ARTIFICIAL BANDWIDTH EXTENSION USING DEEP NEURAL NETWORKS FOR SPECTRAL ENVELOPE ESTIMATION

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

Many artificial speech bandwidth extension (ABE) approaches perform source-filter decomposition of the input narrowband speech, with subsequent computation of upper frequency band (UB) spectral envelope posteriors. In this paper we perform a *direct comparison* of HMM- and deep neural network (DNN)-based modeling of likelihoods or posteriors for ABE UB envelope estimation. DNN-based approaches turn out to significantly exceed GMM-based ones in speech quality. Further analysis reveals that this is *not* due to a better shape of the estimated UB spectral envelope, but primarily due to a much better estimate of the energy ratio of the upper band vs. the lower band – an important result with significant impact on ABE speech quality particularly for fricative sounds.

***Index Terms*—** artificial speech bandwidth extension, deep neural network, spectral envelope estimation

1. INTRODUCTION

Artificial speech bandwidth extension (ABE) algorithms provide speech enhancement in the receive path of telephone calls. ABE tries to compensate for limited acoustical bandwidth, lost somewhere during transmission due to, e.g., speech coding, or restraining device characteristics. Typically, narrowband (NB) speech ($0 < f \leq 4$ kHz) is enhanced by an ABE solution, which estimates and synthesizes the upper band (UB), i.e., frequency components from 4 kHz to 8 kHz and thus leading to an artificial wideband (WB) signal ($0 < f \leq 8$ kHz). Nowadays more and more telephone calls are operated in native WB mode [1], known as HD Voice, improving the perceived speech quality compared to a NB call [2]. For establishing an HD Voice call, however, many requirements need to be met: Mobile handsets need to be WB-capable, underlying infrastructure needs to be WB-capable, and the mobile cells assigned to caller and callee need to be WB-capable. Additionally, inter-operator calls might also prevent a WB call setup. Therefore, ABE solutions will play an important role as fallback to deliver a speech quality as good as possible whenever an HD Voice call cannot be established.

Traditionally, ABE algorithms often make use of the source-filter model for speech production: The problem of finding an UB speech signal is split into finding an UB residual signal (source) and an UB spectral envelope (filter) separately. While for UB residual generation simple modulation techniques are employed, e.g., spectral folding [3], most efforts aim at estimating an UB spectral envelope. The connection between the observable NB signal and the desired UB spectral envelope can be constituted using a simple linear codebook mapping [4]. Research evolved towards more powerful (and non-linear) statistical models, such as Gaussian mixture models (GMMs) [5] and hidden Markov models (HMMs) [6–9].

Bauer et al. successfully employed neural networks for fricative sound classification in ABE [10]. Without explicit separation according to the source-filter model, in [11] the NB spectrum is duplicated and shifted to the UB via spectral folding and then shaped via a time-variant spline function, controlled by the output of a neural network.

Motivated by the estimation power of deep architectures for automatic speech recognition [12], current ABE-related research tends to omit the source-filter model and lets the deep architectures take over. In [13] UB frequency bins are directly estimated by making use of a deep neural network (DNN) and subsequently synthesized to an artificial WB signal using an overlap-add structure. Perharz et. al investigated several statistical models such as Gaussian-binary restricted Boltzman machines (RBMs), conditional RBMs, (deep) autoencoders, and sum-product networks [14, 15] for estimation of log spectrograms. In [16] the log power spectrum is directly estimated using a DNN. Furthermore, DNN-based cepstral estimation for ABE is conducted in [17].

Aiming at practical product employment, however, still the source-filter model-based ABE approaches incorporating a DNN promise to have the best generalization properties (as compared to end-to-end DNNs), and can nicely be parameterized. In this publication a source-filter model-based ABE approach derived from [7] is employed for a direct comparison of GMM- and DNN-based acoustic models for spectral envelope estimation. Several DNN topologies are evaluated in the context of UB spectrum estimation, evaluating different aspects such as number of hidden layers, number of units each hidden layer consists of, and the number of free parameters. A further important aspect of our evaluation will be the cross-database training and test setup, allowing more realistic conclusions w.r.t. generalization aspects of GMMs and DNNs.

