Deep Neural Networks for Shimmer Approximation in Synthesized Audio Signal

Mario Alejandro García and Eduardo Destefanis

Universidad Tecnológica Nacional Facultad Regional Córdoba, Argentina
mgarcia@frc.utn.edu.ar

Abstract. Shimmer is a classical acoustic measure of the amplitude perturbation of a signal. This kind of variation in the human voice allow to characterize some properties, not only of the voice itself, but of the person who speaks. During the last years deep learning techniques have become the state of the art for recognition tasks on the voice. In this work the relationship between shimmer and deep neural networks is analyzed. A deep learning model is created. It is able to approximate shimmer value of a simple synthesized audio signal (stationary and without formants) taking the spectrogram as input feature. It is concluded firstly, that for this kind of synthesized signal, a neural network like the one we proposed can approximate shimmer, and secondly, that the convolution layers can be designed in order to preserve the information of shimmer and transmit it to the following layers.

Keywords: shimmer, voice quality, deep learning, deep neural network, convolutional neural network

1 Introduction

Shimmer is a classical acoustic measure of the amplitude perturbation of a signal. This kind of variations in the human voice allows to characterize some properties, not only of the voice itself, but on the person who speaks [1].

Shimmer value is associated to voice quality [2–7], state of mind [8–13], age [14] and gender [15] of people. There are many research works that use shimmer (among other measures) with goals ranging from pathologies detection [6, 16, 17] to the improvement of human-machine interfaces through the estimation of the intensionality of a spoken phrase [19]. Regarding synthesized voices, Yamasaki et al. show in [18] that a certain shimmer level increases the degree of naturalness.

The application of deep learning techniques is the state of the art in automated audio analysis, with the detection of pronounced phonemes and the identification of the person that speaks as main objectives [20–26], but also used to detect emotions, age, gender, etc. [27–33].

Classifiers based on neural networks can be divided into two groups according to the type of input features, those using previously calculated acoustic measures [10, 14] and those using raw audio [22, 24, 25] or spectral data [21–23, 28–31, 34, 35]. In [26] a hybrid approach is applied by adding shimmer and other measures

to improve the recognition achieved with spectral data. It is important to clarify, for first group of classifiers, that shimmer calculation has a major complication, it depends on the previous detection of the fundamental frequency (f_0) of vocal cords vibration. It is difficult to estimate f_0 in pathological voices [36, 37]. The estimation of the actual f_0 value is still a research topic [36–40]. Regarding the second group of classifiers, it is not possible to know whether the outputs are influenced by the shimmer value of the signal.

1.1 Objectives

The objective of this work is to make an estimation of the value of shimmer in a synthesized audio signal through a neural model. The neural network must combine convolutional layers and feed forward layers. The inputs to the neural model will be the spectral values of the signal.

The main contributions of design a shimmer estimation neural network from the spectral features of an audio signal are, on the one hand, the procurement of a f_0 independent shimmer calculation method, and on the other, to answer the question about the extent to which amplitude disturbances of the original audio can influence the output of a deep learning model with raw audio or spectral data input. In other words, how much shimmer information is preserved to the last layers of the model.

1.2 Shimmer Calculation

There are different versions of shimmer. The most important difference between them is the window size (number of f_0 cycles) used in the calculation. Some versions can be seen in [41].

The chosen version of shimmer for this work is the proposed by Klingholz and Martin [42], also known as *Relative Shimmer*.

Relative Shimmer, hereinafter referred to as "shimmer", is a way of measuring the cycle-to-cycle amplitude perturbations of the fundamental frequency of a signal. It is shown as a *perturbation/total amplitude* relation.

$$shimmer = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N}|A_i|} \tag{1}$$

where N is the number of periods of f_0 that the signal has and A_i is the maximum amplitude into the i period.

2 Methods and Materials

2.1 Neural Models

Deep learning models with ascending complexity were generated for problems of shimmer approximation. First, shimmer was approximated for f_0 variable, k

constant and f_{mod} constant. Then, shimmer was approximated for f_0 variable, k variable, and f_{mod} constant. Finally, a model was found to approximate shimmer with f_0, k and f_{mod} variable.

In all cases, spectral audio data (instead of raw audio) were used as input features. There are two reasons, the improvement of training performance due the dimension reduction, and because in this manner models work in a similar way as the human auditory system, where spectral division is performed in the basilar membrane of the cochlea and not by the neurons of the auditory cortex [43].

