論文 / 著書情報
Article / Book Information

Title	Speaker Adaptation Techniques for Automatic Speech Recognition
Author	Koichi Shinoda
Journal/Book name	Proc. APSIPA ASC 2011, Vol. , No. , pp.
Issue date	2011, 10

Speaker Adaptation Techniques for Automatic Speech Recognition

Koichi Shinoda*
* Tokyo Institute of Technology, Tokyo, Japan
E-mail: shinoda@cs.titech.ac.jp

***Abstract*—Statistical speech recognition using continuous-density hidden Markov models (CDHMMs) has yielded many practical applications. However, in general, mismatches between the training data and input data significantly degrade recognition accuracy. Various acoustic model adaptation techniques using a few input utterances have been employed to overcome this problem. In this article, we survey these adaptation techniques, including maximum a posteriori (MAP) estimation, maximum likelihood linear regression (MLLR), and eigenvoice.**

I. Introduction

In statistical speech recognition, there are usually mismatches between the conditions under which the model was trained and those of the input. Mismatches may occur because of differences between speakers, environmental noise, and differences in channels. They should be compensated in order to obtain sufficient recognition performance. *Acoustic model adaptation* is the process of modifying the parameters of the acoustic model used for speech recognition to fit the actual acoustic characteristics by using a few utterances from the target user. In this paper, we mainly deal with *speaker* adaptation focused on the mismatch caused by the speaker variability.

Speaker-dependent speech recognition systems that were intended to recognize utterances from one target speaker were studied until the early 1980's. In this system, the target speaker registers his/her utterance for each word in the recognition vocabulary beforehand to create its template pattern. In recognition, each template pattern is matched to his/her input speech, and the word whose template has the smallest distance to the input speech is selected as the recognized word. In practice, the number of utterances to be registered should be as small as possible to decrease the load of users to register their voice. On the other hand, if a speaker uses a speaker-dependent system for the other speaker, its recognition accuracy seriously deteriorates, since acoustic characteristics vary much from speaker to speaker. Speech recognition systems which require only a few utterances from a user and have as high recognition performance as speaker-dependent systems were strongly demanded. This was the motivation for researchers to start development of speaker adaptation techniques.

Speech recognition technologies using hidden Markov models (HMMs) have significantly advanced since the late 1980's. In particular, speech recognition algorithms often employ continuous density HMMs (CDHMMs) using triphones as recognition units and a Gaussian mixture distribution as the output distribution. In CDHMMs, a variety of speech signals are represented as a continuous density distribution whose parameters are determined by using the expectation-maximization (EM) algorithm to make a maximum-likelihood (ML) estimation. The use of utterances from many speakers for training enables these models to represent not only phonetic features but also speaker features. Although this ability has made speaker-independent systems practical, the systems still do not perform as well as speaker-dependent systems in which the HMM parameters are estimated from a sufficient amount of utterances from one target user. This means that speaker adaptation techniques are important to speech recognition using CDHMMs.

There have been several surveys on adaptation techniques for speech recognition. For example, Lee and Huo [43], Woodland [73] and Sagayama *et al.* [62] surveyed the adaptation techniques in existence around 2001. Moreover, Furui [17] reviewed generalization techniques for training and adaptation, and Bellegarda [3] surveyed language model adaptation techniques for large vocabulary continuous speech recognition. This overview summarizes our recent comprehensive survey [67]. We aim to give the reader a unified view of present-day speaker adaptation methods.

The rest of this paper is organized as follows. Section II outlines speaker adaptation techniques for speech recognition. Sections III, IV, V explains the three major approaches in speaker adaptation; Section III explains adaptation techniques based on MAP estimation, Section IV explains transformation-based techniques including MLLR, Section V explains techniques using a pool of speakers including eigenvoice. Section VI describes speaker adaptive training techniques, which use adaptation techniques for feature normalization. Section VII briefly explains adaptation techniques for noisy environment, and Section VIII concludes our paper.

II. What is speaker adaptation?

A. Purpose

Let us assume that we have an acoustic model that has high recognition accuracy for one speaker, and let us consider how we can improve its speech recognition accuracy for another speaker by using only a few utterances worth of his/her speech data (*adaptation data*).

The number of parameters in a triphone CDHMM is generally large. It consists of a few thousand states, and each state has a Gaussian mixture distribution with dozens of mixture

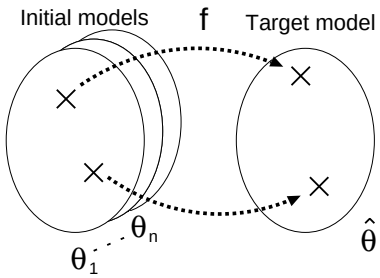

Fig. 1. Speaker adaptation. f is a mapping function (an adaptation model). $\theta_1 \ldots \theta_n$ are the parameter sets of initial models, and $\hat{\theta}$ is the parameter set of the target model.

