Turkish Journal of Electrical Engineering and Computer Sciences

Volume 23 | Number 2 Article 8

1-1-2015

# Bandwidth extension of narrowband speech in log spectra domain using neural network

SARA POURMOHAMMADI

MANSOUR VALI

MOHSEN GHADYANI

Follow this and additional works at: https://journals.tubitak.gov.tr/elektrik

 Part of the Computer Engineering Commons, Computer Sciences Commons, and the Electrical and Computer Engineering Commons

## Recommended Citation

POURMOHAMMADI, SARA; VALI, MANSOUR; and GHADYANI, MOHSEN (2015) "Bandwidth extension of narrowband speech in log spectra domain using neural network," *Turkish Journal of Electrical Engineering and Computer Sciences*: Vol. 23: No. 2, Article 8. https://doi.org/10.3906/elk-1212-109
Available at: https://journals.tubitak.gov.tr/elektrik/vol23/iss2/8

This Article is brought to you for free and open access by TÜBİTAK Academic Journals. It has been accepted for inclusion in Turkish Journal of Electrical Engineering and Computer Sciences by an authorized editor of TÜBİTAK Academic Journals. For more information, please contact academic.publications@tubitak.gov.tr.

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2015) 23: 433 – 446
© TÜBİTAK
doi:10.3906/elk-1212-109

# Bandwidth extension of narrowband speech in log spectra domain using neural network

**Sara POURMOHAMMADI**[1,*], **Mansour VALI**[2], **Mohsen GHADYANI**[1]
[1]Faculty of Electronic Engineering, University of Shahed, Tehran, Iran
[2]Department of Electrical and Computer Engineering, K.N. Toosi University of Technology, Tehran, Iran

**Received:** 20.12.2012 • **Accepted:** 25.03.2013 • **Published Online:** 23.02.2015 • **Printed:** 20.03.2015

**Abstract:**In recent years, there have been significant advances in communication technology, but speech signals still suffer from low perceived quality caused by bandwidth limitations of telephone networks. The bandwidth extension (BWE) approach adds high-frequency components of the speech signal to band-limited telephone speech and increases speech perception significantly. In this work, we develop a new method for representation of vocal tract filter coefficients using log of filter bank energy (LFBE) parameters as an alternative for mel-frequency cepstral coefficients (MFCCs). This approach is based on a strong correlation between the spectral components of low- and high-band spectrums. Furthermore, the performances of Gaussian mixture model and multilayer perceptron neural network methods for estimation of the high-frequency envelope are evaluated. Objective evaluations of the obtained results indicate that the LFBE feature vectors have better performance than the MFCCs. In addition, findings of the objective evaluations showed that using a neural network, which is not common in BWE, achieves a better performance as compared to the Gaussian mixture model.

**Key words:** Bandwidth extension, log spectra domain, narrowband speech, neural network, wideband speech

## 1. Introduction

The bandwidth of the speech signal produced by humans has a frequency range of 0 to 10 KHz. In this range, quality of speech and its perception is very high. In some conditions, however, transmission of such speech signals may lead to relatively band-limited signals. For instance, almost all of the public telephone exchanges are digital, but the existing telephone network transmission bandwidth is still limited to the frequency range of 300–3400 Hz [1]. However, previous studies have shown that acoustic bandwidth reduces the quality of perceived speech dramatically [2].

Bandwidth extension techniques improve speech quality by adding the missed spectral components into the narrowband signal. Most bandwidth extension (BWE) algorithms are based on a human speech production model that is called the source-filter model. The main procedure of the BWE technique can be divided into 2 separate tasks: expansion of the excitation and expansion of the spectral envelope [3]. A block diagram of this procedure is depicted in Figure 1.

The expansion of the spectral envelope is a more challenging task and strictly depends on the features that estimate the spectral envelope. In the BWE procedure, spectral envelope information is usually represented as a set of cepstral coefficients [2], linear predictive coding (LPC) coefficients [4], line spectral frequency coefficients

*Correspondence: s_pourmohammadi@yahoo.com

[5–7], mel-frequency cepstral coefficients (MFCCs) [8–10], a set of autocorrelation coefficients [11], or mel-spectrum coefficients [12].

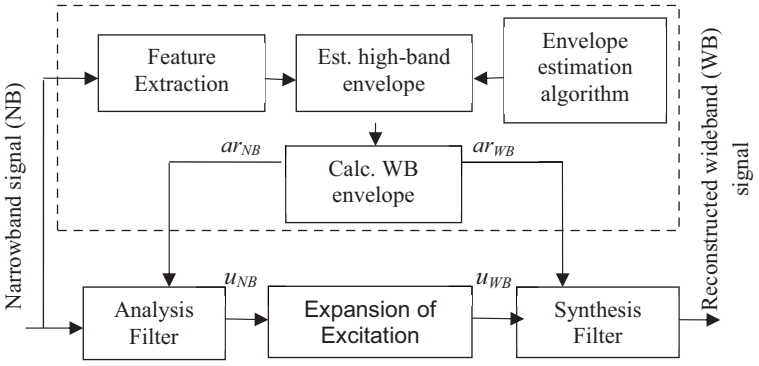


**Figure 1.** Block diagram of BWE algorithm.

The codebook technique is the fundamental envelope prediction method. A codebook of the BWE method contains a predefined set of narrowbands and their corresponding high-band envelopes. The spectral envelope information of the narrowband frame is compared to all codebook entries, and the candidate with the best matching is selected [13].

Using the hidden Markov model (HMM) in the field of speech recognition is common and it is used in the expansion of spectral envelopes. The HMM is able to model hidden information, e.g., how a speech sequence evolves over time. Therefore, it utilizes information about previous frames to estimate the high-band components [14].

