Artificial Bandwidth Extension of Wideband Speech by Pitch-Scaling of Higher Frequencies

Bernd Geiser and Peter Vary

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University
{geiser|vary}@ind.rwth-aachen.de

Abstract: In this paper, a simple DFT-domain pitch-scaling technique is used to extend the audio bandwidth of wideband speech (50 Hz – 7 kHz) to the super-wideband range (50 Hz – 12 kHz). Therefore, the higher frequencies of the wideband signal (6 – 7 kHz) are pitch-scaled with a scaling factor of four and the resulting, scaled signal is inserted into the 8 – 12 kHz band. A subjective listening test has been conducted wherein it could be shown that the new proposal clearly outperforms a previous method for artificial bandwidth extension which is based on statistical estimation techniques.

1 Introduction

Techniques for artificial bandwidth extension (ABWE) of speech aim to improve the quality of *band-limited* speech signals by *artificially* extending the reproduced audio frequency range without making use of any additional side information. The extension can be performed toward the lower and upper end of the audible frequency range. This paper is concerned with the extension toward higher audio frequencies.

ABWE algorithms are particularly interesting for realtime audio communication systems which are based on (possibly older) transmission standards or speech codecs that only support a limited audio bandwidth. In such systems, as no further side information is available, ABWE is usually conducted directly within the *receiving* terminal so that no other network components need to be modified. However, for the extension of typical *telephone speech* (narrowband, NB, 300 Hz – 3.4 kHz) toward the *wideband* (WB) frequency range (50 Hz – 7 kHz), it could be shown that the achievable WB speech quality is limited [NGAK02, JV02]. In this case, other approaches such as *embedded coding* or *steganographic transmission of side information* should be preferred if possible, see [Gei12a, Gei12b] for an overview.

In contrast to the extension of NB speech, much less attention has been paid to the extension of WB speech toward the *super-wideband* bandwidth (SWB, e.g., 50 Hz – 12 kHz). In this scenario much more consistent quality gains can be expected. Therefore, in this paper, a new ABWE approach is proposed which is particularly suited for WB-to-SWB bandwidth extension. It is based on DFT domain pitch scaling techniques and yields clearly better results than previous ABWE methods.

1.1 Outline

After a brief review of previous ABWE methods in Section 2, the proposed, novel ABWE approach, which is based on DFT domain pitch scaling techniques, is described in detail (Section 3). The new system is evaluated and the achieved super-wideband speech quality is compared with a previous method as well as other references (Section 4). The paper is then concluded with Section 5.

2 Previous ABWE Methods

Most modern approaches are based on a parametric representation of the higher frequency band and treat the ABWE task as a classification or statistical estimation problem. The respective algorithms usually follow these four consecutive steps:

- A relevant *feature vector* $\mathbf{x}_\mathrm{f}$ is computed from time frames of the received "baseband" signal $s(k)$ or, if possible, directly from the bitstream of the codec that is used to transmit this signal.
- With the help of a pre-trained statistical model or a specific classification rule, a parameter vector $\mathbf{p}$ for the current time frame is estimated (or derived) from the current (and possibly previous) feature vectors.
- An "excitation" signal for the higher frequencies is generated, often by spectral folding of a spectrally whitened version of the lower frequency band.
- The synthetic high band signal is produced by a suitable high band synthesis algorithm based on the generated "excitation" signal and on the estimated/derived parameter set $\mathbf{p}$.

To establish the connection between the baseband feature vector and the high band parameter set, numerous classification approaches as well as statistical estimation algorithms have been proposed in the literature, e.g., the mapping of the entries of a narrowband codebook to a wideband shadow codebook [CH94], (piecewise) linear mapping [NTN97, CGMS01], artificial neural networks, [KLA07, PA11], and Minimum-mean-square-error (MMSE) estimation based on Gaussian Mixture Models (GMMs) [PK00] or Hidden Markov Models (HMMs) [Jax02, JV03].

