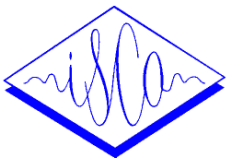

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

FLEXIBLE SPEECH RECOGNITION

Sadaoki Furui
NTT Human Interface Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo, 180 Japan
Tokyo Institute of Technology
2-12-1, Ookayama, Meguro-ku, Tokyo, 152 Japan
furui@splab.hil.ntt.jp

ABSTRACT

This paper overviews the main methods that have recently been investigated for making speech recognition systems more flexible at both the acoustic and linguistic processing levels. Improved flexibility will enable such systems to work well over a wide range of unexpected and adverse conditions by helping them to cope with variations between training and testing speech utterances. This paper focuses on the Bayesian adaptive learning approach, the minimum classification error (MCE) approach, the HMM composition technique, and spontaneous speech recognition techniques.

1. INTRODUCTION

In speech recognition, we always encounter discrepancies between training data and test utterances. These discrepancies are the major factors degrading the performance of speech recognition systems, especially in practical situations. Even if a speech recognition system performs remarkably well in laboratory evaluations and during demonstrations to prospective clients, it often performs not nearly as well in the "real world".

There are two reasons for discrepancies. First, speech signals are subject to variations resulting from linguistic variations as well as such acoustic variables as speaker individuality, environment-dependent speaking styles, additive noise, and microphone and transmission characteristics. Second, it is not practical to collect a large set of speech and text data spoken and written by a large population over all possible combinations of signal conditions.

Flexible speech recognition refers to the problem of designing an automatic speech recognizer that works well over a wide range of unexpected and adverse conditions. There are two major approaches to improving flexibility. One is to apply statistical modeling techniques (e.g., HMM) to characterize the basic speech units and use an integrated training/testing paradigm (invariant methods). The other is to apply automatic adaptation techniques (adaptive methods). Invariant methods assume no explicit knowledge of the testing signal environment, while the adaptive methods estimate some of the characteristics and adjust the signals or models accordingly to achieve a reliable pattern-matching result.

Recognition performance under adverse conditions is often impaired by variations in the degree of speech quality degradation rather than by the degradation itself. Problems are created, for example, by the variation in noise level associated with variations in the distance between the speaker and the microphone. If training can be done under the same noise conditions as those under which the speech is to be recognized, performance is better than that attained when the training is done under noise-free conditions. Conversely, if training is done under noisy conditions and the speech to be recognized is clean, performance is worse than when noisy speech is recognized.

If the ranges of the noise characteristics and speaking manner (such as speaking rate, loudness, and Lombard effect) are known ahead of time, training under the various expected conditions is useful [Lippmann et al., 1987]. This method is limited, however, because it is impossible to train under every condition.

It would therefore be useful to have methods for automatically adapting to and normalizing the effects of adverse conditions. Possible adaptation schemes include equalization, normalization, and adaptation of the signals, features, and models. In this paper I will focus on model adaptation techniques.

2. SPEECH VARIATION

Figure 1 shows the main causes of speech variation resulting from the speech production processes [Cole et al., 1995] Although the physical phenomena of speech variation can be classified as either noise addition or distortion, the distinction between these categories is not clear. The actual phenomena resulting from variation are

(1) Spectral variation
(2) Nonlinear time expansion and contraction
(3) Additive noise, which increases the difficulty of determining the speech period.

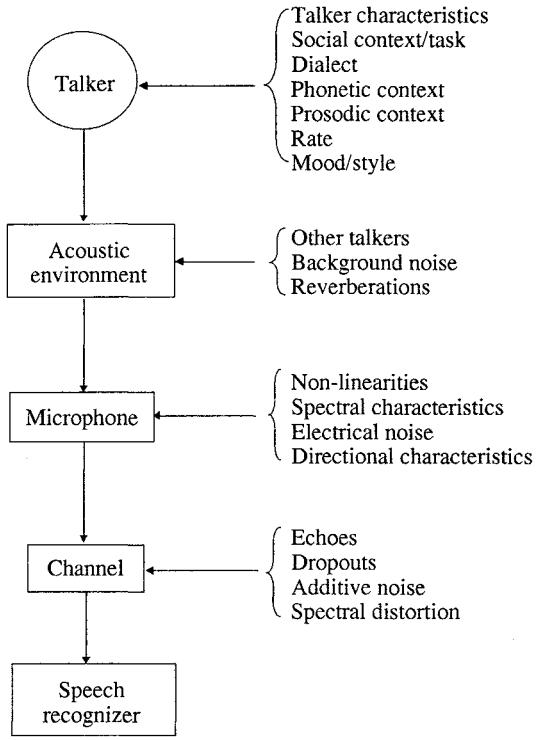

Fig. 1 Main causes of speech variation.

