SPOKEN LANGUAGE MISMATCH IN SPEAKER VERIFICATION: AN INVESTIGATION WITH NIST-SRE AND CRSS BI-LING CORPORA

*Abhinav Misra, John H. L. Hansen**

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
The University of Texas at Dallas (UTD), Richardson, TX 75080-3021, USA
{abhinav.misra, john.hansen}@utdallas.edu

ABSTRACT

Compensation for mismatch between acoustic conditions in automatic speaker recognition has been widely addressed in recent years. However, performance degradation due to language mismatch has yet to be thoroughly addressed. In this study, we address langauge mismatch for speaker verification. We select bilingual speaker data from the NIST SRE 04-08 corpora and develop train/test-trials for language matched and mismatched conditions. We first show that language variability significantly degrades speaker recognition performance even with a state-of-the-art i-vector system. Next, we consider two ideas to improve performance: i) we introduce small amounts of multi-lingual speech data to the Probabilistic Linear Discriminant Analysis (PLDA) development set, and ii) explore phoneme level analysis to investigate the effect of language mismatch. It is shown that introducing small amounts of multi-lingual seed data within PLDA training has a significant improvement in speaker verification performance. Also, using data from the CRSS Bi-Ling corpus, we show how various phoneme classes affect speaker verification in language mismatch. This speech corpus consists of bilingual speakers who speak either Hindi or Mandarin, in addition to English. Using this corpus, we propose a novel phoneme histogram normalization technique to match the phonetic spaces of two different languages and show a +16.6% relative improvement in speaker verification performance in the presence of language mismatch.

Index Terms— speaker verification, language mismatch, i-vector system, phoneme analysis

1. INTRODUCTION

Speech utterances from a given speaker contain information related to the acoustic environment, transmission channel, speaker traits (accent, stress, speaking style, etc.) and spoken language. All this information can be thought of as different dimensions of the speaker acoustic space. If there is a mismatch in any of these aspects between train and test, it results in degraded speaker recognition performance. Most of the previous research on mismatch compensation for speaker recognition focused on acoustic conditions [1, 2, 3, 4], while variability in spoken language has been given less weightage.

Some of the previous works include training speaker models on both languages spoken by the user [5]. Using a language detector to detect the language of the test utterance and then choosing an appropriate speaker model trained on that language for scoring [6]. However, in both of these works, availability and knowledge of train and test utterance languages is required. In [7], authors study how language impacts both discrimination and calibration of a system, but they don't propose any significant solution. In [8], the authors estimate a language dependent sub-space in the Joint Factor Analysis (JFA) [3] framework and then suppress it as a nuisance attribute in order to compensate for language variability. The Oregon Graduate Institute (OGI) multi-lingual data-set was used to train the language sub-spaces.

The OGI corpus was originally developed for language recognition studies and does not contain many bi-lingual speakers, making it hard to use for systematically studying language variability in Speaker Recognition. In NIST SRE'04–08 corpora, there are bi-lingual speakers, but as these data-sets contain channel and handset variability also, the effect of language mismatch alone cannot be studied properly. This prompted us to collect our own corpus containing bi-lingual speakers and analyze the problem in more detail. We looked at the problem from phoneme level.

In [9] authors observed that Phonemes with the same amount of training data gave different errors. The performance of the system remained same when combining the scores due to only 10-15 phonemes , as against combining the scores from all the phonemes. Hence the system uses discrimination based on phonemes for speaker recognition. Also, in [10] it is observed that in broad groupings the nasals and vowels are found to provide the best speaker recognition

This project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

performance, followed by the fricatives, affricates and approximants, with the stops providing the worst performance of all. These works motivated us to look at the problem of language mismatch as being caused by the differences in the phonetic spaces between the train and test segments. The CRSS Bi-Ling corpus was collected in such a way so as to facilitate such a study. Based on experiments using this corpus, we propose a new approach of phoneme histogram normalization to alleviate language mismatch in speaker verification systems.