The multilayer perceptron (MLP) feedforward neural network is used to reconstruct the wideband spectral features, too [2]. Feature vectors are derived from narrowband speech as like the corresponding wideband speech template. These vectors are used as input-output pairs to train a neural network model. The ultimate goal is mapping of a narrowband input signal to its corresponding wideband output.

The Gaussian mixture model (GMM) is able to model the probability density function of data. The GMM has been utilized to estimate a wideband spectral envelope from narrowband features. The GMM is trained using the expectation-maximization (EM) algorithm. The estimator then minimizes the mean squared error between the estimated wideband feature and the real wideband one [7,15–17]. In [18], a wideband excitation was generated by spectral folding from the narrowband linear prediction residual. The high-band of this signal is divided into 4 subbands with a filter bank, and a neural network is used to weight the subbands based on features calculated from the narrowband speech.

The correlation characteristics between the spectral components in the narrowband and the high-band signals, using several preliminary experiments, were investigated. The experiments demonstrated that 2 parts in the narrowband signal are mainly correlated with missing components in the high-band signal. These 2 parts are the area of the first formant (F1) and the boundary of the cutoff frequency. In addition, the corresponding experiments demonstrated that a particular spectral component is highly correlated not only with the spectral information around the first formant frequency, but also with the adjacent components. Figure 2 shows the 50 most highly correlated spectral components in the available frequency band with a particular mel-filter bank index in the cutoff frequency region. For example, the first plot in Figure 2 shows the top 50 spectral components (i.e. mel-filter bank outputs) in the available region, which are highly correlated with the 14th Mel filter banks index. The 14th index is the first component in the missing high-band region [19].

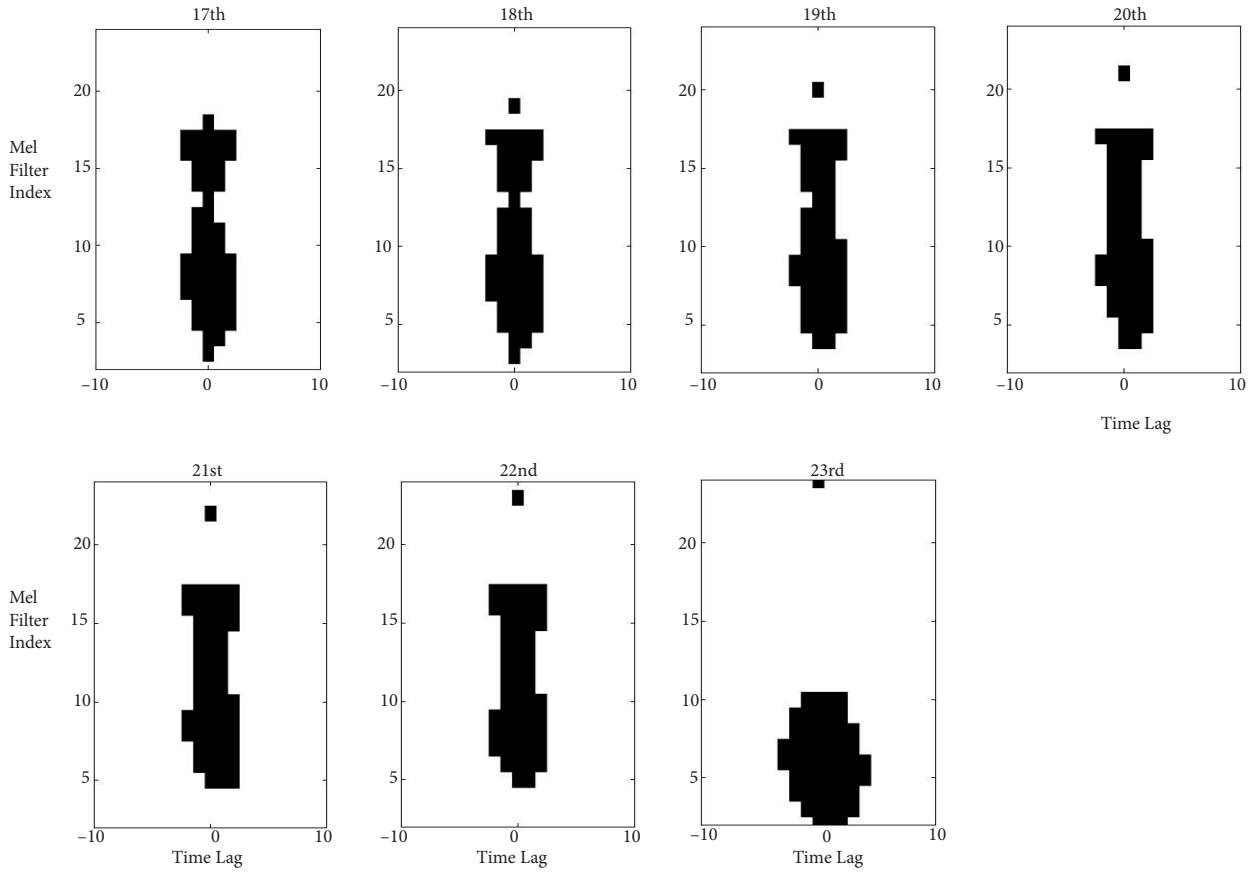


**Figure 2.** Top 50 highly correlated spectral components for narrowband speech [19].

Based on the above research, there are 2 innovative points in our present work: first, choosing feature vectors such as the log of filter bank energy (LFBE), and second, using a neural network for more accurate results. It is also easier to investigate the effect of using several frames to reconstruct wideband speech in neural network. In this study, both neural network and GMM techniques are applied to estimate wideband vocal tract filter coefficients from narrowband LFBE and MFCC vectors to compare their efficiency.

