INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Artificial Bandwidth Extension using H^∞ Optimization

Deepika Gupta, Hanumant Singh Shekhawat

Department of Electronics and Electrical Engg
Indian Institute of Technology Guwahati
Guwahati, Assam, India

deepika.gupta@iitg.ac.in, h.s.shekhawat@iitg.ac.in

Abstract

This work proposes a new method for artificial bandwidth extension (ABE) that aims to extend the bandwidth of speech signals in narrowband voice communications. We extract a signal model which consists of the wideband information. Using the signal model, we obtain an infinite impulse response (IIR) interpolation filter with the help of H^∞ optimization. Interpolation filters are going to be distinct for the speech signals because of their non-stationary (time-variant) nature. In narrowband communications, only narrowband signal is accessible. Hence, a codebook approach is intended to keep the IIR interpolation filters information (wideband feature) together with their corresponding narrowband signal characteristic (narrowband attribute). For that, the Gaussian mixture modeling (GMM) codebook approach is utilized to estimate the wideband feature for a given narrowband attribute of the signal. Performances are assessed for the two sorts of narrowband attributes.

Index Terms: H^∞-optimization, speech production model, signal model, lifting

1. Introduction

The high quality of speech signals is required in voice communication that is highly dependent on the frequency components present in the human speech signal. Practically, the standard sampling rate of the telephone speech used in the Global System for Mobile communications (GSM), is 8 kHz [1]. As per the Nyquist criteria, a signal of the maximum frequency of 4 kHz can be transmitted through the channel. So, frequencies present in the speech signals above 4 kHz are not transmitted. Because of the absence of the high-frequency components above 4 kHz, the naturalness, clarity, and pleasantness in the listening of the speech signal go down. To beat these issues, the narrowband (NB) (0-4 kHz) speech signal sampled at 8 kHz is processed to recuperate the high-frequency components up to 8 kHz. It can be made possible by using the additional information, obtained from the wideband (WB) (0-8 kHz) signal (sampled at 16 kHz frequency). This process is known as artificial bandwidth extension (ABE), as shown in Fig. 1. Here, LPF is a low pass filter. The NB speech signal S_{NB} is generated by low pass filtering of the WB signal followed the downsampling by a factor of 2 at the transmitter side. The NB signal S_{NB} is processed by a bandwidth extension process for recovering the corresponding high-band (HB) ($4 - 8$ kHz) signal S_{HB}. The bandwidth extension process requires some HB information.

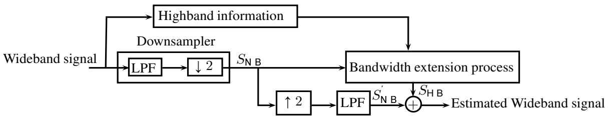


Figure 1: *A fundamental outline for generating the stationary narrowband signal and after that its bandwidth extension*

Many approaches are proposed for ABE based on the speech production model (SPM) in which the speech signal is segregated in two parts: speech production filter (SPF) as a vocal tract filter and excitation signal as a residue signal [2]. In ABE methods based on the SPM, the speech production filter is represented in many different parameters such as linear prediction coefficients (LPC) [3], line spectral frequencies (LSF) [4], linear frequency cepstral coefficients (Cepstrum) [5] and Mel frequency cepstral coefficients (MFCC) [6]. Further, the high-band excitation can be assessed by utilizing the numerous other ways, i.e., the bandpass-envelope modulated Gaussian noise (BP-MGN) [7], harmonic noise model (HNM) [8], spectrum folding [9, 10], pitch adaptive modulation [11], full-wave rectification [12] and spectral translation [9, 11, 12]. In [13, 14], the spectrum of the WB signal is directly used to represent HB as well as NB information for ABE.
According to speech production theory, speech production filter can be accurately represented by the pole-zero model [15]. Many of the existing methods use an all-pole model [5, 4]. It may not be sufficient for some utterances like fricatives, nasals, laterals and the burst interval of stop consonants because of the presence of zeros in SPF's frequency response [15]. In this paper, we used the pole-zero models [15]. Existing methods estimate the high-band (HB) signal only. Hence, they need a low pass filter (LPF) which is non-ideal, but close to the ideal filter. However, this non-ideal nature helps in identifying the high-frequency components specifically for unvoiced speech. Hence, in our approach, we introduce a little more non-ideality in LPF, i.e., NB signal is obtained by direct downsampling of the wideband signal at the transmitter side. It introduces aliasing in the obtained NB signal. Hence, our approach requires the estimation of NB as well as HB of the speech signal. Benefits of our proposed method are that it does not require the energy adjustment between NB and HB and, reduces delay in communication.

