Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 12º Congresso de Engenharia de Áudio
18ª Convenção Nacional da AES Brasil
13 a 15 de Maio de 2014, São Paulo, SP

The Effect of Speech Rate on Automatic Speaker Verification: a Comparative Analysis of GMM-UBM and I-vector Based Methods

Anderson R. Avila,[1,2] Milton Sarria-Paja,[2] Francisco J. Fraga,[1] and Tiago H. Falk[2]

[1] Engineering, Modeling and Applied Social Sciences Center (CECS), Universidade Federal do ABC (UFABC), Santo André, São Paulo, Brazil

[2] INRS-EMT, University of Quebec, Montréal, Quebec, Canada

ABSTRACT

Automatic speaker verification (ASV) performance is known to degrade in mismatched test-train conditions. In this paper, we explore the effects of speech rate variation on both GMM-UBM and the state-of-the-art i-vector based ASV systems. In our experiments, we used normal speech for training and fast speech for testing to represent train-test mismatch. The results showed that, despite both methods being significantly affected by mismatch conditions, the performance degradation caused by speech rate variation can be mitigated by the addition of fast speech into the training set. Moreover, we verified that GMM-UBM outperforms the i-vector based system under mismatch conditions, although the same was not true under matched situations.

0 INTRODUCTION

Given the constantly increasing number of speech applications on smartphones and devices, one can assume that the market for speech technology will become widespread in the near future. According to recent reports [1], the industry is expected to hit over US$ 31 billion by 2017, due to the fast growing demand for three voice applications: automatic speech recognition (ASR), automatic speaker verification (ASV) and text-to-speech synthesis (TTS). Although ASV can achieve good accuracies under matched conditions (i.e., when no discrepancies are encountered between testing and training data), it is still a great challenge to keep performance at acceptable levels for real-world applications where mismatch between training and test conditions are seen. To overcome degradation, most recent research has focussed on channel effects, such as noise and reverberation [2, 3, 4], or on vocal effort (e.g., whisper and loud voices) [5, 6]. Speech rate though, as a form of speaking-style variation, has been overlooked by the ASV research community.

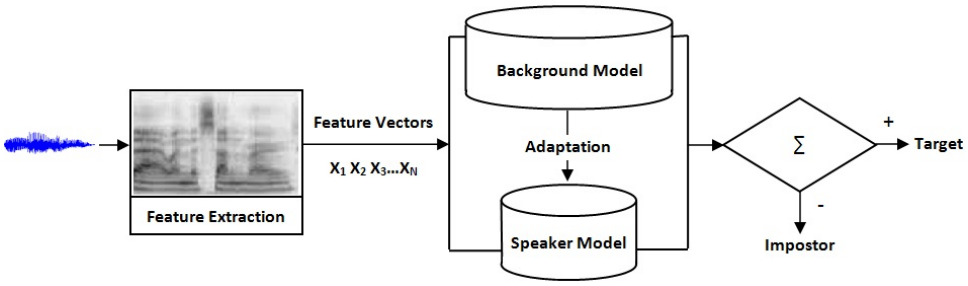

Figure 1: Generic block diagram of an automated speaker verification system.

This paper is concerned with speech rate variation which is an import issue for practical speech applications considering that speakers commonly vary their speed of articulation while producing an utterance. For instance, variation may occur either depending on the vocal effort being used [5] or due to the emotional state of the speaker. Some people may speak faster or even slower than usual just for being distressed while making an emergency call. Based on the assumption that average duration of a sentence increases under the consumption of alcohol, speech rate features have been used to detect intoxicated speech [7]. Moreover, the authors in [8] have investigated how speech rate and speaker's respiration are related. As can be seen, speech rate variability is a common factor in practical speech applications, yet little work has been done to investigate its effects on ASV performance.