In contrast to these classication or estimation based methods, there are several, mostly older approaches that directly manipulate the received baseband signal to regenerate the higher frequencies. For example, the appliction of non-linearities to produce additional harmonic signal content dates back to the year 1933 [Sch33]. Aliasing components which are due to digital-to-analog conversion or digital sampling rate conversion have been exploited to regenerate higher audio frequencies in [Die84, Yas95]. The method of [PX81] operates in the frequency domain, shifts frequency coefficients toward higher frequencies and applies a scaling factor.

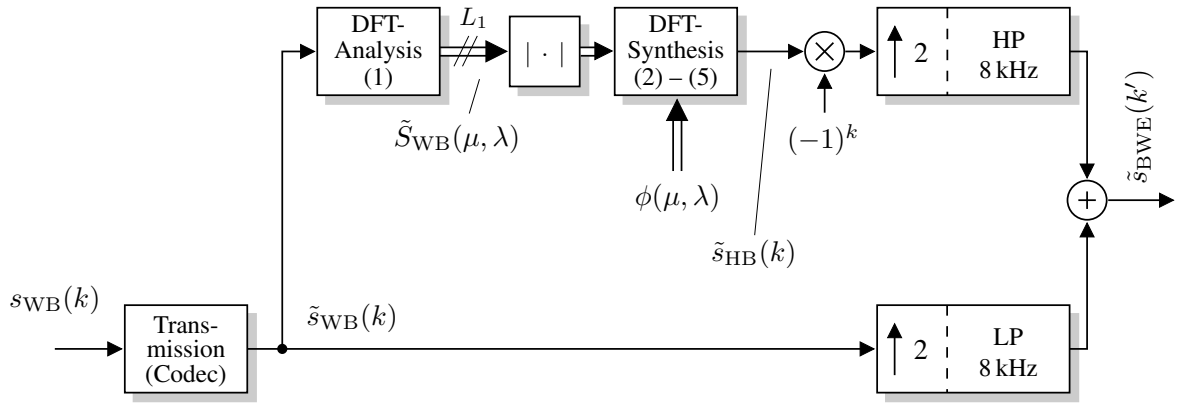

Figure 1: System overview (bracketed numbers reference the respective equations in the text).

The novel ABWE method to be described in the following is also based on a comparatively simple, direct manipulation of DFT coefficients of the baseband signal rather than relying on elaborate and complex statistical estimation techniques.

3 ABWE by Pitch Scaling of Higher Frequencies

In [Gei13], we have proposed a system for backwards-compatible in-band transmission of higher speech frequencies over a narrowband telephone connection. The basic idea was to insert a pitch-scaled version of the higher frequencies (4 – 6.4 kHz in this paper) into the previously "unused" 3.4 – 4 kHz frequency range of standard telephone speech which corresponds to a down-scaling factor of $\rho = (4 - 3.4)/(6.4 - 4) = \frac{1}{4}$. The down-scaling operation is reverted at the decoder side (up-scaling factor $1/\rho = 4$). Of the numerous pitch-scaling methods which are available, cf. [Zö11], a comparatively simple DFT-domain technique turned out to be well-suited for our purposes, because, in this case, the pitch scaling and the required frequency domain insertion/extraction operations can be carried out within the same signal processing framework.

In this work, we will reuse the respective *decoder algorithm* to realize a system for ABWE of *wideband* speech toward the *super-wideband* bandwidth. A block diagram is depicted in Figure 1.

3.1 Algorithm Overview

First, time frames of the wideband signal $\tilde{s}_{\mathrm{WB}}(k)$, sampled at $f_{\mathrm{s}} = 16$ kHz, are transformed into the frequency domain using a windowed DFT. From the resulting spectral coefficients, the high band information, corresponding to the 6 – 7 kHz frequency range,

is extracted[1]. These passband coefficients are further processed and reinterpreted as the DFT coefficients of the high band signal $\tilde{s}_{\mathrm{HB}}(k)$ which is synthesized by inverse DFT and an overlap-add procedure. The regenerated high band signal is finally combined with the original wideband signal to form the bandwidth extended output signal $\tilde{s}_{\mathrm{BWE}}(k')$ at its sampling rate of $f'_{\mathrm{s}} = 32\,\mathrm{kHz}$.

