EFFICIENT SUPER-WIDE BANDWIDTH EXTENSION USING LINEAR PREDICTION BASED ANALYSIS-SYNTHESIS

Pramod Bachhav, Massimiliano Todisco and Nicholas Evans

EURECOM, Sophia Antipolis, France – {bachhav,todisco,evans}@eurecom.fr

ABSTRACT

Many smart devices now support high-quality speech communication services at super-wide bandwidths. Often, however, speech quality is degraded when they are used with networks or devices which lack super-wideband support. Artificial bandwidth extension can then be used to improve speech quality. While approaches to wideband extension have been reported previously, this paper proposes an approach to *super*-wide bandwidth extension. The algorithm is based upon a classical source filter model in which spectral envelope and residual error information are extracted from a wideband signal using conventional linear prediction analysis. A form of spectral mirroring is then used to extend the residual error component before an extended super-wideband signal is derived from its combination with the original wideband envelope. Improvements to speech quality are confirmed with both objective and subjective assessments. These show that the quality of super-wideband speech, derived from the bandwidth extension of wideband speech, is comparable to that of speech processed with the standard enhanced voice services (EVS) codec with a bitrate of 13.2kbps. Without the need for statistical estimation of missing super-wideband components, the proposed algorithm is highly efficient and introduces only negligible latency.

***Index Terms*—** bandwidth extension, super-wideband, voice quality

1. INTRODUCTION

The quality of speech offered by modern communications systems and devices has improved enormously in recent times. Whereas many devices were, and continue to be restricted to narrow and wide bandwidths, today's technology such as the enhanced voice services (EVS) codec [1, 2] developed by the 3rd Generation Partnership Project (3GPP), increasingly supports super-wide bandwidths. When used with other devices and networks with compatible support for super-wideband (SWB) services, such technology offers extremely high quality communications.

Often, though, SWB devices are used with other devices and networks which support only narrowband (NB) or wideband (WB) communications. While they usually offer backward compatibility, users of SWB devices will then be restricted to NB or WB communications. A reduction in bandwidth accompanies a reduction in speech quality. Fortunately, though, there is potential to improve quality in these situations using artificial bandwidth extension (ABE).

The extensive body of ABE research in the literature targets mostly the extension of NB speech signals to WB speech signals. In these cases there is substantial potential to improve quality; significant speech components between the NB limit of 4kHz and the WB limit of 8kHz can be recovered reliably using ABE. SWB speech signals extend the limit to 16kHz. Super-wide bandwidth extension (SWBE) approaches can then be employed to recover missing components between 8kHz and 16kHz.

Only few approaches to SWBE are reported in the literature. This is perhaps because the SWBE task is considerably more challenging than the extension of NB signals to WB signals. This is simply because the potential gain in quality from the extension of WB to SWB is much less than the potential when extending from NB to WB. As a result, significant processing artefacts can no longer be tolerated. Most of the existing solutions are either too computationally demanding or impose levels of latency which prohibit real-time implementations. This paper proposes an efficient, low latency approach to SWBE. It is based upon a classical source-filter model in which a WB signal is extended using conventional linear prediction (LP) analysis.

The remainder of the paper is organised as follows. Section 2 presents a review of related, past work. Section 3 describes the proposed SWBE algorithm. Section 4 describes the experimental setup and both subjective and objective assessments. Conclusions are presented in Section 5.

2. PAST WORK

Many different approaches to bandwidth extension have been reported previously. These can be categorized as either blind or non-blind.

Non-blind methods recover missing frequency components from auxiliary high frequency (HF) side information which is encoded into a data stream together with low frequency (LF) components [3]. The inclusion of side information typically incurs an additional burden of 1-5 kbps [4]. Examples of non-blind approaches to SWBE include the spectral band replication (SBR)-based high-efficiency advanced audio codec (HE-AAC) [5], the extended adaptive multi-rate WB codec (AMR-WB+) [6] and the enhanced voice services (EVS) codec (SWB mode) [1]. Non-blind approaches are codec specific and require a matching decoder in order to recover HF components.

