Design of the Speech Tone Disorders Intervention System Based on Speech Synthesis

Yang Xinyi, Si Boyu*, Meng Qingyun, Huang Kailin

College of Medical Instrument, Shanghai University of Medicine & Health Sciences, Shanghai, China

*Corresponding author: siby@sumhs.edu.cn

Abstract — Tone Disorder is a typical performance of speech impediment. It has become an urgent matter because of affecting patients’ common communications. In this paper, a speech tone disorder intervention system is designed on MATLAB GUI. By sampling the patients’ speech, extracting and adjusting the pitch information and synthesizing new speech, the system can help patients do self-intervention and build a ‘speech-hearing feedback chain’ for them during the rehabilitation process. The system is non-invasive for intervention of speech tone disorder from the uncoordinated movement of articulatory organs or psychological disorders. The results of two single-subject experiments prove that the system can improve the intervention effect of speech tone disorders intervention and increase the comprehensibility of users’ speech.

1. INTRODUCTION

Tone Disorder is a typical performance of speech impediment. It is mostly caused by the uncoordinated movement of articulatory organs, such as phonating disorder, wrong way of respiration, swallow disorder and endocrine disorder, or by some kinds of psychological disorders[1]. In the aspect of acoustic analysis, to assessment tone disorder means to measure the speech pitch. It has something to do with the length, quality and strength of the vocal fold, and under-glottal pressures. The performance of speech tone disorder is that the patients’ voice sounds abnormally sharp or muffled, as a result, the fundamental frequency of the speech signal performs too high or too low[2]. Although the patients with speech tone disorders can make pronunciation fluently, their speech sounds too shrill or too low to be heard and understood[3]. It affects the daily communication severely, and the patients’ abnormal tone should be controlled in a certain range to solve the problem above.

In this paper, we design a speech tone disorders intervention system based on speech synthesis and MATLAB GUI. After recording patients’ speech, the system can process the speech signal and extract the fundamental frequency parameters. According to the normal standard of speech pitch, these parameters are adjusted for synthesizing a new piece of speech in the way of formant synthesis, which will be played to patients during the intervention process and help them build a ‘speech-hearing feedback chain’ to improve the rehabilitation quality.

2. SYSTEM MODULE DESIGN

The speech tone disorders intervention system should be a helpful tool to handle patients’ problems that the patients’ speech tone is too high or too low. It makes use of patients’ own speech, change the parameters of fundamental frequency to the normal level and synthesize the ‘correct’ speech. There are
four modules in the system, including pre-processing module, pitch extracting module, formant
detecting module and speech synthesizing module.

2.1 Pre-processing Module
The pre-processing module mainly contains four steps: pre-emphasis, framing, adding windows
function and terminals detection. After pre-processing, the noises are removed from the original signal,
but the useful characters are kept for the following process[4].

1) Pre-emphasis: The aim of this step is to emphasize the high frequency part of speech signal in
order to get rid of the influence by mouth and lips. The result only shows the vocal influence. This
operation can increase the resolution of high frequency and make the spectrum of signal more flat.
Based on the situation, parameters extracted from the signals will match the original vocal fold model
better. We use a high-pass digital filter of which the transfer function is:

\[ H(z) = 1 - az^{-1} \]  \hspace{1cm} (1)

to realize pre-emphasis. Here, \( a \) is the pre-emphasis coefficient, 0.9<\( a \)<1.0.

2) Framing: As the actual speech signal is very long and the pitch is unstable, the signal needs to
be analyzed in short time. The operation is framing. Generally, one frame is 10-30 ms, because the
short-time signal can be considered stable. In order to keep the whole signal’s continuity and the
fundamental frequency’s stability, the adjacent two frames share an overlapping part.

3) Adding windows function: This step can enhance the samples of the speech signal while weaken
the terminal part of each frame[5]. It makes the signal after framing continuous and periodic. We use
the Hamming window, which is a kind of cosine window function, thanks to its narrow sidelobe and
relatively high solution.

4) Terminals detection: A piece of speech signal includes the voiced segment and silent segment.
The start or the end of each voiced segment is called a terminal. Terminal detection is a necessary step
in the early part of pitch and formant extraction. It can confirm the position of the terminals in the
speech segment. We adopt the double threshold detection based on energy spectrum entropy. Vowels
are considered as the energy concentration area in the voiced segment. The parameter energy-entropy-
ratio is used to detect the terminal and vowels area, of which the equation is:

\[ EEF_i = \sqrt{1 + \frac{E_{Li}}{H_i}} \]  \hspace{1cm} (2)

here, \( E_{Li} \) is energy, \( H_i \) is spectrum entropy.

