Comparison of Human Speaker Identification of Known Voices Transmitted Through Narrowband and Wideband Communication Systems

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

[1]Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany
[2]Faculty of Information Sciences and Engineering, University of Canberra, Australia
Email: {laura.fernandez-gallardo|sebastian.moeller}@telekom.de, michael.wagner@canberra.edu.au

Abstract

The traditional Public Switched Telephone Network (PSTN) is the primary platform for voice communications and is commonly limited to narrowband (NB). It has been shown, however, that wideband (WB) transmission produces a higher quality speech signal compared to conventional NB. Additionally, the channel bandwidth plays a critical role in enabling speaker recognition since important voice features are widely distributed in the frequency domain. The goal of the present study is to determine the benefits on human speaker identification when the speech is transmitted through channels with different characteristics such as bandwidth and coding algorithms.

1 Introduction

Current speech communication services offering NB and WB channels use a variety of coding schemes at different bit rates to effectively compress and transmit the speech signal.

It has been demonstrated that WB transmission (50 – 7,000Hz) offers an improved channel quality compared to NB (300 – 3,400Hz) systems. Moreover, higher speech intelligibility and naturalness is experienced by users due to the enhanced bandwidth [1]. On the other hand, the channel bandwidth has a critical role for speaker recognition since essential voice characteristics, such as those of nasal or breathy voices are determined by higher frequencies, suppressed in NB channels.

The aim of this work is to investigate whether human speaker recognition also improves when moving from NB to WB speech transmission. To that end, we selected five common NB and WB compression schemes to study to which extent the characteristics of such systems enable humans to distinguish known voices.

The task contemplated is Speaker Identification (SI), that is, determining which speaker of a group of target speakers is most likely to have produced a given speech sample. Early research on speaker recognition has evaluated the listeners' ability to detect the speakers' identity varying speech parameters [2] or transmitting the voices through low-rate communication systems [3, 4]. According to these studies, essential cues for identification depend on the target voice and high-frequency components of the speech signal are important for the speaker recognition task.

Similarly, effects of NB and WB transmissions were examined in our experiment and also the length of utterances heard by the listeners was varied. An important factor to take into account is the level of acquaintance of listeners with the voices (familiarity) [5]. In the present study we focus on listeners who know the voices they have to identify from their long-term experience.

2 Experimental Set-up

The goal of the experiment is to investigate to which extent subjects are able to identify familiar speakers from their voices alone when these are transmitted through different NB and WB telephone channels. Additionally, we want to determine the effect of the length of the stimuli presented to listeners on their identification accuracy. Therefore, we selected tokens of different lengths recorded from a group of speakers and subsequently transmitted them through simulated communication channels with well-defined characteristics. These processed voice excerpts were presented to the listeners in the tests.

2.1 Speech material

Utterances in German language from a group of 16 (8 males, 8 females) native speakers were recorded at a 48kHz sampling rate and 16 bit quantization on a personal computer using an AKG C 414B-XLS microphone (frequency range from 20 to 20,000Hz).

The recorded speech consisted of three words, one sentence and one paragraph from every speaker. The three words *auch*, *immer*, and *können* (meaning *also*, *always*, and *be able to*) were chosen among the most common German words. The sentence and the paragraph were extracted from the EUROM texts [6], and their average duration was 2.7s and 11.9s respectively.

All tokens were transmitted through five simulated communication systems examined in the present work. We selected three NB systems using the codecs G.711 at 64kbps, Adaptive Multi-Rate (AMR)-NB at 4.75kbps and Global System for Mobile Communications (GSM)-Enhanced Full Rate (EFR) at 12.2kbps. The WB systems use the codecs G.722 at 64kbps and AMR-WB at 23.05kbps.

2.2 Auditory tests

The 26 (19 males, 7 females) listeners who participated in the experiment were native German speakers and working mates for more than two years at the same department as the speakers. Hence, we can assume that the voices of the experiment were already familiar to the listeners from a long-term exposure.

