論文 / 著書情報
Article / Book Information

Author	SADAOKI FURUI
Journal/Book name	ASRU 2009, , , pp. 1-9
発行日 / Issue date	2009, 12
権利情報 / Copyright	

Generalization problem in ASR acoustic model training and adaptation

Sadaoki Furui

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

***Abstract*— Since speech is highly variable, even if we have a fairly large-scale database, we cannot avoid the data sparseness problem in constructing automatic speech recognition (ASR) systems. How to train and adapt statistical models using limited amounts of data is one of the most important research issues in ASR. This paper summarizes major techniques that have been proposed to solve the generalization problem in acoustic model training and adaptation, that is, how to achieve high recognition accuracy for new utterances. One of the common approaches is controlling the degree of freedom in model training and adaptation. The techniques can be classified by whether a priori knowledge of speech obtained by a speech database such as those spoken by many speakers is used or not. Another approach is maximizing "margins" between training samples and the decision boundaries. Many of these techniques have also been combined and extended to further improve performance. Although many useful techniques have been developed, we still do not have a golden standard that can be applied to any kind of speech variation and any condition of the speech data available for training and adaptation.**

I. INTRODUCTION

Recent advances in automatic speech recognition (ASR) can be attributed to the use of the statistical pattern recognition paradigm [12, 22, 32]. In this framework, the true joint distribution of a word sequence, W, and its corresponding sequence of acoustic vector observations, X, is assumed to be modeled by a true parametric probability density function:

$$P(W,X) = P_{\Lambda}(X|W)\ P_{\Gamma}(W), \qquad (1)$$

and the full knowledge of the parameters, Λ and Γ, of the above distributions is known. Then an optimal decoder (speech recognizer) which achieves the expected minimum word error rate and gives the recognized string, $\hat{W}$, becomes the following maximum a posteriori (MAP) decoder:

$$\hat{W} = \underset{W}{\operatorname{argmax}}\ P(W|X) = \underset{W}{\operatorname{argmax}}\ P_{\Lambda}(X|W)\ P_{\Gamma}(W). \qquad (2)$$

Since we do not know the true parametric form of $P(W, X)$ and we do not have the knowledge about its true parameter values, the parameters are estimated from a large set of labeled speech and text training data.

There is a widely known phenomenon: "There is no data like more data." Since speech implicitly contains a large number of sources of variations, we always have a data sparseness problem. Every time we start a new speech recognition task, we begin with a relatively large recognition error rate, and it decreases with the progress of the project. Figure 1 shows the progress of various DARPA projects. The decrease of the error rate for each task has been achieved by the increase of task-dependent database along with technological progress.

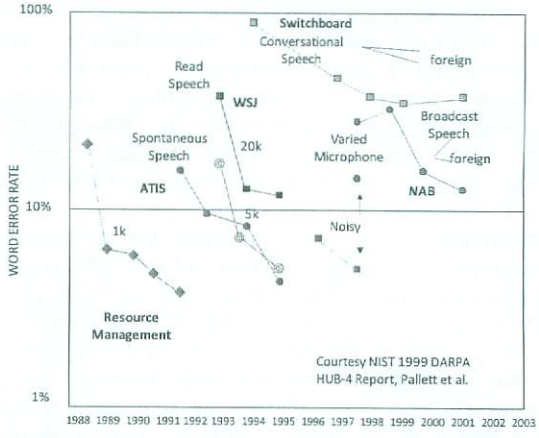

Fig. 1. History of DARPA speech recognition benchmark tests (ATIS: Airline Travel Information System, WSJ: Wall Street Journal, NAB: North American Business).

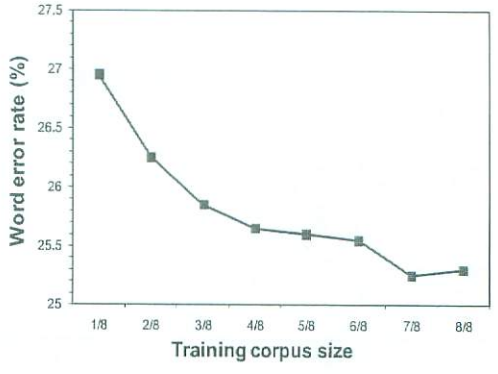

Fig. 2. Word error rate (WER) as a function of the size of acoustic model training data (8/8 = 510 hours) [13].

978-1-4244-5479-2/09/$26.00 © 2009 IEEE

Figure 2 shows the relationship between error rate and the size of the acoustic model training corpus in the case of Japanese spontaneous speech recognition using the Corpus of Spontaneous Japanese (CSJ) [13, 15]. The full size of the CSJ training corpus is approximately seven million words, and the figure shows that the error rate does not converge even when the full size is used for training. It has also been found that spontaneous speech contains substantial variations not present in read speech, and it is therefore more difficult to collect spontaneous speech data with enough coverage [15, 16]. Even if we have a large corpus, we cannot avoid the data sparseness problem in speech recognition, especially for spontaneous speech.

Figure 3 shows an example of a simple 2-category classification problem: salmon and sea-bass classification by two features of lightness and width [9]. The linear model shown in Fig. 3(a) looks too simple to achieve good classification, but the complicated model shown in Fig. 3(b) in which all the training patterns would be separated perfectly is not likely to classify new samples with high accuracy. This is the issue of *generalization*, and it is unlikely that the complex model would provide good generalization – it seems to be too much tuned to the particular training samples. This is called *over-tuning*. The model shown in Fig. 3(c) although having slightly poorer performance on the training samples is preferable for novel patterns to the models in Figs. 3(a) and 3(b).

How to avoid over-tuning and solve the generalization problem is one of the most important and also difficult issues in speech recognition. Unfortunately this problem has no theoretical solution, since in principle "there is no data like more data". One of the common empirical approaches is controlling the degree of freedom in model training. This is a version of Occam's razor, that is, the simplest model that explains data is the one to be preferred. Another approach focuses on maximizing "margins" between the well-classified training samples and the decision boundaries [9].

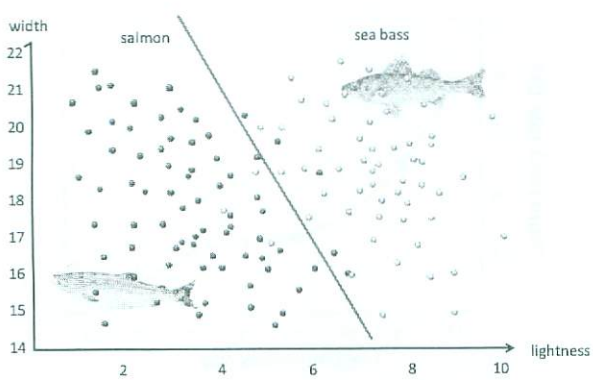

Fig. 3(a). Two features of lightness and width for sea-bass and salmon, and a linear decision boundary [9].

