Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

Spectral Sub-band Analysis of Speaker Verification Employing Narrowband and Wideband Speech

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

It is well known that the speaker discriminative information is not equally distributed over the spectral domain. However, it is still not clear whether that distribution is altered when the speech is transmitted through telecommunication channels, which introduce different kinds of degradations. In this paper we address the analysis of different frequency sub-bands when the speech is distorted with different bandwidth filters and channel codecs, considering narrowband and wideband communications. Our i-vector experiments on different sub-bands with 782 speakers show that standard landline codecs perform generally better than wireless codecs due to their intrinsic coding algorithm, their performance being close to, but slightly worse than that of uncoded speech. Wideband signals offer significant benefits over narrowband for speaker verification. A smaller experiment with 21 speakers leads us to believe that the emerging super-wideband transmissions may provide even better results because it shows important speaker-specific content in the band 8-14kHz.

1. Introduction

The speaker-discriminative properties of different frequency sub-bands are of main interest to select the best performing features to perform speaker authentication. It has been widely asserted [e.g. 7-16] that the speaker individuality information is not equally distributed on the speech spectrum. Based on this assumption, a variety of methods have been developed to conveniently extract the most useful information from the speech signal for further modeling. The main findings of the reviewed literature are presented in the next section. However, most of these studies did not consider coded-decoded speech, present in the majority of today's speaker verification applications, or did not focus on the differences in performance given by clean and distorted speech.

Telecommunication networks have been deployed in the recent years at a rapid pace, allowing automatic speaker verification to be performed remotely after the transmission of speech signals. The drawback is that communication channels introduce various kinds of distortions that degrade the verification performance. Speaker recognition is further hampered if there exists mismatch between the distortion of training data and that of testing utterances.

Two main transmission bandwidths are available nowadays: one is the conventional narrowband (NB), which limits the signal frequency range to 300-3,400Hz and is implemented in the public switched telephone network (PSTN). The second one is wideband (WB), offering the enhanced range 50-7,000Hz and supported by Voice over IP (VoIP) applications among others. Even more extended is the range of super-wideband (SWB) transmissions, 50-14,000Hz, intended for high-quality videoconferencing. It has been shown that the extensions of the NB frequencies contribute to better intelligibility, perceived quality [1], human speaker identification [2], and automatic speaker verification [3]. For efficient transmission, a speech compression algorithm, or codec, is applied to the signals. The bit rate of digital speech representations is reduced after the coding process resulting in loss of signal quality in NB and also in WB [1].

In this paper, different to past studies addressing sub-band analyses, we employ speech segments that have been transmitted through different NB and WB codecs in a controlled manner. NIST speaker recognition evaluation data is not suitable for our experiments because it is mostly limited to NB and the utterances present an uncontrolled variety of distortions caused by different handsets and different modes of transmission. Our aim is to apply only certain controlled distortions to audio segments and to compare their influence on the speaker recognition performance. We attempt to understand the effects of channel filters and codecs on different frequency sub-bands, which we subject to a series of speaker verification experiments and whose speaker discrimination efficacy we evaluate using F-ratios [4, 5] computed. These outcomes are compared to those obtained from identical experiments employing clean speech. The performances offered by NB and by WB signals have not been compared before in terms of the usefulness of the band of frequencies added in WB transmissions. Another novelty is that we employ state-of-the-art i-vector based speaker verification systems [6] in our analyses. After the partition of the whole spectral domain, the verification experiments are performed on each sub-band independently.

2. Related work

Extensive research in the last decades showed evidence that speaker specific information is not equally distributed among the spectral sub-bands, that is, certain sub-bands present more discriminative power than others. Overall, past studies agree that the lower frequency region (below 1kHz) and the higher frequencies (above 3kHz) provide better recognition accuracy than the middle frequencies. This is attributed to the occurrence of different phoneme events. For instance, vowel formants convey speaker individuality [7], nasals present discriminative power in low and mid-high frequencies [8, 9], and other consonants in the upper part of the frequency spectrum, above 6kHz [9]. The most discriminative frequencies found in different studies are shown in Table 1, along with the databases employed, indicating whether the speech was clean or distorted and its bandwidth.

