I-vector Speaker Verification for Speech Degraded by Narrowband and Wide-band Channels

Laura Fernández Gallardo[1,2]*, Michael Wagner*[2,3]*, Sebastian Möller*[1,2]

[1] Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany
[2] Faculty of Education, Science, Technology and Mathematics, University of Canberra, Australia
[3] College of Engineering and Computer Science, Australian National University, Australia
Email: {Laura.Fernandez-Gallardo,Sebastian.Moeller}@telekom.de, Michael.Wagner@canberra.edu.au

Abstract

Voice biometrics are frequently exposed to channel degradations of transmitted speech and to channel mismatch between enrollment and test utterances, which cause speaker recognition systems to perform poorly. In this paper, the influence of channel bandwidth and speech coding on speaker verification is assessed employing the state-of-the-art i-vector technique. Our focus is on the possible benefits of enhanced wideband over narrowband and on the effects of codec mismatch and bandwidth mismatch. Our results on subsets of the NIST SRE (Speaker Recognition Evaluation) 2010 and of the TIMIT corpus show that the performance with wideband data is significantly better than that employing narrowband signals for matched and codec-mismatched conditions. In the presence of bandwidth mismatch, a relative improvement of 40-70% can be obtained by downsampling the wideband signal to 8 kHz.

1 Introduction

A growing number of applications using automatic speaker verification (ASV) require the transmission of the user's voice to remote servers that perform the identity validation, for example, in phone banking or in phone shopping. Efforts in the last decades of ASV research have been concentrated on reducing the within-speaker variations caused by channel mismatch, originated when the utterances for enrollment and the utterances for testing are transmitted through channels of different characteristics. Dehak et al. have recently proposed the i-vector technique [1], an approach to front-end analysis which offers both excellent discriminative capacity and small dimensionality.

The notable verification performance of the i-vector technique has been confirmed, among other investigations, in the last NIST Speaker Recognition Evaluation (SRE) [2] under the effects of session variability, environmental noise, and varying test-sample lengths. There has, however, been little or no consideration of the effects of different signal bandwidths. Most commercial ASV applications operate with speech transmitted through narrowband (NB, 300 – 3,400 Hz) or through wideband (WB, 50 – 7,000 Hz) communication channels, which can degrade the speech to different extents. It has been found that speech intelligibility and quality [3] as well as the human speaker identification ability [4] are enhanced with the migration from NB to WB. Other investigations have shown the benefits of the extended bandwidth for automatic speaker recognition [5, 6]. To the best of our knowledge, only the analysis presented in [7] examines the effects of NB and WB telephone speech on automatic speaker identification. The authors employed HMMs and an evaluation database of only 10 speakers and reported no improvement of WB communications over NB. However, there are no studies examining the effects of bandwidth nor of wideband coded-decoded speech employing the i-vector technique. The effects of codec mismatch have been investigated with traditional classifiers such as Hidden Markov Models (HMM) [8], Gaussian Mixture Model-Universal Background Model (GMM-UBM) [9], and Support Vector Machine (SVM)-GMM systems. These studies conclude that the verification performance is degraded in mismatched conditions, yet only codecs of NB transmissions have been tested. In this paper we also analyze the influence of WB-transmitted speech using the i-vector technique for speaker verification and compare it to that of traditional NB telephone speech. We then investigate the effects of codec mismatch and bandwidth mismatch between enrollment and test utterances.

In the past SRE12 challenge [2], the organizers proposed microphone speech for enrollment and testing which, for the first time, was sampled at 16 kHz, that is, presented an extended range of frequencies rather than the telephone recorded data. However, these data were not processed through any speech codec, and the participants, concerned with the challenging noisy conditions, did not attempt to take advantage of the enhanced bandwidth. In the top-performing systems of the competition, the authors pooled the microphone data together with telephone speech and downsampled the signals to 8 kHz to avoid the bandwidth mismatch [10, 11]. Differently, in our analysis we employ wideband SRE 2010 interview data [12], which was made available for researchers after the scheduled challenge (for which only NB data was proposed) and was recorded through microphones and sampled at 16 kHz. This signal bandwidth allowed us to transmit the speech through simulated NB and WB communication channels, controlling the degradations of the data. We then trained separate i-vector extractors employing different sets of clean and transmitted speech with the same degradations. The systems created were challenged with the processed NIST data, with unseen codecs, and in the presence of codec mismatch and bandwidth mismatch. In addition, we validated the results on a different evaluation database, namely the test partition of the TIMIT dataset, which was also transmitted through the same communication channels.

