latter algorithm is based on the idea of the bagging approach. In both algorithms, the adaptation utterances are split into *K* exclusive subsets, each with roughly the same size. In the CV adaptation, the adaptation utterances used in the decoding step and those used in the model updating step are separated based on the *K*-fold CV technique as shown in Figure 18. *K* sets of recognition hypotheses are made using the separate adaptation utterance sets. Each of the *K* models is then adapted using different *K*-1 sets of hypotheses, and each adapted model is used to decode the utterance set that was not used to adapt the model. This process is repeated until the results converge.

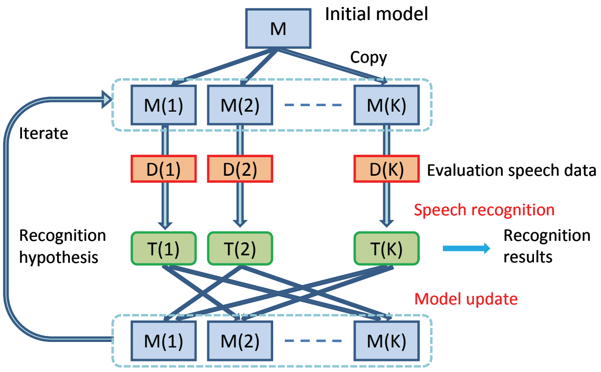


Figure 18: *Unsupervised cross-validation (CV) adaptation.*

In the aggregated adaptation, each adaptation utterance set is decoded *N* times using separate models. Initially, these *N* models are made by copying the initial model. Each of the *N* models is adapted using *N* x *K'* (*K'*<*K*) sets of hypotheses. The *K'* subsets are randomly selected. The adapted *N* models are used to decode each set of adaptation utterances. This process is repeated until the results converge. Any kind of conventional adaptation techniques, such as the MLLR method, can be used in the model adaptation step in both algorithms. Since the proposed methods can suppress the negative effects of recognition errors included in the hypotheses for adaptation, they achieve significantly better results than a normal batch-mode unsupervised adaptation method, and the CV adaptation is more efficient than the aggregated adaptation.

## 6. Future works

Speech is the primary, and the most convenient means of communication between people. Whether due to technological curiosity to build machines that mimic humans or desire to automate work with machines, research in speech and speaker recognition, as a first step toward natural human-machine communication, has attracted much enthusiasm over the past five decades. Fortunately, I have been participating in this research for almost four decades. Although many important scientific advances have taken place, bringing us closer to the "Holy Grail" of automatic speech recognition and understanding by machine, we have also encountered a number of practical limitations which hinder a widespread deployment of applications and services. In most speech recognition tasks, human subjects produce one to two orders of magnitude fewer errors than machines. There is now increasing interest in finding ways to bridge this performance gap. What we know about human speech processing is very limited. Significant advances in speech and speaker recognition are not likely to come solely from research in statistical pattern recognition and signal processing. Although these areas of investigation are important, the most significant advances in next generation systems will come from studies in acoustic-phonetics, speech perception, linguistics, and psychoacoustics. Future systems need to have an efficient way of representing, storing, and retrieving various knowledge resources required for natural conversation [25].

## 7. References

[1] Furui, S., "50 years of progress in speech and speaker recognition," Proc. SPECOM 2005, Patras, Greece, pp. 1-9, 2004.

[2] Furui, S., Itakura, F., and Saito, S., "Talker recognition by longtime averaged speech spectrum," Trans. IECE, 55-A, pp. 549-556, 1972.

[3] Furui, S., "An analysis of long-term variation of feature parameters of speech and its application to talker recognition," Trans. IECE, 57-A, pp. 880-887, 1974.

[4] Furui, S., "Comparison of speaker recognition methods using statistical features and dynamic features," IEEE Trans. Acoustics, Speech, Signal Processing, 29, pp.342-350, 1981.

