Uludag University Journal of The Faculty of Engineering, Vol. 22, No.1, 2017 RESEARCH

DOI: 10.17482/uumfd.309477

EFFECT OF LANGUAGE MISMATCH ON TURKISH SPEAKER VERIFICATION

Cemal HANİLÇİ [*]

Received: 05.12.2016; revised: 06.03.2017; accepted: 21.04.2017

Abstract: In this paper, effect of language mismatch between background data and evaluation data is analyzed for text-independent speaker recognition in particular for Turkish spoken language. Gaussian mixture model with universal background model (GMM-UBM) classifier is utilized using Mel-frequency cepstral coefficients (MFCCs) as speaker-specific features. Experiments conducted on a Turkish speech database consisting of 47 male and 26 female speakers reveals that Turkish speaker recognition performance dramatically degrades in case of language mismatch between UBM and the evaluation data. For example 1.73% and 12.34% equal error rates (EERs) are obtained for male speakers when UBM is trained using Turkish and English data, respectively.

Keywords: Turkish speaker recognition, language mismatch

Türkçe Konuşmacı Doğrulamada Dil Uyumsuzluğunun Etkisi

Öz: Bu çalışmada, arkaplan verisi ile gerçekleştirme verisi arasında konuşulan dil anlamında bir uyumsuzluk olması durumunda Türkçe konuşmalar için konuşmacı tanıma performansı incelenmiştir. Gauss karışım modeli - genel arkaplan modeli sınıflandırıcısı ile mel-frekansı kepstral katsayıları konuşmacılara özgü öznitelikler olarak seçilmiştir. 47 erkek ve 26 bayan konuşmacıdan oluşan Türkçe veritabanı ile yapılan deneylerde görülmüştür ki arkaplan modelini eğitmek için kullanılan seslerin dili ile konuşmacı doğrulama deneylerinde kullanılan dil farklı olduğunda konuşmacı doğrulama performansı dramatik bir şekilde düşmektedir. Örneğin, erkek konuşmacılar için Türkçe ses verileri ile arkaplan modeli eğitildiğinde %1.73 eşit hata oranı elde edilirken, İngilizce sesler ile eğitildiğinde %12.34 eşit hata oranı elde edilmiştir.

Anahtar Kelimeler: Türkçe konuşmacı doğrulama, dil uyumsuzluğu.

1. SECTION 1

Speaker verification is the task of automatically authenticating the speaker's claimed identity using his/her voice sample (Hansen and Hasan, 2015). In recent years, automatic speaker verification systems have found their way to commercial use in real-time applications such as online banking, smart cars etc. However, speaker verification systems have still important challenges to address and solve. Speech signals from a speaker carry information related to transmission channel, speaker's emotion, age, accent and spoken language. Any mismatch of these dimensions between training and test stages of speaker verification systems results considerable degradation on the performance. In recent studies, research mostly focused on compensating the mismatch induced by transmission channels and great improvement have been obtained with the sophisticated *i-vector* approach. However, variability or mismatch in spoken language has been less studied.

[*] Department of Electrical-Electronic Engineering, Bursa Technical University, 16330, Bursa, Turkey
Corresponding Author: Cemal HANİLÇİ (cemal.hanilci@btu.edu.tr)

Spoken language mismatch on speaker verification can be considered as a less important problem for text-dependent speaker verification. This is because in text-dependent verification, a fixed phrase is chosen by the user and it can be in any language (Benesty et. al., 1997). Similarly, in text-prompted applications, a phrase is prompted to user and prompted phrase can be in any language. However, in text-independent speaker verification, the variability in spoken language is an important problem and requires more attention.