Ref.	Datasets (distortion, frequency range)	Findings: most discriminative sub-bands
[7]	TIMIT (clean, 0-8kHz)	Below 0.6kHz and above 2kHz
[10]	Local set of 20 males and 13 females (?,0.3-3.4kHz)	Below 0.6kHz and above 2kHz
[11]	NTT-Voice Recognition (clean, 0-8kHz)	0-2kHz and 6-8kHz
[12]	TIMIT (clean, 0-8kHz) and NTIMIT (narrowband, 0.3-3.4kHz)	Below 0.6kHz and above 3kHz
[13]	TIMIT, 5th dialect region (clean, 0-8kHz)	Below 1kHz and 3-4.5kHz
[14]	TIMIT (clean 0-4kHz)	0.05-0.25kHz for all phoneme classes
[15]	TIMIT, 7th dialect region and Helsinki corpora (ulaw, 0-5.5kHz)	Below 0.2kHz and 2.5-4kHz (TIMIT) and 2-3kHz (Helsinki)
[16]	BT Millar speech database (clean, 0.3-3.4kHz)	1-2.5kHz and 2.5-4kHz
[17]	NTT-Voice Recognition (clean, 0-8kHz)	0.05-0.3kHz, 4-5.5kHz and 6.5-7.8kHz
[8]	NIST SRE 2008 (µ-law, 0.3-3.4kHz).	Around 0.3kHz and above 2kHz
[18]	Accent of British English (clean, 0-11.025 kHz)	Below 0.77 kHz and 3.4-11.025kHz
[9]	RyongNam2006 (clean, 0-11.025kHz)	Below 0.3kHz, 4-5.5kHz, above 9kHz

Table 1: Type of data and main findings of the literature

To detect which frequencies convey speaker information the spectral domain is often partitioned into frequency sub-bands and their effectiveness for speaker recognition analyzed in different manners. For instance, Besacier and Bonastre [7] applied a speaker recognizer to each sub-band separately and then combined their outputs to compute the global decision for text-dependent speaker identification. Some years later, they proposed an on-line feature selection procedure based on their analysis of the most discriminative frequency sub-bands [12]. In [16], the cepstral parameters from different sub-band systems were recombined with sub-band weighting. Optimum band splitting and recombination strategies were addressed in [11]. The authors of [10] employed linear and mel scale filters to analyze sub-band discrimination power and developed a new feature warping function (between linear and mel) that provided optimal speaker identification results employing a relatively small speaker dataset (20 males and 13 females).

The speaker discrimination properties have been determined by means of Hidden Markov Model (HMM) experiments [16] or Gaussian Mixture Models (GMM) [12, 18], although a very popular approach, adopted in [8, 9, 13, 14, 15, 17], is to employ the F-ratio measure [4, 5], which accounts for the relation between the variance of features between speakers and the variance within a speaker. We chose to perform separate i-vector experiments on each sub-band to detect which frequencies enable better speaker verification and to contrast our results with an analysis of F-ratios employing the same clean and coded-decoded datasets, in NB and in WB.

Other investigations were concerned with the design of a custom filterbank as an alternative to the conventional mel-scaled filterbank to extract features that emphasize speaker-specific information. In [13] the sub-band weights were determined using F-ratios and vector ranking.criteria. This work was extended by Kinnunen [14] by adapting the weights of each sub-band depending on the phone detected in the input speech frame, that is, his proposed filterbank emphasized the discriminative power of particular phonemes. In [17], the authors designed sub-band filters with non-uniform bandwidth which was inverse proportional to the F-ratio calculated on each frequency sub-band, whereas the filterbank developed in [9] was based on an F-ratio study considering different phoneme classes. All of these studies showed that the features extracted with custom filterbanks outperformed Mel-frequency cepstral coefficients (MFCCs), which evidences that the latter might not be optimal for the task of speaker recognition. The work in [17] was extended in [8] for telephone speech (NB), demonstrating the superiority of Linear Frequency Cepstral Coefficients (LFCCs) over MFCCs for the nasal and non-nasal consonant regions. Also [19] showed the advantages of LFCCs over MFCCs in NB speech, and that they were accentuated for female speakers. Indeed, the higher resolution of the linear-spaced filters in the higher frequencies, where important speaker individuality is present according to these analyses, can capture more spectral detail and lead to better speaker recognition results compared to the mel-spaced filters. The authors of [17] mentioned that the frequencies outside the telephone bandwidth were more discriminative. In the sub-band analysis of coded-decoded speech presented in this paper we would like to assess the good performance of the higher and lower frequencies of WB signals that are filtered out in NB channels, and also which codec scheme leads to the best results in each bandwidth. From this analysis we could derive a custom filterbank that would be targeted to the corresponding kind of distortion. This is one of our future work plans.

The mentioned studies employed either clean data or coded-decoded data in NB only but no WB codecs were applied. Besides, no comparison between clean and distorted data was attempted. Only [12] detected a decrease of performance between TIMIT and NTIMIT (its NB version), which was attributed to handset, bandwidth filtering and telephone distortions, yet no further explanation was given. A number of other studies have addressed the effects of voice compression on the speaker verification performance but did not provide a sub-band analysis [20-24]. The earlier studies consider only NB coding [20, 21].