2.2 Pitch Extracting Module
Pitch is the key parameter in assessing speech tone disorder. Because Mandarin is made of consonant,
initial consonant and tone, we use the method of the vowels’ main body extension detection. First, we
detect the fundamental frequency of the vowel main body. And then, we check the transitional
extension area of the vowel main body. The advantage of the method is reducing the irregular sampling
points, the detection results can be more accurate. The specific steps are shown in Figure 1.

\[ \text{Adding windows function to the speech signal and framing} \]
\[ \text{Terminals detection and the vowel main body detection} \]
\[ \text{Filter the signal, adding windows and framing again} \]
\[ \text{Detecting the pitch of the front and back of the transitional area} \]
\[ \text{Detecting the pitch of the vowel main body} \]
\[ \text{Calculating the extensive range and length of the vowel main body} \]

Figure1 Flowchart main body - extension pitch detection

To detect the pitch of vowels main body, we adopt the method of auto-correlation function. It is a
time-domain detective way to calculate the fundamental frequency. The auto-correlation function is:
\[ R_i(k) = \sum_{m=1}^{N-m} x_i(m)x_i(m+k) \]  

(3)

Here, \( k \) is the delay time, \( x_i(m) \) is frame \( i \) of the speech signal after adding window function, \( N \) is the length of the frame. The results of the equation are peak values, these points can be used to calculate the fundamental frequency.

But the auto-correlation function may not be suitable for detecting the pitch of the transitional extension arrange, because the periodic of the arrange is not obvious. Hence, we calculate the peak values by the auto-correlation function and seek for the proper ones according to the shortest distance principle. We find two values in the shortest distance and calculate the final results by linear interpolation.

2.3 Formant Detecting Module

Formant detection should be done before speech synthesis, as we use the formant and corrected frequency to synthesize the normal speech. In this paper, we detect the formant frequency based on LPC method, the flowchart is shown in Figure 2.

According to Figure 2, we cut the speech signal into several frames and calculate the first three formant frequency F1-F3 of each frame. Then, the signal is analyzed by linear prediction. If the all-polar transfer function of the vocal tract model is:

\[ H(z) = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \]  

(4)

The prediction error filter is:
The coefficients of $A(z)$ can be used to confirm the central frequency and bandwidth of formant quickly. We suppose 

$$z_i = r_i e^{j\theta}$$

is one root of $A(z)$. $z_i^*$ is its conjugate complex root. The relation between formant frequency and the value of the root is shown in Figure 3.

If $F_i$ is the formant frequency, $B_i$ is the bandwidth, which is 3dB:

$$2\pi T F_i = \theta_i$$  
$$e^{-\theta_i / \pi T} = r_i$$

We can obtain the following results:

$$F_i = \frac{\theta_i}{2\pi T}$$

$$B_i = -\ln r_i / \pi T$$

Finally, the formant frequency of each frame and the speech signal’s formant spectrum are achieved.

### 2.4 Speech Synthesizing Module

The formant speech synthesis method is based on the vocal tract model. It takes the vocal tract as a resonator. The speech generating system can be divided into three parts: excitation model, vocal tract model and radiation model. The excitation produces the vocal vibration. Vocal tract modules the vibration into speech signals. The signals come out from the radiation model, which includes the mouth and the nose, and become human speech. In this paper, we correct the fundamental frequency parameters in the excitation model to change the formant frequency and bandwidth of the vocal tract model. The new signals get through the radiation model to form the normal speech. Before the speech is played, we input it through a high-pass filter to remove the low-frequency elements and noise to improve its quality.

### 3. SOFTWARE DESIGN

The speech tone disorders intervention system is designed to help patients build a ‘speech-hearing feedback chain’ and promote effort of intervention. It can adjust patients’ speech tone to the normal arrange. After all, tones with too high or too low pitch may affect the daily communication, the system will bring positive improvement to patients’ speech quality. There are three parts in the system, including the speech recording part, speech processing and synthesizing part, and training part. We develop the system on MATLAB R2015a, with the toolbox of MATLAB GUI. The software can run on the x64 PC with Microsoft Windows 10 OS.
3.1 Speech Recording Part

The homepage of the system is shown in Figure 4. There is one button, which is ‘start’. When users press the button, they can enter the first part, speech recording part, which is shown in Figure 5. There are two functional buttons on the screen: ‘record’ and ‘play’. When users press ‘record’, the system will record speakers’ voice automatically and draw the signal wave on the screen. Patients can listen to the record by pressing the ‘play’ button.