This paper is organized as follows: Section 2 describes the proposed procedure in order to extract the spectral features for mapping as well as for constructing the synthesis filter. MFCC features are described, too. The reconstruction of the spectral envelope using the GMM and the nonlinear mapping property of MLP neural networks is discussed in Section 3. Section 4 discusses reproducing wideband speech and Section 5 describes the performance of the proposed method.

## 2. MFCC and LFBE feature extraction

The new application of the well-known LFBE parameterization of speech for the narrowband (0–4 KHz) and high-band (4–8 kHz) speech signals (obtained by filtering the wideband speech) is summarized as follows:

1. Preemphasizing the signal with a high-pass filter.
2. Windowing the preemphasized signal with hamming window to minimize the edge effect of discontinuities because of framing. A 20-ms frame size with 50% overlap is used.

3. Applying the fast Fourier transform (FFT) to each frame followed by a magnitude operation to make a magnitude spectrum.
4. Applying mel-scale triangular filters to the magnitude spectrum. Twelve filters are used for the 0–4 kHz narrowband signal and 4 filters for the 4–8 kHz high-band signal.
5. Calculation of the logarithm of signal energies in the filter bank.

Following these steps, one LFBE feature vector for each frame was extracted. Then, using discrete cosine transform (DCT), MFCC features were obtained. Figure 3 depicts the block diagram of the extraction of both MFCC and LFBE features [20].

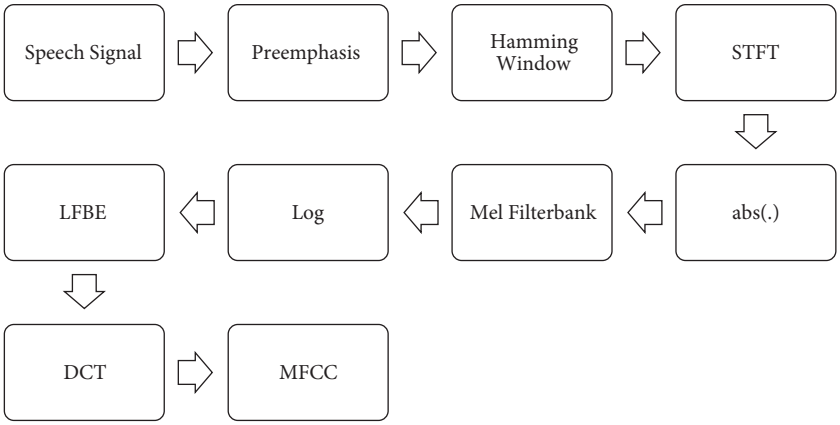


**Figure 3.** Block diagram of LFBE/MFCC extraction from speech signal.

## 3. Expansion of envelope

The main step in the BWE process is the expansion of the envelope. The block diagram of this process is shown in the upper part of Figure1. The estimation of the envelope is usually considered as a more challenging task than the estimation of the excitation. Previous studies proposed several envelope prediction methods such as codebook, GMM, HMM, and finally neural networks. The present study employed the GMM and neural network methods.

### 3.1. The GMM method in envelope extension

Two feature vectors, x and y, can be extracted for each frame of the available speech to form several observations, which will be utilized in the training step. x and y denote the narrowband and corresponding high-band feature vectors, respectively, and we define feature vector z such that $[\mathrm{x\ y}]^T$.

The density function of z can be modeled by the mixture of M Gaussian densities as follows in Eq. (1):

$$f(z) = \sum_{m=1}^{M} c_m g(x|\mu_m, \sum_m), \tag{1}$$

where $c_m$, $\mu_m$, and $\Sigma_m$ are the weighting coefficient, the mean vector, and the covariance matrix of the $m$th Gaussian, respectively. Since high cross-correlation between low- and high-band frequency components is assumed, the covariance matrix can be chosen as a full matrix. No analytical solution exists for finding the parameters of the model. Therefore, the parameters should be estimated by the EM algorithm [15].

### 3.1.1. EM algorithm

The EM algorithm is a widely used method for estimation of parameters of a GMM, given a set of observations. It maximizes the probability of a certain set of observations generated from a distribution with a given set of parameters. This is done by adjusting the parameters so that the likelihood for these parameters is maximized. The EM algorithm performs this estimation iteratively, and it guarantees an increase in likelihood. The objective is to maximize the likelihood by adjusting the parameters, and then the problem is solved as follows in Eq. (2):

$$\theta* = \arg\max \log L(z|\theta) \quad = \arg\max \sum_{t=1}^{T} \log[\sum\nolimits_{m=1}^{M} c_m \, g(z|\mu_m, \sum_{m})], \tag{2}$$

where $z$ represents the observed data, $\theta$ is a set of parameters $\theta = \{\mu, \Sigma\}$ , $L(z|\theta)$ is the likelihood function for the observations, $\theta*$is the set of optimum parameters, and T is the number of frames.

The procedure is divided into 2 parts: the expectation and the maximization steps, which make the name of the algorithm. The expectation step (E-step) is performed by calculating the posterior probability based on observations z and the multivariate Gaussian distributions. The E-step is calculated for all observations and all mixture components. The posterior probability is utilized during the maximization step and the parameters for each mixture are updated in the maximization step (M-step).

The E-step and M-step are conducted iteratively followed by each other until the algorithm reaches convergence. Convergence is achieved when the absolute increase in log-likelihood between 2 iterations is below the threshold. In this case, the algorithm stops and the final parameters $\theta* = \{\mu_z, \Sigma_z\}$ are obtained. Each iteration increases the log-likelihood and it is ensured that the algorithm converges to a local maximum of the log-likelihood function [15]. Using the obtained parameters from the trained GMM and minimum mean square error (MMSE) estimator, we can estimate wideband coefficients.

