I-vector Speaker Verification based on Phonetic Information under Transmission Channel Effects

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

[1] Faculty of Education, Science, Technology and Mathematics, University of Canberra, Australia
[2] College of Engineering and Computer Science, Australian National University, Australia
[3] Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany

(laura.fernandezgallardo|michael.wagner)@canberra.edu.au, sebastian.moeller@telekom.de

Abstract

Past studies have shown evidence of important speaker-specific content in the higher frequencies of the spectrum, which are filtered out by narrowband channels. Besides, wideband transmissions, which are gaining ground over narrowband communications, offer an extended range of frequencies which account not only for better speech quality and intelligibility, but also for an improved speaker recognition performance. In this work, different phoneme classes (fricatives, nasals, and vowels) were removed from speech of different bandwidths, and a series of i-vector based speaker verification experiments were conducted. Our results show that the performance enhancement with clean wideband speech with respect to clean narrowband speech is principally due to the presence of unvoiced fricative consonants. The effects of codec schemes of different bandwidths on the aforementioned speech are discussed.

Index Terms: Automatic speaker verification, i-vectors, phonetic information, channel impairments

1. Introduction

While narrowband (NB, 300 – 3,400 Hz) is still predominant in most of today's communication networks, efforts have been made towards motivating the transition to an extended frequency band, namely wideband (WB, 50 – 7,000 Hz). The superiority of WB over NB has been shown in terms of intelligibility, quality, and also in human and automatic speaker recognition [1, 2, 3]. This last fact indicates the presence of relevant speaker-specific information carried by the frequencies over 3,400 Hz and/or under 300 Hz. The present study intends to elucidate possible reasons for the advantages of WB over NB for automatic speaker verification (ASV), exploring the role of various phonemes in the different frequency bands by using a state-of-the-art i-vector speaker recognizer [4] followed by a Gaussian Probabilistic Linear Discriminant Analysis (G-PLDA) back-end with length normalization [5].

A number of early studies have analyzed the speaker discriminating properties of individual phonemes or phoneme classes [6, 7, 8]. The authors agreed that vowels and nasals provide the best discrimination between speakers. It has also been asserted that fricatives contribute to the speaker recognition performance to a lesser extent and that stop sounds are the least useful phoneme category [8]. These findings have been applied to speaker recognition approaches that take advantage of the most speaker-distinctive sounds, aided by a phoneme detector [9, 10].

Although vowel sounds have proven to be effective for characterizing individual speakers and been widely used for speaker recognition and in forensic analyses [11], there is a growing interest in exploiting also the discriminative properties of fricatives and nasals. Fricative consonants differ among speakers owing to their articulatory and acoustic properties [12] and, due to the complex and relatively fixed nasal and paranasal cavities of talkers, nasal consonants display low within-speaker and high between-speaker variability [13]. Interestingly, a recent study has shown that fricative and nasals can be more useful than vowels for speaker discrimination [14]. The authors examined the speaker discrimination ability of phonemes applying different bandwidth filters and computing the F-ratios, i.e. the ratio of inter- and intra-speaker variance. In this paper we are also interested in the speaker-discriminative potential of sounds with focus on the differences between NB and WB. Especially, we consider fricatives and nasals because they exhibit spectral peaks at high frequencies, from 4 to 7 kHz depending on the particular phoneme [15], which are suppressed in NB channels. NB channel filtering eliminates also the important nasal content below 300 Hz. Unlike other studies with a similar purpose, we analyze the performance offered by i-vector speaker verification, considering clean speech of 8 kHz and of 4 kHz bandwidth and also the transmission through standard telephone communications.

Together with the analysis of channel bandwidth effects on phoneme classes, we also investigate the influence of NB and WB speech coding. The invariance of phonemes through channel transmissions involving coding and decoding processes has been investigated for text-dependent ASV [16], phoneme recognition [17], and some forensic analyses [18]. These studies have reported the unforeseen alteration of formant frequencies [16, 18], and the alterations of consonants and fricatives in particular due to telephone and cellular channels, with respect to clean speech [17]. However, only NB coding has been analyzed.