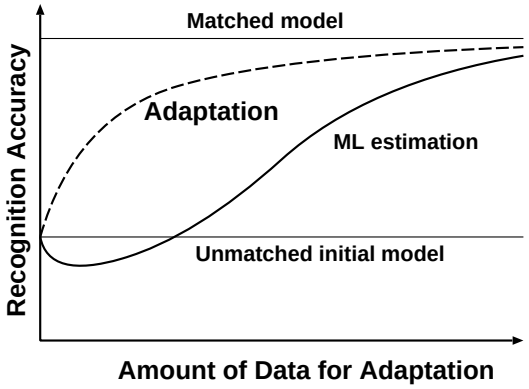

Fig. 2. Requirements of acoustic model adaptation

components. Each mixture component has a mean vector with dozens of elements and its corresponding covariance matrix. Moreover, it also includes parameters for expressing transition probabilities and initial probabilities. In total, the parameters usually number more than 1,000,000. In contrast, the number of data samples that can be obtained from a few utterances is much more limited. Each utterance is typically about 1 sec. long, and one feature vector with dozens of elements can be obtained every 10 msec. That means only a few thousand data samples can be obtained from a few utterances.

In this situation, ML estimation using the EM algorithm cannot precisely estimate the model parameters. As a result, recognition accuracy would be much worse than under the original conditions. This is called the data sparseness problem.

Speaker adaptation aims to overcome the above problems. Let θ_i be the parameter set of an initial model i given beforehand, and let $\hat{\theta}$ be that of the target model to be determined. Speaker adaptation can be defined as a process to find a mapping function f from the space of parameters of the initial models to the space of the target model using adaptation data

$$\hat{\theta} = f(\theta_1, \ldots, \theta_n), \tag{1}$$

where n is the number of initial models provided. We hereafter call this mapping function f an *adaptation model* (Figure 1). When f consists of mapping functions defined for each CDHMM parameter independently, adaptation using it is called *direct adaptation*. Adaptation using adaptation models with parameter sharing are called *indirect* adaptation [43].

An adaptation model should meet the following requirements.

1) It should improve recognition accuracy even with a small amount of adaptation data.
2) As the amount of adaptation data increases, it should make the recognition accuracy asymptotically approach the accuracy of a *matched* model.

Here, a matched model is the CDHMM whose parameters are estimated using a sufficient amount of data collected in the new condition. Figure 2 illustrates these requirements.

The first requirement can be met by designing good adaptation models with only a few free parameters to be estimated. However, such simple models may fail to fulfill the second requirement; they may not improve recognition accuracy once a larger amount of adaptation data becomes available. This happens because the models are too simple to represent the richness of information contained in a large amount of adaptation data. Thus, the adaptation model should have an appropriate number of free parameters (i.e., an appropriate model size).

B. Supervised and unsupervised adaptation

Speaker adaptation techniques are categorized into supervised adaptation and unsupervised adaptation. In supervised adaptation, a transcription exists for each utterance. In unsupervised adaptation, it does not.

In supervised adaptation, the user should register his or her own voice. To do so, the system shows the user predetermined words or sentences and asks him/her to utter them. The speaker registration process used in dictation software is an example of this process. The early speaker adaptation techniques required up to 20 minutes worth of speech data. Nowadays, thanks to the progress of speaker-independent recognition and speaker adaptation, most systems require only one minute worth of data.

While dictation software is intended to be used for a long time, other applications, e.g., airline ticket reservations by telephone, are intended to be used by one person for only a short time. Unsupervised adaptation techniques are needed for such short-period applications since users should not have to spend time registering their voices. Most techniques are related to supervised adaptation in that they use transcriptions obtained from speaker-independent speech recognition as the supervising signal for adaptation[1] These techniques usually perform well when the recognition accuracies of speaker independent speech recognition are high enough to get reliable transcriptions.

Incorrect supervised signals generated by misrecognitions may significantly degrade adaptation performance. Some techniques for alleviating their effect calculate a *confidence mea-*

[1]Although there are unsupervised adaptation methods that are not related to supervised adaptation (e.g., [16]), their recognition accuracies are not as good as the unsupervised methods that are related to supervised adaptation.

sure, such as a posterior probability, for each utterance and use only those utterances with confidence measures larger than a predetermined threshold for adaptation (e.g., [57]). Matsui *et al.* used N-best sentences output from a speech recognizer as the supervising signals [49]. More recently, Shinozaki *et al.* proposed a cross-validation based scheme [68].

C. Batch and on-line adaptation

Speaker adaptation can also be categorized as batch adaptation or on-line adaptation [77]. Batch adaptation is done after all the available utterances are collected, whereas on-line adaptation is done each time one utterance is obtained. Batch adaptation requires sufficient memory to store the statistics to be used for parameter estimation, while on-line adaptation does not require such a large memory.

Batch adaptation performs better than on-line adaptation when both methods use the same adaptation data, since it can simulate any on-line adaptation. Thanks to recent advances in computational technology, high CPU speeds and large memories, batch adaptation can use all the previous utterances each time a new utterance is obtained.

On-line adaptation is preferable for applications such as speech recognition during meetings, where the speakers often change and the change points are not given beforehand. The forgetting parameters [27] of on-line adaptation should be carefully tuned so as not to use utterances obtained before a certain point in time.

The above explanation indicates that supervised batch adaptation is fundamental and unsupervised adaptation and on-line adaptation are its applications. Hence, in what follows, we will discuss supervised batch adaptation unless otherwise noted.