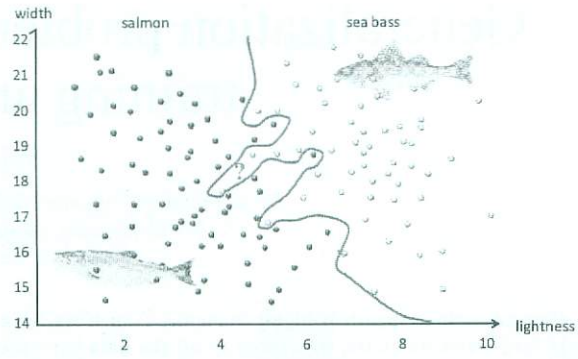

Fig. 3(b). An overly complex decision boundary which leads to perfect classification of the training samples but would lead to poor performance on future patterns. The test point marked "?" is evidently most likely a salmon, whereas the complex decision boundary shown leads it to be classified as a sea-bass [9].

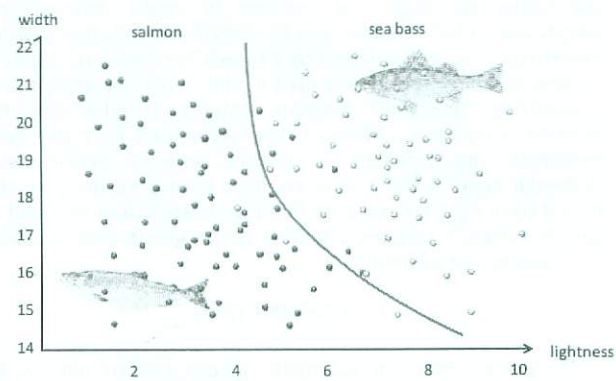

Fig. 3(c). A decision boundary which might represent the optimal tradeoff between performance on the training set and simplicity of classifier, thereby giving the highest accuracy on new patterns [9].

II. Model Adaptation

Because of the variation of speech, there always exists some mismatch which causes a distortion between the trained model and the test data. A conceptual illustration is shown in Fig. 4 [32]. D1, D2 and D3 characterize the distortion in the signal, feature and model spaces, respectively. These mismatches arise from various sources of variations as shown in Fig. 5.

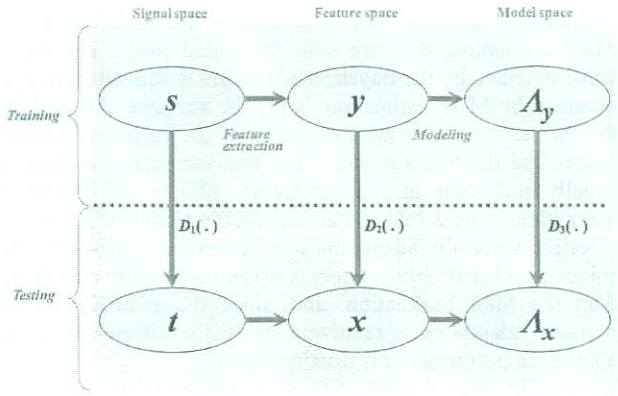

Fig. 4. Mismatch between training and testing [32].

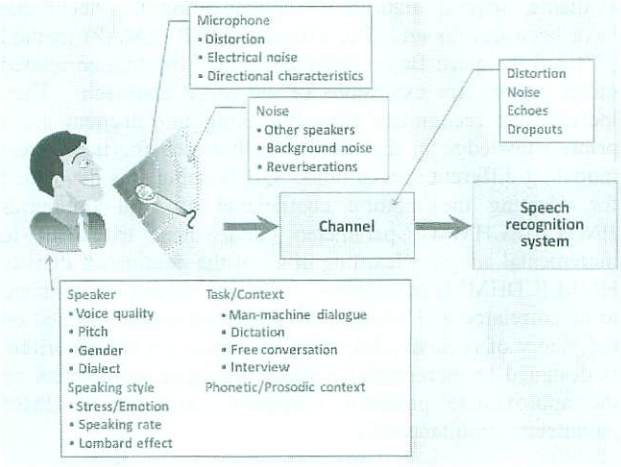

Fig. 5. Main causes of acoustic variation in speech.

In order to reduce the mismatch, two major classes of statistical techniques have been investigated [22]: adapting model parameters with labeled data, and compensating signal, feature and model distortion using testing data, among which the former class is more powerful and flexible than the latter class.

Model adaptation can be classified into supervised and unsupervised adaptation. In the supervised case, we use text or correct transcriptions for the utterances as supervision information, and it is always difficult to obtain a large number of such utterances. In the unsupervised case, since we do not have correct supervision information, we usually use ASR results, that is, recognition hypotheses, as supervision information. Unsupervised adaptation is useful, since we can use the utterances for recognition themselves as utterances for adaptation. However, since the recognition hypotheses include recognition errors, the model parameters are adapted using not only the correct labels but also the errors. Therefore, not only the limited amount of utterances but also the recognition errors included in the hypotheses cause difficulties of the unsupervised adaptation.

Unsupervised adaptation can be classified into on-line adaptation and off-line adaptation. In the on-line adaptation, the model is updated incrementally using each utterance. A common strategy for off-line adaptation is the batch-mode adaptation approach where a set of utterances for recognition are decoded and then a model is updated using all the hypotheses. The process is often iterated several times for higher recognition performance [39]. In both on-line and off-line adaptation, a problem exists in that, when the adaptation is iterated, the errors are reinforced during the iteration. This is more significant for the batch-mode adaptation, since the decoding and update steps are repeated using the same data.

Adaptation methods can also be classified into those for environmental speech variation and others. The environmental variation consists of additive noise and recording as well as transmission distortion, including microphone variation. These can generally be modeled by additive and multiplicative distortion in the spectral domain, except for room reverberation with a long time constant which cannot be modeled independently for each frame. Many useful model adaptation and compensation methods have been proposed for additive and multiplicative distortion in the spectral domain [8, 51]. Major sources of speech variation other than environmental distortion include gender, speech rate, vocal effort, regional accents, and speaking style, and therefore there is no simple model for covering these variations.

This paper targets general adaptation methods which can be applicable to a wide range of variations, mainly focusing on speaker-to-speaker variability.

