INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 4)
Available online at: www.ijariit.com

Speech reconstruction using machine learning approach for speech impaired persons

Kruthika R
rkruthika03@gmail.com
JSS Science and Technology University, Mysuru, Karnataka

Rajeswari P
rajju_p@sjce.ac.in
JSS Science and Technology University, Mysuru, Karnataka

ABSTRACT

The speech disordered persons are able to produce speech which sounds like they are whispering. The main objective of this work is to reconstruct the abnormal to a normal sounding speech by using MFCC coefficients to extract the feature and use these to train cascaded Gaussian Mixture Model (GMM) and Objective measures are used to evaluate the performance of the work. The data used for the work are from WTIMIT online corpus and the speech signals recorded from speech impaired subjects. In this work STRAIGHT toolbox is not employed for its complexity and muffled voice. The obtained SNR is reduced

Keywords— *Speech reconstruction, MFCC, Cascaded GMM*

1. INTRODUCTION

Speech is the major utterance form of communication used by humans. Speech generation is a systematic multi-step process by which emotions, thoughts, and ideas are expressed through spoken utterances. The speech generation starts from the lung which provides pulmonary pressure to produce sound through the glottis present in the larynx then the sound is modified by the vocal tract to generate different speech sounds. Any defect in these parts leads to the abnormal production of sound. These defects may be due to neurological, abnormal growths of the vocal folds, due to surgery involving removal of the larynx.

Chronical dysphonia [1] mainly occurs because of the malfunctioning of the vocal chords. Voice formed this way sounds whisper like characteristics. Larynx-related Dysphonia refers to the patients who have undergone a surgery called laryngectomy whose speech is partial or who have larynx damage, impaired brain function or nerve lesion, disturbance of nervous system [3]. These speech disordered persons are able to produce speech which sounds like they are whispering. Hence these disordered persons don't communicate much as the normal ones. Many Voice output communication aids and prosthetic aids are used to recover the speech but they sound much like artificial speech and the person may hesitate to communicate to the outside world and few aids as to be inserted internally through surgery which may involve a lot of pain and infections and risk factors.

A speech processing technique which provides various methods of reconstructing speech can be used to generate normal sounding speech. These reconstruction methods can be categorized into non-training and training methods. The former tries to reconstruct without any previous information they just try to increase the pitch of the signal as the whispered speech as low pitch compared to the normal one. The latter tries to match the parameters of defected one with the normal one. These are done by training the mapping models which uses algorithms of machine learning. The non-training method includes code excited linear prediction and mixed excitation linear prediction. The training method includes using Hidden Markov Models, Gaussian Mixture models, Restricted Boltzmann machines (RBM) and Semi-Deep Neural networks.

Hamid R Sharifzadeh et.al [2] has proposed speech reconstruction using twin mapping model, Mel-cepstra coefficients are used to train the model and reconstructs the speech using STRAIGHT.

Ian McLoughlin et.al [3] has proposed a Deep Neural Network (DNN) structure that allows a semi-supervised training approach on spectral features from smaller data sets. The proposed semi-supervised DNN (semi-DNN) architecture first uses unsupervised training by using two RBM's to perform coding of the feature, and then this is used to enable supervised training. This method provides good performance however it is computationally expensive to train and also requires expensive GPUs.

Hamid R. Sharifzadeh et.al [4] has proposed an algorithm which relies upon cascaded GMM mapping model and makes use of artificially generated whispers.

Jing-Jie Li et.al [5] has proposed an algorithm which uses multiple Restricted Boltzmann Machines which uses spectral envelops instead of me-cepstra coefficients and also decouples the Voiced/Unvoiced signals from the f0 estimation using Support Vector Machine (SVM). The method shows good performance however it is a high computation cost.

Ian Vince McLoughlin [6] proposed a reconstruction method for the conversion of continuous whispers to a natural sounding speech by reversing Linear Time-Invariant (LTI) source filter speech production model.

