A Preliminary Study on Vowel Recognition via CNN for Disorder People in Malay Language

2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) | 978-1-6654-4930-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISMSIT52890.2021.9604589

Nur Syakirah Muhammad Zamri
Center for Telecommunication Research and Innovation (CeTRI), Fakulti Kejuruteraan Elektronik dan Kejuruteraan Komputer (FKEKK), Universiti Teknikal Malaysia Melaka (UTeM),
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
b021710173@student.utem.edu.my

Nik Mohd Zarifie Hashim
Center for Telecommunication Research and Innovation (CeTRI), Fakulti Kejuruteraan Elektronik dan Kejuruteraan Komputer (FKEKK), Universiti Teknikal Malaysia Melaka (UTeM),
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
nikzarifie@utem.edu.my

Abd Shukur Ja'afar
Center for Telecommunication Research and Innovation (CeTRI), Fakulti Kejuruteraan Elektronik dan Kejuruteraan Komputer (FKEKK), Universiti Teknikal Malaysia Melaka (UTeM),
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
shukur@utem.edu.my

Abd Majid Darsono
Center for Telecommunication Research and Innovation (CeTRI), Fakulti Kejuruteraan Elektronik dan Kejuruteraan Komputer (FKEKK), Universiti Teknikal Malaysia Melaka (UTeM),
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
abdmajid@utem.edu.my

Mohd Juzaila Abd. Latif
Fakulti Kejuruteraan Mekanikal (FKM), Universiti Teknikal Malaysia Melaka (UTeM),
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
juzaila@utem.edu.my

Parathythasan Rajaandra
Pusat Rehabilitasi PERKESO Lot PT7263 Bandar Hijau,
Hang Tuah Jaya, 75450 Melaka, Malaysia
parathythasan.rajaandra@rehabmalaysia.com

***Abstract*— Stroke is one disease showing an increment trend as people live their lives in a stressful manner. Rehabilitation is one of the procedures to recover the patient to a normal condition. The rehabilitation process and activities require an extended period to retrain back the patient's capability, speak, listen, walk, etc. For this, a dedicated physiotherapy procedure was conducted according to the rehab trainer and expertise. One of the rehabilitations is to help the patient to have back their speaking skill and capability. The rehabilitation activities are generally conducted manually through manual listening and teaching the stroke patient periodically by the rehab trainer. The manual rehabilitation activities physically require the rehab trainer's presence, documentation, and manual data recording. This manual activity could be challenging when we face a lack of trainers and the situation of many patients with less trained in the field. Therefore, an intelligent system could be an alternative for rehabilitation to provide the user-friendly and straightforward technique to learn, repeat, and evaluate. In the paper, as the preliminary study, we proposed a smart vowel recognition for Malay Language using Convolutional Neural Network (CNN). We also proposed a new Malay Language dataset consist of 5 vowels, /a/, /e/, /i/, /o/ and /u/ for the use of future research. The result shows that the vowel recognition using this dataset is comparable and suitable for recognizing the vowel type.**

Keywords—CNN, Malay language, rehabilitation, stroke patient, vowel recognition

I. INTRODUCTION

The cerebrovascular disease occurs when blood vessels become plugged or cracked from a failure part of the brain. It will direct to the malfunction of the human brain system. Here, cerebrovascular disease is also well known as stroke by a layman. The primary effect of the stroke is when the brain parts are seriously affected, and the patient could suffer permanent damage to their body system. Untreated stroke disease is the severe effect of the failure part occurred in the human brain. In general view, there are three types of stroke, Ischemic stroke, Hemorrhagic stroke, and Transient ischemic.

Depends on the patient's condition, the rehabilitation is set and determined accordingly as a match and suitable procedure to improve the patient condition to be recovered. Speech is a natural form of human communication. Contrarily, speech disability is one of the disorder effects for a stroke patient. Stroke can affect people's communication in different ways. One of those which could be affected after the stroke is dyspraxia. Dyspraxia refers to difficultly moving and coordinating voice with sound simultaneously. Speaking dyspraxia occurs when the muscles cannot move in a specific order and sequence to produce the sounds required for regular and clear speech. When an average person wants to speak, generally, muscles inevitably create intelligible speech synchronously function well and are free from an error.

