



# The Effect of Highband Harmonic Structure in the Artificial Bandwidth Expansion of Telephone Speech

Hannu Pulakka<sup>1</sup>, Paavo Alku<sup>1</sup>, Laura Laaksonen<sup>2</sup>, Päivi Valve<sup>2</sup>

<sup>1</sup>Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Finland

<sup>2</sup>Nokia Technology Platforms, Finland

hannu.pulakka@tkk.fi, paavo.alku@tkk.fi, laura.laaksonen@nokia.com, paivi.valve@nokia.com

## Abstract

The quality of narrowband telephone speech can be improved by *artificial bandwidth expansion* (ABE), which generates missing frequency components above the telephone bandwidth using only information from the narrowband speech signal. Straightforward bandwidth expansion methods do not reproduce the harmonic structure of voiced sounds properly, but a pitch-adaptive technique can be used to approximate the correct alignment of harmonic frequencies. In this study, pitch-adaptive highband alignment was implemented into an existing ABE method, and the quality of the modified method was studied with formal listening tests in Finnish and Mandarin Chinese. The effect of the highband harmonic structure was found unimportant for the perceived speech quality. Consequently, computationally expensive pitch adaptation was found to be unnecessary for the bandwidth expansion of telephone speech.

**Index Terms:** bandwidth expansion, listening test, harmonic structure

## 1. Introduction

In most of the current telephone systems, including the PSTN and the GSM cellular network, the audio bandwidth of transmitted speech is limited to approximately the band from 0.3 to 3.4 kHz. This bandwidth limitation degrades speech quality and intelligibility because speech contains frequency components far beyond this frequency range. Standards have been developed for speech transmission with a wider audio bandwidth, and wideband coders are likely to become commonplace in the future. For example, the AMR-WB codec with an audio bandwidth of 50–7000 Hz has been selected for the third generation mobile communication system. However, the transition from the narrowband systems of today to wideband systems will probably take a long time. Consequently, narrowband systems are not likely to disappear in the near future.

To improve the quality of narrowband telephone speech and to reduce the quality gap between wideband and narrowband telephone calls, a method called *artificial bandwidth expansion* (ABE) has been developed. The method attempts to reproduce frequency components between 4 and 8 kHz using only information available in the narrowband speech signal. In this paper, *lowband* refers to the frequency band below 4 kHz, and *highband* to the band between 4 and 8 kHz.

Artificial bandwidth expansion of speech has been studied by various authors, e.g. [1], [2], [3]. A common approach is based on the source-filter model of speech production and applies bandwidth expansion separately to the excitation and the spectral envelope. Codebooks are often utilized for the expansion of the spectral envelope, whereas the extended excitation is

typically generated from the narrowband excitation signal using e.g. translation by modulation or spectral folding, see e.g. [2].

The true wideband spectrum of voiced speech contains harmonic peaks at regular intervals, but straightforward methods for the artificial generation of the highband do not reproduce the harmonic structure properly. Peaks in the highband spectrum are not located correctly at harmonics of the fundamental frequency, which has been reported to cause a metallic sound in the resulting speech signal [2]. To regenerate a realistic series of harmonics in the highband, pitch-adaptive spectral shifting or modulation techniques have been developed [4], [1], [2]. These methods estimate the fundamental frequency of speech at each instant of time and construct the excitation in the highband from the lowband by appropriate spectral shifting such that harmonic peaks are located at integer multiples of the fundamental frequency. This requires accurate estimation of the fundamental frequency. The technique has been reported to result in more natural speech quality [2]. However, inconsistent harmonic structure at high frequencies has been found to have a small effect on the subjective quality of bandwidth-expanded speech [1], [2], [3]. Consequently, modulation with a fixed frequency has been considered a reasonable compromise [1], [2].

A robust bandwidth expansion method with moderate computational cost was introduced by [5], and it was shown to perform consistently in various languages [6]. Thus, the method shows potential to be implemented in real-time applications for mobile communications. A modification of this scheme utilizing neural networks was published recently [7]. These algorithms do not explicitly separate the excitation of speech from the spectral envelope. Instead, they use spectral folding directly to generate an initial expansion to the highband and then modify the magnitude spectrum of the highband with a smooth shaping curve. The spectral folding technique is, however, known to produce a spectrum that is different from the correct harmonic structure for voiced sounds. Therefore, this study examines the effect of pitch-adaptive correction of the highband harmonic structure in the method proposed by [5], which is referred to as the ABE method in this paper. The performance of highband correction is evaluated by listening tests in Finnish and Mandarin Chinese. Similar pitch-adaptive techniques have been proposed earlier for other bandwidth expansion approaches, such as in [1] and [2], but neither of these reported testing in different languages or in a tonal language such as Chinese.

