Chapter 1

SPEECH AND SPEAKER RECOGNITION EVALUATION

Sadaoki Furui
Department of Computer Science, Tokyo Institute of Technology
Tokyo, Japan
furui@cs.titech.ac.jp

Abstract This chapter overviews techniques for evaluating speech and speaker recognition systems. The chapter first describes principles of recognition methods, and specifies types of systems as well as their applications. The evaluation methods can be classified into subjective and objective methods, among which the chapter focuses on the latter methods. In order to compare/normalize performances of different speech recognition systems, test set perplexity is introduced as a measure of the difficulty of each task. Objective evaluation methods of spoken dialogue and transcription systems are respectively described. Speaker recognition can be classified into speaker identification and verification, and most of the application systems fall into the speaker verification category. Since variation of speech features over time is a serious problem in speaker recognition, normalization and adaptation techniques are also described. Speaker verification performance is typically measured by equal error rate, detection error trade-off (DET) curves, and a weighted cost value. The chapter concludes by summarizing various issues for future research.

Keywords Speech recognition; Speaker recognition; Objective evaluation; Subjective evaluation; Perplexity; Accuracy; Correctness; Equal error rate; DET curve.

1 Introduction

Given the complexity of the human–computer interface, it is clear that evaluation protocols are required which address a large number of different types of spoken language systems, including speech recognition and speaker recognition components. The majority of research in the area of spoken language system evaluation has concentrated on evaluating system components, such as measuring the word recognition accuracy for a speech recognizer, rather than overall effectiveness measures for complete systems.

L. Dybkjær et al. (eds.), Evaluation of Text and Speech Systems, 1–27.

In the United States, a very efficient evaluation paradigm has been funded by the Defense Advanced Research Projects Agency (DARPA) which includes an efficient production line of "hub and spoke"-style experiments involving the coordination of design, production and verification of data, distribution through Linguistic Data Consortium (LDC), and design, administration and analysis of testing by National Institute of Standards and Technology (NIST). These organizations have strongly advocated the importance of establishing appropriate "benchmarks", either through the implementation of standard tests, or by reference to human performance or to reference algorithms.

In order to give the reader information on how to evaluate the performance of spoken language systems, this chapter first specifies the types of systems and their applications, since this is important for understanding and using the evaluation methods. The chapter next introduces various performance measures, followed by discussions of the parameters which affect the performance. The chapter then goes on to an evaluation framework which includes high-level metrics such as correction and transaction success.

To obtain a detailed description of various evaluation techniques for spoken language systems, readers are suggested to refer to the handbook by (Gibbon et al., 1998).

2 Principles of Speech Recognition

In the state-of-the-art approach, human speech production as well as the recognition process is modelled through four stages: text generation, speech production, acoustic processing, and linguistic decoding, as shown in Figure 1

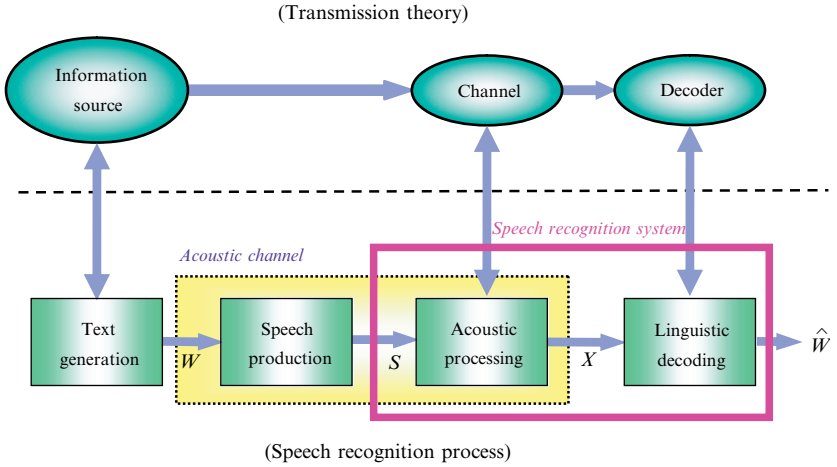

Figure 1. Structure of the state-of-the-art speech recognition system.

(Furui, 2001). A speaker is represented as a transducer that transforms into speech the text of thoughts he/she intends to communicate (information source). Based on the information transmission theory, the sequence of processes is compared to an information transmission system, in which a word sequence W is converted into an acoustic observation sequence X, with a probability $P(W, X)$, through a noisy transmission channel, which is then decoded into an estimated sequence W. The goal of recognition is then to decode the word string, based on the acoustic observation sequence, so that the decoded string has the maximum a posteriori (MAP) probability (Rabiner and Juang, 1993; Young, 1996), i.e.,

$$\hat{W} = \arg\max_{W} P(W\,|X) \tag{1.1}$$

Using Bayes' rule, Eq. 1.1 can be written as

$$\hat{W} = \arg\max_{W} P(X\,|W)P(W)/P(X) \tag{1.2}$$

Since $P(X)$ is independent of W, the MAP decoding rule of Eq. 1.2 is converted into

$$\hat{W} = \arg\max_{W} P(X\,|W)P(W) \tag{1.3}$$

The first term in Eq. 1.3, $P(X|W)$, is generally called the acoustic model as it estimates the probability of a sequence of acoustic observations conditioned with the word string. The second term, $P(W)$, is generally called the language model since it describes the probability associated with a postulated sequence of words. Such language models can incorporate both syntactic and semantic constraints of the language and the recognition task. Often, when only syntactic constraints are used, the language model is called a grammar.

Hidden Markov Models (HMMs) and statistical language models are typically used as acoustic and language models, respectively. Figure 2 shows the information flow of the MAP decoding process given the parameterized acoustic signal X. The likelihood of the acoustic signal $P(X|W)$ is computed using a composite HMM representing W constructed from simple HMM phoneme models joined in sequence according to word pronunciations stored in a dictionary (lexicon).

3 Categories of Speech Recognition Tasks

Speech recognition tasks can be classified into four categories, as shown in Table 1, according to two criteria: whether it is targeting utterances from human to human or human to computer, and whether the utterances have a dialogue or monologue style (Furui, 2003). Table 1 lists typical tasks and data corpora that are representative for each category.

The Category I targets human-to-human dialogues, which are represented by the DARPA-sponsored recognition tasks using Switchboard and Call Home

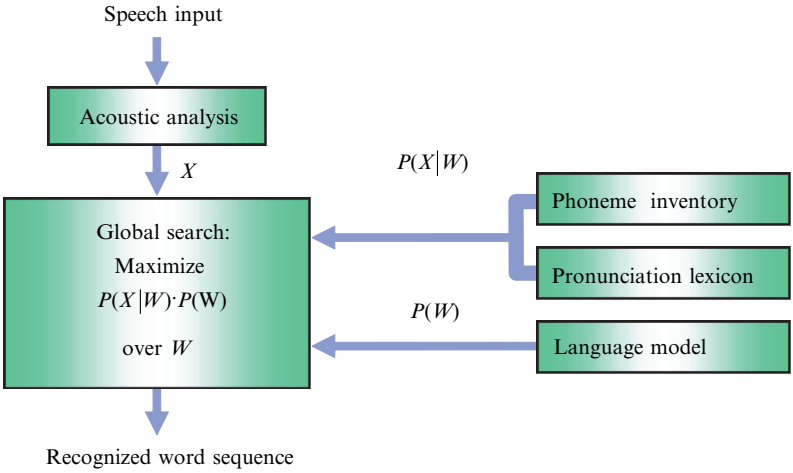

Figure 2. Overview of statistical speech recognition.