In contrast, blind methods estimate missing HF components using only the available LF components. In contrast to non-blind alternatives, blind methods do not incur any additional bit-rate burden and are codec-neutral. The blind approach is often preferred as a result and is that adopted in this work. Very few blind SWBE algorithms are reported in the literature. An approach referred to as efficient high-frequency bandwidth extension (EHBE) [7] estimates missing HF components from those in the highest octave of the WB signal. While improvements in quality are reported, the use of non-linear processing tends to produce audible intermodulation distortion. A small number of attempts, e.g. [4, 8, 9, 10], have been made to improve SWBE performance. However, subjective assessments reported in [4, 9] show that their performance is mostly comparable to that of the EHBE algorithm. These methods also require the

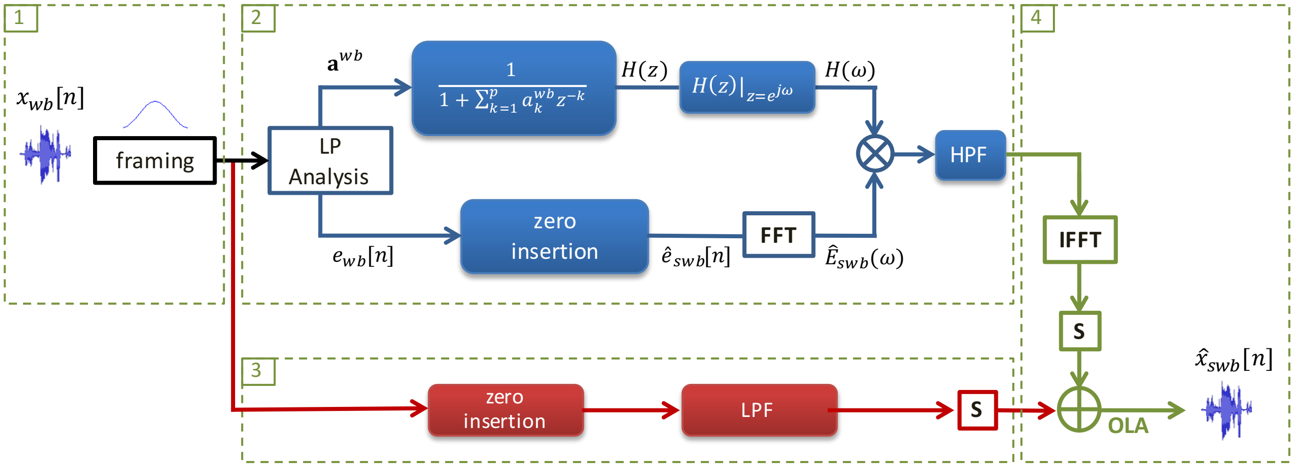

Fig. 1. *A block diagram of the proposed approach to SWBE.*

statistical estimation of missing HF components. Since it performs as well as more recent techniques while not requiring any statistical estimation procedure, the EHBE algorithm is used as a baseline approach in this work.

3. SUPER-WIDE BANDWIDTH EXTENSION (SWBE)

A block diagram of the proposed approach to SWBE is presented in Fig. 1. There are four key components. First, the WB input signal $x_{wb}[n]$ is windowed for subsequent frame-by-frame processing. Second, missing HF components are estimated from available LF components. Third, the original LF components are extracted from the input WB frame. Finally, an extended SWB output signal $\hat{x}_{swb}[n]$ is obtained by combining LF and HF components.

