LANGUAGE NORMALIZATION FOR BILINGUAL SPEAKER RECOGNITION SYSTEMS

Murat Akbacak, John H.L. Hansen

Center for Robust Speech Systems
University of Texas at Dallas
Richardson, TX, 75083, U.S.A
Web: http://crss.utdallas.edu
{murat.akbacak,john.hansen}@utdallas.edu

ABSTRACT

In this study, we focus on the problem of removing/normalizing the impact of spoken language variation in Bilingual Speaker Recognition (BSR) systems. In addition to environment, recording, and channel mismatches, spoken language mismatch is an additional factor resulting in performance degradation in speaker recognition systems. In today's world, the number of bilingual speakers is increasing with English becoming the universal second language. Data sparseness is becoming an important research issue to deploy speaker recognition systems with limited resources (e.g., short train/test durations). Therefore, leveraging existing resources from different languages becomes a practical concern in limited-resource BSR applications, and effective language normalization schemes are required to achieve more robust speaker recognition systems. Here, we propose two novel algorithms to address the spoken language mismatch problem: normalization at the utterance-level via Language Identification (LID), and normalization at the segment-level via multilingual Phone Recognition (PR). We evaluated our algorithms using a bilingual (Spanish-English) speaker set of 80 speakers. Experimental results show improvements over a baseline system which employs fusion of language-dependent speaker models with fixed weights.

Index Terms— bilingual speaker recognition, normalization

1. INTRODUCTION

Speaker recognition technology is based on statistical pattern recognition methods that require a training phase to generate statistical models to represent the speaker identities. Environment, recording, and channel conditions, speaker traits (e.g., dialect/accent, stress, speaking style), and spoken language can be considered as different dimensions in the acoustic space. Mismatch between training and testing in any of these acoustic dimensions results in performance degradation in speaker recognition applications. Previous studies on speaker recognition focus mainly on monolingual applications where the same language is spoken in both training and testing by all speakers. Therefore, within the context of the mismatch problem, the main focus has been on environment/recording/channel mismatch [1], rather than spoken language mismatch.

In a bilingual speaker recognition system, each speaker speaks any one of the two languages (native language or a second language), but not necessarily the same language in both training and testing. One solution towards potential language mismatches in these bilingual applications is to employ language-dependent speaker models after detecting the language spoken during testing. In real-life applications (both military and commercial), speech resources are limited, and for many languages data sparseness (e.g., short-duration train/test condition) is a major obstacle for advancements in speech technology for these countries/languages. In such cases, obtaining the best performance will require leveraging existing resources from resource rich languages with effective normalization schemes to minimize language variation impact, rather than employing language dependent systems.

In [2, 3], a multilingual phonetic string approach (similar to Phone Recognition and Language Modeling (PRLM) approach used in language identification [4]) is applied to speaker recognition. Although these methods provide a degree of language independence, they require extensive training and testing materials. Since the focus here will be on small training (20 sec.) and test set materials (5 sec.), these methods are not applicable towards removing the impact of language variation in speaker recognition systems with small enrollment data.

In [5], authors presented Linguistic Data Consortium's (LDC) data collection efforts to create corpora to support and evaluate systems that meet the challenge of speaker recognition within the context of language variation. In another study [6], a database of 49 bilingual speakers speaking Spanish and Catalan was considered for the problem of language variability. In that study, a method was proposed that uses a language-independent codebook (combination of language-dependent codebooks) in a vector quantization based Speaker ID (SID) system. In an earlier study [7], a series of listener tests were performed, and it was shown perceptually that language familiarity plays a significant role in speaker identification.

In this study, based on the acoustic similarities (e.g., acoustic phonetic space) of two languages, we develop normalization algorithms to leverage resources from these languages in an effective way for BSR applications with limited resources. The first algorithm merges language-dependent system outputs (i.e., likelihood scores) by using language-ID scores of each utterance as fusion weights. In the second algorithm, fusion is done at the segment level (time-synchronous normalization). Acoustic unit segments corresponding to phones that both languages have common in their acoustic phone space will be weighted more heavily during our GMM based accumulative likelihood calculation.

