Text-dependent speaker recognition using speaker specific compensation

Srivatsan Laxman and P.S.Sastry
Dept. Electrical Engineering
Indian Institute of Science
Bangalore 560012, INDIA

ABSTRACT

This paper proposes a new method for text-dependent speaker recognition. The scheme is based on learning (what we refer to as) speaker-specific compensators for each speaker in the system. The compensator is essentially a speaker to speaker transformation which enables the recognition of the *speech* of one speaker through a speaker-dependent speech recognition system built for the other. Such a transformation, adequate for our purposes, may be achieved by a simple vector addition in the cepstral domain. This speaker-specific compensator captures the characteristics of the speaker we wish to recognize. For each speaker who is registered into the system, we learn a unique set of compensators. The speaker recognition decision is then based on which compensator achieves best speech recognition scores.

1. INTRODUCTION

Text-dependent speaker recognition seeks to associate an unknown speaker with a member from a (registered) population, given a textual transcription of the phrases uttered by the speaker [1], [2]. Typically, speaker-dependent word or sub-word models are built for each speaker. Given a labeled utterance from an unknown speaker, the system makes its speaker recognition decision based on the likelihood scores of the appropriate speaker-dependent models.

This paper proposes a new way to build speaker-dependent models using a technique called speaker specific compensation [3]. This technique can be summarized as follows. A speaker-dependent speech classifier is built for one speaker (referred to as the reference speaker). For each new speaker that we want to register, we look for a transformation on his speech which enables its recognition through the reference speaker's speech classifier. This transformation, in combination with the already trained speech classifier (of the reference speaker) together constitute the speaker model for the new speaker. Assuming that a linear transformation in the time domain could be sufficient as the transformation for our recognition purposes, a simple cepstral domain addition achieves the required speaker specific compensation. Thus we only need to learn one cepstral vector for each new speaker whom we wish to register into the system. This way, we do not need to build elaborate speaker-dependent word or sub-word models for each new speaker.

We note here that speaker specific compensation, in general, has several interesting applications in speech and speaker recognition. The compensation idea discussed here lends itself to a variety of recognition frameworks [3]. In this paper however, we restrict the discussion to just the speaker recognition scenario.

2. THE COMPENSATION IDEA

Consider a speaker $\mathcal{R}$. At the heart of our speaker recognition framework is a speaker dependent *speech* classifier $\Phi_{\mathcal{R}}$ for $\mathcal{R}$ that works with LP cepstrum features. $\Phi_{\mathcal{R}}$ is trained only to recognize the speech of $\mathcal{R}$. Suppose that we now wish to register a new speaker, say $\mathcal{S}$, into our speaker recognition system. Let $\mathbf{X}_{\mathcal{S}}$ represent the LP cepstral vector sequence corresponding to a (labeled, training) speech pattern of $\mathcal{S}$. We shall refer to $\mathcal{R}$ as a "reference" speaker and to $\mathcal{S}$ as a "registered" speaker.

The basic compensation idea is to look for a single cepstral vector $\widehat{\mathbf{M}}^*_{\mathcal{SR}}$ that will transform the speech of $\mathcal{S}$ to a representation that renders it recognizable by $\Phi_{\mathcal{R}}$ itself. The "compensated" feature vector sequence $\mathbf{X}_{\mathcal{SR}}$ that results from this transformation is written as

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

where the '$\oplus$' denotes an addition of each constituent vector in the sequence $\mathbf{X}_{\mathcal{S}}$ with the vector $\widehat{\mathbf{M}}^*_{\mathcal{SR}}$. (It may be noted that in Eq. (1) $\widehat{\mathbf{M}}^*_{\mathcal{SR}}$ is a single vector while $\mathbf{X}_{\mathcal{S}}$ is a sequence of vectors.) Such a vector $\widehat{\mathbf{M}}^*_{\mathcal{SR}}$, is referred to as a Cepstral Compensating Vector (CCV) for speaker $\mathcal{S}$ (with respect to speaker $\mathcal{R}$).

