EXPLOITING EXPLICIT MEMORY INCLUSION FOR ARTIFICIAL BANDWIDTH EXTENSION

Pramod Bachhav, Massimiliano Todisco and Nicholas Evans

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

ABSTRACT

Artificial bandwidth extension (ABE) algorithms have been developed to improve speech quality when wideband devices are used in conjunction with narrowband devices or infrastructure. While past work points to the benefit of using contextual information or memory for ABE, an understanding of the relative benefit of explicit memory inclusion, rather than just dynamic information, calls for a comparative, quantitative analysis. The need for practical ABE solutions calls further for the inclusion of memory without significant increases to latency or computational complexity. The paper reports the use of an information theoretic approach to show the potential of benefit of memory inclusion. Findings are validated through objective and subjective assessments of an ABE system which uses memory with only negligible increases to latency and computational complexity. Listening tests show that narrowband signals whose bandwidth is artificially extended with, rather than without the inclusion of memory, are of consistently improved quality.

Index Terms— Artificial bandwidth extension, speech quality, Gaussian mixture model

1. INTRODUCTION

Traditional telephony infrastructure is typically limited to a bandwidth of 0.3-3.4kHz. Such legacy infrastructure supports what is referred to as narrowband (NB) communications. With a bandwidth extending from 50Hz-7kHz, modern devices, systems and infrastructure supporting wideband (WB) communications offer enhanced speech quality.

Since legacy infrastructure will take considerable time to replace or upgrade, artificial bandwidth extension (ABE) algorithms have been developed to improve speech quality when WB devices are used with NB devices or infrastructure. ABE is used to estimate missing highband (HB) components at 3.4-8kHz from available NB components.

ABE solutions exploit the correlation between NB and HB components of speech. Classical solutions estimate missing HB components using a regression model learned from WB training data. In ABE methods based on classical source-filter model, the HB component is usually parameterised with some form of linear prediction (LP) coefficients whereas the NB component can be parameterised by a variety of static and/or dynamic spectral estimates.

In addition to being captured through front-end features, dynamic information, or *memory* can also be captured with specific back-end regression models [1]. Hidden Markov model (HMM) [2, 3, 4, 5, 6], temporally-extended Gaussian mixture model (GMM) [7] and deep neural network (DNN) [8, 9, 10] solutions to ABE are all capable of capturing memory. Some DNN solutions, e.g. [11, 12, 13], capture memory in the front-end instead.

Drawing on the work to optimise front-end feature extraction reported in [14], the first attempt to quantify the importance of front-end memory inclusion is reported in [15, 16, 17]. The work demonstrates the benefit of using memory in the form of delta features with a standard regression model. While this body of work points towards the importance of memory to ABE, it raises the questions of what degree of contextual or explicit memory information is of benefit to ABE and how can it be harnessed without increasing latency and computational complexity. The lack of answers to these questions calls for a quantitative analysis which compares the benefit of utilising memory in an otherwise fixed ABE algorithm. This is the goal of the research reported in this paper.

It is organised as follows. Section 2 describes the approach to assess the benefit of memory. Section 3 describes an ABE algorithm and modification to accommodate the inclusion of memory. Objective and subjective ABE assessments are reported in Section 4. Conclusions are presented in Section 5.

2. ASSESSING THE BENEFIT OF MEMORY TO ABE

Assessment is achieved by evaluating the correlation between the HB component of a speech frame and the NB component of neighbouring frames. The standard information theoretic approach to measure the correlation in terms of mutual information (MI) is described before the approach to analysis and the results.

2.1. Mutual information

The mutual information between two continuous random variables X and Y with joint probability density function (PDF) $f_{XY}(x, y)$ is defined according to:

$$I(X;Y) = \int\int f_{XY}(x,y) \log_2 \left(\frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} \right) dxdy \quad (1)$$

If $f_{XY}(x, y)$ takes the form of a Gaussian mixture model (GMM), then Eq. 1 can be written as an expectation approximated by the sample mean over K samples as follows:

$$I(X;Y) \approx \frac{1}{K} \sum_{k=1}^{K} \log_2 \left(\frac{f_{XY}(x_k, y_k)}{f_{XY}(x_k) f_{XY}(y_k)} \right) \quad (2)$$

As reported in [14, 18], Eq. 2 can be used to estimate the MI between NB and HB components of speech frames parameterised with features X and Y respectively.

