INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Semi-Coupled Dictionary based Automatic Bandwidth Extension Approach for Enhancing Children's ASR

Ganji Sreeram and Rohit Sinha

Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati - 781039, India
{s.ganji, rsinha}@iitg.ernet.in

Abstract

The work presented in this paper is motivated by our earlier work exploring sparse representation based approach for automatic bandwidth extension (ABWE) of speech signals. In that work, two dictionaries one for voiced and the other for unvoiced speech frames are created using KSVD algorithm on wideband data. Each of the atoms of these dictionaries is then decimated and interpolated by a factor of 2 to generate narrowband interpolated (NBI) dictionaries whose atoms have one-to-one correspondence with those of the WB dictionaries. The given narrowband speech frames are also interpolated to generated NBI targets and those are sparse coded over the NBI dictionaries. The resulting sparse codes are then applied to the WB dictionaries to estimate the WB target data. In this work, we extend the said approach by making use of an existing semi-coupled dictionary learning (SCDL) algorithm. Unlike the direct dictionary learning, the SCDL algorithm also learns a set of bidirectional transforms coupling the dictionaries more flexibly. The bandwidth enhanced speech obtained employing the SCDL approach and a modified high/low band gain adjustment yields significant improvements in terms of speech quality measures as well as in the context of children's mismatched speech recognition.

Index Terms: Speech bandwidth enhancement, sparse representation, semi-coupled dictionary.

1. Introduction

The legacy telecom networks employ both narrowband and wideband adaptive multi-rate (AMR) speech codecs with dynamic mode selection to achieve high throughput. In narrowband AMR codec, the speech signal is limited to 0.3-3.4 kHz. Thus, a loss of highband (HB) spectral information occurs when that mode is selected for communicating over the network. This results in degradation of the speech perceptual quality when compared with that of the wideband (WB) speech. Since the unvoiced speech contains significant HB information restoring that is expected to improve the speech quality. The techniques used to estimate the lost HB information from the given lowband (LB) speech data are called as artificial bandwidth extension (ABWE) [1] techniques. The research in ABWE area still has relevance as it helps improve the perceptual quality of speech without actually increasing the bit rate or the format through the addition of side information of the narrow-band speech signal.

The sparse representation (SR) technique has been widely used in many signal processing applications and also resulted in the state-of-the-art performances. The main aim of the SR technique is to approximate the target signals as the linear combinations of a small number of atoms from a data driven dictionary. Lately, the attention has been given towards creating suitable dictionaries for many applications. Dictionary learning methods aims at training an over-complete dictionary in a single feature space for different mapping and classification tasks. Motivated by that, in our earlier work [2], we have proposed a novel SR-based ABWE (SR-ABWE) approach for speech signals. In that work, a WB dictionary is learned from the given WB training speech frames using the KSVD algorithm. Now, the NBI dictionary for sparse coding the target NBI speech frames is derived by first decimating then followed by interpolating the atoms of the learned WB dictionary by a factor of 2. Similar to the existing ABWE approaches, the given lowband (LB) speech has been retained without any modification while the estimated highband (HB) speech has been added to the given LB speech with appropriate amplitude scaling. This approach referred to as the direct-coupled dictionary learning (DCDL) in this work. The earlier proposed DCDL approach for the ABWE task was noted to be quite effective for the voiced speech frames but found to be less effective for the unvoiced speech frames. This is due to the inconsistency in WB and NBI sparse codes. Also, the voiced and the unvoiced speech characteristics differs significantly. So, the dictionaries are to be properly learned to maintain consistency in the WB and NBI sparse codes and also to represent the target voiced and unvoiced speech data.

To address the above mentioned problems, in this work, we have explored an existing semi-coupled dictionary learning (SCDL) [3] algorithm for the ABWE task. This algorithm aims at learning bidirectional transformers iteratively along with the coupled dictionaries. In SCDL approach, the dictionaries are not fully coupled, and hence the two spaces can be mapped more flexibly. This results in more effective modeling of the characteristics of speech signal especially in the case of the unvoiced speech frames. Further to improve the quality of the enhanced speech, the existing scheme for adjusting the gain between LB and HB bands is also modified. Instead of the whole band, now the energies in the narrow higher edge of the estimated and the given LB speech are considered for the gain estimation.

