CONVERSATIONAL EVALUATION OF ARTIFICIAL BANDWIDTH EXTENSION OF TELEPHONE SPEECH USING A MOBILE HANDSET

Hannu Pulakka[1], *Laura Laaksonen*[2], *Ville Myllylä*[2], *Santeri Yrttiaho*[1], *Paavo Alku*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland
[2]Nokia Smart Devices, Finland

hannu.pulakka@aalto.fi

ABSTRACT

Artificial bandwidth extension methods have been developed to improve the quality and intelligibility of narrowband telephone speech. Bandwidth extension methods have typically been evaluated with objective measures or subjective listening-only tests, whereas realistic conversational evaluations have been rare. This paper presents a conversational evaluation of two bandwidth extension methods together with narrowband and wideband speech. The evaluation was performed using a mobile handset with a wired earpiece and microphone both in silence and in simulated street noise. The results indicate that one of the evaluated bandwidth extension methods was significantly preferred over narrowband speech in silence. The results also suggest slight preference for this bandwidth extension method over narrowband speech in street noise. True wideband speech was considered superior to bandwidth-extended and narrowband speech especially in silence.

Index Terms— Speech enhancement, bandwidth extension, conversational evaluation

1. INTRODUCTION

The audio bandwidth of telephone speech is typically limited to approximately the traditional telephone band of 300–3400 Hz. The narrow bandwidth degrades the quality and intelligibility of speech. Narrowband speech with an audio bandwidth of less than 4 kHz is still primarily used in cellular telephone systems, but wideband speech services using an audio band of 50–7000 Hz are becoming available. The transition from narrowband to wideband telephony is likely to take time and, during the transition stage, users of wideband terminals experience a noticeable difference between wideband and narrowband calls. Furthermore, handovers causing switching between wideband and narrowband speech are possible during calls, and they are perceived as abrupt changes of speech quality. To reduce the difference between narrowband and wideband speech, techniques for the artificial bandwidth extension (ABE) of telephone speech have been developed. ABE methods attempt to regenerate missing frequency content outside the telephone band using only the narrowband speech signal as input. The speech bandwidth is most commonly extended above the telephone band in the range 3.4–8 kHz (e.g. [1, 2, 3, 4]) but extension below 300 Hz has also been studied (e.g. [2]).

ABE methods have been evaluated both with objective measures, such as the log spectral distortion (LSD) [1], and subjective listening tests. Subjective evaluation has been almost exclusively conducted using *listening-only* tests such as comparison category rating (CCR) tests [3, 4] and preference tests [2]. In these tests, a subject listens to pre-recorded speech samples and is typically allowed to replay the sample. However, more realistic assessment is possible using *conversational* tests [5] in which subjects evaluate a telephone connection in a conversation task. On the other hand, conversational tests are laborious and each subject typically evaluates the system with the speech of only one talker. To the knowledge of the authors, ABE has been evaluated with conversational tests only in [3] using the speaker phone mode of mobile handsets.

In this study, ABE was evaluated with a conversational test simulating realistic use of a mobile handset. A subject performed conversation tasks with another subject using a handset. To compare connection types, the subject could switch between two different connection types during each conversation. A natural noisy environment was simulated by generating ambient noise with a high-quality multi-channel reproduction system. Two ABE methods based on the earlier work of the authors ([3, 4]) were evaluated. The input to the methods was AMR-coded narrowband speech. Bandwidth-extended speech was compared with AMR-coded narrowband speech and AMR-WB-coded wideband speech.

2. CONVERSATIONAL EVALUATION

The test setting used for the conversational evaluation is illustrated in Fig. 1. Two subjects were seated in different rooms and performed conversation tasks using a simulated telephone connection. The test procedure was based on [5] and modified to enable a direct comparison between two connection types during a conversation.

2.1. Facilities and equipment

Subject 1 was seated in room 1, which conforms to the specifications of the ITU-R recommendation BS.1116-1 [6]. Subject 1 used a mobile handset with a wired earpiece and microphone (Section 2.2). Background noise was played in room 1 through nine Genelec 8260A loudspeakers (Section 2.5). Subject 1 could switch between two connection types at any time using a mechanical custom-made A/B switch. The evaluated connection types are described in Section 2.5. The A/B switch and the handset are shown in Fig. 2.

