Advantages of Wideband over Narrowband Channels for Speaker Verification Employing MFCCs and LFCCs

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

10.21437/Interspeech.2014-286

Abstract

Wideband communications permit the transmission of an extended frequency range compared to the traditional narrowband. While benefits for automatic speaker recognition can be expected, the extent of the contribution of the additional bandwidth in wideband is still unclear. This work compares the i-vector speaker verification performances employing speech signals of 0-4 kHz, 4-8 kHz, and 0-8 kHz and different sets of cepstral features extracted using linearly- and a mel-spaced filterbanks. Analyses of clean speech and of speech transmitted through commonly employed codecs are conducted separately for male and for female speech. Our evaluation on two different datasets shows the improved speaker verification performance with the extended bandwidth, and also that the linear scale can lead to better results for narrowband signals. The advantages of linear- over mel-scaled features for wideband depend on the speakers' gender and on the channel distortion.

Index Terms: speaker verification, channel degradation, MFCC, LFCC

1. Introduction

The emerging wideband (WB) communications, offered by Voice over IP (VoIP) services, permit the transmission of an extended signal bandwidth (0.05-7 kHz) compared to the traditional narrowband (NB, 0.3-3.4 kHz). Benefits of this extended bandwidth have been found for speech quality and intelligibility [1], and also for human speaker identification [2] and automatic speaker verification [3][4]. However, the usefulness of the frequency range beyond the narrowband cut-off frequencies and the best manner to extract the conveyed speaker-specific information have not yet been determined.

The present study has two main objectives. First, to compare the speaker verification performances employing clean speech signals of 0-4 kHz, 4-8 kHz (approximately the range added in WB speech), and 0-8 kHz (full range). We trained and evaluated an i-vector system [5] – state-of-the-art for speaker verification – with these band-limited signals. We also compared the performance with that offered by NB and WB coded-decoded signals, received after telephone transmissions in today's typical scenarios where the speaker authentication is performed in remote servers, such as telephone banking or purchasing. For that, we degraded the original signals transmitting them through simulated communication channels and performed the i-vector experiments in the same manner. Because of the channel impairments degrading the signal quality, we expect lower speaker verification performance compared to that using clean speech [1].

The second objective of our work is to determine whether Mel-Frequency Cepstral Coefficients (MFCCs) are appropriate for i-vector speaker verification in order to take full advantage of the WB signal. They are extensively used in this paradigm and offer an acceptable performance also in WB (improving over NB) [3][4]. However, because of the mel-scale, based on human auditory characteristics, and because of the MFCCs being originally developed for speech recognition and for signals band-limited to 5 kHz [6], this feature set might not offer the best speaker verification performance compared to others. In particular, when signals with a bandwidth of 7 kHz (WB) or above are available, it may be desirable to have a greater resolution of the filters in the filterbank in order to emphasize the higher frequencies instead of following the mel-scale.

Previous studies have confirmed that the speaker-specific information is non-uniformly distributed in frequency bands. Different investigations employing different speech material and methods to detect speaker-discriminative bands, such as speaker recognition experiments [7] and F-ratios [8] [9] [10], agree that important speaker discriminative information is found in the spectral regions 0-0.5 kHz and beyond 3 kHz approximately. The non-uniform distribution of intrinsic speaker characteristics is attributed to the occurrence of phonetic events. For instance, vowel formants convey speaker individuality [7] and are manifested at higher frequencies for female speech due to their shorter vocal tract compared to males [11]. Due to physiological characteristics of speakers, nasals present discriminative power in low and mid-high frequencies [10] [12], and other consonants in the upper part of the frequency spectrum, above 6kHz [10].

Motivated by the presence of important speaker-specific information beyond the NB range, where more emphasis for feature extraction may be needed, we compared the speaker verification performance with MFCCs to that obtained with Linear-Frequency Cepstral Coefficients (LFCCs). The latter implies uniform spacing between the filters and overlap of 50%, thus giving equal importance to each frequency band, which we hypothesized that would lead to an improved performance when speaker discriminative information is present. Other investigations have also addressed this comparison [11][12], yet only the NB telephone data provided by the NIST speaker recognition evaluations (SREs) were employed. The authors of [12] found that LFCCs outperform MFCCs for consonant regions. [11] asserted that the superiority of LFCCs over MFCCs is accentuated for female speech due to the presence of important speaker-specific content at higher frequencies. The analysis in [9], however, shows that MFCCs can offer better results compared to LFCCs considering clean signals of 8 kHz bandwidth, and employing a small database of 22 males and 13 females and GMM speaker identification.