In (Ma and Meng, 2004), bilingual text-independent speaker recognition task was studied where each speaker is trained using English data and tested with Chinese data. In that study, it was reported that language mismatch between training and test data yields significant degradation. To alleviate this degradation, authors proposed to model each speaker using both languages (Ma and Meng, 2004). Another solution for bilingual speaker recognition is training two separate speaker models for each target speaker one with Spanish data and the other using English data (Akbacak and Hansen, 2007). During the recognition phase, first a language detector is used to detect the language of test utterance for choosing the correct speaker model (Akbacak and Hansen, 2007). However, both of these two proposed solutions require knowledge about the languages of training and test utterances. In (Ma et.al., 2007), the effects of device, language and environmental mismatches between training and test data of speaker recognition system is studied and it was found that language mismatch (training each speaker on Chinese data and testing with English speech) brings 288% performance degradation (EER increases to 6.42% from 1.65% whereas environmental mismatch yields 162% degradation. A feature-level solution--combining standard mel-frequency cepstral coefficients (MFCCs) with prosodic features-- was proposed in (Luengo et.al., 2008) for multilingual speaker recognition in which Spanish and Basque languages are used in the experiments. In a more recent study (Misra and Hansen, 2014) the performance of the state-of-the-art i-vector speaker recognition system is analyzed and it was found that language-mismatch significantly reduces the i-vector system performance.

One of the fundamental problem with analyzing the effect of language mismatch on speaker recognition is the lack of speaker recognition databases consisting of utterances in different languages from a particular target speaker. Plus, most of the speaker recognition studies carry out their investigations on English language. This is because of the existence of large English databases from NIST[*] and LDC[**]. The annual NIST Speaker Recognition Evaluation provides large databases to the researchers. Therefore, the researchers mostly reports their results and analysis on NIST corpora.

Although there are few speaker recognition studies on Turkish language (Büyük and Aslan, 2012a, Büyük and Aslan, 2012b), motivated by the fact that there is a lack of speech databases available for Turkish and lack of studies report their findings for Turkish language, in this paper, we analyze the effect of language and environmental mismatch on Turkish speaker verification which is the preliminary results of an ongoing project. To this end, we propose an experimental setup using speakers from Turkish language. Gaussian mixture model with universal background model (GMM-UBM) method is used as the classifier for speaker verification. The UBM for the speaker verification task is trained using English and Turkish data for the investigation of language mismatch on Turkish speaker verification. Although there are more sophisticated algorithms used for speaker verification (e.g. GMM supervector, joint factor analysis and i-vector), we utilize the simple but efficient GMM-UBM method in the experiments because its performance on Turkish speech database is unknown and it requires less data to train hyperparameters in comparison to other methods. Another reason of selecting the GMM-UBM method is that the most of the state-of-the-art techniques require UBM model trained in advance. However, the effect of the training data for UBM is unknown for Turkish

[*] http://www.itl.nist.gov/iad/mig/tests/spk/

[**] https://catalog.ldc.upenn.edu/

speaker verification. Therefore, in order to use other techniques UBM is required and it has a considerable impact on the performance. Thus in this paper, we study the Turkish speaker verification system using GMM-UBM method. Our study differs from previous studies on Turkish language in some manners: First, in (Büyük and Aslan, 2012a, Büyük and Aslan, 2012b), text-dependent speaker recognition using Turkish speech data is considered whereas we study text-independent speaker verification. Second, to the best of our knowledge this is the first study investigating the effect of database/language and recording condition variability on Turkish speaker recognition.

The remainder of the paper is as follows: in Section 2 we briefly explain the speaker verification task using GMM-UBM method. The details of our experimental setup are given in Section 3. In Section 4, the results of our speaker verification experiments are provided and finally in Section 5, we discuss future work and conclude our results.

2. SPEAKER RECOGNITION SYSTEM

Given a speech signal, S, speaker verification, determining whether S belongs to claimed speaker *P*, can be defined as a hypothesis test between two hypotheses (Reynolds et. al. 2000):

- H_0 : S belongs to claimed speaker
- H_1: S does not belong to claimed speaker.

Therefore, likelihood ratio (LR) test can be used to decide between H_0 and H_1. Using the feature vectors $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_T\}$ extracted from S, logarithmic likelihood ratio score is given by,

$$\Lambda(\mathbf{X})=\log p(\mathbf{X}|\lambda_{H_0})-\log p(\mathbf{X}|\lambda_{H_1}), \tag{1}$$

where λ_{H_0} and λ_{H_1} are the acoustic models characterizing the hypotheses H_0 and H_1, respectively.