Subject 2 was located in room 2, which is a silent room next to room 1. No background noise was generated in room 2. Subject 2 used an AKG HSC 271 headset comprising circumaural closed-back headphones and a cardioid condenser microphone with a windscreen. The microphone was placed at a distance of about 6.5 cm from the subject's mouth at the chin level.

The test operator was seated in a control room between the test rooms. He wore an AKG HSC 271 headset, heard all conversations, and instructed the subjects before and between the tasks.

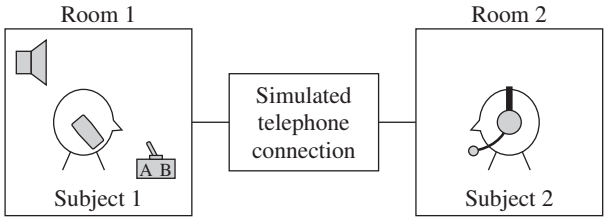

Fig. 1. Schematic illustration of the test setting.

Fig. 2. A/B switch for switching between two connection types and the mobile handset with a wired earpiece and microphone.

The telephone connection was simulated using a custom-made software running in a laptop computer. The software took care of input filtering, codecs, ABE processing, and output equalization of the speech signals in real time. Telephone call simulation and signal level monitoring were also performed with this software. The A/B switch was connected to the serial port of the computer and it controlled the software. The speech signals were routed through a MOTU Traveler mk3 audio interface and a headphone amplifier. The maximum one-way delay was measured to be about 120 ms, which is sufficiently low to provide essentially transparent interactivity [7]. A delay-free sidetone was arranged for both subjects.

2.2. Mobile handset

Since the acoustic front-end, i.e., microphone, earpiece (or loudspeaker), and their housing, was supposed to mimic the real phone as much as possible, an old Nokia 8310 device was used as a starting point (see Fig. 2). First, the earpiece acoustics were modified to support wideband frequencies. Then, both the microphone and the earpiece were wired to enable a connection to external amplifiers. Finally, the resulting acoustic front-end was measured according to the ITU-T P.64 recommendation [8] to make sure it is applicable for wideband testing. The recommendation, which is also followed by the mobile phone industry, describes how to determine sensitivity/frequency characteristics of handsets. The handset has to be measured as it is used – on the ear. The B&K 4128 Head and Torso Simulator, which conforms to [8], was used as measurement equipment. The way how the phone is pressed on the ear may affect the perceived response substantially. The recommendation specifies a so called standard handset position as well as alternatives to be used for devices with a peculiar form factor or other specialties. Since the handset used in this study was a typical one and the acoustic design was such that the earpiece response is robust against positioning, the standard handset position was used.

3GPP specifies the performance requirements for the acoustic characteristics of 3G terminals when used to provide narrowband or wideband telephony [9]. An equalizer was designed for the handset earpiece based on the measured frequency response of the handset and the 3GPP handset receiving sensitivity frequency mask. The measured and equalized frequency response of the handset and the sensitivity frequency mask are shown in Fig. 3.

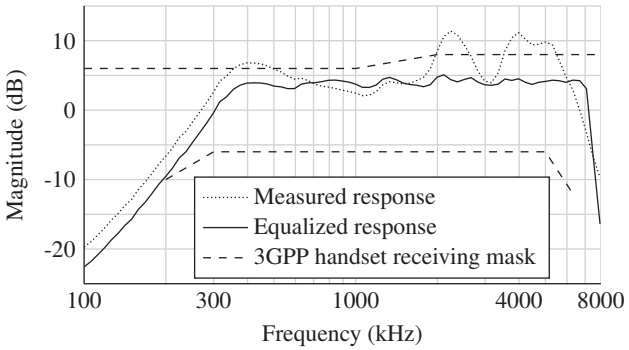

Fig. 3. Magnitude response of the handset earpiece. The curves illustrate the measured response (dotted), the equalized response (solid), and the receiving sensitivity frequency mask [9] (dashed).

Since the microphone characteristics do not affect the perceived ABE quality in the test setting of this study, the microphone response was only measured to verify its functionality, but not equalized.

2.3. Tasks and assessment

The subjects performed asymmetric conversation tasks in Finnish. The same tasks were used also in [3]. Subject 1 called subject 2 to inquire, e.g., hotel options at a holiday resort or details about magazines. Written instructions were given for each task. The tasks were designed so that subject 1 primarily asked a few questions and had the possibility to listen carefully to subject 2, who talked for most of the time. Conversations typically lasted from about 1.5 up to about 5 minutes. Subject 1 could switch between two connection types, A and B, at any time during the conversation. After each conversation, subject 1 answered the following question (in Finnish) using a pen and paper: *Which telephone connection would you prefer to use?* The response alternatives were *A*, *B*, *no opinion*, and *no audible difference*. Subject 1 could also write down comments, but they are not discussed in this paper due to the space limitation.