Another gap in the literature is the study of how super-wideband transmissions affect speaker recognition. It remains still unclear whether the frequencies beyond 11.025kHz are speaker discriminative and how this is altered if a SWB codec is applied. Our past work did not find advantages of SWB over WB regarding human speaker identification performance [2]. The last section of this paper is concerned with a preliminary discrimination study using F-ratios with SWB clean and codec-decoded speech.

The remainder of this paper is as follows. Section 3 details the speech codec algorithms applied to voice segments while Section 4 explains our sub-band analysis procedure. Sections 5 and 6 provide the results and the discussion, respectively. A preliminary super-wideband analysis is included in Section 7. Section 8 concludes this work.

3. Speech coding-decoding

In our analyses we consider speech transmitted through NB and WB channels and clean speech sampled at 8kHz and at 16kHz. Original databases of clean speech were downsampled, bandwidth-filtered, coded and decoded according to each kind of channel degradation and separate sub-band experiments were conducted employing each of the six created speech versions.

3.1. Audio material

As original data we selected only datasets with utterances which were recorded directly through microphones and not transmitted through communication channels so that we could control the degradations of the data. The data should have a sampling frequency of at least 16kHz, which would allow the study of WB codecs. Corpora meeting these requirements are: 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 in this analysis.

The Universal Background Model (UBM) and the total variability matrix $\boldsymbol{T}$ of the i-vector extractor [6] were trained with combined speech from the train partition of TIMIT and the other four datasets, totaling 670 speakers and with approximately 89h of speech. We refer to this combined dataset as *development data* in this paper. The test partition of TIMIT, containing 112 speakers, was set aside for the evaluation of our systems, and thus we refer to it as *evaluation data*. The effects of codec or bandwidth mismatch among background training, enrollment and test segments are not considered in our analyses, that is, the i-vector experiments (background training of the i-vector extractor and enrollment/test via cosine distance scoring) were conducted separately for each kind of distortion.

3.2. Speech codecs

All the speech (for development and for evaluation) was transmitted through the four communication channels listed in Table 2. Besides, also clean, unprocessed speech with sampling frequency of 8kHz (4kHz bandwidth) and of 16kHz (8kHz bandwidth) was considered in our experiments. Hence, we created 6 versions of the data.

The codecs of our study offer different speech quality, which was assessed in [1]. For the transmission through WB codecs the WB filter complying with International Telecommunication Union (ITU)-T Recommendation P.341 was applied to the speech sampled at 16kHz, thus band-limiting the signal to 50-7,000Hz. Differently, for NB transmissions the original speech was downsampled to 8kHz and band-limited to 300-3,400Hz by channel filtering according to ITU-T Recommendation G.712. After applying the WB and NB filters simulated code-decode processes were applied[1]. The operation of each codec is briefly described in next sub-sections along with the expected effects on the frequency sub-bands for speaker recognition. Further information can be found in ITU Recommendations and in European Telecommunications Standards Institute (ESTI) documents.

Bandwidth	Codec	Creator	Bit rate (kbps)	Algorithm
NB	G.711	ITU-T	64	A-law PCM
	AMR-NB	3GPP	12.2	ACELP
WB	G.722	ITU-T	64	SB-ADPCM
	AMR-WB	3GPP	12.65	ACELP

Table 2: Narrowband and wideband telephone channels through which the original clean speech was transmitted.

3.2.1. *G.711*

This NB codec operates at a bit rate of 64kbps, which corresponds to 8kHz sampling rate and 8 bits per sample. Its encoding schemes can be µ-law pulse code modulation (PCM) (in use in North America) or A-law PCM (in use in the other source countries for our data). The difference between them is the method to sample the analog signal (both in a logarithmic way). G.711 encoding/decoding requires little processing (it is a low complexity codec) and produces high quality speech, but consumes more bandwidth than other NB codecs, for instance the AMR-NB. This trade-off between bandwidth, processing power required for the encoding/decoding function, and voice quality is common among the different compression algorithms. Its main applications are digital telephony (it is widely in use in PSTN) and VoIP.

3.2.2. *AMR-NB*

The Adaptive Multi-Rate (AMR) family of codecs was designed for GSM and UMTS cellular networks. AMR can be further categorized as AMR-NB and AMR-WB, depending on the bandwidth employed. These codecs are frequently used in VoIP and wireless telephony. The AMR-NB encodes the 13-bit linear PCM signal at eight different bit rates in the range from 4.75 kb/s to 12.2kb/s and bases its coding scheme on Algebraic Code Excited Linear Prediction (ACELP).

