Analysis of Automatic Speaker Verification Performance over Different Narrowband and Wideband Telephone Channels

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

Faculty of Information Sciences and Engineering, University of Canberra, Australia
Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany
(laura.fernandezgallardo|michael.wagner)@canberra.edu.au, sebastian.moeller@telekom.de

Abstract

Current speaker recognition applications involve the authentication of users by their voices for access to restricted information and privileges. The speech signal is often transmitted to the recognizer through communication channels presenting different transmission characteristics. The aim of this paper is to study the effects of speech bandwidth and coding schemes on speaker verification. We compared the performance of a Gaussian Mixture Model-Universal Background Model (GMM-UBM) classifier in two different conditions: in one condition, the system was trained and tested with speech processed using wideband codecs, and in the other with speech processed using narrowband codecs. Our results show that the verification task improves significantly when the system is trained and tested with speech transmitted through wideband channels.

Index Terms: automatic speaker verification, speech coding, speech bandwidth

1. Introduction

The automatic detection of people's identity from their voices, without requiring the intervention of humans, has attracted the attention of researchers and engineers in the last decades. The typical task of the machine is speaker verification (SV), that is, the binary decision of whether or not one input voice corresponds to a claimed identity. The great majority of SV applications require the speech signal to be transmitted through communication channels like Voice over Internet Protocol (VoIP), mobile telephony (in GSM or UMTS networks), or the Public Switched Communication Network (PSTN). The standardized transmission bandwidth of these channels is either narrowband (NB), containing the frequency range 300 – 3,400Hz, or wideband (WB), offering the extended range of frequencies 50 – 7,000Hz. In order to reach a sufficiently high data compression for these channels, a variety of codecs is used which work at different bitrates.

In contrast to NB speech, the WB signal contains additional frequencies which potentially augment the information about the speaker's identity. The low frequencies include the fundamental and sometimes the first formant of individual's voice while the high frequencies carry higher formants and other important speaker characteristics such as those from a nasal or a breathy voice. Thus, WB speech could enable the system to discriminate better among talkers. It has been demonstrated that the identification performance improves considering the frequencies below 1000Hz and between 3000Hz and 4500Hz [1], which are mostly suppressed in NB channels However, there is still little knowledge of how well WB coding schemes permit reliable speaker recognition and how they compare to NB codecs. The aim of this work is to evaluate the performance of automatic SV over different channels and to determine to which extent the WB transmission - including the relevant coding scheme - benefits this performance compared to NB transmissions.

In our experiments, we transmitted the speech signals through two simulated WB and three simulated NB communication systems, creating five versions of our data set. We then implemented a Gaussian Mixture Model-Universal-Background Model (GMM-UBM) classifier [2], which was first trained with the WB speech and secondly with the NB speech. These two systems were evaluated with test segments of the same signal bandwidth, in order to compare the verification accuracy for a WB system and that for a NB system.

It is indeed a challenge for speaker recognition to overcome the effects of mismatched conditions between training (when the user enrolls into the system) and testing (when the user is being authenticated). Employing different handsets and environmental noise often degrade the system performance severely. In addition, network characteristics in terms of codecs employed, packet loss, etc. degrade recognition performance. In this paper we consider only one of the latter factors, namely the speech codec used, which is associated with a defined audio bandwidth (NB or WB) and a bit rate. Our results are also affected by inter-session variability, since the training data and testing data are obtained from different recording sessions. Nevertheless, because WB signals contain relevant information about talkers, we can expect that an improved performance will be offered by the system trained with WB speech over the system trained with NB speech.

This paper is organized as follows. Related work is reviewed in Section 2. Section 3 provides an insight of the communication systems used to transmit the speech while Section 4 describes the classifier used in our experiments. Results are detailed in Section 5. The conclusions of this work and future research directions are presented in Section 6.

2. Related work

The effect of coded speech on speaker recognition performance has been extensively studied in the last two decades, motivated by the deployment of worldwide communication networks and by the usual client-server architecture of speaker recognition systems. Different classifiers have been employed in order to demonstrate the degraded performance when the speech signal is affected by channel characteristics. The most widely used classifiers are based on Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM).

A HMM-based system was employed to examine the SV performance coding speech with GSM in [3] and later to study the effects of VoIP codecs in [4]. Jokić et al. [5] have recently tested an HTK-based recognizer on a speaker identification task transmitting voice through varied NB systems and one WB system (using the codec G.722, also investigated in our work). Their results, employing a small dataset of 10 speakers, showed no significant identification improvement when the WB codec was used. In this paper, we would like to extend the analysis to other NB and WB codecs, and focus on a SV task which is more representative of real-life applications. Our classifiers are based on the state-of-the-art GMM-UBM instead of HMMs.