The remaining of this paper is organized as follows: In Section 2, the explored SCDL dictionary learning technique for SR-ABWE with the modified gain adjustment approach is presented. A detailed description of the experimental setup for the ABWE technique is given in Section 3. The evaluation of the SCDL-based bandwidth enhanced speech in terms of speech quality measures is presented in Section 4. Also, in Section 5 the effectiveness of the contrast and the explored ABWE approaches are evaluated in context of children's automatic speech recognition (ASR) under mismatched condition. This paper is finally concluded in Section 6.

http://dx.doi.org/10.21437/Interspeech.2016-798

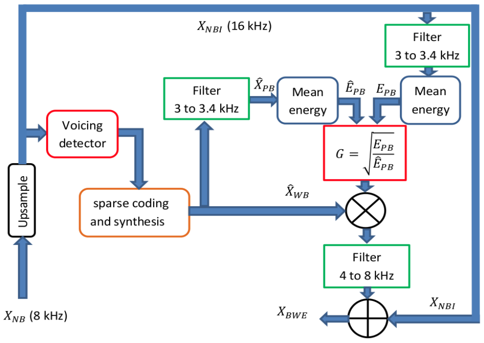

Figure 1: The complete block diagram of the semi-coupled dictionary learning (SCDL) based ABWE approach including the proposed modifications in terms of amplitude scaling.

2. SCDL-based SR-ABWE approach

The basic SR-ABWE approach was reported in our earlier work [2]. In that work, two dictionaries one for voiced and the other for unvoiced speech frames are created using KSVD algorithm on WB data. Each of the atoms of these dictionaries is then decimated and interpolated by a factor of 2 to generate narrowband interpolated (NBI) dictionaries whose atoms have one-to-one correspondence with those of the WB dictionaries. The given narrow (NB) speech frames are also interpolated to generated NBI targets and those are sparse coded over the NBI dictionaries. The resulting sparse codes are then applied to the WB dictionaries to estimate the WB target data. This approach worked well for voiced speech frames but found to be less effective for unvoiced speech frames due to the inconsistency in WB and NBI sparse codes. So, to address this issue, in this work a pair of dictionaries are learned to capture the structural hidden characteristics of the two spaces along with a mapping function. Once the dictionaries and the mapping function are learned, ABWE can be performed using them. Since, the two dictionaries are not fully coupled, this approach is referred to as semi-coupled dictionary learning (SCDL). The detailed block diagram of the proposed SCDL-based ABWE (SC-ABWE) approach for a speech frame is shown in Fig. 1. The block diagram shows that the given speech frame is first classified into two broad classes (V/UV) using a voice detector. For separating passband (PB) and HB portions of the bandwidth enhanced speech, 40-order Chebyshev filters are employed. Further, the creation of semi-coupled dictionaries is elaborated in Sec. 2.1.

2.1. SCDL algorithm

Given the training speech frames $\mathbf{S}_x$ and $\mathbf{S}_y$ corresponding to NBI and WB speech data, respectively. The SCDL algorithm aims at minimizing the energy function given by Eq. 1 to find the coupled dictionaries along with the desired mapping function:

$$\min_{\{\boldsymbol{\Phi}_x,\boldsymbol{\Phi}_y,f(.)\}} E_d(\boldsymbol{\Phi}_x,\mathbf{S}_x)+E_d(\boldsymbol{\Phi}_y,\mathbf{S}_y)+$$
$$\eta E_m(f(\boldsymbol{\theta}_x),\boldsymbol{\theta}_y)+\zeta E_r(\boldsymbol{\theta}_x,\boldsymbol{\theta}_y,f(.),\boldsymbol{\Phi}_x,\boldsymbol{\Phi}_y) \tag{1}$$

where $E_d(.,.)$ and $E_m(.,.)$ represent the data description error, the mapping error between the coding coefficients of two spaces while E_r represents the regularization term. In this approach, the speech data $\mathbf{S}_x$ and $\mathbf{S}_y$ are related to $\boldsymbol{\Phi}_x$ and $\boldsymbol{\Phi}_y$ by a mapping function $f(.)$. Here, the semi-coupled dictionaries ($\boldsymbol{\Phi}_x$ and $\boldsymbol{\Phi}_y$) and the mapping function $f(.)$ are jointly optimized.