Table 1. Categorization of speech recognition tasks.

	Dialogue	Monologue
Human to human	(Category I) Switchboard, Call Home (Hub5), meeting, interview	(Category II) Broadcast news (Hub 4), other programmes, lecture, presentation, voice mail
Human to machine	(Category III) ATIS, Communicator, information retrieval, reservation	(Category IV) Dictation

(Hub 5) corpora. Speech recognition research in this category aiming to produce minutes of meetings (e.g., Janin et al., 2004) has recently started. Waibel and Rogina (2003) have been developing a meeting browser that observes and tracks meetings for later review and summarization. Akita et al. (2003) have investigated techniques for archiving discussions. In their method, speakers are automatically indexed in an unsupervised way, and speech recognition is performed using the results of the indexing. Processing human–human conversational speech under unpredictable recording conditions and vocabularies presents new challenges for spoken language processing.

A relatively new task classified into this category is the Multilingual Access to Large spoken ArCHives (MALACH) project (Oard, 2004). Its goal is to advance the state-of-the-art technology for access to large multilingual collections

of spontaneous conversational speech by exploiting an unmatched collection assembled by the Survivors of the Shoah Visual History Foundation (VHF). This collection presents a formidable challenge because of heavily accented, emotional, and old-age characteristics of the survivor's spontaneous speech. Named entity tagging, topic segmentation, and unsupervised topic classification are also being investigated.

Tasks belonging to Category II, which targets recognizing human–human monologues, are represented by transcription of broadcast news (Hub 4), news programmes, lectures, presentations, and voice mails (e.g., Hirschberg et al., 2001). Speech recognition research in this category has recently become very active. Since utterances in Category II are made with the expectation that the audience can correctly understand what he/she speaks in the one-way communication, they are relatively easier as a target of speech to recognize than utterances in the Category I. If a high recognition performance can be achieved for the utterances in Category II, a wide range of applications, such as making lecture notes, records of presentations and closed captions, archiving and retrieving these records, and retrieving voice mails, will be realized.

Most of the practical application systems widely used now are classified into Category III, recognizing utterances in human–computer dialogues, such as in airline information services tasks. DARPA-sponsored projects including ATIS and Communicator have laid the foundations of these systems. Unlike other categories, the systems in Category III are usually designed and developed after clearly defining the application/task. The machines that have been designed so far are, almost without exception, limited to the simple task of converting a speech signal into a word sequence and then determining from the word sequence a meaning that is "understandable". Here, the set of understandable messages is finite in number, each being associated with a particular action (e.g., route a call to a proper destination or issue a buy order for a particular stock). In this limited sense of speech communication, the focus is detection and recognition rather than inference and generation.

Various researches have made clear that utterances spoken by people talking to computers, such as those in Categories III and IV, especially when the people are conscious of computers, are acoustically, as well as linguistically, very different from utterances directed towards people, such as those in Categories I and II. One of the typical tasks belonging to Category IV, which targets the recognition of monologues performed when people are talking to computers, is dictation, and various commercial softwares for such purposes have been developed. Since the utterances in Category IV are made with the expectation that the utterances will be converted exactly into texts with correct characters, their spontaneity is much lower than those in Category III. Among the four categories, spontaneity is considered to be the highest in Category I and the lowest in Category IV.

Speech recognition tasks can also be classified according to whether it is isolated word recognition or continuous speech recognition and whether it is speaker-dependent or speaker-independent recognition. For isolated words, the beginning and the end of each word can be detected directly from the energy of the signal. This makes word boundary detection (segmentation) and recognition much easier than if the words are connected. However, in real applications where speech is contaminated by noise, it is not always easy to detect word boundaries by simply relying on the energy of the signal. Speaker-independent recognition is more difficult than speaker-dependent recognition, since the speech model must somehow be general enough to cover all types of voices and all possible ways of word pronunciations, and yet specific enough to discriminate between individual words. For a speaker-dependent system, training or adaptation of speech models is carried out by using utterances of each speaker. In speaker adaptation, the system is bootstrapped with speaker-independent models, and then gradually adapts to the specific aspects of the speaker.

4 Evaluation of Speech Recognition Systems

4.1 Classification of Evaluation Methods

Techniques for evaluating speech recognition methods/systems can be categorized depending on whether they use subjective or objective methods. The former directly involve human subjects during measurement, whereas the latter, typically using prerecorded speech, do not directly involve human subjects. Objective methods have the advantage of producing reproducible results and of lending themselves to being automated; thus, they are also more economical. The problem with objective methods for speech recognition application evaluation is that it is difficult to create methods with the capacity to cope easily with the complex processes required for evaluating speech understanding or interaction systems. On the other hand, subjective methods are more suited to evaluating applications with higher semantic or dialogue content, but they suffer from the fact that human subjects cannot reliably perform quality measurement and that they cannot handle fine-grained measurement scales, either. On average, a human subject uses gradation scales with 5–10 levels and no more.

In order to compare performances of different speech recognition systems, it is necessary to normalize the difficulty of the task of each system. For this purpose, the following task difficulty evaluation methods are used.

4.2 Evaluation of Task Difficulty

In order to reduce the effective number of words to select from, recognition systems are often equipped with some linguistic knowledge. This may vary

from very strict syntax rules, in which the words that may follow one another are defined by certain rules, to probabilistic language models, in which the probability of the output sentence is taken into consideration, based on statistical knowledge of the language. An objective measure of the freedom of the language model is *perplexity*, which measures the average branching factor of the language model (Ney et al., 1997). The higher the perplexity, the more words to choose from at each instant, and hence the more difficult the task.

The perplexity is defined by

$$PP = 2^{H(L)} \tag{1.4}$$

where $H(L)$ is the *entropy* of the language model per word, which is defined by

$$H\left(L\right) = -\sum_{w_1 \ldots w_n} \frac{1}{n} P\left(w_1 \ldots w_n\right) \log P\left(w_1 \ldots w_n\right) \tag{1.5}$$

Here, $P(w_1...w_n)$ is the probability of producing a word sequence $w_1...w_n$ given the language model L. $H(L)$ indicates the amount of information (bits) necessary to specify a word produced by the language model. The perplexity defined above is often called language model perplexity.

