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

# Deep Neural Networks for Voice Quality Assessment based on the GRBAS Scale

*Simin Xie* [1, 2], *Nan Yan* [2], *Ping Yu* [3], *Manwa L. Ng* [4], *Lan Wang* [2], *Zhuanzhuan Ji* [2]

[1] School of Information Engineering, Wuhan University of Technology, Wuhan, China
[2] CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences/ The Chinese University of Hong Kong, Shenzhen, China
[3] Department of Otorhinolaryngology Head and Neck Surgery, Chinese People's Liberation Army General Hospital, Beijing 100853, China
[4] Speech Science Laboratory, University of Hong Kong, China

nan.yan@siat.ac.cn

## Abstract

In the field of voice therapy, perceptual evaluation is widely used by expert listeners as a way to evaluate pathological and normal voice quality. This approach is understandably subjective as it is subject to listeners' bias which high inter- and intra-listeners variability can be found. As such, research on automatic assessment of pathological voices using a combination of subjective and objective analyses emerged. The present study aimed to develop a complementary automatic assessment system for voice quality based on the well-known GRBAS scale by using a battery of multidimensional acoustical measures through Deep Neural Networks. A total of 44 dimensionality parameters including Mel-frequency Cepstral Coefficients, Smoothed Cepstral Peak Prominence and Long-Term Average Spectrum was adopted. In addition, the state-of-the-art automatic assessment system based on Modulation Spectrum (MS) features and GMM classifiers was used as comparison system. The classification results using the proposed method revealed a moderate correlation with subjective GRBAS scores of dysphonic severity, and yielded a better performance than MS-GMM system, with the best accuracy around 81.53%. The findings indicate that such assessment system can be used as an appropriate evaluation tool in determining the presence and severity of voice disorders.

**Index Terms**: voice quality, automatic assessment**,** DBN, MLP, GRBAS

## 1. Introduction

Speech production is a complex physiological process in which the larynx serves the very important function of phonation [1]. With the increase in voice use during social interaction, voice disorders, known as dysphonia, as a result of extensive or improper voice use, are becoming more common. This seriously affects the physical and psychological well-being of the voice users [2]. As a first step in managing dysphonia, it is crucial to correctly detect its presence and severity. Currently, two dominant approaches in diagnosing dysphonia and evaluating voice quality are used: perceptual analysis and acoustic analysis.

The use of perceptual analysis in assessing a pathological voice has been an important element of clinical diagnosis of dysphonia, and it has a definite role in the design of an appropriate therapeutic regimen [3, 4]. The GRBAS (Grade, Breathiness, Asthenia, Strain) scale has been serving as a standard for practicing voice clinicians which is recommended by the Japan Society of Logopedics and Phoniatrics [5, 6]. It has been widely recognized as a gold standard for effective and reliable perceptual voice quality evaluation. When using the scale, clinicians provide a score of 0, 1, 2, or 3 for GRBAS traits of the voice perceived, with a "0" representing perceived normal, "1" slightly, "2" moderately, and "3" severely disordered quality. Apparently, such perceptual analysis is considerably subjective and can be unreliable. The accuracy of such rating depends strongly on the experience level and the amount of bias of the assessor. Although the assessor is usually well-trained professional, his/her psychophysical condition and some other subjective factors might affect the results [3]. Perceptual analysis appears to be more accurate and reliable in distinguishing between normal and severely dysphonic voices, but it seems not as sensitive in ranking severity of disordered voices.

Acoustic analysis using signal processing techniques, on the other hand, offers an objective and effective method in early detection and diagnosis of pathological voices, on top of its relatively low cost and non-invasive nature [6]. Attempts have been made to evaluate voice quality by using acoustical measures such as traditional perturbation measures, nonlinear dynamical analysis and cepstral analysis [7, 8]. Using these acoustic parameters, researchers have become increasingly interested in developing automatic detection and rating of pathological voices. They were done by means of acoustic analysis, non-parametric feature extraction, pattern recognition algorithm and various statistical methods [9-13]. Currently, a new array of parameters based on Modulation Spectrum (MS) which were proposed to characterize perturbations of the human voice, with its objective basis to help clinicians detect pathological voice or to perform an automatic pathology classification [14]. An efficiency of 81.6% was obtained for Grade (G) of selected samples using MS and Gaussian Mixture Models (GMM) classifiers [14].