In the experiment session, listeners were asked to select the speaker identity among the 16 possibilities presented, after they heard each utterance. The stimuli were played from shorter to longer duration and the order was randomized for each subject. Only when the tokens heard

	Narrowband codecs						Wideband codecs			
	AMR-NB (4.75 kbps)		GSM-EFR (12.2 kbps)		G.711 (64 kbps)		G.722 (64 kbps)		AMR-WB (23.05 kbps)	
auch	40.62		45.19		50.48		61.30		63.70	
immer **Word**	50.48	47.76	58.89	54.25	62.12	56.65	69.23	66.75	70.91	67.31
können	52.16		58.65		57.45		69.71		67.31	
Sentence	84.37		89.66		89.18		93.99		94.95	
Paragraph	93.03		94.47		93.75		95.67		96.39	

Table 1: Mean accuracy (%) of listeners for different transmission channels and different stimulus lengths.

were paragraphs, the response time from the start of the stimuli until subjects gave their answers was measured.

Since we are considering 16 speakers and 5 communication systems, the samples played in each test session were 240 words (3 words × 16 × 5), 80 sentences (16 × 5), and 80 paragraphs. This results in a total of 400 stimuli.

3 Results and discussion

3.1 Accuracy of listeners

The accuracy of the listeners was calculated as the number of correct answers from all subjects divided by the total number of stimuli. The values of the accuracy reached for different communication channels and different sets of stimuli are given in Table 1.

Our results indicate that listeners achieve a higher accuracy when voices are transmitted through WB communication systems using the codecs G.722 or AMR-WB, while their performance when NB systems using codecs GSM-EFR or G.711 are applied remains significantly lower. Besides, the performance when voices were transmitted through the AMR-NB codec was significantly lower than when transmitting through any other NB or WB system.

3.2 Response time

The response time was measured when the stimuli presented to the listeners were paragraphs. In Figure 1, we observe that listeners needed significantly less time to identify the target speaker when the utterances were transmitted through WB codecs than when they were processed with the NB codecs G.711 and GSM-EFR. A greater significant difference in response time is found when the NB codec AMR-NB is compared with the rest of systems.

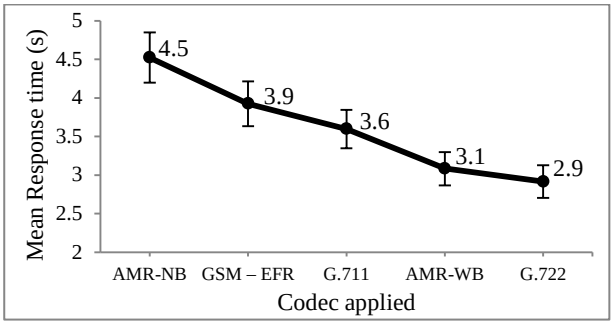

Figure 1. Mean response time with 95% confidence intervals when the stimuli were paragraphs (max. average duration: 11.9s).

4 Conclusions

The results of this work show that standard WB transmission offers significant improvements over NB, allowing listeners to identify known speakers more precisely and faster. The lowest performance is achieved with a NB transmission with the codec AMR-NB, possibly due to the combined effect of low sampling rate and low bit rate coding.

The full paper corresponding to this extended abstract will also include further analysis of our results supporting the improvements of WB over NB.

References

[1] Rodman, J., "The Effect of Bandwidth on Speech Intelligibility", Polycom inc., White paper, 2003.

[2] Van Lanker, D., Kreiman, J. and Emmorey, K., "Familiar voice recognition: Patterns and parameters, Part I: Recognition of Backwards Voices", Journal of Phonetics, 13:19-38, 1985.

[3] Uzdy, Z., "Human Speaker Recognition Performance of LPC voice processors", IEEE Trans. Acoust. Speech, Signal Processing, 33(3):752-753, 1985.

[4] Catellier, A. and Voran, S., "Speaker Identification in Low-Rate Coded Speech", in Proceedings of the 7th International MESAQIN (Measurement of Audio and Video Quality in Networks) Conference, 2008.

[5] Bőhm, T. and Shattuck-Hufnagel, S., "Listeners Recognize Speakers' Habitual Utterance-Final Voice Quality", in Proceedings ParaLing07, 2007, pp. 29-34.

[6] Gibbon, D., "EUROM.1 German Speech Database", ESPRIT Project 2589 Report (SAM, Multi–Lingual Speech Input/Output Assessment, Methodology and Standardization), Universität Bielefeld, D–Bielefeld, 1992.

[7] Fernandez-Gallardo, L., Möller, S. and Wagner, M., "Comparison of Human Speaker Identification Performance for Speech Transmitted Through Different Communications Channels", submitted for publication.