In our experiments, we removed different phoneme classes (voiced and unvoiced fricatives, nasals, and vowels) from clean speech creating different versions of the data. The TIMIT database, although not challenging for i-vector speaker verification, was chosen for this purpose for three reasons. First, it was recorded in clean conditions, which allows us to control the channel distortions of this study. Second, the sampling frequency is 16 kHz, permitting the study of WB (and of NB by downsampling). Third, it is conveniently annotated at phone level, easing the phoneme filtering. Once the phonemes were removed, the original and the phoneme-filtered speech were transmitted through communication channels of different bandwidth and codecs, in order to study the effects of channel degradations.

The remainder of the paper is structured as follows. Section 2 details the procedure to remove certain phonemes from the original speech and Section 3 describes the communication channels through which this speech is transmitted. After the preparation of the speech material, the i-vector experiments were conducted as shown in Section 4. Section 5 presents the results and a discussion, and Section 6 the conclusion of our work.

2. Phoneme filtering

In order to study the influence of phoneme classes on the i-vector performance, several phoneme-filtering conditions were applied to the original speech removing particular phonemes. The database employed to this end was TIMIT, containing speech recorded directly through microphones in clean conditions, with 16 kHz sampling frequency, and including time-aligned phone-level transcriptions. Only the test partition (TIMIT_test) of this dataset was processed, and the train partition (TIMIT_train) was retained for training the i-vector extractors. A total of 112 males belong to TIMIT_test, each of them uttering 10 phonetically rich sentences.

The voiced and unvoiced fricatives, nasals, and vowels that were indicated in the phonetic transcriptions were removed from the original speech. We also kept the original segments with no phoneme-filtering for our experiments, which will be referred as all-phonemes speech. Thus, 6 conditions were applied, resulting in the all-phonemes and five phoneme-filtered versions of TIMIT_test:

a) original speech containing all the phonemes
b) speech with no voiced fricatives /v, dh, z, zh/
c) speech with no unvoiced fricatives /f, th, s, sh/
d) speech with no voiced and no unvoiced fricatives
e) speech with no nasals /m, n, ng, em, en, eng, nx/
f) speech with no vowels /iy, ih, eh, ey, ae, aa, aw, ay, ah, ao, oy, ow, uh, uw, ux, er, ax, ix, axr, ax-h/

More information on the phone codecs can be found on the TIMIT database documentation.

Since some phoneme classes appear more often in speech than others, the phoneme-filtered versions of the utterances had different durations. To avoid the possible effects of different amounts of evaluation data on the i-vector scores, the speech segments were cropped after the voice activity detection stage (VAD) so that their lengths across conditions were the same, including the original speech segments.

3. Speech transmission

The all-phonemes and the phoneme-filtered segments were processed following different procedures for the analysis of the ASV performance with clean speech and with telephone-transmitted speech of different bandwidths.

For the analysis of clean WB speech, the sampling frequency of the utterances was maintained at 16 kHz, while for the analysis of NB clean speech the segments were downsampled to 8 kHz by applying an anti-aliasing filter. No further processing was applied to these two sets.

Differently, for the analyses of WB and NB telephone speech, the segments were subsequently transmitted through the four communication channels listed in Table 1 via software simulation. These channels consisted of a bandwidth filter and a speech compression scheme with a particular bit rate. The bandwidth filters comply with the standard implementation of the International Telecommunication Union (ITU)-T Recommendations G.712 and P.341 for NB and WB, respectively. The coding-decoding processes were applied using standard ITU and 3GPP tools for channel simulation.

Bandwidth	Codec	Bit rate (kbps)	Coding type
Narrowband	G.711	64	A-law companded pulse code modulation (PCM)
	AMR-NB	12.2	Algebraic Code Excited Linear Prediction (ACELP)
Wideband	AMR-WB	12.65	
	G.722	64	Sub-band adaptive differential pulse code modulation (SB-ADPCM)

Table 1. *Four telephone channels studied in our work, besides clean NB and clean WB speech.*

The described processing results in a total of 36 sets with which to evaluate the i-vector/G-PLDA systems. This corresponds to 6 phoneme-filtering conditions transmitted through 6 channel settings, including clean and telephone speech.