Assuming $f(.)$ is a linear transformation, Eq. 1 can be further modified into the following dictionary learning and regression problem:

$$\min_{\{\boldsymbol{\Phi}_x,\boldsymbol{\Phi}_y,\mathbf{T}\}} \|\mathbf{S}_x-\boldsymbol{\Phi}_x\boldsymbol{\theta}_x\|_F^2+\|\mathbf{S}_y-\boldsymbol{\Phi}_y\boldsymbol{\theta}_y\|_F^2$$
$$+\eta\|\boldsymbol{\theta}_y-\mathbf{T}\boldsymbol{\theta}_x\|_F^2+\zeta_x\|\boldsymbol{\theta}_x\|_1+\zeta_y\|\boldsymbol{\theta}_y\|_1+\zeta_T\|\mathbf{T}\|_F^2$$
$$\text{s.t. } \|\boldsymbol{\phi}_{x_i}\|_{l_2}\leq 1, \|\boldsymbol{\phi}_{y_i}\|_{l_2}\leq 1, \forall i \tag{2}$$

where $\eta, \zeta_x, \zeta_y, \zeta_T$ are the regularization parameters and $\boldsymbol{\phi}_{x_i}$, $\boldsymbol{\phi}_{y_i}$ are the atoms of the dictionaries $\boldsymbol{\Phi}_x$ and $\boldsymbol{\Phi}_y$, respectively. The objective function in Eq. 2 is not jointly convex to $\boldsymbol{\Phi}_x$, $\boldsymbol{\Phi}_y$, $\mathbf{T}$. But, it is convex with respect to each of the parameters if others are kept fixed. So, an iterative algorithm can be designed to optimize the variables alternatively. The mapping function $\mathbf{T}$ is initialized as an identity matrix and the coding coefficients $\boldsymbol{\theta}_x$ and $\boldsymbol{\theta}_y$ are assumed to be identical.

2.1.1. *Training*

To minimize Eq. 2, the objective function is partitioned into three sub-tasks namely sparse coding for the training data, updating the dictionary and updating the mapping function. By initializing the mapping function $\mathbf{T}$ and the dictionary pair $\boldsymbol{\Phi}_x$ and $\boldsymbol{\Phi}_y$, the sparse coefficients $\boldsymbol{\theta}_x$ and $\boldsymbol{\theta}_y$ can be calculated as follows:

$$\min_{\{\boldsymbol{\theta}_x\}} \|\mathbf{S}_x-\boldsymbol{\Phi}_x\boldsymbol{\theta}_x\|_F^2+\eta\|\boldsymbol{\theta}_y-\mathbf{T}_x\boldsymbol{\theta}_x\|_F^2+\zeta_x\|\boldsymbol{\theta}_x\|_1$$
$$\min_{\{\boldsymbol{\theta}_y\}} \|\mathbf{S}_y-\boldsymbol{\Phi}_y\boldsymbol{\theta}_y\|_F^2+\eta\|\boldsymbol{\theta}_x-\mathbf{T}_y\boldsymbol{\theta}_y\|_F^2+\zeta_y\|\boldsymbol{\theta}_y\|_1 \tag{3}$$

The Eq. 2 is an l_1-optimization problem and can be solved using the least angle regression (LARS) algorithm [4]. Here, we assume that the mapping function $\mathbf{T}$ is linear and the bidirectional transformation approach can be adopted to learn the transformation from $\boldsymbol{\theta}_x$ to $\boldsymbol{\theta}_y$ and from $\boldsymbol{\theta}_y$ to $\boldsymbol{\theta}_x$ simultaneously.