The paper is structured as follows. First the ABE framework is presented in Section 2, along with the generation of training data for statistical model training. Subsequently, the experimental setup is explained in Section 3. In Section 4 experiments are presented and finally conclusions are drawn in Section 5.

2. THE ABE FRAMEWORK

In this section the ABE framework is explained. Special focus lies on the UB spectral envelope codebook design and the generation of training data for statistical model training.

2.1. Overview

The employed ABE approach is depicted in Fig. 1. Being based on the source-filter model for speech production, the extension process is divided into the steps of identifying an UB residual signal and a suitable UB spectral envelope, respectively.

For UB spectral envelope estimation, first the 15 static features (cf. [7]) are extracted from a 10 ms frame of the 8 kHz input sig-

978-1-5090-2007-2/16/$31.00 ©2016 IEEE

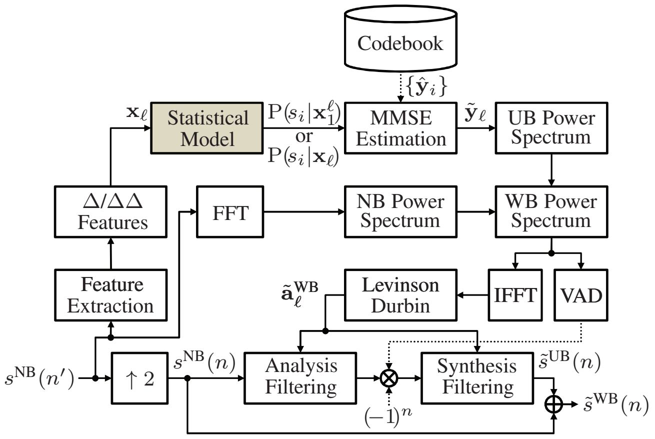

Fig. 1. Block diagram of the ABE approach employing a statistical model block (shaded in gray) as a placeholder for the GMM-based (Sec. 3.1) and DNN-based (Sec. 3.2) acoustic models.

nal $s^{\mathrm{NB}}(n')$, where n' is the 8 kHz sample index. Subsequently, first- and second-order temporal derivatives (Δ/ΔΔ features) are calculated, using only a single frame look-ahead to maintain low latency as required for telephony ABE approaches, leading to the 45-dimensional feature vector $\mathbf{x}_\ell$, with ℓ being the frame index. The feature vector is then subject to a statistical model for calculation of posterior probabilities $\mathrm{P}(s_i|\mathbf{x}_1^\ell)$, with $\mathbf{x}_1^\ell$ denoting the sequence $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\ell\}$. Alternatively, $\mathrm{P}(s_i|\mathbf{x}_\ell)$ is being computed. The HMM state s_i corresponds to a quantized UB spectral envelope vector $\hat{\mathbf{y}}_i$ from the codebook, found in a preceding training step, with i being the state index. Subsequently, $\tilde{\mathbf{y}}_\ell = \sum_{i=1}^{N} \hat{\mathbf{y}}_i \cdot \mathrm{P}(s_i|\mathbf{x}_1^\ell)$ leads to the minimum mean square error (MMSE) estimated UB spectral envelope.

Afterwards, the estimated spectral envelope is converted to the power spectrum and appended to the NB power spectrum, resulting in a WB power spectrum. Autocorrelation coefficients are then calculated by inverse fast Fourier transform (IFFT). Finally, filter coefficients $\mathbf{a}_\ell^{\mathrm{WB}}$ are derived via the Levinson-Durbin recursion.

