Artificial bandwidth extension based on regularized piecewise linear mapping with discriminative region weighting and long-span features

Duy Nguyen Duc, Masayuki Suzuki, Nobuaki Minematsu, Keikichi Hirose

The University of Tokyo, Tokyo, Japan

{ngd_duy, suzuki, mine, hirose}@gavo.t.u-tokyo.ac.jp

1. Abstract

Artificial Bandwidth Extension (ABE) has been introduced to improve perceived speech quality and intelligibility of narrowband telephone speech. Most of the existing algorithms divided ABE into 2 sub-problems, namely extension of the excitation signal and that of the spectral envelope. In this paper, we propose a new method for spectral envelope extension based on REgularized piecewise linear mapping with DIscriminative region weighting And Long-span features (REDIAL). REDIAL is a revised version of SPLICE, a well-known method for speech enhancement. In REDIAL, however, discriminative model is introduced for space division step of the original SPLICE. The proposed REDIAL-based method approximates non-linear transformation from narrowband features to their wideband counterpart by a summation of piecewise linear transformations. The proposed method was compared with the widely used GMM-based method, through objective and subjective evaluations in both speaker-dependent and speaker-independent conditions. Both evaluations showed that the proposed method significantly outperforms the conventional GMM-based method.
Index Terms: Artificial Bandwidth Extension, REDIAL, spectral envelope extension, objective and subjective evaluations

2. Introduction

Although human ears are able to perceived sound at much higher ferquencies than 8 kHz, and often more than 15 kHz, traditional telephone networks were designed to limit the frequency to a lower range, approximately below 3.4 kHz, in order to conserve the bandwidth and increase the number of voice streams transmittable by a transmission channel. This results in degradation of perceptual speech quality of the narrowband speech at receiving end. True wideband transmission is therefore desirable, but this requires a significant amount of cost and time, since the whole transmission chain including terminals and network elements need to be upgraded. This challenge can be overcome with Artificial Bandwidth Extension (ABE) technique. ABE is a technique that tries to recover missing low and high frequency components of the speech signal only from the narrowband speech. By integrating ABE into terminals of the telephone networks, we can easily realize wideband transmission without modifying the networks.

A number of techniques have been proposed over the years for bandwidth extension of narrowband speech signals, including methods based on codebook mapping [1] and statistical approaches [2, 3, 4]. Most of these ABE algorithms are based on the source-filter model [5] of speech production whereby the speech signal is regarded as output of the vocal tract filter which takes excitation source signals as input. This model breaks the problem down into two subtasks: one is to extend the spectral envelope, and the other is to extend the excitation signal. The extension of spectral envelope is typically considered as the main problem of ABE since it had been shown that extension of the spectral envelope has a large effect on speech quality of the reconstructed wideband speech [6].

It is known that the Gaussian Mixture Model (GMM) [7] represents robustly the acoustic space of speech and was successfully applied to the problem of spectral transformation, especially voice conversion [8]. Based on the successes in voice conversion, in [4] an effective approach to the problem of extending the spectral envelope was proposed. In this approach, the spectral envelope of wideband speech was estimated using a GMM trained by parallel data of narrowband speech and its corresponding wideband speech. This approach showed that there was a large improvement in speech quality from the original narrowband speech to the reconstructed wideband speech. However, the gap between the reconstructed wideband and the original wideband speech was still large.

Stereo-based Piecewise LInear Compensation for Environments (SPLICE) [9], in which non-linear transformation between two feature vectors is approximated by the summation of piecewise linear transformations, is an effective and widely used method in speech enhancement. A revised version of SPLICE, in which a discriminative model, long-span features and regularization are introduced into SPLICE, was proposed [10, 11] and has been shown to outperform the original SPLICE. This revised version was named REgularized piecewise linear mapping with DIscriminative region weighting And Long-span features (REDIAL). The aim of spectral envelope extension, which is to make a transformation from spectral envelope of narrowband speech to that of wideband speech, is very similar to the scheme of REDIAL. This suggests that we can apply REDIAL to the problem of spectral envelope extension as it is. In this paper, we propose an approach to the problem of spectral envelope extension based on REDIAL and describe its effectiveness through objective and subjective evaluations.