The remainder of this paper is organized as follows: Sec. 2 presents an overview of our GMM based Speaker ID (SID) system, as well as two different baseline systems, one using a traditional fusion approach where language-dependent system outputs are merged, and another using a multi-style training approach. Sec. 3 explains the proposed approaches which employ normalization schemes

This work was funded by grants from the U.S. Air Force Research Laboratory, Rome NY, under contract number F30602-03-0110, and by the University of Texas at Dallas under Project EMMITT.

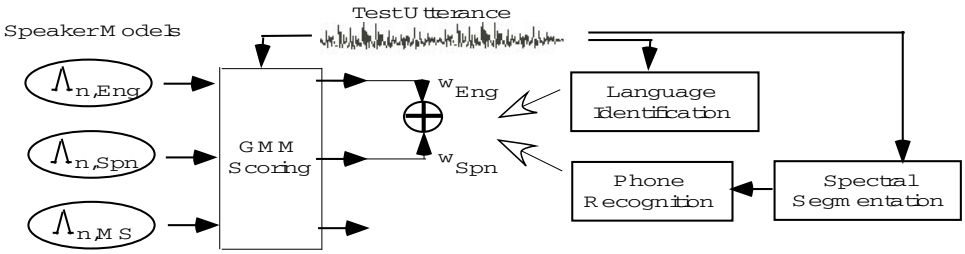

Fig. 1. An overview of segment-based and utterance-based normalization.

either at the utterance-level or at the segment-level. Speaker recognition results and evaluation of the proposed normalization algorithms for bilingual (English and Spanish) speakers are provided in Sec. 4. Discussion and future work are presented in Sec. 5, with a summary and conclusions in Sec. 6.

2. BASELINE SYSTEM

Our speaker recognition system is based on a GMM approach which has been widely studied and used in previous speaker recognition tasks [8]. We do flat training instead of using traditional Maximum a Posteriori (MAP) adapted Universal Background Model (UBM) approach since some model components with little enrollment data would remain unchanged in the derived speaker model, and this will result in weak discriminative capability over the background model, and would impair subsequent recognition performance.

We assume the following is given: a set of N in-set bilingual speakers, and the collected data $X_{n,L}$, corresponding to each enrolled speaker $S_n, 1 \leq n \leq N$, speaking language L. Each language-dependent speaker model $\{\Lambda_{n,L} : \{\hat{\omega}_{n,L}, \hat{\mu}_{n,L}, \hat{\Sigma}_{n,L}\} \in \mathbf{\Lambda}, 1 \leq n \leq N\}$ can be obtained from $X_{n,L} = \{x_{1,L}, \ldots, x_{t_{n,L}}\}$ where $t_{n,L}$ denotes the total number of samples that belong to speaker S_n while speaking language L. If O denotes the sequence of feature vectors extracted from the test utterance, in the recognition stage we classify O as Λ_L^*, to be the most likely in-set speaker model among all language-dependent speaker models. Therefore, Λ_L^* is written as,

$$\Lambda_L^* = \underset{1 \leq n \leq N}{argmax}\, p(O|\Lambda_{n,L}).$$

In our experiments, we use a bilingual language pair of English and Spanish. Therefore language-dependent speaker models for speaker S_n are denoted as $\Lambda_{n,Eng}$ and $\Lambda_{n,Spn}$, respectively. In addition to language-dependent speaker models, we also have a language-independent speaker model $\Lambda_{n,MS}$ trained via multi-style (MS) training using data from both English and Spanish. We employ two different baseline systems: ($\mathcal{B}1$) and ($\mathcal{B}2$). $\mathcal{B}1$ employs language-independent speaker models during recognition,