Gaussian mixture model (GMM) (Reynolds and Rose 1995) is a popular modeling technique for representing acoustic models (H_0 and H_1) in speech applications. In GMM, each class is represented as a weighted sum of M multivariate Gaussians, $p(\mathbf{x}|\lambda)=\sum_{i=1}^{M} w_i p_i(\mathbf{x})$. Here, w_i is the weight of i th mixture component and $p_i(\mathbf{x})$ is a $D-$variate Gaussian density function of the form,

$$p_i(\mathbf{x})=\frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_i|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\}, \tag{2}$$

with mean vector w_i and covariance matrix Σ_i. Thus, each mixture component consists of a mixture weight ($\boldsymbol{\mu}_i$), a mean vector ($\boldsymbol{\mu}_i$) and a covariance matrix Σ_i. Therefore an acoustic model represented by a GMM is denoted by $\lambda=\{w_i,\mu_i,\Sigma_i\}_{i=1}^{M}$.

In speaker verification, each hypothesis is represented by a GMM where the parameters (weights, mean vectors and covariance matrices) of alternative hypothesis (H_1) is trained using the expectation maximization (EM) algorithm (Dempster, et. al. 1977) via maximum likelihood (ML) criterion (Reynolds and Rose, 1995) using a large amount of speech data from many speakers. This GMM model characterizing the H_1, is popularly known as universal background model (UBM) and the model is denoted by λ_{UBM}(Reynolds et. al. 2000). Target (claimed) speaker model parameters representing the hypothesis H_0 in turn (λ_{TGT}), are obtained via maximum a-*posteriori* adaptation (MAP) of the UBM with the feature vectors extracted from target speaker's training utterance. The general framework for GMM-UBM based speaker verification system is depicted in Figure 1.

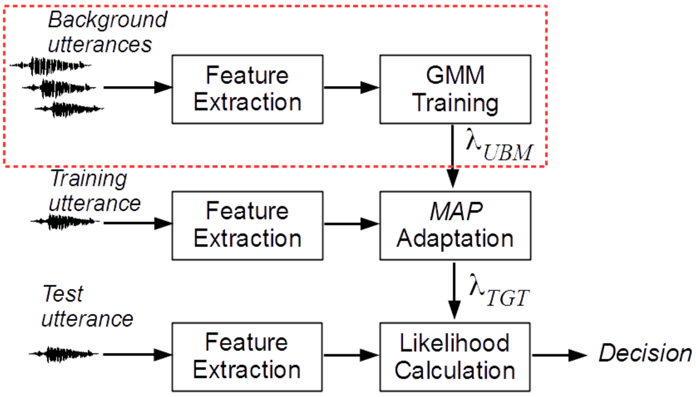

Figure 1:
General structure of a GMM-UBM based speaker verification system.

3. EXPERIMENTAL SETUP

Speaker verification experiments are conducted on TURTEL (TURkish TELephony) speech database consisting of 57 male and 36 female speakers. Each speaker reads the same phonetically balanced 15 Turkish sentences each sampled at 16 kHz and approximately with the duration of 3 seconds. After eliminating the non-speech portions of the speech signal with voice activity detection (VAD) the duration of each utterance reduces approximately to 1.5 seconds. Histogram plots of the duration of speech signals before and after VAD process is shown in Figure 2. Each speaker is trained using his/her randomly selected 5 utterances and the remaining ten utterances are used for verification.

Table 1. Statistics of the TURTEL database

Gender	# Speakers	# Utterances Per Speaker	Total Number of Utterances
Male	57	15	855
Female	36	15	540

Mel-frequency cepstral coefficients (MFCCs) features are extracted from Hamming windowed speech frames of 20 ms with 10 ms overlap. Discrete Fourier transform (DFT) of windowed speech frames are computed to obtain power spectra. Power spectra is then processed through Mel-filterbank consisting of 27 triangular filters in mel-scale. Logarithmic filterbank outputs are converted into MFCCs by taking discrete Cosine transform (DCT).