### 3.1.2. Estimation of wideband signal coefficients

The MMSE estimator minimizes the mean squared error between the estimated and real wideband features of ŷ and y. Because we use the full covariance matrix, the equation can be presented as follows:

$$\begin{matrix} \hat{y}_{MMSE} = \sum\limits_{m=0}^{M} P(m|x, \Theta) \\ [\mu_m^y + \sum\nolimits_m^{yx} (\sum\nolimits_m^{xx})^{-1} (x - \mu_m^x)] \end{matrix}, \tag{3}$$

where $P(m|x, \Theta)$ is a weighting function for each component $m$ as defined in Eq. (4).

$$P(m|x, \Theta) = \frac{c_m g(x|\mu_m^x, \sum_m^{xx})}{\sum\limits_{n=1}^{M} c_n g(x|\mu_n^x, \sum_n^{xx})}, \tag{4}$$

where $\mu_m^x$ and $\mu_m^y$ are parts of the mean vector, and $\sum_m^{xx}$ and $\sum_m^{yx}$ are parts of the covariance matrix for the $m$th component. These terms arise from decomposition of $\mu_m^z$ and $\sum_m^z$ by Eqs. (5) and (6) [15].

$$\mu_m^z = \begin{bmatrix} \mu_m^x \\ \mu_m^y \end{bmatrix} \tag{5}$$

$$\sum_{m}^{z} = \begin{bmatrix} \sum_{m}^{xx} & \sum_{m}^{xy} \\ \sum_{m}^{yx} & \sum_{m}^{yy} \end{bmatrix} \tag{6}$$

### 3.2. Envelope extension with neural network

Using neural networks is common in the field of speech recognition, but not in BWE. This paper exploits a nonlinear neural network to achieve a mapping of narrowband features to wideband spectral features. According to [10], there is high correlation among the current frame and adjacent frames with the upper part of spectrum. Therefore, the use of a time-delay neural network (TDNN) is suggested. The main property of this method is working on continuous data. The TDNN structure should possess several adjacent feature vectors that have been recently used to prepare a single input for a neural network [21].

To use neural network as a mapping function, 2 hidden layers are sufficient. In this case, hidden layer activation functions are nonlinear and activation functions of input and output layers are linear. If the output of the neural network is normalized, one hidden layer is sufficient, and in this case, the output activation function is nonlinear.

The selection of number of neurons is based on the fact that the number of inputs is 4 to 10 times more than the neural network's unknown weights. In this case, the unknown parameters are proportionate to known parameters and weights may be estimated accurately.

## 4. Reconstruction of wideband speech

After estimation of the MFCC or LFBE coefficients with the GMM or neural network methods, it is necessary to convert these obtained feature vectors to vocal tract filter coefficients for the ability to use the linear prediction model for evaluating and reconstructing wideband speech.

Unfortunately, some steps of MFCC generation process are noninvertible. Therefore, some of the useful information of the speech signal will be lost. There is still a fairly broad range of differing estimates that show a possibility of logically estimating the speech power spectrum [22]. Calculation of vocal tract coefficients from MFCC feature vectors is a 2-part process:

Part one: Recovering power spectrum from MFCC feature vectors.

Part two: Estimating linear prediction model coefficients from the power spectrum.

The first step of the above process is to use the indirect DCT as in Eq. (7).

$$\log \hat{Y}_k = \sqrt{\tfrac{2}{N}} \sum_{n=0}^{N-1} c_n \cos\left(\tfrac{(2k+1)n\pi}{2N}\right) \quad , \quad 0 < k < N-1 \tag{7}$$

where $\log \hat{Y}_k$, $c_n$,and $N$ are the $k$th estimated power spectrum, the $n$th MFCC feature vector, and the number of filter banks, respectively.

An exponential operator is the easiest way to invert the logarithm operation. Power spectrum estimation is the next step. Since an inversing process was performed with a limited number of feature vectors, only a limited number of mel-scale filters will be available as a result. Therefore, for reconstructing the spectrum at high quality, interpolation between energies of filter banks parameters (LFBEs) is necessary.

Interpolation is done with high-resolution inverse cosine transform as in Eq. (8).

$$\log \hat{Y}_{k'} = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} c_n \cos\left(\frac{(2k'+1)n\pi}{2iN}\right) \quad 0 < k' < iN-1 \tag{8}$$

Here, $i$ is the interpolation factor and refers to the number of filter banks. As a result, the number of log-energies will be $iN$.

This interpolation results in mel-scales with very accurate resolution that can be used to estimate separate frequency bands with linear (not mel) spacing.

Therefore, by use of the DCT as an interpolation function, the interpolation between the centers of mel-frequency bands is performed. The interpolation factor will be determined by the desired resolution of the mel scale. Using Eq. (9), the linear frequency scale becomes the mel-frequency scale.

$$f_{mel} = 2595 \log_{10}(1 + f_{HZ}/700) \tag{9}$$

Here, $f_{mel}$ and $f_{Hz}$ are the frequency in mel and hertz domains, respectively. The resolution factor is then calculated using Eq. (10).

$$i = \left[\frac{f_{mel2} - f_{mel1}}{N+1}\right] \tag{10}$$

Here, $f_{mel1}$ and $f_{mel2}$ are the start and end of bandwidth of signal in mel spacing. The next step is converting the power spectrum to the predictive linear model coefficients. Computing the inverse Fourier transform of the 2-sided power spectrum results in the autocorrelation coefficients. This can then be used to solve the Yule–Walker equations by means of Levinson–Durbin recursion. Thus, linear prediction model coefficients are obtained by minimizing the forward predictor mean square error. These LPC parameters represent the coefficients of the all-pole vocal tract filter [23,24].