By fixing $\boldsymbol{\theta}_x$ and $\boldsymbol{\theta}_y$, the dictionaries $\boldsymbol{\Phi}_x$ and $\boldsymbol{\Phi}_y$ can be updated by using the following equation:

$$\min_{\{\boldsymbol{\Phi}_x,\boldsymbol{\Phi}_y\}} \|\mathbf{S}_x-\boldsymbol{\Phi}_x\boldsymbol{\theta}_x\|_F^2+\|\mathbf{S}_y-\boldsymbol{\Phi}_y\boldsymbol{\theta}_y\|_F^2$$
$$\text{s.t. } \forall i, \|\boldsymbol{\phi}_{x_i}\|_{l_2}\leq 1, \|\boldsymbol{\phi}_{x_i}\|_{l_2}\leq 1 \tag{4}$$

By fixing the dictionaries and coding coefficients the mapping function $\mathbf{T}$ can be updated as follows:

$$\min_{\{\mathbf{T}\}} \|\boldsymbol{\theta}_y-\mathbf{T}_x\boldsymbol{\theta}_x\|_F^2+(\zeta_T/\eta).\|\mathbf{T}\|_F^2 \tag{5}$$

Eq. 5 is a ridge regression problem. So, it can be solved as follows:

$$\mathbf{T}=\boldsymbol{\theta}_y\boldsymbol{\theta}_x^T(\boldsymbol{\theta}_x\boldsymbol{\theta}_x^T+(\zeta_T/\eta).\mathbf{I})^{-1} \tag{6}$$

where $\mathbf{I}$ represents an identity matrix.

2.1.2. *Synthesis*

By using the resultant dictionaries $\boldsymbol{\Phi}_x$, $\boldsymbol{\Phi}_y$ and the mapping function $\mathbf{T}$, the target NB speech frame $\mathbf{s}_{x_i}$ can be transformed into corresponding enhanced WB speech frame $\mathbf{s}_{y_i}$ by iterating the following equations:

$$\min_{\{\boldsymbol{\alpha}_{x_i},\boldsymbol{\alpha}_{y_i}\}} \|\mathbf{s}_{x_i}-\boldsymbol{\Phi}_y\boldsymbol{\alpha}_{x_i}\|_F^2+\|\mathbf{s}_{y_i}-\boldsymbol{\Phi}_y\alpha_{y_i}\|_F^2+$$
$$\eta\|\boldsymbol{\alpha}_{y_i}-\mathbf{T}_y\boldsymbol{\alpha}_{x_i}\|_F^2+\zeta_x\|\boldsymbol{\alpha}_{x_i}\|_1+\zeta_y\|\boldsymbol{\alpha}_{y_i}\|_1 \tag{7}$$

Algorithm 1 *: Semi-coupled dictionary learning (SCDL) algorithm*

```
Input: Given the training speech data $\mathbf{S}_x$ and $\mathbf{S}_y$, the initial dictionary pair $\mathbf{\Phi}_x$ and $\mathbf{\Phi}_y$ and the initial mapping functions $\mathbf{T}_x$ and $\mathbf{T}_y$ for the two spaces.
For each iteration Until convergence:
    1. By sparse coding, update $\theta_x$ and $\theta_y$ by fixing other variables in Eq. 3.
    2. Update $\mathbf{\Phi}_x$ and $\mathbf{\Phi}_y$ by fixing other variables in Eq. 4.
    3. Update $\mathbf{T}_x$ and $\mathbf{T}_y$ by fixing other variables in Eq. 5.
Output: $\mathbf{\Phi}_x$, $\mathbf{\Phi}_y$, $\mathbf{T}_x$ and $\mathbf{T}_y$
```