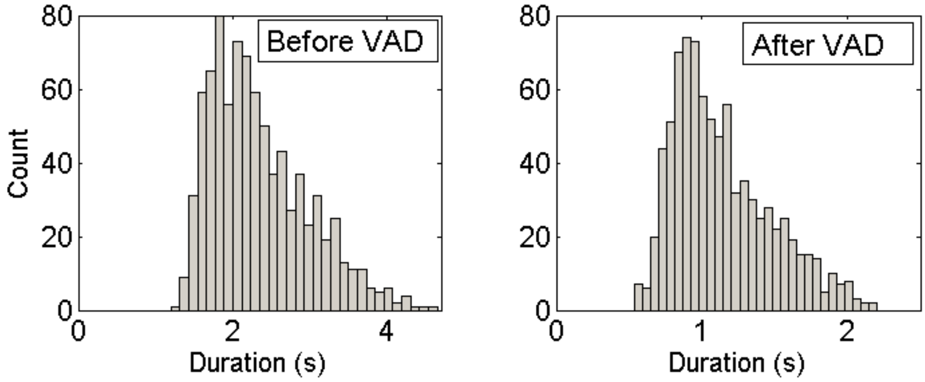

Figure 2:
Duration of the speech signals before and after VAD

Speaker verification performance is evaluated using Gaussian mixture model with universal background model (GMM-UBM) classifier. In order to investigate the effect of language and recording condition mismatch between UBM and evaluation data, two different UBMs trained using different speech data are used in the experiments:

- **TIMIT-UBM** : English microphone speech database TIMIT is used to train UBM which introduces both language and recording conditions mismatch between the TURTEL database and the speech data used to train UBM.
- **Oracle-UBM** : UBM is trained using the speech signals of randomly selected 10 male and 10 female speakers from the TURTEL database. Since the language and the recording conditions of this setup exactly match with the data used in speaker recognition experiments, comparison of the TIMIT-UBM results with Oracle-UBM will help to understand performance differences under language and recording condition mismatch.

Since 10 male and 10 female speakers are excluded from TURTEL database to train Oracle-UBM, in speaker recognition experiments the remaining 47 male and 26 female speakers are used. In both cases, TIMIT-UBM and Oracle-UBM, gender-independent UBMs with different model orders (number of Gaussian components) are trained using 20 EM iterations. Target speaker models are created using five training utterances of each speaker with maximum a-posteriori (MAP) adaptation of UBM model with a relevance factor of 8.

With the aforementioned UBM training cases, we aim to compare both the effect of the language and recording condition mismatch on the performance. Since TIMIT database consists of clean microphone speech collected from American speakers uttering English sentences, TIMIT-UBM introduces both language and recording condition mismatch between UBM and evaluation data. Oracle-UBM in turn, exactly matches with the evaluation data in terms of both recording condition and the language.

We use equal error rate (EER) as the performance criterion of speaker verification experiments. EER is the operational point where the false alarm (P_{fa}) and the false rejection rates (P_{miss}) are equal. The reported EERs in Section 4 are computed using the Bosaris toolkit which uses the convex hulls on receiver operating characteristic curve (ROCCH) (Bosaris Toolkit 2010).

4. RESULTS

This section the results of the above experiments. All the results are represented in terms of EER (%).

4.1. Effect of Number of Features

In the experiments we first study the effect of number of MFCC features on the performance. The EERs (%) for different number of MFCC features on both male and female speakers using the TIMIT-UBM and Oracle-UBM are shown in Figure 3. From the figure, Oracle-UBM yields smaller EERs than TIMIT-UBM as expected. This is because both the language and the recording conditions of the speech data used to train Oracle-UBM exactly matches with the evaluation data. However, using the TIMIT-UBM dramatically degrades the performance because of the mismatch between the UBM and the evaluation data. Using 20 MFCCs yields the smallest EER for both male and female speakers in Oracle-UBM case whereas the smallest EER is obtained by using 18 MFCCs for TIMIT-UBM case. Therefore, 20 MFCCs and 18 MFCCs will be used in the remaining experiments for Oracle-UBM and TIMIT-UBM, respectively