The parameters of the ACELP model are Linear Prediction (LP) filter coefficients, transmitted in the form of Line Spectral Pairs (LSPs), and fixed and adaptive codebook indices and gains, which encode the excitation (residual) signal. After the transmission of these parameters, at the decoder, the waveform is synthesized by filtering the reconstructed excitation signal through the LP synthesis filter. The LP coefficients represent the speech spectrum. ACELP is a more complex algorithm than PCM, hence it is expected that this codec, operating at a lower bitrate than the G.711, would introduce more distortions into the signal.

3.2.3. *G.722*

This ITU codec can operate at 48, 56, and 64kbps, although its main mode is 64kbps. It is used in the Integrated Services Digital Network (ISDN) and in VoIP applications. It applies the Adaptive Differential PCM (ADPCM) algorithm to encode two separate sub-bands (0-4kHz and 4-8kHz). 48kbps are dedicated to the lower sub-band, where most of the voice energy is concentrated, while the remaining 16kbps are dedicated to the higher sub-band. This difference in allocated bandwidth may cause a greater distortion of the high frequency components.

[1] The source code for the ITU-T software tools is available at http://www.itu.int/rec/T-REC-G.191/_page.print

3.2.4. AMR-WB

As AMR-NB, it is mainly used for speech compression in mobile telephony. It belongs to the same family of codecs. The operation bit rate chosen in our experiments was 12.65 kbps. This higher compression implies that it requires more processing cycles than G.722, i.e. more complexity, and it is likely to degrade the speech to a greater extent in comparison with the other WB codec. Voice Activity Detection and Comfort Noise Generation algorithms are adopted to decrease the bit rate, this is also expected to introduce degradations.

Its coding algorithm is ACELP, as for AMR-NB. In this case, two frequency bands, 50–6,400Hz and 6,400–7,000Hz, are coded separately. The parameters of the encoder are: the Immittance Spectral Pair (ISP) vector built from the LP parameters, fractional pitch lags, Long Term Prediction (LTP) filtering parameters, innovative codevectors, and sets of vector quantized pitch and innovative gains [22]. The higher frequency band (6,400–7,000Hz) is reconstructed in the decoder using the parameters of the lower band and a random excitation when the codec operates at a bit rate lower than 23.85kbps. Hence, the higher frequencies might be more distorted than those of the lower band.

4. Spectral sub-band analysis

We conducted a series of independent i-vector experiments considering feature vectors with cepstral coefficients (LFCCs) derived from each of the sub-bands. A linear filterbank of 32 triangular filters with 50% overlap was employed to extract the cepstral coefficients. 28 overlapping groups of 5 filters were considered: the S-th sub-band consisted of the outputs of filters S to S+4, where S = 1,…,28. The spectrum was thus partitioned according to the distribution of the 32 filters, being the low cutoff frequency of the first filter at 0Hz, 0Hz, 300Hz, and 50Hz, and the high cutoff frequency of the 32th filter at 4kHz, 8kHz, 3.4kHz and 7kHz, for clean 4kHz bandwidth, clean 8kHz bandwidth, NB coded-decoded, and WB coded-decoded signals, respectively.

After energy-based voice activity detection (VAD), four LFCCs were extracted from each group of filters with a 25ms Hamming window with 10ms frame shift. These coefficients constituted the feature vector, discarding the 0th coefficient and the log energy. A total of 168 i-vector experiments were performed with these features, resulting from 28 sub-bands x 6 versions of our data.

As baseline, we also performed a separated set of 6 experiments, one with data of each distortion, in which the whole spectrum was considered, limited by the frequencies of the NB and WB filters as described above. 32 linearly-spaced triangular filters were employed and feature vectors of 60 components computed: the first 20 LFCCs excluding the 0th coefficient and the log energy feature, extracted using a 25ms Hamming window with 10ms frame shift, and the corresponding delta and delta^2 coefficients.

We trained the 168+6 i-vector extractors separately, employing different versions of the development data accordingly. Hence, the UBM and the total variability matrix $\boldsymbol{T}$ were estimated from either clean or coded-decoded development data in NB or in WB. The UBMs were built with 1024 Gaussian components 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. Of the 10 utterances per speaker in our evaluation data, 5 were concatenated for speaker enrollment and 5 were used for testing. Confronting 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 112.

Applying the Probabilistic Linear Discriminant Analysis (PLDA) compensation technique did not improve the performance given by cosine distance scoring. The PLDA model, estimated from the same development data as for the UBM and the $\boldsymbol{T}$ matrix, was not adequate for the compensation in our case, where feature vectors of only four components are employed. Further analyses to determine optimal training data and model parameters are needed. The PLDA compensation did benefit the performance in the case of the experiments on the full band, although we omitted these results in this paper.

