

論文 / 著書情報
Article / Book Information

Title	Recent Advances in Robust Speech Recognition
Authors	SADAOKI FURUI
Citation	ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, Vol. , No. , pp. 11-20
Pub. date	1997, 4

RECENT ADVANCES IN ROBUST 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.co.jp, furui@cs.titech.ac.jp

ABSTRACT

This paper overviews the main technologies that have recently been developed for making speech recognition systems more robust at both the acoustic and linguistic processing levels. These technologies are reviewed from the viewpoint of a stochastic pattern matching paradigm for speech recognition. Improved robustness enables better speech recognition over a wide range of unexpected and adverse conditions by reducing mismatches between training and testing speech utterances. This paper focuses on supervised vs. unsupervised adaptation techniques, the Bayesian adaptive learning approach, the minimum classification error (MCE/GPD) approach, the parallel model combination (PMC, HMM composition) technique, and spontaneous speech recognition techniques.

1. STOCHASTIC PATTERN MATCHING PARADIGM

Conversational speech production and recognition are widely modeled as a four-stage process as

shown in Fig. 1: text generation, speech production, acoustic processing, and linguistic decoding [Furui, 1989b]. These stages can be likened to an information transmission system: word sequence w is coded into label sequence x , sent through a noisy transmission channel, and then decoded into estimated sequence w' . In this paradigm, the speech utterances are considered to be stochastic patterns, so a stochastic model is used at each stage.

The speaker is modeled as a transducer, transforming into speech the thoughts (word sequences) he or she wants to communicate w , which corresponds to an information source, with his or her own modifications. The recognition system then uses acoustic processing, i.e., analysis to extract acoustic parameters x from the speech. The combined speaker and acoustic processing are regarded as an acoustic channel, corresponding to a noisy transmission channel in transmission theory.

The linguistic decoder is formulated as a *maximum a posteriori* (MAP) decoder. That is, it estimates w' by maximizing *a posteriori* probability $p(w|x)$. Based on Bayes rule, the speech decoding problem is reformulated as

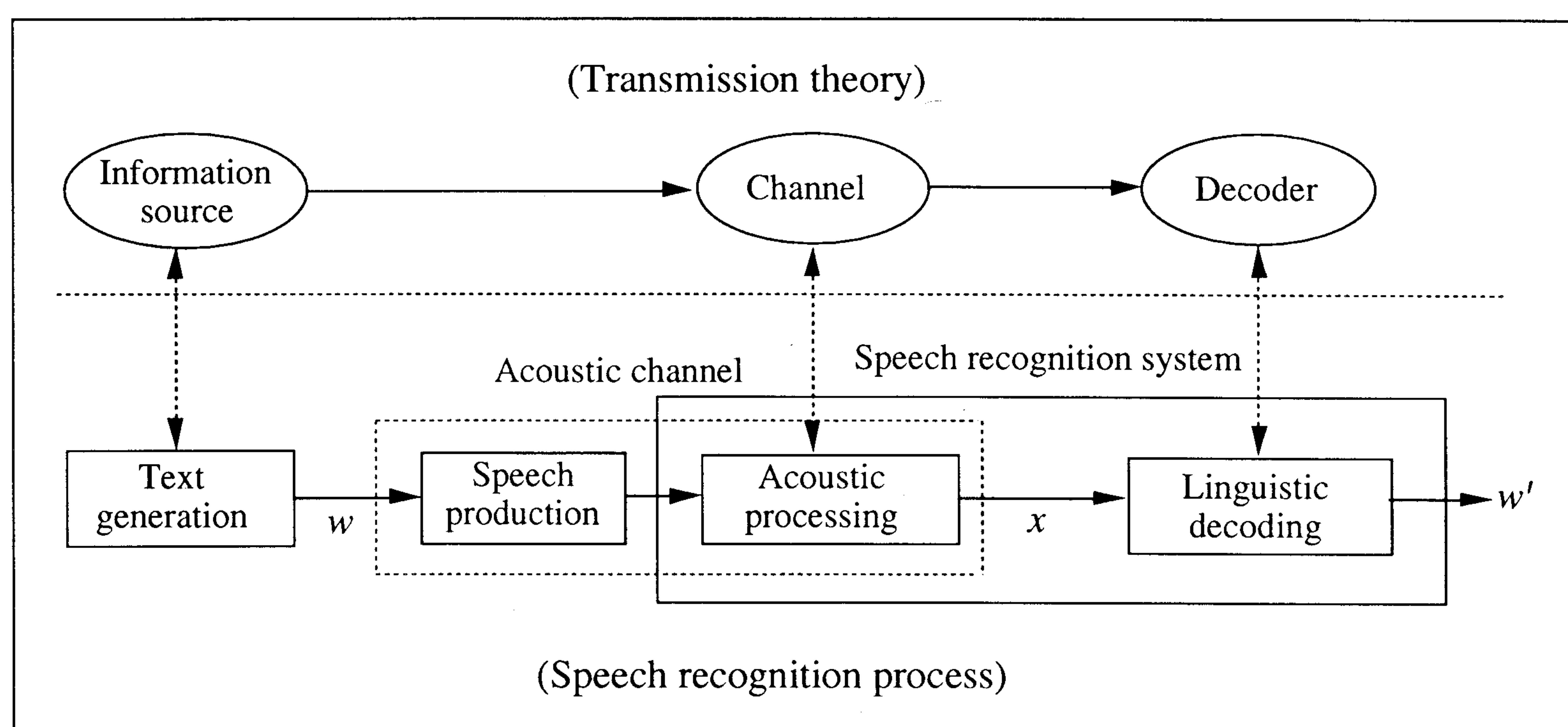


Fig. 1 - Structure of speech production and recognition system based on information transmission theory.

$$w' = \arg \max_{w \in \Gamma} p(w|x) = \arg \max_{w \in \Gamma} p(x|w)p(w) \quad (1)$$

Γ is the set of all possible word sequences, $p(x|w)$ is the conditional probability that output x is produced by the acoustic channel, given a particular sequence of words w , and $p(w)$ is the *a priori* probability that word sentence w is generated at the text generator, i.e., that w is internally constructed by the speaker for utterance. This linguistic decoder corresponds to the decoder in transmission theory, which estimates the information source by using a noisy signal. A *language model* is used to calculate $p(w)$, and an *acoustic model* is used to calculate $p(x|w)$.

Because it is not feasible to have complete knowledge about such a noisy channel, the stochastic pattern matching approach often assumes particular parametric forms for $p(x|w)$ and $p(w)$. Typically, hidden Markov models (HMMs) and stochastic language models, respectively, are used.

2. MISMATCHES BETWEEN TRAINING AND TESTING

2.1 Speech Variation

Mismatches between the training data and the test utterances always occur in speech recognition. These mismatches are the main factor degrading the performance of speech recognition systems in practical situations. Even if a speech recognition system performs remarkably well in laboratory

evaluations and during demonstrations to prospective clients, it often performs much worse in the "real world".

Two factors cause mismatches. First, speech signals are subject to acoustic and linguistic variations. The main causes of acoustic variations resulting from the speech production processes are shown in Fig. 2. They include speaker individuality, additive noise, room reverberation, and microphone and transmission characteristics. Although the physical phenomena of speech variations can be classified as either noise addition or distortion, the distinction between these two categories is not clear. (Additive noise also increases the difficulty of determining the speech period, but this problem is not addressed in this paper.) Nonlinear time expansion and contraction also occur, but these problems can be coped with by using the dynamic programming (DP) technique. 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 [Junqua et al., 1990]. The linguistic variation includes environment-dependent speaking styles.