Ling-Hui Chen, Zhen-Hua Ling, Yan Song, Li-Rong Dai [7] presented a spectral modeling and conversion method for voice conversion. They use restricted Boltzmann machines (RBMs) as probability density models to model the joint distributions of the source and target spectral features.

Christophe Veaux et.al [8] has proposed a Hidden Markov Models (HMM) based speech synthesis, to build personalized synthetic voices to build personalized synthetic voices.

Tomoki Toda et.al [9] has proposed voice conversion methods from Non-Audible Murmur to normal speech and a whispered voice in a probabilistic manner using Gaussian Mixture Models.

Boon Pang Lim [10] has given the information about how the whispers are different from a phonated speech by perceptual and physiological studies. The data for this work is also used by the proposed authors WTIMIT open source for normal and whisper speech corpus.

Hamid R Sharfzadeh et.al [11] as proposed in which the reconstruction is done by using modified Code Excited Linear Prediction (CELP) codec in which the excitation sequence is selected from a codebook and then shaped by Long Term Predictor (LTP) filter to convey the pitch information of the speech.

Tomoki Toda, Alan W. Black and Keiichi Tokuda [12] describe a spectral conversion method for voice conversion (VC). A Gaussian mixture model (GMM) of the joint probability density of source and target features is used for performing spectral conversion between speakers.

Ian McLoughlin [13] Applied Speech and Audio Processing with MATLAB Examples is a Matlab based resource that blends speech and hearing research in describing the key techniques of speech and audio processing.

From the above review of the literature based on the less complexity, cost-effective and performance cascaded GMM mapping model has been employed in this paper MFCC coefficients in the feature extraction and STRAIGHT is not employed for synthesis of speech regular LPC synthesis filter is employed.

2. IMPLEMENTATION

The flow of the work is as shown in Figure 1. It includes Analysis of Speech, Synthesis of speech, Feature Extraction, Cascaded GMM model, Reconstruction of Speech and objective measures.

2.1 Analysis of speech

The data for this work is collected from the whispered TIMIT (wTIMIT) [10] corpus. In our work, both Singaporean English and North American speaker accents are used of 49 speakers including both male and female speakers. For the training 300 parallel whisper and normally spoken utterances are used. The other database is collected locally recorded from Speech impaired subjects and normal speech from normal subjects.

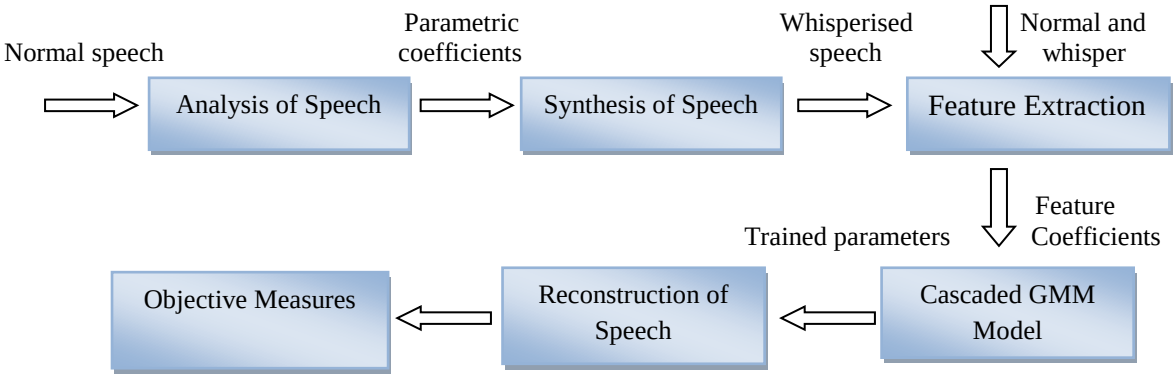

Fig. 1: Block diagram for speech reconstruction

In the analysis and synthesis part, only normal speech are used to synthesize the artificial whisper speech termed as whispered speech this provides the speech which is similar to the whisper as in this process we are going to make the pitch zero and the synthesized speech is time aligned with the normal speech which helps in reducing the complexity of aligning the speech signals during the training stage.