Algorithm 2 *: Bandwidth enhancement using SCDL*

```
Input: Given the target speech frame $\mathbf{s}_{x_i}$, learned dictionaries $\mathbf{\Phi}_x$ and $\mathbf{\Phi}_y$ and the learned mapping functions $\mathbf{T}_x$ and $\mathbf{T}_y$ for the two spaces.
Initialization: Initialize $\mathbf{s}_{y_i} = \mathbf{\Phi}_y(\mathbf{T}_x\boldsymbol{\alpha}_{x_i})$ where $\boldsymbol{\alpha}_{x_i}$ is obtained by sparse coding $\mathbf{s}_{x_i}$ over $\Phi_x$.
For each iteration Until convergence:
    1. Find the optimum $\boldsymbol{\alpha}_{y_i}$ using Eq. 7.
    2. Update $\mathbf{s}_{y_i}$ using Eq. 8.
Output: Enhanced WB speech frame $\mathbf{s}_{y_i}$.
```

$$\hat{\mathbf{s}}_{y_i} = \mathbf{\Phi}_y\boldsymbol{\alpha}_{y_i} \tag{8}$$

Finally, the enhanced WB speech frame $\mathbf{s}_{y_i}$ is the resultant $\hat{\mathbf{s}}_{y_i}$ after the final iteration.

2.2. Modified gain adjustment technique

In the existing DCDL approach, the scaling factor for adjusting the amplitude level estimated highband portions for each frame is determined by finding the ratio of energies of the given signal and the lowpass (0-4 kHz) version of the estimated WB signal. In subsequent investigations, this scheme was noted to result in improper scaling of highband for the unvoiced frames. In Fig. 2, the concerned spectra for an unvoiced frame are plotted to illustrate this problem. From that figure, it can be observed that the spectral profiles of the estimated and the target (original) signals differ substantially for lowband and highband regions. As a result of that the values of the estimated scale factor turns out to be quite low, this in turn leads to loss of the highband energy in the bandwidth enhanced signal. To address this problem, we explored collecting the energy from a narrow passband close to 4 kHz in scale factor computation. For this purpose, a 40-order Chebyshev bandpass filter is designed with higher edge of the passband as 3.4 kHz and its lower edge is tuned on a development dataset. The passband range of 3-3.4 kHz is found to result in the best performance.

3. Experimental setup

In this section, the detailed description for the database used in conducting the experiments, the system parameters tuning and the measures used for evaluating the quality of bandwidth enhanced speech have been discussed.

3.1. Speech database

The evaluation of the SC-ABWE approach has been performed on the speech data obtained from PFSTAR [5] speech corpus, a commonly used dataset used for evaluating children's automatic speech recognition (ASR) performance. In that dataset, the speech data is partitioned into training and test sets. A smaller development set is created from the existing training set for learning the dictionaries by randomly selecting 127 utterances such that at least one speech file is considered for each of 80 speakers (male and female). In the same way, a smaller evaluation set is created from the existing test set by randomly selecting 125 utterances such that at least one speech file is considered for each of 60 speakers (male and female) for evaluating the SC-ABWE approach. The selected training and test speech data is analyzed into frames of 20 ms length keeping a frame-shift of 5 ms. The training and the test speech frames are classified into voiced/unvoiced classes using *FXRAPT* function [6] available in the *voicebox*, a commonly used MATLAB-based speech toolbox. This resulted in a total of $382,148$ voiced and $329,882$ unvoiced frames in the training set and a total of $339,995$ voiced and $681,889$ unvoiced frames in the test set.

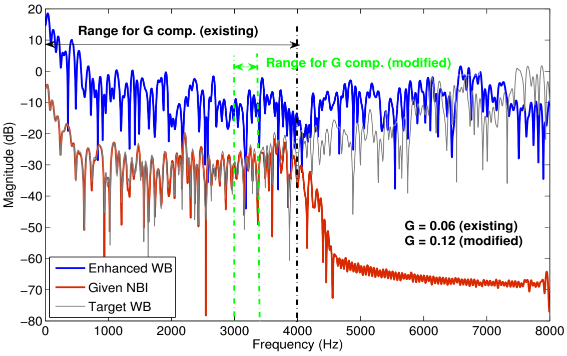

Figure 2: Spectral plots of the observed and the estimated signals for an unvoiced frame, illustrating the procedure followed in the modified gain computation to boost the energy of the estimated highband spectra.