On the other hand, stroke patient cannot move their muscle in a correct and consistent order. As a result, dyspraxia patients may have difficulty pronouncing words. They could attempt by repeating the phrase multiple times and then try to amend them whenever they wanted to speak. Since speech consists of vibrations produced by our throat as voice signal, in a science perspective, these vibrations represent speech waveforms.

Conversely, from the traditional concept of understanding a voice signal as a wave signal, we proposed a new alternative signal observation way in this paper. In a traditional approach in voice research, the voice signal is considered a two-axis signal, represented as time and amplitude. However, utilizing a traditional wave signal could not be utilized and applied directly to any Convolution Neural Network (CNN) approaches. As a result, the wave signal requires a modification into an image form. This modification could be

978-1-6654-4930-4/21/$31.00 ©2021 IEEE

in various forms of images. Here, to the extent of our knowledge, we observe the wave signal as a spectrogram image that will provide sufficient information about each vowel in the Malay language. The Malay language is spoken by more than 33 million speakers [1] from the wide area of South-East Asia, Malaysia, Indonesia, Southern Thailand, Singapore, and Borneo[2]. Since the number of Malay speakers delivers a large number, the stroke patient vowel recognition study is crucial for strengthening knowledge for the rehab activities [3-6]. For this, in the experimental and analysis purposes, we utilized five vowels in Malay language as our analysis main subject, /a/, /e/, /i/, /o/ and /u/. CNN, which is widely implemented in many classification research, provides a reliable and comparable classification result. One of the simple and versatile CNN networks, VGG16[7], is employed as the primary model for vowel classification in the experimental work for representing the experimental performance. Our contribution can be summarized as follows:

- we propose a way to gain voice information by utilizing the spectrogram image,
- we introduce a new Malay language vowel dataset image with a mixed male and female subject, which consists of three lengths of each vowel wave signal, and
- we show that the proposed dataset images are comparable and acceptable in recognizing the vowels correctly via CNN.

II. Related Works

Researchers have done several works towards this stroke patient rehabilitation methods or could for other patient types. Begin with a traditional approach to facilitate the students with articulation disorders using a manual activity called as conventional articulation therapy approach. Articulation disorder [8] is a speech disorder involving various difficulties in articulating specific types of sounds, substituting sound for another, slurring of speech, or indistinct speech. This method was focused on the phonetic placement of the error sound by training the motor skills to produce the sound appropriately. This intervention method uses a hierarchy to help children establish the correct sound manually without any intelligent approach. The research on the severity of speech disorders was proposed from a mild, moderate, severe, and inability to speak level. This traditional approach teaches them how to pronounce accurately every individual sound. This is one of the early speech rehabilitation methods conducted by the professional in this field to fix the severity of speech disorder. Later, Van Riper proposed articulation drills and motor learning, which is also considered a manual rehab approach by training the tongue movement and coordination of the other articulators such as lips and jaw [9].

One of the earliest approaches using Machine Learning (ML) later improved the conventional methods. The ML approach results in clinically useful major disorder disease (MDD) risk-stratification models generated from baseline patient self-reports [10]. Although the techniques are not generally involved in speech rehabilitation, using machine learning can be as guidance while performing for this project. In the Development of Control System for Fruit Classification based on Convolution Neural Network project published in 2018, this project used Convolution Neural Network (CNN) to develop a fruits detection and recognition based on CNN. The project's accuracy is closed to 94 percent for 30 classes of 971 images. The data set consists of 971 images, classified into 30 different fruit classes, and every fruit class contains about 32 additional images. Next, the implementation of CNN for fruit classification is applied. There is a control system design for a vision-based automated decision-making system to develop a computer vision-based control system [9]. This project is about developing a control system using 'Alexnet.' The implementation using a graphic processing unit in Matlab to perform the classification and simulation process. In the end, the accuracy result of the project is closed to 94 percent[11]. This research paper facilitates the implementation of CNN in speech rehabilitation for disorder people project. This work also implements the CNN for evaluating the accuracy for normal and disordered people.