This work utilizes the H^∞ optimization in order to get an interpolation filter for the speech signals. The speech signals are non-stationary [16]. So, a frame-based approach is favoured for non-stationary signals. This approach builds the need of storage for extra data about interpolation filters (wideband feature) with their corresponding narrowband detail. To generalize, a codebook approach is used in [10, 11, 17, 18, 19, 20, 21, 22, 4, 5]. In this paper, we use the Gaussian mixtures model [23] for designing the codebook. Also, we compare our proposed method with the spectrum folding method [22, 9, 10].

Rest of the paper is organized as follows: Section 2 presents the bandwidth extension process for a stationary signal; Section 3 contains its use for speech signals; Section 4 consists of

http://dx.doi.org/10.21437/Interspeech.2019-1580

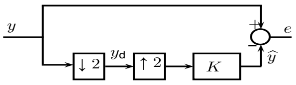


Figure 2: *An error system set-up for recovering of a stationary wideband speech signal.*

the experimental results and analysis using the GMM model.

2. Proposed bandwidth extension process for a stationary speech signal

General set-up of ABE problem is shown in Fig. 2. In Fig. 2, ↓ 2 represents an ideal down-sampler, y represents the original stationary WB signal, and $\widehat{y}$ represents the estimated WB signal from the stationary NB signal y_d using a linear discrete time-invariant (LDTI) interpolation filter K. The K design is based on the reconstruction error minimization using a suitable norm. Every discrete-time signal can be represented by the LDTI system driven by the white noise or an impulse [15]. Therefore, information about the WB original signal y is extracted in the form of a generating model (F), which reflects the signal properties. The modified block diagram is represented in Fig. 3 with y being the output of system F driven by an input w_d. The transfer function of F is represented by $F(z)$. It is further assumed that $F(z)$ is a stable and strictly proper rational transfer function. Such a system can be represented in the frequency domain as $F(z) = C(zI - A)^{-1}B$, where A, B, C are constant matrices of appropriate dimension. The standard Prony's method is used to evaluate the $F(z)$ [24]. A function based on this method is available in MATLAB [25]. Next, our objective is to minimize the error with a suitable norm.

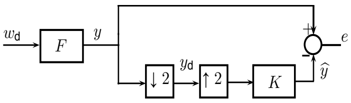


Figure 3: *Proposed architecture of error system for reconstructing a stationary wideband speech signal.*

2.1. Performance Index

We are utilizing the H^∞ system norm to minimize the reconstruction error. Let $\mathbb{T}$ denotes the operator from input w_d to output e in Fig. 3. The H^∞-norm of $\mathbb{T}$ is defined as

$$||\mathbb{T}||_\infty := \sup_{w_d \neq 0} \frac{||e||_2}{||w_d||_2},$$

2.2. Problem Formulation

To design $K(z)$, we have to take care of the following problem.

Problem 1. *Given a stable and strictly proper transfer function $F(z)$, design a stable and causal interpolation filter K_{opt} defined as*

$$K_{opt} := \arg\min_K (||\mathbb{T}||_\infty),$$

where

$$\mathbb{T} := F - K(\uparrow 2)(\downarrow 2)F.$$

Time-variant nature of speech signals yields some uncertainty in the model $F(z)$. It is well known that H^∞-norm optimization gives a robust solution [26]. The solution of the problem H^∞ optimization is given in [27, 28].

