A SIMPLE CEPSTRAL DOMAIN DNN APPROACH TO ARTIFICIAL SPEECH BANDWIDTH EXTENSION

Johannes Abel, Maximilian Strake, and Tim Fingscheidt

Institute for Communications Technology, Technische Universität Braunschweig, Germany

{j.abel, m.strake, t.fingscheidt}@tu-bs.de

ABSTRACT

In this work, we present a simple deep neural network (DNN)-based regression approach to artificial speech bandwidth extension (ABE) in the frequency domain for estimating missing speech components in the range $4 \ldots 7$ kHz. The upper band (UB) spectral magnitudes are found by first estimating the UB cepstrum by means of a DNN regression and subsequent conversion to the spectral domain, leading to a more efficient and generalizing model training rather than estimating highly redundant UB magnitudes directly. As second novelty the phase information for the estimated upper band spectral magnitudes is generated by spectrally shifting the NB phase. Apart from framing, this very simple approach does not introduce additional algorithmic delay. A cross-database and cross-language task is defined for training and evaluation of the ABE framework. In a subjective comparison category rating test, the proposed ABE solution significantly outperforms the competing ABE baseline and was found to improve NB speech quality by 0.80 CMOS points, while the computation time is reduced to about 3 % compared to the ABE baseline.

***Index Terms*—** artificial speech bandwidth extension, speech enhancement, machine learning, deep neural network, regression

1. INTRODUCTION

Artificial speech bandwidth extension (ABE) is a speech enhancement approach, typically located in the downlink path of a telephone call. ABE algorithms aim to enhance narrowband (NB) speech signals, i.e., signals containing only frequency components up to 4 kHz. Compared to wideband (WB) speech signals, NB calls miss acoustic components in the upper band (UB), i.e., the frequency range $4\,\text{kHz} < f < 8\,\text{kHz}$, and thus limited speech quality and intelligibility is resulting. Consequently, recovering the UB has high potential for enhancing the speech quality of a NB telephony call. In several works, the improved subjective speech intelligibility and quality of additional acoustic bandwidth either by employment of WB speech codecs or an ABE solution was shown [1, 2, 3, 4]. Using a WB speech codec and thus having a high-quality phone call often fails for practical reasons, e.g., if one of the participants of a call is not located in a WB-capable cell or if a call is conducted from one operator to another. In all of these cases, ABE can serve as safety net to keep the speech quality as high as possible, even if a WB call is not available. In the age of an increased number of WB calls this is expected to be the major role of ABE (at least) in the next two decades.

The task of extending speech signals is often solved by means of the source-filter model for speech production. In more detail, an UB residual signal as well as an UB envelope needs to be estimated. While the UB residual is often found by simple application of spectral folding [5], the estimation of an UB spectral envelope is a challenging task. One approach is to classify among pretrained UB spectral codebook entries, representing UB spectral envelopes. In [6, 7], classification among the pretrained entries was done by finding the lowest distance to codebook entries of NB speech signals, which directly map to the UB codebook entries. Furthermore, Gaussian mixture models (GMMs) have been employed in [8, 9]. Additionally, considering behavior over time, hidden Markov models (HMMs) with GMMs as acoustic models were employed in [10, 11, 12, 13, 14]. Supporting the HMM/GMM statistical model, Bauer et al. additionally employed neural networks (NNs) [15]. In [16], an NN is used to estimate parameters for UB envelope shaping. In the recent past, deep neural networks (DNNs) have been employed as classifiers for pretrained UB envelopes in [17] or for directly estimating (by regression) the UB envelope in [18, 19, 20, 21].

Opposed to source-filter model-based ABE, the UB can be estimated by finding UB magnitudes and UB phases in the frequency domain, followed by a transformation back to the time domain. In [22] sum-product networks estimate a log spectrogram, while the phase is generated via an iterative algorithm. Furthermore, in [23] the UB log-power spectrum is directly estimated using a DNN, while the UB phase is obtained via inverse mirroring of the phase from the NB signal. Obtaining the phase also from the NB signal, in [24] UB magnitudes are estimated by a recurrent NN with long short-term memory (LSTM) cells.