Firstly in the analysis part, framing and windowing of the normal speech signal are performed. An audio signal will be constantly changing, but it is pseudo-stationary over a short interval of time at this interval the signal doesn't change much hence framing of

the signal is performed. In this work, the frame length is considered to be 20ms. It is obtained by multiplying the sampling frequency of the signal with 20ms.

$$\text{Frame length}=fs*0.02 \quad (1)$$

Where fs is the sampling frequency and 0.02 seconds.
An overlap of 50% is used to prevent the data loss while using the window function.

$$\text{Frame shift}=\text{Frame length}/2 \quad (2)$$

The number of frames for a signal is calculated by

$$N= (\text{length (speech)} - \text{Frame length})/ \text{Frame shift}+1 \quad (3)$$

Where N is the number of frames of a given signal.
By using the number of frames N and window function framing of the signal is performed. A Hamming window function of finite length is used.

The Coefficients of the hamming window are computed using the below equation:

$$w(n) = 0.54 - 0.46\cos(2\pi\frac{n}{N}) \quad 0\leq n \leq N \quad (4)$$

Where N=L-1, L is the length of the frame

Secondly, for each frame of the signal analysis is performed. Firstly the pitch and pitch period is calculated using the Long-term prediction filter [13].

Consider's' audio samples, the relation of it with the pitch component is given by

$$x(n) = s(n) + \beta x(n - M) \quad (5)$$

Where x (n) is the signal that contains speech and β is the pitching amplitude and M is the pitch period.

Finally, The Linear Predictive coefficients (LPC) are calculated, by applying the LPC analysis filter to the speech signal provides the residual signal. Linear Prediction is a mathematical operation where future values of a signal are estimated as a function of the previous sample.

To calculate the LPC coefficients for an Lth order linear prediction filter, we have to predict the coefficients of next sample at time instant n which can be represented by a linear combination of the past L samples. This linear combination is given by

$$x'(n) = w1x(n-1) + w2x(n-2) + \cdots + wLx(n-L) \quad (6)$$

Where w1-wL is the predictor coefficients.
The error between the predicted sample and the next sample is obtained by minimizing given by

$$e(n) = x(n) - x'(n) \quad (7)$$

Minimize the mean-squared error over all n samples to obtain the optimum value

$$E = \sum_n e^2(n) = \sum_n\{x[n] - \sum_{k=1}^{L} a_k x[n-k]\}^2 \quad (8)$$

Differentiating equation (7) and equating to zero determines the set of LPC coefficients obtained by minimizing squared error E:

$$\frac{\delta E}{\delta w_j} = -2\sum_n x[n-j]\{x[n] - \sum_{k=1}^{L} a_k x[n-k]\} = 0 \quad (9)$$

Resulting in an L unknown set of Linear equations, that can be derived from the unknown speech samples x[n] to x [n-L]:

$$\sum_{k=1}^{L} a_k \sum_n x[n-j]\,x[n-k] = \sum_n x[n]x[n-j] \quad (10)$$

Where j=1, ... , L.

The set of equations can be solved by the autocorrelation method.
Firstly, we use the below equations to note the relationships hold:

$$\sum_{n=-\infty}^{\infty} x[n-j]x[n-k] = \sum_{n=-\infty}^{\infty} x[n-j+1]x[n-k+1] \quad (11)$$

and

$$\sum_{n=-\infty}^{\infty} x[n-j+1]x[n-k+1] = \sum_{n=-\infty}^{\infty} x[n]x[n+j-k] \quad (12)$$