When people speak in a noisy environment, not only does the loudness (energy) of their speech increase, but the pitch and frequency components also change. These speech variations are called the Lombard effect [Pisoni et al., 1985; Junqua et al., 1990]. Several experimental studies have indicated that these indirect noise influences have a greater effect on speech recognition than does the direct influence of noise entering microphones [Rajasekaran et al.; 1986, Lippman et al., 1987]. It has also been reported that articulation effects under noisy conditions are context dependent [Junqua et al., 1990].

Figure 2 shows the major methods that have been investigated to cope with speech variation problems [Juang, 1990, 1991; Furui, 1992b], along with the basic sequence of speech recognition processes. Several methods have been used to deal with additive noise: using special microphones, using auditory models for speech analysis and feature extraction, reducing and suppressing noise, using noise masking and adaptive models, using spectral distance measures that are robust against noise, and compensating for spectral deviation resulting from the special speaking manners used in noisy environments (Lombard effect). Various methods have also been used to cope with the problems caused by the different characteristics of different kinds of microphones.

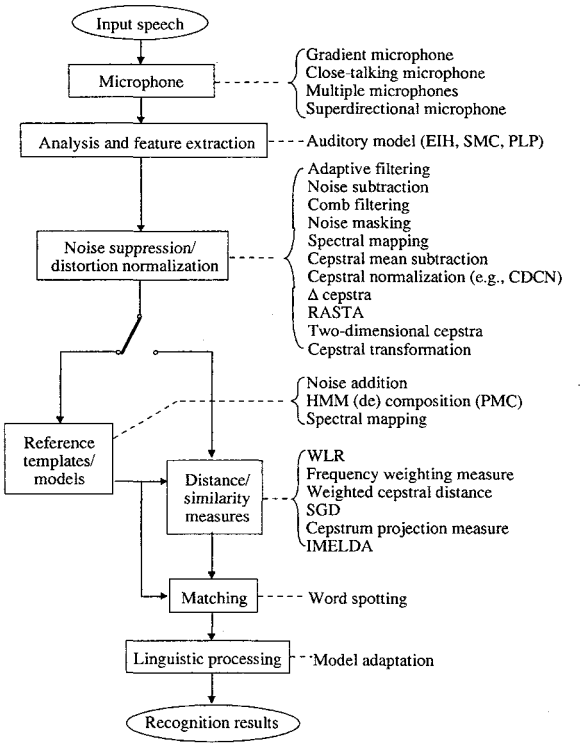

Fig. 2 Major methods for coping with speech variation problems in speech recognition.

Because methods for coping with individual speech variation have already been surveyed [Furui, 1991, 1992a], I will just briefly review the major speaker-independent methods used in speech recognition, concentrating especially on speaker adaptation techniques. Readers wanting more detail should refer to the original papers.

Discourse recognition using spontaneous speech has

recently occupied the attention of many researchers. When recognizing spontaneous speech in dialogues, it is necessary to deal with variations that are not encountered when recognizing speech that is read from text. These variations include out-of-vocabulary words, ungrammatical sentences, botched utterances, restarts, repetitions, style shifts, and instability in the detection of end-points [Furui, 1995a, 1995b].

Another problem related to the variation of speech is how to adapt to a new task. Most advanced speech recognition systems use statistical approaches consistently for acoustic and linguistic processing. To build reliable statistical models, it is necessary to use large speech and language databases. These databases must be changed every time the tasks are changed, but it is difficult to collect a large database for each task. Adapting the model to new tasks using a relatively small database would be useful.

This paper introduces the major techniques that have recently been introduced for automatic adaptation of model parameters. They include Bayesian adaptive learning, the minimum classification error (MCE) approach, and the HMM composition technique. This paper also addresses spontaneous speech recognition problems. For other techniques, readers should refer to the review papers by Juang [1990, 1991] and Furui [1992b].