Differently, in this work we attempt to clarify the differences in performance between MFCCs and LFCCs employing clean and coded-decoded speech of different bandwidths from two different evaluation datasets, and focusing on the gender particularities. Our development and evaluation sets differ from those employed in NIST challenges [11][12] since clean speech of a bandwidth of at least 8 kHz was necessary in our analysis. The controlled transmission conditions of our prepared speech data will permit us the study of only the selected channel impairments, namely bandwidth limitation and codec, which is not possible with the already band-limited and distorted NIST speech. Despite the smaller amount of training data, the results obtained were informative and permitted comparisons between the effects of different bandwidths, channel degradations, and feature sets.

The structure of this paper is as follows. The experimental setup, describing the audio material and the procedures for feature extraction and for the i-vector development and evaluation, is given in Section 2. The results are presented in Section 3 and discussed in Section 4. Section 5 concludes and outlines future work.

2. Experimental setup

2.1. Audio material

The i-vector extractor of our experiments was trained and evaluated on either MFCCs or LFCCs from a combined set of different databases, which had previously been processed according to the different conditions of our study.

The requirements to select the speech material were that the data had not been processed or distorted through the transmission over a handset or a communication channel. Thus, the telephone data from NIST SREs could not be employed. Furthermore, the data should have a sampling frequency of at least 16 kHz for WB transmission, hence microphone NIST SRE speech, which is sampled at 8 kHz, is not suitable for our analysis either. We employed the following databases of microphone speech of American English meeting our requirements: 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).

Four degraded versions of the mentioned datasets were created, transmitting the clean speech though the following NB and WB channels applying a bandwidth filter and a codec:

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

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 using standard ITU and 3GPP tools for transmission channel simulation. In addition to these coded-decoded versions we also kept the clean original data.

2.2. MFCCs and LFCCs extraction

The MFCC and LFCC features were extracted employing mel-spaced filterbanks and linearly-spaced filterbanks, respectively. The different sets of MFCCs and LFCCs were extracted for male and for female speech for each condition of Table 1. In each case, the coefficients were computed using a 25ms Hamming window with 10ms frame shift and the corresponding deltas and delta-deltas. 20 coefficients were extracted if N>=21, or N-1 coefficients if N<21, where N is the number of filters in the filterbank. The 0th coefficient and the log energy were discarded. Delta and delta-delta coefficients were also included in the feature vector in each case.

The output of each filter band accounts for the frequency energy around its central frequency. Note that applying the mel scale implies having more filter resolution in the lower band.

Condition	Mel scale		Linear scale	
	N	fl - fh (Hz)	N	fl - fh (Hz)
Clean 0-4 kHz	24	0 - 4022	16	0 - 4121
Clean 4-8 kHz	8	3675 - 8000	16	3878 - 8000
Clean 0-8 kHz	32	0 - 8000	32	0 - 8000
G.711 (NB)	32	300 - 3400	32	300 - 3400
AMR-NB (NB)	32	300 - 3400	32	300 - 3400
G.722 (WB)	32	50 - 7000	32	50 - 7000
AMR-WB (WB)	32	50 - 7000	32	50 - 7000

Table 1. *Filter scaling, number of filters (N), low end of the first filter (fl), and high end of the last filter (fh) of the filterbanks used to extract cepstral coefficients from data of each bandwidth and distortion.*

2.3. I-vector extractor

The gender-dependent i-vector experiments were conducted separately from each of the 14 feature sets of Table 1. They were performed employing the same total number of recordings and the same total number of speakers for each gender, to eliminate the effects of different amounts of training data in the comparison of performances. In addition, two different datasets were used for the evaluation of the systems in order to assess the consistency of our results, which may be affected by the different speaker populations and speech content. One evaluation dataset was the test partition of the TIMIT corpus (TIMIT_test), containing 56 speakers, and the rest of datasets mentioned in sub-section 2.1 were combined to build the i-vector extractor, totaling in 420 speakers. For a second evaluation the WSJ0 database, with 59 speakers, was employed, and the speech from the remaining 417 speakers of the other datasets pooled for the development of the i-vector extractor.

The Universal Background Models (UBMs) were built with 1024 mixtures, and the T matrix estimated with 400 total factors. The i-vector extraction and the cosine distance scoring processes were implemented in Matlab. Compensation methods, such as probabilistic linear discriminant analysis were not performed in order to provide a better understanding of the effects of the different conditions on raw scores, eliminating the influence of modeling techniques for compensation, commonly applied in the presence of channel mismatch.