Re-formulate equation (10), using the above relationship as follows:

$$\sum_{k=1}^{L} a_k \sum_{n=-\infty}^{\infty} x[n]x[n+j-k] = \sum_{n=-\infty}^{\infty} x[n]x[n-j] \quad (13)$$

$$R(k) = \sum_{n=-\infty}^{\infty} x[n]x[n+k] \quad (14)$$

Equation (12) as similarity with the standard autocorrelation function in (13), where $R(k)$ denotes the *k*th autocorrelation.
The set of L linear equations can be represented as the following matrix:

$$\begin{bmatrix} R(0) & R(1) & R(2) & \dots & R(L-1) \\ R(1) & R(0) & R(1) & \dots & R(L-2) \\ R(2) & R(1) & R(0) & \dots & R(L-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(L-1) & R(L-2) & R(L-3) & \dots & R(0) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_L \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(L) \end{bmatrix} \quad (15)$$

Where, $w_{1,\dots,L}$ are the LPC coefficients.
The Durbin- Levinson –Itakura method is one of the most used techniques for matrix solution given by:

$$k_{n+1} = \frac{e^n{}_{n+1}}{e_0{}^n} \quad \text{for } n = 0, \dots, \ L \quad (16)$$

$$e_0{}^{n+1} = e_0{}^{n} - k_{n+1}e_{n+1}{}^{n} = e^{n}{}_{0}(1 - k^{2}{}_{n+1}) \quad \text{for } i = n, \dots, L \tag{17}$$

Where $e_i{}^0 = R(i)$ are set to be the initial conditions for the recursion for each i in the set of L equations.
The LPC analysis filter removes the vocal tract information from a signal. After analysis, the obtained vector represents the residual signal.

2.2 Synthesis of Speech

The obtained residual signal from the analysis stage is used in the synthesis stage, the obtained pitch parameters from the analysis stage are made zero in the residual signal. Then the residual signal and LPC coefficients are passed to the LPC synthesis filter. The obtained signal from the synthesis is termed as whispered speech [2]. The obtained signal is time aligned with the normal speech which lessens the computation of aligning the signals.

2.3 Feature Extraction

The Feature extraction step is performed to extract the sequence of feature vectors that represent the speech signal. By using these features we are going to train our models.

2.3.1 Mel Frequency Cepstral Coefficients (MFCC)

Mel Frequency Cepstral Coefficients are a widely used method for feature extraction. The MFCC coefficients are in the frequency domain. The Mel scale used in the process represents a hearing procedure similar to the human ear. The relation between Mel scale and frequency of speech [14] is given by:

$$Melscale = \left[1125\ln(1 + \frac{f}{700})\right] \tag{18}$$

The MFCC is calculated for all the three speech signals. Fig 2 shows a block diagram for steps to calculate MFCC. Firstly Framing and windowing of the signal is performed similarly as given in section 2.1, the equations from (1) - (4). In the periodogram estimation, a Fast Fourier Transform (FFT) is performed.

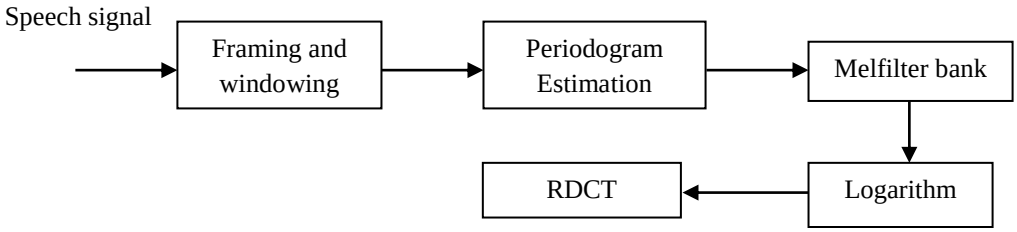

Fig. 2: Steps to calculate MFCC

A 512 point FFT is computed and only the first 257 are used. Next is to use Mel filter bank which contains a set of triangular filters, usually, it is set to 26 filter banks. Each filter bank is multiplied with the FFT coefficients and then the coefficients are added. We take the logarithm of the 26 coefficients which is a channel normalization technique that provides mean subtraction. Finally, DCT is taken to decorrelate the coefficients the obtained coefficient are called Mel Cepstrum Coefficients. Out of 26 cepstral coefficients only 13 coefficients are used.