We also used the F-ratio [4] to measure the discriminative power of different regions of the spectra of our evaluation data. The same filterbank with 32 filters was applied to compute the spectral energy around the central frequency of each filter. They had an uniform bandwidth of 242Hz, 485Hz, 188Hz, and 421Hz for clean 4kHz bandwidth, clean 8kHz bandwidth, NB-processed, and WB-processed signals, respectively.

Given a sub-band k, the F-ratio was computed as:

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

where $x_i^j(k)$is the energy in the k-th sub-band of the j-th speech frame of the i-th speaker with k = 1,…32, j = 1,…,N$_i$, 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:

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

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

The higher the F-ratio, the more speaker-specific information is conveyed by the spectral sub-band. However, the F-Ratio measure also presents some limitations [25]. If the classes – in our case, the speech from different speakers – have the same means or are multimodal the F-ratio discrimination power is weak and can be misleading. Since our data, pooled from all phoneme classes, is a mixture distribution with multiple modes, we do not expect that the F-ratio to reliably indicate the speaker-discriminative regions of the spectra, although we use it thoughtfully in this paper to get an idea of the location of discriminative information.

The F-ratios can also be derived from the divergence, a distance measure based on information theory, assuming the data is normally distributed and equal between-talkers covariance matrices [25].

5. Results

The results of our experiments show the speaker-discriminative properties of each group of frequency sub-bands. The graphs of Figures 1 and 2 show the performance of our i-vector experiments in terms of the Half Total Error Rate (HTER), for clean and degraded signals sampled at 8kHz and 16kHz, respectively. This performance measure assumes equal prior probabilities and detection error costs. The frequency of each of the dots plotted corresponds to the central frequency of the sub-band considered, e.g. for the first sub-band of NB speech (300-864Hz) the central frequency is (864-300)/2+300=582Hz. The extended range of frequencies of the non-filtered (clean) speech can be seen in the graphs.

The superior performance of clean speech with respect to coded-decoded speech can be observed, as well as the consistent better performance of WB codecs compared to NB-transmitted speech. This is in concordance with the overall HTERs of our 6 baseline experiments when the frequency bands were not separated, given in Table 3. A relative decrease of the HTER of 77% has been found comparing clean 4kHz and 8kHz data. The frequencies beyond 4kHz account for this verification error reduction.

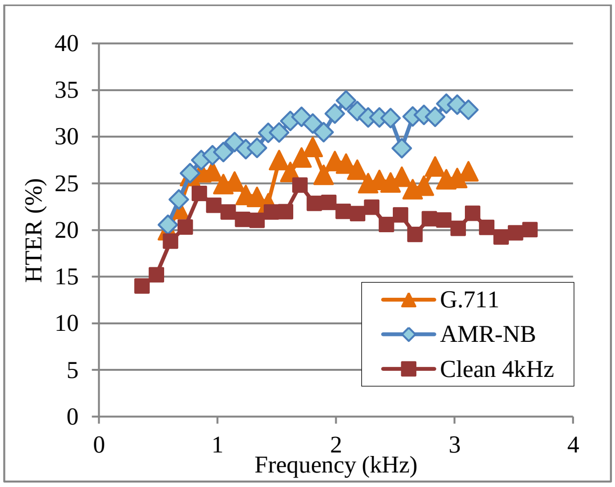

Figure 1: HTER (%) for NB speech and clean (uncoded) speech of 4kHz bandwidth. Feature vectors of 4 LFCCs.

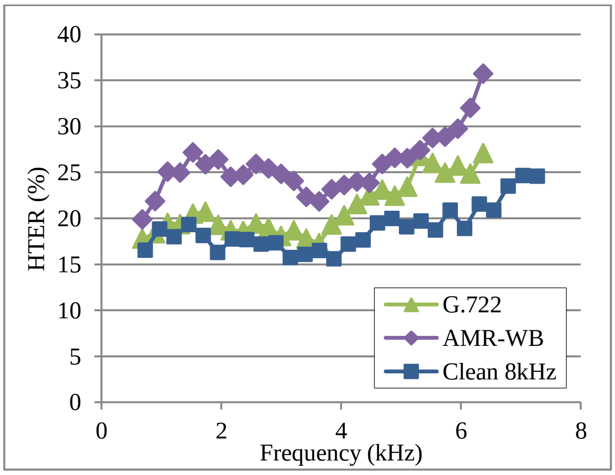

Figure 2: HTER (%) for WB speech and clean (uncoded) speech of 8kHz bandwidth. Feature vectors of 4 LFCCs.

