A Comparison of Normalization and Training Approaches for ASR-Dependent Speaker Identification[1]

Alex Park and Timothy J. Hazen

MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, MA 02139, USA
{malex, hazen}@sls.csail.mit.edu

Abstract

In this paper we discuss a speaker identification approach, called ASR-dependent speaker identification, that incorporates phonetic knowledge into the models for each speaker. This approach differs from traditional methods for performing text-independent speaker identification, such as global Gaussian mixture modeling, that typically ignore the phonetic content of the speech signal. We introduce a new score normalization approach, called phone adaptive normalization, which improves upon our previous speaker adaptive normalization technique. This paper also examines the use of automatically generated transcriptions during the training of our speaker models. Experiments show that speaker models trained using automatically generated transcriptions achieve the same performance as models trained using manually generated transcriptions.

1. Introduction

Traditional methods for performing text independent speaker identification, such as the use of global Gaussian mixture speaker models (GMMs), typically ignore the phonetic content of the speech signal [5]. Although some researchers have sought to address this deficiency through phone-dependent training [1], even these approaches assume no phonetic knowledge of the test utterance. In a previous paper, we described a speaker identification approach that incorporates such phonetic knowledge into the models for each speaker [4]. Specifically, this approach rescores the best sentence hypothesis obtained from a speaker independent speech recognizer using speaker dependent versions of the context-dependent acoustic models used by the recognizer. This method is similar to the LVCSR-based speaker identification approach developed by Dragon Systems and described by Weber *et al.* in [8]. Because our approach relies on the output of an automatic speech recognition (ASR) system, we refer to it as *ASR-dependent speaker identification*. We have incorporated our speaker ID system into conversational systems at MIT to improve both security (to protect confidential user account information) and convenience (to avoid cumbersome login sub-dialogues) [3].

In this paper, we explore two issues associated with ASR-dependent speaker identification. First, we examine the issue of score combination and normalization. Typical speech recognizers have hundreds if not thousands of context-dependent (CD) acoustic models. However, the enrollment data available for any particular speaker may be limited, and therefore many of the CD acoustic models within that speaker's model set may have only a small number of training observations (or even none at all). In this case, appropriate mechanisms for scoring observations of these poorly trained phonetic events during the evaluation of a new utterance must be developed. In this paper we examine a novel score combination and normalization approach called *phone adaptive normalization*, and compare it with the *speaker adaptive normalization* introduced in our previous work.

Second, because our approach relies on the imperfect output of an ASR system, it is important to consider the potential effects of speech recognition errors during both the enrollment and evaluation processes. To this end, we compare the use of accurate manual transcriptions versus inaccurate automatic transcriptions of the enrollment data when training the speaker identification models, and consider how these training approaches are affected by speech recognition errors during evaluation.

2. ASR-Dependent Speaker Identification

We will make use of the following notation when describing the ASR-dependent speaker identification approach and its corresponding normalization methods: Let $\mathbf{X}$ represent the set of feature vectors, $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, extracted from a particular spoken utterance. Let the reference speaker of an utterance $\mathbf{X}$ be $S(\mathbf{X})$. Furthermore, assume that the aligned phonetic transcription of the utterance, $\Phi(\mathbf{X})$, provides a mapping between each feature vector $\mathbf{x}_k$ and it's underlying phonetic unit $\phi(\mathbf{x}_k)$.

In our ASR-dependent approach, each speaker S is represented by a set of models, $p(\mathbf{x}|S,\phi)$, which mirror the CD acoustic models trained for speech recognition, $p(\mathbf{x}|\phi)$, where ϕ ranges over the inventory of phonetic units used in the speech recognizer. The use of a set of context-dependent phonetic models for each speaker is markedly different from global GMM modelling approaches, where the goal is to represent a speaker with a single model, $p(\mathbf{x}|S)$. During evaluation, automatic speech recognition is performed on the utterance producing an automatically generated phonetic transcription, $\hat{\Phi}(\mathbf{X})$, which assigns each vector, $\mathbf{x}_k$, to its most likely phonetic unit, $\hat{\phi}(\mathbf{x}_k)$.

