SPEAKER ADAPTATION THROUGH SPEAKER SPECIFIC COMPENSATION

Srivatsan Laxman and P. S. Sastry

Dept. of Electrical Engineering
Indian Institute of Science, Bangalore
Email: {srivats,sastry}@ee.iisc.ernet.in

ABSTRACT

This paper describes a new speaker adaptation strategy that we term speaker specific compensation. The basic idea is to transform speech of a speaker in a way that renders it recognizable by a speaker dependent classifier built for another speaker. The compensating filter is learnt as a cepstral vector using labeled speech samples of the speaker. Using some ideas about combining multiple pattern classifiers, we present a new speaker independent speech recognition system that uses a few speaker dependent classifiers along with a bank of cepstral compensating vectors learnt for a large number of other speakers. Each of the speaker dependent classifiers is trained on the given speech samples of only one speaker and is never retrained or adapted thereafter. We present some results to illustrate the effectiveness of this speaker specific compensation idea.

1. INTRODUCTION

The primary challenge in Automatic Speech Recognition (ASR) over the years has been one of generalizing over speech patterns from multiple speakers. A generic class of techniques that address this issue goes under the name of *speaker adaptation*. Typically in speaker adaptation, a preliminary speaker independent (SI) recognition system is first trained using speech from a few speakers and is then gradually adapted to new speakers. This can be done either in the feature space or in the model space (and some times in both). Feature space adaptation seeks to *normalize* speech from different speakers to the training space of the initial classifier [1, 2]. Such normalization, it is hoped, will yield speech representations without speaker specific variabilities. In model space adaptation, the prototype SI model is adapted to account for the new speaker variabilities [3, 4]. This can yield either a new classifier for the new speaker or simply an updated version the existing classifier. Either way, it is expected that such adaptation enhances the speaker generalization capability of the eventual ASR system.

This paper proposes a new method for speaker adaptation that is based on what is referred to as *speaker specific compensation* [5]. The approach here, unlike in the techniques cited above, is to start with a few *speaker dependent* speech classifiers and then learn *speaker specific* transforms that take the speech from one speaker space to another. Suppose a speaker dependent speech classifier is built for one speaker (who we shall call as a *reference* speaker). To adapt the system to a new speaker we transform his/her speech to the reference speaker space in a way that enables it's recognition by the initial speaker dependent ASR system built for the reference speaker (without any further adaptation whatsoever of the classifier). This transformation is achieved by learning what is referred to as a Cepstral Compensating Vector (CCV) for the new speaker. This basic compensation structure has many interesting applications. For example, in [6], this framework was used in a text-dependent speaker recognition context. In this paper we present a method to build speaker independent ASR systems by combining the speaker specific CCVs together with a few initial speaker dependent classifiers.

The general idea of using speaker to speaker transforms for building ASR systems, though not extensively explored, is not an entirely new one either. For example, spectral mapping techniques have been used for such transformations in the context of codebook adaptation [7] or data augmentation [8]. Our idea of CCV-based compensation is significantly different from such approaches and yields a much more general purpose recognition system. Another interesting technique that has recently been proposed [9] regards the speaker space as a linear subspace spanned by a few (appropriately learnt) *eigenvoices*. Our speaker specific compensation idea is somewhat related (in spirit) to this eigenvoices approach in that it also seeks a robust representation of the speaker space using a few speaker dependent ASR systems and some speaker specific transforms.

2. SPEAKER SPECIFIC COMPENSATION

Consider a speaker $\mathcal{R}$. Let $\Phi_{\mathcal{R}}$ be a speaker dependent classifier built for the speech of $\mathcal{R}$ that works with LP cepstral features. (In this paper $\Phi_{\mathcal{R}}$ is a set of HMMs, one HMM per speech unit class). Now suppose the speech of another

0-7803-8674-4/04/$20.00 ©2004 IEEE

speaker, $\mathcal{S}$, needs to be recognized. Let $\mathbf{X}$ represent the LP cepstral vector sequence corresponding to a speech pattern of $\mathcal{S}$. $\mathcal{R}$ is referred to as a "reference" speaker and $\mathcal{S}$ as a "registered" speaker.

