Circuits, Systems, and Signal Processing (2022) 41:2898–2922
https://doi.org/10.1007/s00034-021-01925-0

A New Framework for Artificial Bandwidth Extension Using H^∞ Filtering

Deepika Gupta[1] · **Hanumant Singh Shekhawat**[1] · **Rohit Sinha**[1]

Received: 25 February 2021 / Revised: 26 November 2021 / Accepted: 26 November 2021 /
Published online: 26 January 2022

Abstract

This work proposes a new artificial bandwidth extension (ABE) framework for enhancing the quality of narrowband speech signals. This enhancement process recovers missing high-frequency components of the signal. In this regard, a new bandwidth extension process based on H^∞ sampled-data control theory and machine learning models is proposed. In addition, a little non-ideality (aliasing) is allowed in the narrowband signal to get better reconstruction for the missing higher frequencies. The H^∞ sampled-data control theory works on a signal model, representing the already available wideband signal. Direct use of this theory is not possible in the bandwidth extension process as the signal models may not be the same for different phonemes of speech signals, even if uttered by the same speaker due to their non-stationary behavior. Hence, machine learning models are necessary. We have performed experiments with four types of narrowband features and two types of machine learning models approaches. The proposed method improves most of the measures when compared to the existing techniques, such as 12% minimum improvement in log spectral distance (LSD).

Keywords H^∞-norm · Codebook · Speech production filter · Lifting

A portion of this paper is published in Interspeech 2019 held in Graz, Austria, with title "Artificial Bandwidth Extension using H^∞ Optimization".

✉ Deepika Gupta
deepika.gupta@iitg.ac.in

Hanumant Singh Shekhawat
h.s.shekhawat@iitg.ac.in

Rohit Sinha
rsinha@iitg.ac.in

[1] Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India

List of symbols

$\mathbb{Z}$	The set of integers
LDTI	Linear discrete time invariant
$\mathbb{R}$	The set of real numbers
$\mathbb{R}^n$	n-Dimensional vector space over $\mathbb{R}$
$l^2(\mathbb{Z}, \mathbb{R})$	Square summable sequences in $\mathbb{R}^n$
$\|\|.\|\|_2$	l^2-norm of a discrete sequence
$\uparrow N$	Upsampler with an upsampling factor N, i.e., inserting the N-1 zero-valued samples between two consecutive original samples for increasing the sampling rate
$\downarrow N$	Downsampler with a downsampling factor N, i.e., keeping every Nth sample and deleting the remaining samples

1 Introduction

High-fidelity voice communications preserve the quality of message signals. In the Global System for Mobile communications (GSM), message signals are typically sampled at the rate of 8000 samples/sec [1]. According to the Nyquist criterion, the transmission bandwidth in the GSM happens to be narrow, i.e., limited to 0–4 kHz. Hence, frequencies in the human speech signals above the transmission bandwidth get suppressed. As a result, the naturalness, clarity, and pleasantness of the received signals deteriorate. Therefore, digital signal processing techniques are developed, which improve the quality of signal by extending bandwidth. More specifically, a narrowband (NB) telephone signal sampled at 8 kHz is processed to recover the frequency components higher than 4 kHz present in the original wideband (WB) signal sampled at 16 kHz. For this, the high-band (HB) information present in 4–8 kHz range is extracted from the wideband (0–8 kHz) signal. The extracted high-band information is further used in the bandwidth extension of the narrowband signal at the receiver end. This process is called artificial bandwidth extension (ABE) for a stationary narrowband signal. A general ABE process is shown in Fig. 1 in the case of a stationary signal.

Figure 1 consists of the transmitter setup and receiver setup. The transmitter setup generates the narrowband signal sampled at 8 kHz. Conventional transmitter setup has a low pass filter (LPF) followed by a downsampler with a downsampling fac-

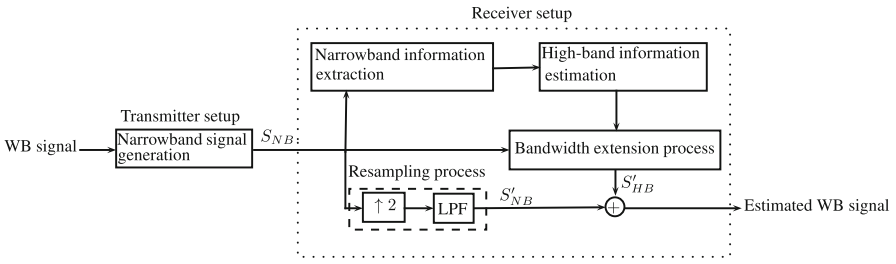

Fig. 1 A basic block diagram depicting the process to produce the narrowband signal and artificial bandwidth extension of a stationary narrowband signal

tor ($\downarrow 2$). The narrowband signal $S_{NB}[n]$ sampled at 8 kHz is the output signal of the transmitter-set. The receiver setup synthesizes the wideband signal. The receiver setup consists of four processes: narrowband information extraction, high-band information estimation, resampling process, and bandwidth extension process. In Fig. 1, $\uparrow 2$ represents an upsampler with an upsampling factor 2, $S'_{NB}[n']$ denotes the narrowband signal sampled at 16 kHz, and $S'_{HB}[n']$ denotes the estimated high-band signal sampled at 16 kHz. A bandwidth extension process is applied to the received narrowband signal $S_{NB}[n]$ for estimating the missing high-band signal at the receiver side. The bandwidth extension process uses high-band information, which is estimated using a machine learning model for given narrowband information/features. The machine learning model is trained offline. The narrowband features are extracted from the narrowband signal. In the resampling process, the resampled narrowband signal $S_{NB}[n']$ is obtained by passing the narrowband signal $S_{NB}[n]$ through the upsampler ($\uparrow 2$) followed by the low pass filter. The wideband signal is estimated by adding the estimated high-band signal $S'_{HB}[n']$ and narrowband signal $S_{NB}[n']$.

Many approaches are proposed for ABE based upon the source-filter model. In this model, the speech signal is segregated into two parts: speech production filter (SPF) as a vocal tract filter and excitation signal as a residue signal [52]. The excitation signal is passed through the speech production filter to produce the speech signal. The excitation signal can be either a white noise for the unvoiced speech or a quasi-periodic impulse train for the voiced speech. The magnitude spectrum of the excitation signal is flat in both cases: white noise and quasi-periodic impulse train. Thus, the vocal tract filter shapes the spectral envelope of the speech signal. The spectral envelope can be accurately modeled using a signal model containing the poles (resonances) as well as the zeros (anti-resonances) [38]. The spectral envelope and excitation of the high-band signal are estimated using an extrapolation process applied on the narrowband signal and some extra information [17,18,32,46,47]. In existing methods for ABE, spectral envelopes of the high-band signal and narrowband signal can be represented by linear prediction coefficients (LPC) [6], line spectral frequencies (LSF) [35], linear frequency cepstral coefficients (Cepstrum) [2], and Mel frequency cepstral coefficients (MFCC) [44,53] features. These features capture poles (formants) information present in the speech spectrum. Further, the high-band excitation can be estimated using many different ways, i.e., bandpass-envelope modulated Gaussian noise (BP-MGN) [47], harmonic noise model (HNM) [56], spectrum folding [17,37], pitch adaptive modulation [28], full-wave rectification [18], and spectral translation [18,28,37]. Another method has been proposed, which is based on the temporal envelope model [30]. It uses the temporal envelope and fine structure of the sub-bands for synthesizing the high-band speech signal. Some approaches are developed without using any modeling, and such approaches use the magnitude spectrum to synthesize the high-band information. A joint dictionary training model is proposed, which utilizes the sparsity of the spectrogram [50]. Log spectra of the wideband signal is directly used to represent narrowband and high-band information for ABE [11,34]. In [3], the Cepstrum feature is used to represent the high-band information for ABE. In [8], CQT (constant-Q transform) feature is used for ABE, but the dimension of this feature has been taken high.