2.3. Solution of the Problem 1

The error system in Fig. 3 is a multi-rate system. It can be transformed into a single rate system by using the lifting technique which converts the one-dimensional signal into a multi-dimensional signal and vice versa by the inverse lifting [27, 29]. z-transform representations of lifting and inverse lifting are [28, 30]

$$\mathbf{L_N} = (\downarrow N) \begin{bmatrix} 1 & z & z^2 & ..... & z^{N-1} \end{bmatrix}^T \tag{1a}$$

$$\mathbf{L_N^{-1}} = \begin{bmatrix} 1 & z^{-1} & z^{-2} & ..... & z^{-(N-1)} \end{bmatrix} (\uparrow N). \tag{1b}$$

Proposition 1. *Let transfer function $G(z)$ be represented in state space as*

$$G(z) := \left[\begin{array}{c|c} A_G & B_G \\ \hline C_G & D_G \end{array}\right] = D_G + C_G(zI - A_G)^{-1}B_G,$$

with $A_G \in \mathbb{R}^{N\times N}, B_G \in \mathbb{R}^{N\times p}, C_G \in \mathbb{R}^{m\times N}, D_G \in \mathbb{R}^{m\times p}$ matrices, m and p being the dimensions of output and input of $G(z)$, respectively. Next, the lifted (by a factor of 2) transfer function of $G(z)$ in state space form is represented as

$$\overline{G(z)} := \mathbf{L_2}G(z)\mathbf{L_2^{-1}} = \left[\begin{array}{c|cc} A_G^2 & A_G B_G & B_G \\ \hline C_G & D_G & 0 \\ C_G A_G & C_G B_G & D_G \end{array}\right],$$

where $\mathbf{L_2}$ and $\mathbf{L_2^{-1}}$ can be obtained by using (1a) *and* (1b)*, respectively.*

Proof. See [29, Theorem 8.2.1]. □

We also have the following results

$$K(z)(\uparrow 2) = \mathbf{L_2^{-1}}\tilde{K}(z) \tag{2}$$

$$K(z) = \begin{bmatrix} 1 & z^{-1} \end{bmatrix} \tilde{K}(z^2) \tag{3}$$

where $\tilde{K}(z) := \overline{K(z)} \begin{bmatrix} 1 & 0 \end{bmatrix}^T_{1\times 2}$ and $\overline{K(z)} := \mathbf{L_2}K(z)$ $\mathbf{L_2^{-1}}$. z-domain representation of the error system $\mathbb{T}$ can be written as

$$\mathbb{T}(z) = F(z) - \mathbf{L_2^{-1}}\tilde{K}(z)(\downarrow 2)F(z).$$

Thus, the sampling rates of $F(z)$ and $\tilde{K}(z)$ are not the same. Hence, the lifted input and output of $\mathbb{T}$ by a factor of 2 give the lifted transfer function of $\mathbb{T}$ as follows

$$\begin{aligned} \overline{\mathbb{T}}(z) =& \mathbf{L_2}\mathbb{T}(z)\mathbf{L_2^{-1}} \\ =& \overline{F(z)} - \tilde{K}(z)S\overline{F(z)}, \end{aligned} \tag{4}$$

with $S = \begin{bmatrix} 1 & 0 \end{bmatrix}$, $\overline{F(z)} := \mathbf{L_2}F(z)\mathbf{L_2^{-1}}$, and $\overline{N(z)} := \mathbf{L_2}N(z)\mathbf{L_2^{-1}}$. The system $\mathbb{T}$ is changed over into a single rate system $\overline{\mathbb{T}}$ after applying the lifting. Note that, the norm is not changed after introducing the lifting, i.e., $||\mathbb{T}||_\infty = ||\overline{\mathbb{T}}||_\infty$ [29]. Equation (4) can be written in the form of standard discrete control system as depicted in Fig. 4 [29]. Here, $\tilde{w}_d = \mathbf{L_2}w_d$, and $\tilde{e} = \mathbf{L_2}e$. Now, an optimal causal and stable filter $\tilde{K}(z)$ is acquired by utilizing the robust control toolbox in Matlab [31, 25]. Finally, the filter $K(z)$ is evaluated by (3). The obtained interpolation filter $K(z)$ is an IIR filter which is converted into an FIR filter (finite impulse response) by truncating its impulse response. The number of terms in FIR filter is taken 21 empirically as see later in Section 4.