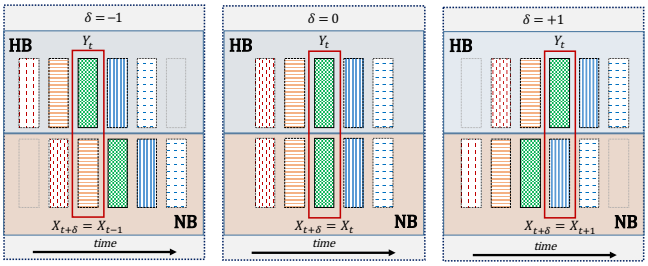

Fig. 1. *An Illustration of MI estimation with contextual information from neighbouring frames. Vertical bars represent NB (bottom) and HB (top) feature vectors. Red boxes represent the pair of NB ($X = X_{t+\delta}, \delta = -1, 0, 1$) and HB ($Y = Y_t$) components used for MI calculations.*

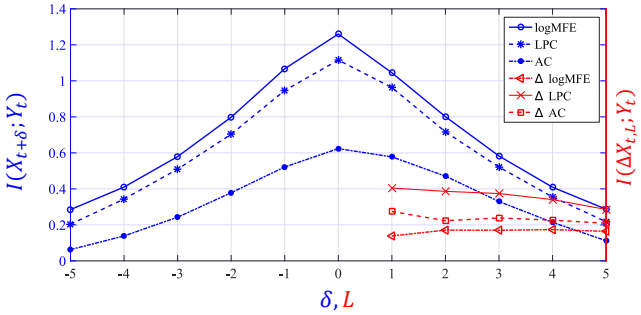

Fig. 2. *An illustration of the variation in MI between static HB features Y_t and static NB features (blue profiles) extracted from neighbouring frames $X_{t+\delta}$, and delta features $\Delta X_{t,L}$ (red profiles).*

2.2. Analysis

The analysis of MI requires a choice of front-end features. Due to the ease in time domain reconstruction, LP coefficients [19] are chosen for HB features. NB features used here include log-Mel filter energy (logMFE) coefficients, LP coefficients and autocorrelation coefficients (ACs) [20].

WB speech signals from the entire TIMIT dataset [21] (excluding dialect (SA) sentences) are low and high-pass filtered and then processed with some form of feature extraction to give NB and HB features X and Y respectively. All signals are processed using 20ms frames with 10ms overlap. NB features X are extracted from the NB power spectrum P_{NB}. logMFE features were calculated by applying a Mel filterbank (MFB) with 10 filters. LP coefficient features of 10 dimensions including the gain parameter are obtained through selective linear prediction (SLP) [22, 20]. Conventional AC features consist of the first 10 normalised coefficients. HB features Y are similarly extracted from the HB power spectrum P_{WB} using SLP, also giving 10 LP coefficients including the gain.

Since phonetic events span in the order of 50ms [23], Eq. 2 is then used with a GMM of 128 components to estimate the MI between instantaneous HB features and NB features spanning a similar duration. This procedure is illustrated in Fig. 1 where Y_t is the instantaneous HB component at time t and where $X_{t+\delta}$ is the NB component at time $t + \delta$ where $\delta \in \mathbb{Z}$.

2.3. Findings

Blue profiles in Fig. 2 show the MI (vertical axis) between instantaneous HB features Y_t and NB features $X_{t+\delta}$ for $\delta \in [-5, +5]$ (horizontal axis). The three profiles correspond to logMFE, LPC and AC features. As expected, for all three profiles, the MI is greatest for $\delta = 0$ for which NB and HB features are extracted from the same frame. For $\delta \neq 0$, the MI is symmetrically lower. The highest MI is obtained with logMFE coefficients. For $\delta = 1, 2$ the MI falls by 17% and 36% relative to that obtained for $\delta = 0$.

