Quality Improvement of Telephone Speech by Artificial Bandwidth Expansion – Listening Tests in Three Languages

Hannu Pulakka[1], *Laura Laaksonen*[2], *Paavo Alku*[1]

[1]Lab. of Acoustics and Audio Signal Processing, Helsinki University of Technology, Espoo, Finland
[2]Multimedia Technologies Laboratory, Nokia Research Center, Helsinki, Finland
hannu.pulakka@tkk.fi, laura.laaksonen@nokia.com, paavo.alku@tkk.fi

Abstract

Quality and intelligibility of narrowband telephone speech can be improved by *artificial bandwidth expansion* (ABE), which expands the speech bandwidth using only information available in the narrowband speech signal. This paper describes an ABE method that generates a high-band expansion using spectral folding and then modifies the magnitude spectrum of the expansion band with spline curves. The performance of the ABE algorithm was evaluated by formal listening tests in three languages: American English, Russian, and Mandarin Chinese. The results of the listening tests indicate that ABE-processed speech was preferred to narrowband speech in all tested languages.

Index Terms: speech enhancement, speech bandwidth expansion, listening test

1. Introduction

The audio bandwidth utilized for speech transmission in most current communications systems, including traditional PSTN and GSM networks, is limited to the frequency range of 300–3400 Hz. This limitation reduces both the quality and the intelligibility of speech because voice sounds contain important spectral contents beyond the upper limit of the transmitted band. The cutoff frequency of 3.4 kHz will be extended up to 7 or 8 kHz in, for example, the 3G wireless system with the AMR-WB speech codec [1]. However, wideband speech transmission requires that both ends of the transmission channel and the link between them support wideband speech coding. The transition from narrowband systems to wideband is likely take a long time. During the transition phase, the quality gap between narrowband and wideband coding can be reduced using a method called *artificial bandwidth expansion* (ABE) that regenerates the missing content of the highband (4–8 kHz) artificially at the receiving end of the transmission link using only information available in narrowband speech. Consequently, speech quality and intelligibility can be improved without any additional transmitted information.

Most artificial bandwidth expansion methods are based on the source-filter model of human speech production [2]. In this model, speech is divided into an excitation signal and a vocal tract filter. Correspondingly, the bandwidth expansion problem is split into two subtasks: the expansion of the excitation, and the expansion of the filter. The excitation can be extended by, e.g., spectral folding, spectral translation, or non-linear processing [3]. For the expansion of the filter, many bandwidth expansion algorithms described in the literature utilize codebooks. An ABE method without explicit expansion of excitation and filter separately was proposed by [4]. This algorithm uses spectral folding for an initial bandwidth expansion. Each speech frame is classified into one of three phonetically motivated categories, and the magnitude spectrum of the expansion band is then shaped according to this classification with a spline curve. This algorithm is the basis of the method applied in the current study.

To the best of our knowledge, artificial bandwidth expansion methods have not been previously evaluated thoroughly by listening tests in several languages. In this study, formal listening tests were conducted in three languages: American English, Russian, and Mandarin Chinese. These languages are among the most widely spoken languages in the world. They also possess different phonetic characteristics. Russian has a rich set of fricative sounds, which are problematic for bandwidth expansion, and Mandarin Chinese is a tonal language.

2. Method

The artificial bandwidth expansion algorithm used in the listening tests was nearly identical to that presented in [4]. A block diagram of the method is shown in Figure 1. The input signal, s_{nb}, is treated in short time domain frames. New initial frequency components are created through spectral folding [5], which is implemented in the time domain by zero-insertion. As a result, the sampling rate of the signal is doubled from 8 kHz to 16 kHz, and in the frequency domain the folded frequency components appear in the highband. The amplitude spectrum, computed by a 256-point FFT, is lowpass filtered in the frequency domain to smoothen the mirror images of the harmonics between 5.5 and 8 kHz. The highband is modified with a shaping function that is constructed based on a feature vector and a speech sound category derived from the original narrowband signal. Finally, the artificial wideband spectrum is converted back to the time domain through inverse FFT.