This paper is organized as follow. Section. 3 describes the conventional ABE method based on GMM. Section. 4 gives a brief view of SPLICE and our proposed REDIAL-based method. The general process of ABE is discussed in Section. 5. Section. 6 describes experiments and results.

3. GMM-based Bandwidth Extension [4, 8]

Let $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, ..., \boldsymbol{x}_N^\top]^\top$ be the feature vectors characterizing the narrowband and $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, ..., \boldsymbol{y}_N^\top]^\top$ be the feature vectors characterizing the wideband speech. $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ and $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ define feature vectors consisting of static and dynamic features at frame t of narrowband and wideband speech, respectively.

In the training step, we model the joint probability density of the source and the target features by a GMM as follows:

$$P(\boldsymbol{Z}_t; \theta) = \sum_{m=1}^{M} \omega_m \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\mu}_m^{(\boldsymbol{Z})}, \boldsymbol{\Sigma}_m^{(\boldsymbol{Z})}) \quad (1)$$

θ defines a parameter set of GMM, which consisting of weights ω_m, mean vectors $\boldsymbol{\mu}_m^{(\boldsymbol{Z})}$ and covariance matrices $\boldsymbol{\Sigma}_m^{(\boldsymbol{Z})}$. M is the total number of mixture components of GMM, and m is

GMM index. The mean vectors and covariance matrices can be decomposed as below:

$$\boldsymbol{\mu}_m^{(\boldsymbol{Z})} = \begin{bmatrix} \boldsymbol{\mu}_m^{(\boldsymbol{X})} \\ \boldsymbol{\mu}_m^{(\boldsymbol{Y})} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(\boldsymbol{Z})} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(\boldsymbol{XX})} & \boldsymbol{\Sigma}_m^{(\boldsymbol{XY})} \\ \boldsymbol{\Sigma}_m^{(\boldsymbol{YX})} & \boldsymbol{\Sigma}_m^{(\boldsymbol{YY})} \end{bmatrix} \quad (2)$$

The conditional probability $P(\boldsymbol{Y}_t^\top | \boldsymbol{X}_t^\top, m; \theta)$ is given by:

$$P(\boldsymbol{Y}_t^\top | \boldsymbol{X}_t^\top, m; \theta) = \mathcal{N}(\boldsymbol{Y}_t; \boldsymbol{E}_{m,t}^{(\boldsymbol{Y})}, \boldsymbol{D}_m^{(\boldsymbol{Y})}) \quad (3)$$

where

$$\boldsymbol{E}_{m,t}^{(\boldsymbol{Y})} = \boldsymbol{\mu}_m^{(\boldsymbol{Y})} + \boldsymbol{\Sigma}_m^{(\boldsymbol{YX})} \boldsymbol{\Sigma}_m^{(\boldsymbol{XX})^{-1}} (\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(\boldsymbol{X})}) \quad (4)$$

$$\boldsymbol{D}_m^{(\boldsymbol{Y})} = \boldsymbol{\Sigma}_m^{(\boldsymbol{YY})} - \boldsymbol{\Sigma}_m^{(\boldsymbol{YX})} \boldsymbol{\Sigma}_m^{(\boldsymbol{XX})^{-1}} \boldsymbol{\Sigma}_m^{(\boldsymbol{XY})} \quad (5)$$

In the conversion step, firstly, we can write the time sequences of feature vectors of narrowband and wideband speech as follow:

$$\boldsymbol{X} = [\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, ..., \boldsymbol{X}_N^\top]^\top \quad (6)$$

$$\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, ..., \boldsymbol{Y}_N^\top]^\top \quad (7)$$

A time sequence of the converted static feature vectors $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \hat{\boldsymbol{y}}_2^\top, ..., \hat{\boldsymbol{y}}_N^\top]^\top$ is then calculated as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P(\hat{\boldsymbol{m}} | \boldsymbol{X}, \theta) P(\boldsymbol{Y} | \boldsymbol{X}, \hat{\boldsymbol{m}}; \theta) \quad \text{subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \quad (8)$$