3.2. System parameter tuning

The dictionaries in the DCDL approach are learned using KSVD algorithm with number of dictionary atoms as 1000, a sparsity value of 10, and iterations 50. In this approach, the orthogonal matching pursuit (OMP) algorithm with representation sparsity value of 50 is used during the sparse coding stage. In the case of the explored SCDL approach, the dictionary learning procedure is iterated 10 times with number of dictionary atoms as 1000. During synthesis, the re-estimation is performed for 10 iterations to obtain the enhanced target WB speech frame. In the SCDL approach, during the sparse coding stage the LARS algorithm is used without any constraint on the sparsity.

3.3. Performance measures

The effectiveness of the SCDL-based ABWE technique is measured by a number of speech quality measures, namely subband log spectral distortion (LSD) [7], log-likelihood ratio (LLR), segmental signal-to-noise ratio (segSNR) and perceptual evaluation of speech quality (PESQ). All the above mentioned speech quality measures are computed using the *Composite* tool downloaded from the website of the authors of [8].

4. Results

The effectiveness of the modified gain adjustment technique applied over the earlier proposed DCDL-based ABWE approach

Table 1: Evaluation of DCDL and SCDL based ABWE approaches in terms of speech quality measures. For highlighting the efficiency of the explored ABWE methods, the measures are also calculated for the interpolated narrowband speech and is referred to as 'No enhancement'.

Type	LSD	LLR	Seg_SNR	PESQ
No enhancement	13.99	4.03	14.62	4.49
DCDL + existing G	9.08	1.69	10.60	4.42
DCDL + modified G	8.75	1.50	10.47	4.42
SCDL + modified G	**7.93**	**1.09**	**10.14**	**4.42**

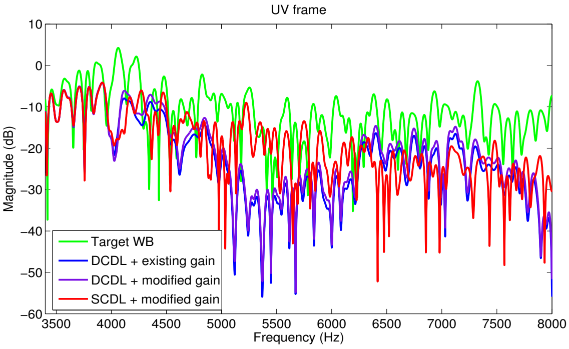

Figure 3: Plots showing the enhanced spectra of an unvoiced frame using SCDL and DCDL based approaches.

is evaluated in terms of various speech quality measures and the results of the same are given in Table 1. It is noted that with the proposed modification in gain adjustment, the overall quality of the bandwidth enhanced speech is significantly improved, perticularly in unvoiced case. To highlight the same, we have compared the enhanced spectra of the UV frames with the modified gain adjustment and default gain adjustment and the same are shown in Fig. 3.

Further, the effectiveness of the explored SCDL-based ABWE approach with the modified gain adjustment technique is also evaluated and the results are listed in the Table 1. This approach results in significant improvement in LSD and LLR, although a slight degradation is noted for segSNR compared to DCDL approach. Note that PESQ measure turned out to be more or less same for all the cases. From the figure, it is evident that, by the SCDL approach with the modified amplitude scaling, a significant improvement in bandwidth enhancement is achieved.