In order to obtain an UB residual signal, input signal $s^{\mathrm{NB}}(n')$ is interpolated to $f_s = 16$ kHz and linear prediction (LP) analysis-filtered using the formerly found filter coefficients $\mathbf{a}_\ell^{\mathrm{WB}}$ leading to an interpolated NB residual signal. Spectral folding is performed by multiplying every sample by factor $(-1)^n$, with n being the sample index for 16 kHz signals, and a further multiplication with a voice activity detection (VAD) output $VAD_\ell \in \{0, 1\}$ [18]. The resulting UB residual signal is then LP synthesis-filtered using again the filter coefficients $\mathbf{a}_\ell^{\mathrm{WB}}$ and thereby applying the estimated UB envelope to the estimated UB residual signal.

Finally, the interpolated NB speech signal $s^{\mathrm{NB}}(n)$ and the estimated UB speech signal $\tilde{s}^{\mathrm{UB}}(n)$ are summed up and constitute the artificial WB signal $\tilde{s}^{\mathrm{WB}}(n)$.

2.2. UB Spectral Envelope Codebook

In general, UB spectral envelopes are represented as vectors $\mathbf{y}_\ell \in \mathbb{R}^{N_\nu}$, with $N_\nu = 9$, composed of a relative UB-to-NB log-energy ratio

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

and 8 UB cepstral envelope values $\mathrm{y}_\ell(1), \ldots, \mathrm{y}_\ell(N_\nu - 1)$. The log-energy ratio $\mathrm{y}_\ell(0)$ relates the UB LP prediction gain g_ℓ^{UB} to the NB LP prediction gain g_ℓ^{NB} during training.

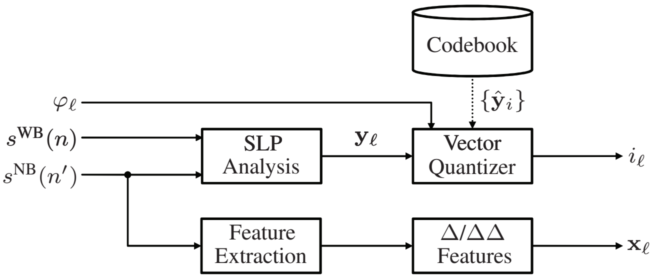

Fig. 2. Generation of input ($\mathbf{x}_\ell$) and target (i_ℓ) pairs for statistical model training.

Only those UB spectral envelopes that can be represented as linear combination of the codebook entries $\hat{\mathbf{y}}_i$ can be the result of the UB MMSE spectral envelope estimation process. A simple data-driven Linde-Buzo-Gray (LBG) algorithm of envelopes extracted from WB speech training data results in an insufficient representation of alveolar fricative sounds /s/ and /z/ [7], which, however, are exactly those sounds that miss most of their characteristical spectral content when limited to a NB-typical acoustical bandwidth and thus might be of most importance to an ABE approach. Therefore, the transcription-based codebook design presented in [7, Sec. 4.1.2], employing the phonetic class information φ_ℓ, is chosen, which designates 8 codebook entries for spectral envelopes of high-energy /s/ and /z/ sounds ($\varphi_\ell = 1$) and further 16 entries for the remaining sounds ($\varphi_\ell = 0$). In total, $N = 24$ envelopes are in the codebook and consequently the HMM consists of $N = 24$ states.

2.3. Training Data for the Statistical Model

The statistical model aims at determining the correct envelope index i_ℓ (target) given the current feature vector $\mathbf{x}_\ell$ (input). Input and target data for training is generated as depicted in Fig. 2. Input to the calculation is the ℓ-th frame of the WB speech $s^{\mathrm{WB}}(n)$, of the NB speech $s^{\mathrm{NB}}(n')$, and a phonetic class information φ_ℓ. Please note that these signals need to be time-synchronous.

The input data is calculated in the bottom branch of the block diagram via static feature extraction and subsequent consideration of the temporal dynamics leading to feature vector $\mathbf{x}_\ell$.