2.2 Data

Audio Audio data without harmonics was generated. As in [1] the amplitude modulation of human voice was approximated by a sinusoidal wave. The expression of each audio signal $y(t)$ was:

$$y(t) = \frac{1}{1+k}\sin(\alpha + 2t\pi f_0)(1 + k\sin(\beta + 2t\pi f_{mod})) \tag{2}$$

where t is time $[sec]$, f_0 is the frequency of vocal fold vibration $[Hz]$, f_{mod} is the modulatory frequency $[Hz]$, k is the constant of the amplitude modulator sensibility, α and β are constant to handle the phase of the signal to be modulated and the modulating signal respectively.

For the training, test and validation data generation, random values were taken with uniform distribution. f_0 got values in $[200, 1000]$ Hz range, f_{mod} in $[5, 10]$ Hz, k in $[0, 0.4]$, α and β in $[0, 2\pi]$.

250 ms of audio generated with $f_0 = 200$ Hz, $f_{mod} = 8$ Hz and $k = 0.4$. is shown in Fig. 1.

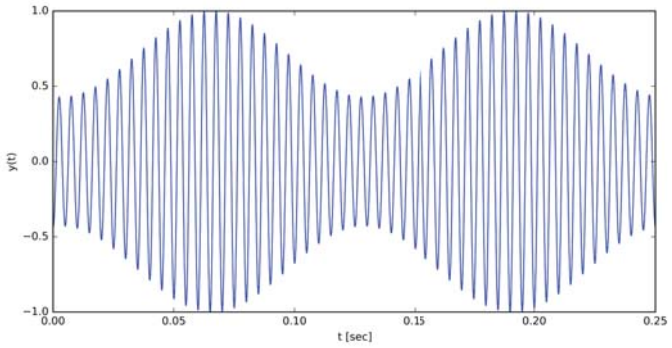

Fig. 1. Generated audio for $f_0 = 200$ Hz, $f_{mod} = 8$ Hz and $k = 0.4$

Training data To train each model, one 2500 elements training dataset, one 500 elements testing dataset and one 500 elements validation dataset were generated. Each element is composed by shimmer (Eq.(1)) value to be estimated and the spectrogram of generated audio.

Due to the fact that f_0 is known at the time of generating the audio, shimmer value can be calculated accurately.

The spectrogram is calculated on 2 *sec* of audio generated with 44100 $samples/sec$. For the calculation a $Tukey(0.25)$ window of width = 256 was used, which determines a structure of shape 129 x 393 (frequency/time) containing the spectral density of the signal.

Fig. 2 shows the values of the second, third and fourth rows (index 1 to 3) of the spectrogram of signal in Fig. 1.

Spectrograms data and shimmer data was normalized between 0 and 1.

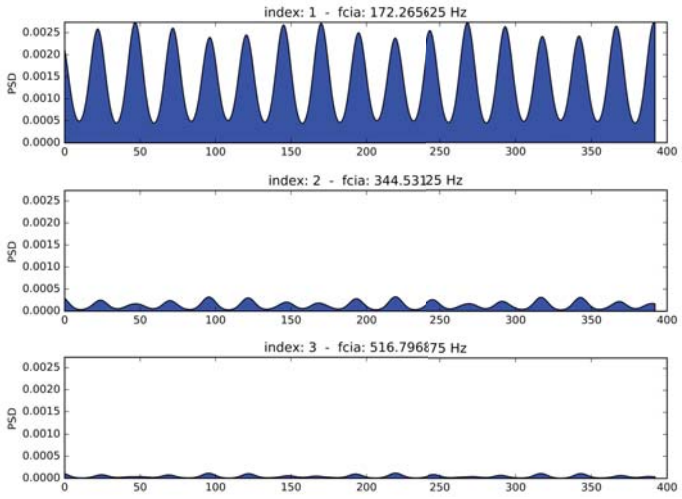

Fig. 2. Three rows with higher average value of Power Spectral Density (PSD) in spectrogram of audio generated with $f_0 = 200\ Hz$, $f_{mod} = 8\ Hz$ and $k = 0.4$

3 Results

An initial analysis was performed with f_0, f_{mod} and k known data. It was found that a neural network with dense connections is able to calculate shimmer value with high precision if it gets f_0, f_{mod} and k as input features. Optimal structure of this network was empirically found. This is a three layer network, two layers of 20 neurons with $tanh()$ activation function and a linear neuron as output. In next models, convolution layers are used at the initial part of the network, and then, dense layers of 20, 20 and 1 neurons. The function of convolution layers

is to calculate the values of f_0, f_{mod} and k in order to dense layers calculate shimmer.