We further study the language mismatch issue utilizing the NIST SRE corpora in the i-vector [4] PLDA [11, 12] system framework. We embed language information in the Universal Bakground Model (UBM), Total Variability (TV) and PLDA spaces of an i-vector speaker recognition system and analyze the improvement in performance. The remaining paper is organized as follows. Section 2 discusses the experiments carried out in the i-vector-PLDA framework. Section 3 analyzes the language mismatch problem from the phoneme level, and section 4 discusses the results as well as the future work.

2. I-VECTOR-PLDA FRAMEWORK

The i-vector based speaker verification systems have become the state-of-the-art in the field of speaker verification during the past few years. In this method the i-vectors are first extracted from the speech utterances [4] and then subjected to various channel compensation techniques, such as Linear Discriminant Analysis (LDA), Within Class Covariance Normalization (WCCN) [13], length normalization [14], and finally classified using a PLDA model [11] or other classifiers [4].

Conventional session variability compensation techniques can be assumed to be general purpose, and thus, processing the i-vectors using techniques such as LDA and WCCN should, in theory, also help in compensating language variability. However, in this study, we demonstrate that even when state-of-the-art compensation techniques are utilized, mismatch due to language persists and degrade the system performance. This, we believe, is an important finding since NIST has decided to not include language variability in the two most recent evaluations, namely SRE 2010 and 2012.

Now, one of the intuitive ways to improve the performance of the system would be to embed multi-lingual information in the i-vector extraction stage (UBM, Total variability space). While, the other way would be to add the multilingual information in the i-vector modelling/scoring stage (PLDA). In this paper we study the effect of both these methods and discuss the results.

2.1. I-Vector System Description

In this section, we describe the i-vector-PLDA based speaker verification system used in this study. The system was made following the protocol of the most recent SRE, that is, SRE-2012. Further details about the system can be found in [15].

Table 1. *Development data list details*

List ID	# files	Description	Amount of multi-lingual data
EN−1	5665	English telephone data	0%
ML−1	9680	Multilingual tel. and mic. data	50.17%
EN−2	28102	English tel, mic and noisy data	0%[15]
ML−2	28102	Multilingual tel, mic and noisy data	17.28%

First we do SAD that generally follows the method in [16], as implemented in the open-source Voicebox toolkit [17].In case of interview recordings, SAD is first performed on both interviewee (A) and interviewer (B) channel, and then, speech segments detected in channel B are removed from channel A. Since channel B is usually corrupted by a noise floor to mask the interviewee speech, spectral subtraction [18] is always performed before SAD on channel B.

Then, we extract 36 dimensional MFCC features followed by Quantile based Cepstral Normalization (QCN) [19] for robustness. A gender dependent 1024-mixture UBM with diagonal covariance matrices is trained on telephone and microphone utterances selected from the Switchboard-II Phase 2 and 3, Switchboard Cellular Part 1 and 2, and the SRE'04-06 enrollment data. We use two different lists for UBM training: EN−1 and ML−1 as specified in Table 1. The initial four iterations per mixture are gradually increased to 15 for higher order mixtures.

For training the i-vector extractor, we use a larger dataset than the one used for training the UBM. It also contains additional noisy speech data. Noisy files contain Heating, Ventilation, and Air Conditioning (HVAC) and crowd noise samples, are as prepared for our SRE-12 evaluation discussed in [15]. This list is identified as EN−2 in Table 1.For analyzing the effect of multi-lingual data in TV matrix training, we replace about 17% of the utterances (4857 files) by speech segments of languages other than English. This list is identified as ML−2 in Table 1. The i-vector extractor is trained using 5 EM iterations.

The 600 dimensional i-vectors are first mean normalized and then length normalized using radial Gaussianization [14]. LDA projection is performed to reduce the i-vector dimension to 400 before the PLDA scoring. A Gaussian PLDA model with a full-covariance residual is used for session variability compensation and scoring [14]. We used two different lists for PLDA model training: EN−2 and ML−2, as specified in Table 1, to observe the effect of multi-lingual speech in training.