For calculation of the target data, first selective linear prediction (SLP) analysis [19, 20] is performed using both the NB and the WB speech signals. By means of the spectral envelope codebook the calculated UB spectral envelope vector (in cepstral representation) $\mathbf{y}_\ell$ is quantized. Please note that the phonetic information φ_ℓ is used for supervising the quantization process, meaning that frames transcribed by /s/ or /z/ sounds are quantized only using the designated 8 codebook entries, while in all other frames the quantizer selects only among the remaining 16 codebook entries. The quantization result, i.e., the envelope index i_ℓ, serves as target for statistical model training.

For finding suitable spectral envelopes of the codebook and statistical model training, P.341-filtered [21] WB speech data and transcriptions were taken from the TIMIT database [22], while the actual ABE processing and thus the evaluation of the statistical models is done using American English and German speech data from the NTT-AT database [23]. Respective NB speech signals were obtained following largely the preprocessing steps as defined in [24], namely the subsequent application of the mobile station input (MSIN) filter [25], decimation to 8 kHz, 16 to 13 bit conversion, adaptive multirate

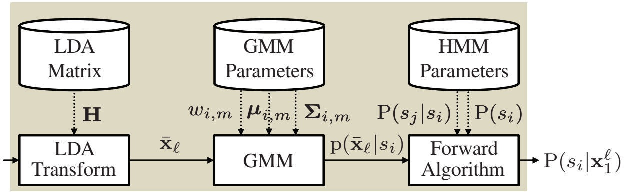

Fig. 3. GMM-driven statistical model processing.

(AMR) coding at 12.2 kbps and immediate decoding [26], followed by a final 16 to 13 bit conversion. Any introduced delay during preprocessing was compensated for.

3. STATISTICAL MODELS: GMM AND DNN

In this section the integration of the two competing statistical models into the ABE algorithm is explained. Furthermore, the underlying training processes are briefly explained.

3.1. Baseline: GMM-Driven Statistical Model

Based on all input and target data pairs $(\mathbf{x}_\ell, i_\ell)$ (Fig. 2), a linear discriminant analysis (LDA) transform matrix $\mathbf{H}$ is calculated and used to reduce the feature vector dimension from 45 to 10. The reduced feature vector is referred to as $\bar{\mathbf{x}}_\ell$. For each of the $N = 24$ UB spectral envelope entries in the codebook the respective reduced feature vectors are input to the expectation maximization (EM) algorithm to obtain a GMM with $M = 8$ modes. Please note that the EM algorithm also uses $\mathbf{y}_\ell$ to control mixture splitting.

As shown in Fig. 3, during ABE processing, the feature vector is reduced using the same LDA transform matrix $\mathbf{H}$ as in the training step. Each of the $N = 24$ GMMs is then evaluated with the current reduced feature vector $\bar{\mathbf{x}}_\ell$. The resulting likelihood $\mathrm{p}(\bar{\mathbf{x}}_\ell|s_i)$ together with the precalculated initial state probability $\mathrm{P}(s_i)$ and the state transition probability $\mathrm{P}(s_j|s_i)$ are subject to the forward algorithm for obtaining the posterior probabilities $\mathrm{P}(s_i|\bar{\mathbf{x}}_1^\ell) = \mathrm{P}(s_i|\mathbf{x}_1^\ell)$.

3.2. DNN-Driven Statistical Model

For posterior probability estimation a hybrid DNN-HMM approach and a direct DNN-only approach are investigated. Both make use of fully connected feed-forward DNNs with sigmoid activation functions and a softmax output layer, using the same training procedure. DNN pre-training and fine-tuning is done according to [27] on the TIMIT corpus training set. The TIMIT test set is used for validation during fine-tuning.

Initial weights and biases were obtained with RBM-based generative pre-training [28] using the 45-dimensional features $\mathbf{x}_\ell$ as inputs. The first layer (Gaussian-binary RBM) is trained for 225 epochs with a learning rate of 0.002 and the remaining layers (binary-binary RBMs) are trained for 75 epochs with a learning rate of 0.02.

The subsequent fine-tuning is done with standard backpropagation using a cross-entropy error function with targets i_ℓ and net inputs $\mathbf{x}_\ell$. The learning rate is chosen to be 0.1 at the start of training. When the validation error increases between epochs the weights and biases are reset to those of the last epoch and the learning rate is halved. Training is stopped once the learning rate gets smaller than 0.001.