III. Generalization Problem

In speech recognition, it is very important to deal with the generalization problem in both model training and adaptation to reduce the effect of hypothesis bias and allow robust estimates using a limited amount of data. Since the amount of data that we can use for adaptation is usually small and also variable from several words or a single sentence to multiple sentences, how to properly solve the generalization problem in the adaptation process is much more difficult than that in the training process.

Adaptation is usually performed based on the maximum likelihood estimation or the maximum a posteriori (MAP) estimation, but the discriminative estimation techniques [22, 24], such as minimum classification error (MCE) [28, 36], maximum mutual information (MMI) [42, 73], minimum phone error (MPE) [46, 72, 76] and closely related minimum word error (MWE), are also used. The discriminative estimation techniques are more heavily biased towards the supervision hypothesis, which causes a serious problem in unsupervised model adaptation.

How well will our system generalize to new patterns? As already described above, a major approach for generalization is controlling the degree of freedom of the model so that the trade-off between a complicated model and a constrained

model is optimized. This can be done with or without using our a priori knowledge. The methods which do not rely on using a priori knowledge can be classified into those directly controlling the degree of freedom and those using some parameter smoothing. Another relatively new approach is to maximize "margins" between the training samples and the decision boundaries.

This paper overviews techniques to deal with the generalization problem that have been investigated in acoustic model training and adaptation, mainly focusing on adaptation techniques.

IV. Constraining the Degree of Freedom by Using a Priori Knowledge

Various methods have been proposed for constraining the degree of freedom in HMM adaptation using a priori knowledge of speech. The a priori knowledge is obtained from our general knowledge about speech or from training data, such as those spoken by a large pool of speakers or under various environmental conditions.

A. VTN

Vocal tract length normalization (VTN) is one of the simplest constrained models for speaker variability, and therefore it has been widely investigated. Two methods have been commonly used; piecewise linear or bilinear warping in the frequency domain [71] and the speaker-specific Bark/Mel scale warping [33]. Although the VTN can significantly reduce the speech recognition error rate, when other general adaptation methods are applied or gender-dependent models are used as a baseline, the effectiveness of the VTN often becomes minor.

B. Correlation

Pair-wise correlation between the mean vectors can be used to enhance estimation of the mean parameters of some speech units even if they are not directly observed in the adaptation data and therefore the recognition rates are significantly improved [10]. This approach has been extended and combined with the MAP approach as described below [26].

C. MAP and Bayesian estimation

Maximum a posteriori (MAP) estimation algorithms [23] have been widely adopted and successfully applied to model adaptation, typically speaker adaptation of HMMs. In this method, the model parameters are regarded as random variables whose joint prior probability density function (pdf) is assumed. The MAP estimate of the parameter vector is defined as the mode of the posterior pdf given the adaptation data. The improvement obtained with MAP estimation is significantly larger than that obtained with maximum likelihood (ML) estimation, especially when the amount of adaptation data is relatively small. It is well known that, since MAP estimates are asymptotically equivalent to ML estimates, the resulting recognition performance is similar to that of speaker-dependent (SD) HMMs when the amount of data becomes large.

Although the Bayesian estimation is not included in the MAP estimation, they are strongly related [68]. The mean value obtained by the Bayesian estimation is equivalent to that obtained by MAP estimation, while the variance obtained by the former method is the summation of the variance of mean values and the true variance. The true variance, however, is usually unknown in real problems. Since the Bayesian estimation is more robust than the MAP estimation, it is more effective when the adaptation data is limited. However, the Bayesian estimation is computationally much more expensive than the MAP estimation, and, since the effectiveness of variance adaptation is relatively small, the difference in the adaptation performance is usually minor.

D. EMAP and QB methods

Since the MAP approach is not capable of improving recognition accuracy when only a small amount of data is available, several algorithms supplementing this technique have been developed. The extended MAP (EMAP) method [77] and the quasi-Bayes (QB) technique [26] with correlated mean vectors are extensions of the MAP approach. They increase the recognition rates by taking into account the a priori knowledge in the correlation between the parameters modeling different speech units. QB learning was developed for adapting the mixture coefficients of semi-continuous HMM (SCHMM) parameters and then extended to incremental adaptive learning of all of the continuous density HMM (CDHMM) parameters. All mean vectors are assumed to be correlated and have a joint prior distribution. Based on the theory of recursive Bayesian inference, the QB algorithm is designed to incrementally update the hyper-parameters on the approximate posterior distribution and the CDHMM parameters simultaneously.

E. Jacobian approach

The Jacobian approach is one of the analytic approaches to adapting models under an initial condition to a target condition, assuming that the variation can be analytically modeled and the difference between the two conditions is relatively small [53, 54]. In this approach, changes in the environment, such as noise, and changes in the resultant acoustic model are related by Jacobian matrices, and the adaptation is performed by simple matrix arithmetic.

F. Eigen-voice

This method was proposed for speaker adaptation using a limited amount of data for each new speaker [31, 41]. In this approach, speaker-dependent models from many speakers are created and the principal component analysis (PCA) is carried out for model parameters of all the speakers. The lower order eigen-vectors are selected as eigen-voices. For a new speaker, weights for each eigen-voice are estimated in a maximum likelihood estimation to be used for model adaptation.

G. Multiple modeling (multi-style training)

Ensemble of models specialized to specific conditions, such as gender, age, speaking rate or spontaneity, can be trained and then be used within a selection, competition or else

combination framework [2, 40]. When multiple modeling is available, all the available models may be used simultaneously during decoding, as done in many approaches, or the most adequate set of acoustic models may be selected from a priori knowledge, or their combination may be handled dynamically by the decoder. Dynamic Bayesian networks (DBN) have been used as described below to handle dependencies of the acoustic models with respect to auxiliary variables, such as local speaking rate, or hidden factors related to a clustering of the data.

The multiple models can be prepared using a clustering technique, and the optimal model for input speech is selected (cluster-based model selection) [30, 44]. Speaker clustering has been mostly employed in this scheme. Various automatic clustering techniques have been used. Clustering training data at the utterance level provides better performance than that at the lecture level [62].

H. Cluster adaptive training (CAT)

In contrast to speaker clustering where a particular cluster is selected as the speaker model, a linear interpolation of all the clusters is used in this approach [19]. To simplify the estimation process, the component weights and variances are tied over all the speaker clusters. For any particular speaker a set of interpolation values, the weight vector, is estimated. An explicit set of means per cluster or cluster dependent MLLR (maximum likelihood linear regression) transforms (see Sec V-B) of some canonical model are used. In both cases simple closed-form ML estimation can be performed. CAT is mathematically similar to the eigen-voice method, since both express the model means of the new speaker as linear combinations of some basis vectors representing "prototypical" speakers.