$$\Lambda_{MS}^* = \underset{1 \leq n \leq N}{argmax}\, p(O|\Lambda_{n,MS}).$$

whereas $\mathcal{B}2$ merges language-dependent systems' outputs via score fusion,

$$\Lambda^* = \underset{1 \leq n \leq N}{argmax}\, [p(O|\Lambda_{n,Eng})\, w_{Eng} \;+\; p(O|\Lambda_{n,Spn})\, w_{Spn}]$$

where fusion weights w_{Eng} and w_{Spn} are optimized using a development data. In this study, $\mathcal{B}2$ is considered as our baseline system since it performs better than $\mathcal{B}1$.

3. ALGORITHM DEVELOPMENT

Here, we present algorithm formulation for the proposed language normalization in bilingual speaker recognition systems. The first algorithm (LID-norm) is somewhat similar to $\mathcal{B}2$, the difference is that LID-norm employs dynamic fusion weights based on the Language Identification (LID) scores of each utterance. In the second algorithm (PR-norm), we employ normalization at the segment level after performing a spectral based segmentation scheme followed by Phone Recognition (PR). Details of these algorithms are explained in the following subsections.

3.1. LID-norm: Normalization at the utterance level

In this algorithm, we employ a GMM based Language ID module to determine an acoustic similarity metric for each utterance to quantify the degree of language variation. We employ GMM based Language ID rather than using phonetic string approaches (e.g., PRLM) since they require extensive data to train phonetic language models.

LID scores corresponding to each language are used as fusion weights. In other words, the probability of the event that the utterance is spoken in language L is used to weight the likelihood score coming from language dependent speaker recognition system $\Lambda_{n,L}$.

$$\Lambda^* = \underset{1 \leq n \leq N}{argmax}\, [p(O|\Lambda_{n,Eng})\, p(Eng|O) + p(O|\Lambda_{n,Spn})\, p(Spn|O)]$$

Generally speaking, acoustic models trained for each language can be interpreted as language-dependent speaker UBMs, and the distance between these UBMs can be used as a measure of acoustic similarity between languages.

3.2. PR-norm: Normalization at the segment level

In this algorithm, language-dependent speaker recognition system outputs are merged at the segment level. By using a spectral-distance based segmentation scheme, we first detect boundaries corresponding to phone boundaries, and then apply the fusion scheme at the segment level. By adjusting the spectral-distance threshold during segmentation, we attempt to force each segment to correspond to a phone.

We analyze the acoustic content of each segment by calculating the acoustic scores using language-dependent phone acoustic models. Phone mapping (e.g., IPA) knowledge between the languages is used in this approach to boost segment likelihoods corresponding to overlapping phones. In other words, segments appearing in the overlapping acoustic spaces of English and Spanish (e.g., based on IPA mapping) will be interpolated more uniformly [11]. Fig. 2 shows an example of language normalization at the segment level. It can be seen that segments corresponding to phones existing in both English and Spanish acoustic spaces are weighted more. Segments