Both pre-training and fine-tuning use a momentum of 0.9 and an additional L2 weight-decay of 0.0002, but during fine-tuning momentum is not applied in the first epoch. Both methods are implemented according to [29].

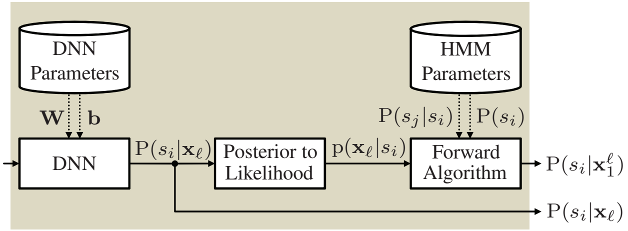

Fig. 4. DNN-driven statistical model processing.

Hybrid DNN-HMM processing is done as shown in the upper signal branch in Fig. 4. Given the input vector $\mathbf{x}_\ell$ and the trained parameters, the DNN outputs the posterior probabilitiy $\mathrm{P}(s_i|\mathbf{x}_\ell)$, which is then converted to the likelihood $\mathrm{p}(\mathbf{x}_\ell|s_i)$ by dividing out the state prior $\mathrm{P}(s_i)$. This likelihood is fed to the forward algorithm to determine the posterior probability $\mathrm{P}(s_i|\mathbf{x}_1^\ell)$ based on the current and all preceding feature vectors. The DNN-only processing path simply uses the posterior probabilities $\mathrm{P}(s_i|\mathbf{x}_\ell)$ as provided by the DNN. In both approaches the next step is the MMSE estimation of the spectral envelope, followed by all subsequent processing steps as depicted in Fig. 1.

4. EXPERIMENTAL EVALUATION

The estimated spectral envelopes are evaluated using the estimated spectral envelope vector $\tilde{\mathbf{y}}$ and the reference envelope vector $\mathbf{y}$, both in cepstral representation, extracted from the WB speech signal using SLP analysis. The cepstral distance is defined as [30]

$$CD = 10\sqrt{2} \cdot \log_{10}(e) \sqrt{\sum_{\nu} \left(\mathrm{y}(\nu) - \tilde{\mathrm{y}}(\nu)\right)^2}, \qquad (2)$$

quantifying the estimation quality. Since the value of $\tilde{\mathrm{y}}_0$ (log-energy ratio (1)) plays such an important role in ABE approaches, we will report separately on CD_0, where only $10\sqrt{2} \cdot \log_{10}(e)|\mathrm{y}(0) - \tilde{\mathrm{y}}(0)|$ is evaluated, and on CD_{1-8}, where the summation index in (2) ranges from $\nu \in \{1, \ldots, N_\nu - 1 = 8\}$. Additionally, the WB perceptual evaluation of speech quality (WB-PESQ) [31] is used for result reporting. An accuracy (ACC) of the statistical model for classifying the correct state/spectral envelope is calculated by applying the maximum posteriori rule to the calculated posterior probabilities, and subsequent comparison to the reference class indices i_ℓ obtained as shown in Fig. 2.

Results are presented in Table 1. Several DNNs were trained and evaluated, covering the number of hidden layers (HLs) from 1 to 6, and for each of them a number of units 256, 512, and 1024 were examined.