Performance of a speech recognition system depends not only on its task but also on texts of a test set, i.e., a set of utterances to be used for a recognition test. Therefore, in order to evaluate the difficulty of the test set, the perplexity is often calculated for the test set, which is called *test set perplexity* or *corpus perplexity*. If we assume that the Ergodic feature exists for language, the entropy per word can be calculated as follows:

$$H(L) = -\frac{1}{Q} \log P_M\left(w_1 \ldots w_Q\right) \tag{1.6}$$

where $P_M(w_1...w_Q)$ is the probability of producing the test set word sequence $w_1...w_Q$. Therefore, the test set perplexity can be calculated as follows:

$$PP = 2^{H(L)} = P_M\left(w_1 \ldots w_Q\right)^{-\frac{1}{Q}} \tag{1.7}$$

The above equations show that the test set perplexity is the geometric average of the reciprocal probability over all Q words. Apart from the constant factor $(-1/Q)$, the perplexity is identical to the average conditional probability or likelihood. Therefore, minimizing the perplexity is the same as maximizing the log-likelihood function. Since the test set for recognition experiments should be separate from the corpus that is used to construct the language model, the language model perplexity and the test set perplexity are usually different.

When a formal grammar, such as finite automaton and context-free grammar, is used as the language model, every partially parsed tree up to word w_i

is made, and the number of words that can follow the word w_i is calculated. The test set perplexity is obtained as a geometric mean of the number of possible following words at each word w_i, assuming that every word is selected with equal probability.

The set of all words the recognition system has been set up to be able to recognize is called *vocabulary* V_L. The vocabulary size is one of the measures indicating the task difficulty. The test vocabulary V_R is defined as the set of words appearing in the evaluation test. A word w is called out-of-vocabulary (OOV) if it is present in the test vocabulary but not in the recognizer's vocabulary. The *OOV rate* is defined as the ratio of the number of words in the test set which are not included in V_L to the total number of words in the test set. In general, the larger the test vocabulary size V_R and the larger the OOV rate, the more difficult the task is.

The perplexity requires a closed vocabulary. If OOV exists, the perplexity definition may become problematic because it then becomes infinitely large. Therefore, usually OOV class <UNK> (unknown word class) is defined and the language model of OOV is calculated by

$$p'(< UNK > |h) = \frac{p(< UNK > |h)}{V_R - V_L} \tag{1.8}$$

where h is the history. Since there are $(V_R - V_L)$ kinds of OOV words to be recognized that are not included in the language model vocabulary, the OOV probability is divided by $(V_R - V_L)$.

The perplexity changes according to the vocabulary size. In general, the perplexity decreases by decreasing the vocabulary size V_L, since the probability allocated to each word becomes larger. However, if the test vocabulary size V_R is fixed and the language model vocabulary size V_L is decreased, the linguistic constraint becomes lower, since the number of OOV in the test set increases. Therefore, the test set perplexity cannot be used for comparing the difficulty of the tasks if the OOV rates of the language models are different. In order to solve this problem, *adjusted perplexity* (APP) has been proposed (Ueberla, 1994). In APP, by using the language model of the OOV words defined above and defining V_R as union of V_L and all the words appearing in the test set, the perplexity is adjusted by the total number of OOV words, o, and the number of different OOV words, m, in the test set as follows:

$$\log APP = -\frac{1}{Q} \log P_M(w_1 \ldots w_Q) + o \log m \tag{1.9}$$

Although the perplexity and the OOV rate measure the test source's complexity from the recognizer's point of view, they refer to written (e.g., transcribed) forms of language only and completely disregard acoustic–phonetic modelling. Difficulty of the recognition task also depends on the length of the sentences (average number of words) and average number of phonemes of which each

word consists. Therefore, task difficulty needs to be measured by a combination of various factors covering both linguistic and acoustic complexity.

4.3 Objective Evaluation of General Recognition Performance

Isolated word scoring. The *error rate* of speech recognition is defined as "the average fraction of items incorrectly recognized". Here, an item can be a word, a subword unit (e.g., a phone), or an entire utterance. For an isolated word recognition system, the error rate is defined as:

$$E = \frac{N_E}{N} \tag{1.10}$$

Here, N is the number of words in the test utterance and N_E the number of words incorrectly recognized. The latter can be subdivided into substitution error, N_S, and deletion (incorrect rejection) error, N_D:

$$N_E = N_S + N_D \tag{1.11}$$

Sometimes the fraction of correctly recognized words, $C = 1 - E$, called *correctness*, is used:

$$C = \frac{N_C}{N} = \frac{N - N_S - N_D}{N} \tag{1.12}$$

These measures do not include so-called insertions, since it is assumed that the beginning and the end of each word can be detected directly from the energy of the signal. However, in real applications where speech is contaminated by noise, it is not always easy to detect word boundaries, and sometimes noise signals cause insertion errors. Therefore, in these practical conditions, the same measure as that used in continuous word scoring, which will be described later, is also used in the isolated recognition task.

For isolated word recognizers, a more specific measure than the various contributions to the error rate, a *confusion matrix*, has also been used, in which the class of substitutions is divided into all possible confusions between words. The confusion C_{ij} is defined as the probability that word i is recognized as word j. The value C_{ii} is the fraction of times word i is correctly recognized. These probabilities are estimated by measuring the number of times the confusion took place:

$$C_{ij} = \frac{N_{ij}}{\sum_{j'} N_{ij'}} \tag{1.13}$$

where N_{ij} is the number of times word j is recognized on the input word i. The confusion matrix gives more detailed information than the error rates. Insertions and deletions can also be included in the matrix by adding a null word $i = 0$ (non-vocabulary word). Then, the row C_{0j} contains insertions,

the column C_{i0} the deletions, and $C_{00} = 0$. Using this expanded confusion matrix, the error rate can be calculated from the diagonal elements, i.e., $E = 1 - \Sigma_i C_{ii} = \Sigma_{i \neq j} C_{ij}$. The elements C_{ij} for $i \neq j$ are called the off-diagonal elements.

Continuous word scoring. In continuous speech recognition, the output words are generally not time-synchronous with the input utterance. Therefore, the output stream has to be aligned with the reference transcriptions. This means that classifications such as substitutions, deletions, words correct and insertions can no longer be identified with complete certainty. The actual measurement of the quantities through alignment is difficult. The alignment process uses a dynamic programming algorithm to minimize the misalignment of two strings of words (symbols): the reference sentence and the recognized sentence. The alignment depends on the relative weights of the contributions of the three types of errors: substitutions, insertions, and deletions. Hunt (1990) discussed the theory of word–symbol alignment and analysed several experiments on alignment. Usually, the three types of errors have equal weights. Depending on the application, one can assign different weights to the various kinds of errors.

Thus, the total number of errors is the summation of three types of errors:

$$N_E = N_S + N_I + N_D \tag{1.14}$$

where N_S, N_I, and N_D are the numbers of substitutions, insertions, and deletions, respectively. The error rate is therefore

$$E = \frac{N_E}{N} = \frac{N_S + N_I + N_D}{N} \tag{1.15}$$

Note that this error measure can become larger than 1 in cases of extremely bad recognition. Often, one defines the *accuracy* of a system as

$$A = 1 - E = \frac{N - N_S - N_I - N_D}{N} \tag{1.16}$$

Note that this is not just the fraction C of words correctly recognized, because the latter does not include insertions.

