A DNN REGRESSION APPROACH TO SPEECH ENHANCEMENT BY ARTIFICIAL BANDWIDTH EXTENSION

Johannes Abel and Tim Fingscheidt

Institute for Communications Technology, Technische Universität Braunschweig, Germany
{j.abel, t.fingscheidt}@tu-bs.de

ABSTRACT

Artificial speech bandwidth extension (ABE) is an extremely effective means for speech enhancement at the receiver side of a narrowband telephony call. First approaches have been seen incorporating deep neural networks (DNNs) into the estimation of the upper band speech representation. In this paper we propose a regression-based DNN ABE being trained and tested on acoustically different speech databases, exceeding coded narrowband speech by a so-far unseen 1.37 CMOS points in a subjective listening test.

Index Terms— speech enhancement, artificial speech bandwidth extension, deep neural networks, regression

1. INTRODUCTION

Artificial speech bandwidth extension (ABE) belongs to speech enhancement approaches aiming at improving quality and intelligibility of narrowband (NB) telephony speech, which is typically limited to an acoustical bandwidth of $0 < f \leq 4$ kHz. ABE approaches aim at restoring missing frequency components in the upper band (UB) ($4 < f \leq 8$ kHz). Wideband (WB) telephony with transmitted frequencies in the range $0 < f \leq 8$ kHz improves both speech quality and intelligibility [1] compared to NB and therefore motivates ABE as a receiver-sided speech enhancement approach.

ABE approaches are often based on classifiers that are used to choose among precalculated UB representations for synthesis. Based on the minimal distance to a codebook containing NB representations, a matching UB representation is chosen [2, 3]. More sophisticated statistical models such as Gaussian mixture models (GMMs) [4] and hidden Markov models (HMM) [5, 6, 7, 8] have also been employed as classifiers for ABE. For ABE speech quality aspects, see also [9].

Motivated by their high phoneme recognition performance [10], deep neural networks (DNNs) have also been employed as classifiers for ABE. In [11], DNNs have been used as acoustic models for the HMM and also replaced the entire HMM.

Furthermore, neural networks have been employed to approximate a regression between features from the NB speech signal and the missing UB spectral components. Without employing the source-filter model for speech production, the UB log-power spectrum was estimated directly using a DNN [12, 13], or using a recurrent neural network in [14].

In [15, 16], DNNs are used for directly estimating a parametric representation of the UB spectral envelope, however, publications are lacking either a thorough investigation of DNN training parameters, an acoustically diverse real-world data set definition, and/or results of a (semi-)formal subjective listening test on acoustically different data.

In this work, we train DNNs to be a regression model for the link between NB speech and the missing UB, and compare the resulting model to a classification-based baseline employing a traditional HMM/GMM statistical model [6]. We define datasets for training and testing the statistical models using several different speech databases, so that results on the test set are really representative for conclusions on practical employment of ABE. Furthermore, we investigate several DNN training parameters, such as the type of unit, dropout and pretraining and report results on independent test data. Finally, we provide results of a subjective comparison category rating (CCR) test, where we compare the best classification-based and regression-based ABE approaches against coded NB and WB speech.

The paper is structured as follows: Sec. 2 presents the ABE framework, explains the baseline classification-based, and then presents the new regression-based UB spectral envelope estimation. The experimental setup, speech databases, preprocessing, statistical model training, and instrumental assessment of the estimated UB spectral envelopes is presented in Sec. 3. Subsequently, experimental results are presented in Sec. 4. A comparison category rating (CCR) test is conducted in Sec. 5. Conclusions are drawn in Sec. 6.

2. ABE FRAMEWORK

The ABE framework based on [6] is presented in Fig. 1. The system inputs a NB speech signal $s^{\mathrm{NB}}(n')$ with n' being the $f_s' = 8$ kHz sample index. Making use of the source-filter model for speech production, the estimation of a suitable UB speech signal $\hat{s}^{\mathrm{UB}}(n)$, with n being the 16 kHz sample index, is achieved by estimating an UB residual and UB spectral envelope separately. The framework outputs $\hat{s}^{\mathrm{WB}}(n)$, which is the superposition of the interpolated NB signal $s^{\mathrm{NB}}(n)$ and the estimated UB signal $\hat{s}^{\mathrm{UB}}(n)$.