Estimating high-dimensional UB spectral magnitudes leads to a psychoacoustically highly redundant estimation output and therefore prevents an efficient and generalizable statistical model training. Accordingly, in this paper, we estimate the UB spectrum by first estimating a cepstral representation which is of low dimension and thus contains less redundancy. Consequently, the proposed approach is simpler in terms of algorithmic complexity. Additionally, no lookahead is employed, thus the solution extends the input signal instantaneously on a frame basis. Special focus lies on practical employment of the presented ABE solution, therefore we trained and tested the solution using different speech databases. Furthermore, the solution is trained on English speech data, while speech quality assessment was conducted including also German speech data. The efficiency and performance of our very simple ABE approach is finally proven by the results of a subjective listening test.

The paper is structured as follows: In Sec. 2 a detailed description of the employed framework is given. Special attention lies on the description of UB magnitude and phase estimation. Subsequently, in Sec. 3 we explain the experimental setup, including speech data sets, preprocessing steps, and an ABE baseline system which we use for comparison. Instrumental assessment of the new simple ABE approach is conducted in Sec. 4, and finally we present the results of a subjective comparison category rating (CCR) listening test.

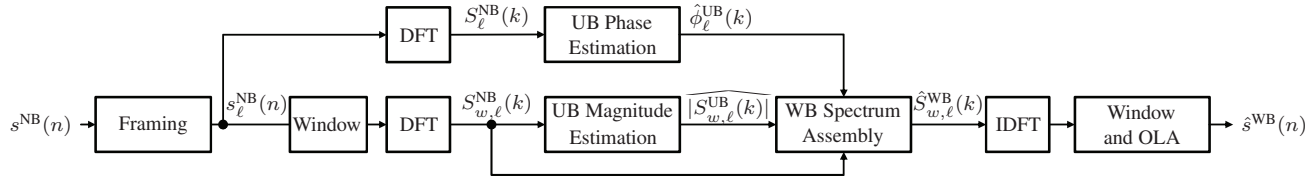

Fig. 1. Block diagram of the proposed ABE algorithm. Sample index n refers to 16 kHz sampling rate.

2. ARTIFICIAL BANDWIDTH EXTENSION

In Fig. 1 a block diagram presenting the ABE framework is depicted. We assume that the input is a speech signal $s^{\mathrm{NB}}(n)$, with n being the sample index for a sampling frequency of $f_s = 16$ kHz, however, containing only energy in the lower band (or NB). This is a typical scenario in modern signal processing paths in the downlink, which are increasingly able to transmit 16 kHz sampled speech signals, even if somewhere in the transmission path only a NB coding has been used. The algorithm outputs an artificially-extended speech signal $\hat{s}^{\mathrm{WB}}(n)$, containing an estimated UB.

2.1. Framework

Following the block diagram in Fig. 1, the input NB speech signal $s^{\mathrm{NB}}(n)$ is framed to a length of $L = 256$ samples (i.e., 16 ms). Subsequently, the speech frames $s_\ell^{\mathrm{NB}}(n)$, with ℓ being the frame index, are subject to a periodic square root Hann window and subsequent discrete Fourier transform (DFT) of size $K = 256 = L$, with k being the frequency bin index. The frame shift is $L/2 = 128$ samples (i.e., 8 ms). The set of indices representing the NB speech components is $\mathcal{K}^{\mathrm{NB}} = \{k | 0 \leq k \leq \frac{K}{4}\}$. Accordingly, the UB is represented by set $\mathcal{K}^{\mathrm{UB}} = \{k | \frac{K}{4} + 1 \leq k \leq \frac{K}{2}\}$. The resulting spectrum $S_{w,\ell}^{\mathrm{NB}}(k) \in \mathbb{C}$ is used as input for the statistical model which estimates the UB magnitude $\widehat{|S_{w,\ell}^{\mathrm{UB}}(k)|}$ (Sec. 2.2) and to provide the NB speech components for the subsequent WB spectrum assembly. Once the UB magnitudes and phases have been estimated, the WB spectrum is assembled according to

$$\hat{S}_{w,\ell}^{\mathrm{WB}}(k) = \begin{cases} S_{w,\ell}^{\mathrm{NB}}(k) & \text{for } k \in \mathcal{K}^{\mathrm{NB}} \\ \widehat{|S_{w,\ell}^{\mathrm{UB}}(k)|} \cdot \exp\left(j \cdot \hat{\phi}_\ell^{\mathrm{UB}}(k)\right) & \text{for } k \in \mathcal{K}^{\mathrm{UB}}. \end{cases} \quad (1)$$

