Intelligibility Assessment of a System for Artificial Bandwidth Extension of Telephone Speech

Florian Heese, Bernd Geiser, and Peter Vary
Institute of Communication Systems and Data Processing (iNAD) RWTH Aachen University, Germany
{heese|geiser|vary}@ind.rwth-aachen.de

Introduction

Techniques for artificial bandwidth extension (ABWE) [1] aim to extend the limited audio bandwidth of current narrowband telephone systems (0.3 \( \text{–} \) 3.4 kHz) towards the wideband frequency range (0.05 \( \text{–} \) 7 kHz). In several studies, a consistent quality advantage of the enhanced signals over narrowband speech is reported, e.g., [2]. In this contribution we evaluate an ABWE algorithm under realistic conditions and assess its impact on the speech intelligibility.

A modified rhyme test has been conducted to compare the intelligibility of narrowband (IRS filtered) telephone speech with its bandwidth extended counterpart under different realistic background noise environments (babble noise and car noise). In the “babble noise” environment, the test stimuli have been encoded with the ITU-T G.722 codec and presented to the subjects using wideband handsets (DECT+/CAT-iq). For the car noise environment (city and Autobahn), a typical hands-free equipment was simulated.

ABWE Algorithm

For the present study, the extension from the 0.3 \( \text{–} \) 3.4 kHz range towards higher audio frequencies was considered. The employed ABWE algorithm follows the source-filter model of speech production. The linear prediction residual of the received narrowband speech signal (0.3 \( \text{–} \) 3.4 kHz, \( f_s = 8 \) kHz) is normalized, mixed with white noise, and then reused as excitation for the artificial extension band. The extension band signal, also sampled at \( f_s = 8 \) kHz, represents the 4 \( \text{–} \) 7 kHz part of the wideband frequency range. The spectral envelope of the extension band signal is shaped with an autoregressive filter of order 4. Afterwards, an interpolated gain factor is applied. The filter coefficients and the gain factor are estimated from 14 narrowband features (zero crossing rate and MFCCs). The employed estimator is based on Hidden Markov Modelling and has been configured with 128 states and 16 Gaussian mixture components per state to model the state-dependent observation PDF of the narrowband features, see [1]. Finally, the narrowband speech signal and the synthesized extension band signal are combined with a QMF synthesis filterbank. The remaining “spectral gap” between 3.4 and 4 kHz has only a minor perceptual impact, cf. [1].

Modified Rhyme Test

A rhyme test according to [3] is used to determine the intelligibility of narrowband telephone speech and bandwidth extended speech in different acoustic environments. Therefore, the subject is asked to identify the target word from an ensemble of rhyming words (including the target word). The original rhyme test [3] consists of 100 ensembles per test which is not applicable due the huge duration that is required for each test cycle. Therefore, we chose a modified shortened rhyme test similar to [4] consisting of 26 ensembles. The used word list represents the phoneme distribution of the German language. Since fricative sounds (/s/, /z/, /.../) are often considered to be particularly relevant in ABWE it makes sense to classify the 26 rhyming ensembles as follows:

Class A: Ensemble does not contain fricative sounds (9 \( \times \)), e.g., Wacht, Macht, dacht, Nacht, Jacht.
Class B: Ensemble consists of words with identical fricative sounds (4 \( \times \)), e.g., Fell, Fall, full, viel, fahl.
Class C: Target word has a fricative sound, some alternative words do not (7 \( \times \)), e.g., Tat, bat, Rat, Naht, fad.
Class D: Target word has no fricative sound, some alternative words do (6 \( \times \)), e.g., hin, drin, Zinn, Sinn, Kinn.

In case of the narrowband speech, the stimuli were obtained by applying the P.48 IRS filter and a normalisation to a level of \(-26\) dB. The ABWE stimuli are processed versions of the narrowband stimuli, thus the energy for the extension band is added to the narrowband energy. To reflect realistic acoustic scenarios, these stimuli are presented in three different setups:

1. DECT+/CAT-iq Gigaset S68H handset with diffuse ambient babble noise. The noise field is generated in line with [5]. All stimuli were encoded with the ITU-T G.722 wideband codec.
2. Hands-free equipment with ambient city driving noise (4-channel recording in a middle class car).
3. Hands-free equipment with Autobahn noise (do.).