The second factor causing mismatches is an incorrect modeling assumption and/or an insufficiency of training data to accurately estimate the model parameters. It is not practical to collect a huge set of speech and text data spoken and written by a large population over all possible

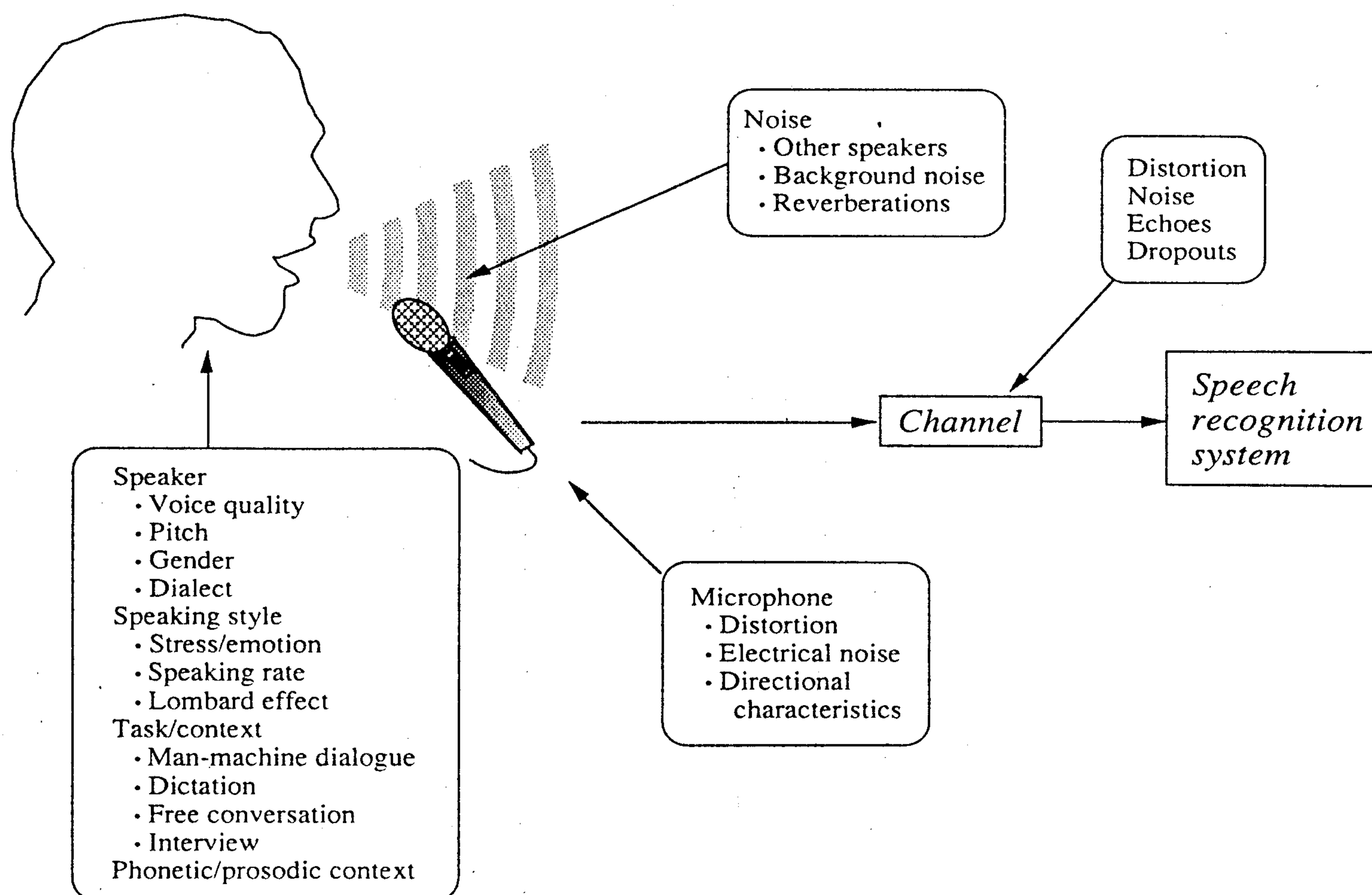


Fig. 2 - Main causes of acoustic variation in speech.

combinations of signal conditions.

Recognition performance under adverse conditions is often affected by the degree of speech quality degradation rather than by the degradation itself. Problems are created, for example, by a mismatch in noise levels 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, better performance is attained than 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, the performance is worse than when noisy speech is recognized.

2.2 Inter-Speaker Variation

The performance of speaker-independent speech-recognition systems has been greatly improved by training them with a large speech database spoken by many speakers and by incorporating stochastic models of speech variations. The acoustic models can be refined by partitioning the training data according to speaker groups, then estimating separate models for each group (for example, gender-dependent models).

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 because of large mismatches between the models and the input speech. This is an example of the “sheep and goats” phenomenon. Inter-speaker variation is one of the most serious sources of variation in speech challenging today’s speech recognition technologies [Furui, 1991, 1992a; Zhao, 1995].

Differences among speakers arise from differences in the anatomy of vocal tracts and vocal cords, in dialects, in speaking idiosyncrasies, etc. Mutual effects have been observed between phonetic and speaker-specific characteristics; specifically, the differences in speaker-specific characteristics depend on the phonemes [Furui, 1986]. Experiments have shown that humans 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 enabling recognition systems to adapt to individual speakers by automatically reducing the mismatches [Furui, 1991, 1995c].

2.3 Spontaneous Speech Variation

Discourse recognition using spontaneous speech has recently captured the attention of many researchers. Recognizing spontaneous speech in dialogues requires dealing with variations in both acoustic and linguistic models 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, and style shifts [Furui, 1995a, 1995b].

2.4 Task-level Variation

Another problem related to variation in speech is adapting to new recognition tasks. As described in Sec. 1, most advanced speech recognition systems use stochastic approaches for both acoustic and linguistic processing. To build reliable stochastic models, it is necessary to use large speech and language databases. These databases must be changed every time the task is changed, but it is difficult to collect a large database for each task. It would thus be useful to be able to adapt the model to new tasks by using a relatively small database.

3. REDUCING MISMATCHES TO IMPROVE SPEECH RECOGNITION

Robust speech recognition refers to the problem of designing an automatic speech recognizer that works well over a wide range of unexpected and adverse conditions. As described in Sec. 1, the models Λ are used in speech recognition to decode acoustic parameters x by using the MAP decoder:

$$w' = \arg \max_{w \in \Gamma} p(x | w, \Lambda) p(w) \quad (2)$$

A mismatch between the training and testing environments causes a corresponding mismatch in the likelihood of x given Λ , as evaluated by Eq. 2, thus causing errors in decoded sequence w' . Decreasing this mismatch should reduce the recognition error rate [Sankar et al., 1996].