Please note that the UB phase estimate $\hat{\phi}_\ell^{\mathrm{UB}}(k)$ is obtained from the spectrum $S_\ell^{\mathrm{NB}}(k) \in \mathbb{C}$, i.e., the DFT of the input speech frame which was not subject to windowing (i.e., rectangular windowing). Estimation of a suitable phase is described in Sec. 2.3.

By means of the inverse discrete Fourier transform (IDFT), the estimated WB spectrum $\hat{S}_{w,\ell}^{\mathrm{WB}}(k)$ is transformed back into the time domain. After windowing with a periodic square root Hann window, overlap-add (OLA) provides the bandwidth-extended output signal $\hat{s}^{\mathrm{WB}}(n)$. At this point in processing, the input NB signal was multiplied twice with a square root Hann window. In addition, the DNN was trained using targets which were also calculated on square root Hann windowed spectra. Consequently the required synthesis properties for a 50% OLA structure are fulfilled. For result reporting, we will refer to this proposed approach as **ABE-Simple**.

2.2. UB Magnitude Estimation

A fully connected feedforward DNN [25] is used for UB spectral magnitude estimation. The input feature vector is defined as

$$\mathbf{x}_\ell = \left(\ln|S_{w,\ell}^{\mathrm{NB}}(k)|\right)\big|_{k \in \mathcal{K}^{\mathrm{NB}}}, \quad (2)$$

i.e., the log-spectral magnitude information of the NB speech signal. The DNN works in regression mode and outputs a 20-dimensional cepstral vector $\hat{\mathbf{c}}_\ell^{\mathrm{UB}}$ as estimates for the UB magnitudes. The network parameters required for this processing were found in a preliminary training phase, which will be described in more detail in Sec. 3.2. To obtain a spectral representation of the UB magnitudes, the estimated cepstral vector is converted by means of the inverse discrete cosine transform (IDCT) [26, 27], as follows:

$$\left(\ln|S'_{w,\ell}(k')|\right)\big|_{k' \in \mathcal{K}'} = \mathrm{IDCT}\{\hat{\mathbf{c}}_\ell^{\mathrm{UB}}\}, \quad (3)$$

with $k' \in \mathcal{K}' = \{0, 1, \ldots, \frac{K}{4} - 1\}$ being the frequency bin index of the critically downsampled UB spectrum. Finally, the estimated UB spectral magnitudes are calculated:

$$\widehat{|S_{w,\ell}^{\mathrm{UB}}(k)|} := \begin{cases} 0 & \text{for } k \in \mathcal{K}^{\mathrm{NB}} \\ \exp\left(\ln|S'_{w,\ell}(k - K/4 - 1)|\right) & \text{for } k \in \mathcal{K}^{\mathrm{UB}}. \end{cases} \quad (4)$$

2.3. UB Phase Estimation

To obtain the estimated UB phases, we simply copy the phase from the NB spectrum to the UB, following

$$\hat{\phi}_\ell^{\mathrm{UB}}(k + K/4) := \arg\left(S_\ell^{\mathrm{NB}}(k)\right), k \in \{1, 2, \ldots, K/4\}, \quad (5)$$

with $\arg()$ returning the phase angle of the complex-valued spectrum $S_\ell^{\mathrm{NB}}(k)$. Although this estimate can be considered as very coarse, it maintains a plausible evolvement of phase, both over time and frequency, which will turn out to provide good quality.