The UB residual signal is found via simple modulation of the interpolated NB signal with $(-1)^n$ [17], also known as spectral folding, which mirrors the NB residual signal into the UB frequency range. To prevent annoying artifacts during speech pauses, a voice activity detection (VAD) [18] which outputs hard-decision values $VAD_\ell \in \{0, 1\}$, with ℓ being the 10 ms frame index, is employed for controlling the residual extension. Subsequent linear prediction (LP) synthesis filtering using WB LP filter coefficients $\tilde{\mathbf{a}}_\ell^{\mathrm{WB}}$ leads to the desired UB speech signal $\tilde{s}^{\mathrm{UB}}(n)$. Finally, spectral floor suppression (SFS) is employed to control the synthesized energy in the UB for sounds with higher and lower UB energy separately to further improve speech quality [19].

A WB power spectrum density (PSD) representation $\tilde{\mathbf{\Phi}}_\ell^{\mathrm{WB}}$ is obtained, consisting of the NB PSD $\mathbf{\Phi}_\ell^{\mathrm{NB}}$, calculated from the interpolated NB speech signal, and the estimated UB envelope PSD $\tilde{\mathbf{\Phi}}_\ell^{\mathrm{UB}}$. Next, auto-correlation function (ACF) coefficients are ob-

978-1-5386-1631-4/17/$31.00 ©2017 IEEE

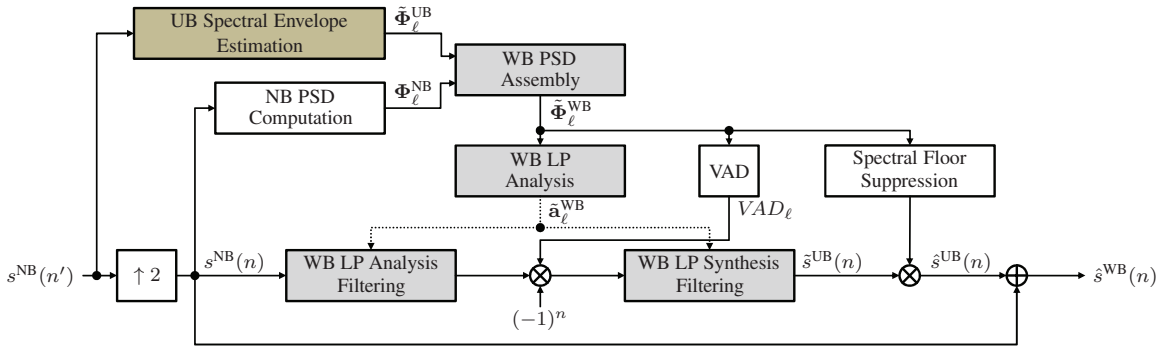

Figure 1: Block diagram of the **artificial bandwidth extension framework**. The input signal $s^{\mathrm{NB}}(n')$ with sampling rate $f_s' = 8$ kHz is artificially extended providing the artificial WB speech signal $\hat{s}^{\mathrm{WB}}(n)$ sampled at $f_s = 16$ kHz as output.

tained via inverse DFT of the WB PSD $\tilde{\mathbf{\Phi}}_\ell^{\mathrm{WB}}$, and subsequent application of the well-known Levinson-Durbin recursion leads to the desired WB LP filter coefficients $\tilde{\mathbf{a}}_\ell^{\mathrm{WB}}$.

Due to the fact that the UB PSD contains a lot of redundant information, we estimate the UB spectral envelope in parametric form, i.e., a cepstral vector $\tilde{\mathbf{y}}_\ell$, which is then converted into the spectral domain. The cepstral vector $\mathbf{y}_\ell$ in general consists of a relative UB energy coefficient

$$\mathrm{y}_\ell(0) = \frac{1}{\sqrt{2}} \cdot \ln\left(\frac{g_\ell^{\mathrm{UB}}}{g_\ell^{\mathrm{NB}}}\right), \tag{1}$$

with g_ℓ^{NB} and g_ℓ^{UB} being the prediction gain factors from LP analysis for the NB and UB PSD, respectively, while the remaining coefficients $\mathrm{y}_\ell(1), \ldots, \mathrm{y}_\ell(N_{\mathrm{UB}})$ are cepstral coefficients describing the UB envelope [20]. For statistical model training, $\mathbf{y}_\ell$ is calculated from WB speech data via selective linear prediction (SLP) analysis [20], while during ABE processing, an estimate $\tilde{\mathbf{y}}_\ell$ is provided.