There are several classification methods used to investigate the effect of voice signal in perceptual ratings and automatic evaluation GRBAS Grade trait. Previous studies have revealed that automatic detection of voice impairments can be carried out by means of Multilayer Perceptron (MLP),



http://dx.doi.org/10.21437/Interspeech.2016-986

Support Vector Machine (SVM), Classification and Regression Tree (CART), Learning Vector Quantization (LVQ), using the well-known Mel Frequency Cepstral Coefficients (MFCCs) acoustical measures based on the GRBAS ratings [11-13].

Although the feasibility of an automated assessment system has been confirmed, improving the classification result is still needed in order for a better and more meaningful clinical application to be made. In the parameterization stage, Smoothed Cepstral Peak Prominence (CPPS) and Long-Term Average Spectrum (LTAS) were adopted to complement the classical MFCCs measures for the advantage of running speech [15-16]. Furthermore, most studies of voice quality focused on sustained vowels instead of running speech. To the best of our knowledge, few of them offered an assessment of voice quality for running speech, limiting their applicability of voice evaluation in the practical realm. The major goal of the present study was to apply the complementary automatic assessment system for voice quality based on running speech. In this case, both CPPS and LTAS parameters can be used as appropriate assessment parameters.

The purpose of this work was to evaluate the pathological voice quality using multidimensional acoustical parameters extracted from running speech samples through DBN-DNN classifier according to the GRBAS scale. Using DNNs for adaptive acoustic model for handling large-scale running speech recognition has attracted extensive attention [17-19].On the one hand, in a continuous space, DBN-DNN can simulate the complicated distribution without making an artificial hypothesis of distribution and help to extract the phonetic characteristics of distributed simulation and has a powerful discriminative for features[17]. On the other hand, DBNs has been applied in the visual attributes learning, so as to effectively improve the multi-level classification performance and generalization accuracy of classifier [17, 19]. Considering these advantages of DBN-DNN, it is believed that it should be suitable for dealing with attribute description and objective evaluation of pathological voice quality.

In the present study, a four-class classification for pathological voice quality based on the well-known “Grade” parameter of GRBAS scale was proposed. A multidimensional features battery including MFCCs, CPPS and LTAS was extracted from running speech samples. Deep Neural Networks (DNNs) was utilized as the classification model which adopted the Deep Belief Nets (DBN) for pre-training and Multi-layer Perceptron (MLP) for fine-tuning. The assessment system based on sustained vowels using MS features and GMM classifier was also employed and used as baseline set.

# 2. Methods

## 2.1. Database

The database used in the present study contained sustained vowel and running speech samples selected from the North Chinese Corpus. For each voice, two recordings were available: production of a sustained vowel /a/ and a short running speech. There are 3162 voice samples for running speech and 732 voice samples for sustained vowel produced by 260 subjects with voice disorders and 106 (48 females and 58 males) normal controls. The disordered voice samples were obtained from 117 females and 143 males, of an age range of 18-60 years recorded by the Department of Otorhinolaryngology of the People’s Liberation Army General Hospital. And the voice samples represented a wide variety of different vocal pathologies, including various types of vocal fold lesions, vocal cord paralysis, arytenoid granuloma, vocal cord pre-cancerous lesions and vocal cord carcinoma. Voice samples were recorded in a professional recording studio with an acoustic sensor (B&K 4189), which was sheathed with a wind-shelter to block the airflow to avoid recording of unnecessary noise. The voice signals were sampled at 22.05 kHz with 16 bits/sample resolution. All speech recordings arranged in three different orders which were perceptually evaluated by five practicing speech pathologists, who rated each recording using the GRBAS scale.

## 2.2. Features extraction

Two parameterization batteries are considered in this study, including multidimensional features extracting from running speech and MS parameters from sustained vowels. For the former, the MFCCs, CPPS and LTAS were included.

### *2.2.1. Mel-Frequency Cepstral Coefficients*

For nonlinear characteristic parameters, the MFCCs describe the energy distribution of a signal in the frequency domain and refer to perceived frequency. They have been widely used as features in automatic speech recognition and in assessing pathological voice quality [7, 8, 11-13]. Taking into account the characteristics of human auditory perception, MFCCs were estimated using a nonparametric FFT-based approach. A 39-dimensional MFCCs feature vector consists of log energy, 12 mel frequency cepstral coefficients, the first-order derivatives and the second-order derivatives of these 13 static features. They were extracted by using Hcopy tools in Hidden Markov Model Toolkit (HTK) [20].