The mismatch can be viewed in the signal space, the feature space, or the model space. Since there is no significant theoretical difference in viewing the mismatch in these three spaces, we will consider the problems in the model space. Let’s consider a transformation with parameter η that maps Λ into transformed models. One approach to decreasing the mismatch between x and Λ is to find the η and the word sequence w that maximize the joint likelihood of x and w in Eq. 2, given the models Λ . We thus need to find η' and w' such that

$$(\eta', w') = \arg \max_{\eta, w} p(x, w | \eta, \Lambda) \quad (3)$$

which is equivalent to

$$(\eta', w') = \arg \max_{\eta, w} p(x | w, \eta, \Lambda) p(w) \quad (4)$$

Trying every possible combination of hypothesis w and η is very costly; three methods have been investigated to reduce this cost: Method I is an iterative method that sequentially estimates w and η , Method II is a joint maximization method that uses the N-best hypotheses, and Method III is a supervised adaptation method that uses given texts.

In Method I, a typical conventional unsupervised adaptation method, the *pseudo*-joint maximization over η and w is performed iteratively by keeping η fixed and maximizing over w , then keeping w fixed and maximizing over η (Fig. 3) [Sankar et al., 1996]. In unsupervised speaker adaptation, a speaker-independent model is used as the initial model Λ and hypothesis w maximizing the likelihood is chosen as an initial hypothesis. Since recognition hypothesis w governs the algorithm, a poor hypothesis can result in suboptimal performance, especially if the degree of freedom in the adaptation is large.

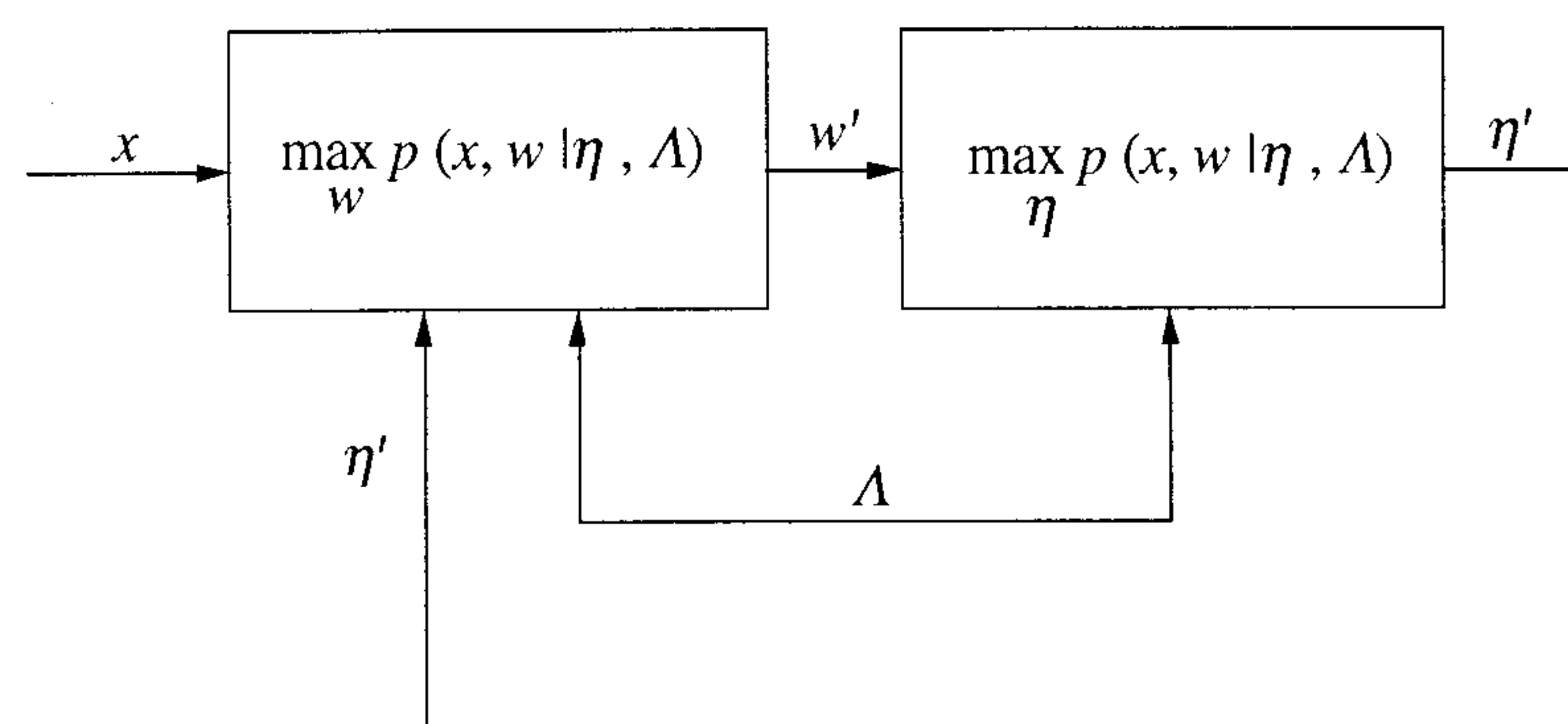


Fig.3 - Iterative procedure for maximizing joint likelihood over η and w .

In Method II, proposed by Matsui et al. [1996], the N-best hypotheses are used in parallel and the joint likelihood is iteratively maximized. This method will be described in detail in the next section.

In Method III, which is widely used for supervised adaptation (training), the texts for the set of word sequences w 's used for adaptation are given. Maximization is thus performed only for η , with the expectation that the transformed models Λ will also decrease the mismatches between unseen input speech and the models. This method has difficulty generalizing the adaptation of the models to input speech not seen in training. More research is thus needed on using these adaptation data more efficiently in creating stochastic models that are capable of dealing with different acoustic variations. A better understanding of speech signals and how they are affected by the speaker characteristics and speaking environments will help researchers develop theories and techniques for coping with this generalization problem.

In general, it is not easy to estimate η directly. However, for some model transformations, we can use the expectation-maximization (EM) algorithm for iteratively improving the current estimate and obtaining new estimates such that the likelihood increases at each iteration.

4. ADAPTIVE SPEECH RECOGNITION

4.1 Main Methods

As described in the previous sections, it is essential to introduce automatic adaptive techniques in order to make speech recognition systems robust against different variations, especially to cope with speaker-to-speaker variability. Various adaptation methods based on the principle of reducing the mismatches have been investigated.

Adaptation methods are generally classified either as supervised (text-dependent), in which training words or sentences are known, or as unsupervised (text-independent), in which arbitrary utterances are used. They are further classified into off-line and on-line modes. The off-line mode requires a new speaker to contribute a certain amount of adaptation speech; the on-line mode takes the (unknown) speech from the speaker as the adaptation data. Supervised on-line adaptation requires user feedback.

The most useful adaptation method is on-line instantaneous (unsupervised) adaptation. In this approach, adaptation is performed at runtime on the input speech in an unsupervised manner. Therefore, the recognition system does not require training speech to estimate the mismatch; it works as if it were a universal (speaker-independent) system. This method is especially useful when the speakers or environmental conditions vary frequently. The most important issue in this method is how to perform phoneme-dependent adaptation without knowing the correct model sequence for the input speech. This is especially difficult for speakers whose utterances are error prone when using universal (speaker-independent) models, that is, for speakers who definitely need adaptation.

Another useful adaptation method is on-line incremental adaptation, in which the recognition system continuously adapts to new adaptation data without using previous training data [Matsuoka et al., 1993].