3. BAYESIAN ADAPTIVE LEARNING

The Bayesian learning framework offers a way to incorporate newly acquired application-specific data into existing models and combine them in an optimal manner. It is therefore an efficient technique for handling the sparse training data problem which is typical in adaptive learning of model parameters [Lee et al., 1995]. Its principle has been used to derive maximum a posteriori (MAP) estimates of the parameters of speech models, including HMM parameters.

The HMM parameters are usually estimated using the maximum likelihood (ML) approach. Although ML estimation has good asymptotic properties, it often requires a large amount of training data to achieve reliable results.

For a given set of training/adaptation data $\mathbf{x}$, the ML estimation assumes that the HMM parameter λ is fixed but unknown and that it satisfies

$$\lambda_{\mathrm{ML}} = \arg\max_{\lambda} f(\mathbf{x}|\lambda), \tag{1}$$

where $f(\mathbf{x}|\lambda)$ is the likelihood of observing $\mathbf{x}$. On the other hand, the MAP formulation assumes λ to be a random vector with a certain distribution. Before making any new observations, the parameter vector is assumed to have a priori density $g(\lambda)$; when new data $\mathbf{x}$ are incorporated, the parameter vector is characterized by a posteriori density $g(\lambda|\mathbf{x})$. The MAP estimate maximizes the a posteriori density:

$$\lambda_{\mathrm{MAP}} = \arg\max_{\lambda} g(\lambda|\mathbf{x}) = \arg\max_{\lambda} f(\mathbf{x}|\lambda)g(\lambda). \tag{2}$$

Since the parameters of the a priori density can also be estimated from an existing HMM λ, λ_0, this framework provides a way to combine λ_0 with newly acquired data $\mathbf{x}$ in an optimal manner.

The a priori density can be used to impose constraints on the values of the parameters. If the parameter is unknown, a priori density $g(\lambda)$, is a constant for the entire parameter region of interest. The MAP estimate obtained by solving eq. (2) is therefore equivalent to the ML estimate obtained by solving eq. (1). When the a priori density of the HMM parameters is assumed to be the product of the conjugate a priori densities for all HMM parameters, the MAP estimates can be solved with the expectation-maximization (EM) algorithm [Gauvain et al., 1994].

The MAP estimates can usually be expressed as a weighted sum of two components; one depends on the information in the a priori density (i.e., λ_0) and the other depends on the new set of adaptation data. It can further be shown that the MAP and the ML estimates are asymptotically equivalent.

The MAP-based adaptive learning algorithms have been applied to a number of applications [Lee et al., 1991; Gauvain et al., 1992; Lee et al., 1993; Matsuoka et al., 1993], including speaker and task adaptation, context adaptation, corrective training, parameter smoothing, speaker group modeling, on-line incremental adaptation, and N-gram and histogram probability smoothing and adaptation. The same approach is expected to be applicable to the problems of transducer and channel adaptation.

4. MCE/GPD APPROACH

In contrast to conventional ML and MAP training, which estimates a model based only on training utterances from the same category, discriminative training takes into account the models of other competing categories and formulates the optimization criterion so that category separation is enhanced and the classification/recognition error rate for the training data is directly minimized. The optimization solution is obtained with a generalized probabilistic descent algorithm. This method is therefore called the MCE (minimum classification error)/GPD (generalized probabilistic descent) method. Unlike the Bayesian framework, this method does not require estimating the probability distributions, which usually cannot be reliably obtained. This method has been applied in various experimental studies for both speech and speaker recognition with good results [Katagiri et al., 1992].

This method can also be used in model adaptation: it was recently confirmed that this method is effective for speaker adaptation [Lin et al., 1994; Matsui et al., 1995].

5. HMM COMPOSITION / PMC

The hidden Markov model (HMM) composition/ parallel model combination (PMC) method creates a noise-added-speech HMM by combining HMMs that model speech and noise [Gales et al., 1992; Martin et al., 1993]. This method is closely related to the HMM decomposition proposed by Varga et al. [1990, 1991]. In HMM composition, observation probabilities (means and covariances) for a noisy speech HMM are estimated by convoluting observation probabilities in a linear spectral domain. Figures 3 and 4 show the HMM composition process. Since a noise HMM can usually be trained by using input signals without speech, this method can be considered as an adaptation process where speech HMMs are adapted on the basis of the noise model.