The phone assignments generated during speech recognition can then be used to calculate speaker-dependent phone-dependent conditional probabilities, $p(\mathbf{x}|S,\hat{\phi}(\mathbf{x}))$. Ideally, these probabilities alone would act as suitable *speaker scores* for making a speaker identification decision. For example, the closed-set speaker identification result might be:

$$\hat{S}(\mathbf{X}) = \arg\max_S p(\mathbf{X}|S,\hat{\Phi}(\mathbf{X})) \quad (1)$$

In practice however, enrollment data sets for each speaker are typically not large enough to accurately determine the parameters of $p(\mathbf{x}|S,\phi(\mathbf{x}))$ for all $\phi(\mathbf{x})$.

[1]This work was sponsored in part by an industrial consortium supporting the MIT Oxygen Alliance.

3. Normalization Approaches

In this section, we present two normalization techniques which address the problem of constructing robust speaker scores when enrollment data for each speaker is unevenly distributed over the library of context-dependent phonetic events. The choice of normalization technique becomes especially important when the system is forced to synthesize an appropriate speaker score for a context-dependent phonetic event that has few or no training tokens in the enrollment data.

3.1. Speaker Adaptive (SA) Normalization

We originally described a speaker adaptive normalization approach in [4]. This technique relies on interpolating speaker dependent (SD) probabilities with speaker independent (SI) probabilities on a per-unit basis. This approach learns the characteristics of a phone for a given speaker when sufficient enrollment data is available, but relies more on general speaker independent models in instances of sparse enrollment data.

Mathematically, the speaker score can be written as:

$$Y(\mathbf{X},S)=\frac{1}{|\mathbf{X}|}\sum_{\mathbf{x}\in\mathbf{X}}\log\left[\lambda_{S,\hat{\phi}(\mathbf{x})}\frac{p(\mathbf{x}|S,\hat{\phi}(\mathbf{x}))}{p(\mathbf{x}|\hat{\phi}(\mathbf{x}))}+(1-\lambda_{S,\hat{\phi}(\mathbf{x})})\frac{p(\mathbf{x}|\hat{\phi}(\mathbf{x}))}{p(\mathbf{x}|\hat{\phi}(\mathbf{x}))}\right] \quad (2)$$

Here $\lambda_{S,\hat{\phi}(\mathbf{x})}$ is the interpolation factor given by:

$$\lambda_{S,\hat{\phi}(\mathbf{x})}=\frac{n_{S,\hat{\phi}(\mathbf{x})}}{n_{S,\hat{\phi}(\mathbf{x})}+\tau} \quad (3)$$

In this equation, $n_{S,\hat{\phi}(\mathbf{x})}$ refers to the number of times the CD phonetic event $\hat{\phi}(\mathbf{x})$ was observed in the enrollment data for speaker S, and τ is an empirically determined tuning parameter that was the same across all speakers and phones.

By using the SI models in the denominator of the terms in Equation 2, the SI model set acts as the normalizing *background model* typically used in speaker verification approaches. The interpolation between SD and SI models allows our technique to capture detailed phonetic-level characteristics when a sufficient number of training tokens are available from a speaker, while falling back onto the SI model when the number of training tokens is sparse. In other words, the system backs off towards a *neutral* score of zero when a particular CD phonetic model has little or no enrollment data from a speaker. If an enrolled speaker contributes more enrollment data, the variance of the normalized scores increases and the scores become more reflective of how well (or poorly) a test utterance matches the characteristics of that speaker's model.

3.2. Phone Adaptive (PA) Normalization

An alternative and equally valid technique for constructing speaker scores is to combine phone dependent and phone independent speaker model probabilities. In this scenario, the speaker-dependent phone-dependent models can be interpolated with a speaker-dependent phone-independent model (i.e., a global GMM) for that speaker. Analytically, the speaker score can be described as:

$$Y(\mathbf{X},S)=\frac{1}{|\mathbf{X}|}\sum_{\mathbf{x}\in\mathbf{X}}\log\left[\lambda_{S,\hat{\phi}(\mathbf{x})}\frac{p(\mathbf{x}|S,\hat{\phi}(\mathbf{x}))}{p(\mathbf{x}|\hat{\phi}(\mathbf{x}))}+(1-\lambda_{S,\hat{\phi}(\mathbf{x})})\frac{p(\mathbf{x}|S)}{p(\mathbf{x})}\right] \quad (4)$$

Here, $\lambda_{S,\phi(\mathbf{x})}$ has the same interpretation as before. The rationale behind this approach is to bias the speaker score towards the global speaker model when little phone-specific enrollment data is available. In the limiting case, this approach falls back to scoring with a global GMM model when the system encounters phonetic units that have not been observed in the speaker's enrollment data. This is intuitively more satisfying than the speaker adaptive approach, which backs off directly to the neutral score of zero when a phonetic event is unseen in the enrollment data.