4. I-vector experiments

Six different i-vector extractors followed by G-PLDA back-ends have been created, separately trained with speech presenting different kind of distortions. The systems are referred as the distortion of the utterances employed for development and for evaluation: CleanNB, CleanWB, G711, AMRNB, G722, and AMRWB.

The following databases were combined for the estimation of the Universal Background Model (UBM), of the total variability matrix T of the i-vector approach, detailed in [4], and of the G-PLDA parameters [5]: TIMIT_train, Resource Management Corpus 2.0 Part 1 (RM1), North American Business News Corpus (CSRNAB1), Wall Street Journal Continuous Speech Recognition Phase I (WSJ0), WSJ Phase II (WSJ1), and a portion of AusTalk [19]. These datasets were employed because they contain clean microphone speech, which allows us to control the telephone channel degradations to be applied, and the bandwidth of the samples is sufficient for the study of WB. The number of speakers combining the six datasets is 770 males, and the approximate total duration of speech is 105 h. Larger amounts of audio data would certainly provide better verification results, yet the performance obtained with our data permitted the comparisons between the effects of the different phoneme-filtering conditions.

These data for development, containing all the phonemes, were transmitted through communication channels as indicated in Section 3. Hence, 6 versions of the originally recorded datasets were created, with different channel degradations. Each of the versions was employed to train one i-vector/GPLDA system. The amount of speech and the number of speakers were thus constant for the six systems.

Aspects shared by the systems are as follows. Feature vectors consisting of 63 components: 20 Mel Frequency Cepstral Coefficients (MFCCs) together with a log energy feature, extracted using a 25ms Hamming window with 10ms frame shift, and the corresponding delta and delta-delta coefficients. UBM models of 1024 Gaussian mixtures. 400 total factors for the estimation of the T matrix. G-PLDA with length normalization [5] employing 120 eigenvoices. The i-vector extraction process was implemented in Matlab and the scripts for the G-PLDA, available online[1], were run for compensation in the i-vector space and for scoring.

The systems trained were evaluated on the 36 different sets described in Section 3. Each system was confronted with evaluation data of the corresponding bandwidth and codec. For each of the evaluation sets there were 10 sentences per speaker; 5 of which were combined to extract one enrolment i-vector, and 5 test i-vectors were extracted from the remaining utterances.

5. Results and discussion

5.1. Clean speech

This subsection presents the performance of the CleanNB and of the CleanWB systems confronted with clean NB and with clean WB speech, respectively, which was processed with the 6 phoneme-filtering conditions a) to f) mentioned in Section 2. Figure 1 shows the Half Total Error Rates (HTERs) for each evaluation and the associated 95% confidence intervals, calculated as indicated in [20]. As stated before, the results are not affected by different lengths of the evaluation utterances but only by the absence of certain phoneme classes on speech, and by the signal bandwidth.

As expected, the CleanWB system offers an improved performance over the CleanNB system, since the data with which it was built and evaluated offered an extended range of frequencies. The differences in performance between NB and WB are statistically significant [20] for the all-phonemes data and for every phoneme-filtering condition except for “No unvoiced fricatives” and for “No fricatives”.

There are three implications of this finding: First, it shows that the presence of fricatives in speech, with important peaks at high frequencies [15], is relevant for the superiority of WB over NB automatic speaker verification. For speech without fricatives, the performances of the NB and the WB systems are comparable, i.e. not significantly different, yet incorporating the fricative consonants causes these performances to differ significantly. This outcome agrees with [14]. The authors asserted that when the frequencies above 4 kHz were removed, the fricative consonants were less useful for speaker discrimination. Second, voiced fricatives are not as meaningful as unvoiced fricatives for the higher performance with WB data. This may be explained by the fact that the unvoiced fricative /s/, which appears more often than other fricatives in the TIMIT speech (about twice more often than its voiced counterpart /z/) and possesses the greatest inter-speaker variability among unvoiced fricatives [12], exhibits spectral peaks above 6.5 kHz [15], which is not present in the clean NB signals. Third, the nasals and the vowels seem not to have an influence on the better accuracy provided by WB signals, contrasting with the effects of fricatives. Their absence causes the decrease of the performance in NB and in WB to approximately the same extent, i.e. the performances under these conditions are significantly different comparing clean NB and clean WB, which indicates that these sounds are equally important for speaker verification in both bandwidths.