The GMM-UBM system offers the best performance for text-independent SV and has also been tested to examine how communication channels affect its performance [6], [7]. However, only limited NB telephone channels, offered by the still predominant PSTN, were considered in most analyses. Pradhan and Prasanna [7], using a GMM-UBM system, studied the effect of bandwidth and demonstrated the significance of WB speech in different environmental conditions. Even for the most favorable conditions for their system (matched headphones and clean speech), WB offers improvements in SV over NB. However, they did not examine the effects of speech coding. A different classifier, based on a GMM-SVM system, was developed by Janicki and Staroszczyk [8]. They stated that their system outperformed a GMM-UBM baseline but the number of Gaussian mixtures was required to be higher (They augmented it from 16 to 256). Two of the codecs they investigated are also used in this paper: GSM-EFR and G.711.

The mentioned studies agreed that speaker classification accuracy decreases with the degree of mismatch between training and testing conditions, or if a low quality codec was employed for speech transmission. Hence, we could expect that our first system, trained with WB speech, would lead to better performance results in comparison to our system trained with NB signals.

Multiple methods and techniques have been proposed to reduce the degradations due to coded speech, like the combination of prosodic and acoustic features [9], handset compensation from coded speech [10], or training the speaker models with the coded-decoded signal [5]. This latter approach reduces the degree of mismatch considerably and can be used in applications if the channel configurations are known beforehand. Differently from these studies, we do not attempt to improve the accuracy of our classifier but to compare the performance over communication systems with different properties. The same handset type was used to record speakers of our database and the background noise is minimal. Only inter-session variability and the coding schemes used for the training and testing segments are sources of mismatch in our experiments.

3. Transmission channels

In our experiments, we evaluated the speaker verification performance to authenticate users when their voice utterances were transmitted through WB and NB communication channels. WB and NB codecs were applied to the utterances for the evaluation of our systems. We simulated the five channels listed in Table 1. They consisted of a bandwidth filter and a speech compression scheme with a particular bit rate. To simulate the NB channels, the speech signal was first downsampled to 8kHz and then filtered according to the standard implementation of the International Telecommunication Union (ITU)-T Recommendation G.712. Differently, for WB channels, the signal was downsampled to 16kHz and filtered complying with ITU-T Recommendation P.341.

The codecs G.711 and G.722 were standardized by the ITU for use in digital telephony. Adaptive Multi-Rate (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 NB codec Global System for Mobile Communications (GSM)-Enhanced Full Rate (EFR) was also included in our study, as it is the standard codec for cellular telephony in Europe. We believe that the five codecs used represent a good selection of NB and WB compression schemes available today and in the near future.

Bandwidth	**Codec**	**Bit rate (kbps)**	**Coding type**
Narrowband	G. 711	64	A-law companded pulse code modulation (PCM)
	AMR-NB	4.75	Algebraic Code Excited Linear Prediction (ACELP)
	GSM-EFR	12.2	
Wideband	AMR-WB	23.05	
	G.722	64	Sub-band adaptive differential pulse code modulation (SB-ADPCM)

Table 1. *Five coding schemes studied in our work.*

4. Implementation of the classifier

We describe here our two speaker verification systems which are based on GMM-UBM. This classifier was chosen because it provides state-of-the-art performance and its simplicity was sufficient for us to compare the effects of speech coding. In order to create the speaker models, two versions of the data were used for training and testing the WB system (processed with the codecs G.722 and AMR-WB, respectively) and three versions were used for training and testing the NB system (processed with the codecs G.711, AMR-NB and GSM-EFR). The data to construct the gender-dependent background models (UBMs) were obtained from a set of 108 (54 males and 54 females) speakers from the ANDOSL database [11]. Each of the speakers uttered 200 phonetically balanced sentences with an average duration of 5s. The first 13 Mel-Frequency Cepstral Coefficients (MFCC) were extracted from the WB or NB speech signals as the feature vector for building the models. For the parameterization, the frame length used was 20ms with a frame shift of 10ms. The UBMs, one for males and one for females and each with 1024 Gaussian mixtures, represent the competing alternative speakers.

The AusTalk database [12], containing three session recordings, was used to train and test the systems. 32 (16 male and 16 female) client models were derived for each system by adapting the parameters of the UBM using training data from each speaker by means of *maximum a posteriori* (MAP) adaptation [2].