Distortion	HTER (%) TIMIT
G.711 (NB) AMR-NB (NB) Uncoded 4kHz	6.29 8.56 3.75
G.722 (WB) AMR-WB (WB) Uncoded 8kHz	2.03 2.89 0.88

Table 3: HTER (%) considering the whole spectrum. Independent experiments for each distortion.

The plots of the popular F-ratio measure computed on our evaluation data are shown in Figure 3, for clean and degraded signals sampled at 8kHz and 16kHz. The frequency on the x axis corresponds to the central frequency of each filter in the filterbank. It can be seen that the F-ratio values corresponding to clean data are higher than for transmitted data. However, only the comparisons of relative values between different frequencies are relevant in our case, since the voice transmission also involved a level-equalization process, characteristic of telephone channels, that “normalizes” the signal energy: the speech was level-equalized 26dB below the overload point of a 16-bit digital system using the voltmeter algorithm of the ITU-T Recommendation P.56.

The frequency regions of clean speech exhibiting higher speaker discriminative ability are, as expected and in concordance with the literature, below 0.7kHz and between 2 and 4 kHz [12, 13, 15]. The behavior found with coded-decoded speech and the relation with our i-vector sub-band experiments are discussed in next section.

6. Discussion

The advantages of WB- over NB-transmitted speech are obvious comparing Figures 1 and 2. The low (50-300Hz) and the high (3,4-7kHz) frequencies bands carried by the enhanced bandwidth contribute to a better speaker verification performance in comparison to NB, as also manifested in the results of Table 3.

From the clean data in Figure 1 it can be concluded that the frequencies filtered out in the NB channels provide better speaker verification accuracy. The performance of the first two sub-bands (0-727Hz and 121-848Hz) is especially good compared to the rest possibly due to the presence of glottal

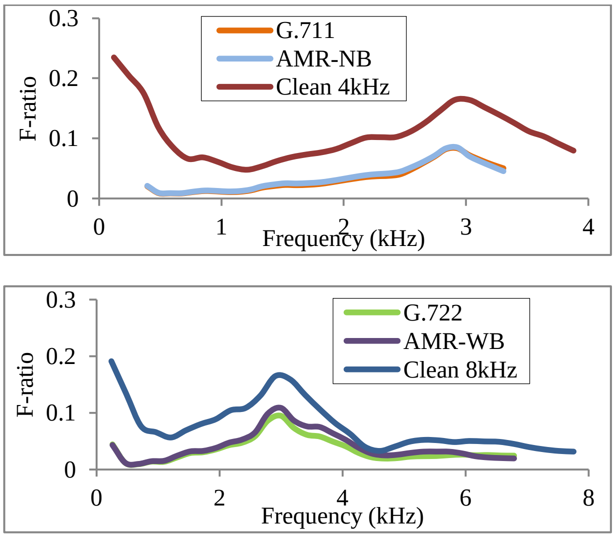

Figure 3: F-ratio values for NB, WB, and clean speech. 112 males of the TIMIT test partition. The G.711 and AMR-NB lines are almost indistinguishable.

information and the first formant in this frequency region [13, 17]. Comparing the performances of clean 4kHz bandwidth speech to that of clean 8kHz bandwidth speech below 4kHz, it can be appreciated that the latter leads to generally lower HTERs because the range of frequencies considered in each sub-band is doubled, that is, an experiment on a wider sub-band results in better accuracy. The lower performance of coded-decoded speech compared to clean speech in individual sub-bands is partially due to this fact.

Considering NB transmissions (Figure1), the G.711 codec exhibits a behavior closer to that of clean speech while the performance is more degraded by the AMR-NB codec, noticeably for frequencies beyond 1kHz. The performances of both codecs decrease from 0.3 to 1kHz (as for clean speech). In the frequencies 1 – 1.5kHz, the performances with clean speech and with G.711 improve while that of AMR-NB continues decreasing. This difference can be justified by the mode of operation of the codecs. As mentioned in Section 3, the low complexity of the G.711 results in higher quality audio compared to the more efficient AMR-NB [1]. It seems that the ACELP algorithm at the low bit rate of 12.2kbps induces some loss of speaker individuality in the signal synthesized from the transmitted LP coefficients and residual, and thus hampers automatic speaker recognition, in contrast to A-law PCM coding.