It was noted that only the first 15 rows of the spectrogram (lower frequencies) would have significant information. Then, only for the scope proposed in this work, the rest of the frequencies were eliminated. Then, spectrogram shape change from 129 x 393 to 15 x 393. This provides a important performance improvement.

3.1 Shimmer approximation with f_0 variable

Without harmonics, the calculation of f_0 from the spectrogram is simple, it is enough to obtain the energy average value weighted by the frequency that each spectrogram row represents. As expected, a network such as Fig. 3, where each complete row of the spectrogram connects to a neuron of an *Average pooling* layer, is able to perform the weighted average of frequencies and calculate the shimmer value in densely connected layers. Tests were performed with f_0 in $[200, 1000]$ Hz range, $k = 0.4$ and $f_{mod} = 8$ Hz. A satisfactory approximation was achieved, with a mean square error (MSE) $< 10^{-4}$.

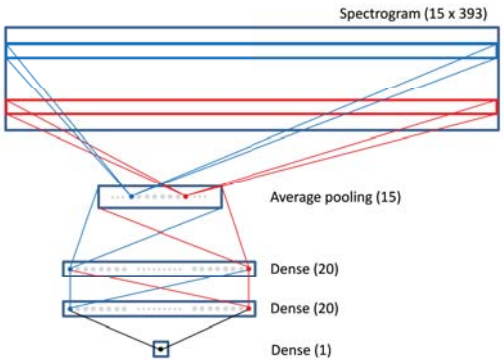

Fig. 3. Shimmer approximation model on signals with f_0 in $[200, 1000]$ Hz range, $f_{mod} = 8$ Hz and $k = 0.4$. Each neuron in the *Average pooling* layer has a complete frequency of the spectrogram as its visual field. The activation function of hidden dense layers neurons is $tanh()$ and the output neuron is linear.

3.2 Shimmer approximation with f_0 and k variable

The value of k affects inversely the area under the energy curve of the spectrogram.Therefore, information about the value of k can be obtained through the energy average of the spectrogram. The model presented in the previous section preserves the information necessary to estimate the energy average. Tests were

performed with audio data for f_0 in $[200, 1000]$ Hz range, k in $[0, 0.4]$ range and $f_{mod} = 8$ Hz. Results were satisfactory again. The model approximates shimmer with an MSE $< 10^{-4}$.

3.3 Shimmer approximation with f_0, k and f_{mod} variable

For f_0 in the range $[200, 1000]$ Hz, k in the range $[0, 0.4]$ and f_{mod} in the range $[5, 10]$ Hz it was necessary to create a more complex model than the previous one. Shimmer depends on the modulation frequency, so a new transformation is necessary (the first one was the transformation from time domain to frequency domain in the spectrogram). The new (second) transformation is performed in a convolution layer at the initial part of the model (Fig. 4).

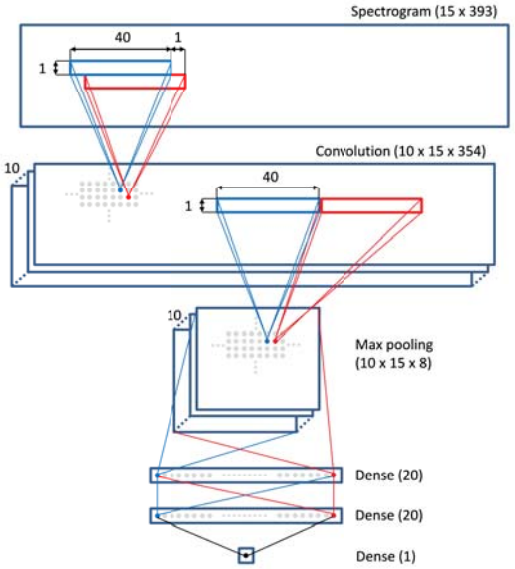

Fig. 4. Shimmer approximation model on signals with f_0 in the range $[200, 1000]$ Hz, k in the range $[0, 0.4]$ and f_{mod} in the range $[5, 10]$ Hz. The shape of convolutonal layer windows is 1 x 40, strides 1 x 1. Convolutional layer has 10 sub-layers. The shape of *max pooling* layer windows is 1 x 40, strides 1 x 40. The network finish with three dense layers of 20, 20 and 1 neurons.