3.1. High frequency component estimation

The HF component of the input WB signal sampled at 16 kHz is estimated frame-by-frame via the blue-coloured components illustrated in Fig. 1 (box 2). Standard linear prediction (LP) coefficients $\mathbf{a}^{wb}$ and the residual component $e_{wb}[n]$ are obtained with conventional LP analysis of order p = 16. The LP coefficients, which characterise the filter/envelope of the WB signal, are used to determine the frequency response $H(\omega)$ from the transfer function $H(z)$. The residual component is extended by zero insertion in the time domain $\hat{e}_{swb}[n]$. As a form of spectral mirroring, the operation is equivalent to an up-sampling operation without an anti-aliasing filter [11]. The complex frequency domain representation of the excitation signal $\hat{E}_{swb}(\omega)$ is obtained from the extended residual $\hat{e}_{swb}[n]$ using the fast Fourier transform (FFT) and then combined by multiplication with the filter/envelope $H(\omega)$. Since the output is a composite of estimated HF components and distorted LF components, the latter are removed via high pass filtering (HPF), thereby preserving HF components only.

3.2. Low frequency component up-sampling

The LF component of the input signal $x_{wb}[n]$ is also extracted frame-by-frame. The processing involved is illustrated by the red-coloured components in Fig. 1 (box 3). Each frame is up-sampled in the time domain using zero insertion. An anti-aliasing low pass filter (LPF) is then applied. The result is an interpolated time domain signal at a sampling rate of 32kHz comprising only frequency components below 8kHz. This operation is common to all bandwidth extension algorithms.

3.3. Re-synthesis

Re-synthesis of the extended output $\hat{x}_{swb}[n]$ is performed via the green-coloured elements of Fig. 1 (box 4). A time domain signal containing only estimated HF components is obtained via the inverse FFT (IFFT). After synchronisation (S) to compensate for delays introduced by the different processes involved in the estimation of LF and HF components, a full-spectrum SWB speech signal with a sampling frequency of 32kHz is obtained from their addition. Synchronisation is also a component of every approach to bandwidth extension. Re-synthesis is accomplished using a conventional overlap-add (OLA) [12, 13] technique in order to avoid discontinuities at frame edges.

3.4. Spectral envelope analysis

Illustrations of the envelope extension process are shown in Fig. 2 for an arbitrary unvoiced (a) and voiced (b) speech frame. Blue and dashed-black profiles show the spectral envelopes of true WB and SWB signals respectively. These are derived with linear prediction of orders 16 (WB) and 32 (SWB).

Extended SWB signals are obtained by combining the original LF components with estimated HF components. As described in Section 3.1, the latter are obtained by passing the extended excitation signal through a filter whose frequency response is defined by the WB spectral envelope, followed by high-pass filtering. The effective frequency response that is combined with the extended excitation for re-synthesis is then a stretched copy of the WB spectral envelope (0-8kHz, blue profiles in Fig. 2), which gives the extended SWB envelope (0-16kHz, red profiles in Fig. 2). Only the HF components, contained within the green boxes in Fig. 2, bear influence on the resulting SWB signal. In this region the extended (red) and true SWB (dashed-black) profiles follow spectral shapes which are sufficiently similar to support SWBE.

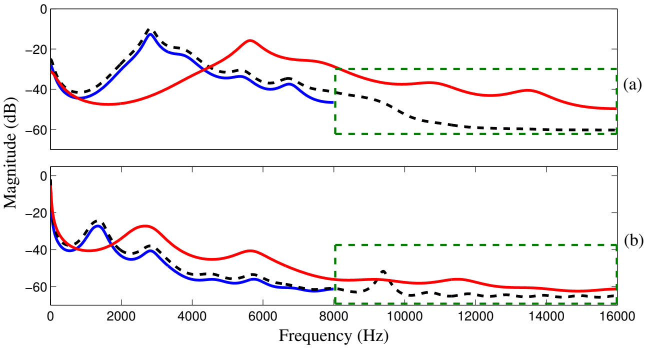

Fig. 2. *A comparison of spectral envelopes for an arbitrary speech frame extracted from a recording in the CMU database. Profiles shown for true WB speech (blue), true SWB speech (dashed-black) and WB-to-SWB extended speech (red). Plots shown for distinct frames of (a) unvoiced and (b) voiced speech.*

4. EXPERIMENTAL SETUP AND RESULTS

This section reports both objective and subjective assessments of the proposed SWBE algorithm.