The training data were obtained from the first and second recording sessions of each speaker. The contents of the

utterances employed were words, digits and sentences, totalling about 11, 1.5, and 5 minutes respectively for each speaker. As testing data we employed word utterances from their third session (100 words from each speaker), processed with the different codecs. We estimated the performance of the speaker verification task for each communication system separately. The experiments were run in Matlab, version 7.13 (R2011b).

5. Results and discussion

To test our systems we computed the log-likelihood of every test segment being produced by the speakers from whom a model was built. Male and female utterances were tested separately against the models of speakers of the same gender. For each communication system the test set consisted of 3,200 client accesses and 48,000 impostor accesses.

The performance of our systems is shown by means of Detection Error Tradeoff (DET) curves. Each curve in Figure 1 represents the corresponding system performance when the speech was transmitted through one communication channel. The speech transmitted using the WB codecs G.722 and AMR-WB was used to test our WB system while the utterances processed with the codecs G.711, GSM-EFR and AMR-NB were used to test our NB system. The WB system offers considerably better performance, as all the DET curves corresponding to the WB codecs are closer to the origin.

In addition to the DET curves, we present measures in Table 2 to summarize the performance of the systems and permit a comparison. The systems have been evaluated according to the detection cost function (DCF)

$$DCF = C_{miss}P_{miss}P_{target} + C_{FA}P_{FA}(1-P_{target}) \quad (1)$$

Communication System	EER (%)	HTER (%)	95% CI (%)
AMR-WB (WB)	12.03	11.83	0.53
G.722 (WB)	12.47	12.28	0.55
G.711 (NB)	16.40	16.12	0.61
GSM-EFR (NB)	16.56	16.45	0.58
AMR-NB (NB)	18.53	18.16	0.65

Table 2. *EERs, HTERs and 95% confidence intervals around HTERs for each evaluation.*

where C_{miss} and C_{FA} are the costs of false rejections and false alarms, respectively, and P_{target} the *a priori* probability that a trial is a target trial. P_{miss} and P_{FA} are the detection error probabilities, plotted in Figure 1 for each possible operating point of the different systems. In this work, we set $C_{miss} = C_{FA} = 1$ and $P_{target} = 0.5$, and choose the minimum detection cost point (min_DFC) as the operating point of the system for the comparisons. In this case, min_DFC corresponds to the *a posteriori* Half Total Error Rate (HTER) as defined in [13].

In order to assess the significance of the differences between the evaluated codecs we applied the HTER significance test [13]. This test is suitable for comparing the performance of our classifier on the different data sets, assuming independent distributions of P_{miss} and P_{FA}, and is based on the confidence intervals around HTER, or in our case, around min_DCF. The test shows that the difference in performance of our two verification systems is statistically significant with 99% of confidence. Therefore training and testing the system with WB signals provides significantly

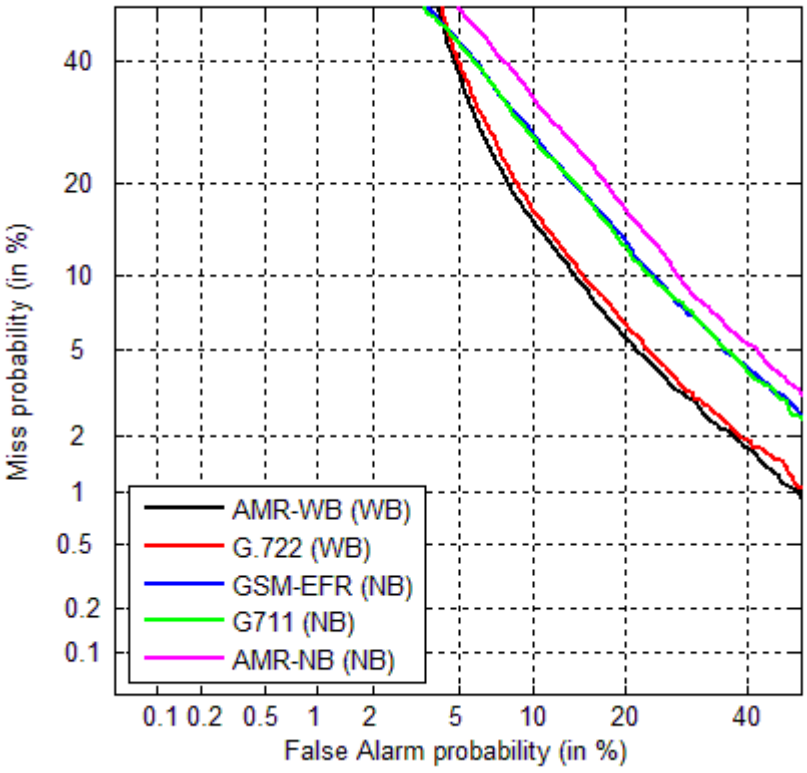

Figure 1. *DET performance of the systems trained and tested with speech transmitted through WB or through NB communication channels.*

better verification performance compared to the system trained and tested with speech transmitted through NB channels. With regard to the individual codecs, no significant difference was found when testing our WB system on speech processed with AMR-WB or with G.722. Likewise, applying the codecs G.711 and GSM-EFR results in statistically similar accuracy for the NB system. The AMR-NB codec, however, causes significantly worse verification rates than the other two NB codecs.