The basic compensation idea is to look for a *linear* filter, $\mathcal{M}^*_{\mathcal{SR}}$, that will transform the speech of $\mathcal{S}$ to a representation that renders it recognizable by $\mathbf{\Phi}_{\mathcal{R}}$ itself, i. e. without retraining $\mathbf{\Phi}_{\mathcal{R}}$ on any (not even the transformed) speech of $\mathcal{S}$. Since a convolution in time domain manifests as an addition in cepstral domain, the compensated feature vector sequence, $\mathbf{X}_{\mathcal{SR}}$, resulting from the action of $\mathcal{M}^*_{\mathcal{SR}}$ on the speech of $\mathcal{S}$ can be written as

$$\mathbf{X}_{\mathcal{SR}} = \mathbf{X} \oplus \widehat{\mathbf{M}}^*_{\mathcal{SR}} \tag{1}$$

where $\widehat{\mathbf{M}}^*_{\mathcal{SR}}$ denotes the cepstral vector corresponding to the linear filter $\mathcal{M}^*_{\mathcal{SR}}$ and the '$\oplus$' denotes an addition of each constituent vector in the sequence $\mathbf{X}$ with the vector $\widehat{\mathbf{M}}^*_{\mathcal{SR}}$. The adaptation strategy uses training examples of speaker $\mathcal{S}$ to learn an "optimal" compensating filter such that the best possible (post compensation) recognition performance is attained (over the training set of $\mathcal{S}$) using the speaker specific classifier $\mathbf{\Phi}_{\mathcal{R}}$. The adaptation process is detailed in Sec. 4 and the (optimal) cepstral vector, $\widehat{\mathbf{M}}^*_{\mathcal{SR}}$, so obtained, is referred to as the Cepstral Compensating Vector (CCV) of $\mathcal{S}$ with respect to $\mathcal{R}$.

Thus, for each speaker $\mathcal{S}$, whose speech the system must recognize, an $\widehat{\mathbf{M}}^*_{\mathcal{SR}}$, is learnt with respect to the reference speaker $\mathcal{R}$, using the adaptation data available for $\mathcal{S}$. Once this is done, any speech pattern of $\mathcal{S}$ may be recognized by first compensating the corresponding cepstral sequence using the CCV $\widehat{\mathbf{M}}^*_{\mathcal{SR}}$ and then recognizing it using $\mathbf{\Phi}_{\mathcal{R}}$.

The next question is how can this basic "compensate and classify" step be used as a building block for designing a robust speech recognition system? We would have as many CCVs as there are registered speakers. When the speaker identity is given, the speech of any of the registered speakers can be recognized by first compensating with the *appropriate* CCV and then classifying it with $\Phi_{\mathcal{R}}$. If the speaker identity is not known, then we can use each CCV to compensate in turn and thus generate an ensemble of classifier decisions. These are to be then combined suitably to arrive at the final classification decision. Another important issue is that of the choice of reference speaker. In general, it is unreasonable to assume that the speech of any registered speaker can be transformed to render it recognizable by the speaker dependent classifier built for one (reference) speaker. Hence we would have a number of reference speakers as well and learn a CCV for every pair of registered and reference speakers. Once again the system would essentially be an ensemble of classifiers and we need a suitable classifier combining strategy. The next section describes a compensation-based speech recognition system designed using these general ideas.

3. COMPENSATION-BASED ASR

Our speech recognition system is organized as follows. Let $\mathcal{J}_{\mathbf{r}} = \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_{N_\mathbf{r}}\}$ denote the set of reference speakers. For each of these reference speakers, individual speaker dependent classifiers, $\mathbf{\Phi}_{\mathcal{R}_1}, \ldots, \mathbf{\Phi}_{\mathcal{R}_{N_\mathbf{r}}}$, are built using LP cepstral features. Suppose we have a set of $N_\mathbf{s}$ registered speakers denoted by $\mathcal{J}_\mathbf{s} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{N_\mathbf{s}}\}$. (Typically, $N_\mathbf{r} \ll N_\mathbf{s}$). For each pair $(s, r) \in \mathcal{J}_\mathbf{s} \times \mathcal{J}_\mathbf{r}$, there is a CCV learnt, thus giving rise to a collection of compensating vectors, $\mathcal{C} = \{\widehat{\mathbf{M}}^*_{sr}, s \in \mathcal{J}_\mathbf{s}, r \in \mathcal{J}_\mathbf{r}\}$. The individual speaker dependent classifiers built for the reference speakers along with the set of CCVs, form an ensemble of classifiers: $\mathcal{D} = \{\mathbf{\Phi}^s_r, \; r \in \mathcal{J}_\mathbf{r}, s \in \mathcal{J}_\mathbf{s}\}$. On input $\mathbf{X}$ (which is a sequence of cepstral vectors) the decision of the classifier $\mathbf{\Phi}^s_r$ is obtained as