First of all, the accuracies on the validation set taken during the last epoch of DNN fine-tuning show improvement for adding hidden layers. This trend stops at a layer size of 5 for 256 units and a layer size of 4 for 512 and 1024 units. We observed a maximal accuracy of 36.19% at layer size 4 and 1024 units per hidden layer, however, the range is rather small with a minimum accuracy of 35.29% achieved by a DNN setup using one layer and 1024 units. In general, we can confirm the results as presented in [28], especially the finding that a higher number of layer is more important than a higher number of hidden units for increasing the estimation performance. For comparison, we evaluated the GMM-HMM on the same data set

#HL	#Units	#Param ($\times 10^6$)	ACC[%]	CD_{1-8} [dB]		CD_0 [dB]	
				DNN direct	DNN-HMM	DNN direct	DNN-HMM
1	256	0.02	35.51	5.38	5.44	7.29	7.33
2		0.08	35.78	5.38	5.40	7.00	6.92
3		0.15	36.01	5.38	5.38	6.91	**6.83**
4		0.22	36.04	5.42	5.46	7.02	6.95
5		0.28	36.10	5.39	5.41	7.23	6.97
6		0.35	36.05	5.39	5.43	7.08	6.98
1	512	0.04	35.41	5.34	5.34	7.13	7.16
2		0.30	35.95	5.41	5.45	7.23	7.23
3		0.56	36.08	5.38	5.40	6.97	6.92
4		0.82	36.15	5.44	5.50	7.13	7.09
5		1.09	36.13	5.40	5.44	7.12	7.04
6		1.35	36.06	5.39	5.42	7.05	6.99
1	1024	0.07	35.29	5.37	5.37	7.23	7.25
2		1.12	35.88	5.45	5.50	7.32	7.31
3		2.17	36.10	5.41	5.44	7.33	7.25
4		3.22	**36.19**	5.41	5.45	7.09	7.09
5		4.27	36.17	5.37	5.41	7.06	7.04
6		5.32	36.12	5.41	5.46	7.18	7.15
GMM-HMM			28.75	**5.31**		9.12	
Oracle model			100	4.44		1.95	

Table 1. Results for the DNN-driven acoustic model, the GMM-HMM acoustic model, and finally an oracle experiment, using always the correct envelope from the codebook. DNN experiments cover the number of hidden layers (HLs), ranging from 1 to 6 (without softmax output layer) in combination with different hidden layer sizes ranging from 256 to 1024 units. The best practical schemes for each of the three quality metrics ACC, CD_{1-8}, and CD_0 are printed in **bold**.

and achieved an accuracy of only 28.75%, attesting the DNN-driven acoustic model a clearly superior estimation quality.

Not shown in Tab. 1, the baseline approach, i.e., the GMM-based acoustic model, results in a mean opinion score (MOS) of 2.73 using WB-PESQ. Results for the DNN-based acoustic models are very close to each other: The DNN-HMM approach led to an MOS of approx. 3.01 (ranging from 2.99 to 3.02) and for the DNN-only approach to an MOS of approx. 3.07 (ranging from 3.05 to 3.08), over all experiments. The improvement of about 0.35 MOS points shows superior speech quality when using a DNN-based approach. Informal subjective listening tests also put DNN-driven acoustic models far ahead of the GMM-based equivalent. The oracle experiment, where always the correct quantized spectral envelope vector is picked from the codebook, achieved a MOS of 3.26.

The following analysis aims at tracing back the speech quality gain to either the spectral envelope estimation, the estimation of log-energy ratio, or both. In the conducted DNN experiments of Tab. 1, the employment of prior knowledge in form of state and transition probabilities, i.e., the DNN-HMM variant, could not improve envelope estimation in terms of CD_{1-8} compared to the DNN-only variant. Surprisingly, the envelope cepstral distance CD_{1-8} of 5.31 dB generated by the GMM-HMM approach is still slightly ahead of any of the DNN experiments. Ranging from 5.34 to 5.45 dB, the CD_{1-8} metric shows almost similar results for DNN-driven acoustic modeling. The oracle experiment resulted in a CD_{1-8} of 4.44 dB thus showing only small potential for further improvement. Consequently, UB envelope estimation cannot be the cause for the observed gain in WB-PESQ of DNN- vs. GMM-based approaches.