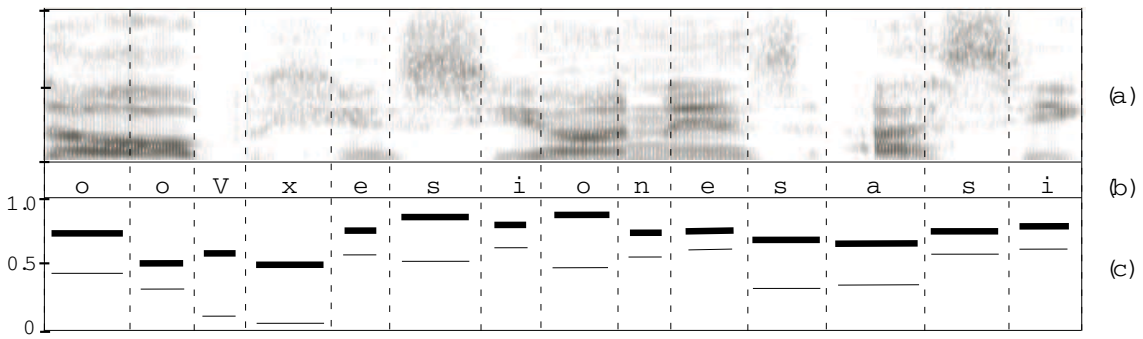

Fig. 2. Example to segment-based normalization for a Spanish utterance with its spectrogram (a), phonetic transcription (b), and normalization weights (c). Thick and thin lines correspond to $w_{i,Spn}$ and $w_{i,Eng}$ values respectively.

with same phonetic representations can be assigned different weights since the weights are calculated by using both acoustic distance and linguistic distance.

Based on the segment boundaries generated by the spectral segmentation module, the sequence of feature vectors extracted from the test utterance can be represented as $O = \{O_1, O_2, \ldots, O_M\}$, where M represents the number of segments. Next, we consider the formulation of score fusion where the fusion weights are updated at every segment,

$$S(n) = \sum_{i=1}^{M} p(O_i|\Lambda_{n,Eng})\, w_{i,Eng} \;+\; p(O_i|\Lambda_{n,Spn})\, w_{i,Spn}$$

In this likelihood calculation scheme, $w_{i,Spn}$ represents the weight for the likelihood score of speaker model $\Lambda_{n,Spn}$ for the i^{th} segment. Although we can use an LID based approach to determine the fusion weights, adapted-monophone models will give more discriminative scores.

In addition to generating fusion weights to merge language dependent speaker recognition outputs, this algorithm can also be used when there is only one language-dependent model (e.g., English speaker models tested on Spanish utterances) employed during recognition. In that case, this scheme can be considered as a weighted sum of likelihood scores for each segment.

4. EXPERIMENTAL RESULTS

In our experiments, we focus on the close-set Speaker ID (SID) task where English and Spanish are the target languages. We use the Miami Corpus [10] since it consists of utterances in Spanish (with two different dialects: Cuban and Peruvian), as well as in English. The speech utterances in the corpus are about 3 minutes in duration captured from an interview of native Spanish speakers. The effective size of the utterances is approximately 1.5 minutes as the interviewer's voice is eliminated from the utterances. There are different parts of the corpus such as read speech, spontaneous speech, digits, etc. Each of these utterances was recorded using a boom microphone and recorded with a Sony digital audio tape with a sampling rate of 48 kHz and 16-bit quantization. We used the downsampled version of the data (16kHz) in our experiments.

We created a set of 80 bilingual speakers from whom at least 35 sec. of spontaneous speech is available in both English and Spanish. Phone recognition based speech activity detection is used to extract the speech-only part of the conversation, and then fixed-length tokens are created. We use 20 sec. for training, and at least 15 sec. for testing (3x5 sec. test tokens). Rather than keeping the number of test tokens fixed for each speaker (total of 3x80=240 test tokens), we use all the test tokens from all speakers resulting in a total of 1000 test tokens for each language. 500 tokens are used as development set to optimize the fusion weights in $\mathcal{B}2$. To be able to leave speaking style variability aside, we only used the spontaneous speech part of the corpus in our evaluations. Speaker models contain 32 Gaussian components, with an intuitive understanding that one component would be used to cover each phone (after removing low energy consonants). 19-dimensional static MFCCs are used as feature vectors. To train language-dependent speaker UBMs for language identification, we use 1024 Gaussian components.