### *2.2.2. Smoothed Cepstral Peak Prominence*

The CPPS measure represents the distance between the first harmonic peak and the point with equal frequency on the regression line through the smoothed cepstrum. This parameter indexes the periodic attribute of voice signal. The more periodic is a voice signal, the more prominent will the cepstral peak be [15]. Consequently, CPPS has been used to reliably evaluate the dysphonia severity in both sustained vowels and running speech samples. In this study, CPPS parameter was estimated using the smoothing algorithm and linear regression analysis.

### *2.2.3. Long-Term Average Spectrum*

LTAS analysis provides spectral information averaged over a long period of time, thereby highlighting aspects of speech or singing voice production over a longer temporal span. One way to quantify the LTAS output is by means of spectral moment analysis, where the shape of the speech spectrum is described according to the spectral mean, standard deviation, skewness, and kurtosis [16] of the curve. In the present study, LTAS parameters were estimated using a method based on a single transformation followed by spectrum size reduction. Fast Fourier transform (FFT) and a uniformly spaced filter bank were used to calculate the LTAS parameters.

### *2.2.4. Measures of modulation spectra*

Modulation spectra provide a visual representation of sound energy spread in the space of acoustic by modulation. They

provide information about perturbations related to amplitude and frequency modulation of the voice signal. With a group of well-defined parameters including Centroids, Dynamic Range per Band, Low Modulation Ratio, Contrast and Dispersion parameters, MS has been used for detecting the presence of dysphonia and the degree of disorder [14]. In this study, the MS parameters were calculated using the Modulation Toolbox library version 2.1 [21], and MS parameters were obtained by calculating mean value from all frames of the input signal, except the Centroids.

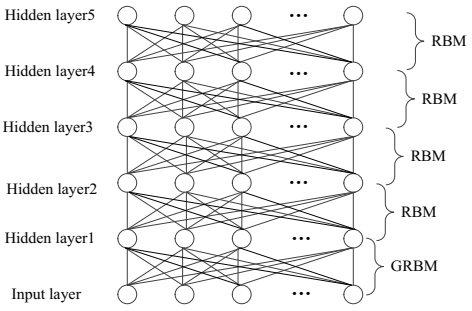
For running speech, the multidimensional features were extracted from each sample of one to three seconds long with 25-ms interval overlapping 10-ms and multiplied by a hamming window. For sustained vowels, only a 1-s segment extracted from the vowel was used for analysis, which was framed and windowed using hamming windows overlapped 50%. The window length was 40ms in 20ms step.

## 2.3. Classification algorithm

The DBN-DNN consisted of DBN and MLP, with DBN used to pre-train the offset value and weight in an unsupervised manner, and MLP used by Bach-Propagation (BP) algorithm for training in classification. A classical supervised learning based on the Gaussian Mixture Model (GMM) paradigm was employed for comparison.

### *2.3.1. DBN model*

DBN is a generative model of probability which is constructed from multiple layers of hidden units. It can be seen as simple learning modules that make up each layer of the Restricted Boltzmann Machine (RBM) [22]. In the present study, the Contrastive Divergence (CD) was used as the optimization method for RBMs [23]. As the input data in the experiment were real and were obtained from real human productions, the energy function of the RBMs was suitable for binary data, the study used the Gaussian–Bernoulli restricted Boltzmann machine (GRBM) [23] instead of the first RBM (Constituted by the input layer and hidden layer1). The DBN-DNN systems were pre-trained using the multidimensional acoustic features. These were used as the input feature to train DBN with 512 neurons in each hidden layer using the Kaldi toolkit [24]. Figure 1 shows the DBN schematic representation that adopted by the experiment.




Figures 1 : *DBN structure*

### *2.3.2. DBN training*

The multidimensional features were fed into the DBN-DNN classification as the input features. In order to release the noise, all data were divided into several mini-batches. The updated value mini-batch was averaged after they were being updated though the stochastic gradient descent algorithm and one-step Contrastive Divergence approximation. To update the parameters according one mini-batch, then repeated until all of mini-batches were updated, and after this the machine has completed an epoch. The training methods of GRBM and RBM were similar. Briefly, the DBNs were first pre-trained with the training dataset in an unsupervised manner though training the layer-layer RBMs by using the Kaldi toolkit [24]. It was then followed by the supervised fine-tuning using the same training dataset and the validation dataset to do early-stopping. The selected DBN model is described in Table 1.