It is known that intra-speaker variability has a negative impact on ASV. In [9], a number of experiments were conducted in order to show the detrimental influence of 10 types of disguised voices on automatic speaker recognition. In [6], the authors presented the impact of five different vocal effort levels (i.e., whisper, soft, normal, loud and shouting voice) on speaker recognition systems and also proposed the use of the so-called multiple model framework (MMF) associated with a specialized vocal effort classifier as a means to confer robustness to the system. The work presented in [10] has confirmed that it is possible to improve ASV performance by combining a fixed length of normal speech with variable lengths of whispered speech into the training set.

Motivated by these findings in related fields, the approach taken in this paper considers a speaking-style independent model where variable lengths of fast speech are used together with a fixed length of normal speech during enrolment. Gaussian mixture models (GMM) based systems with universal background models (UBM) are compared to the new so-called i-vector based system, which has demonstrated state-of-the-art results with normal speech. To the best of the authors' knowledge, these systems have yet to be tested under speech rate variation conditions.

This paper is organized as follows. Section 1 describes the ASV systems based on the GMM-UBM and i-vector frameworks. Section 2 presents the database used in this paper and also gives the details about the experimental setup. Results and discussion are presented in Section 3. Lastly, Section 4 concludes the paper.

1 ASV SYSTEMS

In automatic speaker verification, the goal is to decide whether or not a speech sample belongs to a claimed speaker. Figure 1 depicts a general scheme of such a system. The ASV problem can be approached by applying a likelihood-ratio test to a test utterance in order to decide if the claimer must be accepted as the genuine speaker or rejected as an impostor. The first hypothesis, H_0, states that X belongs to the claimer. The second hypothesis, H_1, states that X belongs to an impostor. According to [2], the likelihood-ratio is given by:

$$\frac{p\left(X \; the \; genuine \; speaker\right)}{p\left(X \; not \; the \; genuine \; speaker\right)} = \frac{p\left(X \mid \lambda_C\right)}{p\left(X \mid \lambda_{\overline{C}}\right)}, \quad (1)$$

where X is a set of feature vectors, λ_C stands for the model corresponding to the claimed speaker identity (H_0) and $\lambda_{\overline{C}}$ correspond to its complement, i.e., it must represent every speaker other than the claimant (H_1).

Transcribing the likelihood-ratio to the log domain corresponds to:

$$\sum = \log p\left(X \mid \lambda_C\right) - \log p\left(X \mid \lambda_{\overline{C}}\right). \quad (2)$$

The decision on accepting or rejecting the speaker identity is based on a threshold θ. The idea is to accept the claimer if $\sum > \theta$ and reject if $\sum < \theta$. In our experiments, the Microsoft Research MSR Identity Toolbox was used [11].

1.1 Pre-processing, feature extraction, and GMM-UBM

As a pre-processing step, each speech recording was downsampled to 8 kHz, pre-emphasized and normalized at -26 dBov (dB overload). Features used were the mel-scale cepstral coeficients (MFCC), obtained every 10-ms using a 20-ms Hamming window. After applying the FFT, a set of 24 triangular band-pass filters obeying the Mel scale were used. Discrete

cosine transformation (DCT) was also applied, resulting in 19 cepstral coefficients and log-energy. Delta and double-delta coefficients were then computed from the MFCCs and log-energy features, resulting in a final 60-dimensional feature vector. After features are extracted, a universal background model is characterized by a GMM (GMM-UBM) via the well-known expectation-maximization (EM) algorithm. Then, via maximum a posteriori (MAP) adaptation, speaker models are obtained for each speaker. In our experiments, 64-component GMMs were used.

