HUMAN SPEAKER IDENTIFICATION OF KNOWN VOICES TRANSMITTED THROUGH DIFFERENT USER INTERFACES AND TRANSMISSION CHANNELS

Laura Fernandez 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

ABSTRACT

Together with the variety of networks, diverse terminals and devices, such as telephones with handset or hands-free mode, mobile phones and headsets, are commonly available for everyday calls. We conducted an auditory test to examine the combined influence of these user interfaces, audio bandwidths, coding schemes and packet loss on human speaker identification of previously known voices. The effects of the user interfaces on transmission and reception were tested separately with the different channel impairments. Our study confirms that the identification task is facilitated if the voices are transmitted through wideband instead of narrowband channels, and that headsets and hands-free phones take greater advantage of the improved bandwidth that is gaining ground rapidly.

Index Terms— Human speaker identification, channel impairments, listening tests

1. INTRODUCTION

In today's telecommunication networks, speech signals are transmitted through a wide variety of channels with different characteristics. The bandwidth offered by the traditional Public Switched Telephone Network (PSTN) is commonly limited to narrowband (NB, 300 – 3,400Hz), while Voice over the Internet Protocol (VoIP) services also support wideband transmissions (WB, 50 – 7,000Hz). Super-wideband services (SWB, 50 – 14,000Hz), offering an extended bandwidth, have been emerging recently as users demand a higher-quality audio experience. An efficient transmission entails the digital data to be compressed at an adequate bit rate, depending on the network bandwidth and application requirements. However, the coding-decoding processes, especially at low bit-rate introduce non-linear distortions that degrade the quality of the speech to some extent.

Multiple investigations assess voice quality in telephony, focusing on the influence of various channel degradations. It has been shown that WB services offer advantages in voice naturalness and intelligibility over NB [1]. Regarding perceived signal quality, an improvement of about 30% has been found when switching from NB to WB [2]. We aim to demonstrate that human speaker identification can be considered as an additional criterion when judging the benefits of WB and SWB over NB and to motivate the deployment of IP-based services offering extended bandwidths for voice communications. Our previous study [3] shows, accordingly, that WB facilitates speaker identification of known voices, indicating that important speaker-specific information is conveyed through the frequencies filtered out in NB channels. However, only the effects of speech compression were contemplated as channel impairments in that study.

The user interfaces employed in communication channels introduce further distortion in sending and in receiving direction, due to the intrinsic characteristics of their microphones and loudspeakers and their integration into the physical device, respectively. The relevant aspects of terminals affecting the transmitted signal are referenced in the European Telecommunications Standards Institute (ETSI) standard method for end-to-end (mouth to ear) speech quality testing [4]. The influence of handsets and headphones in receiving direction in conjunction with that of different bandwidths has been found to be significant regarding signal quality [5]. However, the combined effects of transmission channels and terminals on human speaker recognition still need to be addressed. The goal of this work is to evaluate the effects of the user interface and other channel artifacts (i.e. bandwidth, codec and packet loss) on the performance of listeners identifying previously known voices.

We test the identification performance of voices transmitted over four user interfaces in sending direction: mobile phone, typical phone with handset, hands-free terminal and headphones, and over the last three of them also in receiving direction. Although devices are not consistent between brands from the design and technology point-of-view, we have chosen representative user interfaces typical for use with VoIP services. Only general user interface components are standardized, such as the Send and

Receive Loudness Ratings (SLR and RLR) and Listener Sidetone Rating (LSTR) [6], [7].

We also study the effects of different random packet loss rates on the speaker recognizability. These effects, which result from delay jitter compensated by a receive-side jitter buffer, have been tested for VoIP quality [5] and for speech recognition [8], but their influence on human speaker identification has not yet been investigated.

The remainder of this paper is organized as follows. Section 2 describes the methods for the preparation of the speech stimuli presented in the auditory test, which is detailed in Section 3. Section 4 shows the results of our experiment and discusses the listeners' performance over the different conditions. Finally, Section 5 presents concluding remarks.

2. PREPARATION OF SPEECH MATERIAL

A total of 16 (8 male, 8 female) native German speakers and work colleagues at the Q&U Lab volunteered to participate in our experiment as speakers. Their voices were recorded with a high-quality sound card with 48kHz sampling frequency and 16-bit quantization, with a AKG C 414 B-XLS microphone in an acoustically-isolated room. The segment "Könnten Sie mir", meaning "Could you (…) me" was extracted from two different parts of texts they read [9]. In this manner, two versions of the same segment, with a slightly different prosody from the same speaker, were used to test the speaker identification performance. The length of this segment is considered short enough [3] to get a resulting 60% to 90% accuracy. Our intention is to obtain identification rates in this range, which will enable us to compare different transmission conditions.