$$\mathbf{\Phi}^s_r(\mathbf{X}) \stackrel{\text{def}}{=} \mathbf{\Phi}_r(\mathbf{X} \oplus \widehat{\mathbf{M}}^*_{sr}) \in \Omega, \mathbf{X} \in \mathcal{X}, \tag{2}$$

where $\mathcal{X}$ is the space of cepstral vector sequences and Ω is the set of all speech unit labels. Thus, $\mathbf{\Phi}^s_r(\cdot)$ represents a decision rule that first compensates an input feature sequence $\mathbf{X}$, with the CCV $\widehat{\mathbf{M}}^*_{sr}$, and then classifies using the speaker specific classifier $\mathbf{\Phi}_r$. Next, we have to decide how to combine the decisions from the ensemble of classifiers, $\mathcal{D}$, to arrive at the final classification.

Combining individual opinions to arrive at a final decision has attracted a lot of interest from the pattern recognition community in recent times. In [10], Kittler, et. al. develop a common theoretical framework for combining classifiers and experimentally compare some of the combining schemes that are frequently in use. The simplest of combining strategies is a *majority vote*. Here, each classifier adds a weight of one to the class label corresponding to its decision and zero to all others. The final class label would be the one with highest weight. A minor modification of this is to add weights to votes of individual classifiers. That is, each classifier adds a non-zero weight to the label corresponding to its decision and this weight is, in general, determined by the confidence that the classifier has in its decision. The final class label is once again the one with highest weight. The strategy we employ is such a weighted voting method.

Each component classifier, Φ^s_r, in the ensemble $\mathcal{D}$, is a set of HMMs. For any input pattern, HMM outputs are essentially posterior probabilities, evaluated at the feature sequence corresponding to the given pattern. We use the *winning margin* of the winning HMM as the weight for the decision of each individual classifier, Φ^s_r. This is found to perform well for the recognition tasks that we address.

In the next two subsections we describe the operation of our ASR system with and without the speaker identity information. We refer to these as *supervised* and *unsupervised* ASR respectively.

3.1. Supervised ASR

In supervised ASR, the recognition system is supplied with speaker identity information. Let $\mathcal{S} \in \mathcal{J}_s$ denote the speaker identity corresponding to the given speech feature sequence $\mathbf{X} \in \mathcal{X}$. Let η_r be the class label defined by

$$\eta_r = \mathbf{\Phi}_r^{\mathcal{S}}(\mathbf{X}),\ r \in \mathcal{J}_r. \tag{3}$$

The final class label, ω^*, for $\mathbf{X}$, is obtained as follows:

$$\omega^* = \arg\max_{\omega\in\Omega}\left[\sum_{r=1}^{N_r}\mathcal{I}_{\eta_r}(\omega)\right] \tag{4}$$

where $\mathcal{I}_{\eta_r}(\omega)$ is a function whose value is the winning margin at $\omega = \eta_r$ and 0 for all $\omega \neq \eta_r$.

Thus, when we are given the identity of the speaker (as $\mathcal{S}$), we combine the decisions of only a subset of classifiers in our ensemble given by $\mathcal{D}_{\mathcal{S}} = \left\{\mathbf{\Phi}_1^{\mathcal{S}}, \mathbf{\Phi}_2^{\mathcal{S}}, \ldots, \mathbf{\Phi}_{N_r}^{\mathcal{S}}\right\} \subset \mathcal{D}$, which are the only classifiers that pertain to the speaker under consideration.

This is an interesting new scheme for recognizing the speech of any one of a given collection of speakers in a supervised mode. Limited though this framework might be in its scope for application in speech recognition, it brings out one of the interesting features of our compensation based recognition strategy.