Research using the specific vowel in the Malay language initiated from past decades. Among of them [12,13], their work was done using the phonetic properties of the six Malay vowels, /a/, /e/, /ə/, /i/, /o/ and /u/. [6,7] work by employing image processing techniques utilizing magnetic resonance imaging (MRI) to represent the vocal tract. [13] work shows the usage of four formant frequencies to analyze the sustained six vowels of Malay children aged between 7 and 12 years old. Here, all the Malay language vowel research is still not conducted intelligently and for stroke patient purposes.

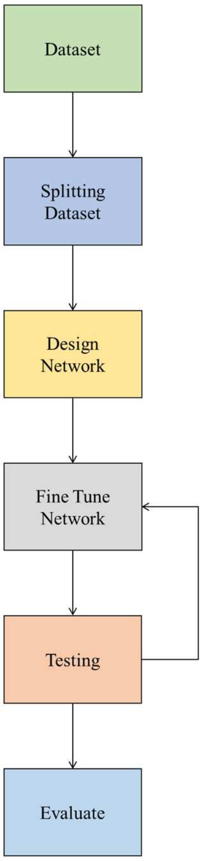

Fig. 1. Proposed method flowchart

III. Proposed Method

In general, the flow of the proposed evaluation method is illustrated in Figure 1 below. A suitable dataset of images needs to be ready to acquire an intelligent classification for stroke patients. For this, we carefully select the wave signal which recorded from the subject with various wave condition. The wave signal is converted into spectrogram images so that it is later suited to be the input for the CNN. VGG16, VGG

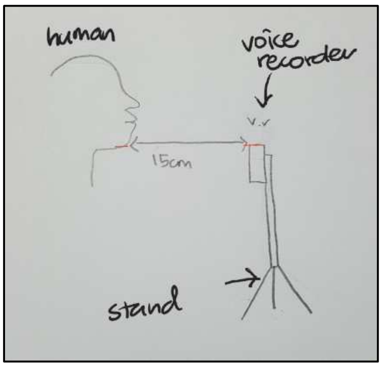

Fig. 2. Voice recording arrangement schematic drawing for the experimental setting

19, and ResNet are employed with the self-trained model with five classes as the network's output.

A. Audio Recording

The wave signals are recorded from nine people from 20-24 years old with mixed gender. To ensure the variety of wave signals for the proposed dataset, we recorded three wave types: short, middle, and long for each vowel /a/, /e/, /i/, /o/, and /u/. Here the short wave is subject required to say /a/ in one second. At the same time, the middle and long signals are at the 2- and 3-seconds length. The voice recordings were done using a voice recorder Remax RP1 8Gb Digital Audio Voice Recorder and the recording arrangement schematic for the experimental setting shown in Figure 2 above.

B. Converting wave signal to image form

The wave signal recorded are later converted into image form, spectrogram image. We set the y-axis as in log frequency for visualizing the upper region of the spectrogram with more information. The transformed spectrogram images are later manually crop as one vowel image from /a/ to /o/ at a specific size 240 x 55 pixels.

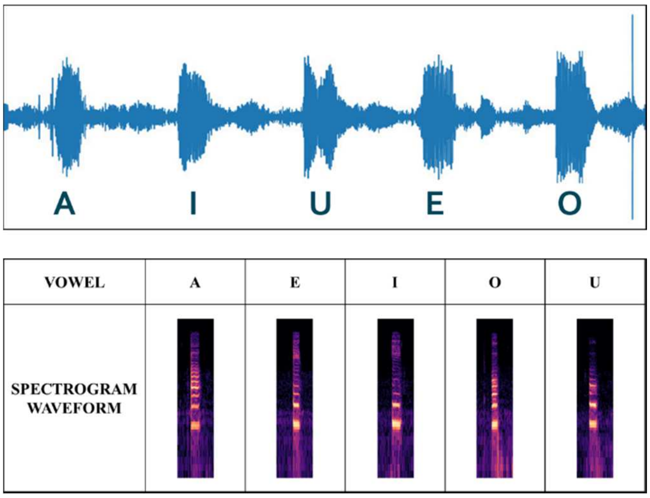

Fig. 3. (Above) Vowel voice signal in amplitude and time, (Bottom) Example of vowel voice signal in image form using spectrogram

C. Dataset Images

Then, the dataset is separated into training and validation in a ratio of 2, as shown in Figure 3, for inferring the suitable network for the stroke patient. The proposed dataset image in a total number of 4050 images. The proposed dataset images are publicly available later on our research website after the publication of the paper. The images were not only for the stroke patient research purpose but also for other suitable research which utilized wave signals with the CNN.