According to [38], speech production filter can be accurately represented by a pole–zero model. Many existing methods use an all-pole model, which may not be sufficient to represent the spectral envelope of speech portions like fricatives, nasals, laterals, and the burst interval of stop consonants due to the presence of valleys in the frequency response of the SPF [38]. In our work, the pole–zero model (we call it the signal model also) is used to represent the spectral envelope of the wideband signal [38]. Moreover, existing methods focus on the estimation of the high-band (HB) signal only as the narrowband signal S_{NB} is available at the receiver side. At the transmitter side, the original wideband signal is passed through a near-ideal low pass filter (LPF) prior to the downsampler to produce the narrowband signal. The decomposition of narrowband and high-band information at the transmitter is a common technique used in many ABE works (see [2,35,58]), including our work reported in [21]. On account of the decomposition of narrowband and high-band information at the transmitter, two challenges arise for the effective ABE of the narrowband speech signal: (i) weaker conditional dependence between narrowband and wideband specifically for the unvoiced frames of speech and (ii) the need for the adjustment of energy levels between the estimated high-band and the retained narrowband speech signals [44,58]. In different unvoiced frames of speech, narrowband information is almost the same, while high-band information varies. Therefore, it is difficult to estimate respective high-band information for given narrowband information of the unvoiced frame. To tackle these challenges, a new ABE framework is proposed in this work. The proposed work differs from the existing works in two aspects. First, the narrowband signal generated at the transmitter is no longer perfect. It can be stated that the transmitted aliased narrowband signals may have less intelligibility, but these are hypothesized to establish the better conditional dependence between narrowband and wideband information. The narrowband signal includes aliasing distortion due to dropping the low pass filter prior to downsampler (a similar approach has been used in [20] also), which helps in the estimation of high-band information of the unvoiced speech. Because the high-band information is reflected in the narrowband region after downsampling, which yields more variations among the narrowband features for the unvoiced speech. This results in a better conditional dependency between narrowband features and proposed wideband features for the unvoiced speech. Second, the interpolation filter for the speech signal is estimated by using the H^{∞} optimization/filtering, which is recommended in the literature (especially in control) to handle variations in system models (in our case, the pole–zero model or signal model) [51]. This has been used in [7,60] for the reconstruction of the orchestral music signal by using a single pole–zero model. However, a single model is not sufficient for a non-stationary signal (orchestral music and speech signal [36,41]). Due to the non-stationary nature of speech signals, a frame-based approach (short-time processing) is applied to speech signals, which increases the necessity of storage for additional information about interpolation filters with their corresponding narrowband details. For this, machine learning models are designed and used to estimate the wideband information [2,9,10,17,27,28,32,35,58,59]. In this work, this problem is solved by using two machine learning models, Gaussian mixtures model [15] and feed-forward DNN [24].

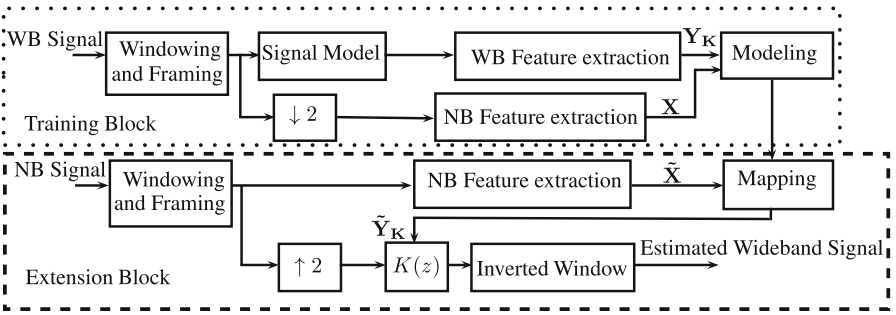

Fig. 2 Block diagram consists of training of a machine learning model and extension of the narrowband signal

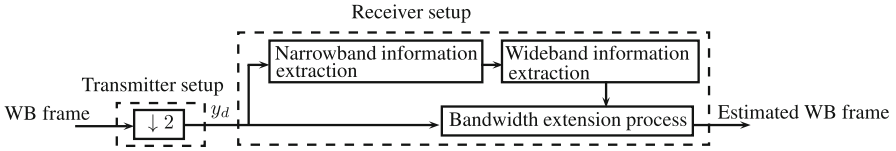

Fig. 3 Generation of the narrowband signal and reconstruction of the stationary wideband speech frame

2 A Proposed Setup for Artificial Bandwidth Extension of Speech Signals

This section discusses the proposed artificial bandwidth extension framework for the narrowband signal sampled at 8 kHz. Figure 2 shows an outline of the proposed ABE framework. It includes the windowing and framing processes, setups used at the transmitter side and receiver side (explained in Sect. 2.1), processes to obtain the wideband feature and narrowband feature for bandwidth extension (explained in Sect. 2.2), estimation of wideband feature (explained in Sect. 2.3), and synthesis of the wideband signal (explained in Sect. 2.4). The windowing and framing processes are performed to get stationary frames/signals from non-stationary speech signals [36]. It is done by using the Hamming window of 25 ms duration with 50% overlapping between adjoining frames. Each subblock of Fig. 2 is further explained in forthcoming subsections.

2.1 Setups Used at the Transmitter Side and Receiver Side

This section discusses the transmitter and receiver setups. These setups are combined and drawn in Fig. 3.

The transmitter (Tx) produces the narrowband signal at the output. A wideband speech frame is downsampled by a factor of 2 at the transmitter side. This leads to an output narrowband speech frame y_d, which is drawn in Fig. 3. This narrowband generation process introduces distortion (aliasing) in the narrowband speech frame. Hence, our work is focused on estimating the full wideband (0–8 kHz) signal at the receiver side. The receiver setup has three processes: narrowband information extraction, wideband information extraction, and bandwidth extension process (see Fig. 3).

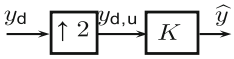

Fig. 4 Bandwidth extension process for a stationary speech frame

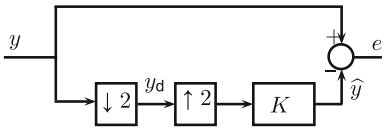

Fig. 5 Error system setup for the reconstruction of a stationary speech frame

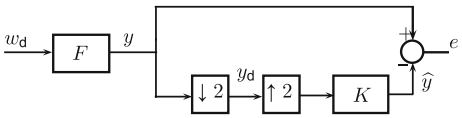

Fig. 6 Proposed architecture of error system for reconstructing a stationary speech frame

These processes are used at the receiver side for estimating wideband speech frames. A bandwidth extension process is applied to the narrowband speech frame at the receiver side, as shown in Fig. 4. In Fig. 4, y_d is upsampled by a factor of 2 and subsequently passed through an interpolation filter K. This leads to an estimated wideband speech frame $\widehat{y}$. The interpolation filter (K) contains the wideband information of a signal.

Designing the filter K is the core of this work. For designing the filter K, an error system is made by combining the wideband speech frame, narrowband generation process, and bandwidth extension process, as shown in Fig. 5.

The synthesis filter K is designed by minimizing the reconstruction error using a suitable norm.

In Fig. 5, $e = y - \widehat{y}$. y and $\widehat{y}$ denote the original/true wideband speech frame and estimated wideband speech frame, respectively.

Every discrete-time stationary speech signal can be represented by a linear discrete time-invariant (LDTI) system driven by a white noise for unvoiced speech or an impulse train for voiced speech [38]. Hence, pole–zero information about the original wideband speech frame y is extracted in the form of a signal model F as the speech production filter, which reflects the signal properties. In other words, the signal model F represents the spectral envelope information of the wideband speech frame. A modified error system containing the signal model F is given in Fig. 6.

In Fig. 6, y is the output of system F driven by an input w_d with known features. 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. F can be represented in the z-domain as

$$F(z) = \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B},$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are constant real matrices of appropriate dimensions. The signal model $F(z)$ is computed by the standard Prony's method-based function available in MATLAB [39,40]. The obtained model is causal and but may be unstable. To make it stable, those poles of the model, lying outside of the unit circle, are emulated inside by reciprocating their magnitudes without altering the phase [38]. Note that the magnitude spectrum of $F(z)$ remains the same, however, the phase spectrum changes. This stabilizing process does not affect too much the perception of a speech signal because the human auditory system is less sensitive to phase information [38].

2.1.1 Performance Index

The H^∞ system norm is used to minimize the reconstruction error. Because this norm handles small modeling errors [51]. The H^∞-norm of a system $\mathcal{G}$ with input $\mathcal{X} \in l^2(\mathbb{Z}, \mathbb{R}^n)$ and output $\mathcal{Y} \in l^2(\mathbb{Z}, \mathbb{R}^m)$ is defined as (see, e.g., [13,51,60])

$$||\mathcal{G}||_\infty := \sup_{\mathcal{X}\neq 0} \frac{||\mathcal{Y}||_2}{||\mathcal{X}||_2}. \tag{1}$$

2.1.2 Problem Formulation

To design optimal $K(z)$, the following optimization problem is solved.

Problem 1 Given a stable and causal $F(z)$, design a stable and causal interpolation filter K_{opt} defined as

$$K_{\text{opt}} := \arg\min_K (||\mathbb{T}||_\infty), \tag{2}$$

where $\mathbb{T} := F - K(\uparrow 2)(\downarrow 2)F$. $\mathbb{T}$ maps w_d to e (see Fig. 6).

As mentioned earlier, the non-stationary behavior of speech signals introduces some uncertainty in the estimation of the signal model $F(z)$. In such a case, H^∞-norm optimization provides a robust solution against small modeling error in $F(z)$ [51]. Solution of Problem 1 is explained in "Appendix A.1." It computes the optimal IIR filter K_{opt}. Henceforth, K_{opt} is denoted by K.

2.2 Speech-Specific Wideband and Narrowband Features

The strategy explained in Sect. 2.1 is used for extending the bandwidth of the narrowband speech frame. Further, interpolation filters are obtained for all speech frames. The interpolation filter K has an infinite impulse response (IIR). Practically, the IIR filter K cannot be modeled directly by machine learning techniques. Therefore, this filter is converted into an approximate finite impulse response (FIR) interpolation filter by truncating its Taylor series at the origin. The number of terms in FIR interpolation

filter is chosen 21 empirically, which is explained in Sect. 3.1. This FIR filter response is taken as the wideband feature $\mathbf{Y_K}$ in this work.