III. MAP ADAPTATION

Maximum *a posteriori* (MAP) estimation (e.g., [10]) is used in statistical modeling and has a wide range of applications. In particular, it has been often used for speaker adaptation (e.g., [21]). It estimates model parameters more robustly than ML estimation when the amount of data is small, and its estimates asymptotically approach ML estimations as the amount of data increases.

Let $f(x|\theta)$ be the probability density function (*pdf*) of variable x. We estimate its parameter θ by using T samples of x, $\mathcal{X} = \{x_1, \ldots, x_T\}$. In ML estimation, the parameter is estimated as follows.

$$\tilde{\theta} = \underset{\theta}{\operatorname{argmax}}\, f(\mathcal{X}|\theta), \tag{2}$$

where $\tilde{\theta}$ is the maximum likelihood estimator of θ. In MAP estimation, θ is regarded as a random variable that follows a certain *pdf*. We expect that our knowledge about it increases as we observe data samples. The parameter distribution before observing the data is called a *prior* distribution. Let $g(\theta)$ be the prior distribution for θ. The pdf of the parameter after observing $\mathcal{X}$, $g(\theta|\mathcal{X})$ is called a posterior distribution, and it is written as follows (Bayes' Theorem),

$$g(\theta|\mathcal{X}) = \frac{f(\mathcal{X}|\theta)g(\theta)}{\int f(\mathcal{X}|\theta)g(\theta)d\theta}. \tag{3}$$

MAP estimation obtains the value of $\hat{\theta}$ that maximizes the mode of the posterior distribution, that is, the value which gives the maximum of the posterior distribution:

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmax}}\, g(\theta|\mathcal{X}) \\ &= \underset{\theta}{\operatorname{argmax}}\, f(\mathcal{X}|\theta)g(\theta). \end{aligned} \tag{4}$$

When we have little knowledge about θ, we should select a uniform distribution over the range of possible θ values as the prior distribution [2]. In this case, the MAP estimator becomes almost identical to the ML estimator.

There are no *theoretically correct* answers as to which class to use for prior pdfs and how to set their parameters. Users may determine them according to their own preference by making observations. However, we can analytically obtain the MAP estimator when $f(x)$ has sufficient statistics with a fixed dimension, and such an $f(x)$ should belong to an exponential family. Furthermore, when $f(x)$ belongs to an exponential family and we choose the prior distribution from the same family of the kernel distribution of $f(x)$ (the distribution whose parameters are sufficient statistics of $f(x)$ only), which we call the conjugate family, the posterior distribution accordingly belongs to the same family. This fact makes our calculation much easier. This type of prior distribution is called a natural conjugate prior distribution, and used in many adaptation techniques.

A. MAP estimation for CDHMMs

Next, let us discuss a MAP estimation method for CDHMM parameters [21]. This method is called *MAP adaptation*.

Let $\Lambda = \{\Pi, A, W, B\}$ be the parameter set of a CDHMM. Here, $\Pi = \{\pi_i\}$ is the set of initial probabilities, $A = \{a_{ij}\}$ is the set of transition probabilities, $W = \{w_{ik}\}$ is the set of mixture weights in the Gaussian mixture distribution, and $B = \{b_{ik}(\boldsymbol{x})\}$ is the set of pdfs in each mixture component, where i, j are state indexes, and k is an index for each mixture component in a state.

In general, models with hidden variables, such as HMMs, do not have natural conjugate priors. Accordingly, the MAP estimator can not be analytically calculated for these models. To overcome this problem, we assume that Π, A, W, and B are independent from each other, and furthermore, their elements are independent from each other. Accordingly, the prior distribution can be defined as the joint probability of the natural conjugate prior for each parameter [41], [21]. Here, the normal-Wishart distribution can be used as the prior for normal distribution, and the Dirichlet distribution can be used as the prior for the initial probability, transition probability,

[2]We call such a prior distribution a non-informative prior distribution. An example is Jeffreys' prior distribution.

and mixture weight. The prior for an HMM can be expressed as follows.

$$g(\Lambda) = g(\Pi)g(A)g(W)g(B) = C\prod_{i=1}^{N}\left[\pi_i^{\eta_i-1}\left(\prod_{j=1}^{N}a_{ij}^{\eta_{ij}-1}\right)\left(\prod_{k=1}^{K}w_{ik}^{\nu_{ik}-1}g(b_{ik})\right)\right]. \quad (5)$$

Here, C is a normalization factor, and η_i, η_{ij}, and ν_{ik} are parameters of a prior pdf for the initial probability π_i, transition probability a_{ij}, and mixture weight w_{ik}, respectively. $g(b_{ik})$ is the prior pdf for the normal distribution $b_{ik}(\boldsymbol{x})$ and is a normal-Wishart distribution.

B. Related methods

Quasi-Bayes adaptation [27], [28] is an application of MAP estimation to on-line adaptation. In this method, the posterior probability is approximated with a normal distribution in the sequential Bayes estimation scheme. The parameter estimation is carried out using the following auxiliary function.

$$R(\Lambda,\bar{\Lambda}) = Q(\Lambda,\bar{\Lambda}) + \rho\log g(\Lambda) \quad (6)$$

Here, the model parameterΛ is estimated from all past samples. ρ is the forgetting factor, which should be optimized for each application. While it is not necessary to memorize sufficient statistics for the past samples, the estimated parameters may not converge to the ML estimator obtained in batch training. MAP adaptations for discrete HMMs and semi-continuous HMMs have also been studied [26].