This problem can be solved by EM algorithm, in which the following auxiliary function is iteratively maximized:

$$Q(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{\text{all } m} P(m | \boldsymbol{X}, \boldsymbol{Y}, \theta) \log P(\hat{\boldsymbol{Y}}, m | \boldsymbol{X}; \theta) \quad (9)$$

Solution to the problem defined in Eqn. 8 is given by,

$$\hat{\boldsymbol{y}} = (\boldsymbol{W}^\top \overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}^{(\boldsymbol{Y})}} \quad (10)$$

where $\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}}$, $\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}^{(\boldsymbol{Y})}}$ are defined as follows (see [8, 12] for more details):

$$\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}} = \text{diag}[\overline{\boldsymbol{D}_1^{(\boldsymbol{Y})^{-1}}}, ..., \overline{\boldsymbol{D}_T^{(\boldsymbol{Y})^{-1}}}] \quad (11)$$

$$\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}^{(\boldsymbol{Y})}} = [\overline{\boldsymbol{D}_1^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}_1^{(\boldsymbol{Y})}}, ..., \overline{\boldsymbol{D}_T^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}_T^{(\boldsymbol{Y})}}] \quad (12)$$

$$\overline{\boldsymbol{D}_t^{(\boldsymbol{Y})^{-1}}} = \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}^{(\boldsymbol{Y})^{-1}} \quad (13)$$

$$\overline{\boldsymbol{D}_t^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}_t^{(\boldsymbol{Y})}} = \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}_{m,t}^{(\boldsymbol{Y})} \quad (14)$$

$$\gamma_{m,t} = P(m | \boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top; \theta) \quad (15)$$

4. REDIAL-based Bandwidth Extension

4.1. Original SPLICE [9]

SPLICE is an effective and widely used approach in speech enhancement. Different from GMM approach, in which posterior probabilities of indexes of GMM of joint feature vectors were used for space division, SPLICE uses posterior probabilities of indexes of GMM of corrupted input feature vectors.

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be N-dimensional feature vectors of clean speech and those of corrupted speech, respectively. In original SPLICE, an estimate $\hat{\boldsymbol{x}}$ of the clean speech feature is calculated as follows:

$$\hat{\boldsymbol{x}} = \sum_{k=1}^{K} p(k | \boldsymbol{y}) \boldsymbol{A}_k \boldsymbol{y}', \quad (16)$$

where $\boldsymbol{y}' = [1, \boldsymbol{y}^\top]^\top$ is an augmented feature vector. $\boldsymbol{A}_k$ is an conversion matrix in region k which is trained as described below.

First, a K-component GMM is trained using corrupted feature vectors $\boldsymbol{y}_i$.

$$p(\boldsymbol{y}) = \sum_{k=1}^{K} \omega_k N(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (17)$$

Next, the conversion matrix $\boldsymbol{A}_k$ is estimated using minimum mean square error criterion as follows:

$$\boldsymbol{A}_k = \arg\min_{\boldsymbol{A}_k} \sum_{i=1}^{I} p(k | \boldsymbol{y}_i) \| \boldsymbol{x}_i - \boldsymbol{A}_k \boldsymbol{y}_i' \|^2 \quad (18)$$

An estimate $\hat{\boldsymbol{x}}$ of the clean speech feature is now calculated by substituting $\boldsymbol{A}_k$ in Eqn. 18 into Eqn. 16.