Only narrowband information is available on the receiver side. The interpolation filter is needed in the bandwidth extension process. For estimating the interpolation filter, a pre-trained model is trained using the interpolation filter information and corresponding narrowband information (narrowband feature). The pre-trained model is further used to estimate the filter information for a given narrowband feature (see Sect. 2.3). The narrowband information (narrowband feature) is taken in four different ways, i.e., linear prediction coefficients (LPC) [5], line spectral frequencies (LSF) [25], linear frequency cepstral coefficients (Cepstrum) [2], and Mel frequency cepstral coefficients (MFCC) [44,53]. These parameters are computed from the narrowband speech frame. The dimension of the narrowband feature is fixed to 10.

2.3 Modeling and Mapping

This section has details of the machine learning models used in this paper. Machine learning models are used to estimate the FIR interpolation filter using the narrowband feature. For this purpose, a pre-trained model is trained using the narrowband and wideband features. In our work, machine learning models such as GMM and DNN are used, which are explained in "Appendixes A.2 and A.3," respectively.

2.4 Wideband Signal Estimation

The entire flow for the training of a machine learning model and extension of the narrowband signal is shown in Fig. 2, which is used for ABE of speech signals. It can be broadly divided into two principal blocks: training and extension. In the training block, windowing of the wideband signal is performed first. Two parallel processes are then performed on the windowed wideband signal. The one process is the computation of signal model and subsequent extraction of wideband feature $\mathbf{Y_K}$ (see Sects. 2.1 and 2.2). The another one performs the downsampling of wideband speech frame and subsequent extraction of narrowband feature $\mathbf{X}$ (see Sect. 2.2). Narrowband and wideband features are modeled by GMM or DNN (see Sect. 2.3). In the extension block, the first step is the windowing of narrowband signal and subsequent extraction of narrowband feature $\tilde{\mathbf{X}}$. Further, $\tilde{\mathbf{X}}$ is mapped to the wideband feature $\tilde{\mathbf{Y}}_K$ by using the pre-trained model. The windowed narrowband signal is upsampled by a factor of 2 and then passed through the interpolation filter $K(z)$. $K(z)$ is obtained by the estimated wideband feature $\tilde{\mathbf{Y}}_K$. The resulting signal is multiplied by the reciprocal of the Hamming window to estimate the wideband speech frame. Further, the overlapped portion of two adjacent frames is estimated by averaging the overlapped parts of the estimated wideband speech frames. In other words, the weighted overlap-add method (WOLA) is applied [16,57].

3 Experimental Analysis and Results

This section has a description of the speech signals, which are taken from the TIMIT database [61] and RSR15 database [33]. Both the datasets contain the recorded speech files at a sampling rate of 16 kHz. TIMIT database consists of two different sets: test set and training set. The training set is used for training the machine learning models, and the test set is taken as a validation set. A new test set is made by some speech files taken from the RSR15 dataset and used for testing the machine learning models. This new test set has the speech files uttered by 4 female and 3 male speakers. The test set from a different database leads to more generalized results.
Section 3.1 has the mathematical formulations of objective measures used for evaluating the proposed method. In Sect. 3.2, the objective measures are analyzed for deciding the dimension of the wideband feature. Further, the proposed method is evaluated using the GMM model in Sect. 3.2.1 and DNN topology in Sect. 3.2.2. In Sect. 3.2.3, the proposed method is compared with the existing methods. In Sect. 3.3, the subjective measure is discussed.

3.1 Objective Measures

In this work, several standard objective speech quality measures such as MSE (mean square error) [43], SDR (signal to distortion ratio) [23], LLR (log likelihood ratio) [36,49], LSD (log spectral distance) [3], MOS-LQO (mean opinion score listening quality objective) estimated from PESQ (perceptual evaluation of speech quality) [26, 49], and STOI (short-time objective intelligibility) [54] are computed for performance analysis. Mathematical formulations of objective measures are written as follows

$$\mathrm{MSE} = \frac{\sum_{i=1}^{L}(s(i) - \tilde{s}(i))^2}{L}. \tag{3}$$

L is the signal length, s is the original wideband signal, and $\tilde{s}$ is the reconstructed wideband signal.

$$\mathrm{SDR(dB)} = 10\log_{10}\frac{\sum_{i=1}^{L}(s(i)^2}{\sum_{i=1}^{L}(s(i) - \tilde{s}(i))^2}. \tag{4}$$

Parameters in (4) are the same as defined in (3).

$$\mathrm{LLR} = \frac{\sum_{i=1}^{M}\log_{10}\left(\frac{\overrightarrow{a_i}_p^T R_{ic} \overrightarrow{a_i}_p}{\overrightarrow{a_i}_c^T R_{ic} \overrightarrow{a_i}_c}\right)}{M}, \tag{5}$$

where M is the number of frames, $\overrightarrow{a_i}_c$ and $\overrightarrow{a_i}_p$ are the LPC vectors of the original ith speech frame and reconstructed ith speech frame, respectively. R_{ic} is the autocor-

Table 1 Performance comparison of speech signals enhanced by applying an upsampler with an upsampling factor 2 (without applying filter K) and the oracle interpolation filter K in Fig. 4 on the speech files taken from the validation set

Output subblock	MSE ($\times 10^{-5}$)	SDR	LLR	MOS-LQO	LSD	STOI
Upsampler	81.1673	3.01	1.4254	3.5044	11.3135	0.9015
Interpolation filter K	4.8634	15.81	0.6547	3.8047	7.6220	0.9403

relation matrix of the original ith speech frame.

$$\text{LSD} = \frac{\sum_{i=1}^{M} \sqrt{\left(\frac{\sum_{j=1}^{N} (20\log_{10}|X(i,j)| - 20\log_{10}|\tilde{X}(i,j)|)^2}{N}\right)}}{M}, \tag{6}$$

with $|X(i, j)|$ and $\tilde{X}(i, j)$ being the absolute values of FFT of ith frame and jth frequency bin of the original and reconstructed speech frames, respectively. M and N denote the number of frames and the number of frequency bins, respectively.

$$\text{MOS-LQO} = a + \frac{b}{(1 + \exp(c * p + d))} \tag{7}$$

with $a = 0.999$, $b = 4.999 - a$, $c = -1.4945$, $d = 4.6607$, and p is PESQ.
These measures are characterized into two major categories based on frequency and time domain. MSE and SDR yield performance with respect to time. LLR and LSD yield information about frequencies. MOS-LQO and STOI measures are suitable to measure quality together in both the time and frequency domain. LLR, SDR, and PESQ measures are computed with the help of a composite tool downloaded from the website of the author. MOS-LQO is estimated from the PESQ [22,26].

3.2 Objective Analysis

Initially, the performance of enhanced speech signals is analyzed. The narrowband speech signal is enhanced by applying the interpolation filter K on the upsampled narrowband signal in the condition of using the oracle filter K directly in the architecture shown in Fig. 4. The objective measures are listed in Table 1 for the wideband speech signals estimated by the output of an upsampler with upsampling factor 2 ($y_{d,u}$) and output of the interpolation filter K ($\widehat{y}$).

In Table 1, the interpolation filter K improves all the objective measures significantly.

Moreover, filter K has an infinite impulse response. It is transformed into an approximate FIR filter by using the Taylor series truncation method. For deciding the length of the truncated FIR filter, the objective measures are computed on some speech files taken from the validation set with the varying length of the filter.

Table 2 Performances evaluation on the speech files taken from the validation set in condition of direct implanting FIR filter K (oracle K) in Fig. 4 for ABE

Number of terms	MSE ($\times 10^{-5}$)	SDR	LLR	MOS-LQO	LSD	STOI
11	8.9405	13.18	0.7925	3.7450	8.2260	0.9308
15	7.4762	13.74	0.7851	3.7521	8.1389	0.9319
21	6.0912	14.79	0.7233	3.7782	7.9339	0.9355
25	5.8136	15.06	0.7065	3.7810	7.8678	0.9367
31	5.6043	15.25	0.6937	3.7854	7.8078	0.9374
∞	**4.8634**	**15.81**	**0.6547**	**3.8047**	**7.6220**	**0.9403**

In Table 2, the objective measures are improved with increasing the number of terms present in the FIR filter, but gradually after the length 21. Hence, the filter length is set to 21. Then, the pre-trained models GMM and DNN are obtained using the training data information. Then, the performance of the test set is analyzed using the pre-trained models, as described in the following subsections.