In particular, vowels seem to be essential for an acceptable speaker recognition performance due to their great speaker-discriminative power in NB and in WB [11]. This is in concordance with the early studies on the importance of phonemes for distinguishing among speakers [6, 7, 8].

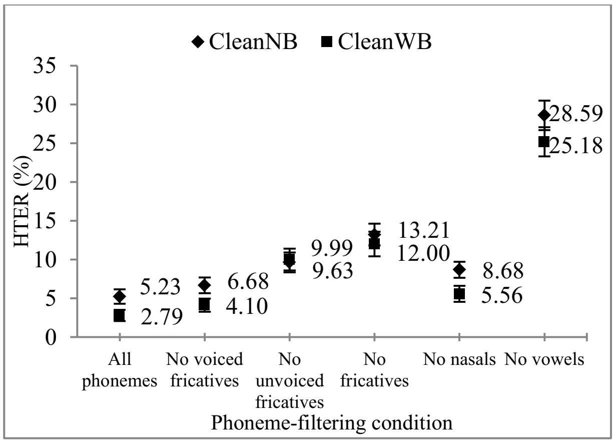

Figure 1: *HTERs and 95% confidence intervals of the CleanNB and of the CleanWB i-vector systems.*

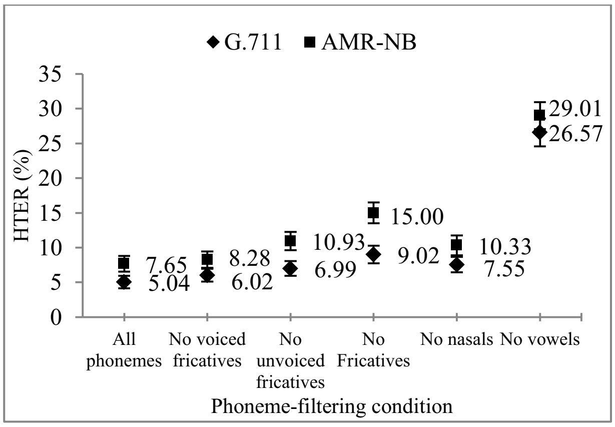

Figure 2: *HTERs and 95% confidence intervals of the G711 and of the AMRNB i-vector systems (NB).*

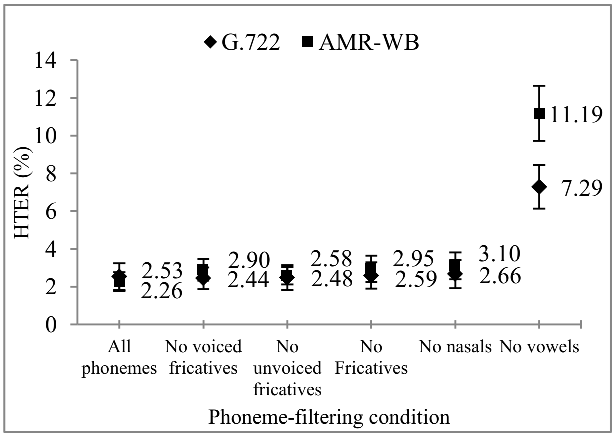

Figure 3: *HTERs and 95% confidence intervals of the G722 and of the AMRWB i-vector systems (WB).*

5.2. Coded-decoded speech

The all-phonemes and phoneme-filtered segments transmitted through different NB and WB telephone channels were employed to evaluate the i-vector/G-PLDA systems built from utterances presenting the same distortion. In all the

[1] https://sites.google.com/site/dgromeroweb/software/

evaluations, WB coding yielded a significantly higher verification performance compared to NB coding, improving by approximately 50 to 70% relative HTER for all-phonemes data.