From a speaker recognition perspective, this $\widehat{\mathbf{M}}^*_{\mathcal{SR}}$ constitutes the speaker model for $\mathcal{S}$. Thus we need one such model for each speaker who is registered into the system. Then, the speaker recognition rule simply asks which (registered) speaker's CCV, when used to compensate the given labeled speech, achieves highest recognition scores on the speech recognition engine $\Phi_{\mathcal{R}}$.

The registration strategy we propose uses training examples of speaker $\mathcal{S}$ to learn an "optimal" compensator such that the best possible (post compensation) speech recognition performance is attained (over the training set of $\mathcal{S}$) using the speaker specific (speech) classifier $\Phi_{\mathcal{R}}$. (It may be noted that $\Phi_{\mathcal{R}}$ is

built using speech samples of only $\mathcal{R}$ and is never retrained) The registration process is detailed later in Section 4. Before that, we first describe the speaker recognition framework in detail, bearing in mind that the registration process will yield a CCV for each registered-reference speaker pair in the system.

3. TEXT DEPENDENT SPEAKER RECOGNITION

Let $\Phi_{\mathcal{R}}$ denote the speaker dependent speech classifier for $\mathcal{R}$ (based on LP cepstral features). In this paper, this classifier is a simple isolated word recognition system. The classifier is built by learning a set of HMMs from some training examples of speaker $\mathcal{R}$. One HMM is trained for each word in the system's vocabulary. Consider the set $\mathcal{J}_{\mathbf{s}} \stackrel{\text{def}}{=} \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{N_{\mathbf{s}}}\}$ of speakers who are *registered* with respect to $\mathcal{R}$. The speaker $\mathcal{S}$ is said to be *registered* with respect to $\mathcal{R}$, if a CCV $\widehat{\mathbf{M}}^*_{\mathcal{SR}}$ has been learnt for the speaker pair $(\mathcal{S}, \mathcal{R})$. Notice that the CCV, $\widehat{\mathbf{M}}^*_{\mathcal{RR}}$, for the speaker pair $(\mathcal{R}, \mathcal{R})$ is simply the *zero* vector. We denote by $\mathcal{C}_{\mathcal{R}}$ (which is a collection of CCVs for each $\mathcal{S} \in \mathcal{J}_{\mathbf{s}}$) the *bank of compensators* with respect to the reference speaker $\mathcal{R}$.

Given the label ω of the uttered speech unit (represented as a sequence of cepstral vectors $\mathbf{X}$) our recognition system would decide the speaker identity as $\mathcal{S}$ if the likelihood of ω (under the classifier $\Phi_{\mathcal{R}}$) for the *compensated* sequence $(\mathbf{X} \oplus \widehat{\mathbf{M}}^*_{\mathcal{SR}})$ is better than that for $(\mathbf{X} \oplus \widehat{\mathbf{M}}^*_{s\mathcal{R}})$ $\forall s \neq \mathcal{S}$. If the best likelihood itself is below a threshold the system would reject the input speaker.

It is easy to see that the problem of speaker recognition when multiple utterances (of the same speaker) are input to the system, is one of combining speaker label opinions after repeated application of this rule for each utterance. A similar problem of combining classifiers has to be tackled when more than one reference speaker is used in the system. Hence, we discuss this issue next.

The recognition system described above is feasible only under the assumption that any speaker that needs to be registered can be compensated for with respect to the single chosen reference speaker. In practice, this assumption need not hold. Speakers with vastly different accents may not be compensable with respect to one another. One way to tackle this is to have more than one reference speaker.