Moreover, the objective measures are analyzed for the voiced speech and unvoiced speech of the test set separately. So, speech signals are segregated into two fundamental parts: voiced speech and unvoiced speech by a glottal activity detection (GAD) method [4,42]. It is a well-known fact that the narrowband region contains higher energy than the high-band region for voiced speech and vice versa for unvoiced speech [36]. Our proposed strategy considers the recovery of full wideband. This is because, information present in the narrowband region is distorted because of aliasing; however, information present in the high-band region is lost because the wideband signal is converted into the narrowband signal. As a result, unvoiced speech and voiced speech are affected in our transmitter setup. The main benefit of direct downsampling is the better estimation of wideband feature for a given narrowband feature of the unvoiced speech. Because the high-band information is reflected in the narrowband region after downsampling, which yields more variations among the narrowband features for the unvoiced speech. This results in the better conditional dependence between narrowband features and proposed wideband features for the unvoiced speech. Later, the performance is analyzed for the voiced speech and unvoiced speech separately.

3.2.1 Performance Evaluation Using Gaussian Mixture Model

The GMM-based regression technique is used to estimate the interpolation filter (wideband feature) for a given narrowband feature. GMM model with 128 mixtures is trained using the narrowband features and proposed wideband features. Further, the performance of the proposed approach using the GMM model is evaluated on the test set for four types of narrowband features: LSF, LPC, Cepstrum, and MFCC, as done in Table 3.

The objective measures are analyzed for these narrowband features. The LSF narrowband feature leads to the best performance in comparison with the other narrowband features.

Table 3 Performance evaluation by using 128 GMMs on the test set

Features	MSE ($\times 10^{-4}$)	SDR	LLR	MOS-LQO	LSD	STOI
LSF+$\tilde{\mathbf{Y}}_K$	**3.4667**	**11.17**	**0.6063**	**3.5653**	**7.9945**	**0.9028**
LPC+$\tilde{\mathbf{Y}}_K$	3.6206	10.73	0.6722	3.5629	8.4141	0.8994
Cepstrum+$\tilde{\mathbf{Y}}_K$	3.4719	10.86	0.7192	3.5524	8.7476	0.8952
MFCC+$\tilde{\mathbf{Y}}_K$	3.6033	10.90	0.6385	3.5642	8.2438	0.9002
FIR filter K ($\mathbf{Y_K}$) used directly	1.7615	12.66	0.4980	3.8010	7.6397	0.9189
IIR filter K used directly	1.4563	13.52	0.4576	3.8310	7.4670	0.9231

Table 4 Performance evaluation by using 128 GMMs for voiced speech extracted from the speech signals belonging to the test set

Features	MSE ($\times 10^{-4}$)	SDR	LLR	MOS-LQO	LSD
LSF+$\tilde{\mathbf{Y}}_\mathbf{K}$	**4.6301**	**12.96**	0.9279	4.1549	7.7356
LPC+$\tilde{\mathbf{Y}}_\mathbf{K}$	5.0784	12.13	0.9784	4.1480	7.9903
Cepstrum+$\tilde{\mathbf{Y}}_\mathbf{K}$	4.8338	11.93	0.9729	4.1465	8.0571
MFCC+$\tilde{\mathbf{Y}}_\mathbf{K}$	4.8117	12.56	**0.8993**	**4.1562**	**7.7075**
FIR filter K ($\mathbf{Y_K}$) used directly	3.5890	13.30	0.9130	4.1840	7.9940
IIR filter K used directly	2.8542	14.26	0.8205	4.2055	7.6405

Table 5 Performance evaluation by using 128 GMMs for unvoiced speech extracted from the speech signals belonging to the test set

Features	MSE ($\times 10^{-4}$)	SDR	LLR	MOS-LQO	LSD
LSF+$\tilde{\mathbf{Y}}_\mathbf{K}$	5.0088	**9.30**	**0.6104**	**3.8540**	**7.8360**
LPC+$\tilde{\mathbf{Y}}_\mathbf{K}$	5.0451	8.40	0.6725	3.8504	8.2167
Cepstrum+$\tilde{\mathbf{Y}}_\mathbf{K}$	**4.8122**	7.72	0.7085	3.8380	8.5267
MFCC+$\tilde{\mathbf{Y}}_\mathbf{K}$	5.2222	8.58	0.6447	3.8481	8.0860
FIR filter K ($\mathbf{Y_K}$) used directly	1.5410	10.84	0.4508	3.9947	7.2843
IIR filter K used directly	1.3833	11.27	0.4186	4.0105	7.1523

Furthermore, objective measures are tabulated in Table 4 for the voiced speech and Table 5 for the unvoiced speech extracted from speech signals belonging to the test set by considering the four types of narrowband feature representations.

MSE and SDR measures produced by using LSF narrowband feature are more close to their respective values obtained by using the oracle FIR filter K ($\mathbf{Y_K}$) directly for the voiced speech. The rest of the objective measures produced by using the MFCC narrowband feature are leading to the lowest difference from their respective values

Table 6 Performance evaluation on the validation set for different DNN topologies with varying the number of hidden layers (N_{HL}) and the number of units (N_U), and ReLU activation function in hidden layers, linear activation function in the output layers, LSF narrowband feature and *AdaMax* optimizer

Topology with ReLU activation functions		Performance on the validation set					
N_{HL}	N_U	MSE ($\times 10^{-5}$)	SDR	LLR	MOS-LQO	LSD	STOI
2	512	3.3349	15.19	0.7082	3.6948	7.7229	0.9328
2	1024	3.3347	15.20	0.7074	3.6964	7.7202	0.9329
3	128	3.3376	15.19	0.7053	3.6935	7.7166	0.9326
3	256	3.3386	15.20	0.7046	3.6966	7.7131	0.9327
3	512	3.3453	15.19	0.7055	3.6963	7.7162	0.9328
3	1024	3.3521	15.19	0.7064	3.6981	7.7207	0.9328
4	128	3.3292	15.21	0.7033	3.6908	7.7113	0.9328
4	**256**	3.3174	15.23	0.7023	3.6916	7.7084	0.9330
4	512	3.3247	15.22	0.7025	3.6928	7.7097	0.9330
4	1024	3.3411	15.20	0.7042	3.6924	7.7175	0.9329
FIR filter K ($\mathbf{Y_K}$)	Used directly	3.6892	14.85	0.7170	3.8086	7.9457	0.9346
IIR filter K	Used directly	2.9609	15.86	0.6527	3.8252	7.6405	0.9398

obtained by using $\mathbf{Y_K}$ directly for the voiced speech. The Cepstrum narrowband feature yields the lowest MSE, and the LSF narrowband feature produces the better remaining objective measures for the unvoiced speech.

3.2.2 Performance Evaluation Using Deep Neural Network

DNN topology is used to estimate the interpolation filter coefficients. Some preliminary experiments are done to decide the parameter values for DNN topology with fixing the narrowband feature. An optimal DNN architecture is designed after optimizing its parameters over the fixed LSF narrowband feature representation. AdaMax (adaptive moment estimation based on the infinity norm) [31] optimizer is used to update the weights of the network by applying L_2 regularization empirically [24]. Experimentally hyper-parameters such as mini-batch size, epoch, learning rate α, decay rates β_1 for the first-moment estimate and β_2 for the second-moment estimate over a broad range are set to 200, 50, 0.01, 0.9, and 0.999, respectively. Mean and variance normalization (MVN) is applied to the features. Also, batch normalization before activation function is applied to each hidden layer. The ReLU activation function is used in hidden layers, and the linear activation function is used in the output layer. Performances of different DNN topologies on the validation set are tabulated in Table 6.

The overall good performance on the validation set is acquired by four hidden layers and 256 hidden units. Next, this architecture is trained by changing mini-batch size. As a result, the mini-batch size is decided 50. These obtained parameters are selected in designing the optimal DNN architecture.

Table 7 Performance evaluation on the test set for the DNN models designed using different activation functions such as ReLU, ELU, tanh, softplus used in hidden layers, and linear in the output layer; Number of hidden layers (N_{HL}) = 4; Number of units (N_U) = 256 in each hidden layer

Features	Activation functions	MSE ($\times 10^{-4}$)	SDR	LLR	MOS-LQO	LSD	STOI
LSF+$\tilde{\mathbf{Y}}_K$	ReLU	3.2783	11.61	0.6350	3.5643	8.1894	0.9002
	ELU	3.2978	11.61	**0.6310**	3.5639	8.1649	0.9021
	tanh	3.2898	11.60	0.6351	3.5655	8.1652	0.9005
	Softplus	3.3138	**11.63**	0.6321	3.5641	**8.1645**	0.9026
LPC+$\tilde{\mathbf{Y}}_K$	ReLU	**3.2677**	11.62	0.6487	3.5660	8.2687	0.9001
	ELU	3.3318	11.59	0.6376	3.5661	8.2022	0.9020
	tanh	3.2886	11.59	0.6392	**3.5676**	8.2105	0.9004
	Softplus	3.3535	11.58	0.6350	3.5644	8.1882	**0.9031**
Cepstrum+$\tilde{\mathbf{Y}}_K$	ReLU	4.2454	9.75	0.9356	3.4481	9.6169	0.8626
	ELU	4.3148	9.55	0.9193	3.4814	9.5817	0.8681
	tanh	3.7744	10.06	0.9033	3.5018	9.5110	0.8708
	Softplus	4.1992	9.71	0.8885	3.4949	9.4726	0.8722
MFCC+$\tilde{\mathbf{Y}}_K$	ReLU	3.5402	11.20	0.6525	3.5579	8.2966	0.8964
	ELU	3.5798	11.21	0.6427	3.5586	8.2379	0.8980
	tanh	3.5088	11.21	0.6442	3.5611	8.2430	0.8969
	Softplus	3.5705	11.21	0.6429	3.5601	8.2372	0.8990
FIR filter K ($\mathbf{Y_K}$)	Used directly	1.7615	12.66	0.4980	3.8010	7.6397	0.9189
IIR filter K	Used directly	1.4563	13.52	0.4576	3.8310	7.4670	0.9231

Moreover, different DNN models are trained with other activation functions in the hidden layers such as ELU, tanh, and softplus. Performance on the test set is analyzed for all the DNN architectures, as shown in Table 7.