The structure of this paper is as follows. Section 2 describes the communication channels employed to transmit the speech of our experiments. Section 3 details the organization of the speech data for the i-vector extractors.

Section 4 provides the results of the different evaluations and Section 5 concludes the work.

2 Speech Transmission

For the purposes of this study microphone speech segments from different datasets were transmitted through simulated communication channels presenting different characteristics. In order to control the different types of distortion applied to the signals, we needed: first, speech recorded directly through microphones in clean conditions, that is, not previously transmitted, and second, the sampling frequency to be of at least 16 kHz, which permits the study of WB communications. Speech corpora meeting these requirements and employed in our experiments are: the wideband SRE 2010 release mentioned in the previous section, TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT), Resource Management Corpus 2.0 Part 1 (RMI), North American Business News Corpus (CSRNAB1), Wall Street Journal Continuous Speech Recognition Phase I (WSJ0), and Phase II (WSJ1). They contain only one language, American English, and only male speakers were considered. The i-vector extractors were trained with combined speech from the train partition of TIMIT and the last four datasets, totaling 670 speakers and with approximately 89h of speech. We refer to this combined dataset as *development data* in this paper. These systems were evaluated on two separate sets. One was the common evaluation condition 1 of the SRE10 challenge, which involved 1,231 speaker models (telephone files for enrollment were discarded as these were only NB) and 29,176 test files of 991 male speakers and only interview speech [12]. The second evaluation set was the test partition of the TIMIT database, containing 112 male speakers, which was set aside from the *development data*. We refer to these sets as *Ev_1* and *Ev_2*, respectively.

All the speech (for development and for evaluation) was transmitted through the following communication channels:

- G.711 speech codec at a bit rate of 64kbps (NB)
- AMR-NB speech codec at 12.2kbps (NB)
- G.722 speech codec at 64kbps (WB)
- AMR-WB speech codec at 12.65kbps (WB)

We believe that these four codecs represent a good selection of NB and WB compression schemes, being current standards for digital telephony, Voice over Internet Protocol (VoIP), and wireless telephony. They offer different speech quality, which was assessed in [3].

The channel transmissions entailed bandwidth-filtering the signals complying with the International Telecommunication Union (ITU) recommendations G.712 and P.341 for NB and WB, respectively. Then, the coding-decoding processes were applied via software simulation applying standard ITU and 3GPP tools. In addition to these four versions of the original data, we also kept the unprocessed data (cleanWB) and a sixth version, which consisted of the speech downsampled to 8 kHz, via an anti-aliasing lowpass FIR filter, with no further processing (cleanNB).

3 I-vector systems

In the i-vector approach [1], speaker and channel variability are modelled in the same low-dimensional total-variability space $\boldsymbol{T}$. An i-vector speaker and channel dependent GMM supervector can be represented by

$$M = m + \boldsymbol{T}w \tag{1}$$

where m is the UBM mean supervector, $\boldsymbol{T}$ the total-variability matrix defining the total variability space, and w an independent normally-distributed random vector representing the total-variability factors.

We trained six codec-specific i-vector extractors separately, employing a different version of the development data for each of them. Hence, the UBM and the total variability matrix $\boldsymbol{T}$ were estimated from either clean or coded-decoded data in NB or in WB. We refer to each extractor with the name of the codec applied to the training utterances: *CleanNB*, *G.711*, *AMR-NB*, *CleanWB*, *G.722*, and *AMR-WB*. The two systems trained with clean data will serve as NB and as WB baseline, respectively.