1.2 I-Vector Based ASV

The i-vectors are obtained via a joint factor analysis (JFA) framework [12, 13]. More specifically, the speaker-dependent GMM mean components (obtained via GMM-UBM MAP adaptation) are combined into the so-called supervector *M*, which is assumed to convey speaker dependent, speaker independent, channel dependent, and residual components. Each component can be represented by a low-dimensional set of factors, which operate along the principal dimensions (also known as eigen-dimensions) of the corresponding component. Mathematically, this is represented as:

$$M = m + Vy + Ux + Dz, \tag{3}$$

where m is the speaker and channel-independent supervector, V the speaker eigenvoice matrix, D the diagonal residual matrix, U is the eigen-session matrix, and y, z, x correspond to the low-dimensional eigenvoice, speaker-specific eigen-residual, and eigen-channel factors, respectively.

Instead of assuming subspaces for modelling speaker and channel variability, as above, the authors in [14] proposed the use of a simple space, referred to as total variability space. The argument for this new approach relies on the fact that channel factors estimated by JFA contains information about speakers, as shown in the experiments performed in [14]. Hence, for a given utterance, both speaker and session components represented by (3) can be rewritten as:

$$M = m + Tw \tag{4}$$

where T corresponds to a retangular low-rank matrix and w a random vector with normal distribution. The so-called hidden variable w contains the component factors and is referred to as the identity vector (i.e., i-vector) [14]. As suggested by Fig. 2, m is the mean supervector extracted from the universal background model. Notice also in the diagram that parameter s defines the number of i-vectors and n its dimension.

Within the i-vector framework, the decision process in the total variability space consists basically in computing the similarity between the target speaker factors and test speaker factors. Support vector machines (SVM) using cosine kernels can be applied to the total variability decision process, but one can base the score only on the cosine kernel. Other techniques such as Within Class Covariance Normalization (WCCN), Linear Discriminant Analysis (LDA) and Nuisance Attribute Projection (NAP), have also been applied in order to remove channel effects [15]. In our experiments, we used 50 total factors defined by the total variability matrix T, and the dimensional reduction obtained by LDA led us to 35 factors. We haven't found any significant improvements by considering higher values for the total factors.

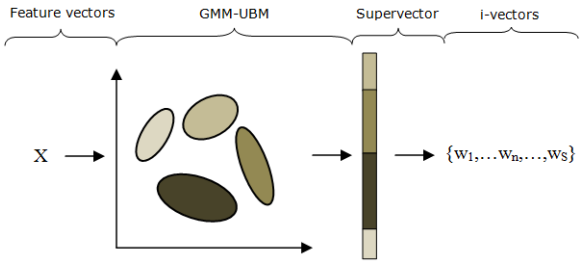

Figure 2: Block diagram of the i-vector extraction.

2 EXPERIMENTAL SETUP

The Chains corpus was used in our experiments [16]. The corpus features recordings of 36 subjects, including male and female with different accents. Six different speaking styles are available but our experiments are based only on the normal and fast speech conditions. Each speaker read 37 prepared texts which generated 37 distinct speech files. The content read was always the same independently of the type of vocal effort or speaking style used by the subject. The normal style was obtained with the subjects reading each text at a comfortable rate. Fast speech recordings, on the other hand, were obtained after subjects were given a speech sample to be taken as an example of the aimed rate.

During the enrolment phase, the first four speech files of each speaker were used for training, which added up to roughly 120 seconds for normal speech and around 60 seconds for fast speech. The remaining 33 speech files were used for testing, as the first four were already used for training. Both styles shared about the same length of data, 50 seconds, for testing. Three experiments were conducted. The aim of the first experiment was to compare the performance of the GMM-UBM and i-vectors based methods in matched train-test conditions. The second, on the other hand, investigated the effects of train-test mismatch. Lastly, the third experiment investigated the benefits obtained by including normal and increasing amounts of fast speech during training. We use two performance figures to gauge system performance, namely equal error rate (EER) and detection error tradeoff (DET) curves.

3 RESULTS AND DISCUSSION

Table 1 shows the performance results obtained for the two ASV systems under matched train-test condi-

tions. As can be seen, almost perfect verification performance is achieved with i-vectors in the normal-normal train-test matched conditions. Training and testing with fast speech decreased performance with both systems, but the i-vector based one still outperformed the GMM-UBM one.