Like Pradhan and Prasanna [7], we show that the accuracy of the verification system increases when WB speech is used for training and testing. This reveals the presence of speaker features in WB signals, necessary for a more reliable speaker authentication.

6. Conclusion and Future Work

In our experiments we have demonstrated that the speaker verification improvement is significant when WB instead of NB signals are employed to train and test the classifier. This evidences that important information about the speaker identity is conveyed through those lower and higher frequencies that are filtered out in NB channels. Thus, WB transmissions offer more reliable automatic speaker verification, assuming the use of the same kind of handsets for the recordings and absence of noise. Similarly, human speaker identification was also shown to improve significantly when moving from NB to WB channels [14].

Nowadays the existing technology does no longer bound transmissions to NB and there is a growing interest in expanding WB communications. This motivates us to further study the effects of other channel characteristics, such as packet loss probability, in conjunction with NB and WB channels. Moreover, future work will concentrate on analyzing the effects of these channel artifacts on other speaker characterization tasks (estimation of gender, age and talker's emotions).

7. Acknowledgements

The authors wish to thank David Vandyke for his assistance with the software used to conduct this research.

8. References

[1] Orman, Ö.D. and Arslan, L.M., "Frequency Analysis of Speaker Identification," in Proc. of Speaker Odyssey: The Speaker Recognition Workshop, pp. 219–222, 2001.

[2] Reynolds, D.A., Quatieri, T.F. and Dunn, R.B., "Speaker Verification Using Adapted Gaussian Mixture Models," Journal of Digital Signal Processing, vol. 10, no. 1-3, pp. 19-41, 2000.

[3] Kuitert, M. and Boves, L., "Speaker Verification with GSM coded Telephone Speech," in Proc. of Eurospeech97, vol. 2, pp. 975-978, 1997.

[4] Fakhr, W., Abdelsalam, A. and Hamdy, N., "Enhancement of Mismatched Conditions in Speaker Recognition for Multimedia Applications," in Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 377-80, 2004.

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

[6] Quatieri, T. F., Singer, E., Dunn, R. B., Reynolds, D. A. and Campbell, J. P., "Speaker and Language Recognition Using Speech Codec Parameters," in Proc. of Eurospeech99, vol. 2, pp. 787-790, 1999.

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

[8] Janicki, A. and Staroszczyk, T., "Speaker Recognition from Coded Speech Using Support Vector Machines," in Proc. of Text, Speech and Dialogue, pp. 291-298, 2011.

[9] Chen, S.H. and Wang, H.C., "Improvement of Speaker Recognition by Combining Residual and Prosodic Features with Acoustic Features," in Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 93-96, 2004.

[10] Yu, E., Mak, M.W. and Kung, S.Y., "Speaker Verification from Coded Telephone Speech Using Stochastic Feature Transformation and Handset Identification," in Proc. of the 3rd IEEE Pacific-Rim Conference on Multimedia, pp. 598-606, 2002.

[11] Vonwiller, J., Rogers, I., Cleirigh, C. and Lewis, W., "Speaker and Material Selection for the Australian National Database of Spoken Language," Journal of Quantitative Linguistics, vol.2, pp. 177-211, 1996.

[12] Burnham, D., Estival, D., Fazio, S., Cox, F., Dale, R., Viethen, J.,Cassidy, S., Epps, J., Togneri, R., Kinoshita, Y., Göcke, R., Arciuli, J,. Onslow, M., Lewis, T., Butcher, A., Hajek, J. and Wagner M., "Building an Audio-Visual Corpus of Australian English: Large Corpus Collection with an Economical Portable and Replicable Black Box," in Proc. of the 12th Annual Conference of the International Speech Communication Association, 2011.

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

[14] Fernández Gallardo, L., Möller, S. and Wagner, M., "Comparison of Human Speaker Identification of Known Voices Transmitted Through Narrowband and Wideband Communication Systems," in Proc. of ITG Conference on Speech Communication, 2012.