Of the 10 utterances per speaker in our evaluation data for TIMIT_test and for WSJ0, 5 were concatenated for speaker enrolment and 5 were used for testing. Using each possible pair of enroll/test utterances, this generated 5 client scores per speaker and (N-1) x 5 impostor scores per speaker, where the

number of speakers N is 56 and 59 for TIMIT_test and for WSJ0, respectively. Identical experiments were performed for male and for female speech independently, with the same number of speakers in each case.

3. Results

This section presents the i-vector experiment results obtained after cosine distance scoring for the TIMIT_test and for the WSJ0 evaluations. The results are given in terms of the Half Total Error Rate (HTER), assuming equal prior probabilities and detection error costs. The statistical significance test presented in [13] was applied to evaluate the significance of the differences in accuracy given by the mel and the linear scales. The differences with a confidence level above 95% within the same evaluation are indicated by * in the tables of the next subsections.

3.1. Male speech

It can be observed in Table 2 that the wider band of frequencies improves the speaker verification results, therefore WB offer better results than NB in every case for male speech and this difference in performance is statistically significant. As expected, the performance drops if transmitted speech instead of clean speech is employed by the system.

For clean speech and the TIMIT_test evaluation, the linear scale offers significantly better accuracy than the mel scale above 4 kHz, which contributes to the overall better performance of LFCCs for the full signal (0-8kHz). For WSJ0, the performances of MFCCs and LFCCs are comparable in each of the bands. For coded-decoded speech, the linear scale leads to an improvement over the mel scale for NB data whereas for WB the results with the mel scale are significantly better that those with the linear scale (they are similar in the case of the G.722 codec and TIMIT_test evaluation).

3.2. Female speech

As in the case of male speech, the enhanced WB leads to superior performance compared to NB also for female speech, as shown in Table 3. This difference in performance is statistically significant.

The verification results in the band 0-4 kHz and in NB (0.3-3.4 kHz) are significantly better in the case of LFCCs compared to MFCCs, revealing important speaker discriminative information conveyed by the higher frequencies of these ranges for female speech. However, the mel scale seems to offer higher accuracy than the linear scale for clean speech of 4 – 8 kHz. For the band 0-8 kHz and for WB coded-decoded speech, (0.05-7 kHz), some inconsistencies between the results of the two evaluation datasets can be observed. However, if we concentrate on the statistical differences between the two sets of features, LFCCs outperform MFCCs for the clean data of WSJ0 and for data transmitted through the AMR-WB codec in the case of the TIMIT_test evaluation.

4. Discussion

From the results presented in the last section it is clear that WB offers an improvement over NB for speaker verification due to the extended range of frequencies transmitted. The verification performance with the band 4-8 kHz is similar yet slightly worse than that with the band 0-4kHz, which evidences important speaker-specific content beyond the NB range, agreeing with past studies [7] [9] [10]. Our main finding is that both sub-bands, 0-4 kHz and 4-8 kHz contribute almost equally in the case of linear scaling to a better performance in the band 0-8 kHz.

Band	TIMIT_test		WSJ0	
	Mel	Linear	Mel	Linear
0 – 4 kHz	3.63	**3.61**	6.88	**5.69**
4 – 8 kHz	5.75*	**3.82***	7.42	**6.88**
0 – 8 kHz	2.14*	**1.33***	**2.43**	3.01
G.711 (NB)	7.96	**7.00**	28.87	**25.88**
AMR-NB (NB)	9.69	**9.52**	25.76	**24.18**
G.722 (WB)	2.31	**2.27**	**9.51***	12.88*
AMR-WB (WB)	**3.45***	4.57*	**8.36***	13.30*

Table 2. *HTERs for clean and coded-decoded male speech.*

Band	TIMIT_test		WSJ0	
	Mel	Linear	Mel	Linear
0 – 4 kHz	8.83*	**5.06***	12.02*	**7.66***
4 – 8 kHz	**5.41**	5.52	**7.44***	9.09*
0 – 8 kHz	**2.18**	2.19	6.03*	**4.48***
G.711 (NB)	16.70*	**11.48***	20.23*	**14.73***
AMR-NB (NB)	16.98*	**12.52***	18.62*	**14.15***
G.722 (WB)	3.98	**3.90**	**5.62**	5.63
AMR-WB (WB)	5.63*	**4.25***	**6.23**	6.85

Table 3. *HTERs for clean and coded-decoded female speech.*

There exist inconsistencies between the results of the TIMIT_test and the WSJ0 evaluations when comparing the effectiveness of MFCCs to that of LFCCs. This is mainly due to the different speaker populations and speech material with different phonetic content in each database. However, our attention is focused on the concordance of certain results and on significant differences from which conclusions can be drawn. The consistent outcomes from both datasets are: 1) LFCCs outperform MFCCs for the 0-4 kHz bandwidth and for NB-transmitted speech, 2) the superiority of LFCCs below 4 kHz is greatly manifested for female speech, 3) LFCCs outperform MFCCs in the band 4-8 kHz for male speech but not for female speech, 4) the performance in the band 0-8 kHz seems to be either better for LFCCs or not statistically different between the two feature sets, 5) for male WB-transmitted speech MFCCs generally outperform LFCCs, 6) similar results are obtained when comparing both feature sets for female WB-transmitted speech except for the case of TIMIT_test distorted with the AMR-WB codec, where LFCCs offer better performance than MFCCs.