Table 1. *DBN model configuration*

| Configuration item | value |
|---|---|
| number of hidden layers | 5 |
| neurons per layer | 512 |
| learning-rate for RBMs | 0.2 |
| learning-rate for GRBM | 0.01 |
| Mini-batch size | 256 |

### *2.3.3. MLP model and training*

MLP is a popular example of feed-forward neural network which can cope with the non-linearly separable problems, as each neuron of the MLP uses a nonlinear activation function [25]. The input features using the same training dataset as in DBN training. The MLP structured included a input layer, five hidden layers (same as DBN) and a output layer. Additionally, each layer of MLP was fully connected to the next level. In the study, the pre-trained result of DBN included weights and offset values were used as the initial data of MLP network; four neurons mapping to the 'G' grade of GRBAS scale though the softmax function [22].

The Back-propagation (BP) algorithm was used in the training of MLP network, as it can calculate the gradient of the error of the network regarding the network's modifiable weights. The steps of MLP training and procedures about the BP algorithm has been described in detail in the literature [23]. In the MLP model, the learning rate was set to 0.008, the outputs were obtained though the softmax layer. The number of hidden layers, mini-batch size and the neurons of hidden layers were the same as DBN setting.

### *2.3.4. GMM training*

Having a data vector of dimension resulting from the features extraction, a GMM is a model of the probability density function defined as a finite of multivariate Gaussian components of the form：

$$\rho(\mathrm{x}|\Theta_{\mathrm{i}}) = \sum_{\mathrm{r}=1}^{\mathcal{G}} \lambda_{\mathrm{r}}\, N(\mathrm{x}; \mu_{\mathrm{r}}, \sum_{\mathrm{r}}) \quad (1)$$

where $\lambda_{\mathrm{r}}$ are scalar mixture weights, $N(\cdot)$ are Gaussian density with mean $\mu_{\mathrm{r}}$ of dimension $d$ and convariances $\sum_{\mathrm{r}}$ of dimension $d \times d$. For each class (i.e., values of the G perceptual levels: 0,1,2, or 3 ), GMM training is performed by estimating the abovementioned parameters with expectation-maximization algorithm (EM) [26]. And a threshold is fixed at the Equal Error Rate (EER) point prior to the final decision. With the aim of an accurate comparison in the state of art, the GMM classifier was used along with MS parameters.

## 3. Results

To evaluate the performance of classification, a 5-fold cross validation scheme [27] was used, in which 85% of dataset were used to train the classifier, 5% for validation and the remaining 10% for testing. The global result for a certain parameterization experiment is the average of the results in all folds. Assessment of the classifier performance was performed in terms of efficiency, the Confidence Interval (CI) [28] and Cohen's Kappa Index (K) [29]. This last indicator provides information about the agreement between results of the classifier and clinician's perceptual labeling. The number of Gaussian components of the GMM was varied from 4 to 48 in order to identify the best performance of classifier. The best results of MS-GMM system for our database were obtained using MS parameters, 20 ms frames, and 8 GMM.

Table 2 shows that the efficiency and Cohen's Kappa Index of G class obtained with DBN-DNN using MFCC only and the multidimensional feature battery, while that of MS-GMM system and MFCC-GMM system are shown as the baseline system for comparison. The best results (81.53%) were obtained using DBN-DNN classifier with multidimensional parameters and value of Kappa index is up to 0.76, indicating that automatic assessment appeared to match well with perceptual assessment. For each G class, similar to total grade classification, better performance was observed, except for G0 class (see Table 3). Growth of the efficiencies was around 1%, the performance of G3 class was markedly improved. In each classifier, the G3 class showed the best performance, followed by G0 class. The G1 and the G2 class exhibited worse classification, which indicate that it is difficult to discriminate between G1 and G2 dysphonia severity.

Table 2. *Results expressed as efficiency ± confidence interval and Cohen's Kappa Index for G class*

| Features | Classifiers | Efficiency (%) | Kappa |
|---|---|---|---|
| Multi-dimensions | DBN-DNN | 81.53 ± 4.21 | 0.761 |
| MFCC+Δ + ΔΔ | DBN-DNN | 80.61 ± 4.3 | 0.734 |
| MFCC+Δ + ΔΔ | GMM | 78.91 ± 4.9 | 0.692 |
| MS | GMM | 73.02 ± 10.61 | 0.687 |