It is analyzed that the LPC narrowband feature yields better MOS-LQO, MSE, and STOI than the other narrowband features. On the other hand, the rest of the objective measures in the majority of the cases are better for the LSF narrowband feature. Among all the activation functions, the softplus function yields the best performance in the majority of the cases using the LSF narrowband feature. Furthermore, Tables 8 and 9 give the objective measures computed for the voiced speech and unvoiced speech taken from the test set, respectively, with different activation functions and different narrowband feature definitions.

The LSF narrowband feature, among all the narrowband features, yields the best performance for voiced speech and unvoiced speech. The LSF narrowband feature yields the best SDR, LLR, and LSD using the ELU, ReLU, and softplus activation functions in the DNN model for the voiced speech, respectively. The LPC narrowband feature yields the best MSE and MOS-LQO using the ReLU and tanh functions in the DNN model for the voiced speech, respectively. For the unvoiced speech, the LSF narrowband feature and ELU, tanh, and softplus functions used in designing of the DNN model yield the closest SDR, LLR, and LSD to their respective values

Table 8 Performance evaluation for voiced speech extracted from speech signals belonging to the test set for the DNN models designed using different activation functions such as ReLU, ELU, tanh, softplus used in hidden layers, and fixed linear activation function in the output layer; Number of hidden layers (N_{HL}) = 4; Number of units (N_U) = 256 in each hidden layer

Features	Activation functions	MSE ($\times 10^{-4}$)	SDR	LLR	MOS-LQO	LSD
LSF+$\tilde{\mathbf{Y}}_K$	ReLU	4.0418	13.63	**0.8924**	4.1548	7.6249
	ELU	4.0361	**13.66**	0.8942	4.1550	7.6243
	tanh	4.0595	13.62	0.8933	4.1558	7.6264
	Softplus	4.0380	13.65	0.8941	4.1544	**7.6215**
LPC+$\tilde{\mathbf{Y}}_K$	ReLU	**4.0240**	13.53	0.8988	4.1556	7.6530
	ELU	4.0643	13.57	0.8960	4.1563	7.6381
	tanh	4.0613	13.53	0.8957	**4.1569**	7.6410
	Softplus	4.0682	13.56	0.8954	4.1555	7.6314
Cepstrum+$\tilde{\mathbf{Y}}_K$	ReLU	7.4300	10.03	1.1916	4.0265	8.9441
	ELU	8.0092	9.83	1.1966	4.0549	9.0048
	tanh	6.4668	10.34	1.1516	4.0870	8.8377
	Softplus	7.5313	9.94	1.1718	4.0668	8.9233
MFCC+$\tilde{\mathbf{Y}}_K$	ReLU	4.4334	12.98	0.9238	4.1508	7.7619
	ELU	4.4389	13.04	0.9191	4.1519	7.7386
	tanh	4.4283	13.00	0.9197	4.1533	7.7441
	Softplus	4.4286	13.04	0.9183	4.1523	7.7307
FIR filter K ($\mathbf{Y_K}$)	Used directly	3.5890	13.30	0.9130	4.1840	7.9940
IIR filter K	Used directly	2.8542	14.26	0.8205	4.2055	7.6405

obtained by using oracle $\mathbf{Y_K}$ directly, respectively. The DNN model designed using the ELU activation function and Cepstrum narrowband feature yield the best MSE for the unvoiced speech. The DNN model designed using the softplus activation function and LPC narrowband feature yield the best MOS-LQO for the unvoiced speech.

3.2.3 Comparisons

Our proposed method is compared with the existing methods based on the conventional source-filter model wherein the excitation signal is extended by two different ways such as spectrum folding [17,37,58] and spectrum translation [37,44]. Experimental conditions such as datasets, dimensions of narrowband and wideband features, windowing, and DNN model are kept the same. The LSF features are used to represent the narrowband feature and wideband feature. Also, these methods require a gain factor, which is calculated by following [58] for spectrum folding and [44] spectral translation. The cepstral domain method is also compared in which the narrowband feature is the narrowband magnitude spectrum and the wideband feature is represented by cepstral coefficients [3].

Moreover, these techniques are implemented by using the low pass filter for generating the narrowband signal. Here, the low pass filter is a non-causal FIR filter defined

Table 9 Performance evaluation for the unvoiced speech extracted from speech files belonging to the test set for the DNN models with considering different activation functions such as ReLU, ELU, tanh, softplus used in hidden layers, and fixed linear activation function in the output layer; Number of hidden layers $(N_{HL}) = 4$; Number of units $(N_U) = 256$ in each hidden layer

Features	Activation functions	MSE ($\times 10^{-4}$)	SDR	LLR	MOS-LQO	LSD
LSF+$\tilde{\mathbf{Y}}_K$	ReLU	5.0273	9.42	0.6365	3.8445	8.0101
	ELU	5.0752	**9.49**	0.6339	3.8453	7.9882
	tanh	5.0406	9.46	**0.6332**	3.8464	7.9888
	Softplus	5.1033	9.44	0.6367	3.8457	**7.9930**
LPC+$\tilde{\mathbf{Y}}_K$	ReLU	5.0134	9.23	0.6473	3.8451	8.0846
	ELU	5.1456	9.33	0.6406	3.8460	8.0306
	tanh	5.0434	9.30	0.6395	3.8464	8.0314
	Softplus	5.1937	9.29	0.6409	**3.8477**	8.0236
Cepstrum+$\tilde{\mathbf{Y}}_K$	ReLU	4.4418	6.02	0.9379	3.7447	9.3435
	ELU	**4.0677**	6.11	0.9120	3.7735	9.2829
	tanh	4.0885	6.24	0.8941	3.7837	9.2438
	Softplus	4.2634	6.25	0.8805	3.7791	9.1869
MFCC+$\tilde{\mathbf{Y}}_K$	ReLU	5.3710	8.80	0.6604	3.8406	8.1386
	ELU	5.4654	8.90	0.6535	3.8413	8.0874
	tanh	5.2900	8.91	0.6510	3.8428	8.0861
	Softplus	5.4495	8.89	0.6543	3.8419	8.0885
FIR filter K ($\mathbf{Y_K}$)	Used directly	1.5410	10.84	0.4508	3.9947	7.2843
IIR filter K	Used directly	1.3833	11.27	0.4186	4.0105	7.1523

Table 10 A comparison of the objective measures computed on the test set speech files for different methods

Methods	MSE ($\times 10^{-4}$)	SDR	LLR	MOS-LQO	LSD	STOI
Spectrum folding method	10.0877	4.95	0.8688	4.4135	9.3018	0.9192
Spectral translation	9.9643	5.00	0.7917	4.3487	9.4874	0.9291
Cepstral domain	9.8040	5.09	0.7136	4.4102	9.4502	0.9208
Proposed method	3.3138	11.61	0.6321	3.5641	8.1645	0.9026

in [1]. Cut off frequency of the LPF filter is 3660 Hz. The length of this filter is 118. Non-causality of this filter introduces a delay in transmission.

As seen in Table 10, the proposed method improves all the objective measures except the MOS-LQO and STOI when compared with the existing methods. MOS-LQO and STOI values are obtained better by the existing methods. It may be due to the available original narrowband information. In the existing methods, the narrowband signal is generated by using the low pass filter. Therefore, the narrowband information does not alter.