In order to justify the differences found between the performances with speech of each gender we examined the speaker-discriminative properties of different frequency subbands by computing F-ratio values from clean speech [14]. The F-ratio measure accounts for the relation between the variance of features between speakers and the variance within a speaker and is computed as:

$$F(k) = \frac{\sum_{i=1}^{M}(u_i(k)-u(k))^2}{\sum_{i=1}^{M}\frac{1}{N}\sum_{j=1}^{N}(x_i^j(k)-u_i(k))^2} \quad (1)$$

Where $x_i^j(k)$is the energy in the kth sub-band of the j-th speech frame of the i-th speaker with k = 1,…32, j = 1,…,N, and i = 1,…M. $u_i(k)$ and $u(k)$ are the averages of the sub-band energy for speaker i and for all speakers, respectively, defined as:

$$u_i(k) = \frac{1}{N}\sum_{j=1}^{N} x_i^j(\mathrm{k}) \quad (2)$$

$$u(k) = \frac{1}{M}\sum_{i=1}^{M} u_i(k) \quad (3)$$

The higher the F-ratio, the more speaker-specific information is conveyed by the spectral sub-band. The F-ratio values are plotted in Figure 1 for the two evaluation datasets and for 32 linearly-spaced sub-bands, along with the mel and linear filterbanks. This gives an intuitive idea of the regions where higher filter resolution might lead to an improvement in the verification results. It can be observed that higher F-ratios are found in the regions below 0.5 kHz and between 2 and 4 kHz, approximately. The higher mel filter resolution in the band 0.5-2 kHz seems to be unnecessary and leads to worse verification results than LFCCs, Also, because of the higher formants of female speech, the discriminative regions are more concentrated at higher frequencies of the spectra compared to male speech, as can be seen in the F-ratios plot. These observations can explain the outcomes 1) and 2), which are in concordance with [11]. According to our analysis of F-ratios of both datasets, male speech is discriminative in the band 6-7 kHz but female speech is not. This may be due to the presence of a formant region, most likely originated by unvoiced fricatives, with energy at high frequencies [10], and to the discriminative power of vowels. It seems that this speaker-specific region could be originated for females at higher frequencies [15], which are above 8 kHz and regrettably already filtered out for both datasets. Because of their higher resolution, LFCCs can gather the discriminative information more efficiently than MFCCs in the case of male speech, while for female speech lesser filters (as in the mel scale) are desirable for a better performance, which explains 3). These facts contribute to an overall superiority of LFCCs over MFCCs for clean speech of 8 kHz bandwidth, as stated in 4).

The channel transmissions had also different effects on the verification performance with the two sets of cepstral coefficients. Worse results are obtained with transmitted speech than with clean speech, due to the band-limiting filter and codec distortions. As asserted before, the performance in NB is better for both genders when LFCCs instead of MFCCs are employed, that is, the speaker-discrimination properties along the sub-bands are not severely degraded. The G.711 codec, operating with low complexity, generates high audio quality compared to the more bandwidth-efficient AMR-NB. It seems that G.711 provides better verification results compared to AMR-NB for TIMIT_test, while for WSJ0 AMR-NB is slightly but not significantly better than G.711. Regarding WB transmissions and the outcomes 5) and 6), the channel coding seems to harm the speaker-specific contents at high-frequencies, specially the AMR-WB codec [4], which results in a better performance of MFCCs over LFCCs for male speech. Contrastingly, these effects seem to be beneficial for LFCCs and female speech, probably mainly due to the removal of the frequencies above 7 kHz with no speaker discriminative information. The performance provided by AMR-WB is generally worse compared to that of G.722, possibly due to the differences in codec complexity.