4.1. Databases

All experiments reported here were performed using speech data from one of three different databases. The **CMU Arctic** database [14] consists of 1132 utterances collected from 3 speakers at a sampling rate of 32kHz. It is used widely in speech synthesis research [15]. The **TSP** database [16] consists of 1378 utterances collected from 12 male and 12 female speakers at a sampling rate of 48kHz. The database has been used previously for BWE [17] [18]. Finally, 6 English utterances collected from 4 speakers with a sampling rate of 48kHz were chosen from the **3GPP** database details of which can be found in ITU-T recommendation P.501 (annexure B and clause 7.3) [19]. These signals are commonly used for the objective evaluation of speech quality in telephonometry. All three databases contain phonetically balanced utterances.

4.2. Data pre-processing

Data pre-processing steps are illustrated in Fig. 3. All data in the TSP and 3GPP databases were first downsampled to SWB signals so that all three databases then have a common sampling rate of 32kHz. Downsampling was performed using the ResampAudio tool contained in the AFsp package [20]. The active speech level of all utterances in all three databases was then adjusted to -26dBov [21] to give SWB data x_{swb}[1]. Enhanced voice services (EVS) [22] encoding with active discontinuous transmission in channel aware mode was then applied to produce reference data x_{evs}

SWB signals x_{swb} were then downsampled to 16kHz and passed through a send-side bandpass filter [23] according to recommendation P.341, thereby limiting the bandwidth to 50Hz-7kHz, gives WB data x$_{wb}$. This data was in turn processed with adaptive multi-rate wideband (AMR-WB) coding [24] in default mode to produce reference data x_{amr} AMR-WB data x_{amr} forms the input to the SWBE algorithm (x_{wb} in Fig. 1 is replaced by x_{amr}).

[1] Indices $[n]$ (as illustrated in Fig. 1) are dropped for convenience.

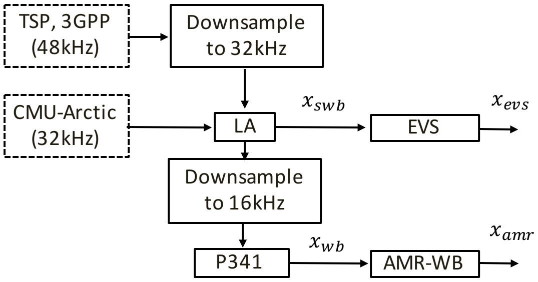

Fig. 3. *Protocol used for data pre-processing. LA = level alignment to -26 dBov.*

4.3. Assessment and baseline algorithm

The proposed bandwidth extension algorithm is assessed against AMR-WB and EVS processed speech signals, with the EHBE algorithm [7] being used as a baseline. Since EVS encodes frequencies up to 14kHz, bandwidth extended signals produced using either the baseline or the proposed approach c are also bandlimited to 14kHz. With a 512-point FFT, the proposed algorithm was implemented with Hann window of 25ms duration and 50% overlap, with OLA conditions necessary for perfect reconstruction [12, 13]. The EHBE baseline algorithm was implemented in the time domain without framing, as described in [7].

Input WB signals are assumed to be AMR-WB signals with a bitrate of 12.65kbps. No significant improvement in quality is obtained beyond this bitrate [25]. Encoding then operates over a frequency range of 0-6.4kHz whereas components up to 8kHz are added during decoding through noise filling [26]. Input signals to both the proposed and baseline algorithms thus extend to 8kHz. The EVS codec operates at a bitrate of 13.2kbps.

4.4. Objective measures

Objective assessment is performed using the standard root mean square log-spectral distortion (RMS-LSD) [27] metric which is known to correlate well with the results of subjective assessments [28]. The average RMS-LSD is determined for estimated HF components only, i.e. in the frequency range 8-14kHz (LF components are not taken into account). It is used to compare EVS-processed and bandwidth-extended speech signals produced using either the proposed algorithm or the EHBE baseline. Comparisons are made with original SWB signals x_{swb}. All signals were time-aligned before evaluation to account for any delay introduced by encoding/decoding.