The speech was parameterized using Mel Frequency Cepstral Coefficients (MFCCs) of 20 dimensions plus energy and deltas and double deltas, resulting in a 63-dimensional feature vector. They were extracted using a 25ms Hamming window with 10ms frame shift. The UBMs were built with 1024 Gaussian mixtures, and the $\boldsymbol{T}$ matrix estimated with 400 total factors. Five iterations were used for the EM training. The i-vector extraction and the cosine distance scoring processes were implemented in Matlab.

Compensation methods, such as probabilistic linear discriminant analysis (PLDA) were not performed as our microphone databases are too small for proper training of the PLDA model parameters. We believe that training the PLDA on extensive data reflecting the same generation process as the evaluation segments would offer acceptable compensated results [13], yet further sets of microphone NIST data sampled at 16 kHz were not available at the time our experiments were conducted.

4 Results and Discussion

4.1 Performance with No Channel Mismatch

The i-vector extractors were first confronted with speech presenting the same distortion as in the development set. The performances of the systems in terms of the Equal Error Rate (EER) for the two evaluation sets are given in Table 1. Overall, worse results are obtained for the Ev_1, that is, the evaluation condition 1 of the NIST SRE 2010 challenge, in comparison to Ev_2, the test partition of the TIMIT database. This difference in performance, consistent throughout this section, can be attributed to the different speaker populations and to the greater session variability of the NIST dataset.

The systems trained with WB data offer a statistically significantly better performance than those trained with NB data, as indicated by the statistical significance test of [14], with 95% confidence. This confirms the advantages of WB communications over NB for i-vector speaker verification. Moreover, the CleanNB and CleanWB systems

Enroll / test	i-vector extractor	Ev_1	Ev_2
CleanNB / CleanNB	CleanNB	8.48	3.41
CleanWB / CleanWB	CleanWB	**5.76**	**1.46**
(NB) G.711 / G.711	G.711	10.10	4.29
(NB) AMR-NB / AMR-NB	AMR-NB	10.99	5.01
(WB) G.722 / G.722	G.722	**7.07**	**1.80**
(WB) AMR-WB / AMR-WB	AMR-WB	7.07	2.52

Table 1: EERs of the codec-specific i-vector extractors confronted with data of the same distortion.

perform better than the systems operating with coded-decoded speech of the same bandwidth. An explanation for this fact can be that the total variability spaces of the clean i-vector extractors (spanned by the respective $\boldsymbol{T}$ matrices) were estimated from unprocessed speech, presenting low channel variability. This causes the performance to be higher than for the other systems built from transmitted speech since the coding-decoding processes introduce non-linear distortions into the signal.

It seems that the G.722 codec, because of its underlying speech synthesis algorithm, can offer better performance than the other WB codec AMR-WB, manifested in the case of the TIMIT evaluation. This suggests that the G.722 introduce less session variability effect.

4.2 Performance under Codec Mismatch

We considered two situations of codec mismatch in this paper. The first one occurs when a codec- specific system is evaluated with data transmitted through an unseen codec, and the second one is the typical case of mismatch between enrollment and test utterances. Both situations could occur in real applications for speaker authentication if they receive data transmitted through various communication channels.

Since it is straightforward to detect the signal bandwidth (simply by measuring the energy of frequency components above 3.4 kHz), we assume that a practical application would be able to select a verification system of the same bandwidth as that of the signals at enrollment and verification time. Thus, for each evaluation, we applied the i-vector extractor built with speech of the same bandwidth as the evaluation data.

Table 2 presents the results of each of the codec-specific systems when the enrollment and test segments are transmitted through an unseen codec of the same bandwidth. The worse performance in WB is significantly superior to the best performance in NB [14]. Employing the same enroll/test codecs for the evaluation of the CleanNB and the CleanWB extractors resulted in a degraded performance with respect to the results of Table 2, because of the mismatch between clean and coded-decoded speech.