I. Bayesian networks

Bayesian networks provide a mechanism whereby different factorizations of a joint distribution can be specified by means of a directed graph [3, 78]. This approach can express all the details of a speech recognition system in a uniform way using only the concepts of random variables and conditional probabilities. Although a powerful set of computational routines complements the representational utility of Bayesian networks, a difficulty exists in that a huge amount of computation is needed for model parameter training to realize high flexibility in the network. So far Bayesian networks have been successfully applied to relatively small ASR tasks [29, 35, 60, 61].

V. Constraining the Degree of Freedom without Using a Priori Knowledge

The following methods have been proposed for constraining the degree of freedom in HMM adaptation without using any a priori knowledge of speech. These methods have an advantage in that they are general enough to be applied to any variation of speech, including the effects of various noise and channel distortion.

Among them, transformation-based approaches are widely used, in which the number of free parameters is limited by tying the HMM parameters or by applying some constraints on the parameters. They include cepstral mean normalization (CMN), Signal bias removal (SBR), maximum likelihood linear regression (MLLR), and vector field smoothing (VFS).

A. Structural approach

The use of structure to aid unsupervised adaptation started with the hierarchical codebook adaptation algorithm which was originally proposed for VQ-based speech coding and recognition [11, 65]. In this method, a set of spectra in adaptation speech and the reference codebook elements (centroids) are clustered hierarchically by increasing the number of clusters as shown in Fig. 6. Based on the deviation vectors between centroids of the adaptation frame clusters and the corresponding codebook clusters, adaptation is performed hierarchically from small to large number of clusters, that is, from the global variation characteristics down to the local ones. The spectral resolution of the adaptation process is therefore improved accordingly.

The proposed method was extended to continuous mixture-density HMM-based speech recognition systems [37]. The mixture-mean bias estimation, in which the biases are shared by the mixture-density distributions in the same cluster, is used for the model transformation, and the number of biases (i.e. the number of clusters) increases as the number of estimation iterations increases. The iteration is stopped when the adaptation becomes sufficiently precise.

In [58, 59], tree-based hierarchical priors were used in bottom-up supervised training of HMMs. This method has been applied to various methods, including SMAP and SMAPLR methods as described in the next section.

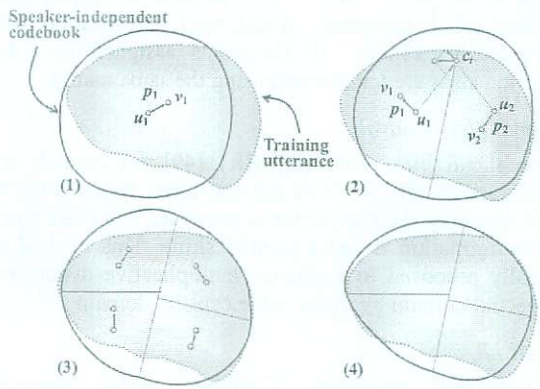

Fig. 6. Hierarchical codebook adaptation algorithm maintaining continuity between adjacent clusters [11] (u_m: centroid of the mth codebook element cluster, v_m: centroid of corresponding training speech cluster, p_m: deviation vector between these centroids, c_i: codebook element).

B. MLLR

MLLR is one of the most widely used transformation-based approaches, in which originally the mean vectors of Gaussian distributions in HMMs were modeled using an affine transformation [34]. The MLLR was extended to also update the Gaussian variances, and re-estimation formulae were

derived for these variance transforms [21]. In the MLLR framework, the Gaussian distributions in HMMs are clustered into several classes, such as phone classes, and one transformation is shared in the distributions in each class. The number of classes needs to be controlled according to the amount of data to avoid the data insufficiency problem.

Rather than using a static prior set of classes, a more robust estimation can be obtained by the use of regression class trees [17, 58]. The regression class tree is constructed by clustering together Gaussians that are close in the acoustic space, in such a way that similar components are transformed using the same transformation. This is based on the same idea as the structural approach described above. A binary regression tree can be constructed using a top-down splitting algorithm with a Euclidean distance or a symmetrized Kullback-Leibler divergence/distance. First, all Gaussians in the model set are assigned to the root node of the tree. At a given level of the tree, each node is split into two child nodes until a desired number of leaves is obtained. The partition of each node is performed by distributing the Gaussians in the two child nodes in such a way that the sum of the distances to the centroid node is minimized for each node. The terminal nodes (leaves) of the tree specify the base regression classes, and each Gaussian in the model set is assigned to one of these base classes.

During adaptation, the available data is aligned with the model set and the occupancy counts (number of observation vectors aligned with a given Gaussian) are computed for each base regression class. If a base regression class has sufficient data, a transform matrix is then estimated. If there is no sufficient data in a given node, observations of child nodes are pooled in its parent node. This process is repeated until sufficient data is collected and then the transformation matrix is estimated. Finally, all Gaussians assigned to a base regression class are transformed using the same matrix.

C. Signal bias removal

The signal bias removal (SBR) [49] corresponds to a special case of MLLR where the transform matrix is the unit matrix and only the bias vector is estimated and used, that is, the transformation is just a parallel shift. This method was originally proposed to normalize multiplicative distortion in the spectral domain by a bias in the cepstral domain.

D. CMLLR

Constrained maximum likelihood linear regression (CMLLR) uses the same transformation matrix for the covariance matrices and the mean vectors of HMMs [6, 18]. This method can be used not only for model adaptation but also as a feature adaptation technique that estimates a set of linear transformations for the features. Note that due to computational reasons, CMLLR is usually implemented for diagonal covariance, continuous density HMMs.

E. Interpolation

In order to solve the problem that features which do not appear in adaptation data cannot be adapted, interpolation techniques have been introduced [56], in which the bias of a parameter having no adaptation data is estimated by interpolating the biases of nearby parameters. This method asymptotically approaches the ML estimation when the amount of adaptation data is increased.

F. Vector field smoothing (VFS)

This method assumes that the correspondence of feature vectors between various conditions, typically between different speakers, can be viewed as a kind of smooth vector field [43]. Based on this assumption, the correspondence obtained from the adaptation is considered to be an incomplete set of observations from the continuous vector field, containing observation errors due to the insufficiency of the adaptation data. To achieve better correspondence as well as reduction in error, both interpolation and smoothing of the correspondence are introduced into the adaptation process. To make this method effective, it is important to control the range of smoothing according to the size of adaptation data.