Figure 3, on the other hand, shows DET curves for the two systems under mismatch conditions. Dashed curves show GMM-UBM (blue) and i-vector (red) performance obtained by training with normal speech and testing with fast speech. The solid curves, on the other hand, show the case where speaker models were obtained with fast speech and tested with normal speech. Comparisons with Table 1 clearly show the negative effects observed with mismatch conditions for both ASV systems. In this mismatched scenario, training with normal speech resulted in best performance and, unlike the matched case, in this mismatched experiment, the GMM-UBM system outperformed the i-vector one.

As mentioned previously, the third experiment investigated the effects of including normal speech and increasing amounts of fast speech during training. In this experiment, we investigated the addition of 20, 40, and 60 seconds of fast speech to the available 120 seconds of normal training speech. Figures 4 and 5 show plots of EER as a function of the amount of fast speech used during training for GMM-UBM and i-vector ASV, respectively. In both cases, addition of fast speech had a subtle detrimental effect on ASV performance with normal test data, but resulted in monotonically improved performance with fast test speech as the amount of training fast speech increased. Overall, the GMM-UBM framework outperformed the i-vector based one for both normal and fast test speech. Table 2 summarizes the EERs obtained with this third experiment. As can be seen, by adding 60 seconds of fast speech to the normal speech training data, decreases in EER of 78% and 77% can be achieved (relative to using normal speech alone for training) by the GMM-UBM and i-vector systems, respectively.

4 CONCLUSION

In this paper, two methods for ASV have been evaluated taking into consideration mismatched condition. The classical GMM-UBM and the state-of-the-art i-vector were compared towards robustness against speech rate variation. Both methods failed to overcome the issue of mismatched condition caused by normal and fast speech. We verify that i-vector outperformed GMM-UBM in matched situation. However, consid-

Train and test mode	GMM-UBM	i-vector
Normal	0.19%	0.04%
Fast	1.38%	0.52%

Table 1: GMM-UBM and i-vector performance in terms of EER (%) for matched train-test conditions

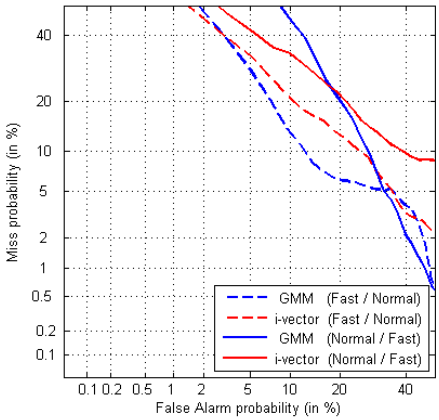

Figure 3: DET curve of GMM-UBM and i-vector ASV under mismatched conditions.

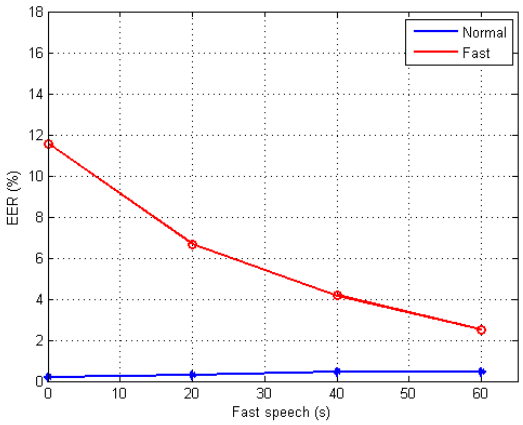

Figure 4: GMM-UBM equal error rates as a function of different lengths of fast speech added to normal speech during training.