Table 3. *Results expressed as efficiency ± confidence interval for each G subclass*

| G class | Efficiency (%) | | |
|---|---|---|---|
| | DBN-DNN | | GMM |
| | MFCC+Δ + ΔΔ | Multi-dimensions | MS |
| G0 | 83.56 ± 8.62 | 81.55 ± 8.34 | 88.14 ± 20.0 |
| G1 | 74.45 ± 9.43 | 75.87 ± 9.37 | 69.67 ± 19.3 |
| G2 | 79.51 ± 8.84 | 80.51 ± 8.57 | 66.18 ± 17.1 |
| G3 | 84.31 ± 7.71 | 89.08 ± 6.83 | 93.63 ± 17.7 |

## 4. Discussion

The present study developed a new automatic system based on the DBN-DNN model to emulate perceptual assessment of voices according to the G feature of the GRBAS scale, using multidimensional features battery based on MFCCs, CPPS and LTAS. The outcomes were compared to the state-of-the-art MS-GMM system. Better performance was obtained with the proposed system, providing 81.53% of efficiency and 0.76 Cohen's Kappa Index. And the agreement with perceptual assessment can be considered as matching well. To the best of our knowledge, this is a first application by using DNNs classifier to utilize such assessment system. Although an accurate comparison with the previous studies is difficult due to the use of different corpus and methodologies, results of the present study promoted a more convincing comparison based on the same database.

Unlike the present study, previous investigations offered no assessment of voice quality associated with running speech samples, mainly because vowels are easily elicited and less affected by articulation and dialectal influences. However, running speech is more representative of a person's daily voice use, and it is an important part of perceptual voice evaluation. The fact that the proposed system yielded a higher quality compared to MS-GMM system for pathological running speech implies that running speech gives more valuable information to assess the degree of dysphonia than sustained vowels.

It is noticeable that the results of total grade classification based on DBN-DNN classifier are much better than GMM classifier. Moreover, DBN-DNN classifier tended to work better in terms of class G1 and G2. The reasoning behind this phenomenon is that the risk of misclassification between class G1 and G2 was reduced. Summary, the integrated accuracy (G in table 2, 3) from the classification and the accuracy of each class by the classifier showed that the present four-class assessment system using the DBN-DNN as a classifier can contribute to achieve a superior classification of pathological voice quality.

Regarding the use of multidimensional features, the improvements are noticeable. The present results showed high classification accuracy, suggesting that it was a meaningful and successful attempt. In future work, for multidimensional features, some other features such as complexity and noise measurements applied to running speech might be of interest. Moreover, a fusion model of automatic assessment based on running speech and sustained vowels should be considered to improve the accuracy of objective evaluation of voice quality.

## 5. Conclusions

The present results suggest that it is an appropriate assessment classifier for evaluating the presence and severity of disordered voices by using the DBN-DNN combing multidimensional acoustic parameters as a tool based on the GRBAS rating scale. Further studies should be involved more objective analyses and nonlinear features in running speech. Besides, more voice samples should be collected to train the DBN-DNN model. In conclusion, the system proposed can objectively, effectively, and reliably evaluate quality of pathological voices.

## 6. Acknowledgements

This study was jointly supported by a grant from National Natural Science Foundation of China (NSFC 61135003, 91420301 and 61401452), Shenzhen Speech Rehabilitation Technology Laboratory.

# 7. References


[1] Throat, *https://en.wikipedia.org/wiki/Throat*

[2] E. Mendoza, and G. Carballo, "Vocal tremor and psychological stress," *Journal of Voice*, vol. 13, no. 1, pp. 105-112, 1999.

[3] J. Kreiman, and B. R. Gerratt, "Listener experience and perception of voice quality," *Journal of Speech and Hearing Research*, vol. 33, no. 1, pp. 103-115, 1990.

[4] C. R. Rabinov, J. Kreiman, B. R. Gerratt, and S. Bielamowicz, "Comparing reliability of perceptual ratings of roughness and acoustic measures of jitter," *Journal of Speech and Hearing Research*, vol. 38, no. 1,pp. 26-32, 1995.

[5] P. Carding, E. Carlson, R. Epstein, L. Mathieson, and C. Shewell, "Formal perceptual evaluation of voice quality in the United Kingdom," *Logopedics, Phoniatrics, Vocoloy,* vol. 25, no. 3, pp. 133-138, 2000.

[6] J. Kreiman, B. R. Gerratt, K. Precoda, and G. S. Berke, "Individual differences in voice quality perception," *Journal of Speech and Hearing Research*, vol. 35, no. 3, pp. 512-520, 1992.