After achieving vocal tract filter coefficients, an excitation signal is obtained using a narrowband speech signal through the analyzing filter (Figure 1). This signal can be used simply by spectral folding method to convert it into a wideband signal. Because of low computation and good results in implementation, this is one of the conventional methods in extension of excitation. In spectral folding, the high band is generated by up-sampling the signal. As a result, the high-band spectrum is a mirror image of the original narrowband spectrum [1]. Then, according to Figure 1, by estimating wideband LPC coefficients, a synthesis filter will be designed. Applying this filter to the excitation signal and also using the overlap-add method, a wideband speech signal is reconstructed.

After reconstructing the wideband signal, it is filtered by a high-pass filter to obtain the missing high-band speech signal. Since the narrowband signal is available, it can already be used as output. As a result, the sampling rate of the narrowband signal increases by interpolation and then the high-band and narrowband signals can be summed together to reconstruct the wideband signal.

## 5. Implementation

First, the database that is used in this work is introduced. An introduction of objective measures that are employed for the evaluation of the results is then presented and, finally, implementation results are compared in various conditions.

### 5.1. Introduction of database

To evaluate the proposed algorithm, a speech database with appropriate training and testing sets is required. In this paper, the TIMIT database was used, in which all audio files are sampled at 16 kHz [25]. The division provided for training and testing is 73% and 27%, respectively. A set of the employed training data consists of 2064 sentences from 258 different speakers. In comparison, the test database contains 760 sentences from 95 speakers. For all audio files (training and testing), the narrowband signal is in the range of 0 to 3400 Hz. Extraction of feature vectors MFCC and LFBE was done from all of them with frame length of 20 ms and 50% overlap. Furthermore, the total number of mel-scale filter bank and coefficients was equal. The number of coefficients used for narrowband LFBEs is 12 plus 4 high-band coefficients. This means that the wideband speech signal is represented with 16 coefficients as a feature vector. The number of MFCC feature vectors for the wideband signal is 16, too. Derivatives of the MFCC coefficients were not used in this research.

### 5.2. Evaluation methods

There are different methods to measure the difference between the original and estimated wideband envelopes. The log spectral distortion (LSD) is the most commonly used measure for evaluating the bandwidth expansion algorithm [1]. This is defined in Eq. (11).

$$d_{LSD} = \sqrt{\frac{1}{2\pi}\int_0^{\pi}(10\ \log_{10}(A(w)) - 10\ \log_{\ 10}(\hat{A}(w)))^2 dw}, \tag{11}$$

where $A(w)$ and $\hat{A}(w)$are the original and estimated wideband envelopes, respectively.

Even though the Itakura distance is not a real measure, since it is not symmetric, it is widely used as a similarity measure between vocal tract filter coefficients. The Itakura distance is heavily influenced by spectral dissimilarity because of the presence of mismatch in formant locations, which is desirable since the auditory system is sensitive to these errors. The idea is to measure the log of the ratio between the total energy of the residual signal for 2 sets of the vocal tract filter coefficients [26]. This is defined in Eq. (12).

$$d_{IS} = \frac{1}{2\pi}\int_0^{\pi}(\frac{A(w)}{\hat{A}(w)} - \log_{10}\frac{A(w)}{\hat{A}(w)} - 1)dw \tag{12}$$

### 5.3. Implementation of the GMM method

The MMSE approach was used to estimate the wideband feature vectors based on the trained GMM parameters. However, first it is necessary that the matrix of the GMM training data be built. After feature extraction, the training matrix $z = [xy]^T$ is built, where x and y denote the narrowband and high or wideband feature vectors, respectively. Common values?? for the total number of mixtures are the integer powers of 2 (16, 32, 64, 128, and 256). In this study, the results of the Gaussian densities per number are investigated for 8, 16, 32, 64, 128, and 256.

Furthermore, a full covariance matrix is used to estimate the parameters. In this case, the input feature vector for the MMSE estimator is a 12-dimension narrowband and its output feature vectors are 16-dimensional for wideband MFCC and 4-dimensional for high-band LFBE. Table 1 reports the results of the bandwidth expansion algorithm that used the GMM model to generate the MFCC wideband spectral envelope. Amounts of Itakura–Saito distance and LSD are shown in terms of RMS and dB in all tables.

**Table 1.** Objective measures?? for the MFCC feature vectors estimated by GMM.

| Number of Gaussian distributions | RMS LSD (dB) | RMS IS (dB) |
|---|---|---|
| 16 | 5.22 | 0.69 |
| 32 | 5.15 | 0.65 |
| 64 | 5.13 | 0.63 |
| 128 | 4.48 | 0.57 |
| 256 | 4.66 | 0.59 |

Table 2 is similar to the previous table that used GMM to produce a high-band spectral envelope, but results for evaluation of LFBE feature vectors are given.

**Table 2.** Objective measures ??for the LFBE feature vectors estimated by GMM.

| Number of Gaussian distributions | RMS LSD (dB) | RMS IS (dB) |
|---|---|---|
| 16 | 4.88 | 0.67 |
| 32 | 4.80 | 0.58 |
| 64 | 4.65 | 0.56 |
| 128 | 4.45 | 0.55 |
| 256 | 4.60 | 0.57 |

Figures 4 and 5 show the results of Tables 1 and 2, analyzing each of the objective measurements, RMS LSD, and RMS IS for MFCC and LFBE feature vectors.