Convolutional layer Each convolution layer neuron is connected to spectrogram through a $height = 1$ and $width = 40$ window. Convolution is performed on a single frequency ($height = 1$) so that the f_0 detail level needed in the dense layers is not lost. The 40-element width is the minimum required to hold a cycle

of $min(f_{mod})$. The number of elements of the spectrogram per modulation cycle (C) for a spectrogram of width W_s and an audio length L is:

$$C = \frac{W_s}{L \times min(f_{mod})} = \frac{393\, elements}{2\, sec \times 5\, Hz} = 39.3\, elements/cycle$$

The window displacement in both directions is 1 step. It imply that on the frequency dimension there is no overlap, and in time dimension there are 39 overlapping elements between the windows of adjacent neurons. Finally, according to these definitions, the shape of each convolution filter or sub-layer is 15 x 354. The convolution layer is formed by 10 sub-layers. This amount is a compromise between performance and the detail level of f_{mod} on the information sent to dense layers. Neurons of this layer have linear activation function. Weights are initialized with orthogonal random values. An attempt was made to initialize them with wavelet families for sinusoidal waves between $5Hz$ y $10Hz$, but no improvement was achieved on the prediction accuracy.

***Max pooling* layer** The neurons in the *max pooling* layer have a 1 x 40 window size on the convolutional layer. Again, $height = 1$ allows allows f_0 information be able to be transmitted to dense layers with no losing details. The 40-element width extends the visual field of this layer neurons to 2 cycles of $min(f_{mod})$ on the spectrogram. In this way, the output value is invariant to the modulation signal translations. There is no overlap between the windows, so the size of each of the 10 sub-layers is 15 x 8 neurons.

The outputs of *max pooling* layer are connected to three layers with dense connections equal to those of the previous model.

For this model, 20 training tests were performed. The size of training dataset was 2500 elements. In all cases, results were compared with a test dataset (500 elements) during training and a validation dataset (500 elements) at the end. The best result, with 150 training cycles, obtained a MSE $= 5.8 \times 10^{-5}$ on the test dataset. In Fig. 5 expected and calculated shimmer values are displayed in ascending order for the 500 elements of test dataset.

4 Conclusion

It was verified that, for simple audio signal modulated in amplitude by a sinusoidal wave, with variable parameters of fundamental frequency, modulating frequency and modulation sensitivity, it is possible to obtain a neural model able to approximate the value of shimmer.

Under the conditions presented in this paper, it is possible to calculate shimmer without knowing f_0. Moreover, it can be affirmed that if the first layers of a deep neural network respects the structure of the second presented model, this neural network is able to use the value of shimmer, internally calculated, to perform other classifications.

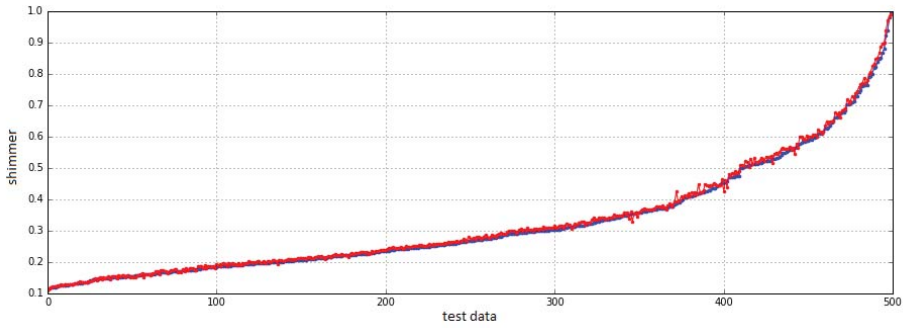

Fig. 5. Normalized shimmer. Expected (blue) vs. calculated (red) for elements in the testing dataset (in ascending order of shimmer value).

5 Future Works

It is planned to extend the analysis, first by expanding the ranges of f_0 y f_{mod}, then adding harmonics and noise to the synthesized signals. Finally, it is planned to analyze the behavior of deep learning models for shimmer calculation on natural voices.