Suppose we choose to have $N_{\mathbf{r}}$ reference speakers. For each reference speaker we build a (speaker specific) speech classifier by training it using only her speech. Let $\mathcal{J}_{\mathbf{s}}$ denote the set of $N_{\mathbf{s}}$ registered speakers and $\mathcal{J}_{\mathbf{r}} \subset \mathcal{J}_{\mathbf{s}}$ denote the set of $N_{\mathbf{r}}$ reference speakers. Typically, it is desired that $N_{\mathbf{r}} \ll N_{\mathbf{s}}$. The reference speakers in the system may be enumerated as $\mathcal{J}_{\mathbf{r}} \stackrel{\text{def}}{=} \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_{N_{\mathbf{r}}}\}$. Now for each registered speaker $s \in \mathcal{J}_{\mathbf{s}}$ we learn (not one CCV, but) a set of CCVs, $\{\widehat{\mathbf{M}}^*_{sr} : r \in \mathcal{J}_{\mathbf{r}}\}$.

The idea behind using multiple reference speakers in the system is that, now in the speaker space, each registered speaker in $\mathcal{J}_{\mathbf{s}}$ is "close" to at least one reference speaker in $\mathcal{J}_{\mathbf{r}}$ (if not more). This will make it more likely for us to obtain at least one satisfactory CCV for every speaker and consequently his speech will be recognizable by one reference classifier or another. Essentially, we plant *sufficiently* many reference speakers and register still many more with respect to those in the reference speaker set, so that we have a very good description of the speaker space in general. This interpretation resembles, in some sense, the *eigenvoice* approach for speech recognition proposed in [4].

Thus there are $N_{\mathbf{s}}$ CCVs associated with each of the $N_{\mathbf{r}}$ speaker-dependent speech classifiers in the system. Each CCV and its associated reference speaker speech recognizer may now together be regarded as one classifier. Consequently, the speaker identity decision in such a scenario of multiple reference speakers must be based on the outputs from many classifiers.

In our speaker recognition system, each *CCV-reference classifier* pair may be used to determine the likelihoods of various speech label possibilities. That is, given an isolated word (speech utterance) its corresponding cepstral sequence $\mathbf{X}$ is obtained and then, for each registered speaker s and reference speaker r, a compensated cepstral sequence $(\mathbf{X} \oplus \widehat{\mathbf{M}}^*_{sr})$ is computed. Each such compensated sequence is input to the appropriate speaker-specific classifier Φ_r. For a text-dependent speaker recognition system, the correct speech label of the input utterance is known. Further, each reference classifier is essentially a system of HMMs and thus, for a given input utterance, the winning HMM is one that outputs the highest likelihood. Using this information, the various post-compensation likelihoods need to be appropriately combined to yield a speaker recognition decision.

Our classifier combining strategy may be described as follows. Each of our speech recognizers, Φ_r, is a HMM-based classifier. In our text-dependent speaker recognition system we use the isolated word recognition framework. Thus, $\Phi_{\mathcal{R}}$ consists of one HMM for each word in the vocabulary. When an input speech utterance is subjected to a particular CCV-reference classifier pair, it may or may not result in the correct speech label (which is known to the text-dependent speaker recognition system). When correct, the winning margin of the winning HMM may be used as a confidence measure (or a real-valued vote) in support of the *registered speaker* associated with the CCV that was used for the compensation. We are looking for that speaker label which accumulates the highest net evidence (obtained after employing all the CCV-reference classifier pairs in the system).

The recognition system

In text-dependent speaker recognition, we are provided with a sequence of labeled speech utterances (of an unknown but registered speaker) as input and the output is a decision on the speaker identity. Each labeled utterance is represented by a sequence of feature vectors.