Results presented in Table 1 show that the proposed algorithm gives a lower RMS-LSD than the EHBE algorithm. An average RMS-LSD of 9.92dB corresponds to an improvement of 1.44dB over the baseline. As expected, EVS processed signals show lower RMS-LSD values. While results for the proposed algorithm are inferior to those of EVS signals, they suggest that it gives a better estimate of the HF spectral shape than the baseline.

4.5. Subjective assessment

Subjective assessments were performed using comparison based mean-opinion score (CMOS) tests [27] following a protocol inspired by the comparison category rating (CCR) assessment method [29]. Each set of tests involves the pairwise comparison of bandwidth ex-

Table 1. *RMS-LSD results in dB (standard deviation).*

	Proposed	EHBE	EVS
CMU Arctic	10.13 (1.68)	11.74 (2.03)	5.00 (0.48)
3GPP	11.06 (1.90)	13.56 (2.30)	4.87 (0.39)
TSP speech	9.29 (0.84)	10.20 (1.04)	4.74 (0.51)
Average	9.92 (1.56)	11.36 (1.96)	4.94 (0.50)

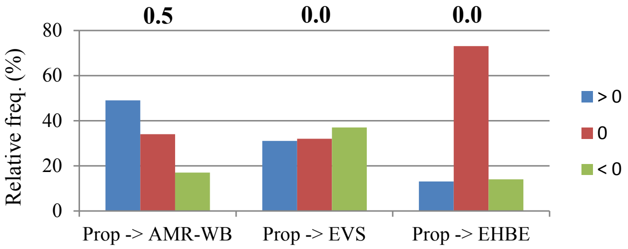

Fig. 4. *Subjective test results in terms of CMOS for bandwidth extended speech generated with the proposed (Prop) algorithm (A) versus either AMR-WB, EVS and EHBE processed speech (B). Each bar indicates the relative frequency that (blue bars) A was preferred to B (score>0), that (green bars) quality was indistinguishable (score=0), or that (red bars) B was preferred to A (score< 0). Scores illustrated to the top are average subjective scores.*

tended signals with (i) AMR-WB signals, (ii) EVS processed signals and (iii) those extended via the EHBE baseline algorithm. Each set of tests was performed by 14 listeners. They were asked to compare the quality of 15 (5 chosen randomly from each of the 3 databases) randomly ordered pairs of speech signals A and B, one of which was treated with the proposed bandwidth extension algorithm. Listeners were asked to rate the quality of signal A with respect to B according to the following scale: -3 (much worse), -2 (slightly worse), -1 (worse), 0 (about the same), 1 (slightly better), 2 (better), 3 (much better). The samples were played using DT 770 PRO headphones. Example speech files used for subjective tests are available online[2].

Subjective assessment results are illustrated in Fig. 4. Each group of three bars shows average listener preferences for each of the three comparisons. Blue bars show the percentage of tests in which signals treated with the proposed bandwidth extension algorithm were judged to be of superior quality (scores>0). Green bars show the percentage of trials where the same signals were judged to be of inferior quality (scores<0). Red bars show the percentage of tests for which relative quality was indistinguishable (scores=0).

Compared to AMR-WB signals, 49% of speech files treated with the proposed algorithm were judged to be of superior quality. As regards comparisons to EVS processed signals, 32% of trials were found to be of equivalent quality, while 31% were judged to be of superior quality. Quality was found to be inferior for 37% of trials. Up to 73% of comparisons to the EHBE baseline showed no discernible difference. CMOS illustrated to the top of Fig. 4 also illustrate the improvement in quality compared to AMR-WB signals and equivalence to EVS and EHBE processed signals. Overall, these results show that the proposed SWBE algorithm improves consistently on speech quality than AMR-WB signals and to the levels comparable with EVS and EHBE processed speech.