The main adaptation methods currently being investigated are

- (1) Bayesian learning,
- (2) spectral mapping
- (3) linear (piecewise-linear) transformation, and
- (4) speaker (environment/noise) cluster selection.

4.2 Speaker Adaptation Methods

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 it is unlikely that all the phoneme units will be observed enough times 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.

In the spectral mapping method, speaker-adaptive parameters are estimated from speaker-independent parameters based on mapping rules. The mapping rules are estimated from the relationship between speaker-independent and speaker-dependent parameters in a supervised or unsupervised way [Furui, 1989a, 1992a; Shikano et al., 1986].

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, 1995]. To improve adaptation efficiency and effectiveness along this line, several techniques have been proposed, including probabilistic spectral mapping [Schwartz et al., 1987], hierarchical spectral clustering and smoothing [Furui, 1989a; Ohkura et al., 1992], cepstral normalization [Acero et al., 1990], and spectrum bias and shift transformation [Sankar et al., 1996].

In addition to clustering and smoothing, a second type of constraint can be given to the model parameters so that all the parameters are adjusted simultaneously according to a predetermined set of transformations, e.g., a transformation based on multiple regression analysis [Furui, 1980]. Various methods have recently been proposed in which a linear transformation (affine 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 using an EM algorithm. The transform parameters can be estimated from supervised adaptation data that form pairs with the training data [Zhao, 1994; Bellegarda et al., 1994; Leggetter et al., 1995; Sankar et al., 1995; Zavaliagos et al., 1995] or estimated from unsupervised adaptation data using existing acoustic models [Cox et al., 1989; Digalakis et al., 1995].

In the speaker cluster selection method, it is assumed that speakers can be divided into clusters, within which the speakers are similar. From many sets of phoneme model clusters representing

speaker variability, the most suitable set for the new speaker is automatically selected. Both supervised and unsupervised adaptation have been investigated. This method is useful for choosing initial models, to which more sophisticated speaker adaptation techniques are applied.

Because inter-speaker variability often interacts with other variabilities, such as allophonic contextual dependency, intra-speaker speech variation, environmental noise, and channel distortion, it is important to create methods that can simultaneously cope with these other variabilities. Inter-speaker variability is generally more difficult to cope with than noise and channel variability, since the former is non-linear whereas the latter can usually be modeled as a linear transformation in the time, spectral, or cepstral domain. Therefore, the algorithms proposed for speaker adaptation can generally be applied to noise and channel adaptation.

4.3 N-Best-Based Unsupervised Adaptation

The N-best-based unsupervised adaptation method [Matsui et al., 1996] uses the N most likely word sequences in parallel and iteratively maximizes the joint likelihood. The N-best hypotheses are created for each input speech by applying speaker-independent models; speaker adaptation based on constrained Bayesian learning is then applied to each hypothesis. Finally, the hypothesis with the highest likelihood is selected as the most likely sequence. Figure 4 shows the overall structure of such a recognition system, and Fig. 5 shows an image of likelihood surface $p(x, w | \eta, \Lambda)$ as a function of w and η . As shown in Fig. 5, conventional iterative maximization, which sequentially estimates η and w , can only reach a local maximum, whereas the N-best-based method can find a global maximum if reasonable constraints on η are applied. Without giving reasonable constraints based on models of inter-speaker variability, an input utterance can be adapted to *any* hypothesis with resulting high likelihood. To reduce this problem, constraints should be placed on the transformation so that it maintains a reasonable geometrical shape.

To achieve such a constraint, we apply constrained MAP estimation, including hierarchical spectral clustering and smoothing; we begin with a small number of bias parameters and increase the number with each iteration. This averages out the hypothesis errors for a smaller number of parameters. As the iterations continue, the hypotheses improve, allowing for an increased number of parameters.

Experimental results using Japanese connected digits recorded over a telephone network showed that the likelihood of a correct hypothesis in a low rank using speaker-independent models is

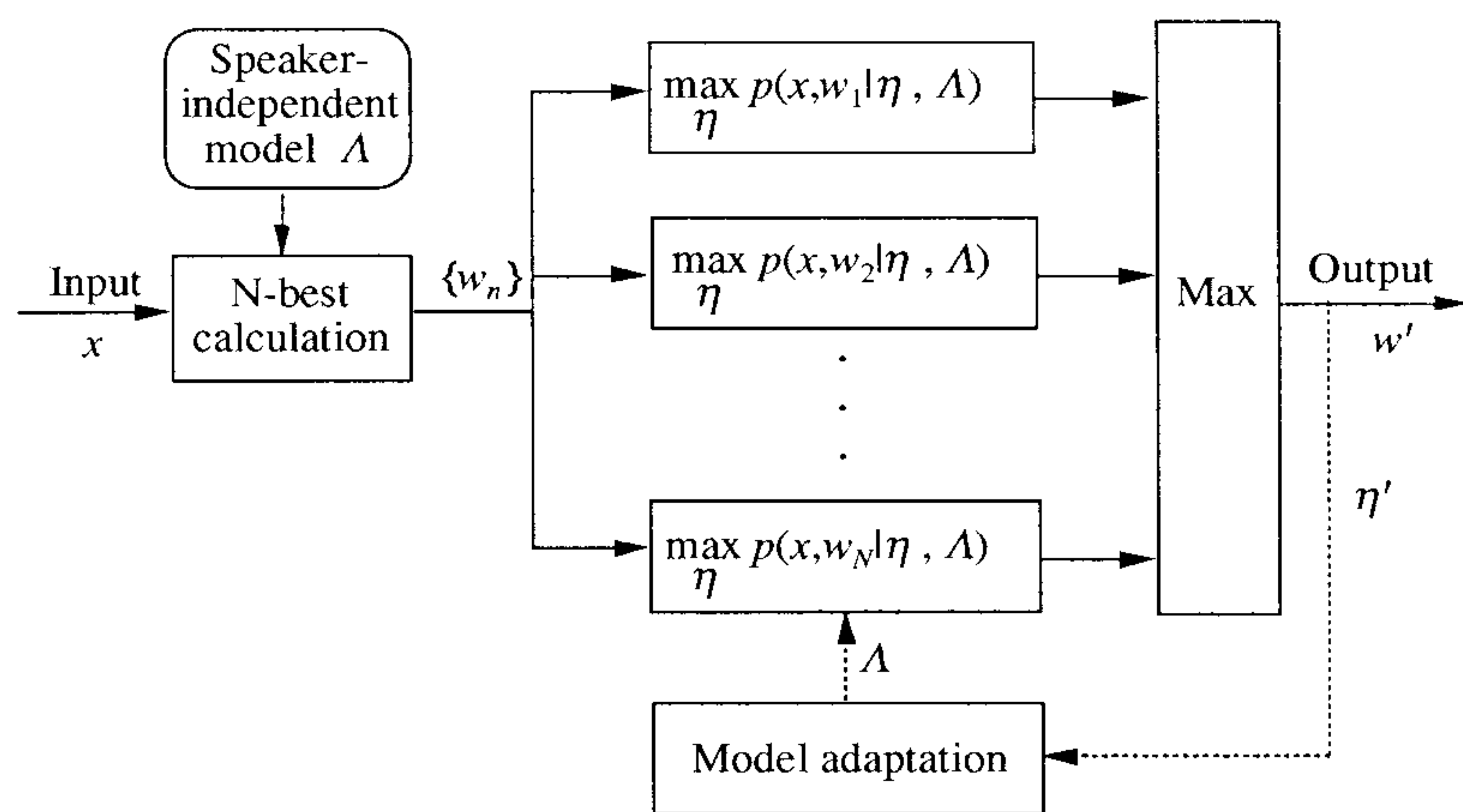


Fig. 4 - Overall structure of recognition system including N-best-based hierarchical model parameter adaptation.