2.4. ABE methods

Two ABE methods were evaluated in this study. The methods are referred to as ABE1 and ABE2 and they have been described in [3] and [4], respectively. The input signal to both methods is a narrowband speech signal sampled at 8 kHz, and the output is a bandwidth-extended signal sampled at 16 kHz. Both methods extend the bandwidth of the narrowband input to the frequency range 4–8 kHz, which is denoted as the *highband*, and process the signal in frames of 10 ms. Both methods generate an excitation signal for the highband, divide it into four sub-bands using a filter bank, and shape the highband spectrum by adjusting the gain coefficients of the sub-bands in each frame. In ABE1, the excitation signal is generated by spectral folding of the input signal directly, whereas the highband excitation in ABE2 is based on the linear prediction residual of the input. The spectral shape of the highband is estimated from time-

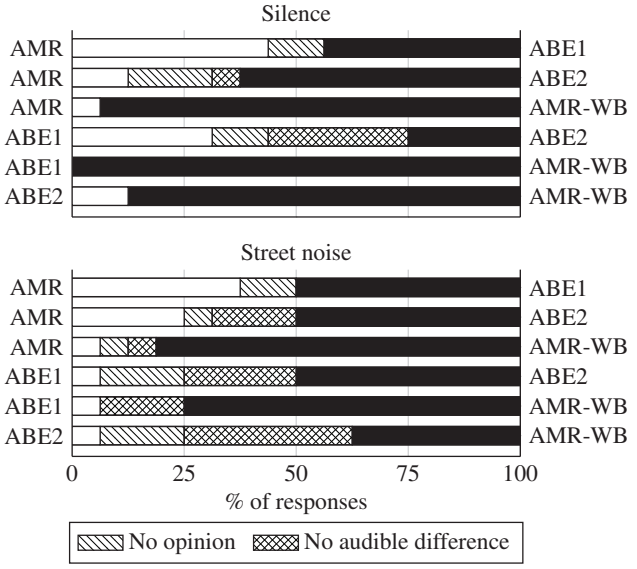

Fig. 4. Preference responses of the subjects.

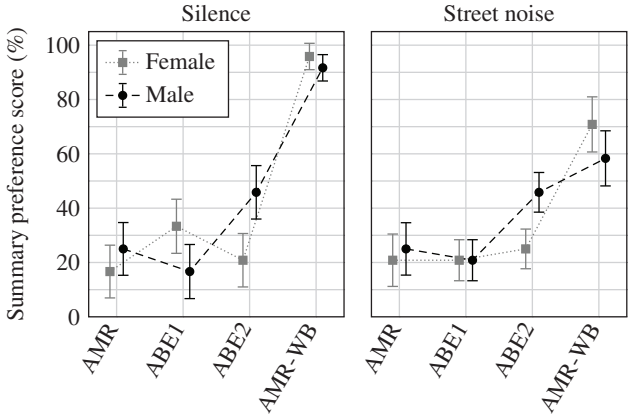

Fig. 5. Summary scores of preference. The scores indicate the percentage of times that a connection type was preferred in comparisons. Error bars indicate the standard error of mean.

domain and frequency-domain features calculated from each input frame. ABE1 classifies each frame into one of three phonetically motivated categories and adjusts the sub-band gain coefficients accordingly. ABE2 utilizes a neural network to estimate the sub-band energy levels and calculates gain coefficients from the energy estimates. Both ABE methods adjust the highband level depending on the estimated noise level in the input speech. ABE2 additionally attenuates the highband during pauses in speech. The near-end noise dependency implemented in ABE1 was disabled for the current study. Finally, in both methods, the weighted sub-bands are summed up and combined with the interpolated narrowband signal to produce the bandwidth-extended output signal.

2.5. Evaluated connection types and conditions

The following four connection types were evaluated for speech transmission from subject 2 to subject 1:

- **AMR**: Narrowband speech coded with the AMR codec [10] at the bit rate of 12.2 kbps.
- **ABE1**: AMR-coded speech processed with ABE1.
- **ABE2**: AMR-coded speech processed with ABE2.
- **AMR-WB**: Wideband speech coded with the AMR-WB codec [11] at the bit rate of 12.65 kbps.