3. EXPERIMENTAL SETUP

3.1. Speech Data and Preprocessing

The definition of data sets and preprocessing steps follow precisely [21]: The speech data used in this work is taken from the TIMIT database [28], Speechdat-Car US (SDC) database [29], and NTT database [30]. The data from TIMIT and SDC was mixed and used to create a training and validation set for DNN training (c.f. Sec. 3.2). For instrumental and subjective performance evaluation of the proposed ABE solution, we take the German and American English parts of the NTT database as test set. Consequently, the DNN uses 5.8 h and 2.4 h of speech material for training and validation, respectively. The test set contains 0.4 h of speech data.

We preprocessed the speech data following [31]: For the NB condition, the available WB speech signals are MSIN-filtered [32], decimated to 8 kHz, coded by the adaptive multirate (AMR) speech codec at bitrate 12.2 kbps [33] and finally decoded. Before and after coding, the speech signal was subject to a 16 to 13 bit conversion. Finally, the NB condition is resampled to 16 kHz and thus serves as input signal to the ABE framework, referred to as **AMR**. The WB condition is obtained by P.341-filtering [34] of the 16 kHz speech

signal provided by the speech databases. This intermediate result serves as input for generating training targets and as reference signal for all employed instrumental metrics. For comparison in the instrumental and subjective evaluation, we further code and decode the speech signal using the AMR WB codec [35] at a bitrate of 12.65 kbps, referred to as **AMR-WB**.

3.2. DNN Training for UB Magnitude Estimation

For DNN training, the features are generated as described in Sec. 2.2. For deriving time-aligned targets, the ground-truth WB speech signal is framed, windowed by a square root Hann window, and subsequently transformed into the frequency domain leading to WB spectra $S_{w,\ell}^{\mathrm{WB}}(k)$ (following exactly the steps as described in Sec. 2.1 for the 16 kHz-sampled NB speech signal). Subsequently, the targets are derived by converting the WB spectra to the cepstral domain by means of the K-point discrete cosine transform (DCT) [26, 27], following

$$\mathbf{c}_{\ell}^{\mathrm{UB}} = \mathrm{DCT}\left\{\left(\ln|S_{w,\ell}^{\mathrm{WB}}(k)|\right)\big|_{k\in\mathcal{K}^{\mathrm{UB}}}\right\}. \tag{6}$$

As targets, we use only the first 20 cepstral coefficients, since only the UB envelope is considered as perceptually relevant for ABE. The DNN is trained with a topology of three hidden layers, each having 256 units. Rectified linear units (ReLUs) are used as activation function in the units [25].

3.3. ABE Baseline Approach

As baseline for our investigations, we employ Bauer's ABE approach presented in [15], which was retrained on the exact same speech data sets as the proposed **ABE-Simple** approach. The baseline approach is based on a source-filter model employing a hidden Markov model with Gaussian mixture model as acoustic model and additional two neural networks, which support the estimation process by contributing detailed information on the UB energy, especially at fricative sounds, such as /s/ or /z/. The baseline approach evolved from years of research and had been found to significantly improve the speech quality of the incoming NB speech signal [15]. We will refer to this ABE approach as **ABE-Baseline**.

3.4. Measures for Instrumental Evaluation

To judge the reconstruction of the UB on signal level, we employ the logarithmic spectral distance (LSD) implemented after [36]. The LSD is calculated in the spectral domain for a set of frequency bins $\mathcal{K}$, following

$$LSD_{\ell} = \sqrt{\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\left[10\log_{10}\left(\frac{|S_{\ell}(k)|^2}{|\widehat{S}_{\ell}(k)|^2}\right)\right]^2}\ [\mathrm{dB}], \tag{7}$$

with $S_{\ell}(k)$ and $\widehat{S}_{\ell}(k)$ being the short-term spectra of the reference and the degraded speech signal as computed from the time domain signals, respectively, and $|.|$ being the number of elements in the set. Please note that the spectra for LSD calculation are computed with a frame size of $L = 256$ samples, plus 50% look-back and lookahead, and were windowed using a Hamming window of length $K' = 512$. The reported LSD values are the result of first averaging over all frames of a speech signal and subsequent averaging over all speech files. Please note that we consider only frames with voice present for calculating the LSD measure. We will report a WB LSD score, where we consider all frequencies up to 7 kHz and an UB LSD score, where only the speech components between 4 and 7 kHz are being accounted for. Please note that the LSD measure in principle is phase-insensitive, however, the phase estimation (5) in this work can lead to constructive or destructive interference in the OLA processing block, thus the LSD can evaluate phase estimation approaches, even though magnitude spectra are compared in (7).