This method can be applied not only to stationary noise but also to time-varying noise, such as another speaker's voice. The effectiveness of this method was confirmed by experiments using speech signals to which noise or other speech had been added. The experimental results showed that this method produced recognition rates similar to those of HMMs trained by using a large noise-added speech database.

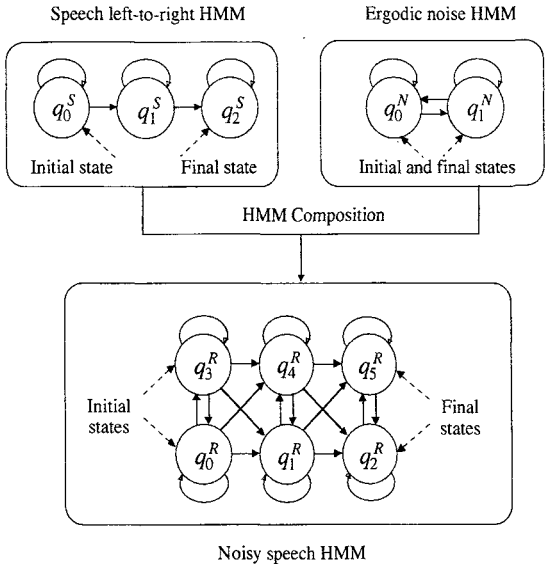

Fig. 3 HMM composition process for creating a noisy speech HMM as a product of two source HMMs.

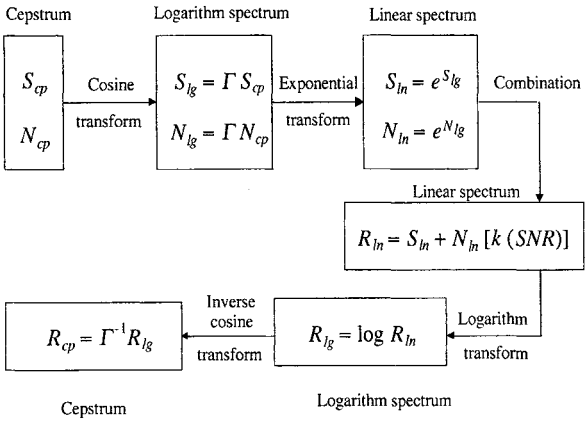

Fig. 4 Transformations in HMM composition process.

This method has recently been extended to simultaneously cope with additive noise and multiplicative distortion [Gales et al., 1993; Minami et al., 1995]. Figure 5 shows a universal model for producing noisy and distorted speech; speech signal S is produced by speech HMMs and noise signal N is produced by a noise HMM. Both S and N are defined in the linear-power spectral domain. Scalar parameter k corresponds to the signal-to-noise ratio (SNR). First, S is multiplied by (multiplicative) distortion G, which could include speaking styles and speaker characteristics. Then (additive) noise N is multiplied by k and

added to speech signal SG. Finally, the noise-added speech signal is multiplied by (multiplicative) distortion H, which includes channel distortion.

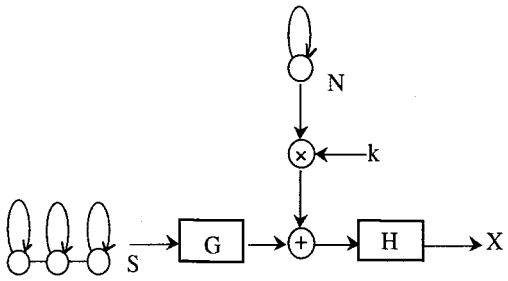

Fig. 5 Universal model for producing noisy and distorted speech.

We thus obtain the final noisy and distorted speech signal: X (= H(GS + kN) = HGS + kHN). By setting W = HG, we get X = WS + kHN. Since absolute gain does not make any difference in the likelihood calculation in the recognition stage, the speech signal can be rewritten as X' = W'S + HN, where W'=W/k. The basic noisy speech model can thus be converted into the model shown in Fig. 6.