Next, spectrograms of the estimated speech signals are shown in Fig. 7, which are estimated by the proposed, spectrum folding, spectral translation, and cepstral

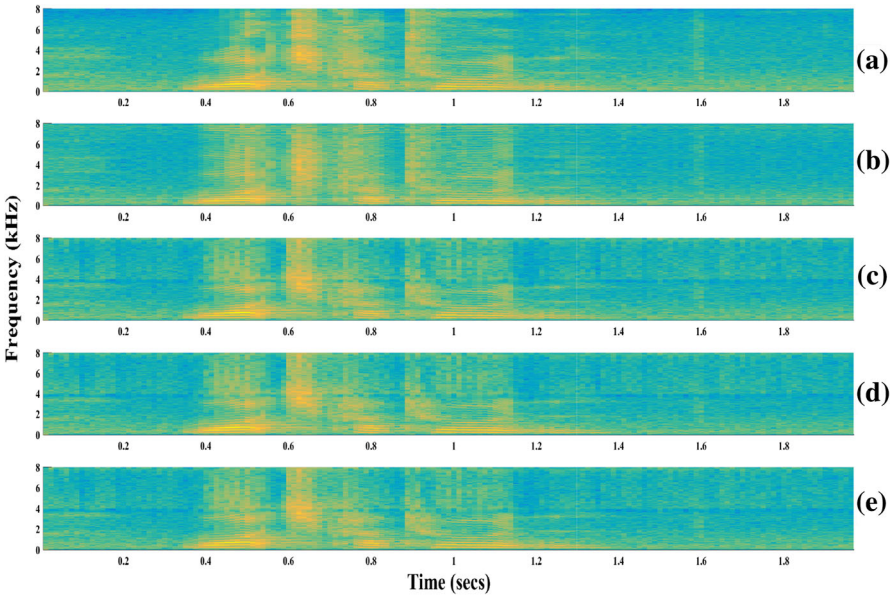

Fig. 7 Spectrogram of **a** original wideband signal, **b**, **c**, **d**, and **e** reconstructed wideband signal by proposed method, spectrum folding, spectral translation (**e**) cepstral domain, respectively

domain methods using the same DNN model. As viewed in Fig. 7, the spectrogram of the extended speech signal has more difference around 4 kHz from the original spectrogram for the existing methods than the proposed method. It has happened because of the energy levels adjustment issue around 4 kHz in the existing methods. It is observed around 0.9 s and 0.77 s in Fig. 7 that the estimated high-band information is more close to the original high-band information by the proposed method than the existing methods. However, the estimated high-band information around 7–8 kHz and 0.40–0.55 s in Fig. 7 is observed more than the original information by the proposed method when compared with the existing methods.

3.3 Subjective Listening Test

Subjective assessment is done according to the ITU-T P.800 [48, Annex E] for examining the speech quality. This task is done for the extended speech signals obtained by the proposed method, spectrum folding method, spectrum translation method, and cepstral domain method using the DNN model with the softplus activation function. Extended wideband speech files by the proposed method are rated with respect to extended wideband speech files by the existing methods. Ten pairs of extended speech signals belonging to the test set are randomly chosen for these methods, i.e., 60 files total. Then, twelve listeners were asked to give a mean opinion score (MOS) value between -3 (much worse) to 3 (much better). The ages of these listeners are between 23 and 32 years. These listeners do not have any hearing impairment and understand well English language. They were permitted to listen to the speech files more than once. Further, 95% confidence interval (CI) and p values are computed for measur-

Table 11 Subjective assessment on artificially extended speech files belonging to the test set by the proposed method with respect to the existing methods

Conditions	CMOS	CI	p
Spectrum folding versus proposed method	1.80	[1.5806 2.0194]	$< .001$
Spectral translation versus proposed method	0.96	[0.7704 1.1546]	$< .001$
Cepstral domain versus proposed method	1.59	[1.3589 1.8161]	$< .001$

ing statistical significance. Then, the comparison mean opinion score (CMOS), 95% confidence interval (CI), and p values are listed in Table 11.

Our proposed method improves CMOS significantly by 1.80, 0.96, and 1.59 points in comparison with the spectrum folding, spectral translation, and cepstral domain, respectively. Unvoiced phonemes are perceived better in the extended speech files using the proposed method than the existing methods. For reference, the speech files are provided for all the conditions that can be accessed using the link.[1]

4 Conclusion

A new framework (which capitalizes on artificially introduced non-ideality in the narrowband signal) is proposed for the artificial bandwidth extension of speech signals. In our proposed framework, the transmitter setup is different from the existing setup, which helps mainly in identifying the high-frequency components for the unvoiced speech. The discrete interpolation filter is obtained by using a signal model with the help of H^{∞} optimization. The obtained rational stable and causal interpolation filter is converted into an FIR filter empirically. This FIR filter is considered as the wideband feature. Experiments are performed by considering four types of narrowband features such as LSF, LPC, MFCC, and Cepstrum. Estimation of wideband feature for a given narrowband feature is conducted by two different modeling techniques such as GMM and DNN with several topologies. Performance is analyzed on the test set speech files taken from the RSR15 database by computing the standard objective measures: SDR, MSE, MOS-LQO, LLR, STOI, and LSD and subjective listening test. Also, the objective measures are analyzed for the voiced speech and unvoiced speech separately. The proposed method gives better results except for the MOS-LQO and STOI objective measures in comparison with the existing methods using the DNN model. In the listening test, CMOS is achieved higher by the proposed method than the existing methods.

Data Availability Two datasets are used in this study. First is the TIMIT database, which can be accessed on the link (https://deepai.org/dataset/timit). Second is the RSR15 database, which is available on the link (https://projets-lium.univ-lemans.fr/sidekit/tutorial/RSR2015.html). However, restrictions are applied to the availability of the RSR15 database. Therefore, it is not publicly available.

[1] https://drive.google.com/file/d/1DFTuI98EUlWb2PJ4fHQzvck3k0Yai0Pd/view?usp=sharing.

A Appendix

A.1 Solution of Problem 1

Solution of Problem 1 is essentially from [12,60]. The error system $\mathbb{T}$ in Fig. 6 is a multi-rate system because of the presence of the upsampler and downsampler. It can be transformed into a single rate system $\overline{\mathbb{T}}$ by using the lifting operation [12,13]. Discrete-time lifting operator $\mathbb{L}$ by a factor of N is defined by $\mathbb{L}_N$ in the time domain and it is defined as [13]

$$\mathbb{L}_N : \quad l^2(\mathbb{Z}, \mathbb{R}) \rightarrow l^2(\mathbb{Z}, \mathbb{R}^N), \tag{8}$$

$$\left\{ v[0], v[1], ., v[N-1], v[N], v[N+1], ., v[2N-1].. \right\} \rightarrow \begin{Bmatrix} v[0] & v[N] \\ v[1] & v[N+1] \\ . & . \quad \dots \\ . & . \\ v[N-1] & v[2N-1] \end{Bmatrix} \tag{9}$$

It is straight forward to see that $\mathbb{L}_N$ is an invertible operator. The z-transform representations of lifting and inverse lifting are [55,60]

$$\mathbf{L_N} = (\downarrow N) \begin{bmatrix} 1 & z & z^2 & ..... & z^{N-1} \end{bmatrix}^T \tag{10a}$$

$$\mathbf{L_N^{-1}} = \begin{bmatrix} 1 & z^{-1} & z^{-2} & ..... & z^{-(N-1)} \end{bmatrix} (\uparrow N). \tag{10b}$$

The following standard result shows effect of input and output lifting on the state space representation of a system.

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

$$G(z) := \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right] = \mathbf{D} + \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}, \tag{11}$$

with $\mathbf{A} \in \mathbb{R}^{N\times N}, \mathbf{B} \in \mathbb{R}^{N\times p}, \mathbf{C} \in \mathbb{R}^{m\times N}, \mathbf{D} \in \mathbb{R}^{m\times p}$ 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

$$\overline{G}(z) := \mathbf{L_2} G(z) \mathbf{L_2^{-1}} = \left[\begin{array}{c|cc} \mathbf{A}^2 & \mathbf{AB} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} & \mathbf{0} \\ \mathbf{CA} & \mathbf{CB} & \mathbf{D} \end{array}\right] \tag{12}$$

where $\mathbf{L_2}$ and $\mathbf{L_2^{-1}}$ can be obtained by using (10).

Proof See [13, Theorem 8.2.1]. □

The following results are

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

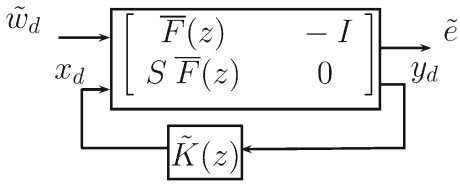

Fig. 8 Standard control unit consists of an open-loop transfer function, along with a feedback system $\tilde{K}(z)$ imposed for stability

$$K(z) = \begin{bmatrix} 1 & z^{-1} \end{bmatrix} \tilde{K}(z^2), \tag{13}$$

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

$$\mathbb{T}(z) = F(z) - \mathbf{L_2^{-1}} \tilde{K}(z)(\downarrow 2) F(z). \tag{14}$$

Thus, $F(z)$ and $\tilde{K}(z)$ are transfer functions at different sampling rates. Hence, lifting the input and output of $\mathbb{T}$ by a factor of 2 gives the lifted transfer function of $\mathbb{T}$. Therefore,

$$\begin{aligned} \overline{\mathbb{T}}(z) &= \mathbf{L_2} \mathbb{T}(z) \mathbf{L_2^{-1}} \\ &= \overline{F}(z) - \tilde{K}(z) S \overline{F}(z), \end{aligned} \tag{15}$$

with $S = \begin{bmatrix} 1 & 0 \end{bmatrix}$ and $\overline{F}(z) := \mathbf{L_2} F(z) \mathbf{L_2^{-1}}$. The lifting converts $\mathbb{T}$ into a single rate system $\overline{\mathbb{T}}$. Note that the norm is not altered after introducing the lifting, i.e., $||\mathbb{T}||_\infty = ||\overline{\mathbb{T}}||_\infty$ [13]. It implies that minimizing the H^∞-norm of system $\overline{\mathbb{T}}$ will automatically minimize the H^∞-norm of system $\mathbb{T}$. Further, the minimum H^∞ gain of $\overline{\mathbb{T}}$ is found by designing the filter $\tilde{K}(z)$. Equation (15) can be written in the form of a standard discrete control system as depicted in Fig. 8, which is an observer-based controller design problem in the control literature [13].