In addition, we instrumentally assess the speech quality of the ABE-processed speech signal using WB-PESQ [37] in terms of the mean opinion score listening quality objective (MOS-LQO). Both, LSD and WB-PESQ use the respective WB speech signals in direct condition as reference.

Condition under test	MOS-LQO	WB LSD	UB LSD
AMR	2.99	17.99	25.41
ABE-Baseline	2.62	10.48	12.61
ABE-Simple	2.83	9.44	11.16
ABE-Simple w/ oracle phase	2.99	8.02	8.27
AMR-WB	3.54	7.35	8.38

Table 1. Instrumental results; LSD values in [dB].

4. INSTRUMENTAL QUALITY ASSESSMENT

The results for instrumentally assessing the ABE-processed speech signals are given in Tab. 1. Regarding the predicted speech quality in terms of MOS-LQO, **AMR-WB** obviously outperforms **AMR** by 0.55 MOS-LQO points. This small difference is due to the fact that in this case WB-PESQ is also used to score the narrowband **AMR** files sampled at 16 kHz. Regarding the ABE approaches, we found the lowest MOS-LQO values at the **ABE-Baseline** condition with 2.62, while the proposed **ABE-Simple** approach obtains a higher score of 2.83. Using the oracle phase during ABE processing, **ABE-Simple** scores 2.99 MOS-LQO points, which is therefore rated by WB-PESQ to have the same speech quality as **AMR**. This prediction of speech quality given by WB-PESQ is inconsistent with the result of preliminary informal subjective tests, conducted in our facilities. This is in line with earlier observations [1, 4, 2] which led to the proposal of the QABE measure [38] as instrumental quality index for ABE systems. In order not to use our own measure for our own approach here, we decided to conduct a subjective listening test in Sec. 5.

The WB LSD measure, on the other hand, attests a huge improvement for the **ABE-Baseline** condition by 7.51 dB, when comparing to **AMR**. Considering **ABE-Simple**, the improvement is even higher with 8.55 dB. Compared to the **AMR-WB** condition, **ABE-Baseline** is 3.13 dB behind, while **ABE-Simple** is deviating only by 2.09 dB.

Looking only at the UB LSD, **ABE-Simple** improves the **AMR** condition by an impressive 14.25 dB, thus also outperforming **ABE-Baseline**. If we use the oracle UB phase information, extracted from the uncoded WB speech signal, the LSD gets as low as 8.27 dB, which is even lower than the UB LSD of the **AMR-WB** condition. This indicates still potential for further improvement by a better phase estimation method.

Finally, we timed both approaches[1] during processing of the test set. Relative to the **ABE-Baseline** approach the **ABE-Simple** method consumes about 3 % of the computational power. The enormous reduction of complexity in the proposed approach comes along with an improvement of more than 1 dB regarding the WB LSD metric, compared to **ABE-Baseline**.

[1]Both approaches have been implemented in MATLAB and were executed on a typical PC platform. File operations were not considered.

CCR Condition	CMOS	CI_{95}
AMR vs. **AMR-WB**	1.63	$[1.48; 1.88]$
ABE-Baseline vs. **AMR-WB**	1.28	$[1.11; 1.44]$
ABE-Simple vs. **AMR-WB**	1.03	$[0.86; 1.20]$
ABE-Baseline vs. **ABE-Simple**	0.15	$[0.01; 0.29]$
AMR vs. **ABE-Baseline**	0.63	$[0.39; 0.87]$
AMR vs. **ABE-Simple**	0.80	$[0.55; 1.05]$

Table 2. Subjective speech quality assessment: Results from a CCR test, evaluating the **ABE-Baseline** baseline and the new **ABE-Simple** approach vs. NB- and WB-coded speech signals.