2.2. Experiments

Data from NIST SRE'04–08 is combined to extract speakers speaking more than one language. While combining the lists of different SRE corpora, it is ensured that all the files contain male multilingual speakers with English as one of their spoken languages. Only 5 minute telephone channel recordings are considered. We did not use any microphone and interview data in order to ensure that the dominant mismatch present in the test-trials is language mismatch. From these utterances we prepare a set of 340 enrollment speakers speaking in English. Two separate test lists are then prepared for these speakers: i) a list containing only English as the spoken language (matched condition), and ii) the other consisting of all the languages other than English (mismatched condition). The number of trials in both these conditions are 267240.

To obtain the multi-lingual development lists, all the speech files from SRE'04–08, other than the ones used in trials, are collected and added to the UBM, TV and PLDA lists (`ML-1` in Table 1). A total of 4,857 such non-English files are obtained. While adding these files, the development lists (`EN-1` and `EN-2`) are pruned in a way that the total number of files remain the same. That is, the size of the TV, UBM and PLDA development lists is kept the same while embedding multi-lingual data into them (Table 1) .This is done to ensure a fair comparison between the trials with English only development lists and the trials with multi-lingual development lists. The results obtained in matched and mismatched conditions using different UBM, TV space and PLDA training datasets (defined in Table 1), are summarized in Table 2.

Table 2. *Results in the language matched & mismatched conditions.*

Development List Specification			Matched	Mismatched
UBM	TV	PLDA	% EER	%EER
EN-1	EN-2	EN-2	**1.745**	**4.395**
EN-1	EN-2	ML-2	0.868	**1.662**
ML-1	EN-2	EN-2	1.495	3.602
ML-1	EN-2	ML-2	1.188	2.291
EN-1	ML-2	EN-2	2.178	6.411
EN-1	ML-2	ML-2	**0.485**	2.288
ML-1	ML-2	EN-2	1.869	5.110
ML-1	ML-2	ML-2	0.526	2.781

2.3. Results and Discussion

From the results in Tables 2, we observe that when only English data is used for UBM, TV space and PLDA training, the EER in the matched and mismatched conditions is 1.745% and 4.395%, respectively. This shows the severe degradation caused by language mismatch alone, dropping the performance metric by a factor of 2.5. Next, we observe the effect of including non-English data in UBM, TV space and PLDA training.

From the results obtained, we make some interesting observations. First, by only adding the non-English data in PLDA training, the EER in the mismatched condition is relatively improved by a significant 62.18%, reaching 1.662% from 4.395%. However, adding non-English data in UBM and TV space training in any combination, is not able to provide such a significant improvement. In the matched condition, a very good performance is obtained in all three performance metric when all the models (UBM, TV and PLDA) are trained on mult-lingual data. This is surprising, since the performance obtained here is superior to a purely English trained system.

To further comprehend the results, error analysis is performed on the mismatched trials with English only PLDA and multi-lingual PLDA. As shown in Fig 1, while using English PLDA list, there were 1378 correct target identifications which increased to 1418 when multi-lingual PLDA was used. Multi-lingual PLDA did not introduce any extra errors as the 40 files that were now correctly identified came all from the 63 erroneous files before. The same phenomenon is observed in analyzing the misses, where around 7260 files were correctly identified as imposters, when switched to multi-lingual PLDA.

Since, the methods used to obtain the performance improvements are very simple in nature, these results can be beneficial for speaker recognition researchers in general, and perhaps be of relevance if language variability is reintroduced in the future NIST evaluations.