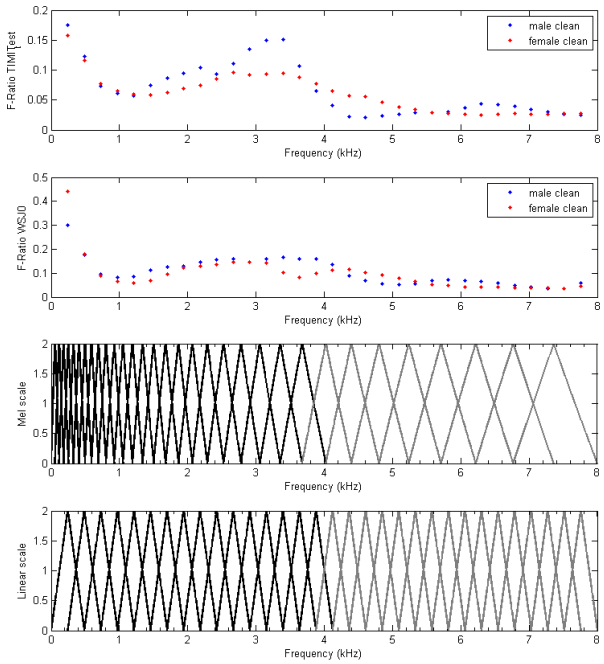

Figure 1: *F-ratio values of the TIMIT_test and of the WSJ0 datasets and mel and linear filterbanks.*

5. Conclusions and future work

We have examined the effects of bandwidth, coding and scaling of the filterbanks for cepstral feature extraction on the performance of i-vector speaker verification. With respect to our first objective, evaluating the effectiveness of the frequency band beyond narrowband, our results show that the band 4-8 kHz provides a performance similar to that obtained with the band 0-4 kHz in the case of LFCCs due to the presence of speaker-discriminative information beyond 4 kHz. Consequently, the band 0-8 kHz and wideband-transmitted speech offer better accuracy compared to that offered by signals of a narrower bandwidth. This difference in performance is statistically significant with at least 95% confidence.

Regarding our second objective, comparing the effects of MFCCs and LFCCs, the latter offer better results than MFCCs for signals of a bandwidth of 4 kHz and for narrowband data, this advantage being more accentuated for female speech. Differently, for the band 4-8 kHz, LFCCs are superior than MFCCs only for male speech, which may be attributable to speaker discriminative properties around 6 kHz, that are not exhibited by female speech. The effects of WB channel transmissions seem to alter the higher frequencies, causing MFCCs to offer better performance for male speech and LFCCs for female speech. Generally, the low-complexity codecs G.711 and G.722 offer better speaker verification performance compared to AMR-NB and AMR-WB, respectively.

Our future work will explore the bark filterbank for feature extraction and examine the speaker recognition performance from signals of a greater bandwidth. In particular, we would like to address the effects of super-wideband transmissions on the verification results and their influence on the spectral regions that carry speaker-discriminative content.

6. References

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

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

[3] Fernández Gallardo, L., Wagner, M. and Möller, S., "I-vector Speaker Verification for Speech Degraded by Narrowband and Wideband Channels," ITG Conference on Speech Communication, 2014.

[4] 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.

[5] 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.

[6] Davis, S. and Mermelstein, P., "Comparison of Parametric Representations for Monosyllable Word Recognition in Continuously Spoken Sentences," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, no. 4, pp. 357-366, 1980.

[7] Besacier, L. and Bonastre, J. F., "Subband Approach for Automatic Speaker Recognition: Optimal Division of the Frequency Domain," Audio and Video based Biometric Person Authentication, 195−202, 1997.

[8] Orman, D., Arslan, L., "Frequency Analysis of Speaker Identification," Speaker Odyssey: the Speaker Recognition Workshop, pp. 219–222, 2001.

[9] Lu, X., and Dang, J.,"Physiological Feature Extraction for Text-independent Speaker Identification Using Non-uniform Subband Processing", in Proc. of ICASSP, 2007.

[10] Hyon S., Wang, H., Wei, J., Dang, J., "An Investigation of Dependencies between Frequency Components and Speaker Characteristics based on Phoneme Mean F-ratio Contribution," Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp.1-4, 2012.

[11] Zhou, X., Garcia-Romero, D., Duraiswami, R., Espy-Wilson, C. and Shamma, S., "Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition," Automatic Speech Recognition and Understanding Workshop (ASRU), pp.559-564, 2011.

[12] Lei, H. and Lopez-Gonzalo, E., "Mel, Linear, and Antimel Frequency Cepstral Coefficients in Broad Phonetic Regions for Telephone Speaker Recognition," Interspeech, 2009.

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

[14] Wolf, J.J., "Efficient Acoustic Parameters for Speaker Recognition," Journal of the Acoustical Society of America, vol.51, no. 6 (Part 2), pp. 2044-2056, 1972.

[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.