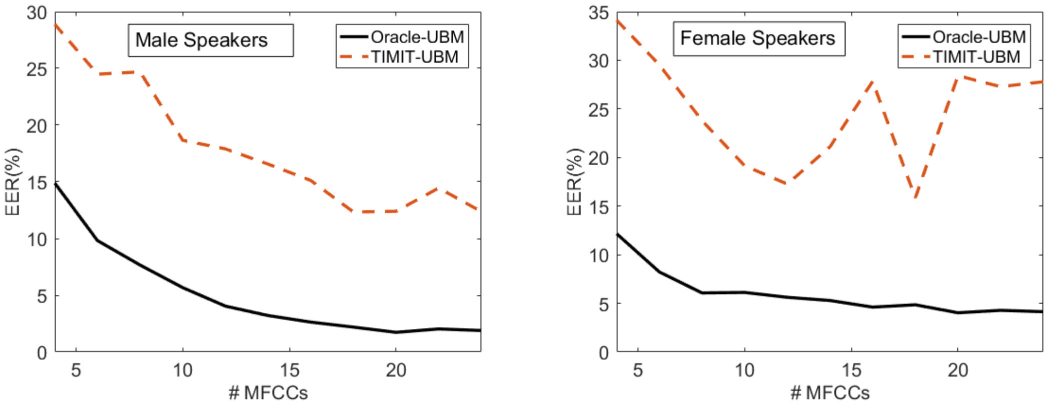

Figure 2:

EERs (%) v.s. number of MFCC features for male and female speakers

4.2. Effect of UBM Size

We next analyze the effect of UBM size - number of Gaussian components - in UBM. To this end, we vary the UBM size between 8 to 2048 and tried to optimize the number of Gaussian components. Table 2, shows the EERs (%) obtained with different UBM sizes using TIMIT-UBM and Oracle UBM. From the Table, using 512 Gaussian components in TIMIT-UBM gives the smallest EER for male speakers whereas 256 UBM size shows the best performance when Oracle-UBM is used. However, the performance difference between 256 and 512 UBM sizes for Oracle UBM is relatively small. 256 Gaussian components gives approximately 2.80% better performance than 512 Gaussians and this can be negligible. For female speakers in turn, independent from the UBM, 256 Gaussians yields the smallest EERs. In the remaining experiments, UBM size is fixed to 256 for both TIMIT-UBM and Oracle-UBM cases.

Table 2. Effect of UBM size on speaker verification performance

		UBM Size (M)								
	UBM	8	16	32	64	128	256	512	1024	2048
Male Speakers	TIMIT	21.57	20.12	18.31	16.37	13.74	12.34	**11.89**	12.49	15.61
	Oracle	4.09	3.60	2.93	2.09	1.99	**1.73**	1.78	1.88	3.76
Female Speakers	TIMIT	30.23	28.87	29.89	30.33	22.94	**15.85**	20.75	19.67	22.44
	Oracle	7.21	5.18	4.35	4.44	4.76	**4.03**	4.78	4.77	4.85

4.3. Effect of Feature Post-Processing

Next we compare the effect of feature post-processing on the speaker recognition performance. To be more precise we study the effect of the 0th MFCC feature (c_0) and the dynamic features (Δ and ΔΔ) - first and second order derivatives of the MFCC features - in addition to static MFCCs. The results are summarized in Table 3. We can see that including the c_0 to the features considerably improves the performance for male speakers in TIMIT-UBM case (EER reduces to 8.82% from 12.34%). However, for female speakers the raw MFCCs yields the best performance. Appending the dynamic features does not bring any improvement on the performance but also increases the EERs. For Oracle-UBM case, the raw features without any additional features yields the best performance. This is probably because, in general dynamic features are helpful to improve the recognition performance when there is a session variability in speech recordings. However, speech recordings in TURTEL database does not have this kind of variability.