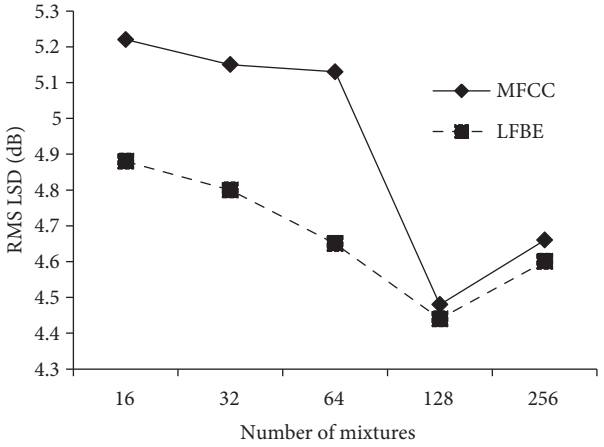


**Figure 4.** Analyzed RMS LSD for MFCC and LFBE feature vectors.

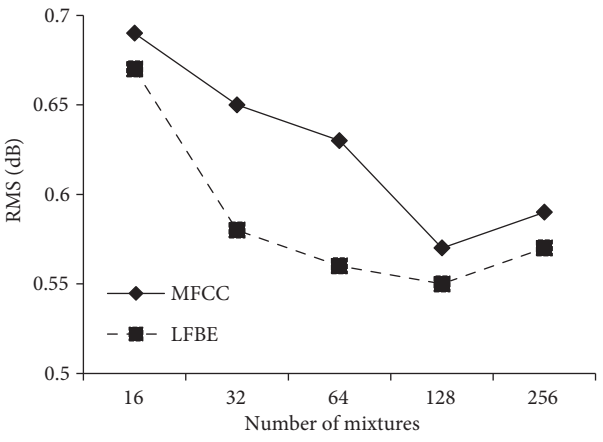


**Figure 5.** Analyzed RMS IS for MFCC and LFBE feature vectors.

### 5.4. Implementation of neural network method

There was significant correlation between the narrow and high-band frequency components [19]. The correlation was not only between the specific frame and high band, but also between the adjacent frames and high band of the current frame. Therefore, the use of the TDNN is recommended. For the preparation of network input, the use of 3 adjacent frames (one frame to the right and another to the left of the current frame) is proposed.

It is important to note that, in this study, a neural network was used as the mapping tool of input to output. The narrowband input is 12-dimensional for MFCC and LFBE, and thus because of using 3 adjacent frames, the number of input neurons will be 36. The number of output neurons for the LFBE is 4 (the number of high-band filters) and for MFCC is 16 (the number of MFCC coefficients in wideband).

After training the network with the scaled conjugate gradient method, the process is continued by testing feature vectors derived from the test database. Tables 3 and 4 evaluate the RMS IS and RMS LSD measures in the case of a neural network used to estimate the spectral envelope. Table 3 shows results for MFCC features and Table 4 corresponds to the LFBE feature vectors.

**Table 3.** Objective measures ??for the MFCC feature vectors estimated by neural network.

| Neural network condition | RMS LSD (dB) | RMS IS (dB) |
|---|---|---|
| Nonnormalized | 4.69 | 0.5 |
| Normalized input | 2.02 | 0.27 |
| Normalized input and output | 3.12 | 0.39 |

**Table 4.** Objective measures ??for the LFBE feature vectors estimated by neural network.

| Neural network condition | RMS LSD (dB) | RMS IS (dB) |
|---|---|---|
| Nonnormalized | 1.94 | 0.26 |
| Normalized input | 1.58 | 0.21 |
| Normalized input and output | 1.87 | 0.25 |

The discussed network was tested in terms of input and output for the best result. A nonnormalized state means that the input and output of the neural network is not normalized. In this state, the network has 2 hidden layers with tangent hyperbolic nonlinear activation function and the numbers of neurons in each layer are 40 and 15, respectively. The activation function of the output layer is linear. Normalized input mode means that normalized features apply to the network, but its output is not normalized. In this condition, the network has 2 hidden layers with nonlinear activation function, with 40 and 15 neurons in the first and second hidden layer, respectively. Furthermore, in this case, the activation function is linear at the output layer.

The normalization method for the training matrix is proportional to its mean and standard deviation. Therefore, from all of the feature vectors, the mean vector can be subtracted and then divided by the standard deviation. There is a single hidden layer of the neural network with nonlinear activation function in the normalized input and output state. In this case, the number of neurons is 60 and the output layer activation function is nonlinear. All the nonlinear functions used are tangent hyperbolic.

In Figures 6 and 7, the results of neural network implementation using MFCC and LFBE feature vectors are illustrated.

In this study, to assess the effect of the number of adjacent frames in the optimum parameter estimation of the current high-band spectral envelope accurately, 2 more cases were evaluated: 1) 5 consecutive frames (2 frames from the right and 2 frames from the left of the current frame) and 2) 1 frame (current frame only). For both of the new cases the optimum state of the neural network (normalized input and nonnormalized output) has been experienced.

This study was performed for both MFCC and LFBE feature vectors. Table 5 shows results of the implementation of only 1 input frame and Table 6 shows the results of implementation of 5 consecutive frames in the neural network input. Table 7 compares 3 input frame conditions for the input networks and the RMS LSD measure.