NIST has developed freely available software for analysis of continuous speech recognition systems. It basically consists of two parts: an alignment program and a statistics package. The alignment program generates a file with all alignment information, which can be printed by another utility in various levels of detail. The statistics program can pairwise compare the results of different recognition systems and decide whether or not the difference in performance is significant.

Other scoring. The objective scores other than accuracy include percentage of successful task completions, the time taken to complete the task, or the number of interactions necessary per task.

Speaker variability. The variety of speech recognition performances is highly dependent on the speaker. Apparently, speakers can be classified as "goats" (low recognition scores) and "sheep" (high recognition scores) (Furui, 2001). Since knowledge of this classification is usually not available a priori, it is necessary to use many speakers for evaluation. A sufficient number of speakers allows estimation of the variance in score due to speaker variability, and significance can be tested using Student's t-test.

4.4 Objective Evaluation of Spoken Dialogue Systems

Performance measures that can be used for evaluating spoken dialogue systems are:

1. Recognition accuracy.
2. OOV rejection: a good system correctly rejects OOV words and asks the users to rephrase, instead of wrongly recognizing them as vocabulary words. This is actually a very difficult issue, since there is no perfect confidence measure for the recognition results.
3. Error recovery: both the system and the user are sources of errors. A good system allows the user to undo actions triggered by previous spoken commands.
4. Response time: important for good usability is the time it takes to respond to a spoken command, i.e., system reaction time. This is defined as the time from the end of the command utterance to the start of the action. Both the average time and the distribution of the response time are important parameters.
5. Situation awareness: users who give commands to a system have certain expectations about what they can say. The active vocabulary usually depends on the internal state of the system but if users are not aware of that state, it is said that they have lost their situational awareness. This can be expressed as the number of times a test subject uttered a command in a context where it was not allowed. A subjective impression by the tester or the subject can also be used as a measure. Suitable questions for the users could be:
   - Is the list of possible commands always clear?
   - Are special skills required?
   - Is on-line help useful?

To learn details of the issues for evaluating telephone-based spoken dialogue systems, readers are recommended to refer to the textbook by Möller (2005).

4.5 Objective Evaluation of Dictation Systems

Various commercial systems (software) for dictation using automatic speech recognition have been developed. The performance measures that can be used for evaluating these systems are:

1. Recognition accuracy
2. Dictation speed: number of words per minute that can be received
3. Error correction strategies: a good measure for the ease of error correction is the average time spent per correction

Dictation systems can be compared to other systems and also to human performance. Error rate and dictation speed are the most obvious performance measures for the human benchmark.

4.6 Subjective Evaluation Methods

In subjective evaluation, the test is designed in such a way that human subjects interact with the system. Subjective measures include level of intelligibility, general impression, annoyance, user-friendliness, intuitiveness, level of difficulty, and the subjective impression of system response time. The ultimate overall measure is: "Can the task be completed?" This is a measure that includes recognition, error recovery, situational awareness, and feedback. In this sense, the time required to complete the entire test might also be indicative of the quality of the system. General impressions of test subjects can be indicative of how the system performs.

5 Principles of Speaker Recognition

A technology closely related to speech recognition is speaker recognition, or the automatic recognition of a speaker (talker) through measurements of individual characteristics existing in the speaker's voice signal (Furui, 1997, 2001; Rosenberg and Soong, 1991). The actual realization of speaker recognition systems makes use of voice as the key tool for verifying the identity of a speaker for application to an extensive array of customer-demand services. In the near future, these services will include banking transactions and shopping using the telephone network as well as the Internet, voicemail, information retrieval services including personal information accessing, reservation services, remote access of computers, and security control for protecting confidential areas of concern.

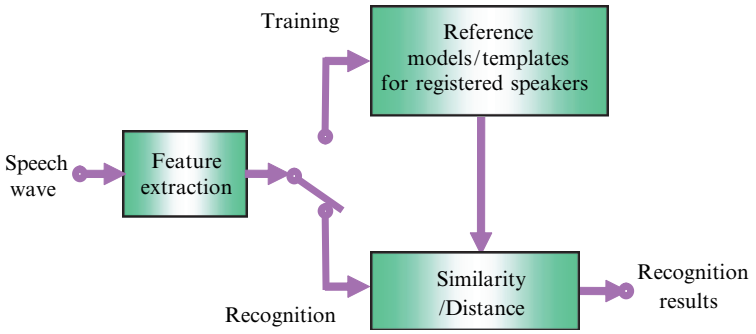

Figure 3. Principal structure of speaker recognition systems.

The common structure of speaker recognition systems is shown in Figure 3. Feature parameters extracted from a speech wave are compared with the stored templates or models for each registered speaker. The recognition decision is made according to the distance (or similarity) values. For speaker verification, input utterances with distances to the reference template/model smaller than the threshold are accepted as being utterances of the registered speaker (customer), while input utterances with distances larger than the threshold are rejected as being those of a different speaker (impostor). With speaker identification, the registered speaker whose reference template/model is nearest to the input utterance among all of the registered speakers is selected as the speaker of the input utterance.

6 Categories of Speaker Recognition Tasks

Speaker recognition can be principally divided into speaker verification and speaker identification. Speaker verification is the process of accepting or rejecting the identity claim of a speaker by comparing a set of measurements of the speaker's utterances with a reference set of measurements of the utterance of the person whose identity is being claimed. Speaker identification is the process of determining from which of the registered speakers a given utterance comes. The speaker identification process is similar to the spoken word recognition process in that both determine which reference model is closest to the input speech.

Speaker verification is applicable to various kinds of services using voice as the key input to confirming the identity claim of a speaker. Speaker identification is used in criminal investigations, for example, to determine which of the suspects produced a voice recorded at the scene of the crime. Since the possibility always exists that the actual criminal is not one of the suspects, however, the identification decision must be made through a combined process of speaker verification and speaker identification.

Speaker recognition methods can also be divided into text-dependent and text-independent methods. The former require the speaker to issue a predetermined utterance whereas the latter do not rely on a specific text being spoken. In general, because of the higher acoustic–phonetic variability of text-independent speech, more training material is necessary to reliably characterize (model) a speaker than with text-dependent methods.

Although several text-dependent methods use features of special phonemes, such as nasals, most text-dependent systems allow words (keywords, names, ID numbers, etc.) or sentences to be arbitrarily selected for each speaker. In the latter case, the differences in words or sentences between the speakers improves the accuracy of speaker recognition. When evaluating experimental systems, however, common keywords or sentences are usually used for every speaker.

Although keywords can be fixed for each speaker in many applications of speaker verification, utterances of the same words cannot always be compared in criminal investigations. In such cases, a text-independent method is essential. Difficulty in automatic speaker recognition varies depending on whether or not the speakers intend to have their identities verified. During actual speaker verification, speakers are usually expected to cooperate without intentionally changing their speaking rate or manner. It is well known, however, and expected that speakers are most often uncooperative in criminal investigations, consequently compounding the difficulty in correctly recognizing their voices.