3.2 Analysis of the Wideband Speech Signal

To realize the pitch scaling (or, equivalently, time scaling) operation, the ratio of the window lengths of the employed IDFT and DFT operations must be equal to the desired scaling factor ρ. In our current implementation we have chosen values of $L_1 = 128$ and $L_2 = 32$, i.e., $\rho = L_2/L_1 = 1/4$. The DFT analysis of the wideband signal $\tilde{s}_{\mathrm{WB}}(k)$ can thus be written as:

$$\tilde{S}_{\mathrm{WB}}(\mu, \lambda) = \sum_{k=0}^{L_1-1} \tilde{s}_{\mathrm{WB}}(k + \lambda S) w_{L_1}(k) \cdot \mathrm{e}^{-2\pi j \frac{k\mu}{L_1}} \tag{1}$$

for frequency bins $\mu \in \{0, \ldots, L_1 - 1\}$ and frame index λ. The window function $w_{L_1}(k)$ has been chosen as the square root of a Hann window of length L_1. The window shift is $S = 8$. Note that the same window shift S must be used as in the IDFT synthesis procedure, see Section 3.4.

3.3 Composition of the High Band Spectrum

The high band information (DFT magnitudes for 6 – 7 kHz) within the spectrum $\tilde{S}_{\mathrm{WB}}(\mu, \lambda)$ is now extracted and a (partly) synthetic DFT spectrum with L_2 bins is formed. For notational convenience, the frame index λ and the (implicit) complex conjugate symmetric extension for $\mu > \frac{L_2}{2}$ are disregarded.

With $\mu_0 = \frac{L_1 - \lceil 2/8 \cdot L_1 \rceil}{2}$ and $\mu_1 = \frac{L_1 - \lceil 1/8 \cdot L_1 \rceil}{2}$, the DFT magnitudes of the newly formed high band spectrum can therefore be written as:

$$\left|\tilde{S}_{\mathrm{HB}}(\mu)\right| = \begin{cases} \left|\tilde{S}_{\mathrm{WB}}(\mu + \mu_0)\right| & \text{for } 0 \le \mu < \mu_1 - \mu_0 \\ 0 & \text{for } \mu \ge \mu_1 - \mu_0, \end{cases} \tag{2}$$

whereby the bin with index $\mu_1 - \mu_0$ corresponds to a frequency of 4 kHz in the high band signal, i.e., 12 kHz in the super-wideband domain.

Compared to the DFT magnitudes, a correct representation of the *phase* is much less important for high-quality reproduction of higher speech frequencies, cf. [JV03]. In fact,

[1] The 6 – 7 kHz range has been chosen instead of the 7 – 8 kHz band because wideband speech is commonly lowpass filtered at 7 kHz. Common wideband codecs also employ such lowpass filtering.

there are several alternatives to obtain a suitable phase $\angle \tilde{S}_{\mathrm{HB}}(\mu)$. For example, an *additional* analysis of $\tilde{s}_{\mathrm{WB}}(k)$ with a window length of L_2 and a window shift of S would facilitate the direct reuse of the *narrowband* phase, an approach which is often used in artificial bandwidth extension algorithms, e.g., [JV03].