The original recordings were subsequently transmitted through different user transmission interfaces (devices tested in sending direction) and applying various codecs and packet loss rates, as listed in Table 1. The telephones employed in our study support NB and WB bandwidths as well as the specified codecs. These codecs are commonly employed in PSTN, ISDN, VoIP and mobile telephony at the indicated bit rates.

The corresponding user interface was connected to an Asterisk server and attached to a head-and-torso simulator, employed to reproduce the speech simulating the acoustic transmission path [4]. The network characteristics of Table 1 were programmed in the server, where the recordings were done in uncompressed audio format, with sampling frequency according to the transmission bandwidth, and 16 bit quantization. In the case of transmission through the headset, no codec was selected for the recordings in the server. Instead, these were made with 44.1kHz sampling frequency and 16-bit quantization and the coding-decoding process was applied later offline, via software simulation. For the processing through the mobile phone device, the set-up was placed in a different room and a different head-and-torso simulator was employed. The network simulator Rohde & Schwarz CMU 200 was employed for the transmission.

Interface	**Codec**	**bit rate (kbps)**	**Packet loss**
Phone with handset (SNOM 870) 	G.711 (A-law) (NB)	64	0, 5, 10, 15
	G.722 (WB)	64	0, 5, 10, 15
Hands-free phone (Polycom IP 7000) 	G.711 (A-law) (NB)	64	0
	G.722 (WB)	64	
Headset (Beyerdynamic DT 790) 	G.711 (A-law) (NB)	64	0
	G.722 (WB)	64	
	G.722.1C (SWB)	32	
	G.722.1C (SWB)	48	
Mobile phone (SONY XPERIA T) 	AMR-NB (NB)	12.2	0
	AMR-WB (WB)	12.65	

Table 1: User interfaces and channel impairments for the analysis in sending direction.

In all cases, the handsets or headset were attached to the head-and-torso simulator in a natural position, with about 3cm of distance from the artificial mouth to the microphone, and the hands-free phone was placed 1m away from the mouth on a desk. The speech level at the mouth reference point of the artificial heads was -4.7dBPa, according to ITU-T recommendations. The head and torso simulator models employed were HEAD acoustics HMS III and B&K 4128C, respectively. The rooms where the set-ups were placed had similar characteristics: office rooms with some furniture and approximate size (and reverberation time): 5m x 3m x 2.7m (280ms RT60) and 4m x 2.6m x 2.7m (200ms RT60).

The handset, the hands-free phone, and the headset were also tested in receiving direction, with the same network conditions as in sending direction except for packet loss, which was not considered in the study of the receive user interface. The processing of the initially recorded segments involved the transmission from the Asterisk server to the device used by the listeners in the auditory test. During the test session, the corresponding network bandwidth and codec were selected in the server before the transmission of each utterance. The stimuli to be heard through the headset were processed offline, transmitting the original recording through the four simulated communication channels.

The signal processing taking place in the devices, such as noise reduction, echo cancellation and voice activity detection is not known (as it is proprietary). However, we consider it not to be dominant in the processing of the entire channel, as background noise in the rooms during the

recordings was minimum, below 30dB(A). Hence, the emphasis of our study is on microphone and loudspeaker type, encapsulation, and interface handling.

3. AUDITORY TESTS

A total of 20 listeners (16 males, 4 females) were the subjects of the auditory test. They were native German speakers and colleagues working at the same department as the test talkers during more than two years. Half of them (6 males, 4 females) participated also as speakers and thus were confronted with their own, processed voice.

A Graphical User Interface (GUI) was written in Java to display the pictures and names of all speakers whose voices appear in the test, to adequately play the stimuli to the listeners, and to register their answers. The tests involved listening to the sets of stimuli with the appropriate user interface and to identify the corresponding speaker by clicking on one picture out of the 16 possibilities right after each audio stimulus. At the beginning of each test session, the voices of the test were trained in clean conditions by listening to one sentence of every speaker, at least once. This also permitted the subject to habituate to the test GUI.

The auditory test involved two individual sessions conducted on separated days: in the first session subjects listened to a total of 256 processed stimuli, resulting from 16 speakers x 16 conditions in sending direction. Listeners employed high quality, closed headphones to listen to this stimulus set: AKG K601 (frequency response 12 – 39,500 Hz) with diotic listening. Differently, in the second session they listened to 128 stimuli (16 speakers x 8 conditions in receiving direction), employing the corresponding user interface in a natural, realistic position. The distance from the hands-free phone to the listener was approximately 0.7m.