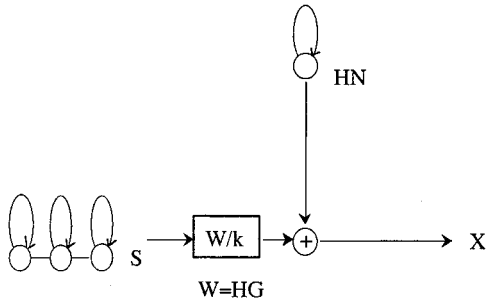

Fig. 6 Conversion of the universal model producing noisy and distorted speech.

The HMM for HN can be trained by using a signal recorded for a period without speech. The HMMs for S can be made from noise-free speech. To estimate the value of W'=W/k, we model X' by combining the HMMs for HN and W'S by using the HMM composition method. W' is thus estimated by maximizing likelihood score P(X|M(W')) for the adaptation data given composed HMMs, M(W'), as a function of W'. For convenience, parameter k corresponding to the SNR is initially estimated using the parallel method shown in Fig. 7. In this method, several sets of models having different W (W') are prepared. From these models, a set of models having the maximum likelihood is selected. W' is then estimated by using the steepest descent method.

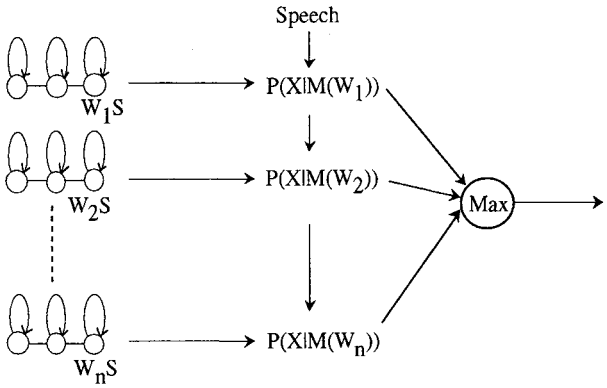

Fig. 7 Parallel method for estimating filter W.

Recognition experiments confirmed that this method greatly improves recognition rates for noisy and distorted speech. The efficiency and flexibility of the algorithm and its adaptability to new noises make it suitable as the basis for a speech recognizer that is flexible to wide variations in conditions.

6. SPEAKER ADAPTATION

Approaches to building speaker-independent recognition systems include [Furui, 1992a]:

(1) Signal processing to extract invariant features
(2) Information theoretical approaches (such as those using discriminant methods)
(3) Pattern recognition (including multiple-template methods)
(4) Artificial intelligence (knowledge engineering)
(5) Statistical approaches (for example, using HMMs)
(6) Neural networks.

The performance of speaker-independent speech-recognition systems has recently been greatly improved by training them with a large speech database spoken by

many speakers and by incorporating statistical models of speech variations. It is still impossible, however, to accurately recognize the utterances of every speaker. A small percentage of people occasionally cause systems to produce exceptionally low recognition rates. This is an example of the "sheep and goats" phenomenon. Experiments have shown that people can adapt to a new speaker's voice after hearing just a few syllables [Kato et al., 1985]. Recent research has therefore explored the possibility of giving recognition systems the ability to automatically adapt to individual speakers [Furui, 1991].

Speaker-adaptation methods are generally classified either as supervised (text-dependent) methods, in which training words or sentences are known, or as unsupervised (text-independent) methods, in which arbitrary utterances can be used. The major speaker-adaptation methods that have recently been investigated are

(1) Speaker cluster selection
(2) Interpolated re-estimation (e.g., Bayesian learning)
(3) Codebook mapping (adaptation/normalization)
(4) Reference-template generation.

In the framework of Bayesian learning, speaker adaptation can be viewed as adjusting speaker-independent models to form speaker-specific ones, using the available a priori information and a small amount of speaker-specific adaptation data. The a priori densities are simultaneously estimated during the speaker-independent training process along with estimating the parameters for the speaker-independent models. Experimental results showed that the speaker-adapted models perform better than speaker-dependent models when relatively small amounts of data were used for training or adaptation. When a large amount of training data was used, the results were comparable, consistent with the Bayesian formulation that the MAP estimate asymptotically converges to the ML estimate [Gauvain et al., 1992].