3.2 Speech Processing and Synthesizing Part

This part is the most important one in the system. We design four buttons on the screen (in Figure 6) according to the speech tone adjusting schedule, including ‘speech pre-processing’, ‘pitch extraction’, ‘tone adjustment’ and ‘speech synthesis’. When users press ‘speech pre-processing’, the record in the first part will be processed to be ready for pitch extraction. And the system will draw the signal waveform on the coordinate. Then, users can press ‘pitch extraction’, ‘tone adjustment’ and ‘speech synthesis’ in sequence to synthesize the new speech with proper fundamental frequency. When the button ‘pitch extraction’ works, a new dialog will appear to suggest users to change the pitch values. When the speech signal is ready, the system will save it to a audio file for the next part.

3.3 Training Part

The last part is used for training patients. There is only one button ‘Play’ on the screen. When users press it, the system will play the synthesis speech through speakers. Patients can follow the doctors’ directions to start the tone disorder training.
4. EXPERIMENT

The speech tone disorders intervention system is designed for single-patient training. In the paper, we adapt the single subject experiment in a period of 45 days. We take patients’ fundamental frequency as the monitoring parameter to observe the effectiveness in rehabilitation.

4.1 Subject

Two 4-year-old children with speech tone disorder are invited to join the experiment. Before intervention, the subjects are assessed to have shrill and abnormally low pronunciation separately. They have received speech training and get used to the way of rehabilitation.

4.2 Method

The experiment is designed in single baseline mode, including baseline period and intervention period. First, in baseline period, we direct the subjects to pronounce /a/ and last for 3 seconds. The subjects are required to repeat the experiment three times. The fundamental frequency F0 of these speech are extracted. Subject A: F0 = 511 Hz, Subject B: F0 = 174 Hz. The normal pitch of children is between 300 and 400 Hz. It is obvious to find that both the subjects’ pitch are not in the normal range. Then, the intervention is held to train the subjects with the system we designed. Speech therapists guarantee the training frequency for 45 days. We collect the subjects’ speech and extract the pitch every 15 days. The patients are still required to pronounce /a/ for 3 seconds, so we obtain 3 groups of pitch data. According to the data, we analyze the improving situation of subjects’ pitch, draw the baseline chart and figure out the significant difference.

4.3 Result

According to the experimental data, the line charts of both the subjects are shown in Figure 7 and Figure 8. We can see that there are significant differences of the pitch data in the intervention for Subject A and B. The pitch of Subject A changed more obvious than B. It proves that the system makes a good effort in improving patients’ tone pronunciation. Their fundamental frequency becomes normal, while the single subject’s baseline data shows significant difference.
5. CONCLUSION
The speech tone disorders intervention system is developed in MATLAB GUI. It consists of the functions of recording, speech signal processing and playing back. The speech signal processing module including pre-processing, pitch extraction, tone modifying and speech synthesis. The system also can show the waveform on the screen. After the single-subject experiment, 2 patients’ tone change to the normal range after the intervention. The results of the experiment prove the feasibility and effectiveness of the intervention system.

ACKNOWLEDGMENT
This work was supported by the educational reform special project for innovation and entrepreneurship training of Shanghai University of Medicine & Health Sciences (B3-0200-20-309010-56, 2020).

REFERENCES
[1] Shfiberg LD, Tomblin JB, McSweeny JL. “Prevalence of speech delay in 6-year—old children and comorbidity with language impairment”. Journal of Speech, Language, and Hearing Research, vol. 1, pp.1461-1481, 1999.
[2] Hui L, Dai B Q, Wei L. “A Pitch Detection Algorithm Based on AMDF and ACF”. Acoustics, Speech and Signal Processing. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, IEEE, 2006.
[3] De Cheveigne, Alain, Kawahara, Hideki. “YIN, a fundamental frequency estimator for speech and music”. Journal of the Acoustical Society of America, vol..111, pp.1917, 2002.
[4] Kato, Akihiro, and T. Kinnunen. "A Regression Model of Recurrent Deep Neural Networks for Noise Robust Estimation of the Fundamental Frequency Contour of Speech." 2018 Odyssey The Speaker and Language Recognition Workshop, 2018.
[5] Ziethe, et al. "Control of Fundamental Frequency in Dysphonic Patients During Phonation and Speech. " Journal of voice, vol. 33(6), pp. 851-859, 2019.