In the case of WB speech (Figure 2), the landline codec G.722 shows better speaker verification performance compared to the wireless AMR-WB in every frequency sub-band. This difference, as in the NB case, can be explained by the difference in coding algorithms and by the lower bitrate of AMR-WB. The precision of the description of the low sub-band by the G.722 (0-4kHz) seems to be more accurate than that of the high sub-band (4-8kHz), where the speaker verification performance is worse in comparison to clean speech. The AMR-WB greatly degrades the performance for the frequencies beyond 6kHz, since the high-band speech signal was reconstructed using a random excitation, in contrast to the lower band, for which the residual was transmitted. However, the discrimination of the frequency band 3-4kHz, which conveys more speaker-specific content than other bands, is enhanced by this codec since the difference between the error of transmitted and the error of clean speech is lower than for other frequency ranges. This fact contributes to a better overall performance, which is close to that of the G.722 codec, as can be seen in Table 3. It has been found in [22] that improved verification results can be obtained from feature vectors with the AMR-WB encoded parameters than with MFCCs from the decoded speech. This implies that most of the distortion is caused by the signal reconstruction in the decoding process.

The introduced frequency distortions are consistent with our F-ratio analysis. Considering NB, the distorted and the clean speech show a similar behavior. In WB, for the G.722, the discriminative power of the frequencies beyond 4.5kHz does not increase greatly from its value at 4.5kHz, as occurs for clean speech of 8kHz bandwidth. For the AMR-WB, the F-ratio values at the frequencies of the separately encoded high-band (6.4-7kHz) tend to decrease. This band is decoded employing a random excitation in the 12.65kbps operation mode, which presumably originates the higher distortion in comparison to other sub-bands.

The speaker verification accuracy found for clean data above 4kHz is not as high as in the bands 0-700Hz and 2-4kHz, as opposed to other results in the literature [7, 12, 17, 18]. Their databases, however, comprised male and female speech and the reported results refer to the mixed set of speakers. It is possible that the female speech carries speaker-discriminative information beyond 4kHz and that this affects the overall results reported in these studies. Differently, our experiments involved only male speakers, and provided results consistent with [13]. The authors of that investigation found little speaker individuality conveyed beyond 4kHz considering only clean male speech of the TIMIT database. For female speech, the F-ratios did not decrease dramatically beyond 4kHz and their female speaker recognition accuracy based on Vector Ranking (VR) was as high in the band 6-8kHz as in the band around 3kHz, which did not occur for male speech. This suggests that female voices carry important speaker-specific high frequency content that may enhance the verification performance to a greater extent than for male voices when signals of bandwidth greater than 4kHz are considered.

7. Preliminary super-wideband analysis

While the frequency regions below 8kHz contain most of the speech energy and their importance has been extensively assessed for speaker recognition, very little attention has been paid to the role of frequency bands above this frequency.

The studies in [9, 18], examining discriminative frequencies of clean speech up to a bandwidth of 11.025kHz, found important speaker-specific content conveyed beyond 8kHz. Extending their studied bandwidth to 16kHz and considering channel transmissions, we employ the newly compiled database of Australian English, (AusTalk) [26], with audio sampled at 44.1kHz. The audio segments were downsampled to 32kHz and transmitted through a SWB channel. Only 21 male speakers were considered in this preliminary analysis. The SWB channel filter limits the bandwidth to 50-14,000Hz (14KBP in ITU-T Recommendation G.191), and the codec applied was G.722.1C, an ITU-T codec at 48kbps. The speech was also transmitted through the codecs described in Section 3.

Figure 4 shows the F-ratio values calculated as indicated in Section 4 from the transmitted speech. They present some differences with respect to the other database considered in this paper. Looking at the clean 8kHz plot, the most discriminative frequencies are found below 0.7kHz (as before) and above 2.5kHz, with an important region between 4.3 and 5.8kHz, approximately. This discriminative range might be originated by the piriform fossa, which is part of the pharynx and causes characteristic spectral structures [17]. This range is not noticeable in the analysis of the TIMIT database, probably due to the different male speaker populations. Interestingly, high F-ratio values can be found at 9kHz and above 11kHz, possibly related to consonants with high-frequency energy such as fricatives. It can also be seen that the G.722.1C codec exhibits a behavior close to that of clean speech, introducing little distortion.