[7] N. Yan, M. L. Ng, D. Wang, L. Zhang, V. Chan, and R. S. Ho, "Nonlinear dynamical analysis of laryngeal, esophageal, and tracheoesophageal speech of Cantonese," *Journal of Voice*, vol. 27, no. 1, pp. 101-10, 2013.

[8] S. S. Liu, N. Yan, M. L. Ng, L. Wang, and Z. J. Wang, "Multidimensional acoustic analysis for evaluation of voice quality of unilateral vocal fold paralysis," in *Information Science and Technology, 2014 – 4th IEEE International Conference on Information Science and Technology*, April 26-28, Shenzhen, Guangdong, China, 2014, pp. 706-709.

[9] M. Hariharan, K. Polat, R. Sindhu, and S. Yaacob, "A hybrid expert system approach for telemonitoring of vocal fold pathology," *Applied Soft Computing*, vol. 13, no. 10, pp. 4148-4161, 2013.

[10] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "An improved method for voice pathology detection by means of a HMM-based feature space transformation.Pattern recognition," *Pattern Recognition*, vol. 23, no. 9, pp. 3100-3112, 2010.

[11] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, M. Blanco-Velasco, and F. Cruz-Roldán, "Automatic assessment of voice quality according to the GRBAS scale," in *Engineering in Medicine and Biology Society, 2006. EMBS 2006 –28th Annual International Conference of the IEEE*, pp. 2478-2481, 2006.

[12] P. Yu, Z. Wang, S. Liu, N. Yan, L. Wang, and M. Ng, "Multidimensional acoustic analysis for voice quality assessment based on the GRBAS scale," in *Chinese Spoken Language Processing (ISCSLP), 2014–9th International Symposium on*, pp. 321-325, Sep. 2014.

[13] Z. Wang, P. Yu, N. Yan, L. Wang, and M. Ng, "Automatic assessment of pathological voice quality using multidimensional acoustic analysis based on the grbas scale," *Journal of Signal Processing System*, vol. 82, no. 2, pp. 241-251, 2016.

[14] L. Moro-Velázquez, J.A. Gómez García, J.I. Godino-Llorente, G. Andrade-Miranda, "Modulation spectra morphological parameters: A new method to assess voice pathologies according to the GRBAS scale,", *BioMed Research Internationa*l, Article ID 259239, 2015.

[15] Y. Maryn, and D. Weenink, "Objective dysphonia measures in the program praat: Smoothed cepstral peak prominence and acoustic voice quality index," *Journal of Voice*, vol. 29, no. 1, pp. 35-43, 2015.

[16] R. Fraile, et al. "Characterization of dysphonic voices by means of a filterbank-based spectral analysis: Sustained vowels and running speech," *Journal of Voice*, vol. 27, no. 1, pp. 11-23, 2012.

[17] JI. Godino-Llorente, P . Gómez-Vilda , "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Trans Biomed Eng,* Vol. 51, no. 2, pp. 380-384. 2004.

[18] Y. Maryn, P. Corthals, P. Van Cauwenberge, N. Roy, and M. De Bodt, "Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels," *Journal of Voice*, vol. 24, pp. 540-55, 2010.

[19] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput*., vol. 18, pp. 1527–1554, 2006.

[20] S. Young, G. Evermann, M. J. F Gales, D. Kershaw, G. Moore, J. J. Odell, D. G. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book Version 3.4.1,* 2009.

[21] L. Atlas, P. Clark, and S. Schimmel, "Modulation Toolbox Version2.1forMATLAB,"*https://isdl.ee.washington.edu/projects /modulationtoolbox/*

[22] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," *Lecture Notes in Computer Science*, *Neural Networks: Tricks of the Trade*, vol. 7700, pp. 599-619, 2012.

[23] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kinsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, pp. 2-17, 2012.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, and N. Goel, *The Kaldi speech recognition toolkit*. Idiap, 2011.

[25] H. Wang, L. Wang and X. Liu, "Multi level adaptive network for accented Mandarin speech recognition", in *Proc. IEEE ICIST*, Shenzhen, China, pp. 602–605, 2014.

[26] T. K. Moon, "The expectation-maxmization algorithm," *IEEM Signal Processing Magazine*, vol. 13, no.6, pp. 47-60, 1996.

[27] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification.* New Jersey: John Wiley & Sons, 1999.

[28] J. Ferreiros, and J. M. Pardo, "Improving continuous speech recognition in Spanish by phone-class semicontinuous HMMs with pausing and multiple pronunciations," *Speech Communication*, vol. 29, no. 1, pp. 65-76, 1999.

[29] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.