Typically, when there is a requirement of such a recognition system, one would build an HMM-based classifier for each speaker, so that when presented with a speech pattern for recognition (along with its associated speaker identity) the appropriate HMM-based classifier may be invoked. Such a framework requires N_s HMM-based classifiers for a system designed to recognize the speech of N_s speakers. Now we can compare the memory requirements of such a system with those of our compensation based strategy. Let the number of memory elements needed to store the model parameters of one HMM-based classifier be denoted by ϑ_{HMM} and similarly, that needed to store a single CCV be ϑ_{CCV}. N_r denotes the number of reference speakers in the CCV-based supervised ASR framework. Consider the ratio of memory requirements of the CCV-based framework, to that of the "one HMM system per speaker" scheme. This ratio can be written as

$$\frac{N_r\vartheta_{\text{HMM}} + (N_s - N_r)N_r\vartheta_{\text{CCV}} + N_r(N_r - 1)\vartheta_{\text{CCV}}}{N_s\vartheta_{\text{HMM}}} = \frac{N_r}{N_s} + N_r\left(1 - \frac{1}{N_s}\right)\frac{\vartheta_{\text{CCV}}}{\vartheta_{\text{HMM}}} \tag{5}$$

With $1 < N_r \ll N_s$ and $\vartheta_{\text{CCV}} \ll \vartheta_{\text{HMM}}$, the efficiency in our scheme becomes immediately apparent. Some results are quoted in Sec. 5 in support of this claim.

3.2. Unsupervised ASR

In the unsupervised ASR framework the speaker identity information is *not* available when making the speech recognition decision. In such a case, the individual decisions of all classifiers in the ensemble $\mathcal{D}$ can be put to use. Define

$$\eta_{sr} = \mathbf{\Phi}_r^{s}(\mathbf{X}),\ r \in \mathcal{J}_r, s \in \mathcal{J}_s. \tag{6}$$

The final class label, ω^*, for $\mathbf{X}$ is given by

$$\omega^* = \arg\max_{\omega\in\Omega}\left[\sum_{r=1}^{N_r}\sum_{s=1}^{N_s}\mathcal{I}_{\eta_{sr}}(\omega)\right] \tag{7}$$

where once again, $\mathcal{I}_{\eta_{sr}}(\omega)$ is a function over ω whose value is the winning margin at $\omega = \eta_{sr}$ and 0 everywhere else.

Consider the classifier combining strategy specified by Eq. (7). We *do not* seek to explicitly solve for the speaker identity or insist that CCVs associated with only a single registered speaker be used. Instead, we prefer to let the robustness in the speaker space handle the absence of speaker identity information in an indirect way. In this context, we note that the strategy of using a weighted voting has some justification. A winning margin associated with the the classifier that makes the "correct" decision is, in general, always much larger than that associated with classifiers that come to a wrong decision. This is because, whenever a speaker dependent classifier is presented with a feature sequence that is very much outside its speaker space, all its HMM outputs (including that of the winning HMM) are bound to be considerably small. In contrast, when a feature sequence belongs to the speaker space associated with the classifier, by virtue of the HMM training process, we can expect that the winning HMM output will indeed take significantly larger values. This, along with the fact that at least one compensation with respect to each reference speaker can be expected to vote correctly, justifies the combining strategy described in Eq. (7).

4. LEARNING THE CCVS

For registering a speaker $\mathcal{S}$, we need to learn a compensating vector for this speaker with respect to each reference speaker $\mathcal{R} \in \mathcal{J}_r$. The central idea is to search for a cepstral vector to compensate utterances of $\mathcal{S}$ so as to yield good recognition rates on $\mathcal{R}$'s speaker specific classifier.

Consider the problem of learning the CCV, $\widehat{\mathbf{M}}_{\mathcal{SR}}^*$, to transform the speech of $\mathcal{S}$ to render it recognizable by $\mathcal{R}$'s speaker dependent classifier, $\mathbf{\Phi}_{\mathcal{R}}$. Let $\{\mathcal{U}_{\mathcal{S}}^j : j = 1, \ldots, N_{\mathcal{S}}^{tr}\}$ be a collection of labeled training examples available for the registration of $\mathcal{S}$. Let $\{\mathbf{X}_{\mathcal{S}}^j : j = 1, \ldots, N_{\mathcal{S}}^{tr}\}$ represent the corresponding set of LP cepstral vector sequences. We define a cost function $\widetilde{\Psi}_{\mathcal{SR}}(\cdot)$ (over cepstral vectors $\widehat{\mathbf{M}}$) as:

$$\widetilde{\Psi}_{\mathcal{SR}}(\widehat{\mathbf{M}}) = \frac{1}{N_{\mathcal{S}}^{tr}} \sum_{j=1}^{N_{\mathcal{S}}^{tr}} \iota^{j}_{\mathcal{SR}}(\widehat{\mathbf{M}}) \tag{8}$$

where $\iota^{j}_{\mathcal{SR}}(\widehat{\mathbf{M}})$ takes value 1 or 0 depending on whether or not utterance $\mathcal{U}^{j}_{\mathcal{S}}$, when compensated with $\widehat{\mathbf{M}}$ was correctly recognized by $\mathcal{R}$'s speaker specific classifier. Now the best compensator for the ordered pair $(\mathcal{S}, \mathcal{R})$, namely the CCV $\widehat{\mathbf{M}}^{*}_{\mathcal{SR}}$, is defined as

$$\widehat{\mathbf{M}}^{*}_{\mathcal{SR}} = \arg\max_{\widehat{\mathbf{M}}} \widetilde{\Psi}_{\mathcal{SR}}(\widehat{\mathbf{M}}). \tag{9}$$

Due to the nature of the cost function in this optimization problem, estimating or computing the gradient information is a difficult task. Therefore, we used ALOPEX [11], a correlation-based optimization technique that does not need any gradient information, to solve the maximization problem specified in Eq. (9).

5. RESULTS

An isolated word database of English digits was collected for a set of 10 speakers, 5 of which were male and the rest female. The recordings were conducted in a closed room but without any special arrangements to cut down ambient noise. Speech was recorded at 16 KHz using a simple microphone connected to a 32-bit sound card on a multimedia computer. Each speaker uttered the ten words (the digits *zero* through *nine*) in sequence with sufficient pause between words. The recordings were segmented manually to yield an isolated digit database. The database comprises a total of 1500 digit utterances (with 15 repetitions per digit per speaker). Of these, for each speaker, 8 utterances (per digit) were used for training (either the speech classifier of a reference speaker or the CCV of a registered speaker, as the case may be) and the remaining were used as test data. We denote the 10 speakers in our database by the speaker labels $\{01, 02, \ldots, 10\}$. Speakers 06, 07, 08, 09 and 10 are the female speakers in the database.

5.1. Supervised mode ASR

Let us first consider the case of supervised recognition where we use the decision rule described by Eq. (4). Table 1 shows the performance of the supervised ASR framework for some chosen reference speaker sets. It can be seen that the recognition rates improve steadily with increasing number of reference speakers. With 3 reference speakers, we were able to get a recognition accuracy of 95.60%[1].

Reference	Recognition accuracy			Memory
speaker set $\mathcal{J}_r$	Train	Test	All	efficiency index
$\{06\}$	90.50	89.57	**90.07**	0.1043
$\{01, 06\}$	95.13	93.86	**94.53**	0.2086
$\{01, 06, 10\}$	96.13	95.00	**95.60**	0.3130

Table 1. Supervised ASR results

The first row of this table shows that we get 90% accuracy when we use a single female speaker as the reference speaker. This is fairly good considering that there are five male speakers in the database. As a matter of fact, the accuracy obtained when the uncompensated speech of any male speaker is recognized by the classifier built for speaker 06 varied from a low of 25% to a high of 65%. However, using our cepstral compensation (and without any retraining of the classifier), the corresponding post-compensation accuracies for the male speakers were between 75% and 97% [5]. .This we feel fully vindicates our idea of CCV-based speaker adaptation.

In Section 3.1 we discussed the ratio of memory requirements of our scheme with that of the "one HMM system per speaker" framework for supervised ASR. This fraction (defined in Eq. (5)) is tabulated under the heading of *memory efficiency index* in Table 1. Using the same notation as in Section 3.1, we have in the simulation experiments, $\vartheta_{\text{CCV}} = 20$ and $\vartheta_{\text{HMM}} = 4200$. Since $\vartheta_{\text{CCV}} \ll \vartheta_{\text{HMM}}$, the memory efficiency index is well approximated by the ratio of the number of reference speakers to the total number of speakers, as can be seen from the table. This index is expected to improve significantly with increase in the number of registered speakers. For instance, with about a hundred registered speakers it is expected that only about ten or fifteen reference speakers would be needed to achieve acceptable recognition performance levels.