Let $\mathcal{Z}$ represent a set of $N_{\mathcal{Z}}$ speech feature sequences (corresponding to $N_{\mathcal{Z}}$ utterances) of some unknown speaker with $\Omega_{\mathcal{Z}}$ being the corresponding set of utterance labels. These sets are described using the following notation:

$$\mathcal{Z} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{N_{\mathcal{Z}}}\} \quad (2)$$

$$\Omega_{\mathcal{Z}} = \{\omega_1^*, \omega_2^*, \ldots, \omega_{N_{\mathcal{Z}}}^*\}. \quad (3)$$

For $s \in \mathcal{J}_s$ and $r \in \mathcal{J}_r$, let η_{sr}^n denote the speech unit label obtained when $\mathbf{X}_n \in \mathcal{Z}$ was subjected to the classifier $\mathbf{\Phi}_r$ after compensation using the CCV, $\widehat{\mathbf{M}}_{sr}^*$. Each of our speaker recognizers is a HMM-based classifier. Hence η_{sr}^n is obtained by supplying the feature vector sequence $(\mathbf{X}_n \oplus \widehat{\mathbf{M}}_{sr}^*)$ to $\mathbf{\Phi}_r$ and getting the identity of the HMM with the largest output. For the classifier decision η_{sr}^n, a score function $\mathcal{I}_{\omega_n^*}(\eta_{sr}^n)$ is defined, which is equal to zero if $\eta_{sr}^n \neq \omega_n^*$ and equal to the winning margin of the winning HMM if $\eta_{sr}^n = \omega_n^*$. For a single speaker s we get N_r such scores (for one utterance) by using each of the reference speakers and the corresponding CCVs. Thus for a single utterance $\mathbf{X}_n$, we take the combined score for speaker s as $\sum_{r=1}^{N_r} \mathcal{I}_{\omega_n^*}(\eta_{sr}^n)$. Since there are $N_{\mathcal{Z}}$ such utterances, the final score for s is obtained by summing over all utterances. This final score is computed for each speaker and the final classification decision is based on which speaker achieves the highest score. Thus, the system would decide s^* as the speaker identity when

$$s^* = \arg\max_{s \in \mathcal{J}_s} \left[\sum_{n=1}^{N_{\mathcal{Z}}} \sum_{r=1}^{N_r} \mathcal{I}_{\omega_n^*}(\eta_{sr}^n) \right] \quad (4)$$

If the value of this maximum score obtained above is below a threshold, the system rejects the input speaker as unknown.

Notice that, here we are indeed insisting on using compensation associated with only a single registered speaker. In other words, if s^* were to be the speaker identity decision, then we demand that, on restricting the system to using only his compensators (with respect to the various reference speakers in $\mathcal{J}_r$) very good speech recognition rates are achieved over the set $\mathcal{Z}$ of input speech patterns.

4. Speaker Registration

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 compensator which transforms the utterances of $\mathcal{S}$ to representations that yield good recognition rates on $\mathcal{R}$'s speaker specific classifier. We use ALOPEX, an optimization technique that works without any gradient information (of the error surface) to obtain the cepstral vector that best compensates $\mathcal{S}$ with respect to $\mathcal{R}$.

Let $\{\mathcal{U}_{\mathcal{S}}^j : j = 1, \ldots, N_{\mathcal{S}}^{tr}\}$ be a collection of independent identically distributed (iid) labeled training examples available for the registration of speaker $\mathcal{S}$. Let $\{\mathbf{X}_{\mathcal{S}}^j : j = 1, \ldots, N_{\mathcal{S}}^{tr}\}$ represent the associated set of corresponding LP cepstral vector sequences. A cost function $\widetilde{\Psi}(\cdot)$ may be specified (as a function of the cepstral vector $\widehat{\mathbf{M}}_{\mathcal{SR}}$) as follows:

$$\widetilde{\Psi}(\widehat{\mathbf{M}}_{\mathcal{SR}}) = \frac{1}{N_{\mathcal{S}}^{tr}} \sum_{j=1}^{N_{\mathcal{S}}^{tr}} \iota_{\mathcal{SR}}^j \quad (5)$$

where $\iota_{\mathcal{SR}}^j$ is a 0-1 function which indicates whether or not utterance $\mathcal{U}_{\mathcal{S}}^j$ was correctly recognized by $\mathcal{R}$'s speech classifier, after cepstral compensation of the corresponding $\mathbf{X}_{\mathcal{S}}^j$ using $\widehat{\mathbf{M}}_{\mathcal{SR}}$. The optimization problem may now be formulated as one that searches for the CCV $\widehat{\mathbf{M}}_{\mathcal{SR}}^*$ in the cepstral vector space such that

$$\widehat{\mathbf{M}}_{\mathcal{SR}}^* = \arg\max_{\widehat{\mathbf{M}}_{\mathcal{SR}}} \widetilde{\Psi}(\widehat{\mathbf{M}}_{\mathcal{SR}}). \quad (6)$$