Regarding NB communications, it can be seen that the AMR-NB system performs better if the G.711 instead of the AMR-NB codec is applied to the enroll/test segments, which indicates the benefits of landline communications with the G.711 codec over AMR-NB for ASV. This difference in performance is significant for the two evaluation datasets. For WB transmissions, an inconsistency between the two evaluations can be observed. However, because there is no statistical difference between the EERs given by the WB systems we can assume that both G.722 and AMR-WB would lead to comparable performance in this situation.

Enroll / test	i-vector extractor	Ev_1	Ev_2
(NB) G.711 / G.711	AMR-NB	10.30	4.82
(NB) AMR-NB / AMR-NB	G.711	13.13	8.75
(WB) G.722 / G.722	AMR-WB	**6.97**	3.07
(WB) AMR-WB / AMR-WB	G.722	7.27	**2.68**

Table 2: EERs for unseen codecs of the same bandwidth.

Enroll / test	i-vector extractor	Ev_1	Ev_2
(NB) G.711 / AMR-NB	G.711	13.13	6.39
	AMR-NB	**11.51**	**4.59**
(NB) AMR-NB / G.711	G.711	12.74	6.41
	AMR-NB	**11.46**	**4.82**
(WB) G.722 / AMR-WB	G.722	**7.27**	**2.34**
	AMR-WB	7.47	3.75
(WB) AMR-WB / G.722	G.722	**7.33**	**3.21**
	AMR-WB	7.52	3.43

Table 3: EERs in the case of codec mismatch between enroll and test segments.

The results in the case of codec mismatch between enroll and test utterances are presented in Table 3. The G.722 and AMR-NB extractors offer the best results against enroll/test mismatch in NB and in WB, respectively, for all the evaluations and consistently for Ev_1 and Ev_2. WB offers significantly better results than NB for these best performing systems [14].

The selection of a codec-specific system based on the codec of the test signals can be problematic in cases where an application does not have access to the communication protocol, e.g. in VoIP, and thus does not have information about the codec employed in the transmission. Hence, the general good behavior in situations of codec mismatch is an important benefit of the i-vector framework, since an ASV application would not require the selection of a codec-specific system (targeted to the codec of the transmission) at verification time for an acceptable performance. Our results, on two separate evaluation databases, suggest that employing a i-vector extractor developed with signals transmitted through AMR-NB and G.722 system for NB and for WB, respectively, would offer the best performance in case of codec mismatch.

4.3 Performance under Bandwidth Mismatch

Situations of bandwidth mismatch often occur in forensic speaker identification scenarios. In a forensic investigation, the voice of the offender is generally obtained from telephone transmissions, while the voice of the suspect, to be compared to that of the offender, is recorded in clean

Enroll / test	i-vector extractor	Ev_1	Ev_2
a) Clean WB / AMR-NB	CleanWB	46.31	48.04
	AMR-NB	48.77	50.00
b) Clean NB / AMR-NB	CleanNB	34.12	19.29
	AMR-NB	**27.47**	**14.97**
c) Clean WB / AMR-NB (16 kHz)	CleanWB	33.64	23.93

Table 4: EERs in a plausible forensic scenario.

conditions during police interrogations. Since NB communications are still predominating, it is plausible that samples transmitted through wireless networks in NB (applying the codec AMR-NB) have to be compared to clean high-quality signals (cleanWB). Table 4 shows the performance in case of a) direct bandwidth mismatch, and when we attempted to reduce it by either b) downsampling the WB signal to 8 kHz, or c) upsampling the NB signal to 16 kHz.

The best performing approach is to employ 8 kHz signals and the AMR-NB system, which permits a 40-70% relative EER reduction from the situation of direct bandwidth mismatch (the performance in this case was chance level). The NB evaluation of approach b) yields, interestingly, better results than the WB evaluation of c), which offers a 27-50% relative EER reduction from a). This is possibly due to undesirable effects of the upsampling process, which degrade the performance of the CleanWB system. Other codec-specific systems offered worse results for the evaluation sets of b) and c).