Fig. 2 also shows the MI between static HB and dynamic or delta NB features (red profiles). Delta features $\Delta X_{t,L}$ are extracted for a frame at time t in the usual way [15] where $L \in [1, 5]$ (same horizontal axis) is the number of static frames considered either side of t. The MI between static HB and delta NB features is considerably less than for static NB features. This observation corroborates the findings reported in [15], namely that NB delta features are of little use to ABE; they provide comparatively little information about static HB features.

This same finding shows that ABE algorithms should use explicit *memory*, i.e. static features extracted from neighbouring frames, instead of dynamic information captured in delta features. The research hypothesis under investigation in the remainder of this paper is that the inclusion of memory in such a way should help to model phonetic events or sequences which span intervals greater than a single frame and should thus give bandwidth extended speech of enhanced quality. Crucial to this work, however, is that the inclusion of such additional information should not impact on latency or computational complexity.

Since the highest level of MI is obtained with logMFE features, they are used as NB representations for all subsequent experiments reported in this paper. Note that, since the aim is to demonstrate the contribution to ABE of memory, the use of energy based coefficients such as those used in [24] is avoided; it is assumed that their use will further enhance the performance of *any* ABE system.

3. ARTIFICIAL BANDWIDTH EXTENSION

The ABE algorithm used for all further work is illustrated in Fig. 3. **Training** uses parallel WB and NB data for feature extraction and GMM modeling. **Estimation** of missing HB LP coefficients is performed from NB data parametrised by logMFE features. **Resynthesis** is performed using original NB data and estimated HB LP coefficients. Details of each step corresponding to the three blocks of Fig. 3 are given in the following.

3.1. Training

NB and WB signals are processed frame-by-frame. x_t and y_t denote NB and WB frames at time t respectively. NB features (X_t^{NB} - top line in training block) consist of 10 logMFE coefficients whereas HB features (X_t^{HB} - bottom line in training block) consist of 9 LP coefficients a^{HB} and a gain coefficient g^{HB}. Both feature sets are mean and variance normalised (mvn$_x$ and mvn$_y$) giving $X_{t,mvn}^{NB}$ and $Y_{t,mvn}^{HB}$.

Memory inclusion: NB features at time t are concatenated with neighbouring features extracted from δ frames either side of t thus giving:

$$X_{t,conc\_\delta} = \left[X_{t-\delta,mvn}^{NB}, ..., X_{t,mvn}^{NB}, ..., X_{t+\delta,mvn}^{NB}\right]^T$$

In order that the complexity of subsequent steps is unaffected, principal component analysis (PCA) is applied to reduce $X_{t,conc\_\delta}$ to 10-dimensional features X_{t,pca_δ}^{NB}. The PCA matrix W_{PCA} is learned from training data and retained for use in the estimation step.