4.2. DIscriminative region weighting And Long-span features (REDIAL) [10, 11]

REDIAL was first proposed in [10] for speech enhancement, in which a joint vector $[\boldsymbol{y}^\top, \hat{\boldsymbol{n}}^\top]^\top$, consisting of a corrupted feature vector $\boldsymbol{y}$ and an estimate vector $\hat{\boldsymbol{n}}$ of noise feature vector, is used instead of the corrupted vector $\boldsymbol{y}$ alone. Moreover, a discriminative model (LDA + GMM) is introduced to space division step to calculate the posterior probabilities of the clean feature GMM using the corrupted features. The estimation of clean feature vector becomes:

$$\hat{\boldsymbol{x}} = \sum_{k=1}^{K} p(k | \boldsymbol{L}[\boldsymbol{y}^\top, \hat{\boldsymbol{n}}^\top]^\top) \boldsymbol{A}_k [1, \boldsymbol{y}^\top, \hat{\boldsymbol{n}}^\top]^\top, \quad (19)$$

where $\boldsymbol{L}$ is a conversion matrix of LDA trained by using joint vectors $[\boldsymbol{y}^\top, \hat{\boldsymbol{n}}^\top]^\top$ with posterior probabilities of indexes $\{k\}$ of clean GMM as their labels. The conversion matrix $\boldsymbol{A}_k$ is estimated as below:

$$\boldsymbol{A}_k = \arg\min_{\boldsymbol{A}_k} \sum_{i=1}^{I} p(k | \boldsymbol{w}_i) \| \boldsymbol{x}_i - \boldsymbol{A}_k [1, \boldsymbol{y}^\top, \hat{\boldsymbol{n}}^\top]^\top \|^2 \quad (20)$$

where $\boldsymbol{w}_i = \boldsymbol{L}[\boldsymbol{y}^\top, \hat{\boldsymbol{n}}^\top]^\top$ are converted vectors of LDA. By using LDA, dimensionality of feature vectors can be reduced effectively. Moreover, using clean GMM indexes as labels of LDA is expected to improve the overall performance, since the purpose of speech enhancement is to estimate feature vectors in clean space. The effectiveness of this has been shown in [10].

In addition to the method described above, the authors also considered using a joint vector of several adjacent frames instead of feature vector of only a single frame. However, to avoid the over-fitting problem that might occur since the vectors dimensionality increases, regularization was used. By concatenating adjacent frames features, the input information increases, therefore an improvement in estimation of clean feature is expected. In [11], the authors have confirmed the effectiveness of this method.

4.3. Proposed method: REDIAL-based Bandwidth Extension

In this research, we adopted the method explained in Section. 4.2, to solve the problem of spectral envelope extension. Its detailed procedure is explained below:

1. Extracting feature vectors $\{\boldsymbol{y}_i\}_{i=1,...,I}$ of wideband speech, and $\{\boldsymbol{x}_i\}_{i=1,...,I}$ of narrowband speech. Define $\boldsymbol{v}_i$ is a joint vector of several feature vectors of frames adjacent to frame i.
2. Training a GMM using wideband feature vectors $\{\boldsymbol{y}_i\}_{i=1,...,I}$ and calculate $\{p(m|\boldsymbol{y}_i)\}_{i=1,..,I}$.
3. Training LDA using joint feature vectors $\{\boldsymbol{v}_i\}_{i=1,...,I}$ with $\{p(m|\boldsymbol{y}_i)\}_{i=1,..,I}$ as their class labels. After obtaining the conversion matrix $\boldsymbol{L}$ of LDA, calculate converted vectors $\boldsymbol{z}_i = \boldsymbol{L}\boldsymbol{v}_i$.
4. Training a GMM using the converted vectors $\boldsymbol{z}_i$ and calculate probability $p(k|\boldsymbol{z}_i)$.
5. The linear conversion matrix $\boldsymbol{A}_k$ is estimated using a weighted minimum mean square error criterion with regularization as below:

$$\begin{aligned}\boldsymbol{A}_k &= \arg\min_{\boldsymbol{A}_k} \sum_{i=1}^{I} p(k|\boldsymbol{z}_i)\|\boldsymbol{y}_i - \boldsymbol{A}_k\boldsymbol{v}_i - \boldsymbol{\mu}_k\|^2 \\ &+ \lambda\|\boldsymbol{A}_k\|^2, \end{aligned} \tag{21}$$

where $\boldsymbol{\mu}_k$ is mean of component k of the GMM of wideband feature vectors and λ is regularization parameter. Solution to this problem is given by:

$$\boldsymbol{A}_k = \boldsymbol{Y'}\boldsymbol{P}\boldsymbol{X'}^{\top}(\boldsymbol{X'}\boldsymbol{P}\boldsymbol{X'}^{\top} + \lambda\boldsymbol{E})^{-1}, \tag{22}$$

where $\boldsymbol{Y'}$ is the sequence of feature vectors $\boldsymbol{y}'_i = \boldsymbol{y}_i - \boldsymbol{\mu}_k$, and $\boldsymbol{X'}$ is the sequence of joint feature vectors $\boldsymbol{v}_i$. $\boldsymbol{P}$ is a diagonal matrix given by $\boldsymbol{P} = \mathrm{diag}([p(k|\boldsymbol{z}_1), ..., p(k|\boldsymbol{z}_I)])$.

6. Finally, estimation of wideband feature vector given the narrowband one is as follow:

$$\hat{\boldsymbol{y}}_i = \sum_{k=1}^{K} p(k|\boldsymbol{z}_i)(\boldsymbol{A}_k\boldsymbol{v}_i + \boldsymbol{\mu}_k) \tag{23}$$

5. Baseline Bandwidth Extension System

The general process of ABE is shown in Fig. 1. First, mel-cepstral coefficients, aperiodic components and F0 of the narrowband speech are extracted using STRAIGHT [13, 14] (*Step 1*). Aperiodic components of the wideband speech are estimated by a simple MMSE-based GMM mapping method [3] (*Step 2*). Mel-cepstral coefficients which represent the spectral envelope are estimated by performing feature conversion as described in Section. 3 and Section. 4.3 (*Step 2*). After that, an estimated wideband speech is generated using the extracted F0 and converted features above (*Step 3*). The estimated wideband speech is now passed through LPF and HPF to generate low-band and high-band speech signals (*Step 4*). For the input narrowband speech, we up-sample it to make an input low-band speech signal (*Step 5*). Then, the power of estimated low-band speech signal is adjusted to that of input low-band speech signal. During the process the power of estimated high-band speech signal is also adjusted to keep its proportion the the low-band counterpart (*Step 6*). Finally, the wideband speech is reconstructed by adding the adjusted high-band speech signal to the input low-band speech signal (*Step 7*).

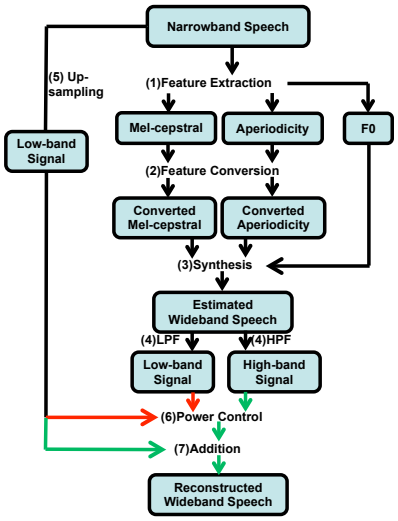

Figure 1: General flowchart of bandwidth extension

Table 1: LPF specifications

Stop band	3.7 - 8kHz
Transition band	3.4 - 3.7kHz
Pass band	0Hz - 3.4kHz

6. Experiments

6.1. Speaker-Dependent Model

6.1.1. Experiment Conditions

We conducted experiments under a speaker-dependent condition using the ATR phonetically balanced corpus [15]. The wideband speeches were 16kHz sampled speeches from subset A (training data) and subset B (evaluation data) of 4 Japanese speakers (2 males and 2 females). The narrowband speech was made by passing the corresponding wideband one through a LPF (described in Table 1), then downsampling the output.