Both text-dependent and text-independent methods have a serious weakness. These security systems can be circumvented, because someone can play back the recorded voice of a registered speaker uttering keywords or sentences into the microphone and be accepted as the registered speaker. Another problem is 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 distance of other people. To cope with these problems, some methods use a small set of words, such as digits as keywords, and each user is prompted to utter a given sequence of keywords which is randomly chosen every time the system is used (Rosenberg and Soong, 1987; Higgins et al., 1991). Yet even this method is not reliable enough, since it can be circumvented with advanced electronic recording equipment that can reproduce keywords in a requested order. Therefore, a text-prompted speaker recognition method has been proposed in which password sentences are completely changed every time (Matsui and Furui, 1993). The system accepts the input utterance only when it determines that the registered speaker uttered the prompted sentence. Because the vocabulary is unlimited, prospective impostors cannot know in advance the sentence they will be prompted to say. This method not only accurately recognizes speakers, but can also reject an utterance whose text differs from the prompted text, even if it is uttered by a registered speaker. Thus, the playback of a recorded voice can be correctly rejected.

7 Normalization and Adaptation Techniques

How can we normalize intraspeaker variation of likelihood (similarity) values in speaker verification? The most significant factor affecting automatic speaker recognition performance is variation in signal characteristics from trial to trial (intersession variability or variability over time). Variations arise from the speakers themselves, from differences in recording and transmission conditions, and from noise. Speakers cannot repeat an utterance in precisely the same way from trial to trial. It is well known that samples of the same utterance recorded in one session are much more correlated than tokens recorded in separate sessions. There are also long-term trends in voices with variation over several months and years (Furui et al., 1972; Furui, 1974).

It is important for speaker recognition systems to accommodate these variations. Adaptation of the reference model as well as the verification threshold for each speaker is indispensable to maintain a high recognition accuracy for a long period. In order to compensate for the variations, two types of normalization techniques have been tried: one in the parameter domain, and the other in the distance/similarity domain. The latter technique uses the likelihood ratio or a posteriori probability. To adapt HMMs for noisy conditions, various techniques, including the HMM composition (or parallel model combination: PMC) method (Gales and Young, 1993), have proved successful.

7.1 Parameter-Domain Normalization

As one typical normalization technique in the parameter domain, spectral equalization, the "blind equalization" method, has been confirmed to be effective in reducing linear channel effects and long-term spectral variation (Atal, 1974; Furui, 1981). This method is especially effective for text-dependent speaker recognition applications using sufficiently long utterances. In this method, cepstral coefficients are averaged over the duration of an entire utterance, and the averaged values are subtracted from the cepstral coefficients of each frame (cepstral mean subtraction: CMS). This method can compensate fairly well for additive variation in the log spectral domain. However, it unavoidably removes some text-dependent and speaker-specific features, so it is inappropriate for short utterances in speaker recognition applications. Time derivatives of cepstral coefficients (delta-cepstral coefficients) have been shown to be resistant to linear channel mismatches between training and testing (Furui, 1981; Soong and Rosenberg, 1988).

7.2 Likelihood Normalization

Higgins et al. (1991) proposed a normalization method for distance (similarity or likelihood) values that uses a likelihood ratio. The likelihood ratio is the ratio of the conditional probability of the observed measurements of the

utterance, given the claimed identity is correct, to the conditional probability of the observed measurements, given the speaker is an impostor (normalization term). Generally, a positive log-likelihood ratio indicates a valid claim, whereas a negative value indicates an impostor. The likelihood ratio normalization approximates optimal scoring in Bayes' sense.

This normalization method is, however, unrealistic because conditional probabilities must be calculated for all the reference speakers, which requires large computational cost. Therefore, a set of speakers, "cohort speakers", who are representative of the population distribution near the claimed speaker, was chosen for calculating the normalization term (Rosenberg et al., 1992). Another approximation of using all the reference speakers is to use speakers who are typical of the general population. Reynolds (1994) reported that a randomly selected, gender-balanced background speaker population outperformed a population near the claimed speaker.

Matsui and Furui (1993, 1994) proposed a normalization method based on a posteriori probability. The difference between the normalization method based on the likelihood ratio and that based on *a posteriori* probability is whether or not the claimed speaker is included in the impostor speaker set for normalization. The cohort speaker set in the likelihood-ratio-based method does not include the claimed speaker, whereas the normalization term for the a posteriori probability-based method is calculated by using a set of speakers including the claimed speaker. Experimental results indicate that both normalization methods almost equally improve speaker separability and reduce the need for speaker-dependent or text-dependent thresholding, compared with scoring using only the model of the claimed speaker.

Carey and Paris (1992) proposed a method in which the normalization term is approximated by the likelihood for a "world model" representing the population in general. This method has the advantage that the computational cost for calculating the normalization term is much smaller than in the original method since it does not need to sum the likelihood values for cohort speakers. Matsui and Furui (1994) proposed a method based on tied-mixture HMMs in which the world model is made as a pooled mixture model representing the parameter distribution for all the registered speakers. The use of a single background model for calculating the normalization term has become the predominate approach used in speaker verification systems.

Since these normalization methods neglect absolute deviation between the claimed speaker's model and the input speech, they cannot differentiate highly dissimilar speakers. Higgins et al. (1991) reported that a multilayer network decision algorithm can make effective use of the relative and absolute scores obtained from the matching algorithm.

A family of normalization techniques has recently been proposed, in which the scores are normalized by subtracting the mean and then dividing by

standard deviation, both terms having been estimated from the (pseudo) impostor score distribution. Different possibilities are available for computing the impostor score distribution: Znorm, Hnorm, Tnorm, Htnorm, Cnorm, and Dnorm (Bimbot et al., 2004). The state-of-the-art text-independent speaker verification techniques combine one or several parameterization level normalizations (CMS, feature variance normalization, feature warping, etc.) with a world model normalization and one or several score normalizations.

8 Evaluation of Speaker Recognition Systems

8.1 Evaluation of Speaker Verification Systems

The receiver operating characteristic (ROC) curve adopted from psychophysics is used for evaluating speaker verification systems. In speaker verification, two conditions are considered for the input utterances: s, the condition that the utterance belongs to the customer, and n, the opposite condition. Two decision conditions also exist: S, the condition that the utterance is accepted as being that of the customer, and N, the condition that the utterance is rejected.

These conditions combine to make up the four conditional probabilities as shown in Table 2. Specifically, $P(S|s)$ is the probability of correct acceptance; $P(S|n)$ the probability of false acceptance (FA), namely, the probabilities of accepting impostors; $P(N|s)$ the probability of false rejection (FR), or the probability of mistakenly rejecting the real customer; and $P(N|n)$ the probability of correct rejection.

Since the relationships

$$P\,(S\,|s\,) + P\,(N\,|s\,) = 1 \tag{1.17}$$

and

$$P\,(S\,|n\,) + P\,(N\,|n\,) = 1 \tag{1.18}$$

exist for the four probabilities, speaker verification systems can be evaluated using the two probabilities $P(S|s)$ and $P(S|n)$. If these two values are assigned to the vertical and horizontal axes respectively, and if the decision criterion (threshold) of accepting the speech as being that of the customer is varied, ROC curves as indicated in Figure 4 are obtained. This figure exemplifies the curves for three systems: A, B, and C. Clearly, the performance

Table 2. Four conditional probabilities in speaker verification.