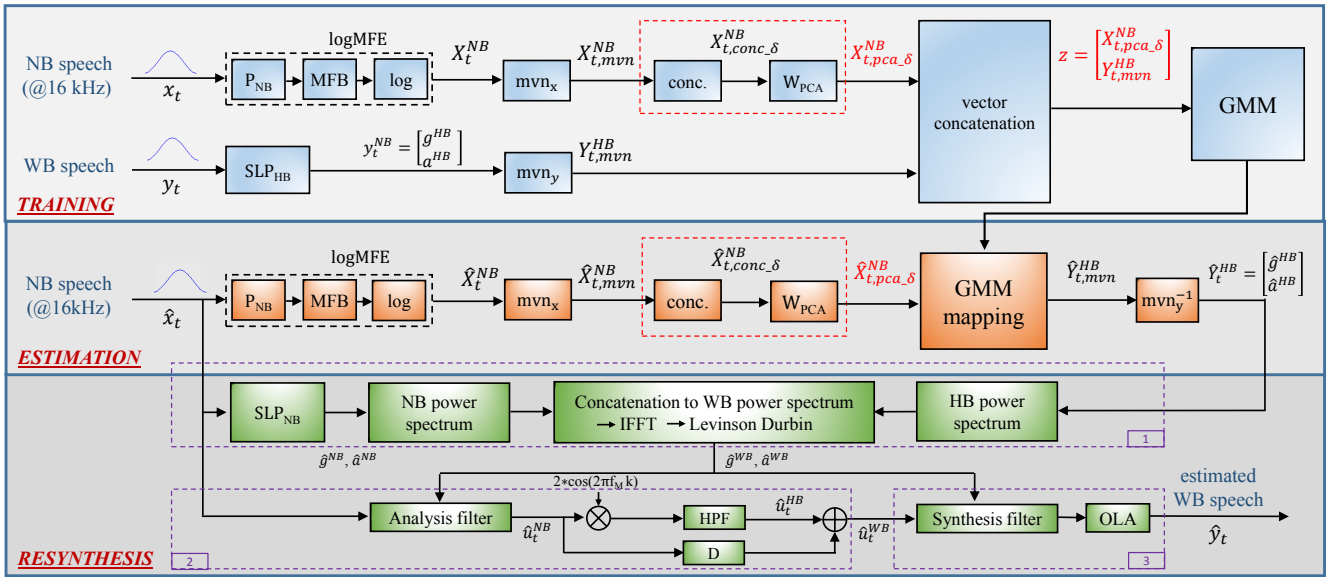

Fig. 3. *A block diagram of the ABE system with memory inclusion.*

Finally, a 128-component, full-covariance GMM is learned from the training data using joint vectors $Z = [X^{NB}_{t,pca\_\delta}, Y^{HB}_{t,mvn}]^T$.

3.2. Estimation

NB signal frames are upsampled to 16kHz signals $\hat{x}_t$ before feature extraction is applied to give $\hat{X}^{NB}_t$. Memory is included according to the same procedure used during training, thereby giving $\hat{X}^{NB}_{t,pca\_\delta}$. The regression model defined by GMM parameters learned during training is then applied in the usual way [19] to estimate HB features $\hat{Y}^{HB}_{t,mvn}$. Using means and variances obtained from the training data, inverse mean and variance normalisation (mvn_y^{-1}) is then applied to estimate HB LP coefficients $\hat{a}^{HB}$ and gain $\hat{g}^{HB}$.

3.3. Resynthesis

Resynthesis involves the three distinct steps illustrated by the numbered blocks in Fig. 3. First, SLP is applied to $\hat{x}_t$ to obtain NB LP coefficients $\hat{g}^{NB}$ and gain $\hat{a}^{NB}$. The NB power spectrum is then determined and concatenated with that of the estimated HB power spectrum obtained from $\hat{g}^{HB}$ and $\hat{a}^{HB}$. Estimated WB parameters $\hat{g}^{WB}$ and $\hat{a}^{WB}$ are then obtained from the inverse fast Fourier transform (IFFT) and application of the Levinson-Durbin recursion to the WB power spectrum. Second, NB speech frames $\hat{x}_t$ are filtered using a LP analysis filter defined by $\hat{g}^{NB}$ and $\hat{a}^{NB}$ in order to obtain NB excitation $\hat{u}_{NB}$. Via spectral translation [24] with a modulation frequency of 6.8kHz, missing excitation components from 3.4-8kHz are estimated followed by a high pass filter (HPF) thereby giving HB excitation components $\hat{u}^{HB}_t$. $\hat{u}_{HB}$ is then added to appropriately delayed (D) $\hat{u}^{NB}_t$ to give extended WB excitation $\hat{u}^{WB}_t$. In the third and final step, $\hat{u}_{WB}$ is filtered using a synthesis filter defined by $\hat{g}^{WB}$ and $\hat{a}^{WB}$ in order to resynthesise extended WB speech $\hat{y}_t$ with an overlap and add (OLA) method.

