Acoustic Correlates of Voice Quality Improvement by Voice Training

Kiyoaki Aikawa¹, Junko Uenuma², and Tomoko Akitake²

¹ Tokyo University of Technology, School of Media Science, Japan
² Bijivo, Japan

aik@media.teu.ac.jp

Abstract

This paper derived four acoustic parameters reflecting voice quality improvement by voice training. Voice training corrects respiration, phonation, articulation, and facial expression. The purpose of this study is to develop acoustic parameters which quantitatively represent the quality of each training item. This paper examined six acoustical feature parameters on the performance as the measure for voice quality improvement. Experimental results indicated that four of the six parameters, the intensity of harmonic structure, the dynamic range of the spectral envelope, formant intensity, and spectral slope were significantly improved after the training. Further analyses were carried out on the correlation between subjective evaluation and above acoustic properties. The results indicated that abdominal respiration was most important for improving these four parameters.

Index Terms: voice training, voice quality, abdominal respiration, harmonic structure, formant intensity

1. Introduction

Voice quality has been discussed on speech synthesis [1], voice transmission such as VoIP [2-3], medical diagnosis [4-6], and emotion recognition [7]. Acoustical analysis on high-quality voice has been mainly studied on singing voice [8-12]. Some papers discuss the acoustical parameters for voice quality [13-15]. However, quantitative study has not been reported on the acoustical correlates reflecting voice quality improvement by voice training. The acoustical analyses for singing voice have been limited within F0 contour and singing formants. Voice training aims acquiring beautiful, powerful, and intelligible voice. Its purpose is not only for singing but also for lecture and business presentation. Voice training includes four training items, respiration, phonation, articulation, and facial expression. So far the voice quality has been evaluated based on the subjective impression by the voice training teachers. However, it is difficult for the teachers to quantitatively and simultaneously evaluate these four voice quality items. It is highly required to develop a set of parameters for quantitatively and simultaneously evaluating these voice quality items. So far, such study using actual speech data recorded at voice training schools has not been reported.

Section 2 describes acoustic features related to voice training. Section 3 derives six acoustic feature parameters for evaluating voice quality. Section 4 and 5 describes experimental results. Section 6 concludes this paper.

2. Voice Training

This section describes acoustical voice properties related to four training items performed at voice training schools.

2.1. Respiration

Abdominal respiration is requested instead of costal respiration in voice training. The abdominal respiration produces powerful and stable air flow. Thorax should be kept stable with enough volume inside even while respiration. This is effective to stabilize voicing. The movement of vocal cords depends on the pressure in the trachea. High tension of the vocal cords makes short glottal impulses. The intensity of high frequency harmonics increases depending on the glottis opening period. Figure 1 compares harmonic component distribution between wide and narrow glottal air flows. The temporal regularity of the vocal cord oscillation is also an important factor for producing clear harmonic structure. If the glottal air flow is the impulse train repeated at a constant interval, the each harmonic component shows a line spectrum. If the glottal pulse interval fluctuates, the harmonic component intensities are decreased and their bandwidths are widened. Jitter and shimmer features are reflected in this harmonic structure.

Preliminary analysis showed that the intensities of high frequency harmonics were increased by abdominal respiration. The low frequency harmonics showed very steep line-spectrum for high quality bel-canto singing voice.

(a) Open 2 ms, Close 3 ms (b) Open 0.5 ms, Close 4.5 ms

Figure 1: Comparison of harmonic component distribution between wide and narrow glottal air flows. (F0=200 Hz)

2.2. Phonation

Phonation is the process of controlling organs for appropriately generating voice source. Coordination of oral, nasal, and pharyngeal cavities is required. The muscle tension around glottis and the physical properties of trachea are related to the voice source quality. Tracheal stricture and glottal air leak result in generating noise. The noise degrades the harmonic structure. Therefore, HNR and SHR[14] features are reflected in the harmonic structure intensity.

2.3. Articulation

Articulation is the process of forming a vocal tract by morphing the mouth cavity and lip shapes. The nasal-cavity
resonance characteristics depend on the properties of soft palate and pharynx. The tongue position and the jaw angle determine the rough cavity shape. Clearly different positioning of articulatory organs is necessary for producing intelligible phonemic sound. The articulation features are represented by the formant frequency allocations and their resonance intensities. The resonance intensities can be represented by the Q-values of the vocal tract filter transfer function. Another possible measure for the resonance intensity is the dynamic range of the spectrum.