7. COPING WITH SPONTANEOUS-SPEECH-SPECIFIC PROBLEMS

As described in Section 2, the recognition of spontaneous speech poses many difficult problems that do not occur in the recognition of speech that is read. One of the most important issues is how to create language models (rules) for spontaneous speech. It is crucial to develop robust and flexible parsing algorithms that match the characteristics of spontaneous speech. How to extract contextual information, predict users' responses, and focus on key words are very difficult and important issues.

A word spotting technique is frequently used to determine the speech period in spontaneous speech - that is, for separating speech from stationary and nonstationary noises in a real environment. Nonstationary noise includes such speech and speech-like sounds as coughing, restarting, and the repetition of utterances. The word spotting method extracts target words from continuous speech by using the acoustical and temporal characteristics of these words in a top-down manner [Wilpon et al., 1984].

In spontaneous speech, correct sentences are not always spoken, and obstacle words and sounds that are unnecessary for understanding the utterances are frequently added. Additionally, the system is requested to respond to the utterance as quickly as possible. To solve these problems, it is necessary to establish a method for detecting the time at which sufficient information has been acquired instead of detecting the end of input speech.

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

8. ADAPTATION TO NEW TASKS AND NEW LANGUAGES

Speech recognition systems basically consist of acoustic and linguistic models, and both models change their characteristics according to the recognition task. For acoustic models, task-independent phone models have been explored [Hon et al., 1991]. However, because automatic training using utterances without phone labels is now possible, it is not difficult to use large speech databases to make task-dependent acoustic models that can be changed every time a task changes.

The adaptation of linguistic models, in contrast, is still a very important issue. This is because a linguistic database needs to be much bigger than an acoustic database, and collecting a large linguistic database for a new task is difficult and costly. Various research efforts have

therefore explored the adaptation of linguistic models [Kuhn et al., 1990; Matsunaga et al., 1992].

It is also desirable that recognition algorithms designed for one language be flexible enough to be easily adapted to other languages.

9. IMPROVED ADAPTATION

Important practical issues in using adaptation techniques include the specification of a priori parameters (information), the availability of supervision information, and the amount of adaptation data needed to achieve effective learning. Since observation of all the phoneme units enough times is unlikely in a small adaptation set, especially in large-vocabulary continuous-speech recognition systems, only a small number of parameters can be effectively adapted. It is therefore desirable to introduce some parameter correlation or tying so that all model parameters can be adjusted at the same time in a consistent manner, even if some units are not included in the adaptation data.

If a correlation structure between parameters can be established and the correlation parameters can be estimated when training the general models, the parameters of unseen units can be adapted accordingly [Furui, 1980; Cox, 1993]. To improve adaptation efficiency and effectiveness along this line, many new techniques have recently been proposed. They include hierarchical spectral clustering and smoothing [Furui, 1989; Ohkura et al., 1992], spectrum bias and shift transformation, cepstral normalization [Acero et al., 1990], probabilistic spectral mapping, acoustic bias normalization and context bias modulation, and stochastic matching.

The second type of improvement is through a set of constraints on the model parameters, so that all the parameters are adjusted simultaneously according to a predetermined set of constraints, e.g., multiple regression analysis [Furui, 1980]. Various methods have recently been proposed in which a linear transformation between the reference and adaptive speaker feature vectors is defined and then translated into a bias vector and a scaling matrix, which can be estimated with an EM algorithm [Zhao, 1993; Bellegarda et al., 1994; Leggetter et al., 1994; Digalakis et al., 1995; Sankar et al., 1995; Zavaliagkos et al., 1995].

The third type of improvement is on-line and incremental adaptation [Matsuoka et al., 1993; Zavaliagkos et al., 1995]. In this approach, adaptation is performed at runtime on the testing data in an unsupervised manner. This method is especially important when the environment, noise, distortion, or speaker vary during the recognition.

10. DISCUSSION

This paper briefly reviewed the major approaches for improving the performance of speech recognition systems under real conditions - when speech signals are subject to many different kinds of variation. Considerable success has been attained by using these approaches. However, continued research on new models of the operating conditions as well as solutions to enhance recognizer flexibility is still necessary to ensure the maximal deployment and usefulness of speech recognition systems.