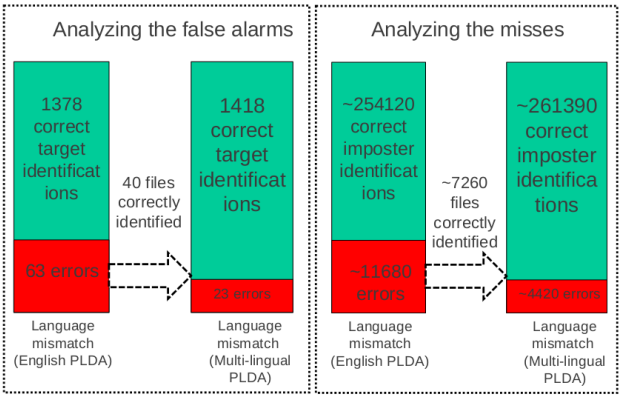

Fig. 1. Error analysis: NIST-SRE Data.

3. PHONEME LEVEL ANALYSIS

There are approximately 6000 different languages spoken all over the world, and we do not have sufficient data for most of these languages. Thus, it becomes crucial to develop techniques that can sustain automatic speaker recognition system

performance even when a low resource language is spoken. The solution proposed in the above section relies on the availability of a multi-lingual data to improve performance. We need a solution that can do language mismatch removal without being dependent on a priori knowledge of the languages or the availability of their data. In this section we try to systematically analyse this aspect of language variability problem in speaker recognition. For this purpose, it is important to ensure that the only source of mismatch is the language being spoken. Other sources of variabilities such as, environment, channel and speaker traits should be minimal. Since such a speech corpora with a sufficient number of unique speakers is not presently available, we choose to undertake the corpus collection ourselves.

3.1. Corpus Description

For Corpus collection, Hindi and Mandarin speaking participants were recruited. Speech is recorded both in English and the native language of the subject. A close-talk microphone, a far-field microphone (table top) and a cell-phone, are used simultaneously to record the speech data. The cell-phone is connected to a server through a telephone line. The subject would make a call to an analog telephone line connected to the motherboard of our server through a PCI telephony interface card. An interactive Voice Response System (IVRS) is set up at the server, which provides instructions to the subject.

In the beginning of the recording session, subjects are asked to read a set of 10 TIMIT sentences. Next, in order to record spontaneous speech from the participant, unfolded copies of the newspaper 'USA Today', and a questionnaire are kept on the table in front of the subject. He/she is allowed to talk about any section in the newspaper (weather, sports, lifestyle, etc.) or answer any of the questions in the questionnaire. Each subject speaks for ten minutes in English and ten minutes in his/her native language. The entire recording duration is about 22-23 minutes, with initial 2-3 minutes of read speech and the remaining 20 minutes of spontaneous speech in two languages. Each participant is called only once so that no session variability is present between recordings. Table 3 summarizes the details of the corpus.

Table 3. *CRSS Bi-Ling speaker corpus statistics. The values in the table indicate number of speakers*

Language	Gender	Microphone		Telephone
		Close-talk	Far-field	
Mandarin	Male	11	11	13
Mandarin	Female	9	9	11
Hindi	Male	21	21	26
Hindi	Female	15	15	20
Total		56	56	70

3.2. Phoneme Histogram Normalization

Different languages have different phoneme structures as illustrated in Fig 2 (a) and (b), where phoneme histograms are plotted for English and Hindi utterances, respectively. Each utterance is 2.5 minutes and spoken by the same speaker from CRSS Bi-Ling database. In the histograms, `pau` indicates silence, while `spk` and `int` indicate speaker and intermittent noise segments, respectively. We use the Brno University of Technology (BUT) phoneme recognizer based on long temporal context to obtain the phonetic transcriptions [**?**]. SAMPA phonetic alphabets are used[1].

The Hungarian phoneme recognizer is used to transcribe Hindi and English phonemes for the following reasons: i) the Hungarian recognizer contains the highest number of unique phonemes (61), and ii) it will not be biased towards either English or Hindi speech. As we can see from the histograms, phonemes `:2` and `l:` are present in the English utterance, but not in Hindi. Similarly, phonemes `b:`, `i:` and `t1:` are present in the Hindi speech segment only. Also, if we consider the common phonemes between the languages, their frequency of occurrence is very different.