4. EXPERIMENTAL SETUP AND RESULTS

This section describes the experimental setup, baselines and results for objective, subjective and mutual information assessments.

4.1. Database

ABE experiments were performed using the TIMIT database [21]. The TIMIT training set (consisting of 3696 utterances spoken by 462 speakers) was used for GMM training with parallel WB and NB speech signals processed according to the steps described in [25] (SA dialect sentences were again removed). Assessment was performed using the TIMIT test set (1344 utterances spoken by 168 speakers). NB TIMIT data at 16kHz was obtained before being extended to WB signals using ABE.

4.2. Assessment

The performance of the ABE algorithm with memory, denoted M_δ where δ indicates the number of neighbouring frames which form the memory, is compared to a baseline algorithm, denoted B_1. System M_δ and B_1 use $\hat{X}^{NB}_{pca\_\delta}$ and $\hat{X}^{NB}_t$ features respectively.

For comparison to past work in [17], assessment includes a second baseline, denoted B_2, which exploits memory in the form of delta features. System B_2 uses 5-dimensional static features appended with second order 5-dimensional delta features for both NB and HB parametrisations (logMFE and LPCs in the context of our implementation). For resynthesis, delta features from the estimated HB feature were eliminated with only the first 5 coefficients being used.

All ABE algorithms were implemented with Hann windows of 20ms duration and 10ms overlap, thereby supporting perfect OLA reconstruction [26, 27]. A 1024-point FFT was used for all frequency domain operations.

4.3. Objective assessment

Objective assessment was performed using well known spectral distortion measures: the root mean square log-spectral distortion (RMS-LSD) and the so-called COSH measure (symmetric version of the Ikatura-Saito distortion) [28]. Defining the power spectra for original $H(f)$ and estimated $\hat{H}(f)$ spectral envelopes as $P(f) = g^2/|H(f)|^2$ and $\hat{P}(f) = \hat{g}^2/|\hat{H}(f)|^2$ where g and $\hat{g}$

Table 1. *Objective assessment results. RMS-LSD and d_{COSH} are distance measures (lower values indicate better performance) in dB, mean and (standard deviation) whereas MOS-LQO$_{WB}$ values reflect quality (higher values indicate better performance).*

ABE method	$d_{\text{RMS-LSD}}$	d_{COSH}	MOS-LQO$_{\text{WB}}$
B_1	9.2 (1.2)	2.4 (0.7)	2.4
B_2	10.1 (1.2)	3.6 (1.2)	2.2
M_1	8.2 (0.9)	2.2 (0.6)	2.8
M_2	**8.1** (0.9)	**2.1** (0.6)	**2.9**
M_3	8.2 (0.9)	2.2 (0.7)	2.8

are the respective LP gains, then RMS-LSD and COSH distance measures are defined as follows:

$$d_{\text{RMS-LSD}} = \sqrt{\frac{1}{\triangle F}\int_{\triangle F}\left[10\log_{10}\left(\frac{P(f)}{\hat{P}(f)}\right)\right]^2 df}$$

$$d_{\text{COSH}} = \frac{1}{2}\left[d_{\text{IS}}(P(f),\hat{P}(f)) + d_{\text{IS}}(\hat{P}(f),P(f))\right]$$

where the Ikatura-Saito distortion (d_{IS}) is defined by

$$d_{\text{IS}}(P(f),\hat{P}(f)) = \frac{1}{\triangle F}\int_{\triangle F}\left[\frac{P(f)}{\hat{P(f)}} - \ln\frac{P(f)}{\hat{P}(f)} - 1\right]^2 df$$

where $\triangle F = [3400, 8000]$Hz. Finally, a WB extension to the perceptual analysis of speech quality algorithm [29] is used to give objective estimates of mean opinion scores (MOS-LQO$_{\text{WB}}$).