Either version of the two segments extracted from the original recordings was randomly selected for every speaker and transmission condition and included in the corresponding stimulus set. The reason of using two different versions of the utterances randomly was to avoid that the listeners' answers are guided by the learnt prosody of the voices. Furthermore, the order of stimuli played was randomized for each listener in both sessions.

The test was administered using a computer with a high quality sound card and the appropriate user interfaces for listening to the stimuli, connected to the Asterisk server for online processing in receiving direction. The sessions were conducted in a quiet office room and each of them took about 20 minutes to complete.

4. RESULTS AND DISCUSSION

The results of the accuracy reached by the group of listeners are presented in this section for the user interfaces in sending and in receiving direction, as well as for the effects of packet loss in sending direction.

4.1. Sending direction without packet loss

Considering the different user transmission interfaces, listeners identify the speakers more accurately when WB stimuli instead of NB stimuli are presented. However, no better identification rates are achieved when subjects are confronted with acoustic signals with a more extended bandwidth (SWB), as can be observed in Figure 1. The statistical significance of these outcomes was analyzed conducting the Mann-Whitney U test, a non-parametric test suitable in this case, when the data does not exhibit a normal distribution. The differences in accuracy are statistically significant for the particular user interfaces regarding bandwidth ($p<0.05$ for handset and for hands-free, and $p<0.001$ for mobile phone and for headset). No differences between the two bit rates of the SWB codec have been found. The accuracy reached with this bandwidth is insignificantly lower than that when listener heard WB stimuli through the same device (headset) but is significantly different from NB ($p<0.005$).

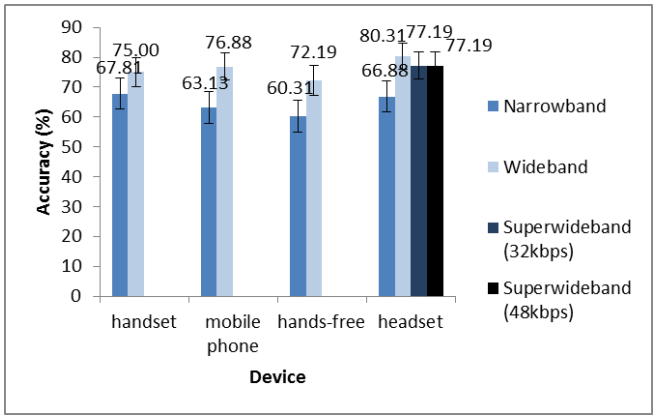

Figure 1: Accuracy reached for each sending interface with 95% confidence intervals.

The identification accuracy is also altered when the speech was transmitted through different devices in sending direction. However, significant differences are only found when the handset and the hands-free telephones are compared in NB ($p<0.05$) and when the headset is compared to the hands-free phone in WB ($p<0.05$).

The optimal user interface to capture the speech signal for WB channels is the headset, while for NB the handset enables a better recognition of the talker; this may be due to the fact that users are more habituated to handset devices in case of NB transmission. The hands-free terminal leads to an inferior accuracy rate in sending direction. Although care was taken to minimize the ambient noise when the speech was acquired by this device, speaker recognizability is influenced by the room and by the distance of the talker to the device. More significant differences between NB and WB accuracies have been found for the mobile phone and for the headset. Hence, we can conclude that higher

advantages from WB transmissions over NB can be obtained when the speech signal is acquired with these kinds of devices.

4.2. Sending direction with packet loss

The influence of packet loss is analyzed when the user transmission interface was the telephone with handset. In Figure 2, a decrease in identification accuracy is detected for both channel bandwidths as the random packet loss rate augments, being more pronounced for NB than for WB.

Considering the enhanced (WB) bandwidth, only the difference in correct answers comparing the loss rates 0% and 15% is statistically significant ($p<0.05$). For NB transmissions, differently, significant differences in accuracy are found between 0% and 5%, and between 0% and 15% rates ($p<0.05$). The effect of the bandwidth examining single packet loss conditions is also significant ($p<0.01$), which confirms the benefits of the WB communication channels.

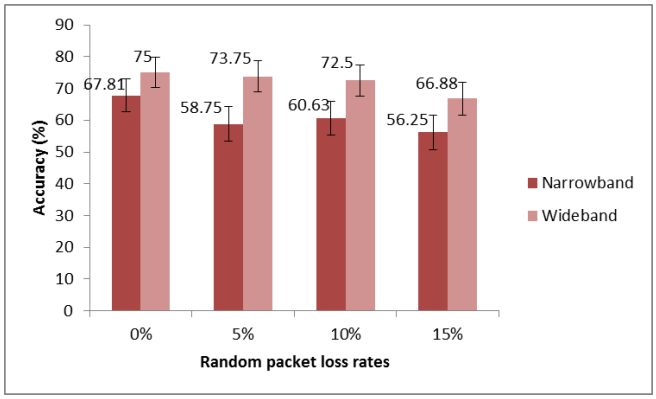

Figure 2: Accuracy reached for different random packet loss rates with 95% confidence intervals.