2.3.2 Dynamic Time Warping (DTW)

As the whisper speech's length is a little lengthier than the normal it needs to be aligned before using them for training. During training, the whisper speech's MFCC is time aligned to the whispered MFCC. Firstly, the Short-term Fourier Transform (STFT) is computed; by using the cosine distance the local match is found between both the STFT's. Then we use dynamic programming to find the lowest cost path. Then the whisper MFCC is aligned using minimum cost path.

2.4 Gaussian Mixture Model Mapping

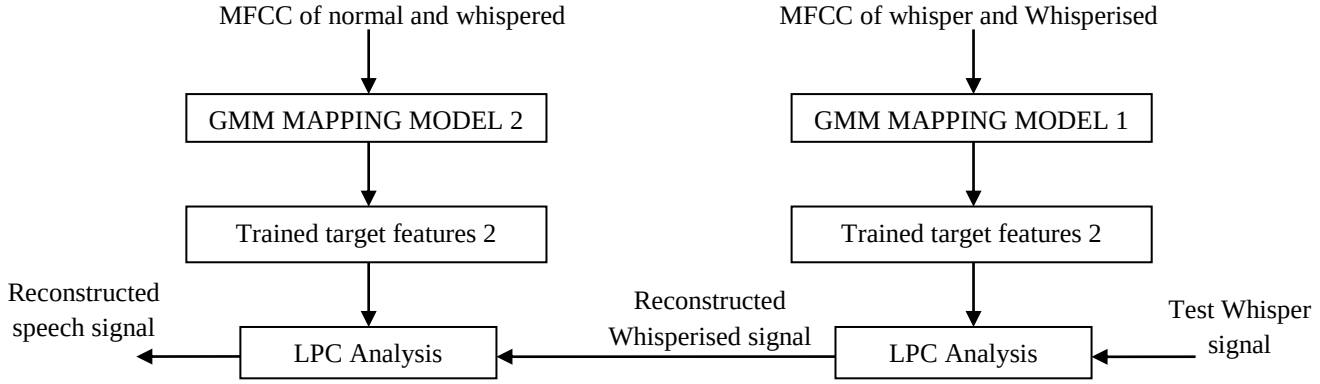

Fig. 3: Block diagram of Cascaded GMM models

The feature vectors obtained from feature extraction are now used to train the model; here the source feature vector is taken to be the abnormal signal which we are going to test. The target feature vector contains a feature which we need to predict according to it. Here in Cascaded GMM [2] as shown in Figure 3 firstly, in Model 1 the test whisper voice is taken as source vector and the whispered speech's feature vector as target one, we should map the source features to target features by maximum-likelihood estimation of a spectral parameter trajectory [12]. The obtained predicted speech vector is used to Model 2, the predicted whispered speech feature vectors are now used as source vector and the normal speech vector as the target vector. Hence the abnormal speech is predicted using normal speech.

2.5 Reconstruction

In this step, we are going to use the trained parameters from the Gaussian Mixture Model to map to the speech signal under test. The test whisper signal is analyzed using Linear Predictive Coding, the residual signal is used to reconstruct with the trained parameters from training model 1. The obtained Whisperised signal is analyzed and the residual signal is used to map the training parameters from the training model 2 to obtain the reconstructed speech. Usually, the STRAIGHT toolbox is used for synthesis it increases the complexity of the system. By using LPC technique frame to frame synthesis of speech is possible and it is not much of complex.

2.6 Signal to noise ratio

To evaluate the performance of the reconstruction SNR ratio is used as given in

$$SNR = 10 * log\left(\frac{rms(normal\ speech)}{rms(Reconstructed\ speech)}\right) \quad (19)$$

3. RESULTS AND ANALYSIS

Figure 4 shows the conversion of normal speech to artificially synthesized one called whispered speech as given in section 2.1 and 2.2 speech analysis and synthesis.