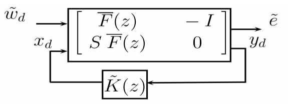


Figure 4: *Standard discrete control unit having an open loop transfer function together with a feedback system $\tilde{K}(z)$.*

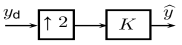


Figure 5: *Artificial bandwidth extension of a stationary narrowband speech signal.*

3. Artificial Bandwidth Extension

In the previous section, we obtained the interpolation filter $K(z)$ to interpolate the up-sampled narrowband signal, as shown in Fig. 5. To apply the results given in Section 2, the speech signals are divided into frames of 25 ms duration. This frame-based approach is needed due to the non-stationary nature of speech signals. As a result of that, the obtained interpolation filters will not be the same for all the frames. Hence, a codebook is required, which stores the interpolation filter information with their corresponding narrowband frame information. The narrowband (NB) information is taken in two different ways, i.e., line spectral frequencies (LSF) [32] and linear prediction coefficients (LPC) [3]. Next, a codebook is trained by GMM [23]. A feature vector $Z \in \mathbb{R}^{31}$ is formed by concatenating the NB feature X of dimension $\mathbb{R}^{10}$ and the corresponding wideband feature Y_K for GMM. $Z \in \mathbb{R}^{31}$ is modeled by GMM to obtain the joint probability distribution function (pdf) of X and Y_K. Parameters of the GMM are estimated using the Expectation-Maximization algorithm [23]. Testing phase requires the estimation of the wideband (WB) feature vector $\tilde{Y}_K$ from the joint pdf for a given narrowband (NB) feature vector $\tilde{X}$. For a given $\tilde{X}$, $\tilde{Y}_K$ is estimated by considering the minimum mean square error (MMSE) criteria [33, 34]. The estimated wideband feature is used in the signal bandwidth extension.

Our proposed method is compared with the spectrum folding technique based upon the source-filter model [22, 9, 10]. Also, the spectrum folding technique is working with an LPF, i.e., it is following the flow drawn in Fig. 1. Here, an LPF is a non-causal FIR LPF filter [35]. The length of this filter is 118. Non-causality of this filter introduces a delay in transmission. We took the same dimensions of features for this technique as in our approach, i.e., 21 and 10 as dimensions of WB feature and NB feature, respectively. LSF represents the NB feature. The WB feature has WB speech frame information in terms of LSF and gain factor [22].

4. Experimental Analysis and Results

The entire flow of training of the GMM model and bandwidth extension of an NB signal is shown in Fig. 6, which is used for the ABE of speech signals. The experiments are performed on the speech signals which are taken from the TIMIT database [36]. It contains two different sets: test and training set. We truncate the actual training set and obtain a new training set. This new training set has the equivalent number of female and male speech files of each dialectical region of the United States. Performances are computed on 400 speech files having equally female and male speech files of each dialectical region from the test set by using the GMM approach. Wideband speech signals are divided into wideband frames of 25 ms length utilizing the hamming window with 50% overlapping between nearby frames. A signal model is computed for each frame, and the corresponding discrete interpolation filter is acquired. Here, The truncated impulse response of the interpolation filter is stored as a wideband feature. NB feature is computed from the NB frame, that is obtained by down-sampling the wideband frame by a factor of 2.

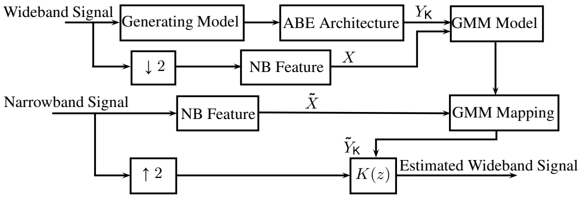


Figure 6: *Block diagram constituting of training the GMM and extension of the narrowband signal.*

4.1. Objective Measures

We use MSE (mean square error) [37], SDR (signal to distortion ratio) [38], LLR (log likelihood ratio) [16], LSD (log spectral distance) [39] and MOS-LQO (mean opinion score- listening quality objective) [39] as the standard objective measures for examining the quality of reconstructed speech signals. Firstly, we compute the performances produced by IIR filter K in Table 1 and, see how much it improves the upsampled narrowband signal in Fig. 3. As seen in Table 1, the interpolation filter K yields notable improvement in all objective measures used in our work.