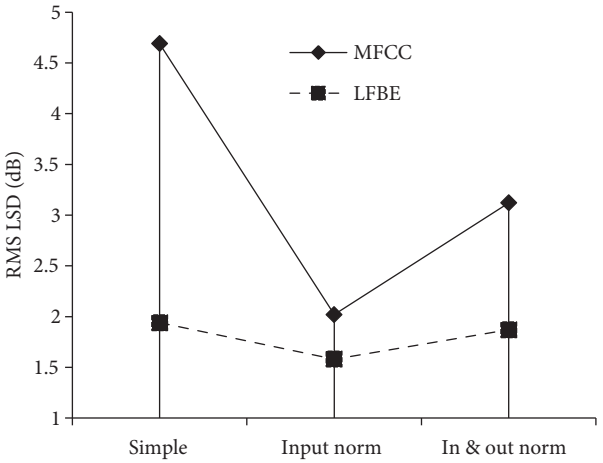


**Figure 6.** Analyzed RMS LSD for MFCC and LFBE feature vectors.

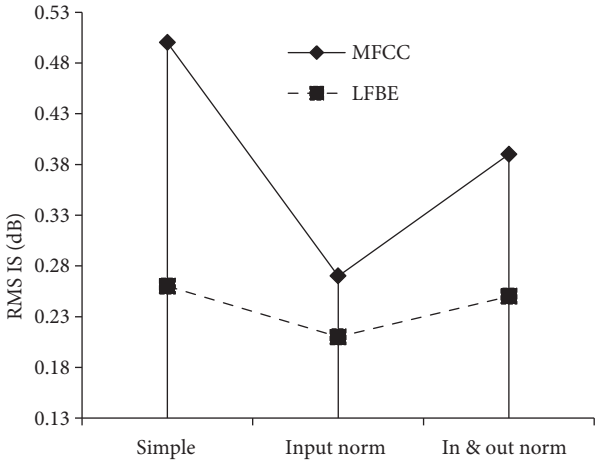


**Figure 7.** Analyzed RMS IS for MFCC and LFBE feature vectors.

**Table 5.** Results of the implementation of only 1 input frame in neural network input.

| Neural network with 1 input frame | RMS LSD (dB) | RMS IS (dB) |
|---|---|---|
| MFCC feature vectors | 2.21 | 0.33 |
| LFBE feature vectors | 1.65 | 0.22 |

**Table 6.** Results of implementation of 5 consecutive frames in neural network input.

| Neural network with 5 consecutive input frames | RMS LSD (dB) | RMS IS (dB) |
|---|---|---|
| MFCC feature vectors | 1.95 | 0.27 |
| LFBE feature vectors | 1.59 | 0.22 |

**Table 7.** Comparison of 3 input frame conditions for the input networks and the RMS LSD measure.

| Neural network input conditions | RMS LSD (MFCC) | RMS LSD (LFBE) |
|---|---|---|
| Only 1 frame | 2.21 | 1.65 |
| Three adjacent frames | 2.02 | 1.58 |
| Five adjacent frames | 1.95 | 1.59 |

## 6. Discussion

As the tables and figures suggest, objective measures of RMS LSD and RMS IS are correlated to each other and their results follow the same trend. This property implies the usefulness of both mentioned criteria in evaluating implementation of the bandwidth extension algorithm.

In this work, we showed that the results of implementation of bandwidth extension algorithm by GMM for LFBE feature vectors are better than implementation with corresponding MFCC feature vectors. The optimal number of Gaussian distributions for MFCC and LFBE feature vectors is equal to 128. Results of the different assessments indicate that extending the model by increasing the number of GMM components does not result in more improvement of the objective measures; it only adds more complexity to the calculation.

The results of neural network implementation show the advantage of LFBE feature vectors against MFCCs in the bandwidth extension algorithm. The best result for the best neural network training is 0.44 dB in RMS

LSD measurements. As shown in Figures 6 and 7, when implementing BWE using a neural network, the best result is produced when the input is normalized and the output is nonnormalized.

For the best result obtained from LFBE feature vectors, the neural network resulted in an improvement of 2.87 dB in LSD measurement and improvement of 0.33 dB in IS assessment as compared to the GMM. These improvements for MFCC feature vectors will be 2.47 dB and 0.29 dB, respectively.

As the results suggest, implementation of a neural network leads to a significant improvement. Using LFBE feature vectors has a better effect than using MFCCs, highlighting the advantage of LFBE coefficients. Results show that adjacent frames are effective in estimating the current frame spectral envelope. These results also show that using 3 adjacent frames is the best choice; therefore, it is not suggested to increase the number of frames to 5, because it will only add more complexity to the calculations and will not improve the results.

## 7. Conclusion

Bandwidth extension increases the quality of narrowband speech signals by adding the lost information to it. In this research, 2 methods, GMM and neural network, were used to estimate a high-band spectral envelope from narrowband information. Furthermore, 2 MFCC and LFBE feature vectors were used. The MFCC parameters were widely used in BWE approaches in previous studies, but the LFBEs for implementation of artificial bandwidth extension with a neural network were used for the first time in this research. Objective evaluations of both results of implementation with the GMM and neural network indicate the superiority of LFBE over MFCCs.

Moreover, implementation of the BWE approach with a neural network under the conditions and with the methods proposed in this study shows more promising results than the GMM technique. When LFBE feature vectors were used in both methods, the best result was obtained from a neural network with 3 normalized adjacent frame inputs. In this condition, the neural network has the advantage of 64.5% in RMS LSD and 61.8% in RMS IS against the GMM method. In addition, as compared to the feature vectors in implementation, the neural network shows improvement of 21.8% in RMS LSD and 22.2% in RMS IS for LFBE over MFCC coefficients.

## References

[1] Nels R, Vedstesen SA. Artificial bandwidth extension of narrowband speech. MSc, Aalborg University, Aalborg, Denmark, 2007.

[2] Shahina A, Yegnanarayana B. Mapping neural networks for bandwidth extension of narrowband speech. In: Annual Conference of the International Speech Communication Association, Interspeech; 17–21 September 2006; Pittsburgh, PA, USA. pp. 1435–1438.

[3] Iser B, Schmidt G. Bandwidth extension of telephony speech. In: Adali T, Haykin S, editors. Adaptive Signal Processing: Next Generation Solutions. New York, NY, USA: Wiley, 2010. pp. 349–391.