References

1. Jafari, M., Till, J. A., Law-Till, C. B.: Interactive effects of local smoothing window size and fundamental frequency on shimmer calculation.Journal of Voice,vol. 7.3. pp. 235–241. (1993)
2. Nieto, R. G., Marín-Hurtado, J. I., Capacho-Valbuena, L. M., Suarez, A. A., Bolaños, E. A. B.: Pattern recognition of hypernasality in voice of patients with Cleft and Lip Palate. Image, Signal Processing and Artificial Vision. pp. 1-5. IEEE. (2014)
3. Holi, M. S.: A hybrid model for neurological disordered voice classification using time and frequency domain features.Artificial Intelligence Research, vol5.1, pp. 87. (2015).
4. Freitas, S. V., Pestana, P. M., Almeida, V., Ferreira, A.: Integrating voice evaluation: correlation between acoustic and audio-perceptual measures. Journal of Voice,vol 29.3, 390–e1. (2015)
5. Little, M. A., Costello, D. A., Harries, M. L.: Objective dysphonia quantification in vocal fold paralysis: comparing nonlinear with classical measures. Journal of Voice, vol 25.1. pp 21–31. (2011)
6. Lopes, L. W., Simões, L. B., da Silva, J. D., da Silva Evangelista, D., e Ugulino, A. C. D. N., Silva, P. O. C., Vieira, V. J. D.: Accuracy of Acoustic Analysis Measurements in the Evaluation of Patients With Different Laryngeal Diagnoses.Journal of Voice,vol. 31.3. pp. 382-e15. (2017)
7. Hillenbrand, J.: Perception of aperiodicities in synthetically generated voices.The Journal of the Acoustical Society of America, vol83.6. pp. 2361-2371. (1988)

8. Li, X., Tao, J., Johnson, M. T., Soltis, J., Savage, A., Leong, K. M., Newman, J. D.: Stress and emotion classification using jitter and shimmer features. ICASSP, vol. 4. pp. IV–1081. IEEE. (2007)
9. Kotti, M., Stylianou, Y.: Effective emotion recognition in movie audio tracks. ICASSP. pp. 5120–5124. IEEE. (2017)
10. Jacob, A.: Speech emotion recognition based on minimal voice quality features. ICCSP.pp. 0886–0890). IEEE. (2016)
11. Sondhi, S., Vijay, R., Khan, M., Salhan, A. K.: Voice analysis for detection of deception. KICSS. pp. 1–6. IEEE. (2016)
12. Palo, H. K., Mohanty, M. N., Chandra, M.: Sad state analysis of speech signals using different clustering algorithm. NGCT. pp. 714–718. IEEE. (2016)
13. Schuller, B., Steidl, S., Batliner, A.: The interspeech 2009 emotion challenge. Tenth Annual Conference of the International Speech Communication Association. (2009)
14. Kim, H. J., Bae, K., Yoon, H. S.: Age and gender classification for a home-robot service. In Robot and Human interactive Communication. RO-MAN. pp. 122–126. IEEE. (2007)
15. Teixeira, J. P., Fernandes, P. O.: Jitter, shimmer and hnr classification within gender, tones and vowels in healthy voices.Procedia Technology,vol 16, 1228–1237. (2014)
16. Tsanas, A., Little, M. A., Fox, C., Ramig, L. O.: Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease.Neural Systems and Rehabilitation Engineering, IEEE Transactions, vol. 22.1. pp 181–190. (2014)
17. Gómez-Coello, A., Valadez-Jiménez, V. M., Cisneros, B., Carrillo-Mora, P., Parra-Cárdenas, M., Hernández-Hernández, O., Magaña, J. J.: Voice Alterations in Patients With Spinocerebellar Ataxia Type 7 (SCA7): Clinical-Genetic Correlations.Journal of Voice,vol. 31.1. pp. 123–e1. (2017)
18. Yamasaki, R., Montagnoli, A., Murano, E. Z., Gebrim, E., Hachiya, A., da Silva, J. V. L., Tsuji, D.: Perturbation Measurements on the Degree of Naturalness of Synthesized Vowels.Journal of Voice, vol 31.3, 389-e1. (2017)
19. Kotti, M., Patern, F.: Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema.International journal of speech technology,vol. 15.2. pp. 131–150. (2012)
20. Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine,vol. 29.6, 82–97. IEEE. (2012)
21. Mitra, V., Sivaraman, G., Nam, H., Espy-Wilson, C., Saltzman, E., Tiede, M.: Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition.Speech Communication,vol. 89. pp 103–112. (2017)
22. Collobert, R., Puhrsch, C., Synnaeve, G.: Wav2letter: an end-to-end convnet-based speech recognition system.arXiv preprint arXiv:1609.03193. (2016)
23. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Chen, J.: Deep speech 2: End-to-end speech recognition in english and mandarin. International Conference on Machine Learning. pp. 173–182. (2016)
24. Palaz, D., Collobert, R.:Analysis of cnn-based speech recognition system using raw speech as input(No. EPFL-REPORT-210039). Idiap. (2015)
25. Sainath, T. N., Kingsbury, B., Mohamed, A. R., Ramabhadran, B.: Learning filter banks within a deep neural network framework. IEEE Workshop onASRU. pp 297–302. IEEE. (2013)
26. Farrús, M.: Jitter and shimmer measurements for speaker recognition. 8th Annual Conference of ISCA. pp. 778–781. (2007)