Shinoda and Lee proposed structural maximum a *posteriori* (SMAP) adaptation [65], [66]. This method shares parameters by using a tree structure when the data amount is small, yet retains the asymptotic nature of MAP estimation. In this method, a tree of Gaussian distributions is first constructed using the Kullback-Leibler pseudo distance as the distance between distributions. The root node represents the whole acoustic space, and each of its leaf nodes corresponds to a Gaussian distribution in an HMM.A Gaussian distribution is assigned to each node, and it is estimated as a shared parameter among its descendant leaf nodes. The parameter of the parent node is used as the prior, and the MAP estimation is carried out from the root node to leaf nodes in a cascade manner. This method was proved to be effective even when only a few utterances were available as adaptation data.

IV. Transformation-based techniques

A. Shift

A *shift* is the difference between mean vectors before adaptation and after adaptation. The method of sharing one shift for the mean vectors of all mixture components in a CDHMM is called signal bias removal (SBR) [60]. It corresponds to parallel displacement in the parameter space. It has been used to adapt to multiplicative noise when cepstral coefficients are used as features. It is a special case of the MLLR method explained later, where $A = I$ (I is an identity matrix). It becomes identical to ML estimation of mean vectors when a shift is provided for each mean vector.

Between SBR and ML estimation lies spectral interpolation [64] and vector field smoothing (VFS) [55]. Spectral interpolation [64] estimates the shifts for the parameters without corresponding data samples in the adaptation data by interpolating the shifts in the neighborhood in the parameter space. Its estimates asymptotically become close to ML estimators. VFS [55] applies smoothing to shifts close to each other after interpolation. The smoothing is effective when the amount of adaptation data is small. The recognition accuracy of VFS cannot reach that of ML estimation when the amount of adaptation data is sufficient. To avoid this, the degree of smoothing has to be controlled according to the amount of available data . Tonomura *et al.* used MAP estimation to estimate shifts in VFS [72].

Stochastic matching (SM) [63] estimates not only the shift but also its variance to improve robustness against noise. Chien *et al.* [7] used MAP estimation for stochastic matching (SM) [63].

B. Linear mapping

Maximum likelihood linear regression (MLLR) [44] uses a linear mapping between the acoustic feature spaces of different speakers as the adaptation model. It is one of the most popular model adaptation methods since it is easy to use and performs well in most cases.

In MLLR, the mean vectors of the Gaussian distributions in the HMMs, $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_n)'$ where n is the dimension of a feature vector, are updated according to the following transformation:

$$\hat{\boldsymbol{\mu}} = A\boldsymbol{\mu} + \boldsymbol{b},, \quad (7)$$

Here, A is an $n\times n$ matrix, and $\boldsymbol{b}$ is a n-dimensional vector.

There are two major methods of transforming covariance matrices: constrained MLLR [11] and unconstrained MLLR [18]. Constrained MLLR transforms features in feature space. The covariance matrix is transformed as follows:

$$\hat{\Sigma} = A\Sigma A'. \quad (8)$$

A cannot be analytically calculated since it is inside the Jacobian of the variable transformation. Instead, it can be numerically calculated by using Newton's method, etc. LU decomposition can also be used in some situations [54]. Unconstrained MLLR, on the other hand, assumes that the covariance matrix represents speaker characteristics that are different from the mean. It results in a different transformation matrix from that of the mean vector and estimates its parameters independently. Although it increases the number of parameters, unconstrained MLLR is useful in noisy environments where the variances of parameters are usually large. As in the case of MAP adaptation, adaptation of variances does not bring much improvement to MLLR. For example, Gales [19] evaluated MLLR on the Wall Street Journal (WSJ) task. Adaptation of only mean vectors reduced the error rate

by 13%, but further adaptation of the mean and covariance amounted to only a 2% reduction.

Several studies have applied MLLR to on-line adaptation [8], [35]. For example, Chien *et al.* applied Quasi-Bayes estimation to affine mapping parameters [8].

MAP is a method to estimate parameters and MLLR provides a class of adaptation models. Using MAP estimation within the MLLR framework is thus expected to yield a larger improvement than using them independently. Digalakis *et al.* [12] used the mean vectors obtained by MLLR as the mean vectors of the prior distribution for MAP adaptation. Different from MLLR alone, this method performs as well as ML estimation even when the amount of data is large. Chesta *et al.* [6] and Chou [9] separately proposed maximum a *posteriori* linear regression (MAPLR); the algorithm refines the MLLR algorithm in the same way as MAP estimation does ML estimation. That is, the problem whereby the MLLR estimate becomes unstable when the amount of data is extremely small can be solved by using an elliptically symmetric matrix variant prior, which is a natural conjugate prior for a linear mapping. In addition, Siohan *et al.* proposed SMAPLR, a combination of SMAP with MLLR [69].