To correctly evaluate speech recognition technologies and to achieve steady technical progress, it is important to comprehensively evaluate technologies under actual field conditions instead of under controlled laboratory conditions. It is also important to evaluate and improve the technologies from the viewpoint of the human interface, for which evaluation under actual and adverse conditions is essential. It is therefore important to collect databases under field conditions (actual or simulated) and to use these databases to compare the effectiveness of the proposed methods. It is also necessary to clarify the appropriate application areas for each major method, to set forth guidelines for combining these methods, and to improve the most promising methods.

REFERENCES

Acero, A. and Stern, R. M. (1990): "Environmental robustness in automatic speech recognition," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Albuquerque, S15b.11, pp. 849-852.

Bellegarda, J. R., De Sousa, P. V., Nadas, A. J., Nahamoo, D., Picheny, M. A. and Bahl, L. R. (1994): "The Metamorphic Algorithm, A Speaker Mapping Approach to Data Augmentation," IEEE Trans. Speech and Audio Processing, Vol. 2, No. 3, pp. 413-420.

Cole, R. et al. (1995): "The challenge of spoken language

systems: Research directions for the nineties," IEEE Trans. Speech, Audio Processing, Vol. 3, No. 1, pp. 1-21.

Cox, S. J. (1993): "Speaker Adaptation Using a Predictive Model," Proc. Eurospeech, Berlin, Vol. 3, pp. 2283-2286.

Digalakis, V. and Neumeyer, L. (1995): "Speaker Adaptation Using Combined Transformation and Bayesian Methods," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Detroit, pp. I-680-683.

Furui, S. (1980): "A Training Procedure for Isolated Word Recognition Systems," IEEE Trans. Acoust., Speech Signal Processing, Vol. 28, No. 2, pp. 129-136.

Furui, S. (1989): "Unsupervised Speaker Adaptation Method Based on Hierarchical Spectral Clustering," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Glasgow, pp. 286-289.

Furui, S. (1991): "Speaker-dependent-feature extraction, recognition and processing techniques," Speech Communication, Vol. 10, Nos. 5-6, pp. 505-520.

Furui, S (1992a): "Speaker-independent and speaker-adaptive recognition techniques," in *Advances in Speech Signal Processing*, edited by S. Furui and M. M. Sondhi, pp. 597-622.

Furui, S. (1992b): "Toward robust speech recognition under adverse conditions," Proc. ESCA Workshop on Speech Processing in Adverse Conditions, Cannes-Mandelieu, France, pp. 31-42.

Furui, S. (1995a): "Prospects for spoken dialogue systems in a multimedia environment," Proc. ESCA Workshop on Spoken Dialogue Systems, Vigso, Denmark, pp. 9-16.

Furui, S. (1995b): "Recent advances in speech processing technology," Proc. ICA, Trondheim.

Gales, M. J. F. and Young, S. J. (1992): "An improved approach to the hidden Markov model decomposition of speech and noise," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, San Francisco, pp. 233-236.

Gales, M. J. F. and Young, S. J. (1993): "Parallel model combination for speech recognition in noise," Technical Report CUED/F-INFENG/TR135.

Gauvain, J.-L. and Lee, C.-H. (1992): "Bayesian Learning for Hidden Markov Models with Gaussian Mixture State Observation Densities," Speech Communication, Vol. 11, Nos. 2-3, pp. 205-214.

Gauvain, J.-L. and Lee, C.-H. (1994): "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," IEEE Trans. Speech and Audio Processing, Vol. 2, No. 2, pp. 291-298.

Hon, H. W. and Lee, K. F. (1991): "CMU robust vocabulary-independent speech recognition system," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto, S14.3, pp. 889-892.

Juang, B. H. (1990): "Recent developments in speech recognition under adverse conditions," Proc. Int. Conf. Spoken Language Processing, Kobe, 25.1, pp. 1113-1116.

Juang, B.-H. (1991): "Speech recognition in adverse environments," Computer Speech and Language, Vol. 5, pp. 275-294.

Junqua, J. C. and Anglade, Y. (1990): "Acoustic and perceptual studies of Lombard speech: Application to isolated-words automatic speech recognition," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Albuquerque, S15b.9, pp. 841-844.

Katagiri, S., Lee, C.-H. and Juang, B.-H. (1992): "New discriminative algorithm based on the generalized probabilistic descent method," Proc. IEEE Workshop on Neural Networks for Signal Processing, Princeton, pp. 299-309.