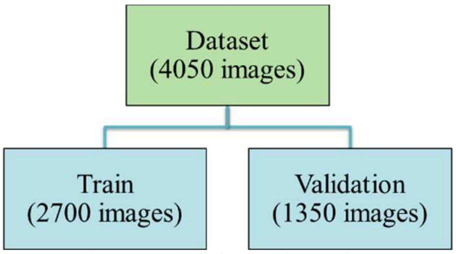

Fig. 4. Dataset images in number

D. Training and Evaluating the Network Model

CNN architecture consists of a convolutional layer, pooling layer, fully connected layer, dropout, and activation functions. In this project, the input size of image is (240, 55). The input for the image of dimensions is (240, 55, 3). Next convolution layer (Conv1), Conv1_1using 32 filters with image of dimensions is (238, 53) and Conv1_2 using 64 filters and max pooling with image of dimensions is (236, 51). Then, at the convolution layer (Conv2), the filters increase from 32 and 64 filters to 128 filters. Conv2_1 using 128 filters and max pooling with the image of dimensions of (116, 23) and Conv2_2 using 128 filters and max pooling with image of (9, 56). After that, for dense, using a dense layer of 1024 units and a dense softmax layer of 5 units. The image size reduction for each layer can be seen in Figure 5 below.

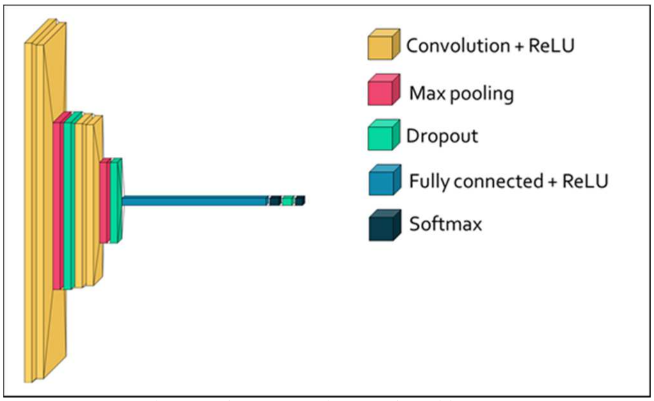

Fig. 5. The proposed network architecture

The model was designed by constructed CNN with input image is (240,55), ADAM classifier as optimizer, and Softmax for activation function. Then, the model was running with different batch sizes and epochs to be compared to each other.

IV. RESULTS

We conducted several experimental settings to present the model's effectiveness for recognizing the Malay language vowel for stroke patients. The simple batch size, epoch size, and network model comparison study were conducted utilizing the proposed dataset images. Since all the experiments were conducted on Google Colab by Google Inc with a free account, it requires plenty of time to finalize the result.

TABLE I. THE RESULTS WHICH OBTAINED BY USING A DIFFERENT NUMBER OF BATCH SIZES WITH EPOCH = 80 UTILIZING THE DESIGNED MODEL.

Batch Size	Accuracy Percentage (%)	
	Training Accuracy (%)	*Validation Accuracy (%)*
5	100.00	73.04
10	100.00	**73.56**
15	100.00	72.74
20	100.00	72.89

TABLE II. THE RESULTS WHICH OBTAINED BY USING A DIFFERENT NUMBER OF EPOCH WITH BATCH SIZE = 50 UTILIZING THE DESIGNED MODEL

Epoch Size	Accuracy Percentage (%)	
	Training Accuracy (%)	*Validation Accuracy (%)*
20	99.78	73.93
50	100.00	**76.59**
80	100.00	72.89
100	100.00	73.04

A. Comparison analysis with batch size

In this comparison analysis, all the results for batch size were collected and compared to each other in Table I. From the table, the best batch size is 10, with the highest validation accuracy in percentage than different batch sizes with 73.56%. However, the higher number of batch sizes will reflect training speed but deliver a poor generalization. Here, the 10 of batch size delivers a good performance on validation accuracy compared to the batch size using the proposed design network model.