2.4. Facial expression

Facial expression emphasizes the mouth shape for pronouncing phones. The sound output efficiency is a function of the opening area of mouth. Therefore, the quality of facial expression is considered to appear in the dynamical feature of vocal power time sequences.

3. Voice Quality Parameters

This section formulates acoustic parameter candidates which can represent voice quality especially for voiced portion in speech. The effectiveness of each parameter for voice quality evaluation will be examined in the next section.

3.1. Harmonic Structure Intensity

According to the analysis on respiration in the previous section, the quality of voice sources can be measured by the intensity of harmonic structure. This feature indicates how clearly the harmonics of the fundamental frequency is observed in the spectrum. Cepstral coefficient corresponding to the fundamental frequency (F0) is available for measuring the intensity of the harmonic structure. Cepstrum is defined as the inverse Fourier transform of log spectrum. The F0 is obtained by the order of the maximum cepstral coefficient. Therefore, the maximum cepstrum value just represents the intensity of harmonic structure. The intensity of harmonic structure at frame n is given by

\[ h(n) = \max_{k \text{ min}-54 \leq k \leq \text{ max}} c_k(n). \]  

(1)

\[ F_0(n) = f_c / \arg \max_{k \text{ min}-54 \leq k \leq \text{ max}} c_k(n), \]  

(2)

where \( c_k(n) \) denotes the k-th order cepstral coefficient at frame n. \( k_{\text{min}} \) and \( k_{\text{max}} \) are determined by the minimum and maximum acceptable fundamental frequencies. In the following experiments, the minimum and the maximum frequencies were 80 Hz and 400 Hz, respectively. \( F_0 \) and \( f_c \) denote fundamental frequency and sampling frequency, respectively. If high harmonic components decrease, the harmonics structure intensity decreases.

3.2. Variance of Power Time Series

Vocal power increases depending on the opening area of mouth. The opening area depends on the phoneme to pronounce and the effort to open mouth. Therefore the effort to widely open mouth is considered to appear in the dynamic range of power time series. This paper uses the variance of log power time series for representing the power dynamic range.

3.3. Spectral Dynamics

If phonemes are clearly pronounced, the spectrum will drastically change while speaking. The k-th order delta-cepstrum is defined by the slope of the linear regression of k-th order cepstrum time series as follows

\[ d_k(n) = \frac{\sum_{i=0}^{L} w(i)c_k(n + i)}{\sum_{j=0}^{L} i^2}, \]  

(3)

where \( n \) is the frame number and \( w(i) \) is the triangular window. \( c_k(n) \) denotes k-th order cepstrum. Therefore, the L2-norm of delta-cepstrum

\[ D(n) = \sum_{k=1}^{N} d_k(n)^2, \]  

(4)

can represent an amount of spectrum change reflecting clear articulation.

3.4. Spectral Dynamic Range and Spectral Slope

The appearance of clear formants is related to intelligible articulation. Spectral dynamic range can be a measure for representing the formant resonance intensity. The variance of spectrum will be more appropriate for the measure of spectral dynamic range than the difference between the maximum and minimum power of the spectrum. This paper uses LPC-based spectral envelope given by

\[ S(z) = \frac{1}{\sum_{k=0}^{p} a_k z^{-k}}, \]  

(5)

where \( a_k \) is the k-th order linear predictive coefficient and \( p \) is the number of poles. Preemphasis given by

\[ H(z) = 1 - 0.98 z^{-1}, \]  

(6)

is operated for improving LPC-based spectral estimation. This paper uses the variance of the log spectral envelope for representing the dynamic range of the spectrum. The spectral slope should be removed in advance. This paper obtains the spectral slope at frame n by linear regression of the log of the spectral envelope \( S(n) \) as follows

\[ \nu(n) = \frac{\sum_{i=0}^{L} \log(S(i,n))(i-\frac{L}{2})}{\sum_{j=0}^{L}(i-\frac{L}{2})^2}, \]  

(7)

where \( i=l \) corresponds to the maximum frequency. The spectral slope value \( \nu(n) \) itself represents the high to low frequency component ratio. Therefore, this paper examines the effectiveness of the spectral slope for the measure of voice quality improvement.