In Fig. 8, I is an identity matrix of 2×2, 0 is a vector of dimension 1×2, $\tilde{w}_d = \mathbf{L_2} w_d$, and $\tilde{e} = \mathbf{L_2} e$. Now, an optimal causal and stable filter $\tilde{K}_{\text{opt}}(z)$ is obtained using the robust control toolbox in MATLAB [14,19]. Finally, $K_{\text{opt}}(z)$ is computed from the filter $\tilde{K}_{\text{opt}}(z)$ by using (13).

Remark 1 For downsampling by a factor N, see [12,60].

A.2 Gaussian Mixture Model

A feature vector $\mathbf{Z} \in \mathbb{R}^{31}$ is formed by concatenating the narrowband feature $\mathbf{X}$ of dimension $\mathbb{R}^{10}$ and the corresponding wideband feature $\mathbf{Y_K}$ of dimension $\mathbb{R}^{21}$. The feature vector $\mathbf{Z}$ is modeled by the Gaussian mixture model (GMM) for obtaining the joint probability distribution function (pdf) of the narrowband feature $\mathbf{X}$ and wideband feature $\mathbf{Y_K}$ [15]. The pdf of $\mathbf{Z}$ is modeled by the summation of the weighted multivariate

Gaussian distributions as written

$$p(\mathbf{Z}|\lambda) = \sum_{i=1}^{M} w_i\, p(\mathbf{Z}|\boldsymbol{\mu}_{\mathbf{z}_i}, \Sigma_{\mathbf{zz}_i}), \tag{16}$$

with w_i being the contribution of the ith Gaussian distribution out of M clusters and $p(\mathbf{Z}|\mu_{\mathbf{z}_i}, \Sigma_{\mathbf{zz}_i})$ denotes the corresponding Gaussian pdf of Z. It is written as

$$p(\mathbf{Z}|\mu_{\mathbf{z}_i}, \Sigma_{\mathbf{zz}_i}) = \frac{1}{(2\pi)^{d/2}|\Sigma_{\mathbf{zz_i}}|^{1/2}} e^{-\frac{(\mathbf{Z}-\bar{\mathbf{z}}_\mathbf{i})^T \Sigma_{\mathbf{zz_i}}^{-1}(\mathbf{Z}-\bar{\mathbf{z}}_\mathbf{i})}{2}}, \tag{17}$$

with d dimension of feature vector $\mathbf{Z}$, and $\mu_{\mathbf{z}_i}$ and $\Sigma_{\mathbf{zz}_i}$ being the mean vector and covariance matrix of Gaussian pdf, respectively, and they are defined as

$$\mu_{\mathbf{z}_i} = \begin{bmatrix} \mu_{\mathbf{x}_i} \\ \mu_{\mathbf{y_k}_i} \end{bmatrix}, \tag{18}$$

$$\Sigma_{\mathbf{zz}_i} = \begin{bmatrix} \Sigma_{\mathbf{xx}_i} & \Sigma_{\mathbf{xy_k}_i} \\ \Sigma_{\mathbf{y_k x}_i} & \Sigma_{\mathbf{y_k y_k}_i} \end{bmatrix}, \tag{19}$$

where $\mu_{\mathbf{x}_i}$ and $\mu_{\mathbf{y_k}_i}$ are mean vectors of $\mathbf{X}$ and $\mathbf{Y_K}$, respectively. $\Sigma_{\mathbf{xx}_i}$ and $\Sigma_{\mathbf{y_k y_k}_i}$ are covariance matrices of $\mathbf{X}$ and $\mathbf{Y_K}$, respectively. $\Sigma_{\mathbf{xy_k}_i}$ and $\Sigma_{\mathbf{y_k x}_i}$ are cross-covariance matrices of $\mathbf{X}$ and $\mathbf{Y_K}$, respectively. For estimating parameters of the GMM, the expectation–maximization [15] algorithm is used that gives the maximum likelihood solutions, i.e., maximize the probability of generating the feature vectors from the model. This leads to a joint pdf of $\mathbf{X}$ and $\mathbf{Y_K}$.

In testing phase, the wideband feature vector is estimated using the joint pdf for a given narrowband feature vector $\tilde{\mathbf{X}}$. For this, a mapping function $f(\tilde{\mathbf{X}})$ is found by considering the minimum mean squared error (MMSE) criteria [29]. Mean squared error

$$\varepsilon_{mse} = E[||\mathbf{Y_K} - f(\tilde{\mathbf{X}})||^2], \tag{20}$$

is computed, where $\mathbf{Y_K}$ and $f(\tilde{\mathbf{X}})$ represent the original and corresponding estimated wideband feature for a given narrowband feature vector $\tilde{\mathbf{X}}$, respectively. To solve (20), Bayesian estimation theory is used that gives a mapping function. This mapping function is a conditional mean of $\tilde{\mathbf{Y}}_K$ given $\tilde{\mathbf{X}}$ and defined as [45]

$$f(\tilde{\mathbf{X}}) = E(\tilde{\mathbf{Y}}_K|\tilde{X}), \tag{21}$$

$$= \sum_{i=1}^{M} \alpha_i(\tilde{\mathbf{X}})[\mu_{\mathbf{y_k}_i} + \Sigma_{\mathbf{y_k x}_i} \Sigma_{\mathbf{xx}_i}^{-1}(\tilde{\mathbf{X}} - \mu_{\mathbf{x}_i})], \tag{22}$$

where

$$\alpha_i(\tilde{\mathbf{X}}) = \frac{w_i\, p(\tilde{\mathbf{X}}|\mu_{\mathbf{x}i},\, \Sigma_{\mathbf{xx}_i})}{\sum_{l=1}^{M} w_l\, p(\tilde{\mathbf{X}}|\mu_{\mathbf{x}_l},\, \Sigma_{\mathbf{xx}_l})}. \tag{23}$$

The weighting function $\alpha_i(\tilde{\mathbf{X}})$ is a posterior probability of ith component in the Gaussian mixture distribution from which, feature vector $\tilde{\mathbf{X}}$ is generated. $f(\tilde{X})$ is the mapping function, which maps the given narrowband feature vector $\tilde{\mathbf{X}}$ to $\tilde{\mathbf{Y}}_K$. E denotes the expectation. $p(\tilde{\mathbf{X}}|\mu_{\mathbf{x}i},\, \Sigma_{\mathbf{xx}_i})$ denotes the Gaussian pdf of $\tilde{X}$ corresponding to ith cluster. $\tilde{\mathbf{Y}}_K$ denotes the estimated wideband feature vector, which is used in the artificial bandwidth extension of speech signal.

A.3 Deep Neural Network

Deep neural network (DNN) is used to estimate the wideband feature vector $\tilde{\mathbf{Y}}_K$ for a given narrowband feature vector $\tilde{\mathbf{X}}$ [24]. DNN model has a variety of different parameters, such as activation functions, number of hidden layers, number of units in each hidden layer, learning rate, regularizations, optimizers, loss functions, and mini-batch size, which need to be checked empirically to design an optimal DNN model. A DNN feed-forward topology architecture is made up of N number of layers, consisting of $N-1$ hidden layers and one output layer. The output of the ith layer for sample index n is defined as

$$\mathbf{h_n^i} = f_i(\mathbf{W^i h_n^{i-1}} + \mathbf{b^i}), \quad 1 \le i \le N, \tag{24}$$

where $\mathbf{W^i}$ and $\mathbf{b^i}$ signify the weight and bias parameters, respectively. $f_i(.)$ represents the nonlinear activation function, and $\mathbf{h_n^i}$ is the output of ith layer. An output ($\mathbf{h_n^N}$) of the Nth layer yields the estimated wideband feature vector and an input ($\mathbf{h_n^0}$) to the first layer is the narrowband feature vector. In (24), the weight and bias are unknown parameters, which are initialized with some random value. Further, the mean squared error is considered as a loss function (α), which is minimized to obtain the optimal weight $\mathbf{W^i_{opt}}$ and bias $\mathbf{b^i_{opt}}$ values of each layer as defined

$$\alpha = \frac{1}{T}\sum_{n=1}^{T} ||\mathbf{h_n^N} - \mathbf{Y_K^n}||_2^2, \tag{25}$$

$$(\mathbf{W^i_{opt}}, \mathbf{b^i_{opt}}) = \arg\min_{\mathbf{W^i},\mathbf{b^i}}(\alpha), \tag{26}$$

with T being the mini-batch size, and $\mathbf{Y_K^n}$ denotes the original wideband feature vector.