In fact, for our application, even a random phase $\phi(\mu) \sim \mathrm{Unif}(-\pi, \pi)$ already delivered a high speech quality, i.e.:

$$\angle \tilde{S}_{\mathrm{HB}}(\mu) = \begin{cases} \angle \mathrm{Re}\{\tilde{S}_{\mathrm{HB}}(\mu_0)\} & \text{for } \mu = 0 \\ 0 & \text{for } \mu = \frac{L_2}{2} \\ \phi(\mu) & \text{else.} \end{cases} \tag{3}$$

3.4 Super-Wideband Speech Synthesis

The regenerated high band DFT spectrum $\tilde{S}_{\mathrm{HB}}(\mu, \lambda)$ is now transformed into the time domain via an IDFT with the short window length $L_2 = L_1 \cdot \rho$:

$$\tilde{s}_{\mathrm{HB}}(k, \lambda) = \frac{1}{L_2} \sum_{\mu=0}^{L_2-1} \tilde{S}_{\mathrm{HB}}(\mu, \lambda) \cdot \mathrm{e}^{2\pi j \frac{k\mu}{L_2}} \tag{4}$$

for $k \in \{0, \ldots, L_2 - 1\}$ and 0 outside the frame interval. Now, for overlap-add, the same window shift S as for analysis (1) is applied, i.e.:

$$\tilde{s}_{\mathrm{HB}}(k) = \sum_{\lambda} \tilde{s}_{\mathrm{HB}}(k - \lambda S, \lambda)\, w_{L_2}(k - \lambda S) \tag{5}$$

for all k. Here, a value of $S = L_2/4 = 8$ has been chosen so that the sequence of synthesis windows sums up to a constant. Half-overlapping windows, i.e., $S = L_2/2$, are possible as well. The window function $w_2(k)$ is, again, the square root of a Hann window of length L_2.

With $\tilde{s}_{\mathrm{HB}}(k)$ and the corresponding low band signal $\tilde{s}_{\mathrm{WB}}(k)$, the final subband synthesis, e.g., with a QMF synthesis filterbank, can be carried out, giving the bandwidth extended output signal $\tilde{s}_{\mathrm{BWE}}(k')$. Thereby, an optional attenuation factor can be applied to the regenerated high band signal so that small residual artifacts therein fall below the perception threshold.

4 Quality Evaluation

The super-wideband speech quality that is achieved with the proposed ABWE algorithm is evaluated in this section. First, spectrograms of an example speech signal are shown and discussed and then the results of a subjective listening test are presented.