These acoustic environments have been created in a sound-attenuated booth, where the ambient noise fields are generated by four loudspeakers. An additional speaker has been used to simulate the hands-free car equipment. The signal-to-noise-ratio (SNR) for Setup 1 was measured as \(-14\) dB, for Setup 2 as 5.5 dB, and for Setup 3 as 0.5 dB.

Each setup was assessed in two separate experiments. The first experiment comprises 13 narrowband target words and 13 different ABWE processed target words. In the second experiment the roles of narrowband and ABWE target words have been interchanged. Each subject had to participate in two experiments with two different noise setups, beginning with the most “difficult” one (e.g., handset). To eliminate any residual learning effect, at least three days passed between the test sessions, and the presentation order of the ensembles was randomized for each experiment. Each of the 3 \( \times \) 2 experiments was completed by 10 subjects, leading to 10 \( \times \) 26 = 260 votes per condition (IRS or ABWE) and setup (1/2/3). Altogether 30 normal-hearing subjects participated the listening test.

Results

For the present rhyme test, the intelligibility \( \nu \) is defined as:

\[
\nu = \frac{C}{N} \cdot 100\% - K, \quad K = \frac{W}{N(A-1)} \cdot 100\%.
\]

where \( C \) represents the number of correct answers, \( N \) the number of ensembles. \( K \) is a correction value to incorporate the statistical guessing-probability, whereby \( W \) is the number of wrong answers and \( A \) is the ensemble size [3].

The results of the listening test, as listed in Table 1, were obtained by averaging the subject-individual intelligibility \( \nu_n \).
Objective Intelligibility Measurement

For reference, the intelligibility of each condition/setup-combination has also been measured objectively with the Speech Intelligibility Index (SII) [6]. For this evaluation, noise recordings, taken from Setups 1-3, at varying input SNR between −15 dB and 15 dB with an increment of 2.5 dB were generated for the target words of the ensembles (IRS processed and its ABWE proc. version). Based on these quantities, the SII was calculated. The averaged results are depicted in Figure 2. For each setup, the performance of the ABWE version is, compared to the IRS version, consistently better over the complete SNR range. While Setups 2 and 3 (hand-free setups) behave similar, Setup 1 (handset) has a relatively low SII, especially at lower SNR values. As a tendency, the SII benchmark confirms the results of the rhyme test.

References


Figure 1: Spectrograms of “Tier” and “fiel”

Figure 2: Speech Intelligibility Index (SII) plotted over SNR

Table 1: Listening test results for each setup

<table>
<thead>
<tr>
<th>Setup</th>
<th>Intelligibility in %</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRS</td>
<td>ABWE</td>
<td></td>
</tr>
<tr>
<td>1. Handset babble</td>
<td>66.35</td>
<td>73.08</td>
</tr>
<tr>
<td>2. Hands-free city</td>
<td>94.47</td>
<td>96.88</td>
</tr>
<tr>
<td>3. Hands-free Autobahn</td>
<td>80.29</td>
<td>80.29</td>
</tr>
</tbody>
</table>

Table 2: Ensemble class results for Setup 1 (handset)

<table>
<thead>
<tr>
<th>Ensemble Class</th>
<th>Intelligibility in %</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRS</td>
<td>ABWE</td>
<td></td>
</tr>
<tr>
<td>Class A (9x)</td>
<td>65.63</td>
<td>73.96</td>
</tr>
<tr>
<td>Class B (4x)</td>
<td>66.67</td>
<td>95.83</td>
</tr>
<tr>
<td>Class C (7x)</td>
<td>39.38</td>
<td>57.50</td>
</tr>
<tr>
<td>Class D (6x)</td>
<td>86.25</td>
<td>37.50</td>
</tr>
</tbody>
</table>

Conclusions

A listening test to determine the speech intelligibility of narrowband telephone speech vs. its ABWE processed counterpart has been carried out. By applying the ABWE, an improvement of the speech intelligibility could be found, particularly for instationary noise (even at better SNR, see Autobahn vs. city). This result is confirmed by the SII benchmark. In addition, the analysis of the ensemble classes showed that fricatives are indeed clearly improved (Classes B and C) and also non-fricatives (Class A) benefited from ABWE. Caution is advised for Class D, so that non-fricatives (e.g., plosives, stops) are not mistaken for other speech sounds after applying ABWE. In this context, further algorithmic optimization could be useful. Also the impact on the speech intelligibility by additional artificial low frequencies (0.05–0.3 kHz) promises to be an interesting question for further research.