## 2. Method

The ABE method used in this study is first briefly described in this section. Then, the implementation of a pitch-adaptive highband alignment into this system is explained.

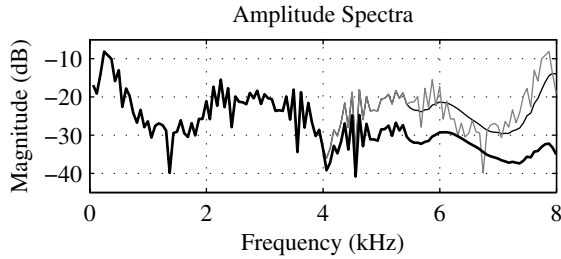


Figure 1: *Amplitude spectrum of /i/ spoken by a male speaker. Gray curve is the folded spectrum, thin black curve denotes the spectrum after smoothing the harmonic structure, and bold black curve shows the final spectrum given by ABE processing.*

### 2.1. Bandwidth expansion algorithm

The ABE algorithm described in [5] and [6] is used as the baseline implementation of bandwidth expansion in this study. The input signal is processed in short time domain frames. New initial frequency components are created through spectral folding [4], which is implemented in the time domain by zero-insertion. As a result, the sampling rate of the signal is doubled from 8 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 smoothed in the frequency domain between 5.5 and 8 kHz to mimic the less prominent harmonic structure of natural speech at high frequencies. Each frame is classified into one of three categories: voiced sounds, sibilants, and plosives. The classification is based on a vector of a few time-domain and frequency-domain features extracted from the original narrowband signal. The highband magnitude spectrum is then modified with a shaping function that is a cubic spline curve constructed around five control points. The control point values depend on the narrowband slope feature and predefined constants, which are different for each frame category. An additional noise dependent gain is also added to the shaping curve. Finally, the artificial wideband spectrum is converted back to the time domain through an inverse FFT. A frequency-domain example of a speech segment processed with the algorithm is shown in Figure 1.

### 2.2. Pitch-adaptive modulation

In this study, the ABE algorithm described in section 2.1 was modified to include the alignment of the harmonic structure in the highband. The modified algorithm is referred to as the pitch-adaptive artificial bandwidth expansion, or PA-ABE. Because the goal of the study was to examine the effect of the alignment of the harmonic peaks in the highband spectrum, only minimal changes were made to the ABE method to avoid affecting the output signal in other ways. This is also the reason why the chosen pitch-adaptive approach approximates spectral folding instead of implementing spectral shift from the lowband to the highband, as described by [1] and [2].

The spectral folding and FFT steps of the algorithm were replaced by a more complicated processing that involves pitch-adaptive modulation. The modified scheme produces an initial expansion of the narrowband speech to a wideband signal such that the harmonic peaks of the highband are located at integer multiples of the fundamental frequency.

First, the fundamental frequency contour  $f_0(t)$  of the narrowband speech signal is estimated using the YIN algorithm [8]. The accepted range of fundamental frequency estimates is limited between 80 and 500 Hz, and unreliable estimates are

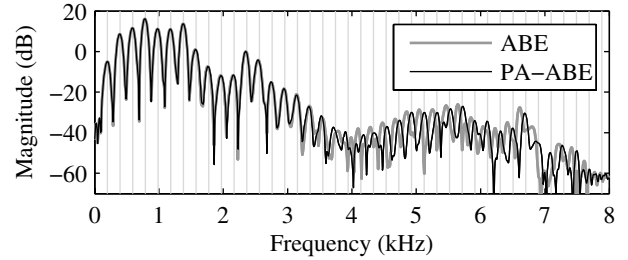


Figure 2: *The spectrum of a segment of voiced speech processed with ABE and PA-ABE. Gray vertical lines denote integer multiples of the fundamental frequency (197 Hz). The peaks of the spectrum processed with PA-ABE coincide with multiples of the fundamental, unlike the peaks of the ABE spectrum.*