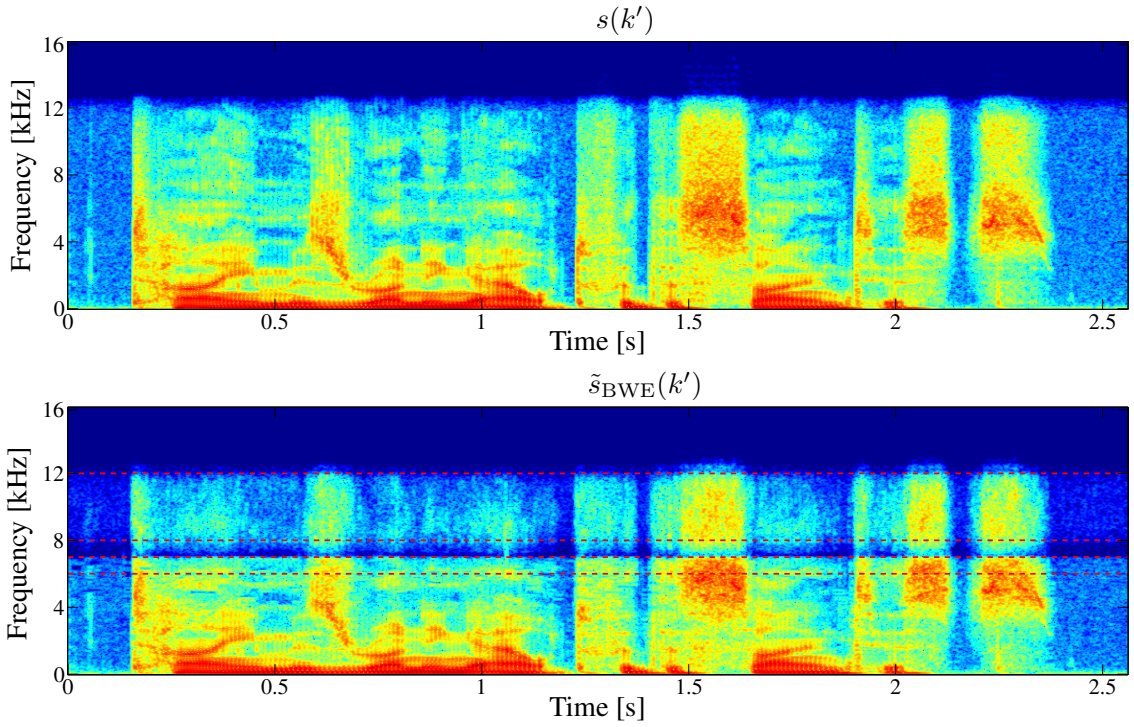

Figure 2: Top: Spectrogram of an exemplary input speech signal. Bottom: Spectrogram of the corresponding bandwidth extended signal (red lines are placed at 6, 7, 8 and 12 kHz).

4.1 Example Spectrograms

Example spectrograms of the bandwidth extended signal $\tilde{s}_{\mathrm{BWE}}(k')$ and, for comparison, of the original super-wideband signal $s(k')$ are shown in Figure 2. Note that $s(k')$ is lowpass filtered at 12 kHz.

The lower plot shows the band-limited (wideband) signal in the range 0 – 7 kHz and the regenerated high band signal in the range 8 – 12 kHz. The spectrograms reveal that the regenerated high band exhibits a virtually correct temporal energy contour. Also the coarse spectral characteristics are represented reasonably well. This is explained by observing that the characteristics of the 6 – 7 kHz (source) passband are very similar to the characteristics of the 8 – 12 kHz (target) passband. The remaining spectral gap between 7 and 8 kHz only has a very small perceptual effect which has also been found by [PLV+08] and [JV03], albeit for a stopband between 3.4 and 4 kHz.

4.2 Listening Test

To judge the super-wideband speech quality that can be obtained with our ABWE proposal, we have conducted a subjective MUSHRA listening test [ITU03] using the test software described in [SSGV11].

The test samples have been taken from the NTT corpus [NTT94], whereby only the

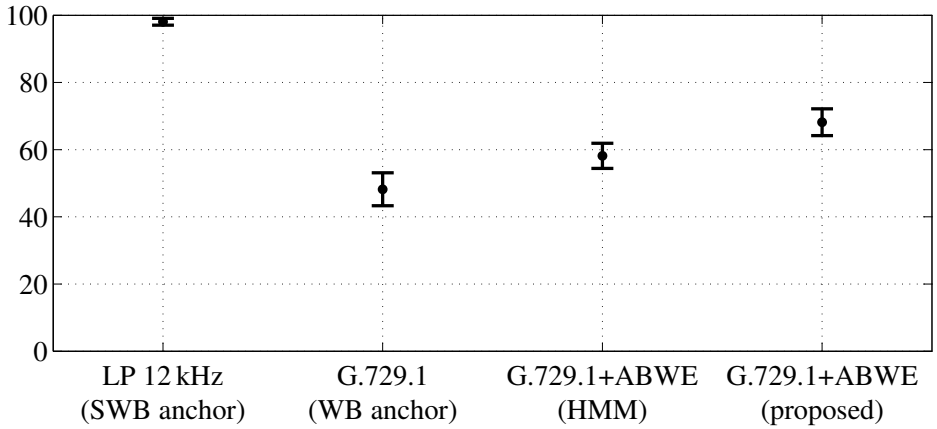

Figure 3: Results of the MUSHRA listening test.

first sentence of the double-sentence utterances of the database has been used. For the MUSHRA test, eight English language samples have been low-pass filtered with a cut-off frequency of 12 kHz. These "original" super-wideband samples have been presented as the reference sample. Then, the listeners should compare four processing variants (conditions) to the respective reference. The quality degradation of these processed samples should be judged on a scale between 0 and 100. The four processing variants that have been presented (in randomized order) for each test sample were:

- **LP 12 kHz** — The reference sample is included again as a hidden upper anchor.
- **G.729.1** — The wideband codec ITU-T G.729.1 at 32 kbit/s [IT06, RKT+07] has been used to encode the speech sample at the wideband sampling rate of 16 kHz. This variant is used as the lower (wideband) quality anchor.
- **G.729.1+ABWE (HMM)** — The Hidden-Markov-Model based ABWE method of [JV03] is applied to the G.729.1 output signal. The corresponding MMSE estimation algorithm has been configured with a 7-bit codebook and 16 Gaussian mixture components per state. See [Gei12a] for a more detailed description.
- **G.729.1+ABWE (proposed)** — This condition corresponds to our new method as described in Section 3 when applied to the G.729.1 output signal.

The test was conducted in a quiet environment (studio box) using Sennheiser HD600 open stereo headphones (diotic presentation). The headphones were driven by a dedicated amplifier with calibrated equalization. A comfortable presentation level was set by the subjects. Nine listeners participated in the test. They had to judge eight audio samples (in the four variants listed above). In total, $9 \cdot 8 = 72$ votes have been received per test condition.

The results, i.e., mean scores and 95% confidence intervals, are shown in Figure 3. It can be seen that, in fact, both ABWE methods yield a statistically significant quality improvement. However, our new proposal clearly outperforms the previous ABWE method despite the fact that it requires much less computational power. Note that the MUSHRA scores of the latter three conditions also include the degradation of the 32 kbit/s G.729.1 codec. A slight improvement of these scores can be expected if no codec is used.

5 Conclusions

The presented ABWE technique constitutes a viable alternative to previous algorithms which are based on statistical estimation and parametric high band synthesis, at least for the WB-to-SWB extension scenario. As shown by the subjective listening test, the proposed pitch-scaling based algorithm is able to regenerate a high-band speech signal of high quality. The computational complexity of the approach is relatively low.

Yet, the method is not directly applicable to NB-to-WB extension scenario because, in this case, the 3.4 – 4 kHz passband does mostly not exhibit a spectrotemporal structure that is similar enough to the extension band above 4 kHz. Instead, as an alternative, the solution presented in [Gei13] can be used if a modification of the transmitter side is allowed. A modification of network components is unnecessary in both cases.

References

[CGMS01] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter. Speech enhancement via frequency bandwidth extension using line spectral frequencies. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 665–668, Salt Lake City, UT, USA, 2001.

[CH94] Holger Carl and Ulrich Heute. Bandwidth Enhancement of Narrow-Band Speech Signals. In *Proceedings of European Signal Processing Conference (EUSIPCO)*, pages 1178–1181, Edinburgh, Scotland, September 1994.

[Die84] Manfred Dietrich. Performance and Implementation of a Robust ADPCM Algorithm for Wideband Speech Coding with 64 kbit/s. In *Proc. of Intl. Zürich Seminar on Digital Communications*, Zürich, Switzerland, March 1984.

[Gei12a] Bernd Geiser. *High-Definition Telephony over Heterogeneous Networks*. PhD thesis, IND, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany, 2012. Volume 33 in "Aachener Beiträge zu Digitalen Nachrichtensystemen (ABDN)", Verlag Mainz, Aachen, Germany.

[Gei12b] Bernd Geiser. Paths toward HD-Voice Communication. In *Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, September 2012. Keynote talk.

[Gei13] Bernd Geiser. Speech Bandwidth Extension Based on In-Band Transmission of Higher Frequencies. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, BC, Canada, May 2013.

[IT06] ITU-T. ITU-T Rec. G.729.1: G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729, 2006.

[ITU03] ITU-R Recommendation BS.1534-1. Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems, 2003.