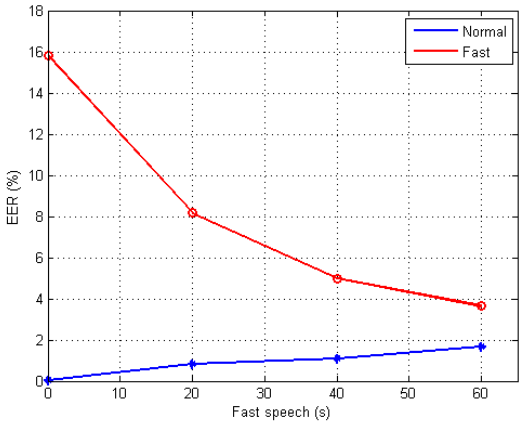

Figure 5: I-vector equal error rates as a function of different lengths of fast speech added to normal speech during training.

duration of fast speech (s)	Test (Fast)		Test (Normal)	
	GMM-UBM	i-vector	GMM-UBM	i-vector
0s	11.57%	15.83%	0.19%	0.04%
20s	6.66%	8.17%	0.31%	0.83%
40s	4.16%	5.00%	0.46%	1.08%
60s	2.5%	3.64%	0.45%	1.66%

Table 2: GMM-UBM and i-vector performance in terms of EER (%) considering variable lengths of fast speech into the training set.

ered mixing up a fixed length of normal speech and a variable length of fast speech, GMM-UBM surpasses i-vector in circumstances of mismatch. We have confirmed our hypothesis that speaking-style independent system based on normal and fast speech can mitigate performance degradation even when speakers vary their speech rate between training and testing phase.

ACKNOWLEDGEMENTS

The authors acknowledge funds from the Centre for Advanced Systems and Technologies in Communications (SYTACOM) and the Emerging Leaders in the Americas Program (ELAP).

REFERENCES

[1] "Speech technology: A global strategy business report globol industry analysis, inc.," Tech. Rep., 2012.

[2] D. A. Reynolds, "Automatic speaker recognition using gaussian mixture speaker models," *The Lincoln Laboratory Journal*, vol. 8, no. 2, pp. 173–192, 1995.

[3] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2023–2032, 2007.

[4] T.H. Falk and W.-Y. Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 90–100, 2010.

[5] C. Zhang and J.H.L. Hansen, "Analysis and classification of speech mode: whispered through shouted," in *Proc. Interspeech*, 2007, pp. 2289–2292.

[6] P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Communication Journal*, vol. 54, no. 6, pp. 732–742, 2012.

[7] D. Bone et. al., "Intoxicated speech detection by fusion of speaker normalized hierarchical features and gmm supervectors," in *Proc. Interspeech*, 2011, pp. 3217–3220.

[8] C. Dromey and L.O. Ramig, "Intentional changes in sound pressure level and rate - their impact on measures of respiration, phonation, and articulation," *Journal of Speech, Language, and Hearing Research*, vol. 41, no. 5, pp. 1003–1018, 1998.

[9] C. Zhang and T. Tan, "Voice disguise and automatic speaker recognition," *Forensic Science International*, vol. 175, no. 2, pp. 118–122, 2008.

[10] M. Sarria-Paja, T.H. Falk, and D. O'Shaughnessy, "Whispered speaker verification and gender detection using weighted instantaneous frequencies," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7209–7213.

[11] S.O. Sadjadi, M. Slaney, and L. Heck, "Msr identity toolbox - a matlab toolbox for speaker recognition research," Tech. Rep., Microsoft Research, Conversational Systems Research Center (CSRC), 2013.

[12] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," Tech. Rep., 2005.

[13] N. Dehak et. al., "Comparison between factor analysis and GMM support vector machines for speaker verification," in *Proc. IEEE-Odyssey of the Speaker and Language Recognition Workshop*, 2008.

[14] N. Dehak et. al., "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[15] N. Dehak et. al., "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proc. Interspeech*, 2009, vol. 9, pp. 1559–1562.

[16] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The chains speech corpus: Characterizing individual speakers," in *Proc. Int. Conf. on Speech and Computer (SPECOM)*, 2006, pp. 1–6.