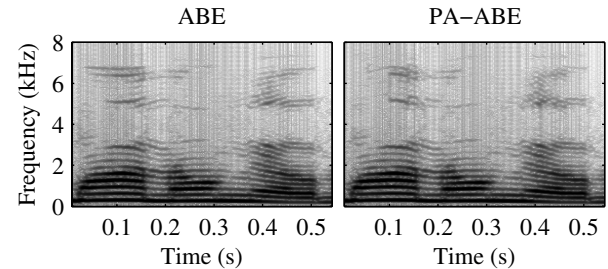


Figure 3: *Spectrograms of a voiced speech segment with changing fundamental frequency processed with ABE (left) and PA-ABE (right). In the PA-ABE processing, harmonic peaks in the highband follow the changes of the fundamental frequency.*

replaced by the last sufficiently reliable estimate.

A modulation frequency contour  $f_m(t)$  is then constructed such that  $7900 \text{ Hz} \leq f_m(t) \leq 8400 \text{ Hz}$  and the modulation frequency is an integer multiple of the fundamental frequency:  $f_m(t) = m(t) \cdot f_0(t)$ . This implies that the harmonics of the fundamental are located at integer multiples of the fundamental frequency also after modulation. The value of  $m(t)$  is changed only when necessary to keep  $f_m(t)$  within the allowed range.

The narrowband signal is upsampled to 16 kHz and lowpass filtered with the cutoff at 4 kHz. This lowband signal is modulated with a complex exponential having the frequency  $f_m(t)$  to construct a complex highband signal, which is then added to the lowband signal. The sum signal is windowed using the framing scheme of the ABE method and transformed to the frequency domain with FFT. Finally, the frame spectra are modified to represent a real-valued signal by replacing frequency bins above the Nyquist frequency by complex conjugates of the corresponding frequency bins below the Nyquist frequency. The obtained frequency-domain representation is then used as an initial expansion of the narrowband signal in the rest of the ABE algorithm.

Figure 2 shows the spectrum of a short segment of voiced speech processed using the original ABE method and PA-ABE. Spectrograms in Figure 3 illustrate the behavior of highband harmonic peaks in a longer voiced speech segment with changing fundamental frequency.

## 3. Experiments

Formal listening tests were arranged to evaluate the difference between ABE and PA-ABE. A Comparison Category Rating (CCR) test [9] was organized to assess the quality difference.

Because the audible difference between the two methods was found to be small in informal listening tests, a formal ABX test [10] was also conducted. Both tests were arranged in two languages: Finnish and Mandarin Chinese. Finnish was selected because the ABE method was originally developed for Finnish and because of the availability of native Finnish listeners. Mandarin Chinese was chosen because it is a tonal language, which suggests that native listeners might be more sensitive to differences in the harmonic structure.

Twelve native speakers of Finnish (four females) and twelve native speakers of Mandarin Chinese (seven females) participated in the listening tests. The ages of the listeners were between 19 and 36 years. All listeners reported normal hearing but this was not verified with audiometric tests. A reward of 20 euros was paid to each participant.

Speech material was obtained from the NTT database [11]. Six sentences spoken by different speakers (3 females and 3 males) were chosen randomly in both languages. The duration of each sentence was approximately two seconds. Pre-recorded office noise was added to the speech signals so that the signal-to-noise ratio (SNR) was 35 dB, and the speech samples were filtered with a model of the input characteristics of a GSM mobile station. The samples were then processed by the following processings:

- Narrowband reference: 2×AMR-NB (12.2 kbps)
- ABE: 2×AMR-NB followed by ABE
- PA-ABE: 2×AMR-NB followed by PA-ABE
- Wideband reference: 2×AMR-WB (12.65 kbps)

Finally, the samples were filtered with an estimated response of a wideband mobile terminal. This estimate was improved from the one used previously in [6].

### 3.1. CCR test

The Comparison Category Rating (CCR) method was used to compare the perceived quality of ABE and PA-ABE as well as narrowband and wideband reference samples.

Listeners compared the speech samples pairwise so that each test item contained two instances of the same sentence. The task of the subject was to evaluate the quality of the second sentence compared to the quality of the first sentence on a seven-point scale: *much worse* (-3), *worse* (-2), *slightly worse* (-1), *about the same* (0), *slightly better* (1), *better* (2), and *much better* (3). Listeners were allowed to repeat the sample pairs as many times as they wanted. The samples were played to both ears with Sennheiser HD 580 headphones.