4. Training Approaches

One of the basic assumptions of the ASR-dependent speaker identification approach is that knowledge about the phonetic content of an utterance $\mathbf{X}$ can be gained by performing the assignment $\mathbf{X}\rightarrow\hat{\Phi}(\mathbf{X})$ through automatic speech recognition. If the phonetic recognition error rate is low enough, then we can assume $\hat{\Phi}(\mathbf{X})\approx\Phi(\mathbf{X})$. Under this assumption, we can use $\Phi(\mathbf{X})$, as derived from manual transcriptions, when training our speaker models. This is the approach we took in our previously reported experiments.

Alternatively, instead of using $\Phi(\mathbf{X})$, we could use automatically derived transcriptions $\hat{\Phi}(\mathbf{X})$ for each enrollment utterance. In order to compare the effect of training from assignments derived from $\hat{\Phi}(\mathbf{X})$ versus $\Phi(\mathbf{X})$, we trained speaker models using transcriptions automatically generated by the same speech recognizer used during testing. For the remainder of this paper, we will refer to the speaker models trained from manual transcriptions as manually trained (MT) models and the speaker models trained from automatic transcriptions as automatically trained (AT) models.

5. Experimental Conditions

5.1. Corpus Description

We conducted our experiments using a corpus of speaker-labeled data collected using the MIT MERCURY airline travel information system [7] and the MIT ORION task delegation system [6]. The 44 most frequent registered users of MERCURY and ORION were selected to represent the set of "known" users. Each of these users spoke a minimum of 48 utterances in the calls representing the enrollment data for our experiments. As would be expected in real-world applications, the amount of enrollment data available for each known user varied greatly based on the frequency with which they used the systems. Of the 44 speakers, 15 had less than 100 utterances available for training, 19 had between 100 and 500 enrollment utterances, and 10 speakers had more than 500 utterances for enrollment. Within the 44 speakers, 21 were females and 23 were males.

For the test set, all calls made to the MERCURY system during a 10-month span were held out for our evaluation set. Only calls containing at least five utterances were included in the evaluation set. The evaluation set was further broken down into two sub-sets: a set of 304 calls containing 3705 utterance from members of the set of 44 known users, and a set of 238

calls containing 2946 utterances from speakers outside of the known speaker set. Each call had a variable number of utterances with an average of 12 utterances per call (stdio=6 utts) and an average utterance duration of 2.3 sec (stdev=1.4 sec).

Within the enrollment and evaluation data, the first two utterances of each call are typically restricted to be the caller's *user name* and *date password*, the remainder of each call is usually comprised of unconstrained user requests mixed with user responses to system initiated prompts. These utterances can be highly variable in their length and phonetic content, ranging from simple one word utterances (such as "yes" or "no") to lengthy requests for flight information (e.g., "I need a flight from Boston to Miami on United next Tuesday afternoon.").

5.2. ASR Acoustic Modeling

For the ASR-dependent speaker identification approaches described in this paper, the segment-based SUMMIT system was used for speech recognition [2]. The recognizer performed word recognition using a vocabulary and language model derived specifically for whichever system was being used (i.e., either MERCURY or ORION). For these experiments, the recognizer's feature vectors (i.e., the acoustic observations $\mathbf{X}$) were derived at acoustic landmarks hypothesized to be potential phonetic segment boundaries. These feature vectors were constructed by concatenating 14-dimension mean normalized MFCC vectors averaged over eight different segments surrounding each landmark. Principal components analysis was then used to reduce the dimensionality of these feature vectors to 50 dimensions. The acoustic model set scored these landmarks using 1388 different context-dependent models. It should be noted that the general principles described in this paper are not specific to segment-based recognition and are applicable to frame-based speech recognition systems as well.

6. Results

6.1. Evaluation Scenarios

For our experiments we have examined both the closed-set speaker identification and speaker verification problems. Because our data is collected via individual calls to our system, we can evaluate speaker identification at both the utterance-level and the call-level. In our case the utterance-level evaluation could be quite challenging because any single utterance could be quite short (such as the single word utterance "no") or ill-matched to the caller's models (as might be the case if the caller uttered a new city name not observed in his/her enrollment data).