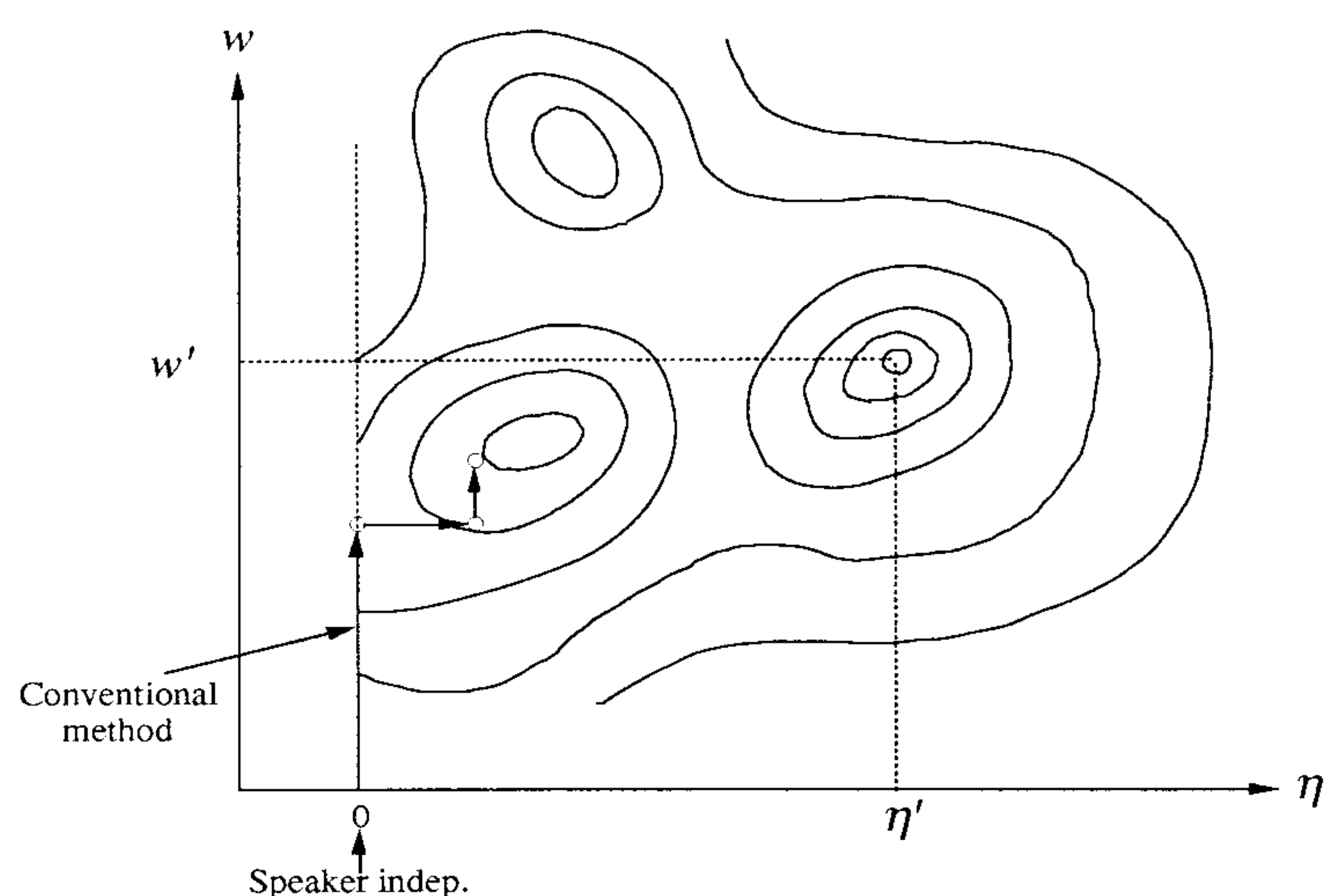


Fig. 5 - Likelihood surface $p(x, w | \eta, \Lambda)$ as a function of w and η .

increased by the adaptation, so that recognition accuracy is greatly improved even for speakers having very low accuracy. Other constraints, including linear transformation (linear regression), have also been investigated [Matsui et al., 1997].

This method can also be interpreted as a multiple-pass search algorithm, where speaker-independent and speaker-adaptive models, respectively, are used in the first and second passes.

4.4 Bayesian (MAP) Adaptive Learning

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

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 x , the ML estimation assumes that HMM parameter Λ is fixed but unknown, and that it satisfies

$$\Lambda_{\text{ML}} = \arg \max_{\Lambda} p(x | \Lambda) \quad (5)$$

where $p(x | \Lambda)$ is the likelihood of observing x .

In contrast, 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 $p(\Lambda)$; when new data x are incorporated, the parameter vector is characterized by *a posteriori* density $p(\Lambda | x)$. The MAP estimate maximizes this *a posteriori* density:

$$\Lambda_{\text{MAP}} = \arg \max_{\Lambda} p(\Lambda | x) = \arg \max_{\Lambda} p(x | \Lambda) p(\Lambda) \quad (6)$$

Because the parameters of the *a priori* density can also be estimated from an existing HMM Λ_0 , this framework provides a way to combine Λ_0 with newly acquired data x in an optimal manner.

The *a priori* density can be used to impose constraints on the values of the parameters. If a parameter is unknown, *a priori* density $p(\Lambda)$ is a constant for the entire parameter region of interest. The MAP estimate obtained by solving Eq. (6) is therefore equivalent to the ML estimate obtained by solving Eq. (5). 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 using the 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 ML estimates are asymptotically equivalent.

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; Matsui et al., 1996], 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 should be applicable to the problems of transducer and channel adaptation.

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 estimated during the speaker-independent training process while simultaneously estimating the parameters for

the speaker-independent models. Experimental results have shown that speaker-adapted models perform better than speaker-dependent models when relatively small amounts of data are used for training or adaptation. When a large amount of training data is used, the results are comparable, consistent with the Bayesian formulation that the MAP estimate asymptotically converges to the ML estimate [Gauvain et al., 1992; Lee et al., 1995].

5. ROBUST SPEECH RECOGNITION TECHNIQUES AGAINST NOISE AND CHANNEL DISTORTION

5.1 Overview

Figure 6 shows the main methods for reducing mismatches that have been investigated to cope with speech variation problems [Juang, 1990, 1991; Furui, 1992b, 1995c], along with the basic sequence of speech recognition processes. These methods can be classified into three levels: signal level, feature level, and model level.