Table 1: *Comparisons of the performances obtained by an ideal upsampler with an upsampling factor 2 (without applying filter K) and straight forwardly utilizing the interpolation IIR filter K in Fig. 3 for speech files belonging to test set.*

Output subblock	MSE	SDR	LLR	MOS-LQO	LSD
Upsampler	8.1167×10^{-4}	3.01	1.4254	3.5044	11.3135
Interpolation filter K	4.8634×10^{-5}	15.81	0.6547	3.8047	7.6220

Moreover, we truncate the IIR interpolation filter K into the FIR filter and, and see the effect of truncation on performances in Table 2. Here, it is easily said that the performances are improving when we increase the length of the FIR filter, but slowly after the length 21. So, we select 21 as the filter length. Then, GMM performances are computed on the testing set as tabulated in Table 3 for our proposed approach and the spectrum folding method. The LSF NB feature yields better results

Table 2: *Performances figured out on the test set an occurrence of direct utilization of FIR filter K in Fig. 3 for ABE.*

Number of terms	MSE	SDR	LLR	MOS-LQO	LSD
11	8.9405×10^{-5}	13.18	0.7925	3.7450	8.2260
15	7.4762×10^{-5}	13.74	0.7851	3.7521	8.1389
21	6.0912×10^{-5}	14.79	0.7233	3.7782	7.9339
25	5.8136×10^{-5}	15.06	0.7065	3.7810	7.8678
31	5.6043×10^{-5}	15.25	0.6937	3.7854	7.8078
∞	$\mathbf{4.8634\times10^{-5}}$	**15.81**	**0.6547**	**3.8047**	**7.6220**

Table 3: *Performances computation for the GMM model on the test set by number of GMM (\$) =128;*

Features	MSE	SDR	LLR	MOS-LQO	LSD
LSF+K \$	$\mathbf{6.5449\times10^{-5}}$	**14.38**	**0.7751**	3.4967	**8.0205**
LPC+K \$	7.3768×10^{-5}	13.82	0.8050	3.4640	8.1997
Spectrum folding (LSF+LSF+gain)\$	4.4493×10^{-4}	5.60	0.8440	**3.8148**	9.6288

as a comparison of LPC for our proposed method. So, we implement the spectrum folding method with LSF features. Then, the objective measures except the MOS-LQO are improved by our proposed method.

Moreover, we analyzed the magnitude spectrum of two types of speech, for example, unvoiced speech and voiced speech by utilizing the GMM model. For this, we plot the magnitude spectrum of original speech and extended speech obtained by the proposed method and also the spectrum folding method for each case: voiced speech and unvoiced speech. In Fig. 7a for unvoiced speech, our proposed approach get back the better magnitude spectrum rather than the spectrum folding approach. In Fig. 7b, the magnitude spectrum of the estimated voiced speech is better for our proposed approach in frequency range $0-4.5$ kHz and the spectrum folding in frequency range $4.5-8$ kHz. Somehow, the magnitude spectrum in a range from 3.5 kHz to 4.5 kHz acquired by our proposed approach is more close to the original spectrum for both the speech.

4.1.1. *Subjective listening test*

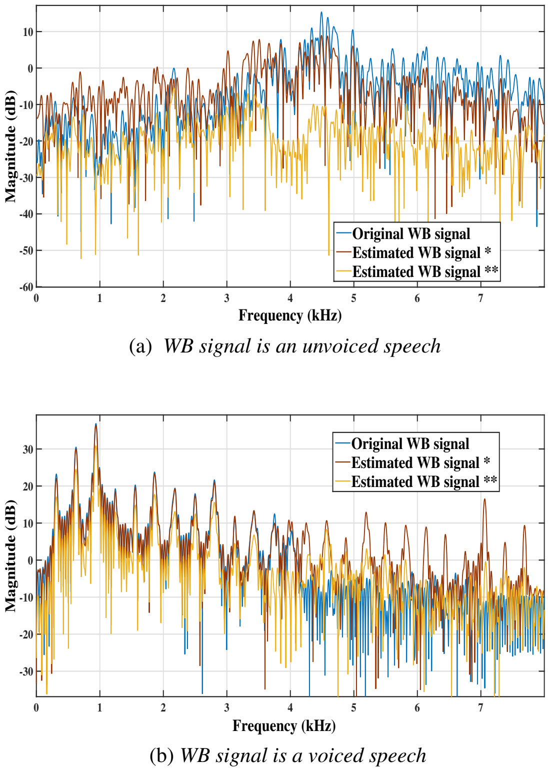
Subjective assessment [40] is done to check the quality of extended speech signals obtained by our proposed method and the spectrum folding method utilizing the GMM as a statistical model and the LSF as an NB feature. For the listening test, arbitrary ten extended speech signals are chosen from the test set, and ten speakers give a mean opinion score (MOS) value between 1 to 5 to these signals with respect to the original WB speech files [40]. Then, the comparison mean opinion score (CMOS) is computed in Table 4 for the proposed method and the spectrum folding method using the same GMM model. Our proposed method improves the CMOS significantly by 0.76 in comparison to the spectrum folding method using the GMM.