The shaping of the highband has a substantial influence on the quality of the artificial wideband signal. The frames are classified into three categories; voiced sounds, sibilants, and plosives. The classification has been optimized using Finnish speech samples. In Finnish, there is only one sibilant, */s/*. The category of plosives comprises */k/*, */p/*, and */t/*, and the rest of Finnish phonemes are classified as voiced sounds in the algorithm.

The classification of the frames is based on a feature vector extracted from the original narrowband signal. The vector comprises the following features:

- Gradient index, which is defined as the sum of the magnitudes of the gradient of the speech signal at each change of direction [3].
- Gradient count, which is a feature describing how long the level of gradient indices has exceeded a predefined level.

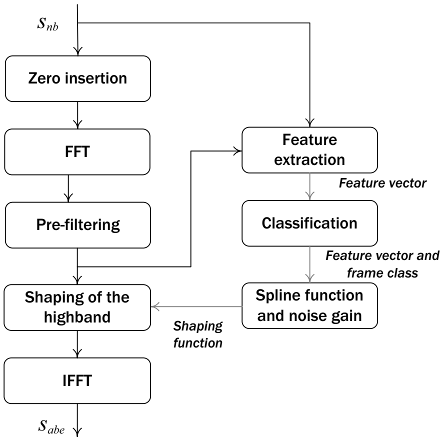

Figure 1: *Block diagram of the artificial bandwidth expansion algorithm. The input signal, s_{nb}, is a narrowband signal with sampling rate of 8 kHz. The output signal, s_{abe}, is the artificial wideband signal with sampling rate of 16 kHz.*

- Energy ratio, which is the energy of the current frame divided by the energy of the previous frame.
- Energy quotient, which is the ratio between a short term frame energy and a long term frame energy.
- Narrowband slope, which is the slope of the narrowband amplitude spectrum estimated between 0.3 and 3.0 kHz.

After the frame has been classified into one of the speech sound categories, the shaping function is constructed. The cubic spline was chosen as a magnitude shaping function because of its good characteristics, such as being smooth and local. The spline function is constructed around five control points located at 4, 5, 6, 7, and 8 kHz. The magnitudes of the control points are calculated from the following equation:

$$C_k = b_k + a_k \cdot n_{nb}, \qquad 1 \leq k \leq 5 \tag{1}$$

where b_k and a_k are predefined control point constants for control point k, and n_{nb} is the narrowband slope. C_1 is always zero, to guarantee a smooth transition around 4 kHz.

The values of the the control point constants are different for each speech sound category and they were optimized using a genetic algorithm (GA) [6] based search. A MSE criterion between the artificially expanded signal and a corresponding true wideband signal was used in the GA search.

After the shaping function is defined, an additional noise dependent gain is added to it. The gain is defined in decibels. As a rule of thumb, the tuning gain is negative if the noise level of the expanded signal is high, and the tuning gain is positive if the background noise level at the listener is high. The rationale behind this is that as the narrowband spectrum is folded to the highband, the noise in the original signal is also folded. Since the noise in the highband easily sounds annoying, the highband is attenuated. On the other hand, if the listener is in a noisy environment, the high frequency components can be amplified resulting in intelligibility improvement, as the noise masks possible artifacts. A frequency-domain example with different processing steps is shown in Figure 2.

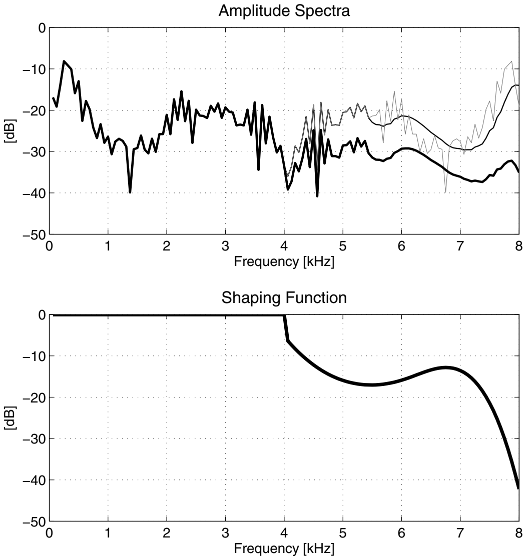

Figure 2: *Amplitude spectrum of /i/ spoken by a male speaker (top) and the corresponding shaping function (bottom). Gray curve is the folded spectrum. Thin black curve is the spectrum after pre-filtering, and bold black curve denotes the shaped spectrum. The highband part of the shaping function has been attenuated by 7 dB due to background noise level estimates.*

3. Experiments

Formal listening tests were arranged to evaluate the performance of the ABE method. The main goal was to study how listeners assessed the quality of ABE-processed speech in comparison to narrowband and wideband sounds.