In many applications, it is acceptable to delay the decision on the speaker's identity for as long as possible in order to collect additional evaluation data. For example, when booking a flight, the system could continue to collect data while the caller is browsing for flights, and delay the decision on the speaker's identity until the caller requests a transaction requiring security, such as billing a reservation to a credit card. To simulate this style of speaker identification, we can evaluate the system using all available utterances from each call in the evaluation data.

6.2. Comparison of normalization schemes

We performed several experiments. First, we compared the performances of the two normalization approaches on the task of closed-set speaker identification using the 3705 in-set utterances. The identification error rates are shown in Table 1. We see that using the full amount of enrollment data per speaker, both techniques perform equally well. On limited enrollment data per speaker, the phone-adaptive normalization approach performs better. This is presumably because it retains the ability to discriminate between speakers even when there are many instances of sparsely trained phonetic units.

Amount of Enrollment Data	*Speaker ID Error Rate (%)*	
	SA Norm	PA Norm
Max 50 utts	26.4	22.5
Max 100 utts	18.4	15.9
All available	9.6	9.6

Table 1: Closed-set speaker identification error rates on individual utterances for different amounts of enrollment data per speaker for speaker adaptive vs. phone adaptive normalization.

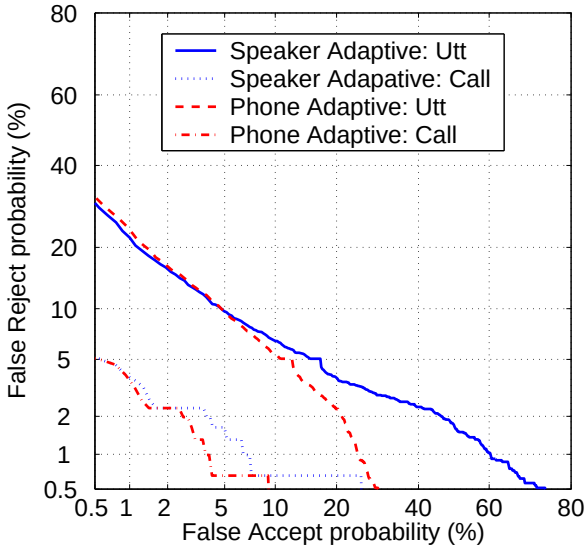

Figure 1: DET curves showing false rejection probability versus false accept probability for speaker adaptive vs. phone adaptive normalization.

For the speaker verification task, we used the 2946 out-of-set utterances to perform same-gender imposter trials (i.e., each utterance was only used as an imposter trial against reference speakers of the same gender). The detection error trade-off (DET) curve of the two approaches is shown in Figure 1. For all four curves, the models were trained on all available data. From the region of low false acceptances through the equal error rate region of the DET curve, the two normalization techniques have very similar performances. However, in the "permissive" operating point region with low false rejection rates, the phone-adaptive approach has significantly lower false acceptance rates than the speaker adaptive approach. This observation is important for a conversational dialogue system where convenience to frequent users is a factor. For example, if we want to maximize convenience by ensuring that only 1% of true speakers are falsely rejected, then the speaker adaptive method will result in a 60.3% false acceptance rate, while the phone adaptive method will result in a 24.6% false acceptance rate.

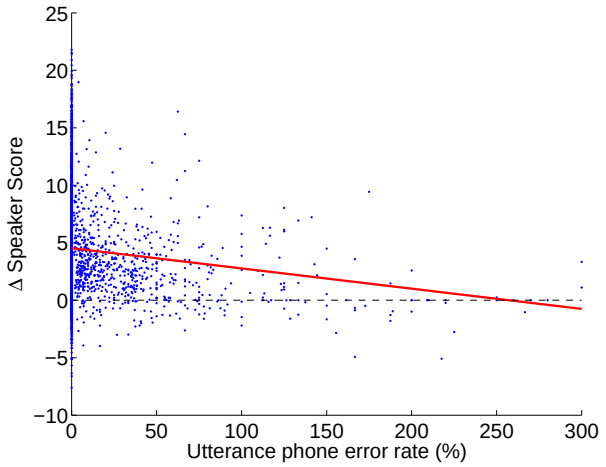

Figure 2: Plot of correct speaker discrimination versus utterance phone error rate using MT models. Each point represents a single utterance. The horizontal axis shows the phone recognition error rate on the utterance. The vertical axis indicates the difference between the scores of the reference speaker and the best competitor (negative values indicate identification error). A best-fit linear approximation of the data is superimposed.