5. Evaluation of bandwidth enhanced speech for ASR

It is well known that the acoustic attributes between adults' and children's speech differ significantly [9, 10]. As the result, a high degradation in the recognition performance is noted when the children's speech recognition is performed over adults' speech trained ASR systems, compared to that of children speech trained ASR systems. Further, instead of WB speech data, when NB speech data is used for developing the ASR systems, the recognition performances for both adults' and children's undergo significant degradation. Also, the extent of degradation is larger for the recognition under children's mismatched condition, i.e., recognition of children speech data w.r.t., the adults' speech trained acoustic models. The reason behind this behavior is significant loss of the highband spectral information with reduction in the bandwidth of speech for children's speech unlike adults' speech. The effectiveness of the SR-ABWE approach using SCDL algorithm and the modified gain adjustment technique explored in this work seems to be encouraging in the speech domain. So, it would be interesting to evaluate whether the bandwidth enhancement explored are also helpful in bridging the gap between recognition performance for WB and NB speech in case of children's mismatched speech recognition. For this purpose, a conventional context-dependent hidden Markov model (CD-HMM) based ASR system is developed using the HTK toolkit [11]. The adults' (male and female) speech data from WSJCAM0 [12] speech corpus is used to train the ASR system parameters in mismatched condition.

Table 2: The ASR performance for different SR-ABWE methods. In this study, separate ASR systems are developed on adults' speech data and tested for children's speech recognition in varying test and training data conditions.

Test data condition	ASR system trained on	WER
NB	NB	51.02
WB	WB	36.68
NBI	WB	59.36
AWBE (DCDL + existing G)	WB	48.10
AWBE (DCDL + modified G)	WB	47.46
ABWE (SCDL + modified G)	WB	**46.41**

The ASR systems in this work are modeled by following the procedure described in the earlier work [13]. The speech analysis is carried out using a Hamming window of length 25 ms, a pre-emphasis factor of 0.97 and frame rate of 100 Hz. The 13-dimensional MFCC base feature vector is computed using a 21-channel mel-filterbank. The first and the second-order derivatives computed over a span of 5 frames are appended to the MFCC base features. Thus, resulting in a 39-dimensional feature vector that is used for the ASR modeling. Also, the cepstral mean subtraction is performed to all the MFCC features during the training and the testing phase. The feature extraction is performed using the HTK toolkit.

The ASR performances for the default and the proposed approaches are measured and noted in Table 2. It has been observed that the SCDL-based ABWE approach with modified gain adjustment technique results in improved ASR performance when compared to the other approaches.

6. Conclusion

In this work, we have explored SCDL-based ABWE technique. Also, a slight modification for the existing gain adjustment technique for ABWE task has been proposed. The SCDL algorithm jointly optimizes the coupled dictionaries and the mapping function resulting in proper coupling between the two spaces, particularly for the unvoiced speech. The speech quality performance measures for the explored SCDL-based ABWE approach have been calculated and found to result in significant improvements compared to that of DCDL-based ABWE approach. Also, the bandwidth enhanced speech files are evaluated in context of children's mismatched speech recognition. The SCDL-based enhanced files results in improved ASR performance compared to that or DCDL-based enhanced files.

7. References

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

[2] Y. Sunil and R. Sinha, "Sparse representation based approach to artificial bandwidth extension of speech," in *Signal Processing and Communications (SPCOM), International Conference on*, July 2014, pp. 1–5.

[3] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2012, pp. 2216–2223.

[4] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani *et al.*, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[5] M. Russell, "The PF-STAR British English childrens speech corpus," December 2006.

[6] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[7] A. H. Gray Jr and J. D. Markel, "Distance measures for speech processing," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 5, pp. 380–391, 1976.

[8] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, 2008.

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

[10] ——, "Exploration of MFCC based ABWE for robust children's speech recognition under mismatched condition," in *Signal Processing and Communications (SPCOM), International Conference on*, July 2014, pp. 1–5.

[11] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, *The HTK book*. Entropic Cambridge Research Laboratory Cambridge, 1997.

[12] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition," in *Acoustics, Speech, and Signal Processing, (ICASSP)., International Conference on*, vol. 1, May 1995, pp. 81–84 vol.1.

[13] S. Ghai and R. Sinha, "Pitch adaptive MFCC features for improving childrens mismatched ASR," *International Journal of Speech Technology*, vol. 18, no. 3, pp. 489–503, 2015. [Online]. Available: http://dx.doi.org/10.1007/s10772-015-9291-7