V. Adaptation method using a speaker pool

A. Speaker clustering

Speaker clustering clusters speakers and prepares an HMM for each resulting cluster. In the recognition phase, a few utterances from a speaker are used to identify the cluster to which he or she belongs, and the corresponding HMM is used to recognize his/her voice.

The measure of the distance between speakers is the key issue in speaker clustering. Popular measures are the Bhattacharyya distance between output probabilities [37] and the probability of generating one speaker's data from another speaker's model after clustering [56]. Yoshizawa *et al.* used sufficient statistics to measure the distance [74], [22]. Hazen *et al.* used a soft selection method in which a speaker model is represented by a weighted sum of more than one speaker cluster [24].

Although gender-dependent models, in which a cluster is made for each gender, are effective, clustering inside the same gender results in little improvement. This is mainly because the data used for making a cluster model becomes smaller as the number of clusters increases. That is, there is a trade-off between the detailed representation of speaker characteristics and the amount of data to make a precise cluster model.

Speaker clustering decreases the size of recognition models without incurring a large degradation in speech recognition accuracy, and thus, it decreases the computational cost for recognition.

B. Eigenvoice

Kuhn *et al.* recently proposed the eigenvoice [38], [40] method for speaker adaptation. This name *eigenvoice* is in analogy to the eigenface [36] method, which employs principal component analysis for face image recognition. Eigenvoice uses principal component analysis to project a speaker-supervector to a subspace of much smaller dimension.

In the training phase, training data from a large number of speakers are prepared and a speaker-dependent model is built for each speaker. Then, for each speaker, a speaker-supervector is constructed by concatenating all the mean vectors of his/her speaker-dependent HMM. Next, principal component analysis is done on the set of speaker-supervectors, and the principal components (eigen vectors) are extracted. Each set of eigen vectors is called an eigenvoice, and it forms a subspace of much smaller dimension than that of the speaker-supervectors.

A linear mapping of a new speaker's supervector to the subspace is estimated by using ML estimation on a small amount of his/her speech data. Let M be the dimension of a speaker-supervector and J be the number of its eigen vectors ($J < M$), which are expressed as:

$$\boldsymbol{e}(j) = (e_1(j), \ldots, e_M(j))', \quad j = 1, \ldots, J. \tag{9}$$

The speaker-supervector of a new speaker is approximated by the weighted sum of the eigen vectors as follows.

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)' = \sum_{j=1}^{J} w(j)\boldsymbol{e}(j), \tag{10}$$

where the weight for each eigen vector $w(j), j = 1, \ldots, J$ is ML estimated with the EM algorithm. This estimation procedure is called maximum likelihood eigen-decomposition (MLED).

The dimension of a speaker-supervector is usually very large but the amount of data for each speaker used for training is usually relatively small. Many techniques have been proposed to deal with the data insufficiency problem. The original eigenvoice paper proposed to use mean vectors estimated by eigenvoice adaptation as the priors for MAP adaptation, and this proved to be effective [38]. Other approaches have used probabilistic PCA (PPCA) in eigenvoice adaptation [29], [33], [34], [32]. Mak *et al.* applied non-linear PCA using kernel methods [46], [47]. Tanji *et al.* explored the way to efficiently cluster the speaker-phone matrix [71].

There is an alternative approach that combines MLLR and eigenvoice [5] wherein a transformation matrix for each speaker in the training data is used to form a speaker-supervector and eigenvoice adaptation is applied to the set of speaker-supervectors.

Some other studies the other multivariate analysis method than PCA to obtain the subspace. For example, Duchateau *et al.* employed non-negative matrix factorization (NMF) [13], Hahm *et al.* used probabilistic latent semantic analysis (PLSA) [23].

VI. Speaker adaptive training

Adaptation updates model parameters to fit the speaker's acoustic features. Normalization, on the contrary, modifies the feature space to fit a prepared model. Sometimes, these two

approached can be combined (e.g., [63]). We shall discuss normalization in this section, in particular, speaker adaptive training, which is normalization for speaker differences.

A. Feature compensation

Feature compensation methods, for example, cepstrum mean normalization (CMN) [2] and vocal tract length normalization (VTLN) [14], try to exclude from input features factors caused by the mismatches in speaker characteristics, environmental noise, and channels.

In CMN, the long-time average of the cepstrum coefficients is subtracted from the cepstrum coefficients. Influences from surrounding noise or channel variations, whose rates of change are much slower than phonetic features in speech, are removed from the features. CMN is a standard method in practical applications.

The formant frequencies in the power spectrum vary from speaker to speaker, since vocal tract lengths vary. VTLN estimates the vocal tract length of each speaker from his/her spectrum of speech data and transforms the spectrum to that of a *canonical* speaker. It is difficult to estimate the vocal tract length precisely, so some methods have uses ML estimation (ML-VTLN) [42], [76]. These methods prepare several different-length vocal tract models and choose the model maximizing the likelihood for the speaker's utterances. McDonough *et al.* approximated the warping function in VTLN by using all-pass transforms [50]. VTLN is a special case of speaker adaptive training using MLLR (explained in the next subsection) where the transformation matrix has free parameters only in its diagonal elements and their neighborhood [15], [58].