[4] Peter J, Peter V. On artificial bandwidth extension of telephone speech. Signal Process 2003; 83: 1707–1719.

[5] Nour-Eldin AH, Kabal P. Objective analysis of the effect of memory inclusion on bandwidth extension of narrowband speech. In: Annual Conference of the International Speech Communication Association, Interspeech; 27–31 August 2007; Antwerp, Belgium. pp. 2489–2492.

[6] Liu Ch, Fu Q, Narayanan S. Effect of bandwidth extension to telephone speech recognition in cochlear implant users. J Acoust Soc Am 2009; 125: 77–83.

[7] Qian Y, Kabal P. Dual-mode wideband speech recovery from narrowband speech. In: Annual Conference of the International Speech Communication Association, Interspeech; 1–4 September 2003; Geneva, Switzerland. pp. 1433–1436.

[8] Nour-Eldin AH, Kabal P. Combining frontend-based memory with MFCC features for bandwidth extension of narrowband speech. In: IEEE 2009 International Conference on Acoustics Speech and Signal Processing, ICASSP; 19–24 April 2009; Taipei, Taiwan:. pp. 4001–4004.

[9] Morales N, Hansen JHL, Toledano DT. MFCC compensation for improved recognition of filtered and band-limited speech. In: IEEE 2005 International Conference on Acoustics Speech and Signal Processing, ICASSP; 18–23 March 2005; Philadelphia, USA. pp. 521–524.

[10] Kamal Omar M, Hasegawa-Johnson M. Maximum conditional mutual information projection for speech recognition. In: IEEE 2002 International Conference on Acoustics Speech and Signal Processing, ICASSP; 13–17 May 2002; Orlando, FL, USA. pp. I81–84j.

[11] Jax P, Vary P. Feature selection for improved bandwidth extension of speech signal. In: IEEE 2004 International Conference on Acoustics Speech and Signal Processing, ICASSP; 17–21 May 2004; Montreal, Canada. pp. 697–700.

[12] Pulakka H, Remes U, Palomaki K, Kurimo M, Alku P. Speech bandwidth extension using Gaussian mixture model-based estimation of the high-band mel spectrum. In: IEEE 2011 International Conference on Acoustics Speech and Signal Processing, ICASSP; 22–27 May 2011; Prague, Czech Republic. pp. 5100–5103.

[13] Soon IY, Yeo CK. Bandwidth extension of narrowband speech using cepstral analysis. In: International Symposium on Intelligent Multimedia, Video and Speech Processing, ISIMP; 20–22 October 2004; Hong Kong. pp. 242–245.

[14] Bauer P, Fingscheidt T. An HMM-based artificial bandwidth extension evaluated by cross-language training and test. In: IEEE 2008 International Conference on Acoustics Speech and Signal Processing, ICASSP; March 30–April 4 2008; Las Vegas, NV, USA. pp. 4589–4592.

[15] Seltzer ML, Acero A, Droppo J. Robust bandwidth extension of noise- corrupted narrowband speech. In: Annual Conference of the International Speech Communication Association, Interspeech; 4–8 September 2005; Lisbon, Portugal. pp. 1509–1512.

[16] Nilsson M, Kleijn WB. Avoiding over-estimation in bandwidth extension of telephony speech. In: IEEE 2001 International Conference on Acoustics, Speech, and Signal Processing, ICASSP; 7–11 May 2001; Salt Lake City, UT, USA. pp. 869–872.

[17] Park KY, Kim HS. Narrowband to wideband conversion of speech using GMM based transformation. In: IEEE 2000 International Conference on Acoustics Speech and Signal Processing, ICASSP; 5–9 June 2000; İstanbul, Turkey. pp.1843–1846.

[18] Pulakka H, Myllyla V, Laaksonen L, Alku P. Bandwidth extension of telephone speech using a filter bank implementation for highband mel spectrum. In: 18th European Signal Processing Conference; 23–27 August 2010; Aalborg, Denmark. pp. 599–602.

[19] Kim W, Hansen JHL. Time-frequency correlation-based missing-feature reconstruction for robust speech recognition in band-restricted conditions. IEEE T Audio Speech 2009; 17: 1292–1304.

[20] Blumer A, Ehrenfeucht A, Haussler D, Warmuth M. Learnability and the Vapnik–Chervonenkis dimension. J Assoc Comput Mach 1989; 36: 929–965.

[21] Waibel A, Hanazawa T, Hinton G, Shikano K, Lang K. Phoneme recognition using time-delay neural networks. IEEE T Audio Speech 1989; 37: 328–339.

[22] Milner B, Shao X. Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model. In: Annual Conference of the International Speech Communication Association, Interspeech; 16–20 September 2002; Denver, CO, USA. pp. 2421–2424.

[23] Chazan D, Hoory R, Cohen G, Zibulski M. Speech reconstruction from mel frequency cepstral coefficients and pitch frequency. In: IEEE 2000 International Conference on Acoustics Speech and Signal Processing, ICASSP; 5–9 June 2000; İstanbul, Turkey. pp. 1299–1302.

[24] Nour-Eldin AH, Kabal P. Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech. In: Annual Conference of the International Speech Communication Association, Interspeech; 22–26 September 2008; Brisbane, Australia. pp. 53–56.

[25] Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, Dahlgren NL. DARPA TIMIT acoustic-phonetic continuous speech corpus. US Department of Commerce, NIST Speech Disc 1-1.1 Edition, Documentation for the TIMIT Database. Washington, DC, USA: US Department of Commerce; 1993.

[26] Laaksonen L, Pulakka H, Myllylä V, Alku P. Development evaluation and implementation of an artificial bandwidth extension method of telephone speech in mobile terminal. IEEE T Consum Electr 2009; 55: 1062–1078.