5. SUBJECTIVE SPEECH QUALITY ASSESSMENT

In a semi-formal CCR listening test [39, Annex E], where two conditions are compared to each other at once and rated on the comparison MOS (CMOS) scale from -3 (much worse) to +3 (much better), we evaluate four conditions:

- **AMR**: Coded NB speech, processed as described in Sec. 3.1
- **AMR-WB**: Coded WB speech, processed as described in Sec. 3.1
- **ABE-Baseline**: Baseline ABE approach as referred to in Sec. 3.3, with **AMR** speech data input
- **ABE-Simple**: Newly proposed ABE solution as described in Sec. 2.1, with **AMR** speech data input

In total, 12 German native speakers without known hearing impairment judged the conditions under test. Each condition was subject to P.341-conformant bandpass-filtering to a frequency range of $0.2 \ldots 7$ kHz [34, 16, 15], active speech level scaling to -26 dBov [40], and final conversion to 48 kHz sampling rate. The sentences under test were taken from the German part of the NTT database, thus they were also included in the test data set for instrumental assessment. Speech files of two female and two male speakers were chosen, each speaker providing 4 sentences for further processing using the conditions under test.

Participants listened to the signals under test in diotic fashion using two PCs with `RME Fireface 400` sound cards using `AKG K-271 MKII` headphones. A preliminary familiarization test was conducted, including all test conditions. Subsequently, the subjects judged 36 speech file pairs, in both orders, leading to 72 comparisons in total. The presentation order was randomized and split into two sets of comparisons, balanced over speakers and conditions.

The results of the listening test are presented in Tab. 2 in terms of CMOS and a respective 95% confidence interval (CI_{95}) for each of the CCR conditions. First of all, the **AMR-WB** condition was found to improve the speech quality compared to **AMR** by 1.63 CMOS points. Furthermore, compared to **AMR-WB**, the **ABE-Baseline** approach is found to be 1.28 CMOS points worse, while **ABE-Simple** was found to be only 1.03 CMOS points worse. This already indicates a superior performance of the new ABE solution w.r.t. the baseline approach.

In a direct comparison, the new **ABE-Simple** approach outperforms the baseline by 0.15 CMOS points, with a confidence interval starting above zero, thus just proving significance of the result.

Furthermore, the baseline approach was confirmed to significantly outperform **AMR** by 0.63 CMOS points. An even higher gain in speech quality could be shown by our new **ABE-Simple** with 0.80 CMOS points, compared to the underlying NB condition. Considering also the first CCR condition, comparing NB to WB, we conclude that the **ABE-Simple** approach bridges about 50% of the gap between NB and WB speech quality, achieved by a quite simple yet effective DNN-based approach.

6. CONCLUSIONS

In this paper we presented a simple approach to artificial speech bandwidth extension (ABE). The approach is characterized by a low-complexity deep neural network-based estimation scheme of upper band (UB) spectral magnitudes and a simple yet effective UB phase estimation which reuses the phase information obtained from the incoming NB signal. In terms of UB log-spectral distance, the proposed ABE framework improves the underlying NB speech condition by 14.25 dB, therefore also outperforming the ABE baseline system. This result was confirmed in a subjective comparison category rating test which revealed an improvement in terms of speech quality by 0.80 CMOS points compared to NB. Compared to the baseline, the proposed approach only takes about 3 % of the computational power, indicating a huge complexity reduction.

Instrumental assessment of the ABE solutions suggests that further quality improvement could be achieved by improving the UB phase estimation.

7. ACKNOWLEDGMENT

We thank Timo Gerkmann, Universität Hamburg, for fruitful discussions on speech phase.

8. REFERENCES

[1] S. Möller, E. Kelaidi, F. Köster, N. Côté, P. Bauer, T. Fingscheidt, T. Schlien, H. Pulakka, and P. Alku, "Speech Quality Prediction for Artificial Bandwidth Extension Algorithms," in *Proc. of Interspeech*, Lyon, France, Aug. 2013, pp. 3439–3443.