Objective assessment results are illustrated in Tab. 1. While all ABE systems with memory outperform both baselines B_1 and B_2, system M_2, which uses memory contained within two neighbouring frames, performs best. Surprisingly, baseline system B_2 gives worse performance than B_1. This is caused by the inclusion of memory through delta features while under the constraint of fixed dimensionality. The latter necessitates the loss of informative higher-order static features in order to accommodate dynamic delta features. On account of these findings, subjective assessments were performed with systems B_1 and M_2.

4.4. Subjective assessment

Subjective assessments were performed using comparison mean-opinion score (CMOS) tests [30]. Tests were performed by 14 listeners who were asked to compare the quality of 14 pairs of speech signals A and B. They were asked to rate the quality of signal B with respect to A according to the following scale: -3 (much worse), -2 (slightly worse), -1 (worse), 0 (about the same), 1 (slightly better), 2 (better), 3 (much better). The samples were played using DT 770 PRO headphones. All speech files used for subjective tests are available online[1].

CMOS results of 0.69 and 0.51 presented in Tab. 2 show that bandwidth extended speech produced by system M_2 is preferred to original NB speech and that produced by system B_1. Further informal listening tests showed that the inclusion of memory helps to reduce the presence of processing artifacts in extended speech, thereby resulting in enhanced quality.

[1] http://audio.eurecom.fr/content/media

Table 2. *Subjective assessment results in terms of CMOS.*

Comparison B $\rightarrow$ A	CMOS
$M_2 \rightarrow$ NB	0.69
$M_2 \rightarrow B_1$	0.51
$M_2 \rightarrow$ WB	-0.78

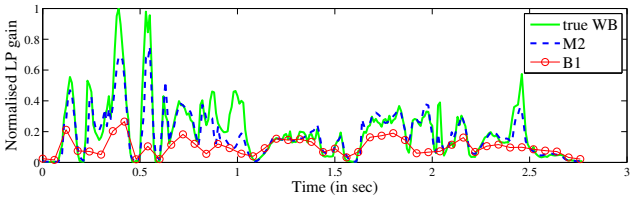

Fig. 4. *A comparison of true WB LP gain $\hat{g}^{WB}_{true}$ to estimated WB LP gain $\hat{g}^{WB}$ for systems M_2 and B_1.*

Table 3. *Mutual information assessment results*

Comparison	logMFE
$I(X_t;Y_t)$ (System B_1)	1.24
$I(\hat{X}^{NB}_{pca\_2};Y_t)$ (System M_1)	1.34

An illustration of the improvement over the baseline B_1 in gain estimation as a result of memory inclusion is shown in Fig. 4. Improvements in gain estimation reduce processing artifacts, improvements which are confirmed by reductions in RMS-LSD.

4.5. Mutual information assessment

A final set of experiments aims to further validate the findings of both objective and subjective assessments by showing the improvement in mutual information (MI) brought by the inclusion of memory. Tab. 3 compares the MI between features X_t and Y_t with that between X^{NB}_{t,pca_2} and Y_t. Results show that the inclusion of memory results in notably higher MI; memory helps to better model missing HB information.

5. CONCLUSIONS

This paper reports an approach to artificial bandwidth extension that incorporates the use of memory in the estimation of missing high-band speech components. The paper extends prior work that studied the benefit of capturing memory via dynamic delta features or of incorporating memory into hidden Markov model and deep neural network solutions. New to this contribution is a study of the explicit inclusion of memory captured through static features extracted from neighboring speech frames. The potential of this approach is first demonstrated through information theoretic analyses and then validated though both objective and subjective assessments. The inclusion of memory captured from two neighbouring frames leads to artificially extended speech of enhanced quality. Key to this contribution and a further differentiator to prior work is the use of feature dimensionality reduction which ensures only negligible impacts on latency and complexity. Future work should investigate dimensionality reduction techniques designed to preserve quality rather than feature variance. Finally, it is hoped that the study reported in this paper may help to shed light on the use of memory in deep learning solutions to artificial bandwidth extension.

6. REFERENCES

[1] A. Nour-Eldin, "Quantifying and exploiting speech memory for the improvement of narrowband speech bandwidth extension," Ph.D. Thesis, McGill University, Canada, 2013.