Due to the nature of the cost function in the optimization problem (specified in Eq. (6)) estimating or computing the gradient information is a difficult task. Therefore, we used ALOPEX, a correlation-based optimization technique, to solve the maximization problem in Eq. (6) [5].

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 under low ambient noise conditions. Speech was recorded at 16 KHz using a simple microphone connected to a 32-bit sound card on a multimedia computer. A 10 word vocabulary of the digits *zero* through *nine* was used. Each speaker uttered these digits in sequence with sufficient pause between words. In all, 15 such sets were collected for every speaker. Thus, the database comprises a total of 1500 digit utterances (with 15 repetitions per digit per speaker). The recordings were segmented manually to yield an isolated digit database. To facilitate easy description of our experiment and results, we denote the 10 speakers in our database by the speaker labels $\{01, 02, \ldots, 10\}$. The speakers 06, 07, 08, 09 and 10 are the female speakers in the database.

We presented the text-dependent speaker recognition system with sets of 10 digit utterances, all from the same registered speaker. There are 15 such sets for each registered speaker, 8 of which were used to train the CCVs for that speaker.

The results obtained are tabulated in Table 1. These results, although achieved on a small database, are quite impressive.

Reference speaker set $\mathcal{J}_r$	Recognition accuracy		
	Train	Test	**All**
$\{06\}$	84.38	75.00	**80.00**
$\{01, 06\}$	100.0	100.0	**100.0**
$\{01, 06, 10\}$	100.0	100.0	**100.0**

Table 1. Speaker recognition results

With one registered speaker in the system, the speaker recognition rate was only 80.00%. But with multiple reference speakers, we were able to achieve 100% speaker recognition accuracy. These results demonstrate that CCVs can very well be used as the speaker models in speaker recognition.

6. Discussion

This paper illustrates the role that speaker specific compensation can play in text-dependent speaker recognition. The CCVs constitute the concise speaker models needed in such an application. Further, building speaker-dependent ASR models for a single (or at most a few) reference speaker(s) is also relatively easy.

For our text-dependent speaker recognition, we considered the specific case of working with an *isolated word recognition* framework. Restricting the discussion to this simple case facilitated easy development of the compensation idea. However, we feel that our ideas about speaker specific compensation would be equally useful for speaker identification based on continuous speech input as well.

We wish to note here that there is a much wider role that speaker specific compensation can play in speech and speaker recognition [3]. Essentially, the process of learning a CCV can be regarded (in the context of speech recognition) as a speaker adaptation process. With a robust description of the speaker space through a few speaker specific classifiers and several CCVs, impressive speaker independent speech recognition performances may be achieved. In fact, the compensation idea together with some classifier combining ideas, may be exploited to build a simultaneous speech-cum-speaker recognition system as well. We will address these and other applications of speaker specific compensation in our future work.

References

[1] J. P. Campbell, "Speaker ecognition: A tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437–1462, Sept. 1997.

[2] Q. Li, B. H. Juang, C. H. Lee, Q. Zhou, and F. K. Soong, "Recent advancements in automatic speaker authentication," *IEEE Robotics and Automation Magazine*, pp. 24–56, Mar. 1999.

[3] L. Srivatsan, "Speaker specific compensation for speech recognition," Master's thesis, Indian Institute of Science, 2002.

[4] 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, pp. 695–706, Nov. 2000.

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