The regions with speaker-specific content in the upper part of the spectrum are expected to improve speaker verification over WB. However, we could not perform i-vector experiments with SWB data since all the speech datasets we have available are either too small, or sampled at a rate lower than 32kHz (required for SWB transmissions), or present already some sort of distortion. However, we are conducting more research in this direction employing the AusTalk database [26]. We foresee that SWB transmission

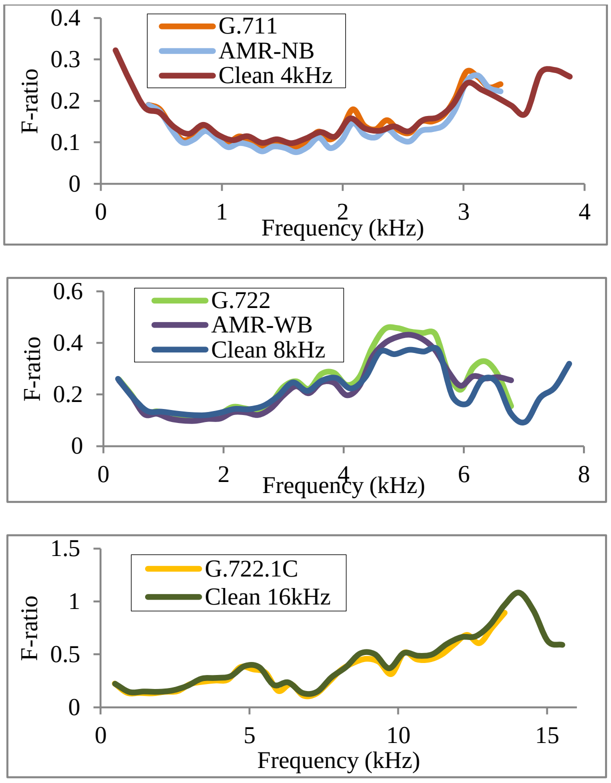

Figure 4: F-ratio values for NB, WB, SWB, and clean speech. 21 males of the AusTalk dataset [26]. The G.711 and AMR-NB lines are almost indistinguishable.

channels will be widely deployed in the future and intend to show that speaker verification can be considered as an additional criterion when judging the benefits of the extended bandwidths in comparison to the traditional NB.

8. Conclusions

In the present work we have examined the outputs of i-vector speaker verification experiments performed on different spectral sub-bands when the speech was degraded by NB and WB transmissions. In agreement with the results in the literature, the most discriminative frequency regions of our clean speech dataset are below 0.7kHz and between 2 and 4kHz. When this dataset was transmitted through telephone channels, the landline codecs offered better results than the more complex mobile telephony codecs but performed slightly worse than clean speech. Each of the codecs caused various effects on the frequency bands, attributable to the different coding algorithms. The wireless codec AMR-NB provides significantly lower recognition results than G.711 and than clean speech in the band 1-1.5kHz. G.722 generates some loss of speaker individuality content in the band 4-8kHz. AMR-WB emphasizes the discrimination of the region 3-4kHz but degrades the verification accuracy beyond 6kHz to a greater extent in comparison to other frequencies. Clearly, the frequencies beyond the NB telephone bandwidth improve speaker recognition. These effects are consistent with the F-ratio measures on degraded speech, although the latter are weak indicators of speaker-discriminative regions in our case [25].

Our preliminary F-ratio analysis on clean and transmitted SWB signals reveals that important speaker-specific content is found in the region 8-14kHz. This finding is encouraging and indicates that speaker recognition may benefit from the frequencies beyond WB. Due to the limitations of the F-ratios, it would be necessary to conduct such an experiment to verify the usefulness of an even more extended transmission bandwidth.

In future work we would like to conduct similar experiments with different datasets including female speech and, if available, with an extended signal bandwidth. We will also examine the design of a custom filterbank that enhances the most important frequencies of transmitted speech.

9. 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," *Acoustics, Speech and Signal Processing*, pp. 7775-7779, 2013.

[3] Fernández Gallardo, L., Wagner M. and Möller, S., "Analysis of Automatic Speaker Verification Performance over Different Narrowband and Wideband Telephone Channels," *Australasian International Conference on Speech Science and Technology*, pp. 157-160, 2012.

[4] 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.

[5] Wagner, M., "The application of a learning technique for the identification of speaker characteristics in continuous speech," *PhD thesis, Australian National University*, 1978.

[6] 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. 4, pp. 788-798, 2010.

[7] Besacier, L., Bonastre, J. F., "Subband approach for automatic speaker recognition: optimal division of the frequency domain," *Audio and Video based Biometric Person Authentication*, pp.195−202, 1997.

[8] Lei, H. and Lopez-Gonzalo, E., "Mel, Linear, and Antimel Frequency Cepstral Coefficients in Broad Phonetic Regions for Telephone Speaker Recognition," *Interspeech*, pp. 2323-2326, 2009.

[9] 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*, pp.1-4, 2012.

[10] Auckenthaler R. and Mason J. S., "Equalizing Sub-band Error Rates in Speaker Recognition," *Eurospeech*, pp. 2303-2306, 1997.

[11] Yoshida, K., Takagi, K. and Ozeki, K., "Speaker Identification Using Subband HMMs", *Eurospeech*, pp. 1019-1022, 1999.