Exp.	Train Language	Test Language	SID perf.
1	Spn	Spn	84.89%
2	Eng	Eng	77.95%
3	Spn	Eng	83.49%
4	Eng	Spn	70.31%
5	Eng + Spn via $\mathcal{B}2$	Eng	81.22%
6	Eng + Spn via $\mathcal{B}2$	Spn	80.05%

Table 1. *Speaker ID performances with different train/test conditions (spoken language) with 20 sec. of total training data.*

Although experiments can be performed within a tiered structure where changing percantages of training material from English and Spanish are used (i.e., $X\%$ of the training data from English, $(100-X)\%$ from Spanish), due to limited resources, the range of available data sizes and corresponding performances are not seperated well from each other. Here, we consider three cases: $X = \{0, 50, 100\}$. Table 1 shows the baseline speaker recognition results. As can be seen, in the matched case, Spanish system (exp. 1) yields better performance than the English system (exp. 2). When $\mathcal{B}2$ is evaluated on English, replacing part of English training data with Spanish (exp. 2 and exp. 5) yields 3.27% absolute improvement. Same experiment on Spanish utterances (exp. 1 and exp. 6) degrades the performance by 4.84%. These can be explained by the fact that all speakers are native Spanish speakers, and there is a broad range of speakers in terms of proficiency level in English (e.g., rate of hesitations/pauses is more).

We employed English and Spanish phone recognizers trained for our previous study [11] using English Wall Street Journal and Spanish Latino-40, respectively. The number of phones for English and Spanish recognizers is 51 and 30, respectively. In [11], we also created a phone mapping (between English and Spanish phones) which was used to bootstrap English acoustic models ini-

tially to align Spanish speech data with its text transcription, and then iteratively train Spanish acoustic models. During monophone recognition using the test utterances from the Miami corpus, acoustic models are adapted using a single-class MLLR adaptation. Due to a lack of transcriptions, we cannot report phone recognition accuracy for the Miami corpus.

Language ID experiments yield language classification rates of 77.50% and 87.93% when 10 sec. and 20 sec. of training data, respectively, is used from each speaker. Table 2 compares baseline results with the results from LID-norm and PR-norm algorithms. Also, for exp. 3 and 4, LID-norm does not change the baseline results since same LID-based weights are applied to all speaker models for a specific language.

Exp.	Train	Test	$\mathcal{B}2$	LID-norm	PR-norm
3	Spn	Eng	83.49%	83.49%	84.82%
4	Eng	Spn	70.31%	70.31%	74.31%
5	Eng + Spn	Eng	81.22%	82.13%	83.21%
6	Eng + Spn	Spn	80.05%	81.32%	82.37%

Table 2. *Comparison of SID performances of baseline system $\mathcal{B}2$ and proposed algorithms with 20 sec. of training data (total).*

According to these results, we can say that PR-norm performs well in terms of compensating the impact of language variation between training and testing in case there is only one language to train the speaker recognition system (exp. 3 and 4), and another language is spoken during recognition. When score fusion is employed using language-dependent systems, LID-norm achieves an absolute improvement of 1.13%, whereas PR-norm achieves 2.15% absolute improvement in the overall (both English and Spanish utterances) over baseline results. Although there is still a performance gap between the system employing PR-norm (exp. 6 with 82.37% accuracy) and the matched system (exp. 1 with 84.89% accuracy), reported results are very promising, and suggest a viable procedure to follow for future advances in language normalization for bilingual speaker recognition systems as well as other speaker recognition tasks such as speaker verification and in-set/out-set speaker ID.

5. DISCUSSIONS

We note that while the results presented here are encouraging, several issues exist for effective crosslingual speaker recognition. One issue is bilingual speakers have varying degrees of fluency in their L1 versus L2 languages. Quantifying this issue is important for demonstrating the general trends. Second, limited corpora exist which effectively document history and language experience of bilingual speakers. This study has taken an initial step in considering effective speaker recognition in a crosslingual scenario. The present Miami Corpus provides a reasonable way to illustrate general improvement. however a more comprehensive corpus with expanded conversational content and sessions is needed to accurately assess the true benefits of proposed algorithms.