In our experiments, we used STRAIGHT to extract mel-cepstral coefficients (spectral envelope) and F0, aperiodic components (*mixed excitation signal*). For both narrowband and wideband speeches, 24-dimensional mel-cepstral coefficients were used. Regarding to aperiodic components, the averaged components on 3 frequency bands (0 - 1, 1 - 2 and 2 - 4 kHz) for narrowband, and those on 5 frequency bands (0 - 1, 1 - 2, 2 - 4, 4 - 6 and 6 - 8 kHz) for wideband were used. In this paper, we adopted a simple MMSE-based GMM mapping method [3] for extension of the excitation signal. The number of mixture components of the GMM was set to 8.

For extension of the spectral envelope, we used a 64-component GMM in both conventional and proposed methods. The number of frames to be concatenated was set to 5 by referring to the results of our preliminary experiments. The regularization parameter was chosen by 5-fold cross validation: training data was divided equally into 5 subsets, then 4 subsets were used as training data and the left one was used as testing data. The optimal regularization parameter for each speaker was shown in Table 2.

6.1.2. Objective Evaluation

In this experiment, the Mel-Cepstral Distortion (MCD) defined below was used to evaluate the performance of conventional and

Table 2: Optimal regularization parameters in a speaker-dependent condition

Speaker	ftk	fws	mmy	msh
λ	0.003	0.009	0.003	0.002

Table 3: Objective evaluation (Speaker-dependent): Mel-cepstral distortion between regenerated speech and original speech

	Speaker	ftk	fws	mmy	msh
MCD [dB]	GMM	3.59	3.70	3.51	3.43
	REDIAL	1.95	1.88	1.86	1.87

proposed methods.

$$MCD[dB] = \frac{10}{\ln 10}\sqrt{2\sum (mc_i^X - mc_i^Y)^2} \qquad (24)$$

where mc^X, mc^Y are mel-cepstral coefficients of regenerated wideband speech and natural wideband speech, respectively. Objective evaluation results for 4 speakers are shown in Table 3.

An approximate 50% reduction in MCD can be seen for every speaker. This demonstrates the superiority of proposed method to the conventional one.

6.1.3. Subjective Evaluation

Subjective evaluation was conducted by using Mean Opinion Score method, which is defined in ITU-T Recommendation P.800 [17]. Opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). Listening results from 21 listeners (2 females, 19 males; age 19 to 22) using SONY MDR-900ST headphones are shown in Fig. 2.

The reconstructed wideband speeches in both approaches showed better perceptual quality than the original narrowband. Moreover, listening test results also demonstrate that the proposed method significantly outperforms the conventional GMM approach (at significance level of 5%).

6.2. Speaker-Independent Model

6.2.1. Experiment Conditions

The effectiveness of the proposed method within a speaker-dependent condition was described in the previous section. In this section, we further verify the effectiveness of proposed method in a more practical condition, speaker-independent condition, using the TIMIT database [16]. The training set contains a total of 4620 utterances of 462 speakers, and the test set contains 1680 utterances of 168 speakers. Feature extraction and other analysis conditions were the same as those in speaker-dependent experiment, except the number of mixture components of GMM for spectral envelope was set to 256 instead of 64.

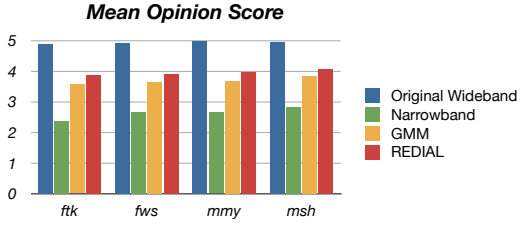

Figure 2: Speaker-dependent: Listening test results

Table 4: Objective evaluation(Speaker-independent): Mel-cepstral distortion between regenerated speech and original speech

Method	GMM	SPLICE	REDIAL
MCD[dB]	4.127	3.485	2.231

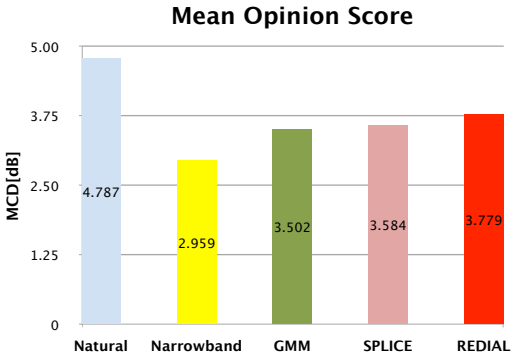

Figure 3: Speaker-independent: Listening test results

In objective evaluation, the regularization parameter of the proposed method was set to 0.1 due to results of by 8-fold cross validation. In subjective evaluation, we used 40 sets of speech samples (each contained the original wideband, narrowband, GMM-based wideband, SPLICE-based wideband and REDIAL based wideband speeches).