5 Conclusions

This paper has examined the effects of bandwidth limitation and of speech coding by means of a set of experiments employing the state-of-the-art i-vector technique, with emphasis on codec mismatch and on bandwidth mismatch. Two different datasets, the evaluation condition 1 of the past NIST SRE 2010 challenge and the test partition of the TIMIT database were employed to validate the results. We show that WB communications lead to a significantly superior performance compared to NB, in matched conditions and under codec mismatch. Codec-specific i-vector extractors built with utterances transmitted through the AMR-NB and the G.722 codecs perform better than the other systems tested under codec mismatch. The effects of bandwidth mismatch, which typically occur in forensic speaker identification scenarios, can be better reduced by downsampling the WB signal rather than by upsampling the NB signal.

In future work we would also like to examine the possible benefits of super-wideband communications for speaker recognition, which offer an even more extended bandwidth of 50 – 14,000 Hz and higher signal quality. However, it may still take some time until large speaker recognition datasets of sufficient bandwidth are made available.

References

[1] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. and Ouellet, P., "Front-End Factor Analysis for Speaker Verification," IEEE Trans. Audio, Speech, and Lang. Process., vol. 19, no. 99, pp. 788-798, 2010.

[2] Greenberg, C.S., Stanford, V.M., Martin, A.F., Yadagiri, M., Doddington, G.R., Godfrey, J.J. and Hernandez-Cordero, J., "The 2012 NIST Speaker Recognition Evaluation," Interspeech, 2013.

[3] Möller, S., Raake, A., Kitawaki, N., Takahashi, A. and Wältermann, M., "Impairment Factor Framework for Wideband Speech Codecs," IEEE Trans. Audio, Speech and Language Processing, vol. 14, no. 6, pp.1969–1976, 2006.

[4] Fernández Gallardo, L., Möller, S. and Wagner, M., "Human Speaker Identification of Known Voices Transmitted Through Different User Interfaces and Transmission Channels," ICASSP, 2013.

[5] Besacier, L., Grassi, S., Dufaux, A., Ansorge, M. and Pellandini, F., "GSM Speech Coding and Speaker Recognition," ICASSP, vol. 2, pp. 1085-1088, 2000.

[6] Pradhan, G. and Prasanna, S. R. M., "Significance of Speaker Information in Wideband Speech," National Conference on Communication, 2011.

[7] Jokić, I., Jokić, S., Gnjatović, M., Sečujski, M. and Delić V., "The Impact of Telephone Channels on the Accuracy of Automatic Speaker Recognition," Telfor Journal vol. 3, no. 2, 2011.

[8] Fakhr, W., Abdelsalam, A. and Hamdy, N., "Enhancement of Mismatched Conditions in Speaker Recognition for Multimedia Applications," in ICASSP, vol. 1, pp. 377-80, 2004.

[9] Quatieri, T. F., Singer, E., Dunn, R. B., Reynolds, D. A. and Campbell, J. P., "Speaker and Language Recognition Using Speech Codec Parameters," Eurospeech, vol. 2, pp. 787-790, 1999.

[10] Ferrer, L., McLaren , M., Scheffer, N., Lei, Y., Graciarena, M. and Mitra, V., "A Noise-Robust System for NIST 2012 Speaker Recognition Evaluation," Interspeech, 2013.

[11] Saeidi, R. et al., "I4U Submission to NIST SRE 2012: A Large-Scale Collaborative Effort for Noise-Robust Speaker Verification," Interspeech, 2013.

[12] A. F. Martin, A.F. and Greenberg, C.S., "The NIST 2010 Speaker Recognition Evaluation," ICASSP, vol. 1, pp. 2726–2729, 2010.

[13] Garcia-Romero, D., Zhou, X. and Espy-Wilson, C.Y. "Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," ICASSP, pp. 4257-4260, 2012.

[14] Bengio, S. and Mariéthoz, J., "A Statistical Significance Test for Person Authentication," The Speaker and Language Recognition Workshop (Odyssey), pp. 237-244, 2004.