[2] J. Abel, M. Kaniewska, C. Guillamé, W. Tirry, H. Pulakka, V. Myllylä, J. Sjöberg, P. Alku, I. Katsir, D. Malah, I. Cohen, M. A. T. Turan, E. Erzin, T. Schlien, P. Vary, A. H. Nour-Eldin, P. Kabal, and T. Fingscheidt, "A Subjective Listening Test of Six Different Artificial Bandwidth Extension Approaches in English, Chinese, German, and Korean," in *Proc. of ICASSP*, Shanghai, China, Mar. 2016, pp. 5915–5919.

[3] P. Bauer, M.-A. Jung, J. Qi, and T. Fingscheidt, "On Improving Speech Intelligibility in Automotive Hands-Free Systems," in *Proc. of IEEE International Symposium on Consumer Electronics*, Braunschweig, Germany, June 2010, pp. 1–5.

[4] P. Bauer, C. Guillaumé, W. Tirry, and T. Fingscheidt, "On Speech Quality Assessment of Artificial Bandwidth Extension," in *Proc. of ICASSP*, Florence, Italy, May 2014, pp. 6082–6086.

[5] J. Makhoul and M. Berouti, "High-Frequency Regeneration in Speech Coding Systems," in *Proc. of ICASSP*, Washington, DC, USA, Apr. 1979, vol. IV.

[6] H. Carl and U. Heute, "Bandwidth Enhancement of Narrow-Band Speech Signals," in *Proc. of EUSIPCO*, Edinburgh, UK, Sept. 1994, pp. 1178–1181.

[7] T. Unno and A. McCree, "A Robust Narrowband to Wideband Extension System Featuring Enhanced Codebook Mapping," in *Proc. of ICASSP*, Philadelphia, PA, USA, Mar. 2005, pp. 805–808.

[8] A. H. Nour-Eldin and P. Kabal, "Memory-Based Approximation of the Gaussian Mixture Model Framework for Bandwidth Extension of Narrowband Speech," in *Proc. of Interspeech*, Florence, Italy, Aug. 2011, pp. 1185–1188.

[9] Y. Wang, S. Zhao, Y. Yu, and J. Kuang, "Speech Bandwidth Extension Based on GMM and Clustering Method," in *Proc. of International Conference on Communication Systems and Network Technologies*, Gwalior, India, Apr. 2015, pp. 437–441.

[10] P. Jax and P. Vary, "Wideband Extension of Telephone Speech Using a Hidden Markov Model," in *Proc. of IEEE Workshop on Speech Coding*, Delavan, WI, USA, Sept. 2000, pp. 133–135.

[11] P. Bauer and T. Fingscheidt, "A Statistical Framework for Artificial Bandwidth Extension Exploiting Speech Waveform and Phonetic Transcription," in *Proc. of EUSIPCO*, Glasgow, Scotland, Aug. 2009, pp. 1839–1843.

[12] C. Yagli, M. A. T. Turan, and E. Erzin, "Artificial Bandwidth Extension of Spectral Envelope Along a Viterbi Path," *Speech Communication*, vol. 55, pp. 111–118, Jan. 2013.

[13] M. A. T. Turan and E. Erzin, "Synchronous Overlap and Add of Spectra for Enhancement of Excitation in Artificial Bandwidth Extension of Speech," in *Proc. of Interspeech*, Dresden, Germany, Sept. 2015, pp. 2588–2592.

[14] I. Katsir, D. Malah, and I. Cohen, "Evaluation of a Speech Bandwidth Extension Algorithm Based on Vocal Tract Shape Estimation," in *Proc. of IWAENC*, Aachen, Germany, Sept. 2012, pp. 1–4.

[15] P. Bauer, J. Abel, and T. Fingscheidt, "HMM-Based Artificial Bandwidth Extension Supported by Neural Networks," in *Proc. of IWAENC*, Juan les Pins, France, Sept. 2014, pp. 1–5.

[16] 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*, vol. 19, no. 7, pp. 2170–2183, Sept. 2011.

[17] J. Abel, M. Strake, and T. Fingscheidt, "Artificial Bandwidth Extension Using Deep Neural Networks for Spectral Envelope Estimation," in *Proc. of IWAENC*, Xi'an, China, Sept. 2016, pp. 1–5.