5.2. Unsupervised mode ASR

Next we present the results of the unsupervised ASR framework described in Section 3.2. So as to appreciate the role of multiple reference speakers, every subset of $\{01, 06, 10\}$ has been considered as a choice for the reference speaker set and the results are shown in Table 2.

As is to be expected, in the unsupervised mode, the recognition accuracy is poor when we have only a single reference speaker. However, recognition rates improve with increasing cardinality of the set $\mathcal{J}_r$. What is significant to note here is that any two reference speakers perform better together than any single reference speaker. Similarly,

[1] We note that by adding more features and doing specialized training it is possible to improve these recognition rates. But since the objective here is to explore effectiveness of the compensation idea, we used 15-dimensional LP cepstrum as the only features and simple discrete HMMs for the speaker dependent classifiers.

Reference	Recognition accuracy		
speaker set $\mathcal{J}_r$	Train	Test	**All**
$\{01\}$	72.66	75.18	**73.83**
$\{06\}$	79.06	78.57	**78.83**
$\{10\}$	75.78	77.86	**76.75**
$\{01, 06\}$	88.28	86.43	**87.42**
$\{01, 10\}$	85.31	83.93	**84.67**
$\{06, 10\}$	86.72	86.43	**86.58**
$\{01, 06, 10\}$	90.62	90.18	**90.42**

Table 2. Unsupervised ASR results

when all three reference speakers are used, the recognition rate is better than that achieved with every possible reference speaker pair. Such improvements in performance with increasing number of reference speakers, is not only useful in our system design, but is also very intuitively satisfying. The robustness in speaker space representation (achieved through the collection of CCVs and speaker dependent classifiers) provides the needed speaker generalization.

6. CONCLUSIONS

This paper describes how speaker specific compensation can be used to achieve speaker independent speech recognition using only speaker dependent classifiers. In this framework, each speech classifier is trained on the speech of only one speaker and is never retrained. The CCVs allow us to transform speech of other speakers into the target feature space of the speaker dependent classifiers. Thus, the ASR system proposed in this paper uses an ensemble of speaker dependent classifiers and speaker specific compensators in a *classifier combining* framework to achieve significant speaker robustness, which we believe is interesting and novel. Further, we think that by judiciously choosing reference speakers and then learning CCVs for many more (well-chosen) registered speakers, our ensemble of classifiers can cover the speaker space well enough to give a truly speaker independent classifier system. As was mentioned earlier, in [6], we described a text-dependent speaker recognition system based on this compensation idea. Consequently, probably the most interesting application of this idea of speaker specific compensation is the possibility of simultaneous speech-cum-speaker recognition achievable within a single unified framework. This and other aspects of the compensation idea will be described in our future publications.

7. REFERENCES

[1] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, Jan. 1998.

[2] P. C. Loizou and A. Spanias, "Improved speech recognition using a subspace projection approach," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 343–345, May 1999.

[3] C. H. Lee, C. H. Lin, and B. H. Juang, "A study on speaker adaptation of the parameters of continuous density Hidden Markov Models," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 806–814, Apr. 1991.

[4] Lutz Welling, Hermann Ney, and Stephan Kanthak, "Speaker adaptive modeling by vocal tract normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 415–426, Sept. 2002.

[5] Srivatsan Laxman, "Speaker specific compensation for speech recognition," M.S. thesis, Indian Institute of Science, Bangalore, 2002.

[6] Srivatsan Laxman and P. S. Sastry, "Text-dependent speaker recognition using speaker specific compensation," in *TENCON 2003, IEEE Region 10 Conference on Convergent Technologies for the Asia-Pacific, Oct. 15-17, India*, 2003.

[7] Francis Kubala, Richard Schwartz, and Chris Barry, "Speaker adaptation from a speaker independent training corpus," in *ICASSP*, 1990, pp. 137–140.

[8] J. R. Bellegarda, P. V. de Souza, A. Nadas, D. Nahamoo, M. A. Picheny, and L. R. Bahl, "The metamorphic algorithm: A speaker mapping approach to data augmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 413–420, July 1994.

[9] R. Kuhn, J. C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–706, Nov. 2000.

[10] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[11] P. S. Sastry, M. Magesh, and K. P. Unnikrishnan, "Two timescale analysis of the alopex algorithm for optimization," *Neural Computation*, vol. 14, pp. 2729–2750, 2002.