6.2.2. Experiments

Results of objective evaluation and subjective evaluation of 16 listeners (7 females, 9 males; age 20 to 25) using Sony MDR-ZX100 headphones are shown in Table 4 and Fig. 3 respectively.

It can be concluded that the original SPLICE showed slightly better performance than the conventional GMM-based method, while the proposed REDIAL-based method significantly outperforms both of them.

Similar to the speaker-dependent case, in this subjective evaluation we also observed a remarkable improvement in speech quality of reconstructed wideband compared to the original narrowband speech in all of three methods. More importantly, with the proposed method we achieved reconstructed wideband speech with significantly better speech quality compare to conventional GMM and SPLICE-based methods (at significance level of 5%).

7. Conclusions

In this paper, we proposed a new approach to the problem of spectral envelope extension of ABE based on REDIAL, a revised version of SPLICE. The merit of the proposed method to the conventional GMM-based method was confirmed on objective and subjective speech quality evaluations in both speaker-dependent and speaker-independent conditions.

For future work, we want to solve the problem of bandwidth extension in noisy environment by applying the proposed method to both problems of speech enhancement and bandwidth extension. Furthermore, we also consider the possibility of applying our proposed method into real systems.

8. References

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

[2] P. Jax, P. Vary, "Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden Markov model," Proceedings of the ICASSP 2003, pp.680-683, 2003.

[3] K. Park, *et al.*, "Narrowband to wideband conversion of speech using GMM based transformation," *Proc. ICASSP*, pp. 1843-1846, 2000.

[4] T. Toda, *et al.*, "Bandwidth Extension of Cellular Phone Speech based on Maximum Likelihood Estimation with GMM," *Proc. 2008 NCSP*, pp. 283-286, March 2008.

[5] L.R. Rabiner, *et al.*, "Digital Processing of Speech Signals," *Prentice Hall*, 1978.

[6] P. Jax, *et al.*, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, pp. 1707-1719, 2003.

[7] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech and Audio Processing, Vo1.3, no. 1, pp. 72-83, 1995.

[8] T. Toda, *et al.*, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *IEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222-2234, 2007.

[9] J. Droppo, *et al.*, "Evaluation of SPLICE on the Aurora 2 and 3 Tasks," *Proc. ICSLP*, pp. 29-32, 2002.

[10] M. Suzuki, *et al.*, "MFCC enhancement using joint corrupted and noise feature space for highly non-stationary noise environments," *Proc. ICASSP*, pp. 4109-4112, 2012.

[11] M. Suzuki, *et al.*, "Feature enhancement with Joint Use of Consecutive Corrupted and Noise Feature Vectors with Discriminative Region Weighting," *IEEE Transactions on Audio, Speech and Language Processing* (submitted).

[12] K. Tokuda, *et al.*, "Speech parameter generation algorithms for HMM-based speech synthesis," *in Proc. ICASSP*, pp. 1315-1318, Jun. 2000.

[13] H. Kawahara, *et al.*, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," *MAVEBA 2001*, September 2001.

[14] H. Kawahara, *et al.*, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a reptitive structure in sounds," *Speech Communication*, 27, pp.187-207, 1999.

[15] A. Kurematsu, *et al.*, "ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis," Speech Communication, 9, 357-363 (1990).

[16] L. D. Consortium, "Timit acoustic-phonetic continuous speech corpus," CD-ROM, ISBN 1-58563-019-5.

[17] ITU-T Rec. P. 800, "Methods for subjective determination of transmission quality," International Telecommunication Union, Geneva, 1996.