[18] Y. Li and S. Kang, "Artificial Bandwidth Extension Using Deep Neural Network-Based Spectral Envelope Estimation and Enhanced Excitation Estimation," *IET Signal Processing*, vol. 10, no. 4, pp. 422–427, 2016.

[19] Y. Wang, S. Zhao, W. Liu, M. Li, and J. Kuang, "Speech Bandwidth Expansion Based on Deep Neural Networks," in *Proc. of Interspeech*, Dresden, Germany, Sept. 2015, pp. 2593–2597.

[20] J. Abel and T. Fingscheidt, "A DNN Regression Approach to Speech Enhancement by Artificial Bandwidth Extension," in *Proc. of WASPAA*, New Paltz, NY, USA, Oct. 2017, pp. 219–223.

[21] J. Abel and T. Fingscheidt, "Artificial Speech Bandwidth Extension Using Deep Neural Networks for Wideband Spectral Envelope Estimation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Jan. 2018, pp. 71–83.

[22] R. Peharz, G. Kapeller, P. Mowlaee, and F. Pernkopf, "Modeling Speech with Sum-Product Networks: Application to Bandwidth Extension," in *Proc. of ICASSP*, Florence, Italy, May 2014, pp. 3699–3703.

[23] B. Liu, J. Tao, Z. Wen, Ya Li, and D. Bukhari, "A Novel Method of Artificial Bandwidth Extension Using Deep Architectures," in *Proc. of Interspeech*, Dresden, Germany, Sept. 2015, pp. 2598–2602.

[24] Y. Gu, Z.-H. Ling, and L.-R. Dai, "Speech Bandwidth Extension Using Bottleneck Features and Deep Recurrent Neural Networks," in *Proc. of Interspeech*, San Francisco, CA, USA, Sept. 2016, pp. 297–301.

[25] G. Montavon, G. B. Orr, and K.-R. Müller, Eds., *Neural Networks: Tricks of the Trade*, Lecture Notes in Computer Science. Springer, 2nd edition, 2012.

[26] Daniel P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005.

[27] H. Hermansky and N. Morgan, "RASTA Processing of Speech," in *IEEE Transactions on Speech and Audio Processing*, Oct. 1994, vol. 2, pp. 578–589.

[28] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium (LDC), Philadelphia, 1993.

[29] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, "SpeechDat-Car: A Large Database for Automotive Environments," in *Proc. of LREC*, Athens, Greece, May 2000, pp. 1–6.

[30] "Multi-Lingual Speech Database for Telephonometry," NTT Advanced Technology Corporation (NTT-AT), 1994.

[31] "EVS Permanent Document EVS-7c: Processing Functions for Characterization Phase (3GPP S4 141126, V. 1.0.0)," 3GPP; TSG SA, Aug. 2014.

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

[33] "Mandatory Speech Codec Speech Processing Functions: AMR Speech Codec; Transcoding Functions (3GPP TS 26.090, Rel. 6)," 3GPP; TSG SA, Dec. 2004.

[34] "ITU-T Recommendation P.341, Transmission Characteristics for Wideband Digital Loudspeaking and Hands-Free Telephony Terminals," ITU, Mar. 2011.

[35] "Speech Codec Speech Processing Functions: AMR Wideband Speech Codec; Transcoding Functions (3GPP TS 26.190, Rel. 6)," 3GPP; TSG SA, Dec. 2004.

[36] I. Katsir, I. Cohen, and D. Malah, "Speech Bandwidth Extension Based on Speech Phonetic Content and Speaker Vocal Tract Shape Estimation," in *Proc. of EUSIPCO*, Barcelona, Spain, Aug. 2011, pp. 461–465.

[37] "ITU-T Recommendation P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," ITU, Nov. 2007.

[38] J. Abel, M. Kaniewska, C. Guillamé, W. Tirry, and T. Fingscheidt, "An Instrumental Quality Measure for Artificially Bandwidth-Extended Speech Signals," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Feb. 2017, vol. 25, pp. 384–396.

[39] "ITU-T Recommendation P.800, Methods for Subjective Determination of Transmission Quality," ITU, Aug. 1996.

[40] "ITU-T Recommendation P.56, Objective Measurement of Active Speech Level," ITU, Dec. 2011.