References

1. ITU-T Software Tool Library 2009 User's Manual. ITU-T Recommendation G.191 (2009)
2. J. Abel, T. Fingscheidt, Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(1), 71–83 (2018)

3. J. Abel, M. Strake, T. Fingscheidt, A simple cepstral domain DNN approach to artificial speech bandwidth extension, in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (IEEE, 2018), pp. 5469–5473
4. N. Adiga, S. Prasanna, Detection of glottal activity using different attributes of source information. IEEE Signal Process. Lett. **22**(11), 2107–2111 (2015)
5. K. Aida-Zade, C. Ardil, S. Rustamov, Investigation of combined use of MFCC and LPC features in speech recognition systems. World Acad. Sci. Eng. Technol. **19**, 74–80 (2006)
6. B. Andersen, J. Dyreby, B. Jensen, F.H. Kjærskov, O.L. Mikkelsen, P.D. Nielsen, H. Zimmermann, Bandwidth expansion of narrow band speech using linear prediction. Web Source **26** (2015)
7. S. Ashida, M. Nagahara, Y. Yamamoto, Audio signal compression via sampled-data control theory, in *SICE 2003 Annual Conference (IEEE Cat. No. 03TH8734)*. (IEEE, 2003), vol. 2, pp. 1744–1747
8. P.B. Bachhav, M. Todisco, M. Mossi, C. Beaugeant, 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
9. P. Bauer, T. Fingscheidt, 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
10. P. Bauer, 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
11. L. Bin, T. Jianhua, W. Zhengqi, L. Ya, D. Bukhari, et al. A novel method of artificial bandwidth extension using deep architecture (2015)
12. T. Chen, B.A. Francis, Design of multirate filter banks by H^∞ optimization. IEEE Trans. Signal Process. **43**(12), 2822–2830 (1995)
13. T. Chen, B.A. Francis, *Optimal Sampled-data Control Systems*, vol. 124 (Springer, Berlin, 1995)
14. R.Y. Chiang, M.G. Safonov, MATLAB: Robust Control Toolbox User's Guide. Math Works (1997)
15. M.B. Christopher, *Pattern Recognition and Machine Learning* (Springer, New York, 2016)
16. R. Crochiere, A weighted overlap-add method of short-time Fourier analysis/synthesis. IEEE Trans. Acoust. Speech Signal Process. **28**(1), 99–102 (1980)
17. N. Enbom, 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
18. J.A. Fuemmeler, R.C. Hardie, W.R. Gardner, Techniques for the regeneration of wideband speech from narrowband speech. EURASIP J. Appl. Signal Process. **2001**(1), 266–274 (2001)
19. K. Glover, J.C. Doyle, State-space formulae for all stabilizing controllers that satisfy an H^∞-norm bound and relations to relations to risk sensitivity. Syst. Control Lett. **11**(3), 167–172 (1988)
20. D. Gupta, H.S. Shekhawat, Artificial bandwidth extension using h^∞ optimization. Proc. Interspeech **2019**, 3421–3425 (2019)
21. D. Gupta, H.S. Shekhawat, Artificial bandwidth extension using H^∞ optimization and speech production model, *Presented at the 29th International Conference Radioelektronika (RADIOELEKTRONIKA), Pardubice, Czech Republic* (April 16–18, 2019)
22. Y. Hu, P.C. Loizou, Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. **16**(1), 229–238 (2008)
23. A. Hurmalainen, J.F. Gemmeke, 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
24. I. Goodfellow, Y. Bengio, *Deep Learning* (MIT Press, Cambridge, 2016)
25. F. Itakula, Line spectrum representation of linear predictive coefficients of speech signal. *J. Acoust. Soc. Am.* (1975)
26. ITU-T, R.P.: 862.1: Mapping function for transforming P. 862 raw result scores to MOS-LQO, *International Telecommunication Union, Geneva, Switzerland* (2003)
27. P. Jax, 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*. (IEEE, 2003), vol. 1, pp. I–I
28. P. Jax, P. Vary, On artificial bandwidth extension of telephone speech. Signal Process. **83**(8), 1707–1719 (2003)

29. A. Kain, M.W. Macon, Spectral voice conversion for text-to-speech synthesis, in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*. (IEEE, 1998), vol. 1, pp. 285–288
30. K.T. Kim, M.K. Lee, H.G. Kang, Speech bandwidth extension using temporal envelope modeling. IEEE Signal Process. Lett. **15**, 429–432 (2008)
31. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
32. U. Kornagel, Techniques for artificial bandwidth extension of telephone speech. Signal Process. **86**(6), 1296–1306 (2006)
33. A. Larcher, K.A. Lee, B. Ma, H. Li, Text-dependent speaker verification: classifiers, databases and rsr2015. Speech Commun. **60**, 56–77 (2014)
34. K. Li, C.H. Lee, A deep neural network approach to speech bandwidth expansion, in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (IEEE, 2015), pp. 4395–4399
35. Y. Li, S. Kang, Artificial bandwidth extension using deep neural network-based spectral envelope estimation and enhanced excitation estimation. IET Signal Proc. **10**(4), 422–427 (2016)
36. P.C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd edn. (CRC Press, Boca Raton, 2007)
37. J. Makhoul, M. Berouti, High-frequency regeneration in speech coding systems, in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, Cambridge, UK*. (IEEE, 1979), vol. 4, pp. 428–431
38. D. Marelli, P. Balazs, On pole-zero model estimation methods minimizing a logarithmic criterion for speech analysis. IEEE Trans. Audio Speech Lang. Process. **18**(2), 237–248 (2010)
39. J.D. Markel Jr., A. Gray, *Linear Prediction of Speech. Communication and Cybernetics 12*, 1st edn. (Springer, Berlin, 1976)
40. MathWorks, M.: the mathworks. Inc., Natick (1992)
41. J. Meyer, *Acoustics and the Performance of Music: Manual for Acousticians, Audio Engineers, Musicians, Architects and Musical Instrument Makers* (Springer, Berlin, 2009)
42. K.S.R. Murty, B. Yegnanarayana, M.A. Joseph, Characterization of glottal activity from speech signals. IEEE Signal Process. Lett. **16**(6), 469–472 (2009)
43. P. Nizampatnam, K.K. Tappeta, Bandwidth extension of narrowband speech using integer wavelet transform. IET Signal Proc. **11**(4), 437–445 (2016)
44. A.H. Nour-Eldin, P. Kabal, Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech, in *Proceedings 9th Annual Conference of the International Speech Communication Association* (2008)
45. H.V. Poor, *An Introduction to Signal Detection and Estimation* (Springer, Berlin, 2013)
46. N. Prasad, T.K. Kumar, Bandwidth extension of speech signals: a comprehensive review. Int. J. Intell. Syst. Appl. **8**(2), 45 (2016)
47. Y. Qian, P. Kabal, Dual-mode wideband speech recovery from narrowband speech, in *8th European Conference on Speech Communication and Technology, GENEVA, Switzerland* (2003), pp. 1433–1436
48. I. Rec, P. 800: methods for subjective determination of transmission quality, *International Telecommunication Union, Geneva* (1996), p. 22
49. A.W. Rix, J.G. Beerends, M.P. Hollier, A.P. Hekstra, Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2001), vol. 2, pp. 749–752
50. J. Sadasivan, S. Mukherjee, 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
51. U. Shaked, Y. Theodor, H^{∞} optimal estimation: a tutorial, in *Proceedings 31st IEEE Conference on Decision and Control*. (IEEE, 1992), pp. 2278–2286
52. X. Shao, Robust Algorithms for Speech Reconstruction on Mobile Devices. Ph.D. thesis, University of East Anglia (2005)
53. Y. Sunil, 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
54. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136 (2011)

55. P.P. Vaidyanathan, *Multirate Systems and Filter Banks* Prentice-Hall Signal Processing Series. (Prentice Hall, Hoboken, 1993)
56. S. Vaseghi, E. Zavarehei, 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)*. (IEEE, 2006), vol. 3, pp. III–844–III–847
57. W. Verhelst, Overlap-add methods for time-scaling of speech. Speech Commun. **30**(4), 207–221 (2000)
58. Y. Wang, S. Zhao, W. Liu, M. Li, J. Kuang, Speech bandwidth expansion based on deep neural networks, in *Proceedings 16th Annual Conference of the International Speech Communication Association* (2015)
59. C. YağLı, M.T. Turan, E. Erzin, Artificial bandwidth extension of spectral envelope along a Viterbi path. Speech Commun. **55**(1), 111–118 (2013)
60. Y. Yamamoto, M. Nagahara, P.P. Khargonekar, Signal reconstruction via H^∞ sampled-data control theory beyond the Shannon paradigm. IEEE Trans. Signal Process. **60**(2), 613–625 (2012)
61. V. Zue, S. Seneff, J. Glass, Speech database development at MIT: TIMIT and beyond. Speech Commun. **9**(4), 351–356 (1990)