G. Ensemble methods

Ensemble methods in machine learning that use multiple classifiers can also be used to alleviate the generalization problem in adaptation. Recently, cross-validation (CV) adaptation and aggregated adaptation algorithms have been proposed for batch-mode unsupervised adaptation [63, 64]. The latter algorithm is based on the idea of the bagging approach. In both algorithms, the adaptation utterances are split into K exclusive subsets, each with roughly the same size. In the CV adaptation, the adaptation utterances used in the decoding step and those used in the model updating step are separated based on the K-fold CV technique as shown in Fig. 7. K sets of recognition hypotheses are made using the separate adaptation utterance sets. Each of the K models is then adapted using different K-1 sets of hypotheses, and each adapted model is used to decode the utterance set that was not used to adapt the model. This process is repeated until the results converge.

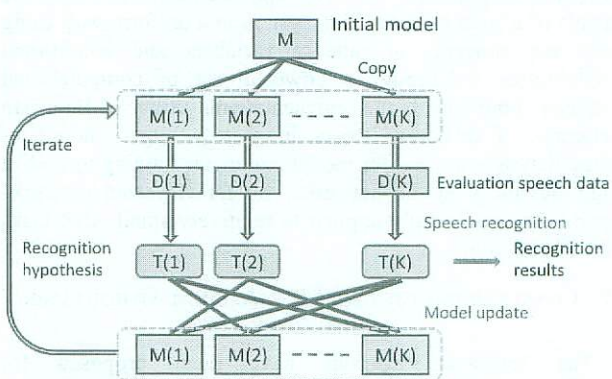

Fig. 7. Unsupervised cross-validation (CV) adaptation [63].

In the aggregated adaptation, each adaptation utterance set is decoded N times using separate models. Initially, these N models are made by copying the initial model. Each of the N

models is adapted using N x K' ($K'<K$) sets of hypotheses. The K' subsets are randomly selected. The adapted N models are used to decode each set of adaptation utterances. This process is repeated until the results converge.

Any kind of conventional adaptation techniques, such as the MLLR method, can be used in the model adaptation step in both algorithms. Since the proposed methods can suppress the negative effects of recognition errors included in the hypotheses for adaptation, they achieve significantly better results than a normal batch-mode unsupervised adaptation method, and the CV adaptation is more efficient than the aggregated adaptation.

VI. Combinations and Extensions

Many of the above mentioned methods have been combined and extended to further improve the performance of adaptation using a limited amount of data.

A. ML-based combinations

The ML-based adaptation techniques, such as MLLR, have been extended to incorporate the MAP estimation criterion. The maximum a posteriori linear regression (MAPLR) algorithm [5, 66] improves MLLR in a way similar to MAP enhancement over ML for HMM parameter estimation. The SMAP approach combines MAP and a structural approach. Combination of MLLR and MAP [7] and MAP and VFS [69, 70] have also been investigated. SMAPLR is the combination of MAP, affine transform, and a structural method [67]. Combination of MLLR and the eigen-voice method has also been investigated [4].

B. SMAP

Shinoda et al. [57] proposed a structural maximum a posteriori (SMAP) approach to improve the MAP estimates obtained when the amount of adaptation data is small. A hierarchical structure in the model space is made, and the priors corresponding to child nodes in the tree are derived from the parent node making it possible to specify all the priors for all the parameters in a large collection of HMMs to perform efficient and effective adaptation as shown in Fig. 8. Results of supervised adaptation experiments showed that SMAP estimation significantly reduced error rates when short utterances were used for adaptation and that it yielded the same accuracy as MAP and ML estimation when the amount of data was sufficiently large. The recognition results obtained in unsupervised adaptation experiments showed that SMAP estimation was effective even when only one utterance from a new speaker was used for adaptation.

C. N-best based method

In order to reduce the effects of recognition errors in the hypothesis obtained for an input utterance in the instantaneous unsupervised adaptation, an N-best list based scheme was proposed [37], which uses MAP estimations of mean biases. Smoothed estimation and utterance verification were also introduced. Experimental results show that this method is effective in improving the recognition accuracy especially for difficult speakers.

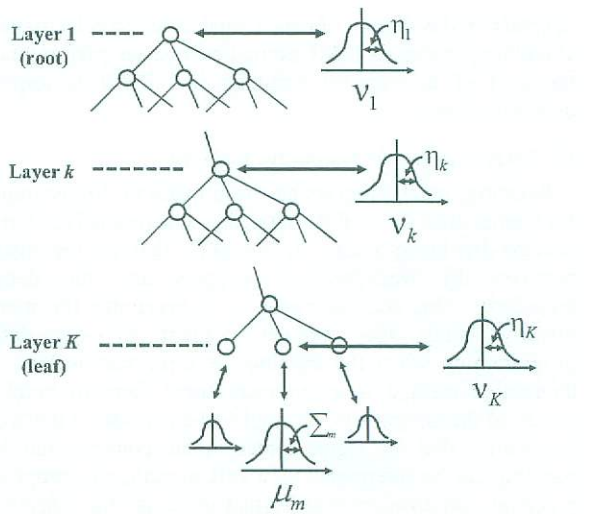

Fig. 8. Tree structure for Gaussian pdfs in continuous density HMMs used in the SMAP method. For simplicity, the case when the dimension is one (scalar) is shown [57].

An N-best list based Bayesian framework for MLLR affine transforms has been investigated in [75] for the unsupervised instantaneous adaptation.

D. Discriminative approach based combination

A supervised speaker adaptation method combining MAP and MCE estimation has been proposed to alleviate the over-tuning problem [36]. In this method, speaker-independent HMM parameters are first adapted to a new speaker by MAP estimation and then modified using MCE estimation. By using this combination, the HMM parameters after the MCE estimation fall into one of the local minima near the parameters adapted by MAP estimation. Since the MCE estimation directly aims at minimizing the recognition error but the results tend to be heavily biased towards the adaptation data, the combination is effective in improving the recognition accuracy.

E. Bayesian discriminative adaptation (Discriminative MAP estimation)

Raut and Gales [50] has investigated a MAP Bayesian approach for unsupervised discriminative adaptation. The Bayesian framework can reduce the hypothesis bias and makes the discriminative adaptation less sensitive to supervision hypothesis errors. Moreover, this Bayesian approach allows robust estimation of discriminative transforms even with a limited amount of data. This makes it possible to use them for instantaneously adapting model parameters.

F. I-smoothing

I-smoothing was proposed for smoothing discriminative training criteria using statistics for ML estimation (MLE) [47]. I-smoothing is a way of applying an interpolation between MLE and a discriminative objective function in a way which

depends on the amount of data available for each Gaussian. I-smoothing improves MMI estimation test-set performance at the cost of training set accuracy i.e., it yields improved generalization.