For each test sentence, all combinations of the four processings were presented in both orders. The experiment also included 8 null pairs, i.e., sample pairs with two identical samples. This resulted in a total of 80 comparisons. The sequence of test items was randomized for each listener.

### 3.2. ABX test

The ABX test method was used to test whether the difference between ABE and PA-ABE processings is audible. In an ABX test the listener is presented with three samples denoted by A, B, and X. The samples A and B are different, whereas X is exactly the same as either A or B. The task of the listener is to determine if X is A or B.

The ABE and PA-ABE samples from the CCR test were used as test material. The ABX test comprised 24 trials. The listeners could play the three samples in any order and any number of times before choosing either A or B as the answer. The same

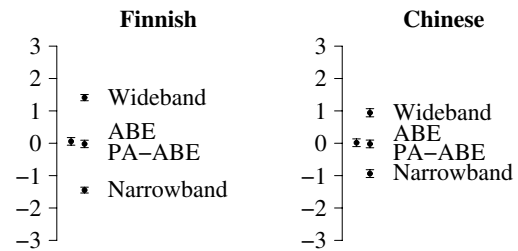


Figure 4: The order of preference of the processings in Finnish and Chinese. Mean scores and 95 % confidence intervals are shown. The average score of ABE is slightly higher than that of PA-ABE in both languages.

listeners participated in the CCR and ABX tests. The CCR test was performed first, and the ABX test was then started after a short break.

## 4. Results

### 4.1. CCR test

The mean score for each processing was calculated from all comparisons in which the processing was involved, except for the null pairs. The presentation order of the sample pairs was taken into account so that the opposite of the given score was used whenever the processing in question was presented as the first sentence. This method yields the order of superiority and distances between the processings.

The mean scores and confidence intervals are shown in Figure 4. The same order of preference was obtained in both languages: Wideband samples were rated the best and narrowband samples the worst. The mean scores of ABE and PA-ABE are between these extremes. No significant difference was found between ABE and PA-ABE in either language.

Pairwise comparisons between the processings are illustrated in Figure 5. The bars indicate the relative frequencies of each score. The presentation order has been normalized according to the title of each illustration. For example, the bar at score 3 refers to responses in which the latter processing in the title was considered *much better*, or the first processing was graded *much worse*. The comparison mean opinion scores (CMOS) and 95 % confidence intervals are shown on the horizontal axes.

Two-tailed t-tests were performed to examine if the pairwise comparisons showed a statistically significant preference to either direction. The  $p$  values are shown in Figure 5. In Finnish, all other pair comparisons yielded a significant difference except for the comparison between ABE and PA-ABE ( $p = 0.37$ ). In Chinese, also the case ABE vs. PA-ABE is significant ( $p = 0.026$ ) with the chosen significance level of 5 %.

Thus, the CCR test indicates no quality difference between ABE and PA-ABE in Finnish. In Chinese, the difference between ABE and PA-ABE is small but statistically significant. The rest of the pairwise comparisons confirm the preference order shown in Figure 4.

### 4.2. ABX test

The proportion of correct identifications in Finnish was  $169/288 = 58.7\%$ . The  $p$  value, which indicates the probability of reaching this or higher number of correct answers by chance, was 0.0019. In Chinese, the fraction of correct responses was  $146/288 = 50.7\%$  ( $p = 0.43$ ). Thus, the result was significantly different from random selection in Finnish but not in Chinese.