3.1. Speech samples

Listening tests were arranged in three languages: American English, Russian, and Mandarin Chinese. Speech samples were taken from the NTT database, which contains high-quality recordings of short sentences in many languages [7]. For each of the three languages, eight sentences were included in the test. Every sentence was spoken by a different speaker, and all test sentences had different content. Four of the speakers were females and four males. The duration of each sentence was approximately two seconds.

The sound level of each test sentence was normalized to 26 dB below overloading. Pre-recorded office noise was added so that the signal-to-noise ratio (SNR) was 35 dB. The speech samples were then filtered with a model of the input characteristics of a GSM mobile station. All the samples obtained in this way were then processed by the following processings:

- Narrowband reference: AMR-NB at 12.2 kbps
- ABE: AMR-NB at 12.2 kbps followed by ABE
- Wideband reference: AMR-WB at 12.65 kbps

Three ABE versions were included in the test. The only difference between these ABE processings was the gain of the expansion band. Only the ABE version with the highest expansion band gain is discussed in this paper.

All the processing chains involved tandem coding, i.e., the speech signal was coded and decoded twice using the same codec.

No level corrections were made after the processings. All processed samples were finally filtered with an estimated response of a wideband mobile terminal.

3.2. Listening test

The Comparison Category Rating (CCR) method [8] was used in the listening test. Processings were compared pairwise so that each test item contained two instances of the same sentence processed with two different processing types. The task of the subject was to evaluate the quality of the second sentence compared to the quality of the first sentence on the following seven-point scale: *much worse* (-3), *worse* (-2), *slightly worse* (-1), *about the same* (0), *slightly better* (1), *better* (2), and *much better* (3).

Test samples were played to both ears with Sennheiser HD 580 headphones. The listening environment was a room with low background noise level, and only the listener was present in the room during the test.

The sample pairs were graded by the listeners using a graphical user interface based on the GuineaPig software [9]. Each sample pair could be repeated an unlimited number of times. The comparison score was given to each sample pair with a slider bar component on the screen.

For each test sentence, all processing pairs were included in both presentation orders. With 5 processing types, 8 sentences, and 2 presentation orders, this resulted in a total of 160 pair comparisons. The experiment also included 20 null pairs, i.e., sample pairs with two identical samples. The sequence of test items was randomized separately for every listener.

Each listener had a short practice session before starting the actual test. Subjects were also instructed to adjust the volume setting to a suitable level during the practice session and not the change this setting later. The test itself was divided into three parts with 60 sentence pairs in each. Short breaks of a few minutes were held between the sessions. Finally, listeners were asked to comment on the samples and differences in the sample pairs after the test. The duration of the entire listening test per participant was approximately one hour, including the practice session, test sessions, breaks, and the short interview at the end.

3.3. Listeners

Native speakers of the three languages were recruited to the listening tests. Only listeners with normal hearing were allowed to participate. A reward of 20 euros was paid to each participant. The number of listeners was 13 (5 females), 21 (8 females), and 19 (9 females) for American English, Russian, and Mandarin Chinese, respectively.

4. Results

The results of the listening test were analyzed in two ways: the preference order of all examined processings was calculated, and pairwise comparisons of the processings were studied.

4.1. Order of preference

An average score for each processing was computed from all comparisons in which the processing was involved (except for null pairs). The presentation order was taken into account. For example, if a score of -2 was given in the comparison "wideband vs. narrowband", then the score 2 was included in the calculation of the mean score for "wideband", and -2 was used for the mean score of "narrowband". This method yields the order of superiority and distances between the processings, but the explanation of the numerical scale does not correspond to that of the CCR test.