[5] Furui, S., "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust. Speech, Signal Processing, 29, pp. 254-272, 1981.

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

[7] Furui, S., "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoustics, Speech, Signal Processing, 34, pp. 52-59, 1986.

[8] Matsui, T. and Furui, S., "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, San Francisco, pp. II-157-160, 1992.

[9] Ohtsuki, K., Matsuoka, T., Mori, T., Yoshida, K., Taguchi, Y., Furui, S. and Shirai, K., "Japanese large-vocabulary continuous-speech recognition using a newspaper corpus and broadcast news," Speech Communication, 28, pp. 155-166, 1999.

[10] Furui, S., “Unsupervised speaker adaptation based on hierarchical spectral clustering,” IEEE Trans. Acoustics, Speech, Signal Processing, 37, pp.1923-1930, 1989.

[11] Matsui, T. and Furui, S. “A study of speaker adaptation based on minimum classification error training,” Proc. Eurospeech, Madrid, Spain, pp. 81-84, 1995.

[12] Matsui, T. and Furui, S., “N-best-based unsupervised speaker adaptation for speech recognition,” Computer Speech and Language, 12, pp. 41-50, 1998.

[13] Matsui, T. and Furui, S., “Concatenated phoneme models for text-variable speaker recognition,” Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Minneapolis, pp. II-391-394, 1993.

[14] Furui, S., “Recent advances in spontaneous speech recognition and understanding,” Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan, pp. 1-6, 2003.

[15] Furui, S., “Recent progress in corpus-based spontaneous speech recognition,” IEICE Trans. Inf. & Syst., E88-D, pp. 366-375, 2005.

[16] Furui, S., Nakamura, M., Ichiba, T. and Iwano, K., “Why is the recognition of spontaneous speech so hard?” Proc. 8th Int. Conf. Text, Speech and Dialogue (TSD 2005), Katlovy Vary, Czech Republic, pp. 9-22, 2005.

[17] Furui, S., “Speech-to-text and speech-to-speech summarization of spontaneous speech,” IEEE Trans. Speech & Audio Processing, 12, pp. 401-408, 2004.

[18] Hirohata, M., Shinnaka, Y., Iwano, K. and Furui, S., “Sentence extraction-based presentation summarization techniques and evaluation metrics,” Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Philadelphia, pp. I-1065-1068, 2005.

[19] Hori, C. and Furui, S., “Advances in automatic speech summarization,” Proc. Eurospeech, Aalborg, Denmark, pp. 1771-1774, 2001

[20] Dixon, P. R., Caseiro, D. A., Oonishi, T. and Furui, S., “The Titech large vocabulary WFST speech recognition system,“ Proc. IEEE Automatic Speech Recognition and Understanding (ASRU 2007), Kyoto, Japan, pp. 1301-1304, 2007.

[21] Oonishi, T., Dixon, P. R., Iwano, K. and Furui, S., “Generalization of specialized on-the-fly composition,” Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Taipei, Taiwan, pp. 4317-4320, 2009.

[22] Dixon, P. R., Oonishi, T. and Furui, S., “Fast acoustic computation using graphics processors,” Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Taipei, Taiwan, pp. 4321-4324, 2009.

[23] Jensson, A. T., Oonishi, T., Iwano, K. and Furui, S., “Development of a WFST-based speech recognition system for a resource deficient language using machine translation,” Proc. 2009 APSIPA Annual Summit and Conference, Sapporo, Japan, 2009.

[24] Shinozaki, T., Kubota, Y. and Furui, S., “Unsupervised cross-validation and aggregated adaptation for improved spontaneous speech recognition,” Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Taipei, Taiwan, pp. 4377-4380, 2009.

[25] Furui, S., “21st century COE program ‘Framework for systematization and application of large-scale knowledge resources’,” Proc. Symposium on Large-scale Knowledge Resources (LKR 2005), Tokyo, Japan, pp. 3-8, 2005.