The UB spectral envelope PSD

$$\tilde{\Phi}_\ell^{\mathrm{UB}}(k) = g_\ell^{\mathrm{NB}} \cdot \frac{\tilde{g}_\ell^{\mathrm{UB}}}{\tilde{g}_\ell^{\mathrm{NB}}} \cdot \exp\left(2\mathrm{Re}\left\{\sum_{\nu=1}^{N_{\mathrm{UB}}} \tilde{\mathrm{y}}_\ell(\nu) e^{-j2\pi \frac{\nu k}{|\mathcal{K}^{\mathrm{UB}}|}}\right\}\right), \tag{2}$$

with $\mathcal{K}^{\mathrm{UB}}$ being the set of bins corresponding to the UB frequency range and $|.|$ the cardinality of the set, is calculated from the estimated cepstral vector $\tilde{\mathbf{y}}_\ell$, where g_ℓ^{NB} is obtained during ABE processing from the NB speech signal and $\tilde{g}_\ell^{\mathrm{UB}}/\tilde{g}_\ell^{\mathrm{NB}}$ stems from the estimated $\tilde{\mathrm{y}}_\ell(0)$ (c.f. (1)).

In this work, we compare two methodologies for estimating $\tilde{\mathbf{y}}_\ell$: The baseline approach is referred to as **ABE-C** (Sec. 2.1) and employs an HMM/GMM-based classifier, which chooses an UB spectral envelope among a set of clustered UB cepstral vectors. Furthermore, we evaluate a new regression-based approach, referred to as **ABE-R** (Sec. 2.2), to directly determine an estimate of the UB cepstral vector $\tilde{\mathbf{y}}_\ell$ without calculating probabilities or having the need for a codebook.

2.1. Baseline: UB Estimation via Classification ("ABE-C")

The baseline classification-based ABE approach **ABE-C** estimates a posteriori probabilities $\mathrm{P}(s_\ell = i|\mathbf{x}_\ell)$ with $i = 1, \ldots, N$ being the state index and $\mathbf{x}_\ell$ the feature vector. The underlying statistical model is an HMM with a GMM as the acoustic model, decoded via the forward algorithm [21]. Input to the HMM/GMM is the feature vector $\mathbf{x}_\ell$, which includes a parametric representation of the NB spectral envelope in form of ten ACF coefficients as well as five additional features being highly relevant for the ABE task [20]. In addition, temporal dynamics (Δ and $\Delta\Delta$) are calculated, leading to a feature vector $\mathbf{x}_\ell$ with dimension $(10 + 5) \cdot 3 = 45$ per frame ℓ.

Each HMM state corresponds to a pretrained quantized UB cepstral vector $\hat{\mathbf{y}}_i$ in the codebook [6]. Using minimum mean square error estimation, the *a posteriori* probabilities $\mathrm{P}(s_\ell = i|\mathbf{x}_\ell)$ serve as weights for the $N = 24$ codebook entries $\{\hat{\mathbf{y}}_i\}$, which leads to an estimated UB cepstral vector of dimension $N_{\mathrm{UB}} + 1 = 9$:

$$\tilde{\mathbf{y}}_\ell = \sum_{i=1}^{N} \mathrm{P}(s_\ell = i|\mathbf{x}_\ell) \cdot \hat{\mathbf{y}}_i. \tag{3}$$

The estimated UB cepstral vector is then converted to the spectral power domain via (2).

2.2. New: UB Estimation via Regression ("ABE-R")

The regression-based UB spectrum estimation approach directly outputs an estimate of the UB cepstral vector ($N_{\mathrm{UB}} + 1 = 9$ again)

$$\tilde{\mathbf{y}}_\ell = \mathbf{f}(\mathbf{x}_\ell; \Theta), \tag{4}$$

where function $\mathbf{f}$ is realized by a feed-forward DNN with feature vector $\mathbf{x}_\ell$ input and the parameters Θ. The feature vector definition employed for regression uses 29 log-mel filterbank coefficients [21] instead of 10 ACF coefficients as parametric representation of the NB spectral envelope, which leads to a dimension of $(29+5) \cdot 3 = 102$. The estimated UB cepstral vector is then converted to the spectral power domain via (2).

3. EXPERIMENTAL SETUP

3.1. Speech Data Sets

The TIMIT database [22] and the Speechdat-Car US (SDC) database [23] provide speech data for training of the statistical models. From both databases, approx. 30% are jointly used as a validation set for DNN training (c.f. Sec. 3.3). Speech data for the test set is taken from the acoustically different German and American English parts of the NTT database [24].