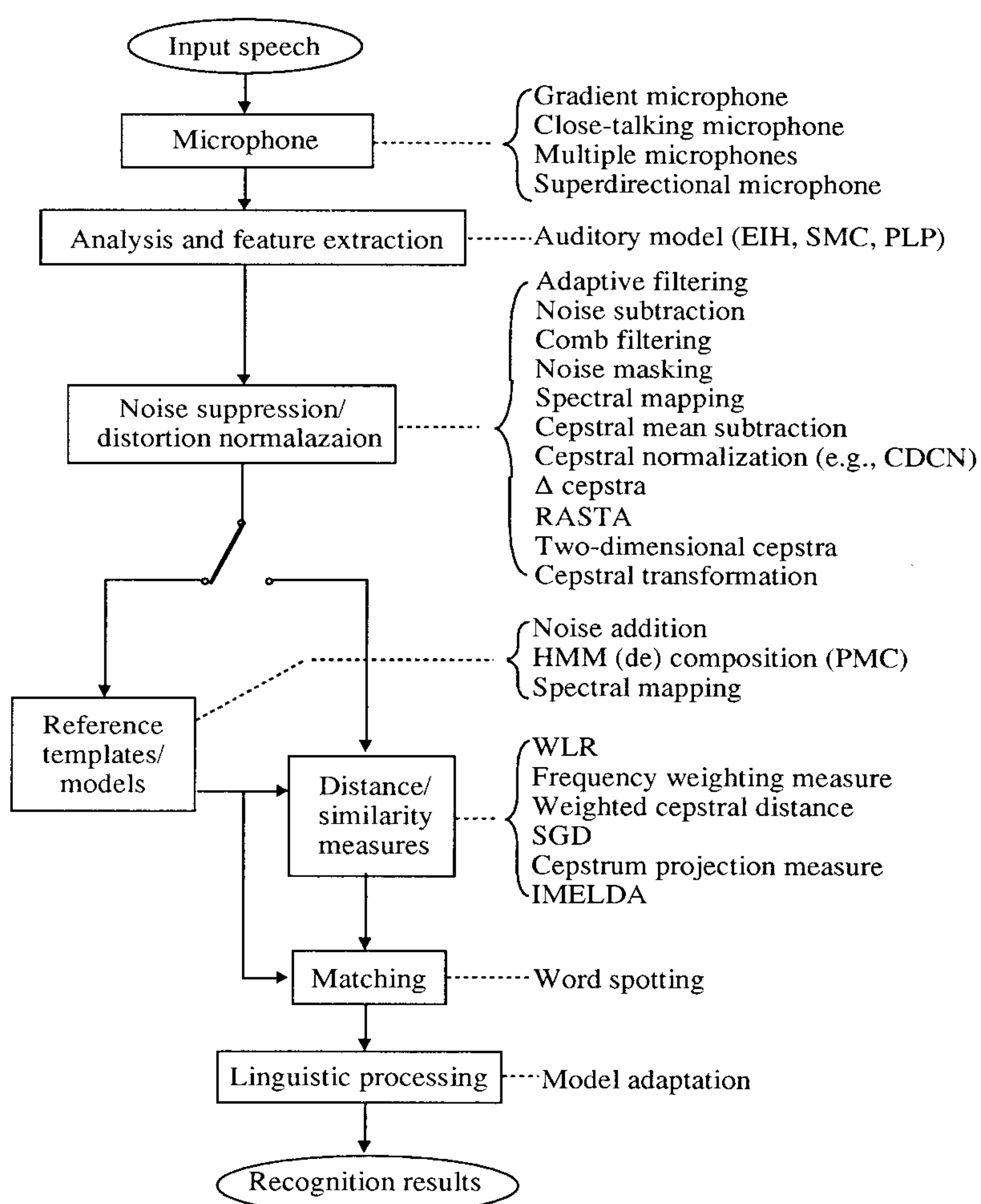


Fig. 6 - Main methods for reducing mismatches to cope with speech variation in speech recognition.

Several methods have been used to deal with additive noise: using special microphones, using auditory models for speech analysis and feature extraction, subtracting 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 differences in characteristics between different kinds of microphones and transmission lines. A commonly used method is cepstral mean subtraction (CMS), in which the long-term cepstral mean is subtracted from the utterance. This method has been extended based on Method I (Sec. 3): a database was used in which the mismatch was due to different transducers and channels. The results showed that this algorithm reduces word error rate about 70% compared to using mismatched conditions [Sankar et al., 1996].

5.2 HMM Composition / PMC

The 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 the observation probabilities in a linear spectral domain. Figures 7 and 8 show the HMM composition process. Since a noise HMM(N) can usually be trained by using input signals without speech, this method can be considered as an adaptation process where speech HMMs(S) are adapted on the basis of the noise model.

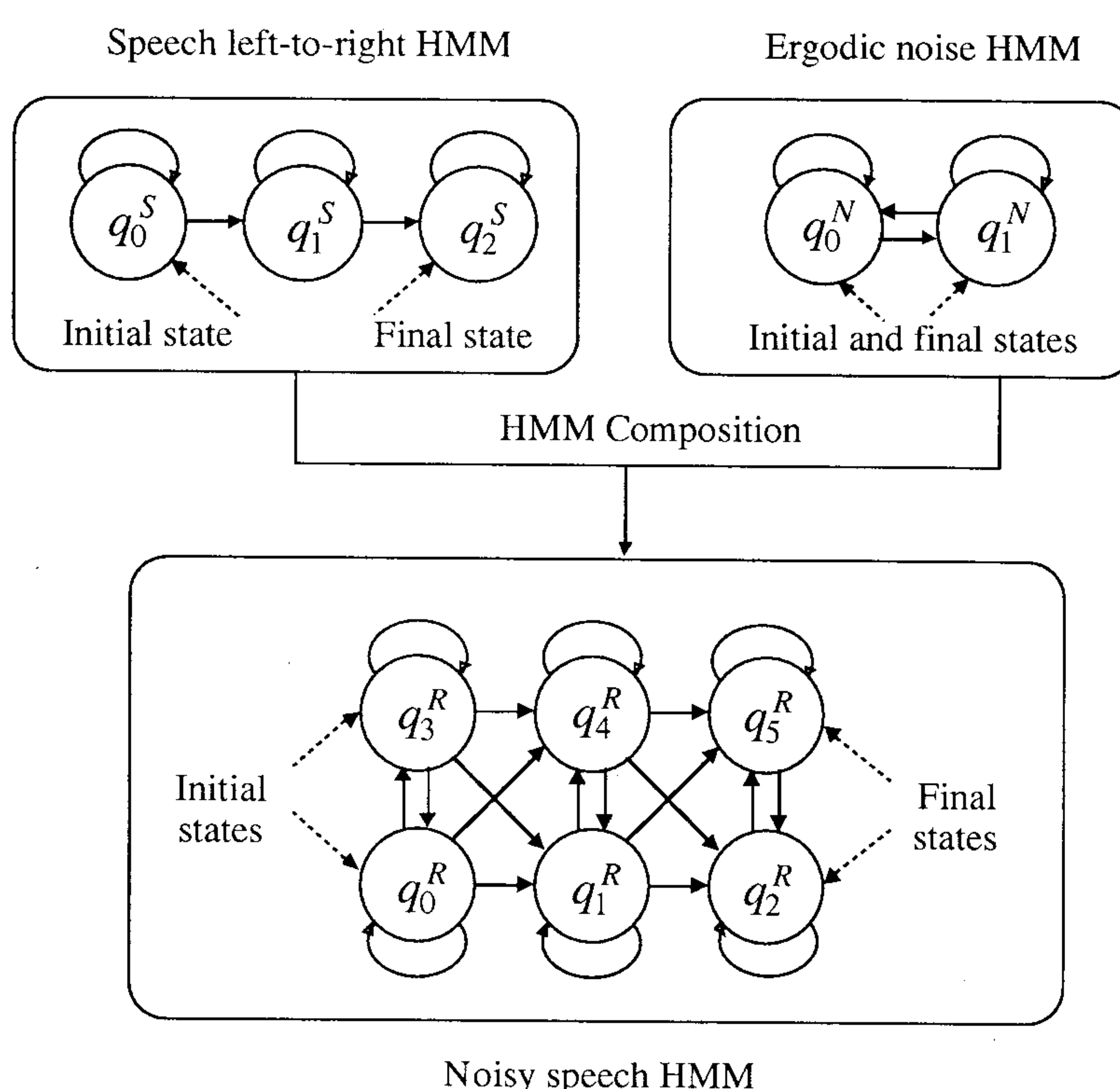


Fig. 7 - Process for creating a noisy speech HMM by combining two source HMMs.