Decision condition	Input utterance condition	
	$s(customer)$	$n(impostor)$
$S(accept)$	$P(S\|s)$	$P(S\|n)$
$N(reject)$	$P(N\|s)$	$P(N\|n)$

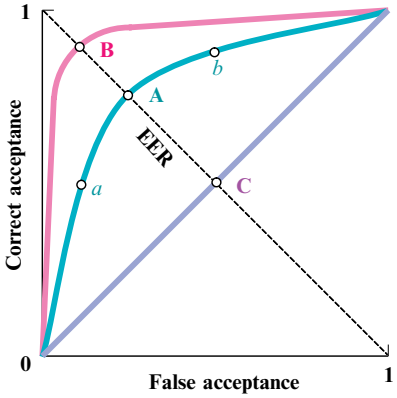

Figure 4. Receiver operating characteristic (ROC) curves; performance examples of three speaker verification systems: A, B, and C.

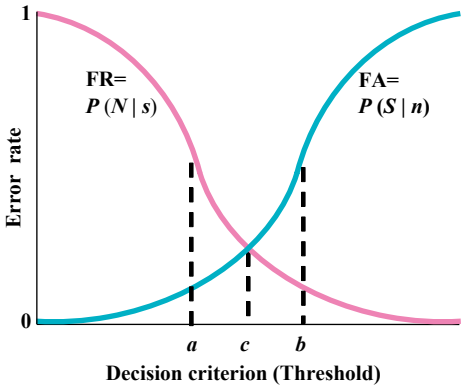

Figure 5. Relationship between error rate and decision criterion (threshold) in speaker verification.

of curve B is consistently superior to that of curve A; and C corresponds to the limiting case of purely chance performance. On the other hand, the relationship between the decision criterion and the two kinds of errors is presented in Figure 5. A "tight" decision criterion makes it difficult for impostors to be falsely accepted by the system. However, it increases the possibility of rejecting customers. Conversely, a "loose" criterion enables customers to be consistently accepted, while also falsely accepting impostors. Position a in Figures 4 and 5 corresponds to the case in which a strict decision criterion is employed, and position b corresponds to that wherein a lax criterion is used. To set the threshold at the desired level of FR and FA, it is necessary to know the distribution of customer and impostor scores as baseline information. The decision criterion in practical applications should be determined according to the

consequences or risk of decision errors. This criterion can be determined based on a priori probabilities of a match, $P(s)$, on the cost values of the various decision results, and on the slope of the ROC curve. If the FR rate is specified, the corresponding FA rate is obtained as the intersection of the ROC curve with the vertical line indicating the FR rate. In experimental tests, equal error rate (EER), is a commonly accepted summary of system performance. It corresponds to a threshold at which the FR rate is equal to the FA rate as indicated by c in Figure 5. The criterion is usually set a posteriori for each individual speaker or for a set of test speakers. The EER point corresponds to the intersection of the ROC curve with the straight line of 45 degrees, indicated in Figure 4. Although the EER performance measure rarely corresponds to a realistic operating point, it is quite a popular measure of the ability of a system to separate impostors from customers. Another popular measure is the half total error rate (HTER), which is the average of the two error rates FR and FA. It can also be seen as the normalized cost function assuming equal costs for both errors.

It has recently become standard to plot the error curve on a normal deviate scale (Martin et al., 1997), in which case the curve is known as the detection error trade-offs (DETs) curve. With the normal deviate scale, a speaker verification system whose customer and impostor scores are normally distributed, regardless of variance, will result in a linear scale with a slope equal to -1. The better the system is, the closer to the origin the curve will be. In practice, the score distributions are not exactly Gaussian but are quite close to it. The DET curve representation is therefore more easily readable and allows for a comparison of the system's performances over a large range of operating conditions. Figure 6 shows a typical example of DET curves. EER corresponds to the intersection of the DET curve with the first bisector curve.

In NIST speaker recognition evaluations, a cost function defined as a weighted sum of the two types of errors has been chosen as the basic performance measure (Przybocki and Martin, 2002). This cost, referred to as the C_{DET} cost, is defined as:

$$C_{DET} = (C_{FR} \times P_{FR} \times P_C) + (C_{FA} \times P_{FA} \times (1 - P_C)) \qquad (1.19)$$

where P_{FR} and P_{FA} are FR and FA rates, respectively. The required parameters in this function are the cost of FR (C_{FR}), the cost of FA (C_{FA}), and the a priori probability of a customer (P_C).

8.2 Relationship between Error Rate and Number of Speakers

Let us assume that Z_N represents a population of N registered speakers, $X = (x_1, x_2, ..., x_n)$ is an n-dimensional feature vector representing the speech sample, and $P_i(X)$ is the probability density function of X for speaker

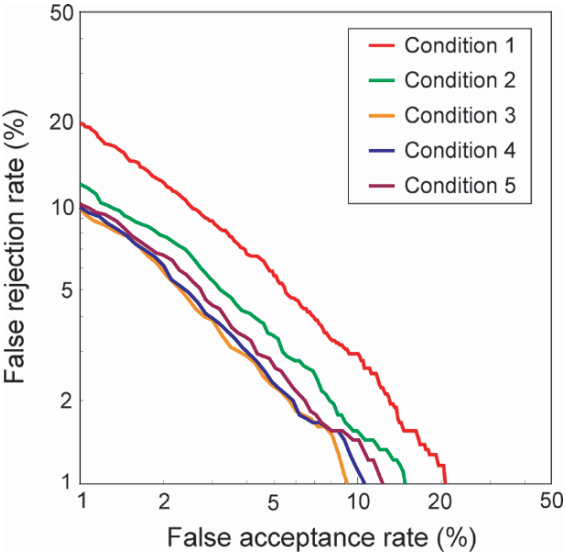

Figure 6. Examples of the DET curve.

i ($i \in Z_N$). The chance probability density function of X within population Z_N can then be expressed as

$$\begin{aligned} P_Z(X) &= \underset{i \in Z_N}{\mathrm{E}} [P_i(X)] \\ &= \sum_i P_i(X)\ Pr[i], \quad i \in Z_N \end{aligned} \tag{1.20}$$

where $Pr[i]$ is the a priori chance probability of speaker i and E indicates expectation (Doddington, 1974).

In the case of speaker verification, the region of X which should be accepted as the voice of customer i is

$$R_{Vi} = \{ X \,|P_i(X) \quad > \quad C_i P_Z(X) \} \tag{1.21}$$

where C_i is chosen to achieve the desired balance between FA and FR errors. With Z_N constructed using randomly selected speakers, and with the a priori probability independent of the speaker, $Pr[i] = 1/N$, $P_Z(X)$ will approach a limiting density function independent of Z_N as N becomes large. Thus, FR and FA ratios are relatively unaffected by the size of the population, N, when it is large. From a practical perspective, $P_Z(X)$ is assumed to be constant since it is generally difficult to estimate this value precisely, and

$$R_{Vi} = \{ X \mid P_i(X) \quad > \quad k_i \} \tag{1.22}$$

is simply used as the acceptance region.