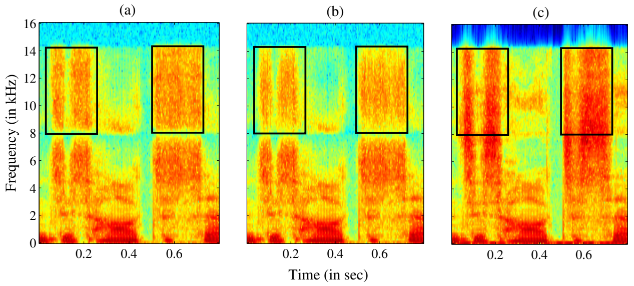

Fig. 5. *Spectrograms of a AMR-WB processed speech segment extended by the proposed algorithm (a) and the EHBE baseline (b) compared to true SWB speech (c). LF components (0-8kHz) in plots (a) and (b) are different than those in plot (c) due to AMR-WB processing.*

4.6. Discussion

Fig. 5 shows a comparison of spectrograms for speech signals after bandwidth extension using (a) the proposed and (b) the baseline algorithms with the true SWB spectrogram illustrated in (c). The spectral gap in both (a) and (b) around 8kHz which arises through AMR-WB processing is generally imperceptible [30]. The comparison of spectrograms in (a) and (b) shows that HF components estimated by the proposed method reflect more reliably the HF components in the true SWB spectrogram (c). This finding confirms the improvements found with objective RMS-LSD assessments. However, subjective assessments show that time domain processing without framing can lead to fewer processing artefacts.

Even though RMS-LSD objective assessment results show that the proposed SWBE algorithm produces speech of lower quality than that produced by the EVS codec, subjective assessment results show only marginal difference. This is because the level discrimination reduces drastically at higher frequencies (especially beyond 8kHz) [31]. As a result re-synthesized SWB speech is perceived to be of similar quality.

Lastly, whereas the EHBE algorithm operates on the speech signal directly, the proposed algorithm is based on a classical source filter model. Therefore, when used in combination with a WB codec which employs some form of linear prediction (e.g. AMR-WB codec), the proposed SWBE algorithm avoids an additional re-synthesis step and therefore introduces lower latency.

5. CONCLUSIONS

This paper proposes an approach to super-wide bandwidth extension that is based on a classical source filter model. With no need for the statistical estimation of high-frequency spectral envelope information, the algorithm is efficient, introduces negligible latency and is thus well suited to real time implementation. Results of both objective and subjective assessment show that the proposed super-wide bandwidth extension algorithm produces speech of notably higher quality than wideband input signal. Super-wideband output signals are furthermore of comparable quality to speech signals processed with the latest super-wideband enhanced voice services codec. Being codec neutral, the proposed algorithm can be used to improve the speech quality offered by wideband networks and devices and can also be used to preserve quality when super-wideband devices are used alongside wideband services.

[2] `http://audio.eurecom.fr/content/media`

6. REFERENCES

[1] "Codec for Enhanced Voice Services; Detailed algorithmic description (3GPP TS 26.445 ver. 13.4.0 rel. 13)," 2016.

[2] "Codec for Enhanced Voice Services; General overview (3GPP TS 26.441 ver. 13.0.0 rel. 13)," 2016.

[3] E. Larsen and R. Aarts, *"Audio bandwidth extension: application of psychoacoustics, signal processing and loudspeaker design,"*. John Wiley & Sons, 2005.

[4] X. Liu and C. Bao, "Blind bandwidth extension of audio signals based on non-linear prediction and hidden markov model," *APSIPA Transactions on Signal and Information Processing*, vol. 3, p. e8, 2014.

[5] P. Ekstrand, "Bandwidth extension of audio signals by spectral band replication," in *Proc. of the 1st IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA02)*, 2002, pp. 53–58.