6.3. Comparison of training approaches

In our previously reported experiments, we used models derived from the manual transcriptions for $\Phi(\mathbf{X})$ of the enrollment data. Under this condition, a high phonetic error rate within $\hat{\Phi}(\mathbf{X})$ of the test data would result in a mismatch between the testing and training conditions. Therefore, we might expect higher phonetic error rates to be positively correlated with higher speaker identification error rates.

To demonstrate this, we plotted speaker discriminability versus phonetic error rate on a per utterance basis in Figure 2. On this graph, the vertical axis indicates the difference between the scores of the reference speaker and the best competitor on a particular utterance, with negative values denoting an identification error. The horizontal axis represents the phonetic recognition error rate of each utterance. As expected, we observe a negative correlation between speaker discriminability and phonetic error rate, which confirms that higher phonetic error rates are correlated with poorer speaker discriminability (and hence higher speaker identification error).

It is reasonable to believe that the correlation between phonetic error rate and speaker error rate could be reduced by achieving a better match between training and testing conditions. In this case, using automatically derived transcriptions during training should provide a match to the testing conditions. However, when we repeated the above plot using the AT models, we observed a nearly identical distribution of points resulting in a nearly identical correlation between the phone error rate and speaker score difference. In other words, the speaker identification performance of the AT models was almost identical to that of the MT models, regardless of the phone error rate.

The overall results of performing closed-set identification using models trained with each approach, which are shown in Table 2, confirm our findings that there is no significant gain or penalty from using the automatically transcribed enrollment data. Although there is no clear benefit in speaker identification accuracy when using the automatic transcriptions, their use is still beneficial because the speaker models can be trained in an unsupervised fashion (i.e., without requiring manual transcriptions of the enrollment data).

Enrollment Transcriptions	*Speaker ID Error Rate (%)*	
	SA Norm	PA Norm
Manual	9.61	9.61
Automatic	9.77	9.36

Table 2: Comparison of closed-set speaker identification error rates on individual utterances for models trained from manually and automatically transcribed enrollment data

7. Conclusion

In this paper, we addressed the issues of speaker score normalization and of using automatically generated transcriptions for training speaker models when performing ASR-dependent speaker identification.

We found that using a phone-adaptive approach is beneficial for normalizing speaker scores compared to a speaker-adaptive approach. Although both methods have similar speaker identification performance, the phone-adaptive method generates scores that are more stable on speaker verification tasks, yielding fewer false acceptances of imposters at permissive operating points where low false rejection of known users is desirable.

In comparing the models trained from manually and automatically generated transcriptions, we found no significant differences in speaker discriminability between the two approaches. This discovery indicates that we can take an unsupervised approach to training speaker models without adversely affecting our speaker identification results.

8. References

[1] U. Chaudhari, *et al.*, "Multi-grained modeling with pattern specific maximum likelihood transformations for text-independent speaker recognition," *IEEE Trans. Sp. Aud. Proc.*, 11(2), January 2003, pp. 100–108.

[2] J. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, 17(2-3), April-July 2003, pp. 137–152.

[3] T. J. Hazen, *et al.*,"Integration of speaker recognition into conversational spoken dialogue systems," in *Proc. EUROSPEECH*, Geneva, Switzerland, September 2003, pp. 1961–1964.

[4] A. Park and T. J. Hazen, "ASR dependent techniques for speaker identification," in *Proc. ICSLP*, Denver, Colorado, September 2002, pp. 2521–2524.

[5] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communications*, 17(1-2), August 1995, pp. 91–108,

[6] S. Seneff, C. Chuu and D. S. Cyphers, "Orion: From online interaction to off-line delegation," in *Proc. ICSLP*, Beijing, China, October 2000, pp. 767–770.

[7] S. Seneff and J. Polifroni, "Formal and natural language generation in the Mercury conversational system," in *Satellite Dialogue Workshop of the ANLP-NAACL Meeting*, Seattle, WA, April 2000.

[8] F. Weber, *et al.*, "Speaker recognition on single- and multispeaker data," *Digital Signal Processing*, 10(1), January 2000, pp. 75–92.