With speaker identification, the region of X, which should be judged as the voice of speaker i, is

$$R_{Ii} = \{\, X \mid P_i(X) \quad > \quad P_j(X)\,, \;\; \forall j \neq i \,\} \tag{1.23}$$

The probability of error for speaker i then becomes

$$P_{Ei} = 1 - \prod_{\substack{k=1 \\ k \neq i}}^{N} Pr\left(P_i(X) \quad > \quad P_k(X)\right) \tag{1.24}$$

With Z_N constructed by randomly selected speakers, the equations

$$\begin{aligned} \underset{Z_N}{\mathrm{E}}\left[P_{Ei}\right] &= 1 - \underset{Z_N}{\mathrm{E}} \left\{ \prod_{\substack{k=1 \\ k \neq i}}^{N} Pr\left(P_i(X) \quad > \quad P_k(X)\right) \right\} \\ &= 1 - \underset{i}{\mathrm{E}} \left\{ \prod_{\substack{k=1 \\ k \neq i}}^{N} \mathrm{E}\left[Pr\left(P_i(X) \quad > \quad P_k(X)\right)\right] \right\} \\ &= -\underset{i}{\mathrm{E}} \left\{ P_{Ai}^{N-1} \right\} \end{aligned} \tag{1.25}$$

can be obtained, where P_{Ai} is the expected probability of not confusing speaker i with another speaker. Thus, the expected probability of correctly identifying a speaker decreases exponentially with the size of the population.

This is a consequence of the fact that the parameter space is bounded. Therefore, when the population of speakers increases, the probability that the distributions of two or more speakers are very close increases. Consequently, the effectiveness of speaker identification systems must be evaluated according to their targeted population size.

8.3 Long-Term Variability of Speech

As described in the previous section, even if the same words or sentences are spoken by the same speaker, speech characteristics are always varying, and there are also long-term trends. Samples recorded together in one session are much more highly correlated than those recorded in separate sessions. Therefore, the number of training sessions for making speaker models or templates and the time interval between those sessions, as well as training and testing

sessions, are important factors. Several training sessions over a long period of time help to cope with long-term variability of speech. It is crucial to leave a gap of at least several weeks between the last training session and the testing session to obtain meaningful results in evaluating speaker recognition systems.

8.4 Individual Variability

A desirable feature for a practical speaker recognition system is a reasonably uniform performance across a population of speakers. Unfortunately, it is typical to observe in speaker recognition experiments a substantial discrepancy between the best performing individuals, the "sheep", and the worst, the "goats". This problem has been widely observed, but there are virtually no studies focusing on the cause of this phenomenon. Speakers with no observable speech pathologies, and for whom apparently good reference models have been obtained, are often observed to be "goats". It is possible that such speakers exhibit large amounts of trial-to-trial variability, beyond the ability of the system to provide adequate compensation.

This means that large test sets are required to be able to measure error rates accurately. For clear methodological reasons, it is crucial that none of the test speakers, whether customers or impostors, be in the training and development sets. This excludes, in particular, using the same speakers for the background model and for the tests. It may be possible to use speakers referenced in the test database as impostors. However, this should be avoided whenever discriminative training techniques are used or if cross-speaker normalization is performed since, in this case, using referenced speakers as impostors would introduce a bias in the results.

9 Factors Affecting the Performance and Evaluation Paradigm Design for Speech and Speaker Recognition Systems

There are several factors affecting the performance of speech recognition and speaker recognition systems. First, several factors have an impact on the quality of the speech material recorded. Among others, these factors are the environmental conditions at the time of the recording (background noise etc.), the type of microphone used, and the transmission channel bandwidth and compression if any (high bandwidth speech, landline and cell phone speech, etc.). Second, factors concerning the speakers themselves (see Sections 4.3 and 8.4) and the amount of training data available affect performance. The speaker factors include physical and emotional states (under stress or ill), speaker cooperativeness, and familiarity with the system. Finally, the system performance measure depends strongly on the test set complexity. Ideally, all these factors should be taken into account when designing evaluation

paradigms or when comparing the performance of two systems on different databases. The excellent performance obtained in artificially good conditions (quiet environment, high-quality microphone, and consecutive recordings of the training and test material) rapidly degrades in real-life applications. Therefore, it is important to make an inventory of the acoustic environment in which the system is typically used. It is also important to know that for high noise conditions, such as higher than 60 dB(A), the Lombard effect (Furui, 2001) may change the level and voice of a speaker. In comparative testing, only a common subset of capabilities should be compared quantitatively.

10 System-Level Evaluation of Speech and Speaker Recognition

There are two complementary approaches to evaluate speech and speaker recognition systems: evaluation of the system components, and system-level evaluation. Evaluation of the system components can be performed using the methods described in the previous sections. Depending on the goal of evaluation, there are three broad categories of system-level evaluation (Cole et al., 1995):

1. Adequacy evaluations: determining the fitness of a system for a purpose: does it meet the requirements, and if so, how well and at what cost? The requirements are mainly determined by user needs.

2. Diagnostic evaluations: obtaining a profile of system performance with respect to possible utilization of a system.

3. Performance evaluations: measuring system performance in specific areas. There are three basic components of a performance evaluation that need to be defined prior to evaluating a system.

    - Criterion: characteristics or quality to be evaluated (e.g., speed, error rate, accuracy, learning)
    - Measure: specific system property for the chosen criterion (e.g., word accuracy)
    - Method: how to determine the appropriate value for a given measure (e.g., counting the number of substitution, insertion, and deletion errors after alignment)

For evaluation of multimodal human–computer dialogue systems, readers are recommended to refer to Dybkjær et al. (2005) and Whittaker and Walker (2005). Readers are also recommended to refer to textbooks on general designing and assessment of human–computer interaction (e.g., Dix et al., 1998).

11 Conclusion

Technology development and evaluation are two sides of the same coin; without having a good measure of progress, we cannot make useful progress. However, since the human–computer interface using speech is very complex, it is not easy to establish evaluation strategies. Although various investigations on evaluation methods have been conducted and various measures have been proposed, a truly comprehensive tool has not yet been developed. Since the target of speech recognition is now shifting from clean read speech to natural spontaneous speech contaminated by noise and distortions, evaluation of system performance is becoming increasingly difficult. The target is also shifting from recognition to understanding. Evaluation of speech understanding systems is far more difficult than that of speech recognition systems. Speech summarization is one interesting research domain that has recently emerged (Furui et al., 2004), but it is very difficult to find a way to objectively measure the quality of automatic summarization results (Hirohata et al., 2005). Thus, continued efforts are required to advance evaluation strategies for speech and speaker recognition systems.

References

Akita, Y., Nishida, M., and Kawahara, T. (2003). Automatic Transcription of Discussions Using Unsupervised Speaker Indexing. In *Proceedings of the IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 79–82, Tokyo, Japan.