This method can be applied not only to stationary noise but also to time-variant 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 produces recognition rates similar to those of HMMs trained by using a large noise-added speech database.

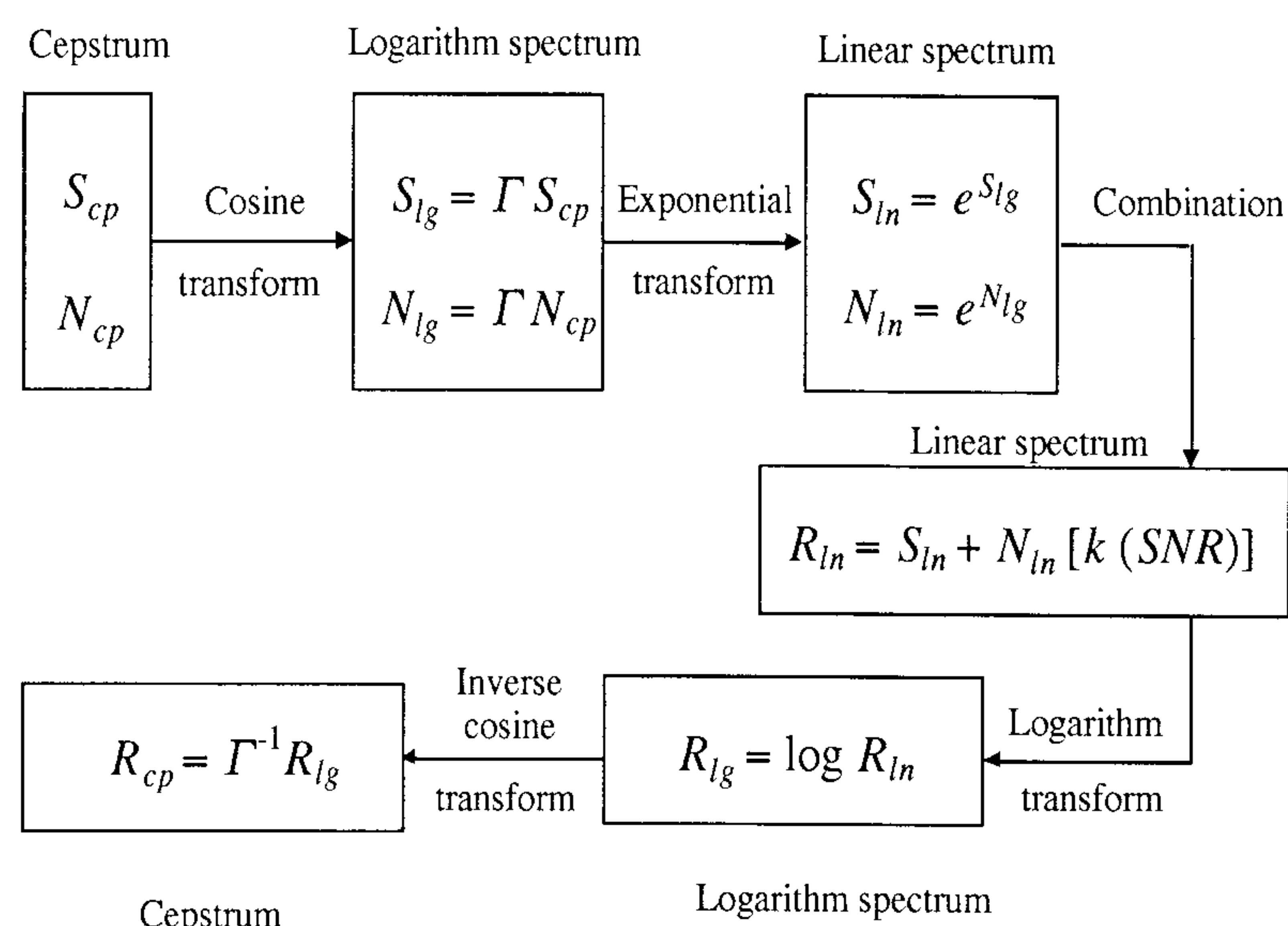


Fig. 8 - Transformations in HMM composition process.

This method has recently been extended to simultaneously cope with additive noise and convolutional (multiplicative) distortion [Gales et al., 1993; Minami et al., 1995]. Figure 9 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 can 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.