Kato, K. and Furui, S. (1985): "Listener adaptability for individual voice in speech perception," Trans. Committee of Hearing Research, H85-5.

Kuhn, R. and De Mori, R. (1990): "A cache-based natural language model for speech recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 12, No. 6, pp. 570-583.

Lee, C.-H. and Gauvain, J.-L. (1995): "Bayesian adaptive learning and MAP estimation of HMM," in *Advanced Topics in Automatic Speech and Speaker Recognition*, edited by C.-H. Lee, K. K. Paliwal and F. K. Soong, Kluwer Academic Publishers (to be published).

Lee, C.-H., Lin, C.-H. and Juang, B.-H. (1991): "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-39, No. 4, pp. 806-814.

Lee, C.-H. and Gauvain, J.-L. (1993): "Speaker Adaptation Based on MAP Estimation of HMM Parameters," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. II-652-655.

Leggetter, C. J. and Woodland, P. C. (1994): "Speaker Adaptation of Continuous Density HMMs Using Linear Regression," Proc. Int. Conf. Spoken Language Processing, Yokohama.

Lin, C.-H., Chang, P.-C. and Wu, C.-H. (1994): "An initial study on speaker adaptation for Mandarin syllable recognition with minimum error discriminative training," Proc. Int. Conf. Spoken Language Processing, Yokohama, pp. I-307-310.

Lippmann, R. P., Martin, E. A. and Paul, D. B. (1987): "Multi-style training for robust isolated-word speech recognition," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Dallas, Texas, 17.4, pp. 705-708.

Martin, F., Shikano, K. and Minami, Y. (1993): "Recognition of noisy speech by composition of hidden Markov models," Proc. Eurospeech, Berlin, pp. 1031-1034.

Matsui, T. and Furui, S. (1995): "A study of speaker adaptation based on minimum classification error training," Proc. Eurospeech, Madrid.

Matsunaga, S., Yamada, T. and Shikano, K. (1992): "Task adaptation in stochastic language models for continuous speech recognition," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, San Francisco, S25.3, pp. I-165-168.

Matsuoka, T. and Lee, C.-H. (1993): "A Study of On-line Bayesian Adaptation for HMM-Based Speech Recognition," Proc. Eurospeech, Berlin, pp. 815-818.

Minami, Y. (1995): "Universal adaptation method based on HMM composition," Proc. ICA, Trondheim.

Ohkura, K., Sugiyama, M. and Sagayama, S. (1992): "Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs," Proc. Int. Conf. Spoken Language Processing, Banff, pp. 369-372.

Pisoni, D. B., Bernacki, R. H., Nusbaum, H. C. and Yuchtman, M. (1985): "Some acoustic-phonetic correlates of speech produced in noise," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tampa, Florida, S41.10, pp. 1581-1584.

Rajasekaran, P. K., Doddington, G. R. and Picone, J. W. (1986): "Recognition of speech under stress and in noise," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo, 14.10, pp. 733-736.

Sankar, A. and Lee, C.-H. (1995): "Robust speech recognition based on stochastic matching," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Detroit, pp. 121-124.

Varga, A. P. and Moore, R. K. (1990): "Hidden Markov model decomposition of speech and noise," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Albuquerque, S15b.10, pp. 845-848.

Varga, A. P. and Moore, R. K. (1991): "Simultaneous recognition of concurrent speech signals using hidden Markov model decomposition," Proc. Eurospeech, pp. 1175-1178.

Wilpon, J. G., Rabiner, L. R. and Martin, T. (1984): "An improved word-detection algorithm for telephone-quality speech incorporating both syntactic and semantic constraints," AT&T Bell Laboratories Technical J., Vol. 63, No. 3, pp. 479-498.

Zavaliagkos, G., Schwartz, R. and Makhoul, J. (1995): "Batch, Incremental and Instantaneous Adaptation Techniques for Speech Recognition," Proc. Int. Conf. Acoust., Speech, Signal Processing, pp. I-676-679.

Zhao, Y. (1993): "A New Speaker Adaptation Technique Using Very Short Calibration Speech," Proc. Int. Conf. Acoust., Speech, Signal Processing, pp. II-592-595.