G. Large-margin discriminative training

Recently, much progress has been made to further improve the generalization ability into the discriminative training process by incorporating a "margin", that is, the distance between the well-classified samples and the decision boundary. One such approach is to maximize the margins directly using the gradient descent or semi-definite programming when the training set error rate is low. An alternative method is to optimize some form of combined scores of the margin and the empirical error rate. Yu et al. [74] has shown that the sigmoid bias in the conventional MCE training can be interpreted as a soft margin, and proposed a practical optimization strategy that increases the margin (the sigmoid bias) incrementally over epochs in the MCE training process so that a desirable balance between the empirical error rates on the training set and the margin can be achieved and verified by cross validation.

Heigold et al. [25] and Saon et al. [55] have recently proposed formulations of modified MPE and MMI that bring these methods into the space of large-margin based methods, while preserving standard optimization methods suitable for large-scale discriminative training. In these formulations, hypothesis strings with *high* error are artificially given *better* scores during the training procedure. Intuitively, "good" strings with low errors will have to work harder to beat the high error "bad" strings during training, but as a result such good strings will be more likely to be recognized during testing [38]. It has been shown that the proposed criteria are equivalent to Support Vector Machines with suitable smooth functions, approximating the non-smooth hinge loss function or the hard error (e.g. phone error) [25].

VII. Confidence Measures

Confidence measures (CMs) are widely used in un-supervised adaptation to select more reliable speech segments from a recognizer's output [27]. One important issue is that the operating point during the verification stage should be set up to guarantee a low false acceptance rate. A CM can be computed for every recognized word to indicate the likelihood that it has been correctly recognized, or for an utterance to indicate how much we can trust the results for the utterance as a whole. The posterior probability in the standard MAP decision rule is widely used as a CM, since it is an absolute measure of how good/reliable the decision is. However, it is very hard to estimate the posterior probability in a precise manner due to its normalization term in the denominator. In practice, many different approaches have been proposed to approximate it, ranging from simple filler-based methods to complex word-graph-based approaches.

The CM problem is sometimes formulated as a statistical hypothesis testing problem, especially under the framework of utterance verification which is a post-processing stage to examine the reliability of the hypothesized recognition result. In this framework, two complementary hypotheses, namely the null hypothesis H_0 and the alternative hypothesis H_1 are proposed. Then H_0 is tested against H_1 to determine whether the recognition result should be accepted or rejected. According to Neyman-Pearson Lemma, under some conditions, the optimal solution to this testing is based on a likelihood ratio testing (LRT). Similarly to the posterior probability, the major difficulty with LRT is how to model the alternative hypothesis which usually represent a very complex and composite event, where the true distribution of data is unknown. In practice, a general background model, hypothesis-specific anti-model, a set of competing models, or a combination of all the above is adopted to model the alternative hypothesis.

VIII. Special Training Methods for the Models Used for Adaptation

In order to make the adaptation processes more effective or to keep the consistency between the training and the adaptation processes, special training methods, instead of simply using a speaker/environment-independent model, have been investigated.

A. Adaptive training

Originally a speaker/environment-independent model was commonly used as the canonical model. Recently, since the majority of training databases have multiple speakers or acoustic environments, the adaptation scheme to be used in recognition has also been used during training. This is known as adaptive training [1, 48]. In the case of speaker adaptive training (SAT), the parameters for speaker-dependent model are estimated in the following process. First, a mapping from the parameters of the model created for each individual speaker to those of an initial model (speaker-independent model) is estimated. Second, this estimated mapping is used to map the utterance data for each speaker. Third, this mapped data is used to train the speaker-dependent model. This process is iterated until convergence. By using such adaptive training it is possible to build canonical models which only represent variability from individual speakers rather than the variability over all speakers in the training database.

B. Acoustic factorization

This technique attempts to explicitly model all the factors that affect the acoustic signal [52]. By explicitly modeling all the factors the trained model set is expected to be used in a more flexible fashion than in standard adaptive training schemes. Since an individual model is trained for each factor, it is possible to factor-in only those factors that are appropriate to a particular target domain, for example the distribution over all training speakers. The target domain specific factors are simply estimated from limited target specific data, for example the target acoustic environment. Gales [20] investigated a particular form of acoustic factorization that uses MLLR as the speaker transform and CAT as the noise transform.

IX. Conclusion and Future Works

Although many important scientific advances have taken place in automatic speech recognition research, we have also encountered a number of practical limitations which hinder a widespread deployment of applications and services. In most speech recognition tasks, human subjects produce one to two orders of magnitude fewer errors than machines [45]. One of the most significant differences exists in that human subjects are far more flexible and adaptive than machines against various variations of speech, including individuality, speaking style, additive noise, and channel distortions. How to train and adapt statistical models for speech recognition using limited amount of data is one of the most important research issues.

This paper has summarized major techniques that have been proposed to solve the generalization problem in acoustic model training and adaptation, that is, how to properly control the degree of freedom of the model or maximize "margins" to achieve high recognition accuracy for new utterances. Although various useful techniques have been proposed, we have not yet obtained a universal method which can be used for every condition of variations and the amount and quality of data for training and adaptation.

What we know about human speech processing and natural variation of speech is very limited. It is important to spend more effort to clarify especially the mechanism of speaker-to-speaker variability, and build a method of simultaneously modeling multiple sources of variations based on the statistical analysis using a large-scale database. It is also important to develop a seamless adaptation method which is applicable to a wide range of the amount of adaptation data.

Significant advances in speech recognition are not likely to come solely from research in statistical pattern recognition and signal processing. Although these areas of investigation are important, the most significant advances in next generation systems will come from studies in acoustic-phonetics, speech perception, linguistics, and psychoacoustics. Future systems need to have an efficient way of representing, storing, and retrieving various knowledge resources required for natural speech conversation [14].

X. Acknowledgment

The author would like thank Dr. Koichi Shinoda and Dr. Takahiro Shinozaki for their valuable comments and suggestions which have greatly improved this paper.

References

[1] T. Anastasakos, J. McDonough, R. Schwarts, and J. Makhoul, "A compact model for speaker-adaptive training," Proc. ICSLP96, 2, FrP2L1.3, 1996.

[2] M. Benzeghiba, et al., "Impact of variabilities on speech recognition," Proc. SPECOM, St Petersburg, pp. 3-16, 2006.

[3] J. Bilmes, "Graphical models and automatic speech recognition," Technical Report UWEETR-2001-05, University of Washington, Dept. of EE, Seattle WA, 2001.