[Jax02] Peter Jax. *Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds*. PhD thesis, IND, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany, 2002. Volume 15 in "Aachener Beiträge zu Digitalen Nachrichtensystemen (ABDN)", Verlag Mainz, Aachen, Germany.

[JV02] Peter Jax and Peter Vary. An upper bound on the quality of artificial bandwidth extension of narrowband speech signals. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 237–240, Orlando, FL, USA, 2002.

[JV03] Peter Jax and Peter Vary. On Artificial Bandwidth Extension of Telephone Speech. *Signal Processing*, 83(8):1707–1719, August 2003.

[KLA07] Juho Kontio, Laura Laaksonen, and Paavo Alku. Neural Network-Based Artificial Bandwidth Expansion of Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):873–881, 2007.

[NGAK02] Mattias Nilsson, Harald Gustaffson, Søren Vang Andersen, and W. Bastiaan Kleijn. Gaussian mixture model based mutual information estimation between frequency bands in speech. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 525–528, Orlando, FL, USA, 2002.

[NTN97] Y. Nakatoh, M. Tsushima, and T. Norimatsu. Generation of Broadband Speech from Narrowband Speech using Piecewise Linear Mapping. In *Proceedings of EUROSPEECH*, pages 1643–1646, Rhodes, Greece, September 1997.

[NTT94] NTT. NTT Advanced Technology Corporation: Multi-Lingual Speech Database for Telephonometry. online: `http://www.ntt-at.com/products_e/speech/`, 1994.

[PA11] H. Pulakka and P. Alku. Bandwidth Extension of Telephone Speech Using a Neural Network and a Filter Bank Implementation for Highband Mel Spectrum. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2170–2183, 2011.

[PK00] Kun-Youl Park and Hyung Soon Kim. Narrowband to wideband conversion of speech using GMM based transformation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1843–1846, Istanbul, Turkey, 2000.

[PLV+08] H. Pulakka, L. Laaksonen, M. Vainio, J. Pohjalainen, and P. Alku. Evaluation of an Artificial Speech Bandwidth Extension Method in Three Languages. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1124–1137, 2008.

[PX81] Peter J. Patrick and C. S. Xydeas. Speech Quality Enhancement by High Frequency Band Generation. In *Proc. of Intl. Conf. on Digital Processing of Signals in Communications*, pages 365–373, Loughborough, England, April 1981.

[RKT+07] S. Ragot, B. Kövesi, R. Trilling, D. Virette, N. Duc, D. Massaloux, S. Proust, B. Geiser, M. Gartner, S. Schandl, H. Taddei, Yang Gao, E. Shlomot, H. Ehara, K. Yoshida, T. Vaillancourt, R. Salami, Mi Suk Lee, and Do Young Kim. ITU-T G.729.1: An 8-32 kbit/s Scalable Coder Interoperable with G.729 for Wideband Telephony and Voice Over IP. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 529–532, Honolulu, Hawai'i, USA, April 2007.

[Sch33] K.-O. Schmidt. Neubildung von unterdrückten Sprachfrequenzen durch ein nicht-linear verzerrendes Glied. *Telegraphen- und Fernsprech-Technik*, 22(1):13–22, January 1933.

[SSGV11] Magnus Schäfer, Christopher Schnelling, Bernd Geiser, and Peter Vary. A Listening Test Environment for Subjective Assessment of Speech and Audio Signal Processing Algorithms. In *Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, volume 61 of *Studientexte zur Sprachkommunikation*, pages 237–244. ITG, DEGA, TUDpress Verlag der Wissenschaften GmbH, September 2011.

[Yas95] Hiroshi Yasukawa. Enhancement of Telephone Speech Quality by Simple Spectrum Extrapolation Method. In *Proceedings of EUROSPEECH*, volume 2, pages 1545–1548, Madrid, Spain, September 1995.

[Zö11] Udo Zölzer, editor. *DAFX: Digital Audio Effects*. John Wiley & Sons Ltd., Chichester, UK, second edition, 2011.