In each test case, speech from subject 2 was processed with two connection types in parallel. The position of the A/B switch determined which of the two connection types was heard by subject 1. In all cases, the speech signal was first highpass filtered with the MSIN filter [12], which simulates the input response of a mobile station, and finally filtered with the earpiece equalization filter (Fig. 3). All four connection types were compared pairwise with each other in two different noise conditions in room 1: (1) silence and (2) street noise with an average sound pressure level of 61 dB (A weighting). A street noise recording was converted from the B-format to loudspeaker signals for the 9-loudspeaker setup using directional audio coding (DirAC) [13]. The test thus contained a total of 12 test cases and one additional practice case in the beginning. The order of the test cases and the order of the connection types in each comparison were randomized separately for each test. The AMR-WB codec was always used for speech transmission from subject 1 to subject 2.

The subjects were allowed to adjust the volume to a suitable listening level before starting the test. During the volume adjustment, both silent and soft street noise (51 dB, A weighting) conditions were presented in room 1. However, there was some acoustical leakage from the earpiece to the microphone in the handset. Therefore, the volume adjustment of subject 1 was limited to moderate levels to avoid disturbing echo for subject 2. A simple echo attenuation technique was also implement in software.

Sixteen conversation tests were arranged. Native Finnish subjects between 20 and 38 years (average 25 years) of age participated in the tests. Sixteen different subjects (8 females and 8 males) served as subject 2 and, similarly, sixteen different subjects evaluated the connection types as subject 1. Some of the subjects participated twice, both as subject 1 and subject 2.

3. RESULTS

The responses of the subjects are illustrated in Fig. 4. The mean values of the preference scores were compared with repeated measures analyses of variance (ANOVA) where two factors (connection type and presence/absence of noise) and a categorical predictor (speaker gender) where included. Pairwise comparisons between different levels of the ANOVA factors were conducted with Tukey's honestly significant difference (HSD) post-hoc tests. All statistically significant effects are reported below.

The summary score of preference was calculated as the percentage of times that a given connection type was preferred in the course of all comparisons. The summary scores are shown in Fig. 5. The summary score depended on the connection type [$F(3,42) = 23.80$, $p < 0.001$]. The most preferred connection type was AMR-WB (79.2 %), which received a higher score than the rest of the connection types (p-values < 0.001). The scores of ABE2, ABE1, and AMR were 34.4 %, 22.9 %, and 21.9 %, respectively. The summary scores depended on the presence of background noise [$F(1,14) = 7.89$, $p < 0.05$] as higher score values were obtained in silence (43.2 %) than in street noise (35.9 %). This indicates a larger proportion of neutral responses in noise. Further, the contrast between AMR-WB and the other connection types was accentuated in silence relative to the street noise condition [$F(3,42) = 4.11$, $p < 0.05$].

Table 1. Preference scores of pairwise comparisons between connection types. Preference for the latter connection type was coded as $+1$, preference for the former as -1, and the neutral responses as 0. t-tests were computed for the score values against the zero value. The pairs where one of the connection types was preferred consistently are indicated with statistically significant p-values in boldface.

Noise	Methods	Mean	t-value	p
silence	AMR – ABE1	0.00	0.00	n.s.
silence	AMR – ABE2	0.50	2.74	**< 0.05**
silence	AMR – AMR-WB	0.88	7.00	**< 0.001**
silence	ABE1 – ABE2	−0.06	−0.32	n.s.
silence	ABE1 – AMR-WB	1.00	> 6.00	**< 0.001**
silence	ABE2 – AMR-WB	0.75	4.39	**< 0.001**
street	AMR – ABE1	0.13	0.52	n.s.
street	AMR – ABE2	0.25	1.17	n.s.
street	AMR – AMR-WB	0.75	5.20	**< 0.001**
street	ABE1 – ABE2	0.44	2.78	**< 0.05**
street	ABE1 – AMR-WB	0.69	4.57	**< 0.001**
street	ABE2 – AMR-WB	0.31	2.08	n.s.

The A/B comparisons were analyzed also in a pairwise manner between the connection types as shown in Table 1. These pairwise scores depended on the pair of connection types [$F(5, 70) = 4.74$, $p < 0.001$] and on the interaction between the presence/absence of noise and the pair or connection types [$F(5, 70) = 2.87$, $p < 0.05$].