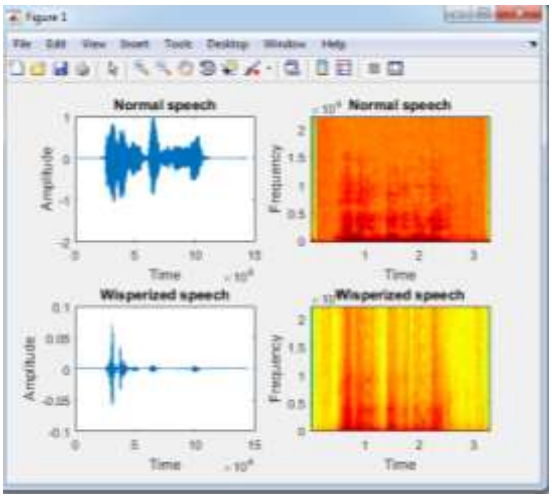

Fig. 4: Waveform and spectrogram of Normal speech and Whispered Speech

3.1 MFCC Coefficients obtained

Table 1: MFCC coefficients of the data used

Speech data	MFCC coefficients
Speech Impaired subjects	4983
Parallel normal subjects speech	4361
Synthesized Intermediate Speech	4361
Wtimit Whisper speech	6899
Wtimit Normal speech	5752
Synthesized Intermediate Speech	5752

3.2 Dynamic time alignment

Table 2: MFCC coefficients after DTW

Speech data	MFCC coefficients
Speech Impaired subjects	4361
Parallel normal subjects speech	4361
Synthesized Intermediate Speech	4361
Wtimit Whisper speech	5752
Wtimit Normal speech	5752
Synthesized Intermediate Speech	5752

3.3 Reconstruction after GMM training and mapping

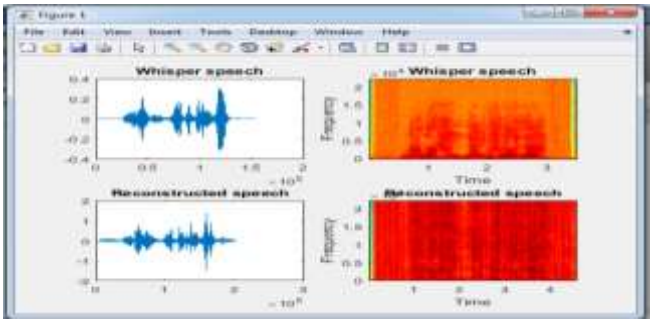

Fig. 5: Input Wtimit whisper speech and output reconstructed speech

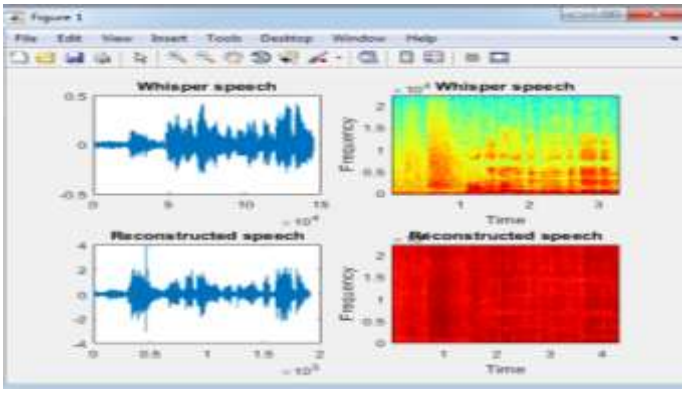

Fig. 6: Input Impaired speech and output Reconstructed speech

3.4 SNR calculation

Table 3: SNR ratio for Wtimit and Speech Impaired data

Speech	Speech	SNR in dB	SNR in dB
Whisper 1	Impaired 1	10.76	6.9
Whisper 2	Impaired 2	10.98	8.86
Whisper 3	Impaired 3	10.19	6.89
Whisper 4	Impaired 4	9.48	6.58
Whisper 5	Impaired 5	9.94	7.8

The Signal to noise ratio value is low. The objective measures used considers the lowest distance between normal and reconstructed [2]. Hence there is an improvement in the work.