Figure 2 and Figure 3 display the HTERs for each condition under NB and WB coding, respectively. Coded speech offers slightly better performance than clean speech for all-phonemes data of the corresponding bandwidth, with the exception of AMR-NB. This is due to the PLDA compensation, which eliminates the effects introduced by the coding-decoding processes. The HTERs given by cosine distance scoring (not compensated) were 7.97, 2.81, 8.72, 10.15, 3.66, and 4.55, for the CleanNB, CleanWB, G711, AMRNB, G722, AMRWB systems, respectively. This shows certain degradation in the results comparing clean with transmitted speech, although in this paper we only consider the results given by the PLDA back-end.

Interestingly, our results show that WB coding causes the performance to be almost unaltered despite the phoneme-filtering conditions as opposed to NB and to clean data. For NB telephone degradations, there is a statistically significant difference between the HTER of all-phonemes and the HTERs of the phoneme-filtered data (with the exception of the "No voiced fricatives" segments), for both the codecs. Hence, the absence of the different phonemes in speech causes a detriment of the performance under NB coding, revealing their importance for speaker verification with NB-transmitted signals. The G.711 codec performs significantly better than the AMR-NB for all the evaluations except for the "No vowels" condition, where both performances are comparable.

Differently, however, the suppression of the same phoneme classes does not affect the performance under WB coding with the exception of the "No vowels" condition. No significant differences have been found between the performance with all-phonemes data and that with data without fricatives or without nasals for any of the WB codecs. It seems that the absence of fricative and nasal phonemes is somehow compensated by the WB coding process. The speech codecs generally employ some sort of underlying speech synthesis model to obtain the reconstructed signal in the decoding process [3]. It is likely that the complex synthesis algorithms of WB codecs, operating with signals of extended bandwidth, compensate for the absence of phoneme classes. In general, G.722 coding leads to better verification results than AMR-WB, although the differences in HTERs between these codecs are only statistically significant for the "No vowels" condition.

Thus, the reason for the superiority of WB- over NB-transmitted speech for speaker recognition is not as clear as for clean speech. The suppression of vowels in WB degrades the performance of all-phonemes speech to a greater extent than the suppression of other phoneme classes. Nevertheless, the "No vowels" condition also causes a great detriment of the performance in NB, which indicates that vowels exhibit important speaker discrimination in both bandwidths and that the superior WB speaker verification is not only due to the presence of vowels. At this stage, we can only assume that both the extended bandwidth and the efficient coding schemes in WB, able to preserve and even to enhance the speaker-discriminative characteristics for the phoneme-filtering conditions, are the causes for the better performance in WB compared to NB. More investigation in this respect is needed.

6. Conclusions

In this work we attempt to clarify the reasons for the improved automatic speaker recognition performance with WB data compared to that with NB data. We have investigated the i-vector/G-PLDA speaker verification performance with and without the presence of certain phonemes in NB and in WB speech, in clean conditions and transmitted through degraded channels.

Our results show that, for clean speech, the presence of fricative consonants, particularly of the unvoiced fricatives /f, th, s, sh/, permits a significantly better speaker discrimination in WB than in NB, since their suppression leads to a comparable performance in both bandwidths. This reveals that unvoiced fricatives are effective for speaker discrimination at high frequencies. The same effect is not manifest when nasals or vowels are suppressed, although the removal of vowels cause the greatest detriment of performance among all the phoneme-filtering conditions due to their speaker-discriminative power, essential in both bandwidths. With respect to NB- and WB-transmitted speech, the absence of different phoneme classes causes very little effect on the performance of with WB speech, while for NB speech the results are degraded to different extents depending on the phoneme-filtering condition applied. This result suggests that the WB codecs tested are able to preserve and even to enhance the speaker-discriminative characteristics for the phoneme-filtering conditions in comparison to uncoded speech and to NB. The characteristics of the WB codec algorithms, along with the extended bandwidth with which the codecs operate, might be among the reasons for the superiority of WB over NB.

In future work we would like to conduct a similar analysis considering individual phonemes instead of phoneme classes to explore their role for NB and WB transmissions. In addition, subject to the availability of speech corpora of enough bandwidth, this analysis will be extended to female speech and to super-wideband communications, which are currently gaining adoption in the marketplace.