B. Comparison analysis with epoch size

Based on epoch number as a comparison study, the results are represented in Table II above. From this table, the best epoch is 50, which gained the highest validation accuracy percentage compared to other epoch sizes. Here, a larger epoch size can overtrain the model; the model will lose generalization capacity by overfitting the training data. As the number of epochs increases, the training data loss decreases.

C. Comparison analysis with other network model

In the third comparison analysis, all the network models were compared with the batch size at 50 and two epoch sizes, 80 and 10. The result gained from different types of network models are tabulated in the table below. Table III below shows that the designed model has the highest percentages of train and validation accuracy than other models with 73.04%. Although the proposed design model is simple, the classification performance is still comparable to the existing CNN network model. Although the wave signal classification is not a direct approach of conventional CNN, with a minor modification from wave signal to an image form, we showed that this idea could be discussed further in the future in various aspects and experimental settings.

TABLE III. THE RESULTS OBTAINED BY USING DIFFERENT TYPES OF MODEL WITH SAME VALUE OF BATCH SIZE AND DIFFERENT VALUE OF EPOCH

Network Model	Batch Size	Epoch	Accuracy Percentage (%)	
			Training Accuracy (%)	*Validation Accuracy (%)*
Designed	50	80	100.00	72.89
		100	100.00	**73.04**
VGG16		80	93.48	61.11
		100	94.04	61.04
VGG19		80	93.37	57.93
		100	93.67	57.48
ResNet50		80	84.89	55.70
		100	87.22	59.33

Figure 6 shows the model accuracy and loss graph based on the epoch size number as a reference when using the new dataset images with the proposed network model. In the proposed network model with the epoch of 100 and batch size of 50, the percentage of final train accuracy is 100 %, while validation accuracy is 73.04%. The result is still considered preliminary, as the arrangement of the dataset images is still in the development stage for a complete and comprehensive dataset.

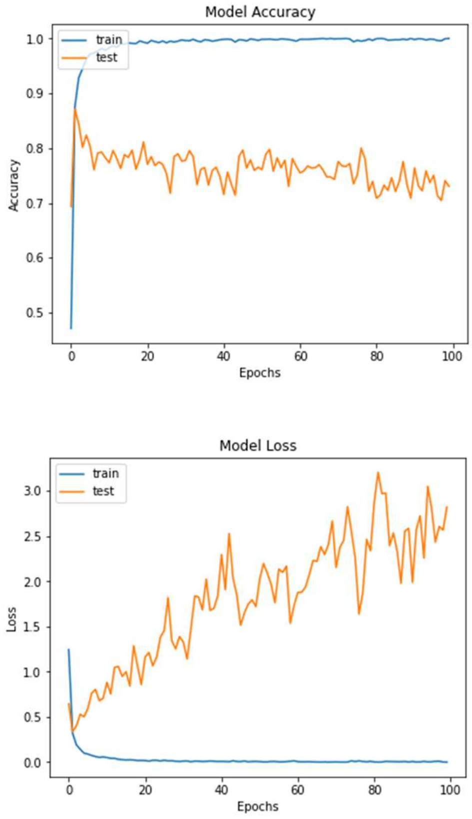

Fig. 6. The proposed network model accuracy (above) and model loss (bottom)

For the comparison study with other existing network models, using Figure 7 below, we show the model accuracy compared with the epoch size number. All three comparative network models infer the validation accuracy in general at below 70.0%.