4. DISCUSSION AND CONCLUSIONS

Two artificial bandwidth extension methods were evaluated against AMR-coded narrowband speech and AMR-WB-coded wideband speech with a conversational test using a mobile handset. The connection types were evaluated both in silence and in simulated street noise. The summary preference scores illustrated in Fig. 5 indicate that AMR-WB was found to be superior to the other connection types in the silent environment. In street noise, however, there is more variation in the responses and the superiority of AMR-WB is not as pronounced. The summary scores suggest that ABE2 performs well compared with AMR for male voices both in silence and in street noise, but the differences are not statistically significant.

The analysis of pairwise comparisons between the connection types (Table 1) indicated that ABE2 was significantly preferred over AMR in silence. Also, according to the pairwise analyses, AMR-WB was considered superior to ABE2 in silence, but in street noise the preference was less pronounced and, in fact, not statistically significant. The distribution of subjects' responses shown in Fig. 4 suggest a slight preference for both ABE methods over AMR in street noise, but the differences are not statistically significant.

Earlier listening-only tests have reported significant preference for both ABE1 [3] and ABE2 [4] over narrowband speech. The small-scale conversation test reported in [3] also indicated clear preference for ABE1 over narrowband speech in speaker phone use. In the conversational evaluation of the current study, preference for ABE over narrowband speech could be statistically confirmed only for ABE2 in silence. The test arrangement including active conversation, the use of a handset, and the realistic background noise are likely to partly explain the difference in results compared with earlier studies. Furthermore, the 16 conversation tests of this study resulted in a much smaller amount of data for statistical analyses than listening-only tests with a comparable effort and the same number of subjects would have produced. More tests are needed to obtain statistically significant results on preference in various conditions.

Conversational evaluation of ABE methods provided valuable information about user preference and the evaluation methods. The test procedure with asymmetric tasks and the A/B switch allowed a direct comparison of connection types and was found to be a useful extension to the conversation test methods described in [5].

5. ACKNOWLEDGEMENTS

The work of Hannu Pulakka is funded by Nokia, the GETA graduate school, the Academy of Finland (LASTU research programme 135003), and Aalto University (Mide/UI-ART). The authors are grateful to M.Sc. Mikko-Ville Laitinen and M.Sc. Tapani Pihlajamäki for their invaluable help with the test facilities. Finally, the authors would like to thank the participants of the conversation tests.

6. REFERENCES

[1] Peter Jax and Peter Vary, "On artificial bandwidth extension of telephone speech," *Signal Process.*, vol. 83, no. 8, pp. 1707–1719, 2003.

[2] H. Gustafsson, U. A. Lindgren, and I. Claesson, "Low-complexity feature-mapped speech bandwidth extension," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 2, pp. 577–588, 2006.

[3] Laura Laaksonen, Ville Myllylä, and Riitta Niemistö, "Evaluating artificial bandwidth extension by conversational tests in car using mobile devices with integrated hands-free functionality," in *Proc. Interspeech*, 2011, pp. 1177–1180.

[4] Hannu Pulakka and Paavo Alku, "Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2170–2183, 2011.

[5] Int. Telecommun. Union, *ITU-T Recommendation P.805, Subjective evaluation of conversational quality*, 2007.

[6] Int. Telecommun. Union, *ITU-R Recommendation BS.1116-1, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*, 1997.

[7] Int. Telecommun. Union, *ITU-T Recommendation G.114, One-way transmission time*, 2003.

[8] Int. Telecommun. Union, *ITU-T Recommendation P.64, Determination of sensitivity/frequency characteristics of local telephone systems*, 2007.

[9] 3rd Generation Partnership Project (3GPP), *Terminal acoustic characteristics for telephony; Requirements; 3GPP TS 26.131*, 2011, Version 10.2.0.

[10] 3rd Generation Partnership Project (3GPP), *ANSI-C code for the floating-point Adaptive Multi Rate (AMR) speech codec; 3GPP TS 26.104*, 2009, Version 9.0.0.

[11] 3rd Generation Partnership Project (3GPP), *Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; ANSI-C code; 3GPP TS 26.204*, 2009, Version 9.0.0.

[12] Int. Telecommun. Union, *ITU-T Recommendation G.191, Software tools for speech and audio coding standardization*, 2005.

[13] Ville Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, 2007.