[4] K. Chen and H. Wang, "Eigenspace-based maximum a posteriori linear regression for rapid speaker adaptation," Proc. ICASSP, p3.2, 2001.

[5] C. Chesta, O. Siohan, and C. Lee, "Maximum a posteriori linear regression for hidden Markov model adaptation," Proc. Eurospeech, 1, pp. 211-214, 1999.

[6] V. V. Digalakis, D. Rtischev, and L.G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," IEEE Trans, Speech, Audio Processing, 3, pp. 357-366, 1995.

[7] V. V. Digalakis and L.G. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," IEEE Trans, Speech Audio Processing, 4, pp. 294-300, 1996.

[8] J. Droppo and A. Acero, "Environmental robustness," Springer Handbook of Speech Processing, Springer, pp. 653-679, 2008.

[9] R. O. Duda, P. E. Hart and D. G. Stork, "Pattern classification," 2nd Edition, John Wiley & Sons, Inc., 2001.

[10] S. Furui, "A training procedure for isolated word recognition systems," IEEE Trans. Acoustics, Speech and Signal Processing, 28, 2, pp. 129-136, 1980.

[11] S. Furui, "Unsupervised speaker adaptation based on hierarchical spectral clustering," IEEE Trans. Acoustics, Speech and Signal Processing, 37, 12, pp. 1923-1930, 1989.

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

[13] S. Furui, "Recent progress in corpus-based spontaneous speech recognition," IEICE Trans. Information and Systems, E88-D, 3, pp.366-375, 2005.

[14] S. Furui, "Selected topics from 40 years of research on speech and speaker recognition," Proc. Interspeech, Brighton, UK, Mon-Ses1-K, 2009.

[15] S. Furui, et al., "Why is the recognition of spontaneous speech so hard?" Proc. 8th Int. Conf. on Text, Speech and Dialogue (TSD 2005), Karlovy Vary, Czech Republic, pp. 9-22, 2005.

[16] S. Furui, M. B. J. Hirschberg, S. Itahashi, T. Kawahara, S. Nakamura, and S. Nakayanan, "Introduction to the special issue on spontaneous speech processing," IEEE Trans. Speech, Audio Process, 12, pp. 349-350, 2004.

[17] M. J. F. Gales, "The generation and use of regression class trees for MLLR adaptation," CUED/F-INFENG/TR263, Cambridge University Engineering Department, 1996.

[18] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Computer, Speech and Language, 12, pp. 75-98, 1998.

[19] M. J. F. Gales, "Cluster adaptive training for speech recognition," Proc. ICSLP, pp. 1783-1786, 1998.

[20] M. J. F. Gales, "Acoustic factorization," Proc. ASRU 2001, Madonna di Campiglio, Italy, 2001.

[21] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," Computer Speech and Language, 10, pp. 249-264, 1996.

[22] M. J. F. Gales and S. J. Young, "The application of hidden Markov models in speech recognition," Foundations and Trends in Signal Processing, 1, 3, pp. 195-304, 2007.

[23] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains," IEEE Trans. Speech, Audio Process, 2, 2, pp.291-298, 1994.

[24] X.-D. He and L. Deng, "Discriminative learning in speech recognition," Technical Report, MSR-TR-2007-129, Microsoft, 2007.

[25] G. Heigold, T.Deselaers, R. Schlueter and H. Ney, "Modified MMI/MPE: a direct evaluation of the margin in speech recognition," Proc. Int. Conf. Machine Learning (ICML), Helsinki, Finland, 2008

[26] Q. Huo and C.-H. Lee, "On-line adaptive learning of the correlated continuous-density hidden Markov model for speech recognition," IEEE Trans. Speech Audio Processing, 6, pp. 386-397, 1998.

[27] H. Jiang, "Confidence measures for speech recognition: A survey," Speech Communication, 45, pp. 455-470, 2005.

[28] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," IEEE Trans. Signal Processing, 40, 12, pp. 3043-3054, 1992.

[29] F. Korkmazsky, M. Deviren, D. Fohr, and I. Illina, "Hidden factor dynamic Bayesian networks for speech recognition," Proc. ICSLP, Jeju Island, Korea, 2004.

[30] T. Kosaka, S. Matsunaga and S. Sagayama, "Tree-structured speaker clustering for speaker-independent continuous speech recognition," Proc. ICSLP, Yokohama, pp.1375-1378, 1994.

[31] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Finke, K. Field, and M. Contoloni, "Eigenvoices for speaker adaptation," Proc. ICSLP, pp. 1771-1774,1998.