Table 3. Effect of Feature Post-Processing on speaker recognition performance

Features	**TIMIT-UBM**		**Oracle-UBM**	
	Male	Female	Male	Female
MFCCs	12.34	**15.85**	**1.73**	**4.03**
MFCCs+ c_0	**8.82**	20.47	3.01	3.86
MFCCs $+ c_0 + \Delta$	12.83	19.68	2.50	4.80
MFCCs $+ c_0 + \Delta + \Delta\Delta$	12.38	22.95	2.41	4.66

5. CONCLUSION

In this paper, we study the effect of language and recording condition mismatch on speaker verification using Turkish speech database. There exist studies addressing this challenge in the literature but most of the studies report their findings using English speech corpora. The experiments carried out in this study showed that speaker verification performance dramatically degrades in case of language and recording condition mismatch between the UBM and the evaluation data. It was found that appending dynamic features does not boost the speaker recognition performance independent from the UBM data. Analyzing the effect of such mismatch using more state-of-the-art speaker modeling technique such as *i-vector* probabilistic linear discriminant analysis (PLDA) would be more interesting as a future work.

ACKNOWLEDGEMENT

This work has been supported by Bursa Technical University, Scientific Research Projects under project no 2016-01-012.

REFERENCES

1. Akbacak M. and Hansen, J. H. L, (2007) Language normalization for bilingual speaker recognition systems, *IEEE International Conference on Acostics, Speech and Signal Processing*, 257-260. doi:10.1109/ICASSP.2007.366898

2. Benesty, J. Sondhi, M. M. and Huang, Y. A., (2007) Springer Handbook of Speech Processing, Springer-Verlag, New York.

3. Bosaris Toolkit (2010). Access address: https://sites.google.com/site/bosaristoolkit/ (Accessed in 17.11.2016)

4. Büyük, O., and Arslan, L. M., (2012a) Model selection and score normalization for text-dependent single utterance speaker verification, *Turkish Journal of Electrical Engineering and Computer Science*, 20(2), 1277-1295. doi:10.3906/elk-1103-35

5. Büyük, O., and Arslan, L. M., (2012b) Combining log-spectral mean subtraction at different frequency resolutions for handset-channel compensation in single utterance speaker verification, *IET Signal Processing*, 6(9), 824-828. doi:10.1049/iet-spr.2011.0270

6. Dempster, A. P., Laird, N. M., and Rubin, D. B., (1977) Maximum likelihood from incomplete data via EM algorithm, Journal of the Royal Statistical Society, 39(1), 1-38. doi:10.2307/2984875

7. Hansen J. H.L and Hasan, T. (2015) Speaker recognition by machines and humans: A tutorial review, *IEEE Signal Processing Magazine*, 32(6), 74-99. doi:10.1109/MSP.2015.2462851

8. Luengo, I., Navas, E., Sainz, I, Saratxaga, I., Sanchez, J., Odriozola, I and Hernaez, I. (2008) Text independent speaker identification in multilingual environments, *LREC*, 1814-1817.

9. Ma, B. and Meng, H., (2004) English-Chinese bilingual text-independent speaker verification, *IEEE International Conference on Acostics*, *Speech and Signal Processing,* 293-296. doi: 10.1109/ICASSP.2004.1327105

10. Ma, B., Meng, H. M., and Mak, M. -W., (2007) Effects of device mismatch, language mismatch and environmental mismatch on speaker verification, *IEEE International Conference on Acostics, Speech and Signal Processing,* 301-304. doi:10.1109/ICASSP.2007.366909

11. Misra, A. and Hansen, J. H. L., (2014) Spoken language mismatch in speaker verification: An investigation with NIST-SRE and CRSS bi-ling corpora, *Spoken Language Technology*, 372-377. doi:10.1109/SLT.2014.7078603

12. Reynolds, D. A., Rose, R. C., (1995) Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Transactions on Speech and Audio Processing, 3(1), 72-83. doi:10.1109/89.365379

13. Reynolds, D. A., Quatieri, T. F., and Dunn, R. B., (2000) Speaker verification using adapted Gaussian mixture models, Digital Signal Processing, 10(1), 19-41. doi:10.1006/dspr.1999.0361