4.3. Receiving direction

The effects of different bandwidths and receiving interfaces are depicted in Figure 3. The impact of transmitting with different bandwidths is also evidenced in receiving direction. Nevertheless, the channel bandwidth has less influence for the phone handset, while the differences between NB and WB are statistically significant for the hands-free phone and for the headset ($p<0.001$). These are also the user interfaces preferred for longer calls, specifically multi-party, as they do not require occupying the hands holding the device.

Similar to the outcomes in sending direction, processing the stimuli with SWB has no effects on the accuracy compared to WB and no statistical differences are found comparing the two SBW bit rates (although the SWB scores are a little lower than WB). This finding was unexpected because SWB has been proven to offer higher signal quality [10]. However, it is probable that human listeners are not yet used to hear voices in the extended bandwidth or that the channel frequency response is less appropriate for the conservation of certain speaker's characteristics related to voice quality; this finding needs to be analyzed further.

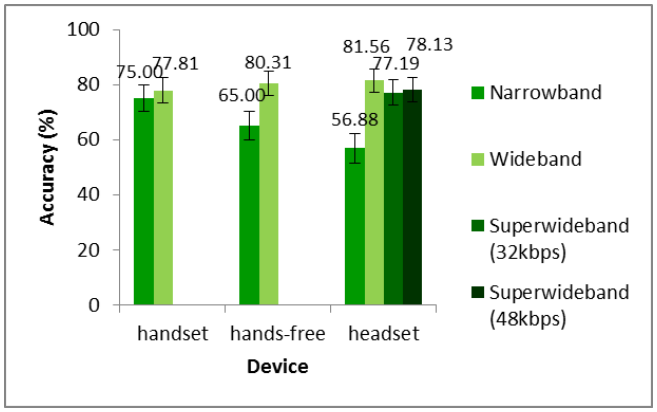

Figure 3: Accuracy reached for each receiving interface with 95% confidence intervals.

There are statistical differences among the three user interfaces ($p<0.05$) in NB, i.e. the identification accuracy decreases from handset towards hands-free and headset, which is not manifest in WB. This reinforces the advantages of WB communications.

5. CONCLUSIONS AND FUTURE WORK

We have analyzed the effects of channel bandwidth, channel coding, random packet loss and electro-acoustic user interfaces on human speaker identification performance, when the listeners are already familiar with the voices they listen to. It has been found that switching from NB to WB improves the identification accuracy for all the user interfaces evaluated, and to a larger extent if the voices are transmitted through mobile phones or headsets in sending direction. WB channels offer also significant advantages over NB if the speech is received through a hands-free phone or headsets, being less substantial the impact for a traditional handset. Regarding communication channels affected by random packet loss, WB permits a higher identification performance compared to NB, starting to decrease significantly at 15% packet loss, whereas the decrease in accuracy for NB channels is already noticeable at 5% packet loss.

Interestingly, SWB offers no improvements in speaker recognizability over WB. In future work we will compare different SBW codecs and study their impact on speaker recognition in more detail, focusing on different coding schemes, length of segments, and range of frequencies included in the communication channel. The effects on automatic speaker verification will also be investigated.

6. REFERENCES

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

[2] Wältermann, M., Raake, A. and Möller, S., "Quality dimensions of narrowband and wideband speech transmission," Acta Acustica united with Acustica, 96(6), pp. 1090-1103, 2010.

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

[4] ETSI EG 201 377-2: Specification and Measurement of Speech Transmission Quality; Part 2: Mouth-to-Ear Speech Transmission Quality Including Terminals, 2004.

[5] A. Raake, "Speech Quality of VoIP – Assessment and Prediction," John Wiley & Sons Ltd, Chichester, West Sussex, UK, 2006.

[6] ITU-T Recommendation G.121 (1993): "Loudness ratings (LRs) of national systems".

[7] ITU-T Recommendation G.111 (1993): "Loudness Ratings (LRs) in an international connection".

[8] Quercia, D., Docio-Fernandez, L., Garcia-Mateo, C., Farinetti, L. and De Martin, J.C., "Performance Analysis of Distributed Speech Recognition Over IP Networks on the AURORA Database," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 4, pp. 3820–3823, 2002.

[9] 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.

[10] Wältermann, M., Tucker, I., Raake, A. and Möller, S., "Extension of the E-Model Towards Super-Wideband Speech Transmission," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4654-4657, 2010.