Table 4: *Subjective listening test for artificially extended speech files belonging to the test set using the GMM model.*

Conditions	CMOS
Spectrum folding with GMM vs Proposed method with GMM	0.76

5. Conclusion

Our proposed framework for ABE of speech signals is not quite the same as the existing framework. The LPF has been dropped off from the transmitter side as well as the receiver side. Therefore, we are estimating the full wideband signal. An advantage of this is that energy level adjustment between the narrowband and the high-band is not required. Another advantage is that it reduces the delay in transmission because of dropping off LPF. The IIR interpolation filter obtained by H^{∞} optimization is truncated into an FIR filter which is taken as a wideband feature. We carry out experiments with two types of NB features such as LSF and LPC. The GMM conducts the estimation of WB feature for a given NB feature. Extended speech signal quality is analyzed by utilizing the objective measures as SDR, MSE, MOS-LQO, LLR, and LSD and the subjective listening test. Our proposed method improves the objective measures except the MOS-LQO and, the subjective measure CMOS in comparison to the spectrum folding method.



(a) *WB signal is an unvoiced speech*

(b) *WB signal is a voiced speech*

Figure 7: *Magnitude spectrum of the original WB signal (blue), the estimated WB signal * (red) by the* **proposed method** *and the estimated WB signal ** (yellow) by the* **spectrum folding method** *using LSF NB feature and* 128 *GMM.*

6. References

[1] H. Pulakka, L. Laaksonen, M. Vainio, J. Pohjalainen, and P. Alku, “Evaluation of an artificial speech bandwidth extension method in three languages,” *IEEE transactions on Audio, Speech, and Language processing*, vol. 16, no. 6, pp. 1124–1137, 2008.

[2] X. Shao, “Robust Algorithms for Speech Reconstruction on Mobile Devices,” Ph.D. dissertation, University of East Anglia, 2005.

[3] B. Andersen, J. Dyreby, B. Jensen, F. H. Kjærskov, O. L. Mikkelsen, P. D. Nielsen, and H. Zimmermann, “Bandwidth Expansion of Narrow Band speech using Linear Prediction,” *web source*, vol. 26, 2015.

[4] 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.

[5] J. Abel and T. Fingscheidt, “Artificial Speech Bandwidth Extension Using Deep Neural Networks for Wideband Spectral Envelope Estimation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 71–83, 2018.

[6] Y. Sunil and R. Sinha, "Exploration of class specific ABWE for robust children's ASR under mismatched condition," in *Proceedings International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2012, pp. 1–5.

[7] Y. Qian and P. Kabal, "Dual-mode wideband speech recovery from narrowband speech," in *Eighth European Conference on Speech Communication and Technology, GENEVA, Switzerland*, 2003, pp. 1433–1436.

[8] S. Vaseghi, E. Zavarehei, and Q. Yan, "Speech bandwidth extension: Extrapolations of spectral envelop and harmonicity quality of excitation," in *Proceedings International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3. IEEE, 2006, pp. III–844–III–847.

[9] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, Cambridge, United Kingdom*, vol. 4. IEEE, 1979, pp. 428–431.

[10] N. Enbom and W. B. Kleijn, "Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients," in *Proceedings IEEE Workshop on Speech Coding*. IEEE, 1999, pp. 171–173.

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

[12] J. A. Fuemmeler, R. C. Hardie, and W. R. Gardner, "Techniques for the regeneration of wideband speech from narrowband speech," *EURASIP Journal on Applied Signal Processing*, vol. 2001, no. 1, pp. 266–274, 2001.