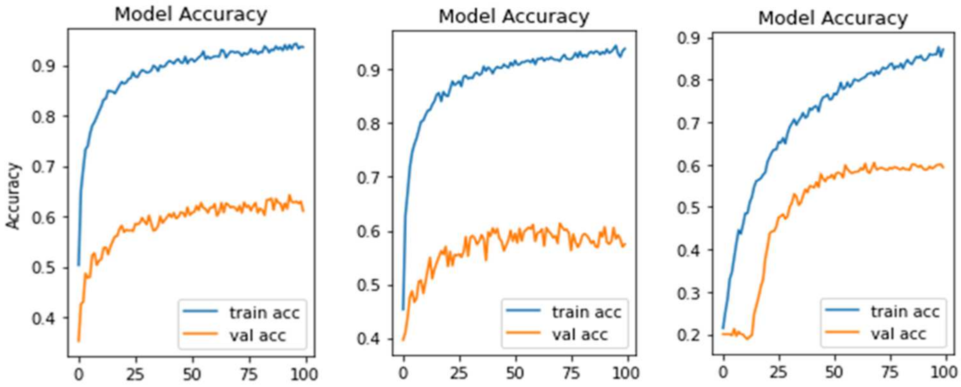

Fig. 7. The model accuracy for VGG16 (left), VGG19 (middle), and ResNet (right)

V. Conclusion

We designed a suitable network model for the stroke patient purpose rehabilitation activities. The paper's contribution could help the rehab center deliver more systematic and intelligent rehab activities for ensuring people with stroke pronounce more clearly after attending the rehab sessions. We trained the model using the new proposed dataset captured from the nine personals of various gender, which is vowel sound, into spectrogram images by utilizing the CNN model. This paper's simple CNN architecture model consists of convolutional, pooling, fully connected, dropout, and activation functions layers. The purpose of the designed model is to classify the sound vowel to be used to recognize the vowel sound from the disorder peoples (patients). For future work, we expected to do more analysis with the wave signal from the stroke patient. The study on noised-add wave signal could also be considered for future work in observing network reliability in many aspects and condition settings.

Acknowledgment

The authors would like to thank Perkeso Rehab Center, Melaka for the Research Project Collaboration, and Universiti Teknikal Malaysia Melaka for the financial support through PJP/2020/FKEKK/PP/S01792.

References

[1] F.M. Onn, Aspects of Malay phonology and morphology: A generative approach. Universiti Kebangsaan Malaysia, 1980.

[2] I. Zahid and M. Shah Omar, *Phonetics and phonology in Malay*. Kuala Lumpur: PTS Professional, 2006.

[3] M. Donohue, "Malay as a mirror of austronesian: Voice development and voice variation," *Lingua,* vol. 118, no. 10, pp. 1470–1499, 2008.

[4] P. Cole and G. Hermon, "Voice in Malay/Indonesian," *Lingua*, vol. 118, no. 10, pp. 1500–1553, 2008.

[5] H. Thamrin, *et al.,* "Crowdsourcing in developing repository of phrase definition in Bahasa Indonesia," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 17, no. 5, pp. 2321-2326, 2019.

[6] H-N. Ting, *et al.,* "Formant frequencies of Malay vowels produced by Malay children aged between 7 and 12 years," *Journal of Voice*, vol. 26, no. 5, pp. 664.e1-664.e6, 2012.

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[8] C. Van Riper, Speech correction : an introduction to speech pathology and audiology / Charles Van Riper, Robert L. Erickson. — 9th ed. p. cm. Needham Heights, MA: A Simon Schuster Company, 1995.

[9] S. P. Rosenbaum S, Speech and Language Disorders in Children: Implications for the Social Security Administration's Supplemental Security Income Program, Washington (DC): National Academies Press (US), 2016.

[10] A. Einstein, "Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports," Molecular Psychiatry, vol. 21, no. 10, pp. 1366–1371, 2016.

[11] Y. H. P. H. Khaing, Zaw Min Naung, "Development of control system for fruit classification based on convolutional neural network," Proceedings of the 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2018, vol. 2018-January, no. 10, pp. 1805–1807, 2018.

[12] H. Thamrin, *et al.,* "Crowdsourcing in developing repository of phrase definition in Bahasa Indonesia," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 17, no. 5, pp. 2321-2326, 2019.

[13] H-N. Ting, *et al.,* "Formant frequencies of Malay vowels produced by Malay children aged between 7 and 12 years," *Journal of Voice*, vol. 26, no. 5, pp. 664.e1-664.e6, 2012.