Preprocessing of the speech data follows [11]. Input to the ABE is a preprocessed NB speech signal, which was subject to MSIN

filtering [25], decimation to 8 kHz, 16 to 13 bit conversion, adaptive multirate (AMR) coding and decoding [26] at 12.2 kbps and final 16 to 13 bit conversion. WB preprocessing is conducted by P.341-filtering [25] of the available WB speech data. Delay introduced in this preprocessing scheme was compensated for.

3.2. HMM/GMM Training ("ABE-C")

All UB cepstral envelope vectors $\mathbf{y}_\ell$ calculated via SLP analysis of the training data are vector-quantized and thereby assigned an index $i_\ell \in \{1, \ldots, N\}$, which represents a codebook entry [6]. Besides calculating state and state transition probabilities on the training data, for each state a GMM with G mixture components is trained via the expectation-maximization (EM) algorithm [27] on data extracted from the training set. A linear discriminant analysis (LDA) of the extracted features reduces the amount of parameters to train for the GMMs. The LDA transformation matrix $\mathbf{H}$ is also found on data derived from the training set, reducing the feature vector from 45 to 9 dimensions. For **ABE-C**, we will investigate the influence of the number of modes per GMM on the UB estimation quality.

3.3. DNN Training ("ABE-R")

In extensive preliminary investigations [19], we found that topology of the feed-forward DNN only has a moderate influence on the performance of the statistical model, resulting in a DNN with a total of 4 hidden layers, each with 512 units for all DNN experiments in this work. The network is trained using backpropagation (BP) [27] with learning rate $\eta = 0.1$, momentum and L2-regularization. The mean-squared error (MSE) criterion is used for training. A maximum of 50 epochs or an increasing error on the validation data leads to early stopping of the BP algorithm.

We employ sigmoid units or rectified linear units (ReLUs). Weights and biases are either initialized randomly [28] or via a restricted Boltzmann machine (RBM) pretraining [29] (sigmoid units only). For further increasing the generalization power of the DNN we employ dropout [30], which will include only a subset of units during an epoch of BP. Zero mean and unit variance is fulfilled for the training data, while validation and test data is normalized using statistics from training data. We investigate the influence of pretraining, dropout, and the choice of the unit type for the DNN employed in **ABE-R**.

3.4. Metrics for UB Spectral Envelope Quality

Estimated UB spectral envelopes are evaluated instrumentally by calculating several cepstral distances between the estimated cepstral vector $\tilde{\mathbf{y}}$ and the correct cepstral vector $\mathbf{y}$ from an SLP analysis of the WB speech signal. The overall cepstral distance is calculated [31]:

$$D = 10\sqrt{2} \cdot \log_{10}(e) \sqrt{\sum_{\nu=0}^{N_{\text{UB}}} \left(\mathrm{y}(\nu) - \tilde{\mathrm{y}}(\nu)\right)^2} \text{ [dB]}. \quad (5)$$

Considering only the cepstral envelope (w/o relative UB energy coefficient (1)),

$$D_{\text{env}} = 10\sqrt{2} \cdot \log_{10}(e) \sqrt{\sum_{\nu=1}^{N_{\text{UB}}} \left(\mathrm{y}(\nu) - \tilde{\mathrm{y}}(\nu)\right)^2} \text{ [dB]} \quad (6)$$

is employed for evaluation. Focusing only on the very important UB relative energy ratio $\tilde{\mathrm{y}}(0)$,

$$D_0 = 10\sqrt{2} \cdot \log_{10}(e) \cdot \left|\mathrm{y}(0) - \tilde{\mathrm{y}}(0)\right| \text{ [dB]} \quad (7)$$

reports a cepstral distance independent of the estimated UB cepstral envelope. The reported metrics have been averaged over file means. Instrumental assessment of speech quality is conducted using WB-PESQ [32] and reported in terms of mean opinion score - listening quality objective (MOS-LQO).

4. EXPERIMENTS

Table 1 presents the resulting cepstral distances and MOS-LQO scores for all conducted experiments.

The classification-based ABE approach **ABE-C**, which uses an HMM/GMM statistical model as described in Sec. 2.1, performs better for a higher number of modes G in terms of lower cepstral distances and higher MOS-LQO points on the validation data. On the test data, however, the smallest number of modes ($G = 4$) leads to the lowest cepstral distances. This trend cannot be observed for the MOS-LQO points, which increase with a higher number of modes on the test data. Still, the higher cepstral distances on the test data for a higher number of modes per GMM indicate overfitting of the acoustic model to the training data. Using the GMM with $G = 32$ modes leads to the lowest cepstral distances and highest MOS-LQO points in the validation data over all of the **ABE-C** experiments, henceforth serving as baseline.