[13] J. Sadasivan, S. Mukherjee, and C. S. Seelamantula, "Joint dictionary training for bandwidth extension of speech signals," in *Proceedings International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5925–5929.

[14] P. B. Bachhav, M. Todisco, M. Mossi, C. Beaugeant, and N. Evans, "Artificial bandwidth extension using the constant-Q transform," in *Proceedings International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5550–5554.

[15] D. Marelli and P. Balazs, "On pole-zero model estimation methods minimizing a logarithmic criterion for speech analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 237–248, 2010.

[16] P. C. Loizou, *Speech enhancement: theory and practice*, 2nd ed. CRC press, 2007.

[17] P. Jax and P. Vary, "Artificial bandwidth extension of speech signals using MMSE estimation based on a Hidden Markov model," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003*, vol. 1. IEEE, 2003, pp. I–I.

[18] U. Kornagel, "Techniques for artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 86, no. 6, pp. 1296–1306, 2006.

[19] 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.

[20] P. Bauer and T. Fingscheidt, "A statistical framework for artificial bandwidth extension exploiting speech waveform and phonetic transcription," in *Proceedings 17th European Signal Processing Conference, 2009*. IEEE, 2009, pp. 1839–1843.

[21] ——, "An HMM-based artificial bandwidth extension evaluated by cross-language training and test," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4589–4592.

[22] Y. Wang, S. Zhao, W. Liu, M. Li, and J. Kuang, "Speech bandwidth expansion based on deep neural networks," in *Proceedings Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[23] M. B. Christopher, *Pattern recognition and machine learning*. Springer-Verlag New York, 2016.

[24] J. D. Markel and A. G. Jr., *Linear Prediction of Speech*, 1st ed., ser. Communication and Cybernetics 12. Springer-Verlag Berlin Heidelberg, 1976.

[25] MathWorks, "http://www.mathworks.com/."

[26] U. Shaked and Y. Theodor, "H^∞ optimal estimation: a tutorial," in *Proceedings 31st IEEE Conference on Decision and Control*. IEEE, 1992, pp. 2278–2286.

[27] T. Chen and B. A. Francis, "Design of multirate filter banks by H^∞ optimization," *IEEE Transactions on Signal Processing*, vol. 43, no. 12, pp. 2822–2830, 1995.

[28] Y. Yamamoto, M. Nagahara, and P. P. Khargonekar, "Signal Reconstruction via H^∞ Sampled-Data Control Theory Beyond the Shannon Paradigm," *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 613–625, 2012.

[29] T. Chen and B. A. Francis, *Optimal sampled-data control systems*. Springer, 1995, vol. 124.

[30] P. P. Vaidyanathan, *Multirate systems and filter banks*, ser. Prentice-Hall signal processing series. Prentice Hall, 1993.

[31] K. Glover and J. C. Doyle, "State-space formulae for all stabilizing controllers that satisfy an H^∞-norm bound and relations to relations to risk sensitivity," *Systems & control letters*, vol. 11, no. 3, pp. 167–172, 1988.

[32] F. Itakula, "Line spectrum representation of linear predictive coefficients of speech signal," *Journal of Acoustic Society of America*, 1975.

[33] N. Vaswani, "Jointly Gaussian random variables, MMSE and linear estimation," 2012.

[34] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1998, pp. 285–288.

[35] "ITU-T Software Tool Library 2009 Users Manual," *ITU-T Recommendation G.191*, Nov. 2009.

[36] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[37] P. Nizampatnam and K. K. Tappeta, "Bandwidth extension of narrowband speech using integer wavelet transform," *IET Signal Processing*, vol. 11, no. 4, pp. 437–445, 2016.

[38] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Detection, separation and recognition of speech from continuous signals using spectral factorisation," in *Proceedings 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 2649–2653.

[39] J. Abel, M. Strake, and T. Fingscheidt, "A Simple Cepstral Domain DNN Approach to Artificial Speech Bandwidth Extension," in *Proceedings International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5469–5473.

[40] L. Malfait, J. Berger, and M. Kastner, "P. 563The ITU-T standard for single-ended speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.