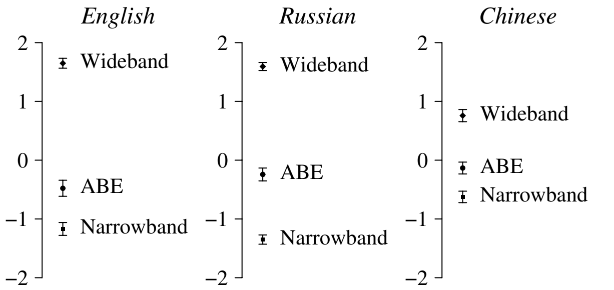

Figure 3: *The order of preference of the processings in the three languages. Mean scores and 95 % confidence intervals are shown.*

A 95 % confidence interval for each mean score was calculated based on the Student's t-distribution as follows:

$$\mathrm{CI}_{95} = t_{N-1,0.05}\frac{S}{\sqrt{N}} \tag{2}$$

where N is the number of scores used for calculating the mean, S is the standard deviation of the scores, and $t_{N-1,0.05}$ is the inverse of the t-distribution with $N-1$ degrees of freedom and probability 0.05.

The mean scores of narrowband, wideband, and ABE processing are shown in Figure 3 for all three languages. The illustration also includes 95 % confidence intervals. The order of preference was found to be the same in all three languages: Wideband samples were rated the best and narrowband samples the worst. ABE-processed samples received a mean score between these extremes but closer to the mean score of the narrowband samples.

4.2. Pairwise comparisons

The processings were also compared pairwise by calculating the comparison mean opinion score (CMOS) from the comparisons between each pair of processings. Before calculating the mean, the scores were recoded so that the order of presentation was normalized and half of the scores were inverted correspondingly. The explanations of the numerical values of the mean scores can be directly taken from the CCR grading scale. A 95 % confidence interval for each CMOS was calculated using the formula (2).

The comparisons between narrowband, wideband, and ABE processings are illustrated in Figure 4. The bars indicate the relative frequencies of each score. The presentation order of each sample pair was normalized so that the scores correspond to the order shown in the title of each illustration. For example in the comparison "Narrowband vs. Wideband", the bar at "2" shows the relative number of listener responses in which a wideband sample was rated "better" in comparison to the corresponding narrowband sample. Mean scores and 95 % confidence intervals are shown on the horizontal score axes.

Two-tailed T-tests were computed to see if the comparisons showed a statistically significant preference to either direction. The results of the T-tests were given as p values indicating the probability that the result could have occurred by chance. The p values in the comparisons between narrowband, wideband, and ABE processings were much lower than 0.001 in all three languages. Thus, the obtained differences are statistically significant.