[6] "Audio Codec Processing Functions - Extended Adaptive Multirate Wideband AMR-WB+ Codec; Transcoding functions (3GPP TS 26.290," 2004.

[7] E. Larsen, R. M. Aarts, and M. Danessis, "Efficient high-frequency bandwidth extension of music and speech," in *Audio Engineering Society Convention 112*. Audio Engineering Society, 2002.

[8] X. Liu and C.-C. Bao, "Audio bandwidth extension based on temporal smoothing cepstral coefficients," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–16, 2014.

[9] C.-C. Bao, X. Liu, Y.-T. Sha, and X.-T. Zhang, "A blind bandwidth extension method for audio signals based on phase space reconstruction," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–9, 2014.

[10] Y. Wang, S. Zhao, K. Mohammed, S. Bukhari, and J. Kuang, "Superwideband extension for AMR-WB using conditional codebooks," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3695–3698.

[11] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 1979, pp. 428–431.

[12] J. Benesty, M. Sondhi, and Y. Huang, *"Springer handbook of speech processing"*. Springer, USA, 2007.

[13] T. Dutoit and F. Marques, *"Applied Signal Processing: A MATLAB-Based Proof of Concept"*. Springer, USA, 2010.

[14] J. Kominek and A. Black, "CMU ARCTIC databases for speech synthesis," 2003. [Online] : http://festvox.org/cmu_arctic/index.html.

[15] A. Black and K. Tokuda, "The blizzard challenge 2005: Evaluating corpus-based speech synthesis on common databases," in *Proc. of INTERSPEECH*, 2005.

[16] P. Kabal, "TSP Speech Database," *McGill University, Database Version : 1.0*, pp. 02–10, 2002. [Online]: http://mmsp.ece.mcgill.ca/Documents/Data/.

[17] Y. Qian and P. Kabal, "Dual-mode wideband speech recovery from narrowband speech." in *Proc. of INTERSPEECH*, 2003.

[18] P. Bachhav, M. Todisco, M. Mossi, C. Beaugeant, and N. Evans, "Artificial bandwidth extension using the constant Q transform," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5550–5554.

[19] "ITU-T Recommendation P. 501, Test signals for use in telephonometry," *ITU*, 2012. [Online]: https://www.itu.int/rec/T-REC-P.501-201201-I/en.

[20] P. Kabal, "The AFsp package," *http://www-mmsp.ece.mcgill.ca/Documents/Downloads/AFsp/*.

[21] "ITU-T Recommendation P. 56, Objective measurement of active speech level," *ITU*, 2011.

[22] "Codec for Enhanced Voice Services; ANSI C Code (fixed point) (3GPP TS 26.442 ver. 13.3.0 rel. 13)," 2016.

[23] "ITU-T Recommendation G. 191, Software Tool Library 2009 User's Manual," *ITU*, 2009.

[24] "ANSI-C Code for the AMR-WB Speech Codec (3GPP TS 26.173 ver. 13.1.0 rel. 13)," 2016.

[25] A. Rämö, "Voice quality evaluation of various codecs," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4662–4665.

[26] "Speech Codec Speech Processing Functions; AMR-WB codec; Transcoding functions (3GPP TS 26.190 ver. 13.0.0 rel. 13)," 2016.

[27] D. Zaykovskiy and B. Iser, "Comparison of neural networks and linear mapping in an application for bandwidth extension," in *Proc. of Int. Conf. on Speech and Computer (SPECOM)*, 2005, pp. 1–4.

[28] P. Jax and P. Vary, "An upper bound on the quality of artificial bandwidth extension of narrowband speech signals," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, pp. I–237.

[29] "ITU-T Recommendation P. 800: Methods for subjective determination of transmission quality," *ITU*, 1996.

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

[31] M. Florentine, S. Buus, and C. Mason, "Level discrimination as a function of level for tones from 0.25 to 16 khz," *The Journal of the Acoustical Society of America*, vol. 81, no. 5, pp. 1528–1541, 1987.