B. Speaker Adaptive Training

Speaker adaptive training (SAT) and related techniques are intended to provide a good initial model for speaker adaptation [16], [1], [59]. If we assume that speaker adaptation is always carried out, an initial model of a *canonical* speaker who has the average nature of all speakers is preferable to a speaker-independent model representing the difference between phonemes and the difference between speakers.

The canonical speaker model is estimated as follows. First, an initial model is prepared and the mapping between its parameters and the parameters of the speaker-dependent model for each training speaker is estimated. Next, this mapping is used to transform speech data of each training speaker. The canonical speaker model is trained with the transformed data of many training speakers and set as the initial model for the next step. This process is repeated several times. The recognition phase estimates the mapping from the canonical model to the target speaker by using the speaker's utterances and the mapping to adapt the model. The mapping should be carefully selected so that it can be precisely estimated even when there is only a small amount of data for each training speaker. Affine mapping as in MLLR is often used.

Cluster adaptive training (CAT) uses several models made by speaker clustering for SAT, and it simultaneously estimates the model parameters and the weight among the models [20]. When only the weight coefficients for speakers are estimated in adaptation, the number of free parameters is very small and thus has a similar tendency with eigenvoice. Yu *et al.* introduced discriminative learning to CAT [75]. Arindam *et al.* combined CAT with MLLR [48]. Tang *et al.* discussed the relationship between CAT and eigenvoice [70].

VII. Adaptation to noisy speech

Since speech recognition accuracy significantly deteriorates in noisy environments, many studies have sought ways to lessen the effect of noise. The methods can be roughly classified into three categories: feature compensation, model adaptation, and missing feature theory. Many techniques have been applied to model adaptation in noisy environments. For example, Zhang *et al.* applied tree-structure-based adaptation [78], and Nguyen *et al.* applied MLLR and eigenvoice [53]. Adaptation to noisy speech is different from speaker adaptation in that noisy speech has not only convolutive factors, but also additive factors in the spectral domain. The vector Taylor series based approach [51], Jacobian adaptation [61], and their extensions (e.g., [4], [45]) have been extensively studied.

VIII. Conclusion and Future work

We surveyed acoustic model adaptation techniques for speech recognition using CDHMMs. In the future, as more speech data recorded in different noisy environments and channels becomes available, we expect that the adaptation techniques using the a speaker pool will become especially prominent. Such techniques will include ones that can efficiently exploit transcriptions that vary largely among speakers [74], [22], those based on multivariate analysis such as eigenvoice, subspace-based methods to separate phonetic features and speaker features [52], and unified approaches for speaker and phonetic features [30], [31].

Lastly, the analysis of speaker variety remains as an important challenge. A large amount of data from many speakers is now available, so we believe it is time to tackle this problem (e.g., [39], [25]).

References

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

[2] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304-1312, 1974.

[3] J. R. Bellegarda, "Statistical language model adaptation: review and perspectives," *Speech Communication*, vol. 42, no. 1, pp. 93-108, 2004.

[4] C. Cerisara, L. Rigazio, and J. C. Janqua, "α-Jacobian environmental adaptation," *Speech Communication*, vol. 42, pp. 25-41, 2004.

[5] K. Chen, W. Liau, H. Wang and L. Lee, "Fast speaker adaptation using Eigenspace-based maximum likelihood linear regression,", Proc. ICSLP-2000, 2000.

[6] C. Chesta, O. Siohan, and C.-H. Lee, "Maximum a posteriori linear regression for hidden Markov model adaptation," *In Proc. EuroSpeech99*, pp. 211-214, 1999.

[7] J.-T. Chien, C.-H. Lee, and H.-C. Wang, "Improved Bayesian learning of hidden Markov models for speaker adaptation," *Proc. ICASSP-97*, pp. 1027-1039, 1997.

[8] J.-T. Chien, "Quasi-Bayes linear regression for sequential learning of hidden Markov models," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 4, pp. 268-278, 2002.

[9] W. Chou, "Maximum *a posteriori* linear regression with elliptically symmetric matrix variance priors," *Proc. Eurospeech-99*, vol. 1, pp. 1-4, 1999.

[10] M. H. DeGroot, *Statistical Decision Theory and Bayesian Analysis*, McGraw-Hill, 1970.

[11] V. V. Digalakis, D. Rtishev and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. on Speech and Audio Processing*, vol. 3, No. 5, pp. 357-366, 1995.

[12] V. V. Digalakis and L. G. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods, " *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 4, pp. 294-300, 1996.

[13] J. Duchateau, T. Leroy, K. Demuynck, and H. Van homme, "Fast speaker adaptation using non-negative matrix factorization," *ICASSP '08*, pp. 4269-4272, 2008.

[14] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. ICASSP96*, vol. 1, pp. 346-3483, 1996.

[15] T. Emori, K. Shinoda, "Rapid vocal tract length normalization using maximum likelihood estimation," *Proc. Eurospeech-2001*, pp. 1649-1652, 2001.

[16] S. Furui, "Unsupervised speaker adaptation method based on hierarchical spectral clustering," *Proc. ICASSP-89*, pp. 286-289, Glasgow, 1989.

[17] S. Furui, "Generalization Problem in ASR Acoustic Model Training and Adaptation," *IEEE ASRU Workshop*, Merano, pp. 1-10, 2009.