[32] C.-H. Lee, "Adaptation and compensation for speech recognition – learning from extra data to improve robustness," Proc. Int. Workshop on Hands-Free Speech Communication (HSC 2001), Kyoto, Japan, pp. 27-30, 2001.
[33] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," Proc. ICASSP, 1, pp.353-356, 1996.
[34] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models", Computer, Speech and Language, 9, pp. 171-185, 1995.
[35] S. Matsuda, T. Jitsuhiro, K. Markov, and S. Nakamura, "Speech recognition system robust to noise and speaking styles," Proc. ICSLP, Jeju Island, Korea, 2004.
[36] T. Matsui and S. Furui, "A study of speaker adaptation based on minimum classification error training," Proc. Eurospeech, Madrid, Spain, pp. 81-84, 1995.
[37] T. Matsui and S. Furui, "N-best-based unsupervised speaker adaptation for speech recognition," Computer Speech and Language, 12, pp. 41-50, 1998.
[38] E. McDermott and A. Nakamura, "Recent advances in discriminative training for speech recognition," Proc. Spring Meeting, Acoustical Society of Japan, Tokyo, Japan, 2-5-10, 2009.
[39] S. Nakagawa, T. Watanabe, H. Nishizaki and T. Utsuro, "An unsupervised speaker adaptation method for lecture-style spontaneous speech recognition using multiple recognition systems," IEICE Trans. Information and Systems, E88-D, 3, pp.463-471, 2005.
[40] H. Nanjo and T. Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition," Proc. ICASSP, pp. 725-728, 2002.
[41] P. Nguyen, R. Kuhn, J.-C. Junqua, N. Niedzielski, and C. Wellekens, "Eigenvoices : a compact representation of speakers in a model space," Annales des Telecommunications, 55, 2000.
[42] Y. Normandin, "Maximum mutual information estimation of hidden Markov models," in Automatic Speech and Speaker Recognition, C.-H. Lee, F. K. Soong and K. K. Paliwal eds., Kluwer Academic Publishers, Norwell, MA, 1996.
[43] K. Ohkura, M. Sugiyama and S. Sagayama, "Speaker adaptation based on transfer-vector-field smoothing with continuous mixture density HMMs", Proc. ICSLP, Alberta, pp.369-372, 1992.
[44] M. Padmanabhan, L. R. Bahl, D. Nahamoo and M. A. Picheny, "Speaker clustering and transformation for speaker adaptation in speech recognition systems," IEEE Trans. Speech Audio Processing, 6, pp. 71-77, 1998.
[45] M. Picheny and D. Nahamoo, "Towards superhuman speech recognition," Springer Handbook of Speech Processing, Springer, pp. 597-616, 2008.
[46] D. Povey, Discriminative training for large vocabulary speech recognition, Ph. D. thesis, Cambridge University, 2003.
[47] D. Povey and P. C. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," Proc. ICASSP, Orlando, pp. I-105-108, 2002.
[48] D. Pye and P. C. Woodland, "Experiments in speaker normalization and adaptation for large vocabulary speech recognition," Proc. ICASSP, 2, pp. 1047-1050, 1997.
[49] M. Rahim and B.-H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," IEEE Trans. Speech, Audio Process., 4, 1, pp.19-30, 1996.
[50] C. K. Raut and M. J. Gales, "Bayesian discriminative adaptation for speech recognition," Proc. ICASSP, Taipei, Taiwan, pp. 4361-4364, 2009.
[51] R. Rose, "Environmental robustness in automatic speech recognition," Proc. COST278 and ISCA Workshop on Robustness Issues in Conversational Interaction, Norwich, UK, 2004.
[52] A-V. I. Rosti and M. J. F. Gales, "Factor analysed hidden Markov models," Proc. ICASSP, Orlando, pp. I-949-952, 2002.
[53] S. Sagayama, K. Shinoda, M. Nakai and H. Shimodaira, "Analytic methods for acoustic model adaptation: a review," Proc. ISCA ITR-Workshop, Sophia-Antipolis, pp.67-76, 2001.
[54] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, "Jacobian approach to fast acoustic model adaptation," Proc. ICASSP, pp.835-838, 1997.
[55] G. Saon and D. Povey, "Penalty function maximization for large margin HMM training," Proc. Interspeech, Brisbane, Australia, pp. 920-923, 2008.
[56] K. Shinoda, K. Iso, and T. Watanabe, "Speaker adaptation for demi-syllable based continuous density HMM," Proc. ICASSP, S13.7, pp. 857-860, 1991.
[57] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," IEEE Trans. Speech, Audio Process., 9, 3, pp. 276-287, 2001.
[58] K. Shinoda and T. Watanabe, "Speaker adaptation with autonomous control using tree structure," Proc. Eurospeech, pp. 1143-1146, 1995.
[59] K. Shinoda and T. Watanabe, "Speaker adaptation with autonomous model complexity control by MDL principle," Proc. ICASSP, pp.717-720, 1996.
[60] T. Shinozaki and S. Furui, "Hidden mode HMM using bayesian network for modeling speaking rate fluctuation," Proc. of ASRU, US Virgin Islands, pp.417-422, 2003.
[61] T. Shinozaki and S. Furui, "Time adjustable mixture weights for speaking rate fluctuation," Proc. Eurospeech, pp. 973-976, 2003.
[62] T. Shinozaki and S. Furui, "Spontaneous speech recognition using a massively parallel decoder," Proc. ICSLP, Jeju Island, Korea, pp. 1705-1708, 2004.
[63] T. Shinozaki, Y. Kubota and S. Furui, "Unsupervised cross-validation adaptation algorithms for improved adaptation performance," Proc. IEEE ICASSP, Taipei, Taiwan, pp. 4377-4380, 2009.
[64] T. Shinozaki, Y. Kubota and S. Furui, "Unsupervised cross-validation and aggregated adaptations for improved speech recognition," IPSJ SIG Technical Report, 2009-SLP-75, pp. 1-6, 2009. (in Japanese)
[65] Y. Shiraki and M. Honda, "Speaker adaptations algorithms for segment vocoder," Trans, Committee of Speech Res., Acoust. Soc. Japan, SP87-67, 1987. (in Japanese)
[66] O. Siohan, C. Chesta and C.-H. Lee, "Hidden Markov model adaptation using maximum a posteriori liner regression," Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions, pp.147-150, Tampere, Finland, 1999.
[67] O. Siohan, T.-A. Myrvoll and C.-H. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," Proc. ISCA ITRW Workshop on ASR, 2000.
[68] A. C. Surendran and C.-H. Lee, "Transformation-based Bayesian prediction for adaptation of HMMs," Speech Communication, 34, pp. 159-174, 2001.
[69] J. Takahashi and S. Sagayama, "Vector-field-smoothed Bayesian learning for incremental speaker adaptation," Proc. ICASSP, Detroit, pp. 696-699, 1995.
[70] M. Tonomura, T. Kosaka, and S. Matsunaga, "Speaker adaptation based on transfer-vector-field-smoothing using maximum a posteriori probability estimation," Proc. ICASSP, Detroit, pp. 688-691, 1995.
[71] H. Wakita, "Normalization of vowels by vocal tract length and its application to vowel identification," IEEE Trans. Acoustics, Speech and Signal Processing, 25, 2, pp. 183-192, 1977.
[72] L. Wang and P. C. Woodland, "MPE-based discriminative linear transforms for speaker adaptation," Computer Speech and Language, 22, 3, pp. 256-272, 2008.
[73] P. Woodland and D. Povey, "Large scale discriminative training for speech recognition," Proc. ITRW ASR, 2000.
[74] D. Yu, L. Deng, X.-D. He and A. Acero, "Large-margin minimum classification error training: A theoretical risk minimization perspective," Computer Speech and Language, 22, pp. 415-429, 2008.
[75] K. Yu and M. J. F. Gales, "Bayesian adaptive inference and adaptive training," IEEE Trans. Audio, Speech, Language Process., 15, 6, pp. 1932-1943, 2007.
[76] K. Yu, M. J. F. Gales, and P. C. Woodland, "Unsupervised discriminative adaptation using discriminative mapping transforms," Proc. ICASSP, pp. 4273-4276, 2008.
[77] G. Zavaliagkos, R. Schwartz, and J. McDonough, "Batch, incremental and instantaneous adaptation techniques for speech recognition" Proc. ICASSP, Detroit, pp. 676-679, 1995.
[78] G. Zweig, "Speech recognition with dynamic Bayesian networks," Ph.D. Thesis, University of California, Berkeley, 1998.