We consider reducing this difference between phoneme distribution of train and test segments. We normalize the test utterance phoneme histogram to match it with the train utterance phoneme histogram. This is accomplished using a simple algorithm that dynamically weighs each phoneme at the scoring stage.

1. Let p_i be the total number of occurrences of i^{th} phoneme in the test utterance. Then, the initial weight w of this phoneme is given by:

$$w = \frac{p_i}{\sum_{i=1}^{N} p_i} \tag{1}$$

   where N is the total number of phonemes present in the test utterance.

2. In a similar manner, the weight is calculated for this phoneme in the train utterance.

3. Next, train and test utterance weights are compared. If the test utterance weight is lower than the train utterance weight an additional weight δw is added to w. Alternatively, if the test utterance weight is more than the train utterance, the same additional weight is subtracted from the current weight.

4. δw is calculated as:

$$\delta w = \frac{|p_i(test) - p_i(train)|}{\sum_{i=1}^{N} p_i} \tag{2}$$

Weights calculated from the above algorithm are used to weigh the scoring of each phoneme in a GMM-UBM set-up .

[1]The following web-page discusses the mapping between SAMPA and IPA: http://www.phon.ucl.ac.uk/home/sampa/

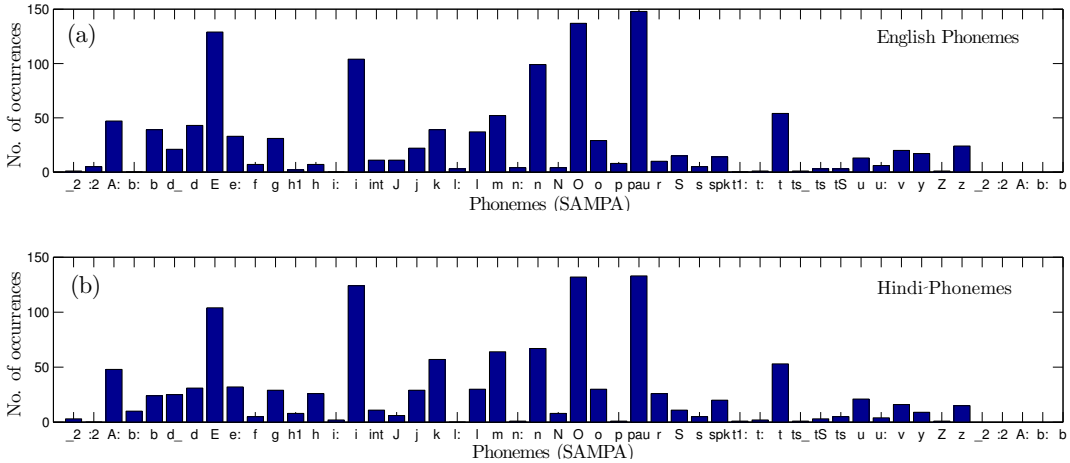

Fig. 2. Histogram of phonemes detected from an English and Hindi spoken utterance.

3.3. Speaker Verification Experiments on Bi-Ling Corpus

Using the collected data we employ a GMM-UBM system to establish a production space justification, which would then be extended to our i-vector system. Telephone channel data is used and labeled as English, Hindi/Mandarin. Since only telephone channel data is taken, there is no mismatch other than the language mismatch. All data from Mandarin speakers is used to train a 512 mixture UBM. Data from 36 native speakers of Hindi (18 male & 18 female), is used to conduct language mismatch studies. 36 dimensional MFCC feature vectors (12 static, with delta and acceleration coefficients) are used as the front-end of the system.