4. CONCLUSION

Reconstruction of abnormal speech to normal sounding is performed. Initially, normal speech is used to construct the whispered speech to obtain time-aligned speech similar to normal speech. The MFCC features are extracted from all the three signals and used the train the cascaded Gaussian Mixture Model. From the trained parameters the whispered speech is reconstructed. The obtained result of the work has a lower SNR ratio.

5. REFERENCES

[1] H. Irem Turkmen and M. Elif Karsligil Dysphonic Speech Reconstruction, Construction of a Novel System for Effective and Efficient Communication IEEE ENGINEERING IN MEDICINE AND BIOLOGY MAGAZINE, 2010.

[2] A training-based speech regeneration approach with cascading mapping models Hamid R. Sharifzadeh, Amir HajiRassouliha, Ian V. McLoughlin, Iman T. Ardekani, Jacqueline E. Allen, Abdolhossein Sarrafzadeh Computers and Electrical Engineering 62 (2017) 601–611 0045-7906/© 2017 Elsevier Ltd.

[3] Ian McLoughlin, Jingjie Li, Yan Song, Hamid R. Sharifzadeh, "Speech reconstruction using a deep partially supervised neural network", Healthcare Technology Letters, 2017, Vol. 4, Iss. 4, pp. 129–133 (2017).

[4] Hamid R. Sharifzadeh, Amir HajiRassouliha, Ian V. McLoughlin, Iman T. Ardekani, Jacqueline E. Allen "Phonated Speech Reconstruction Using Twin Mapping Models", IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) (2015).

[5] Jing-Jie Li, Ian V. McLoughlin, Li-Rong Dai and Zhen-Hua Ling, "Whisper-to-speech conversion using restricted Boltzmann machine arrays" ELECTRONICS LETTERS 20th November 2014 Vol. 50 No. 24 pp. 1781–1782 (2014).

[6] Ian Vince McLoughlin, Jingjie Li, Yan Song, "Reconstruction of continuous voiced speech from whispers" Proc. Interspeech, August 2013, pp. 1022–1026 (2013).

[7] Ling-Hui Chen, Zhen-Hua Ling, Yan Song, Li-Rong Dai, "Joint Spectral Distribution Modeling Using Restricted Boltzmann Machines for Voice Conversion", Proc. Interspeech, Lyon, France, pp.3052-3056 (2013).

[8] Christophe Veaux, Junichi Yamagishi, Simon King, "Towards Personalized Synthesized Voices for Individuals with Vocal Disabilities: Voice Banking and Reconstruction" SLPAT 2013, 4th Workshop on Speech and Language Processing for Assistive Technologies, pages 107–111, (2013).

[9] Tomoki Toda, Member, Mikihiro Nakagiri, and Kiyohiro Shikano, "Statistical Voice Conversion Techniques for Body-Conducted Unvoiced Speech Enhancement", IEEE Transactions On Audio, Speech, And Language Processing, Vol. 20, No. 9, November 2012.

[10] Boon Pang Lim (2010) "Computational differences between whispered and non whispered speech," Ph.D. dissertation, University of Illinois, 2010.

[11] Hamid R. Sharifzadeh, Ian V. McLoughlin, F Ahmadi(2009), "Regeneration of Speech in Voice loss patients", ICBME, Proceedings 23, pp 1065-1068,2009.

[12] Tomoki Toda, Alan W. Black and Keiichi Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory", IEEE Transactions On Audio, Speech, And Language Processing, Vol. 15, No. 8, November (2007).

[13] McLoughlin I.V.: 'Speech and audio processing: a MATLAB-based approach' (Cambridge University Press, 2009).

[14] Davis, S. Mermelstein, P. (1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357-366.