[18] M. J. F. Gales and P. C. Woodland, "Mean and covariance adaptation within MLLR framework," *Computer Speech and Language*, Vol.10, pp. 249-264, 1996.

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

[20] M. J. F. Gales, "Cluster adaptive training for hidden Markov models," *IEEE Trans. on Audio and Speech Processing*, vol. 8, no. 4, pp. 417-428, 2000.

[21] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*, pp. 291-298, Vol. 2, No. 2, April 1994.

[22] R. Gomez, T. Toda, H. Saruwatari, and K. Shikano, "Techniques in rapid unsupervised speaker adaptation based on HMM-sufficient statistics," *Speech Communication*, vol. 51, pp. 42-57, 2004.

[23] S. Hahm, A. Ito, S. Makino, and M. Suzuki, "A fast speaker adaptation method using aspect model," *Interspeech '08*, pp. 1221-1224, 2008.

[24] T. J. Hazen, "A comparison of novel techniques for rapid speaker adaptation," *Speech Communication*, vol. 31, pp. 15-33, 2000.

[25] C. Huang, T. Chen, S. Li, E. Chang and J. Zhou, "Analysis of speaker variability," *EuroSpeech-2001*, 2001.

[26] Q. Huo and C. Chan, "Bayesian adaptive training of the parameters of hidden Markov model for speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 5, pp. 334-345, 1995.

[27] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans. on Audio and Speech Processing*, Vol. 5, No. 2, pp. 161-172, March 1997.

[28] Q. Huo and C.-H. Lee, "On-Line adaptive learning of the correlated continuous-density hidden Markov model for speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 4, pp. 386-397, 1998.

[29] P. Kenny, G. Boulianne and P. Dumouchel, "Bayesian adaptation revisited," *In Workshop on ISCA ITRW ASR2000*, pp. 112-119, 2000.

[30] P. Kenny, G. Boulianne, and P. Dumouchel, "Inter-speaker correlations, intra-speaker correlations and Bayesian adaptation,", in Proc. Isca ITR-Workshop2001, Sophia-Antipolis, 2001.

[31] P. Kenny, G. Boulianne and P. Dumouchel, "What is the best type of prior distribution for EMAP speaker adaptation," *EuroSpeech-2001*, pp. 1207-1210, 2001.

[32] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 3, pp. 345-354, 2005.

[33] D. K. Kim and N. S. Kim, "Maximum *a posteriori* adaptation of HMM parameters based on probabilistic component analysis," *Proc. ISCA ITR-Workshop 2001*, pp. 25-28, 2001.

[34] D. K. Kim and N. S. Kim, "Maximum a posteriori adaptation of HMM parameters based on speaker space projection," *Speech Communication*, vol. 42, pp. 59-73, 2004.

[35] D. K. Kim and N. S. Kim, "Rapid online adaptation based on transformation space model evolution," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 2, pp. 194-202, 2005.

[36] M. Kirby and L. Sirovich, "Application of the Karhunen-Lo'eve procedure for the characterization of human faces,", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 103-108, 1990.

[37] T. Kosaka and S. Sagayama, "Tree-structured speaker clustering for fast speaker adaptation," *ICASSP-94*, vol. 1, pp. 245-248, Adelaide, 1994.

[38] R. Kuhn, P. Nguyen, J.-C. Janqua, L. Goldwasser, N. Niedzielski, S. Finke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," *in Proc. ICSLP-98*, pp. 1771-1774, 1998.

[39] R. Kuhn, J.-C. Janqua, R. Boman, N. Niedzielski, S. Fincke, K. Field and M. Contolini, "Fast speaker adaptation using *a priori* knowledge," *Proc. ICASSP-99*, 2001.

[40] R. Kuhn, J.-C. Janqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in Eigenvoice space robust speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 6, pp. 695-707, 2000.

[41] C.-H. Lee, C.-H. Lin and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," in *IEEE Trans. Signal Proc.*, Vol. SP-39, No. 4, pp. 806-814, April 1991.

[42] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP96*, vol. 1, pp. 353-356, 1996.

[43] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proc. IEEE*, vol. 88, pp. 1241-1269, 2000.

[44] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous-density hidden Markov models," *Computer Speech and Language*, Vol. 9, pp. 171-185, 1995.

[45] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions," *Computer Speech and Language*, vol. 23, pp. 389-405, 2009.

[46] B. Mak, J. T. Kwak, and S. Ho, "Kernel eigenvoice speaker adaptation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 13, no. 5, pp. 984-992, 2005.

[47] B. K.-W. Mak, R. W.-H. Hsiao, S. K.-L. Ho, and J. T. Kwak, "Embedded kernel eigenvoice speaker adaptation and its implication to reference speaker weighting," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1267-1280, 2006.

[48] A. Mandal, M. Ostendorf, A. Stolcke, "Improving robustness of MLLR adaptation with speaker-clustered regression class trees," *Computer Speech and Language*, vol. 23, pp. 176-199, 2009.

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

[50] J. McDonough, T. Schaaf, A. Waibel, "Speaker adaptation with all-pass transforms," *Speech Communication*, vol. 42, pp. 75-91, 2004.