7. References

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

[2] Pradhan, G. and Prasanna, S. R. M., "Significance of Speaker Information in Wideband Speech," National Conf. on Communication (NCC), 2011.

[3] Fernández Gallardo, L., Wagner, M. and Möller, S., "Spectral Sub-band Analysis of Speaker Verification Employing Narrowband and Wideband Speech," Speaker Odyssey: the Speaker and Language Recognition Workshop, 2014.

[4] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. and Ouellet, P., "Front-End Factor Analysis for Speaker Verification," Audio, Speech, and Language Processing, vol. 19, no. 99, pp. 788 – 798, 2010.

[5] Garcia-Romero, D. and Espy-Wilson, C.Y., "Analysis of I-vector Length Normalization in Speaker Recognition Systems," Interspeech, 2011.

[6] Wolf, J.J., "Efficient Acoustic Parameters for Speaker Recognition," Journal of the Acoustical Society of America, vol.51, no. 6, pp. 2044-2056, 1972.

[7] Sambur, M.R., "Selection of Acoustic Features for Speaker Identification," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 23, pp. 176–182, 1975.

[8] Eatock, J.P. and Mason, J.S., "A Quantitative Assessment of the. Relative Speaker Discriminative Properties of Phonemes," ICASSP, pp. 1133-1136, 1994.

[9] Auckenthaler, R., Parris, E.S. and Carey, M. J, "Improving a GMM Speaker Verification System by Phonetic Weighting," ICASSP, pp.313-316, 1999.

[10] Hansen, E.G., Slyh, R.E. and Anderson, T.R., "Speaker Recognition Using Phoneme-Specific GMMs," Speaker Odyssey: the Speaker Recognition Workshop, 2004.

[11] Rose, P., "Forensic Speaker Identification", New York: Taylor & Francis, 2002.

[12] Gordon, M., Barthmaier, P. and Sands, K., "A Cross-linguistic Acoustic Study of Voiceless Fricatives," Journal of the International Phonetic Association, vol. 32, no. 2, pp. 141-174, 2002.

[13] Stevens, K. N., "Acoustic Phonetics". Cambridge, MA: MIT Press, 1999.

[14] Schindler, C. and Draxler, C., "The Influence of Bandwidth Limitation on the Speaker Discriminating Potential of Nasals and Fricatives," International Association for Forensic Phonetics and Acoustics (IAFPA), 2013.

[15] Jongman, A., Wayland, R. and Wong, S., "Acoustic Characteristics of English fricatives," Journal of the Acoustical Society of America, vol. 108, pp. 1252-1263, 2000.

[16] Phythian, M., Ingram, J. and Sridharan, S., "Effects of Speech Coding on Text-dependent Speaker Recognition," Region 10 Conference (TENCON), vol 1, pp. 137-140, 1997.

[17] Junqua, J. C., "Impact of the Unknown Communication Channel on Automatic Speech Recognition: A Review," Eurospeech, 1997.

[18] Guillemin, B. J. and Watson, C. I., "Impact of the GSM AMR Speech Codec on Formant Information Important to Forensic Speaker Identification," Australasian International Conference on Speech Science and Technology, 2006.

[19] Wagner, M., Tran, D., Togneri, R., Rose, P., Powers, D., Onslow, M., Loakes, D., Lewis, T., Kuratate, T., Kinoshita, Y., Kemp, N., Ishihara, S., Ingram, J., Hajek, J., Grayden, D., Göcke, R., Fletcher, J., Estival, D., Epps, J., Dale, R., Cutler, A., Cox, F., Chetty, G., Cassidy, S., Butcher, A., Burnham, D., Bird, S., Best, C., Bennamoun, M., Arciuli, J. and Ambikairajah. E., "The Big Australian Speech Corpus (The Big ASC)," Australasian Int Conf on Speech Science and Technology, pp 166-170, 2010.

[20] Bengio, S. and Mariéthoz, J., "A Statistical Significance Test for Person Authentication," Speaker Odyssey: the Speaker Recognition Workshop, pp. 237-244, 2004.