The regression-based ABE approach **ABE-R**, which uses a DNN as statistical model as described in Sec. 2.2, outperforms all **ABE-C** conditions. Regarding the use of dropout, we cannot observe any increased generalization power in the presented experiments, especially w.r.t. the test set. The validation set performance of **ABE-R** when using ReLUs (w/o dropout) is comparable to using pretrained sigmoid units. The best **ABE-R** experiment in terms of lowest cepstral distances on validation data was achieved by a DNN with pretrained sigmoid units.

In line with the lowest cepstral distances, WB-PESQ attests the **ABE-R** approach also the highest MOS-LQO scores. For comparison, we included the results of an oracle experiment, where always the correct $\mathbf{y}_\ell$ from SLP analysis of the original WB speech data is used. Being only 0.18 MOS-LQO points behind the oracle experiment, **ABE-R** outperforms the **ABE-C** baseline on the validation data by an impressive 0.6 MOS-LQO points. On the test data, still 0.23 MOS-LQO points improvement were noted with a gap of 0.43 points to the oracle.

In summary, both WB-PESQ as well as cepstral distance measures attest the **ABE-R** approach a superior performance compared to the baseline **ABE-C** approach. On the test set, the best **ABE-R** approach reduces cepstral distance D w.r.t. the **ABE-C** baseline ($G = 32$) by $10.33\text{ dB} - 9.15\text{ dB} = 1.18\text{ dB}$. Both approaches will be evaluated subjectively in the next section using the respective parameter setting which led to the lowest cepstral distances on the validation data.

5. SUBJECTIVE SPEECH QUALITY ASSESSMENT

Subjective speech quality is tested in a semi-formal CCR test [33, Annex E]. In such a test, 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) in steps of 1. From the German part of the NTT database two female and two male speakers, each

Statistical Model		Validation Set				Test Set			
		D	D_0	D_{env}	MOS-LQO	D	D_0	D_{env}	MOS-LQO
ABE-C	HMM/GMM with $G = 4$ modes	8.12	6.29	4.12	2.56	**9.98**	**7.79**	**5.32**	2.53
	HMM/GMM with $G = 8$ modes	7.95	6.09	4.11	2.62	10.31	8.11	5.45	2.54
	HMM/GMM with $G = 16$ modes	7.90	6.06	**4.10**	2.66	10.11	**7.79**	5.53	2.54
	HMM/GMM with $G = 32$ modes	**7.86**	**6.01**	**4.10**	**2.67**	10.33	8.04	5.50	**2.57**
ABE-R	DNN with sigmoid units	6.41	4.59	3.72	**3.27**	9.39	7.30	5.16	**2.80**
	DNN with sigmoid units and dropout	6.83	4.97	3.89	3.13	9.93	7.81	5.41	2.68
	DNN with pretrained sigmoid units	**6.35**	**4.56**	**3.70**	**3.27**	**9.15**	**7.10**	**5.05**	**2.80**
	DNN with ReLUs	6.36	**4.56**	3.71	3.26	9.37	7.24	5.19	2.79
	DNN with ReLUs and dropout	7.13	5.29	3.88	3.12	10.01	7.84	5.41	2.68
Oracle		0	0	0	3.45	0	0	0	3.23

Table 1: **Cepstral distances** [dB] and **instrumental speech quality assessment** by WB-PESQ in various settings. Bold numbers mark the best results for each of the statistical models.

CCR Condition	CMOS	CI_{95}
AMR vs. **AMR-WB**	2.15	$[2.03; 2.26]$
ABE-C vs. **AMR-WB**	1.48	$[1.35; 1.61]$
ABE-R vs. **AMR-WB**	1.31	$[1.18; 1.44]$
ABE-C vs. **ABE-R**	0.13	$[0.01; 0.24]$
AMR vs. **ABE-C**	0.81	$[0.60; 1.03]$
AMR vs. **ABE-R**	1.37	$[1.22; 1.51]$

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

with 4 sentences, have been chosen. Twelve native-speaking Germans without known hearing impairment participated in the CCR test. Four conditions are evaluated in the CCR test:

- **AMR**: Coded NB speech, processed as described in Sec. 3.1
- **AMR-WB**: Coded WB speech, derived from the WB preprocessed speech signal as described in Sec. 3.1 with subsequent coding and decoding using AMR-WB operating at 12.65 kbit/s [34]
- **ABE-C**: Classification-based ABE as described in Sec. 2.1 with $G = 32$ modes and **AMR** data as input signal
- **ABE-R**: New regression-based ABE as described in Sec. 2.2 with pretrained sigmoid units and **AMR** data as input signal

In total, six CCR comparisons are presented in the subjective test. P.341-conformant bandpass-filtering to a frequency range of $0.2 \ldots 7$ kHz [35, 36, 37], active speech level scaling to -26 dBov [38], and conversion to 48 kHz sampling rate were conducted before presenting the speech signals to the subjects.