27. Gu, Y., Li, X., Chen, S., Zhang, J., Marsic, I.: Speech Intention Classification with Multimodal Deep Learning. Canadian Conference on Artificial Intelligence. pp. 260–271). Springer. (2017)
28. Chang, J., Scherer, S.: Learning Representations of Emotional Speech with Deep Convolutional Generative Adversarial Networks.arXiv preprint arXiv:1705.02394. (2017)
29. Ghosh, S., Laksana, E., Morency, L. P., Scherer, S.: Representation Learning for Speech Emotion Recognition. INTERSPEECH. pp. 3603–3607. (2016)
30. Mao, Q., Dong, M., Huang, Z., Zhan, Y.: Learning salient features for speech emotion recognition using convolutional neural networks.IEEE Transactions on Multimedia,vol. 16.8. pp 2203–2213. (2014)
31. Ma, X., Yang, H., Chen, Q., Huang, D., Wang, Y.: DepAudioNet: An Efficient Deep Model for Audio based Depression Classification. Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge.pp. 35-42. ACM. (2016)
32. Abumallouh, A., Qawaqneh, Z., Barkana, B. D. : Deep neural network combined posteriors for speakers' age and gender classification. CT-IETA. pp. 1–5. IEEE. (2016)
33. Qawaqneh, Z., Mallouh, A. A., Barkana, B. D.: Deep neural network framework and transformed MFCCs for speaker's age and gender classification.Knowledge-Based Systems, vol.115. pp. 5–14. (2017)
34. Liu, Y., Wang, X., Hang, Y., He, L., Yin, H., Liu, C.: Hypemasality detection in cleft palate speech based on natural computation. ICNC-FSKD. pp. 523–528. IEEE. (2016)
35. Cummins, N., Epps, J., Ambikairajah, E.: Spectro-temporal analysis of speech affected by depression and psychomotor retardation.Acoustics, Speech and Signal Processing. pp. 7542-7546. IEEE. (2013)
36. Teixeira, J. P., Gonçalves, A.: Algorithm for jitter and shimmer measurement in pathologic voices.Procedia Computer Science, vol.100, pp. 271–279 (2016)
37. Shahnaz, C., Zhu, W. P., Ahmad, M. O.: A new technique for the estimation of jitter and shimmer of voiced speech signal. Electrical and Computer Engineering Canadian Conference. pp. 2112–2115. IEEE (2006)
38. Dong, B.: Characterizing resonant component in speech: A different view of tracking fundamental frequency.Mechanical Systems and Signal Processing, vol.88. pp. 318–333. (2017)
39. Liu, B., Tao, J., Zhang, D., Zheng, Y.: A novel pitch extraction based on jointly trained deep BLSTM Recurrent Neural Networks with bottleneck features. Acoustics, Speech and Signal Processing International Conference. pp. 336–340. IEEE. (2017)
40. Schlotthauer, G., Torres, M. E., Rufiner, H. L.: A new algorithm for instantaneous F0 speech extraction based on ensemble empirical mode decomposition. 17th European Signal Processing Conference. pp. 2347–2351). IEEE (2009)
41. Buder, E. H.: Acoustic analysis of voice quality: A tabulation of algorithms 19021990. Voice quality measurement. pp. 119–244. (2000)
42. Klingholz, F., Martin, F.: Quantitative spectral evaluation of shimmer and jitter. J Speech Hear Res, vol 28.2. pp 169–174. (1985)
43. Schnupp, J., Nelken, I., King, A.: Auditory neuroscience: Making sense of sound. MIT press (2011)