[51] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition,", *Proc. ICASSP '96*, pp. 733-736, 1996.

[52] N. Nishida and Y. Ariki, "Speaker recognition by separating phonetic space and speaker space," *in Proc. Eurospeech-2001*, 2001.

[53] P. Nguyen, C. Wellekens, and J.-C. Janqua, "Maximum likelihood eigenspace and MLLR for speech recognition in noisy environment," *Proc. Eurospeech-99*, pp. 2519-2522, 1999.

[54] P. Nguyen, L. Rigazio, C. Wellekens and J. C. Junqua, "LU factorization for feature transformation," *in Proc. ICSLP-2002*, pp. 73-76, 2001.

[55] K. Ohkura, M. Sugiyama, and S. Sagayama, "Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs", *Proc. ICSLP '92*, 369-372, 2002.

[56] 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*, vol. 6, pp. 71-77, 1998.

[57] M. Padmanabhan, G. Saon, and G. Zweig, "Lattice-based unsupervised MLLR for speaker adaptation," *Proc. ISCA ITR-Workshop 2000*, pp. 128-131, 2000.

[58] M. Pitz, S. Molau, R. Schluter, and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *Proc. Eurospeech-2001*, 2001.

[59] D. Pye and P .C. Woodland, “Experiments in speaker normalization and adaptation for large vocabulary speech recognition,” in *Proc. ICASSP97*, vol. 2, pp. 1047-1050, 1997.
[60] M. Rahim and B.-H. Juang, “Signal bias removal by maximum likelihood estimation for robust telephone speech recognition,” *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 1, pp.19-30, 1996.
[61] S. Sagayama, Y. Yamaguchi, S. Takahashi, J. Takahashi, “Jacobian approach to fast acoustic model adaptation,” *Proc. ICASSP ’97*, pp. 835-838, 1997.
[62] S. Sagayama, K. Shinoda, M. Nakai, H. Shimodaira, “Analytic methods for acoustic model adaptation: a review,” *ISCA ITR-Workshop*, Sophia-Antiplois, pp. 67-76, 2001.
[63] A. Sankar and C-.H. Lee, “A Maximum likelihood approach to stochastic matching for robust speech recognition,” *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 3, pp. 190-202, 1996.
[64] K. Shinoda, K. Iso, and T. Watanabe, “Speaker adaptation for demi-syllable-based continuous-density HMM,” *Proc. ICASSP-91*, pp. 857-860, Toronto, 1991.
[65] K. Shinoda and C.-H. Lee, ”Structural MAP speaker adaptation using hierarchical priors,” *Proc. of IEEE Workshop on Speech Recognition and Understanding*, 1997.
[66] K. Shinoda and C.-H. Lee, “A structural Bayes approach to speaker adaptation,” *IEEE Trans. Speech Audio Processing*, vol. 9, no. 3, pp. 276-287, 2001.
[67] K. Shinoda, “(Invited Paper) Acoustic Model Adaptation for Speech Recognition,” IEICE Transaction on Information and Systems, vol. E93-D, no. 9, pp. 2348-2362, 2010.
[68] T. Shinozaki, Y. Kubota, and S. Furui “Unsupervised cross-validation adaptation algorithms for Improved adaptation performance,” *Proc. ICASSP ’09*, pp.4377-4380, 2009.
[69] O. Siohan, T.-A. Myrvoll, and C.-H. Lee, “ Structural maximum a posteriori linear regression for fast HMM adaptation,” *In Workshop on ISCA ITRW ASR2000*, 2000.
[70] Y. Tang and R. Rose, “Rapid speaker adaptation using clustered maximum-likelihood linear basis with sparse training data,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 607-616, 2008.
[71] S. Tanji, K. Shinoda, S. Furui, and A. Ortega, “Improvement of Eigenvoice-Based Speaker Adaptation by Parameter Space Clustering,” *Proc. Interspeech ’08*, pp. 1229-1232, 2008.
[72] M. Tonomura, T. Kosaka, and S. Matsunaga, “Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation,” *ICASSP-95*, vol. 1, pp. 688-691, Detroit, 1995.
[73] P. C. Woodland, “Speaker adaptation for continuous density HMMs: a review,” *ISCA ITR-Workshop*, Sophia-Antiplois, pp. 11-19, 2001.
[74] S. Yoshizawa, A. Baba, K. Matsunami, Y. Mera, M. Yamada, and K. Shikano, “Unsupervised speaker adaptation based on the sufficient HMM statistics of selected speakers,”, *Proc. ICASSP2001*, 2001.
[75] K. Yu and M. Gales, “Discriminative cluster adaptive training,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1694-1703, 2006.
[76] P. Zhan, M. Westohal, “Speaker normalization based on frequency warping,” in *Proc. ICASSP97*, pp.1039-1042, 1997.
[77] G. Zavaliagkos, R. Schwartz, and J. Makhoul, “Batch, incremental and instantaneous adaptation techniques for speech recognition,” *Proc. ICASSP-95*, pp. 676-679, Detroit, May. 1995.
[78] Z. Zhang and S. Furui, “Piecewise-linear transformation-based HMM adaptation for noisy speech,” *Speech Communication*, vol. 42, pp. 43-58, 2004.