The log-energy ratios $\hat{y}_i(0)$, contained in the codebook, control the synthesized energy in the artificial UB during ABE processing. In former studies, a direct relation between overestimation of UB energy and a rise in annoying artifacts could be shown [32]. While the GMM-HMM approach leads to a CD_0 of 9.12 dB, DNN-driven approaches diminish the cepstral distance CD_0 to as low as 6.83 dB for the DNN-HMM approach with 3 HLs each with 256 units. Obviously, the formerly mentioned gain in speech quality can therefore be explained by the improved energy estimation. In other words, these results also emphasize the importance of UB energy estimation, however, compared to the oracle experiment exhibiting a CD_0 of 1.95 dB there is still a huge potential for further improvement. In 13 of the 18 experiments the DNN-HMM approach presents better UB energy estimation capabilities than the DNN-only variant.

The formerly mentioned DNN-HMM with 3 HLs each with 256 units is investigated in more depth. Since the NTT database does not provide time-aligned phonetic transcriptions, we once again take the TIMIT test data set for a phoneme-specific analysis in terms of relative accuracy compared to the GMM-HMM. A phoneme-specific accuracy quantifies how well a statistical model predicts the correct envelope, given a certain phoneme. The here used *relative* phoneme-specific accuracy quantifies the predictive power of the DNN over the GMM. On average, *all* phonemes take profit from the DNN-based statistical model for UB spectral envelope estimation. The smallest improvement with 8% is given for the /s/ sound. The phonemes that benefit most are /zh/, /t/, /dh/, /th/, and /f/ with 52%, 54% 56%, 59%, and 83% relative improvement, respectively. Except for the /t/ sound, all of these sounds belong to the group of fricatives.

In the context of this publication, we could not observe an obvious dependency between accuracy and the number of training parameters, e.g., the DNN with 6 HLs each with 1024 units (no. of parameters: 5.32 M) showed an accuracy of 36.12%, while the DNN with 5 HLs each with 256 units (no. of parameters: 0.28 M) was able to correctly estimate 36.10% of the class indices.

The alert reader might ask, why the DNN experiments overall led to such similar results. Three possible reasons come to mind: (1) The information contained in the features is almost completely exploited by a DNN with rather simple topology, (2) the codebook, defining the UB spectral envelopes and subsequently also the training and target data is sub-optimally chosen, and/or (3) the estimation capabilities of the investigated topologies of the DNN are at their maximum. Either way, investigating and solving these potential problems will be in the focus of future research.

5. CONCLUSIONS

In this paper, Gaussian mixture models (GMMs) were compared to deep neural networks (DNNs), serving as statistical model for upper band (UB) spectral envelope estimation in a source-filter model-based artificial bandwidth extension algorithm. The experiments were conducted using different speech databases for training and evaluation of the statistical models and thus allowing for more realistic conclusions in this study.

The obtained results clearly state that DNNs significantly outperform GMMs. Particularly fricative sounds benefit from the increased predictive power of the here employed deep architectures. Overall, the speech quality enhancement could be traced back to a superior UB *energy* estimation, while the fidelity of the UB spectral envelope *shapes* turned out to play a marginal role only.

6. REFERENCES

[1] Global mobile Suppliers Association, "Mobile HD Voice: Global Update report," Information Papers, Dec. 2015.

[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] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proceedings of ICASSP*, Washington, DC, USA, Apr. 1979, vol. IV.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] G. E. Hinton, Li 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.

[13] 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.

[14] 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.

[15] M. Zöhrer, R. Peharz, and F. Pernkopf, "On Representation Learning for Artificial Bandwidth Extension," in *Proc. of Interspeech*, Dresden, Germany, Sept. 2015, pp. 791–795.

[16] 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.

[17] 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.

[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. E. Markel and A. H. Gray, *Linear Prediction of Speech*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1982.

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

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

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

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

[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] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic Modeling Using Deep Belief Networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[28] G. E. Hinton, S. Osindero, and Y. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, July 2006.

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

[31] "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.

[32] M. Nilsson and W. B. Kleijn, "Avoiding Over-Estimation in Bandwidth Extension of Telephony Speech," in *Proc. of ICASSP*, Salt Lake City, UT, USA, May 2001, vol. 2, pp. 869–872.