Atal, B. (1974). Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification. *Journal of the Acoustical Society of America*, 55(6):1304–1312.

Bimbot, F., Bonastre, F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., and Reynolds, D. (2004). A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Applied Signal Processing*, 2004(4):430–451.

Carey, M. J. and Paris, E. S. (1992). Speaker Verification Using Connected Words. *Proceedings of Institute of Acoustics*, 14(6):95–100.

Cole, R., Mariani, J., Uszkoreit, H., Zaenen, A., and Zue, V., editors (1995). *Survey of the State of the Art in Human Language Technology*. Center for Spoken Language Understanding (CSLU), Oregon, USA.

Dix, A. J., Finlay, J. E., Abowd, G. D., and Beale, R. (1998). *Human-Computer Interaction*. Prentice Hall, London, UK, 2nd edition.

Doddington, G. (1974). Speaker Verification. Technical Report RADC 74–179, Rome Air Development Center.

Dybkjær, L., Bernsen, N. O., and Minker, W. (2005). Overview of Evaluation and Usability. In Minker, W., Bühler, D., and Dybkjær, L., editors,

Spoken Multimodal Human-Computer Dialogue in Mobile Environments, pages 221–246. Springer, Dordrecht, The Netherlands.

Furui, S. (1974). An Analysis of Long-Term Variation of Feature Parameters of Speech and its Application to Talker Recognition. *Transactions of IECE, 57-A*, 12:880–887.

Furui, S. (1981). Cepstral Analysis Technique for Automatic Speaker Verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2):254–272.

Furui, S. (1997). Recent Advances in Speaker Recognition. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication*, pages 237–252, Crans-Montana, Switzerland.

Furui, S. (2001). *Digital Speech Processing, Synthesis, and Recognition*. Marcel Dekker, New York, USA, 2nd edition.

Furui, S. (2003). Toward Spontaneous Speech Recognition and Understanding. In Chou, W. and Juang, B.-H., editors, *Pattern Recognition in Speech and Language Processing*, pages 191–227. CRC Press, New York, USA.

Furui, S., Itakura, F., and Saito, S. (1972). Talker Recognition by Longtime Averaged Speech Spectrum. *Transactions of IECE, 55-A*, 1(10):549–556.

Furui, S., Kikuchi, T., Shinnaka, Y., and Hori, C. (2004). Speech-to-text and Speech-to-speech Summarization of Spontaneous Speech. *IEEE Transactions on Speech and Audio Processing*, 12(4):401–408.

Gales, M. J. F. and Young, S. J. (1993). HMM Recognition in Noise Using Parallel Model Combination. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 2, pages 837–840, Berlin, Germany.

Gibbon, D., Moore, R., and Winski, R., editors (1998). *Spoken Language System Assessment. Handbook of Standards and Resources for Spoken Language Systems*, volume 3. Mouton de Gruyter, Berlin, Germany.

Higgins, A., Bahler, L., and Porter, J. (1991). Speaker Verification Using Randomized Phrase Prompting. *Digital Signal Processing*, 1:89–106.

Hirohata, M., Shinnaka, Y., Iwano, K., and Furui, S. (2005). Sentence Extraction-based Presentation Summarization Techniques and Evaluation Metrics. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1065–1068, Philadelphia, USA.

Hirschberg, J., Bacchiani, M., Hindle, D., Isenhour, P., Rosenberg, A., Stark, L., Stead, L., S., S. W., and Zamchick, G. (2001). SCANMail: Browsing and Searching Speech Data by Content. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 2377–2380, Aalborg, Denmark.

Hunt, M. (1990). Figures of Merit for Assessing Connected-word Recognizers. *Speech Communication*, 9:329–336.

Janin, A., Ang, J., Bhagat, S., Dhillon, R., Edwards, J., Macias-Guarasa, J., Morgan, N., Peskin, B., Shriberg, E., Stolcke, A., Wooters, C., and Wrede, B. (2004). The ICSI Meeting Project: Resources and Research. In *Proceedings of the NIST ICASSP Meeting Recognition Workshop*, Montreal, Canada.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M. (1997). The DET Curve in Assessment of Detection Task Performance. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 4, pages 1895–1898, Rhodes, Greece.

Matsui, T. and Furui, S. (1993). Concatenated Phoneme Models for Text-Variable Speaker Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 391–394, Minneapolis, USA.

Matsui, T. and Furui, S. (1994). Similarity Normalization Method for Speaker Verification Based on a Posteriori Probability. In *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pages 59–62, Martigny, Switzerland.

Möller, S. (2005). *Quality of Telephone-Based Spoken Dialogue Systems*. Springer, New York, USA.

Ney, H., Martin, S., and Wessel, F. (1997). Statistical Language Modeling Using Leaving-one-out. In Young, S. and Bloothooft, G., editors, *Corpus-based Methods in Language and Speech Processing*, pages 174–207. Kluwer Academic Publishers, The Netherlands.

Oard, D. W. (2004). Transforming Access to the Spoken Word. In *Proceedings of the International Symposium on Large-scale Knowledge Resources*, pages 57–59, Tokyo, Japan.

Przybocki, M. and Martin, A. (2002). NIST's Assessment of Text Independent Speaker Recognition Performance. In *Proceedings of the Advent of Biometrics on the Internet, A COST 275 Workshop*, Rome, Italy. http://www.nist.gov/speech/publications/index.htm.

Rabiner, L. R. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall, New Jersey, USA.

Reynolds, D. (1994). Speaker Identification and Verification Using Gaussian Mixture Speaker Models. In *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pages 27–30, Martigny, Switzerland.

Rosenberg, A. E., DeLong, J., Lee, C.-H., Juang, B.-H., and Soong, F. (1992). The Use of Cohort Normalized Scores for Speaker Verification. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 599–602, Banff, Canada.

Rosenberg, A. E. and Soong, F. K. (1987). Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes. *Computer Speech and Language*, 22:143–157.

Rosenberg, A. E. and Soong, F. K. (1991). Recent Research in Automatic Speaker Recognition. In Furui, S. and Sondhi, M. M., editors, *Advances in Speech Signal Processing*, pages 701–737. Marcel Dekker, New York, USA.

Soong, F. K. and Rosenberg, A. E. (1988). On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-36(6):871–879.

Ueberla, J. (1994). Analysing a Simple Language Model: Some General Conclusion for Language Models for Speech Recognition. *Computer Speech and Language*, 8(2):153–176.

Waibel, A. and Rogina, I. (2003). Advances on ISL's Lecture and Meeting Trackers. In *Proceedings of the IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 127–130, Tokyo, Japan.

Whittaker, S. and Walker, M. (2005). Evaluating Dialogue Strategies in Multimodal Dialogue Systems. In Minker, W., Bühler, D., and Dybkjær, L., editors, *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*, pages 247–268. Springer, Dordrecht, The Netherlands.

Young, S. (1996). A Review of Large-vocabulary Continuous Speech Recognition. *IEEE Signal Processing Magazine*, 9:45–57.