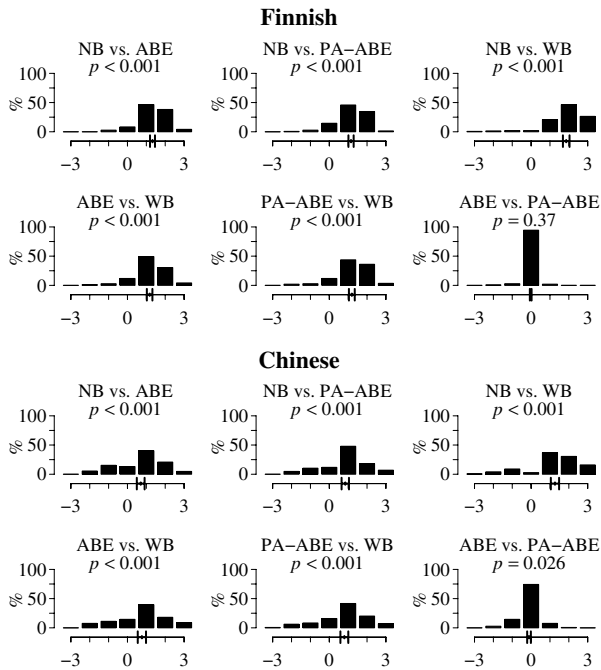


Figure 5: Results of pairwise comparisons in Finnish (top) and in Chinese (bottom) between narrowband (NB), ABE, PA-ABE, and wideband (WB). The horizontal axis denotes the score given to the second processing compared to the first processing. The bars indicate relative frequencies of the scores. The mean score with a 95 % confidence interval is shown on the horizontal axis. The  $p$  values indicate if there is a preference to either direction.

## 5. Conclusions

An artificial bandwidth expansion method, ABE, which is based on spectral folding and has been designed to be applicable for real-time implementations, was modified in order to correctly align the highband harmonic structure. The performance of this modified method, called PA-ABE, was compared with the original ABE implementation by means of CCR and ABX listening tests in Finnish and in Mandarin Chinese. The CCR test in the Finnish language did not show any significant quality difference between ABE and PA-ABE. In the ABX test, Finnish listeners were able to choose the similar sample in less than 60 % of the cases, which is statistically significant but also indicates that the difference is very hard to detect. In Chinese, the CCR test showed a minor but statistically significant difference between ABE and PA-ABE but the ABX test with the same speech samples and listeners failed to prove that a difference could be heard between the processings.

ABE and PA-ABE were found to be practically identical in quality both in Finnish and in Mandarin Chinese. The additional effort required for the adaptive alignment of harmonic peaks in the highband spectrum does not yield such an improvement in quality that would justify the increased complexity. This result is important from the practical point of view. The computational cost and memory requirements of the ABE algorithm are relatively low, and the method is therefore applicable for real-time implementation in devices with limited computational resources. The correction of highband harmonic structure, which would have made the algorithm much more complicated, was now found to be unimportant for the perceived quality of ABE-processed speech.

The current study also replicates the results of our previ-

ous research [6], in which the quality of artificial bandwidth expansion was compared with coded narrowband and wideband references in three languages. The test setting was similar in both studies, but the speech samples, majority of the listeners, and some details of the ABE implementations were different. Both studies indicate that ABE-processed samples are consistently preferred over narrowband samples. The quality of true wideband speech is not reached, but the quality gap between narrowband and wideband speech is reduced significantly. The same finding has now been obtained with formal listening tests in four different languages.

## 6. Acknowledgements

Hannu Pulakka is supported by the Graduate School in Electronics, Telecommunications and Automation (GETA).

## 7. References

- [1] J. A. Fuemmeler, R. C. Hardie, and W. R. Gardner, "Techniques for the regeneration of wideband speech from narrowband speech," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 266–274, 2001.
- [2] P. Jax, "Enhancement of bandlimited speech signals: Algorithms and theoretical bounds," Ph.D. dissertation, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany, Oct. 2002.
- [3] U. Kornagel, "Techniques for artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 86, pp. 1296–1306, 2006.
- [4] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, Apr. 1979, pp. 428–431.
- [5] L. Laaksonen, J. Kontio, and P. Alku, "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)*, vol. 1, Philadelphia, USA, May 2005, pp. 809–812.
- [6] H. Pulakka, L. Laaksonen, and P. Alku, "Quality improvement of telephone speech by artificial bandwidth expansion – listening tests in three languages," in *Proceedings of Interspeech 2006*, Pittsburgh, Pennsylvania, USA, September 2006, pp. 1419–1422.
- [7] J. Kontio, L. Laaksonen, and P. Alku, "Neural network-based artificial bandwidth expansion of speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 873–881, March 2007.
- [8] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, April 2002.
- [9] *ITU-T Recommendation P.800, Methods for subjective determination of transmission quality*, International Telecommunication Union, 1996.
- [10] D. Clark, "High-resolution subjective testing using a double-blind comparator," *Journal of the Audio Engineering Society*, vol. 30, no. 5, pp. 330–338, May 1982.
- [11] NTT Advanced Technology Corporation, "Multilingual speech database for telephony 1994," [http://www.ntt-at.com/products\\_e/speech/index.html](http://www.ntt-at.com/products_e/speech/index.html).