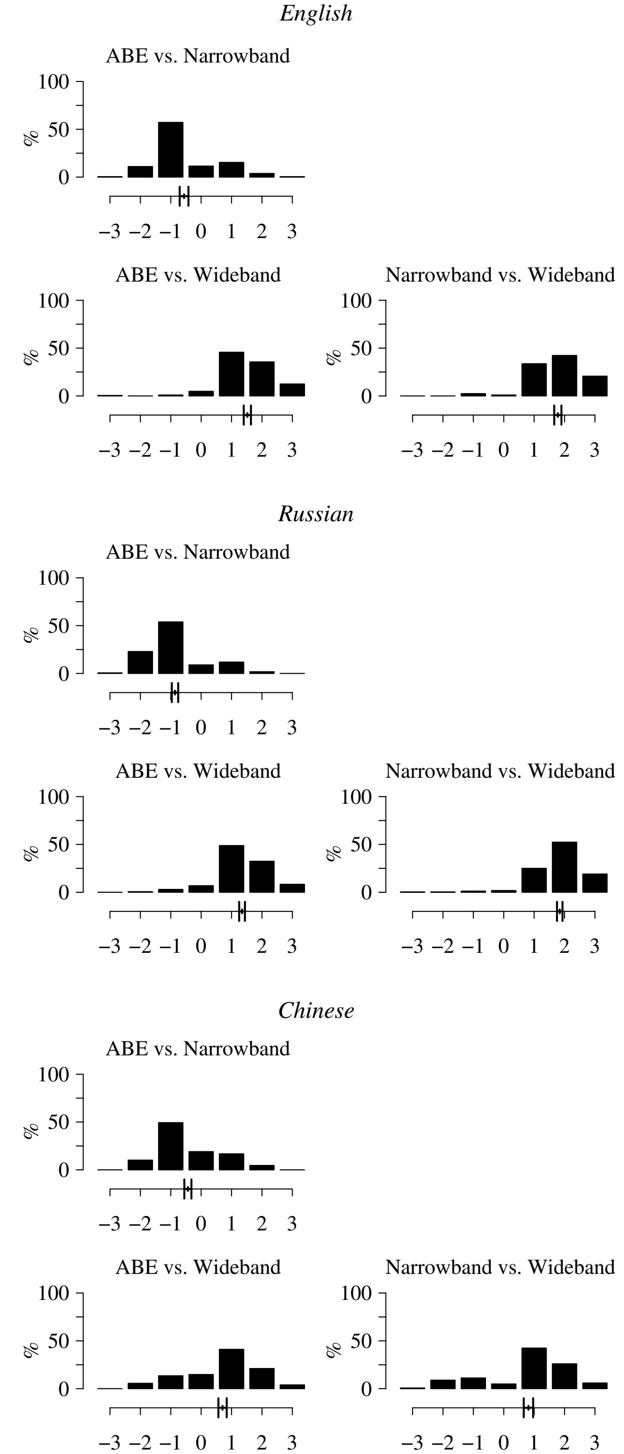

Figure 4: *Results of pairwise comparisons of narrowband, wideband, and ABE-processed samples in American English (top), Russian (middle), and Mandarin Chinese (bottom). The horizontal axis denotes the score given to the second processing compared to the first processing. The range is from much worse (-3) to much better (3). The bars indicate relative frequencies of the scores. The mean score and its 95 % confidence interval are shown on the horizontal axis.*

5. Conclusions

The performance of an artificial bandwidth expansion method was evaluated by formal listening tests in three widely spoken languages: American English, Russian, and Mandarin Chinese. Russian has a varied set of fricatives, both voiced and unvoiced, which are challenging for artificial bandwidth expansion. Mandarin Chinese, in turn, is a tonal language. Despite these challenges, promising results were obtained. The quality of narrowband speech was significantly improved by the ABE processing for all the tested languages.

To the best of our knowledge, the present study is the first one to evaluate artificial bandwidth expansion with formal listening tests in different languages. Furthermore, methods of artificial bandwidth expansion have not typically been tested with low bit-rate compressed speech. In the present study, however, ABE processing was applied to narrowband speech coded with AMR-NB, widely used today in mobile communications. The computational complexity and memory requirements of the evaluated ABE algorithm are also reasonable. Therefore, the method is feasible to be implemented in real-time applications.

6. References

[1] 3rd Generation Partnership Project, *Technical Specification 3GPP TS 26.171, Speech Codec speech processing functions; AMR Wideband Speech Codec; General Description*, 2001, Release 5.

[2] Fant, G., *Acoustic Theory of Speech Production*, Mouton, The Hague, 1960.

[3] Jax, P., *Enhancement of Bandlimited Speech Signals: Angorithms and Theoretical Bounds*, Ph.D. thesis, Rheinisch-Westflische Technische Hochschule Aachen, Aachen, Germany, Oct. 2002.

[4] Laaksonen, L., Kontio, J., and Alku, P., "Artificial bandwidth expansion method to improve intelligibility and quality of AMR-coded narrowband speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA, May 2005, vol. 1, pp. 809–812.

[5] Makhoul, J. and Berouti, M., "High-frequency regeneration in speech coding systems," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 1979, vol. 4, pp. 428–431.

[6] Pohlheim, H., *GEATbx: Genetic and Evolutionary Algorithm Toolbox for use with MATLAB, Documentation*, Nov. 2005, http://www.geatbx.com/docu/index.html.

[7] NTT Advanced Technology Corporation, "Multi-lingual speech database for telephonometry 1994," http://www.ntt-at.com/products_e/speech/index.html.

[8] International Telecommunication Union, *ITU-T Recommendation P.800, Methods for subjective determination of transmission quality*, 1996.

[9] Hynninen, J. and Zacharov, N., "GuineaPig – a generic subjective test system for multichannel audio," in *Proceedings of the 106th AES Convention*. Audio Engineering Society, 1999, preprint 4871.