Each speaker's utterance is split into four 2.5 minute files. Out of these four files, two are English and two are Hindi. A train list of speakers speaking 2.5 minutes of English, is created. Two test lists are prepared, with one containing only English while the other only Hindi utterances, thereby, creating two sets of test-trials having language matched and mismatched conditions. Considering all combinations between train and test utterances, a total of 2386 trials are obtained in both conditions. The experimental results are summarized in Table 4.

From the table, we observe that even when everything else remains consistent (channel, environment, session), and only the spoken language is different, system EER performance is degraded by almost 135% relative to the matched condition. After applying the proposed phoneme histogram normalization method, the EER of the mismatch case improves by a relative +16%. This is encouraging, since it shows that we can leverage the phonetic content of utterances to suppress language mismatch.

Table 4. *GMM-UBM System performance*

Train	Test	EER(%)
English	English	4.718
English	Hindi	11.108
English (phoneme histogram normalization)	Hindi	9.260

4. CONCLUSION

In this study, we considered a systematic study on language variability for the problem of speaker verification. We showed improvemment using state-of-the-art i-vector-PLDA based system on language mismatched test-trials. We also investigated the problem at the phoneme level using a GMM-UBM system. We observed that by making the phonetic profile of two different languages similar to each other, we can achieve an improvement in system performance in the case of language mismatch. For future work, we intend to explore further data normalization techniques to reduce the impact of language mismatch.

5. REFERENCES

[1] D.A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. InterSpeech*, 1997.

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus Eigenchannels in speaker recognition," *IEEE Trans. Audio Speech Lang. Process.*, May 2007.

[3] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Trans. Audio Speech Lang. Process.*, July 2008.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio Speech Lang. Process.*, May 2010.

[5] B. Ma and H. Meng, "English-chinese bilingual text-independent speaker verification," in *Proc. IEEE ICASSP*, May 2004.

[6] M. Akbacak and J.H.L. Hansen, "Language normalization for bilingual speaker recognition systems," in *Proc. IEEE ICASSP*, April 2007.

[7] N. Brummer, L. Burget, J.H. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D.A van Leeuwen, P. Matejka, P. Schwarz, and A Strasheim, "Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006," *Audio, Speech, and Language Processing, IEEE Transactions on*, Sept 2007.

[8] L. Lu, Y. Dong, X. Zhao, J. Liu, and H. Wang, "The effect of language factors for robust speaker recognition," in *Proc. IEEE ICASSP*, April 2009.

[9] R. Auckenthaler, E.S. Parris, and M.J. Carey, "Improving a gmm speaker verification system by phonetic weighting," in *Proc. IEEE ICASSP*, 1999.

[10] J. P. Eatock and J.S. Mason, "A quantitative assessment of the relative speaker discriminating properties of phonemes," in *Proc. IEEE ICASSP*, 1994.

[11] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. Odyssey*, Brno, Czech Republic, 2010.

[12] Jesus Villalba and Niko Brummer, "Towards fully Bayesian speaker recognition: Integrating out the between-speaker covariance," in *Proc. InterSpeech*, Florence, Italy, Oct. 2011.

[13] A.O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. InterSpeech*, Pittsburgh, Pennsylvania, 2006.

[14] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-Vector length normalization in speaker recognition systems," in *Proc. InterSpeech*, Florence, Italy, Oct. 2011.

[15] Taufiq Hasan, Seyed Omid Sadjadi, Gang Liu, Navid Shokouhi, Hynek Boril, and John H.L. Hansen, "CRSS Systems for 2012 NIST Speaker Recognition Evaluation," in *Proc. IEEE ICASSP*, Vancouver, Canada, May. 2013.

[16] J. Sohn, N.S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, 1999.

[17] Mike Brooks, "VOICEBOX: Speech Processing Toolbox for MATLAB," .

[18] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Process.*, Apr 1979.

[19] H. Bořil and J. H. L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments," *IEEE Trans. Audio Speech Lang. Process.*, Sep. 2010.