As mentioned, the experiments we performed use a small set (80 speakers) of bilingual speakers since larger bilingual corpora is not available at the moment. In the future, we are planning to collect a larger corpus of bilingual speakers under clean conditions without any channel or recording mismatch. In addition to closed-set speaker ID task, performance evaluation of the proposed algorithms in an open-set speaker recognition task will be performed. One interesting aspect to consider would be language normalization during generating UBMs to model out-of-set speakers using speakers from resource-rich languages. Applying PR-norm to channel/recording/environment mismatch problems (which were out of the scope of this paper) within speaker recognition systems is another aspect to consider in the future. We would like to employ PR-norm also for small enrollment/test data case where training data can be analyzed with phone acoustic models, and during recognition, segments can be normalized based on the phone histogram generated from training data. In this way, the impact of unseen phones during speaker recognition can be normalized. Evaluation of the proposed algorithms for different language pairs will be under consideration if available test platforms are provided.

6. CONCLUSIONS

In this study, we have focused on the research problem of removing/normalizing the impact of spoken language variation in Bilingual Speaker Recognition (BSR) systems. We proposed two novel algorithms towards the spoken language mismatch problem: LID-norm and PR-norm which are shown to be effective towards solving language variation problem in speaker ID systems. Experimental results using a bilingual (Spanish-English) speaker set of 80 speakers show promising improvements over a baseline system using language dependent speaker models in a fusion scheme with fixed weights. 1.13% and 2.15% absolute improvements are obtained with LID-norm and PR-norm, respectively. The proposed methods provide sufficient flexibility to suggest a viable procedure to follow for other normalization/mismatch problems in speaker recognition systems.

7. ACKNOWLEDGEMENTS

The authors thank Vinod Prakash of CRSS for helpful discussions during this study.

8. REFERENCES

[1] D. Reynolds, "Comparison of Background Normalization Methods for Text-Independent Speaker Verifi cation", Eurospeech, 1997.

[2] W.D. Andrews, M.A. Kohler, J.P. Campbell, "Phonetic Speaker Recognition", Eurospeech, 2001.

[3] Q. Jin, T. Schultz, A. Waibel, "Phonetic Speaker Identifi cation", ICSLP, Denver, CO, 2002.

[4] M.A. Zissman, E. Singer, "Automatic Language Identifi cation of Telephone Speech Messages Using Phoneme Recognition and N-Gram Modeling", ICASSP, Vol. 1, pp.305-308, 1994.

[5] J. P. Campbell, H. Nakasone, C. Cieri, D. Miller, K. Walker, A. F. Martin, M. A. Przybocki, "The MMSR Bilingual and Cross-channel Corpora for Speaker Recognition Research and Evaluation", Odyssey Speaker & Lang. Recog. Workshop, 2004.

[6] A. Satue-Villar, M. Faundez-Zanuy, "On the Relevance of Language in Speaker Recognition", Eurospeech, vol. 3, pp. 1231-1234, 1999.

[7] J. Goggin, C. Thompson, G. Strube, L. Simental, "The role of language familiarity in voice identifi cation", Memory and Cognition 19, pp. 448-458, 1991.

[8] D.A. Reynolds, R.C. Rose, "Robust Text-Independent Speaker Identifi - cation Using Gaussian Mixture Speaker Models", IEEE Trans. Speech & Audio Proc., Vol. 3, No. 1, pp. 72-83, 1995.

[9] D.A. Reynolds, T. Quatieri, R. Dunn, "Speaker verifi cation using adapted Gaussian mixture models", Digital Signal Processing, 10:19-41, 2000.

[10] M. A. Zissman, T. P. Gleason, D. M. Rekart, and B. L. Losiewicz, "Automatic Dialect Identifi cation of Extemporaneous, Conversational, Latin American Spanish Speech", ICASSP, 1996.

[11] M. Akbacak, J.H.L. Hansen, "Spoken Proper Name Retrieval in Audio Streams for Limited-Resource Languages via Lattice Based Search Using Hybrid Representations", ICASSP, 2006.