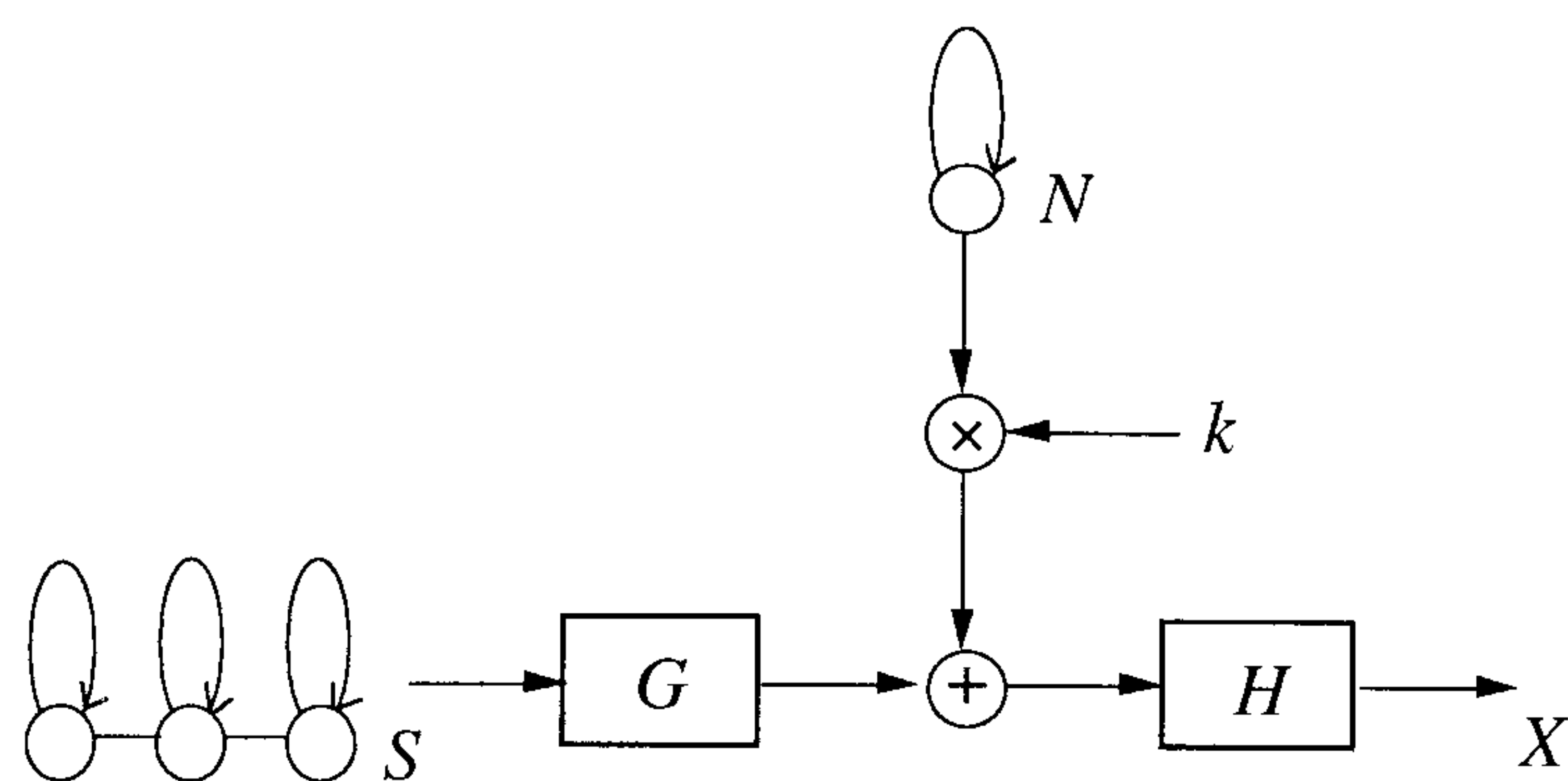


Fig. 9 - Model for producing noisy and distorted speech.

The final noisy and distorted speech signal is thus $X (= H(GS + kN) = HGS + kHN)$. By setting $W = HG$, we get $X = WS + kHN$. Since the 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 HMM for HN can be trained by using a signal recorded for a period without speech. To estimate the value of W' , we model X' by combining the HMMs for HN and $W'S$ by using the HMM composition method. The W' is thus estimated by maximizing the likelihood score $p(X|\Lambda(W'))$ for the adaptation data, given composed HMMs, i.e., $\Lambda(W')$, as a function of 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. MCE/GPD APPROACH

In contrast to conventional ML and MAP training, which estimate 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 [Katagiri et al., 1992]. The optimization solution is obtained using 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.

As has been shown for speaker adaptation [Lin et al., 1994; Matsui et al., 1995], this method can also be used for model adaptation.

7. COPING WITH LINGUISTIC MISMATCH PROBLEMS

7.1 Overcoming Spontaneous Speech-Specific Problems

As described in Sec. 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 an actual 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 should respond to utterances 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 the input speech.

7.2 Adapting to New Tasks and New Languages

Both acoustic and linguistic models for speech recognition 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 possible, it is not difficult to use large speech databases to make task-dependent acoustic models that can be changed each time a task changes. The adaptation of linguistic models, however, is not as easily solved, because a linguistic database needs to be much bigger than an acoustic database, and collecting a large linguistic database for each 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].

Adaptation to shifts in speaking style is also an important issue. In typical laboratory experiments, speakers read lists of words rather than try to accomplish an actual task. Users trying to accomplish a task usually use a different linguistic style.

Additionally, recognition algorithms designed for one language should be flexible enough to be easily adapted to other languages.

8. CONCLUSION

This paper has briefly reviewed the main approaches to improving the performance of speech recognition systems under actual conditions, in which speech signals are subject to many different kinds of variation. The approaches were reviewed from the viewpoint of reducing mismatches

between the input speech and the acoustic models in a stochastic pattern matching framework. They include the supervised/unsupervised speaker adaptation techniques, the Bayesian adaptive learning approach, the minimum classification error approach, the HMM composition technique, and spontaneous speech recognition techniques. Considerable success has been attained by using these approaches.

Because many variations can simultaneously occur under actual conditions, it is crucial to create a universal framework that can cope with them. From the human-interface point of view, it is desirable to develop an unsupervised instantaneous adaptation technique that can simultaneously reduce multiple mismatches introduced by several variation sources. A better understanding of speech signals and how they interact with speaker characteristics and speaking environments is thus needed.

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.
- Cox, S. J. and Bridle, J. S. (1989): "Unsupervised speaker adaptation by probabilistic fitting," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Glasgow, Scotland, S6.11, pp. 294-297.
- Cox, S. J. (1995): "Predictive speaker adaptation in speech recognition," Computer Speech and Language, Vol. 9, pp. 1-17.
- 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. (1986): "Research on individuality features in speech waves and automatic speaker recognition techniques," Speech Communication, Vol. 5, No. 2, pp. 183-197.
- Furui, S. (1989a): "Unsupervised speaker adaptation method based on hierarchical spectral clustering," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Glasgow, S6.9, pp. 286-289.
- Furui, S. (1989b): *Digital Speech Processing, Synthesis and Recognition*, Marcel Dekker, New York.
- 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, pp. 497-504.
- Furui, S. (1995c): "Flexible speech recognition," Proc. Eurospeech, Madrid, pp. 1595-1603.
- 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, Cambridge University.
- 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., 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, Minneapolis, pp. II-652-655.
- 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, pp. 83-107.
- Leggetter, C. J. and Woodland, P. C. (1995): "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, Vol. 9, pp. 171-185.
- 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, S07.31, pp. I-307-310.
- 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, pp. 81-84.
- Matsui, T. and Furui, S. (1996): "N-best-based instantaneous speaker adaptation method for speech recognition," Proc. Int. Conf. Spoken Language Processing, Philadelphia, pp. 973-976.
- Matsui, T., Matsuoka, T. and Furui, S. (1997): "Smoothed N-best-based speaker adaptation for speech recognition," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Munich. (To be published)
- 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, pp. 105-108.
- 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, We.fPM.1.1, pp. 369-372.
- Sankar, A. and Lee, C.-H. (1995): "Robust speech recognition based on stochastic matching," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Detroit, pp. I-121-124.
- Sankar, A. and Lee, C.-H. (1996): "A maximum-likelihood approach to stochastic matching for robust speech recognition," IEEE Trans. Speech and Audio Processing, Vol. 4, No. 3, pp. 190-202.
- Schwartz, R., Chow, Y.-L. and Kubala, F. (1987): "Rapid speaker adaptation using a probabilistic spectral mapping," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Dallas, 15.3, pp. 633-636.
- Shikano, K., Lee, K.-F. and Reddy, R. (1986): "Speaker adaptation through vector quantization," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo, 49.5, pp. 2643-2646.
- 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, Detroit, pp. I-676-679.
- Zhao, Y. (1994): "An acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition," IEEE Trans. Speech and Audio Processing, Vol. 2, No. 3, pp. 380-394.
- Zhao, Y. (1995): "Robust speaker characterization," Proc. IEEE Automatic Speech Recognition Workshop, Snowbird, pp. 101-102.