The signals under test were presented in diotic fashion using two PCs with `RME Fireface 400` sound cards using `AKG K-271 MKII` headphones. After familiarization, which included all test conditions, the subjects judged 36 speech file pairs, each in both orders, leading to 72 comparisons in total. Two randomized sets of comparisons were presented to the subjects, balanced over speakers and conditions.

The results of the listening test are presented in Tab. 2 in terms of CMOS and respective 95% confidence interval (CI_{95}) for each of the CCR conditions. Obviously, **AMR-WB** outperformed **AMR** indicated by 2.15 CMOS points. Compared to the ABE approaches, **AMR-WB** is better than **ABE-C** by 1.48 CMOS points and better than **ABE-R** by 1.31 CMOS points. **ABE-C** is outperformed by **ABE-R** by a significant 0.13 CMOS points. Compared to **AMR**, the **ABE-C** baseline improves the underlying NB condition by 0.81 CMOS points, while the new **ABE-R** approach exceeds the **AMR** condition by an impressive 1.37 CMOS points. With **ABE-R** improving **AMR** by 1.37 CMOS points and **AMR-WB** exceeding **ABE-R** by 1.31 CMOS points, we can conclude that the **ABE-R** approach fills half of the gap between NB- and WB-coded speech.

6. CONCLUSIONS

In this work, we employed a deep neural network (DNN) as regression-based statistical model, directly estimating the upper band spectral envelope in the context of artificial speech bandwidth extension (ABE) and compared the results to a typical classification-based HMM/GMM ABE baseline. Using pretrained sigmoid units, the DNN improved the UB cepstral distance on the test set by more than 1.18 dB, and resulting ABE-processed speech signals were found to be improved by 0.23 MOS points vs. the HMM/GMM baseline. A yet unreported DNN ABE advantage of distinctive 1.37 CMOS points compared to the underlying narrowband telephony speech signals could be shown in a subjective CCR listening test.

7. REFERENCES

[1] N. R. French and J. C. Steinberg, "Factors Governing the Intelligibility of Speech Sounds," *Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

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

[3] J. Sadasivan, S. Mukherjee, and C. S. Seelamantula, "Joint Dictionary Training for Bandwidth Extension of Speech Signals," in *Proc. of ICASSP*, Shanghai, China, Mar. 2016, pp. 5925–5929.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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 Listenining 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.

[10] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[11] 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.

[12] K. Li and C.-H. Lee, "A Deep Neural Network Approach to Speech Bandwidth Expansion," in *Proc. of ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4395–4399.

[13] B. Liu, J. Tao, Z. Wen, Y. 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.

[14] 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, USA, Sept. 2016, pp. 297–301.

[15] 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.

[16] 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.

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

[18] B. Fodor and T. Fingscheidt, "Reference-free SNR Measurement for Narrowband and Wideband Speech Signals in Car Noise," in *Proc. of 10th ITG Conference on Speech Communication*, Braunschweig, Germany, Sept. 2012, pp. 199–202.

[19] J. Abel and T. Fingscheidt, "Artificial Speech Bandwidth Extension Using Deep Neural Networks for Wideband Spectral Envelope Estimation," *submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing*

[20] P. Jax, "Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds," Ph.D. dissertation, vol. 15 of P. Vary (ed.), Aachener Beiträge zu digitalen Nachrichtensystemen, 2002.

[21] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall, 2001.

[22] 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.

[23] 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.

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

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

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

[27] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

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

[29] G. E. Hinton, "A Practical Guide to Training Restricted Boltzmann Machines," Tech. Rep. UTML TR 2010-003, Dept. Comput. Sci., Univ. Toronto, 2010.

[30] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[31] R. Hagen, "Spectral Quantization of Cepstral Coefficients," in *Proc. of ICASSP*, Adelaide, Australia, Apr. 1994, pp. 509–512.

[32] "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.

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

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

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

[36] 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.

[37] 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.

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