[2] G. Chen and V. Parsa, "HMM-based frequency bandwidth extension for speech enhancement using line spectral frequencies," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, 2004, pp. I–709.

[3] C. Yağli and E. Erzin, "Artificial bandwidth extension of spectral envelope with temporal clustering," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2011, pp. 5096–5099.

[4] C. YağLı, M. T. Turan, and E. Erzin, "Artificial bandwidth extension of spectral envelope along a viterbi path," *Speech Communication*, vol. 55, no. 1, pp. 111–118, 2013.

[5] I. Katsir, D. Malah, and I. Cohen, "Evaluation of a speech bandwidth extension algorithm based on vocal tract shape estimation," in *Proc. of Int. Workshop on Acoustic Signal Enhancement (IWAENC)*. VDE, 2012, pp. 1–4.

[6] P. Bauer, J. Abel, and T. Fingscheidt, "HMM-based artificial bandwidth extension supported by neural networks," in *Proc. of Int. Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2014, pp. 1–5.

[7] A. Nour-Eldin and P. Kabal, "Memory-based approximation of the gaussian mixture model framework for bandwidth extension of narrowband speech," in *Annual Conf. of the Int. Speech Communication Association*, 2011.

[8] Y. Wang, S. Zhao, D. Qu, and J. Kuang, "Using conditional restricted boltzmann machines for spectral envelope modeling in speech bandwidth extension," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2016, pp. 5930–5934.

[9] Y. Gu, Z.-H. Ling, and L.-R. Dai, "Speech bandwidth extension using bottleneck features and deep recurrent neural networks." in *Proc. of INTERSPEECH*, 2016, pp. 297–301.

[10] Y. Wang, S. Zhao, J. Li, J. Kuang, and Q. Zhu, "Recurrent neural network for spectral mapping in speech bandwidth extension," in *Proc. of IEEE Global Conf. on Signal and Information Processing (GlobalSIP)*, 2016, pp. 242–246.

[11] B. Liu, J. Tao, Z. Wen, Y. Li, and D. Bukhari, "A novel method of artificial bandwidth extension using deep architecture," in *Sixteenth Annual Conf. of the Int. Speech Communication Association*, 2015.

[12] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2015, pp. 4395–4399.

[13] J. Abel, M. Strake, and T. Fingscheidt, "Artificial bandwidth extension using deep neural networks for spectral envelope estimation," in *Proc. of Int. Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2016, pp. 1–5.

[14] P. Jax and P. Vary, "Feature selection for improved bandwidth extension of speech signals," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2004, pp. I–697.

[15] A. Nour-Eldin, T. Shabestary, and P. Kabal, "The effect of memory inclusion on mutual information between speech frequency bands," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, 2006, pp. III–III.

[16] A. Nour-Eldin and P. Kabal, "Objective analysis of the effect of memory inclusion on bandwidth extension of narrowband speech," in *Proc. of INTERSPEECH*, vol. 1, 2007, pp. 2489–2492.

[17] ——, "Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech,," in *Proc. of INTERSPEECH*, 2008, pp. 53–56.

[18] M. Nilsson, H. Gustaftson, S. Andersen, and W. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2002, pp. I–525.

[19] K.-Y. Park and H. Kim, "Narrowband to wideband conversion of speech using gmm based transformation," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 3, 2000, pp. 1843–1846.

[20] P. Jax, "Enhancement of bandlimited speech signals: Algorithms and theoretical bounds," Ph.D. Thesis, Aachen University (RWTH), Germany, 2002.

[21] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report N*, vol. 93, 1993.

[22] J. Markel and A. Gray, *Linear prediction of speech*. Springer Science & Business Media, 2013, vol. 12.

[23] D. O'shaughnessy, *Speech communication: Human and Machine*. Universities press, USA, 1987.

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

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

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

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

[28] R. Gray, A. Buzo, A. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 367–376, 1980.

[29] "ITU-T Recommendation P.862.2 : Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," *ITU*, 2005.

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