3.5. Formant Intensity

The formant intensity is related to the quality of articulation. It corresponds to the resonance of the vocal tract filter. This paper uses Q-values of the vocal tract filter for representing formant intensities. The transfer function of the vocal tract filter given by Eq. (5) can be represented using the k-th pole \( u_k \) as Eq. (8). The frame number is omitted, hereafter.
\[ S(z) = \frac{z^p}{\sum_{k=1}^{p} (z - \lambda_k)} , \]  \hspace{1cm} (8)

where \( \lambda_k \) is the \( k \)-th pole given by
\[ \lambda_k = \gamma_k e^{j\theta_k} . \]  \hspace{1cm} (9)

The vocal tract filter shows multi-band-pass characteristics. The -3dB bandwidth \( b_k \) of the \( k \)-th spectral peak is given by
\[ b_k = 2 \arcsin \left( \frac{1 - r_k}{2r_k} \right) \sqrt{r_k^2 + 2r_k - 1} . \]  \hspace{1cm} (10)

Then the Q-value at the \( k \)-th spectral peak is given by
\[ Q_k = \frac{\theta_k}{b_k} . \]  \hspace{1cm} (11)

### 4. Parameter Selection Experiment

This section examines which acoustic parameter is effective for measuring the voice quality improvement.

#### 4.1. Data 1 for parameter selection

The subjects were ten males and two females from 20 to 54 years old. The speech data were sampled at 8 kHz and recorded in AMR format. The spectra were analyzed every 10 ms using 32ms Hanning window (256 samples). The number of poles for LPC analysis was 10. Utterances were isolated phrase sequences or continuous short sentences. The segmental acoustic parameter values were averaged for all voiced frames, subject by subject. The voiced frame was defined as one whose log power was greater than the average. Total number of frames analyzed for pre and post training were 9341 and 9811, respectively. This paper used a database which had already been recorded at a voice training school using a voice recorder of 8 kHz sampling frequency. It would take a long time to collect data for pre and post voice training. This was the reason why the sampling frequency was a little bit lower than that conventionally used for speech analysis.

#### 4.2. Results

Acoustical feature parameter values were compared between before and after the voice training. Table 1 summarizes the experimental results for the Data 1. The second column shows the p-value of ANOVA comparing pre and post training sets. The third column shows the number of subjects whose acoustic feature parameter value was increased after the training. It may sounds strange that a parameter value degraded after the voice training. However, even if a parameter value decreased, other parameter values increased. This depended on the instruction mainly given to the subject by the voice training teacher. The right column shows the p-value of T-test. The T-test analyzed whether the post-pre difference of the acoustical feature values obtained subject by subject was significantly positive or not.

ANOVA indicated that harmonics intensity, spectral dynamic range, and formant intensity showed statistically significant difference between pre and post training. T-test indicated further significant results. While spectral slope did not show significant difference by ANOVA, it showed significant difference by T-test.

Figures 2 and 3 show the results for the parameters which achieved most significant improvements. Figure 2 compares the harmonics intensities between pre and post training for each subject. The parameter value clearly increased for most of the subjects. Figure 3 compares the spectral dynamic range between pre and post training for each subject. The parameter value increased for all subjects. However, the increase was very small for subjects 4 and 9.

### Table 1. P-values of ANOVA and T-test (bold italic: significant). UP: number of subjects whose acoustic feature value was increased. ANOVA: comparison between pre and post training. T-TEST: T-test on the individual difference between pre and post training.

<table>
<thead>
<tr>
<th>Acoustic Feature</th>
<th>ANOVA</th>
<th>UP</th>
<th>T-TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harmonics Intensity</td>
<td>0.0084</td>
<td>11</td>
<td>0.0002</td>
</tr>
<tr>
<td>Power Variance</td>
<td>0.8288</td>
<td>7</td>
<td>0.3542</td>
</tr>
<tr>
<td>Spectral Dynamics</td>
<td>0.9628</td>
<td>5</td>
<td>0.4795</td>
</tr>
<tr>
<td>Spectral Dynamic Range</td>
<td>0.0217</td>
<td>12</td>
<td>0.0005</td>
</tr>
<tr>
<td>Formant Intensity</td>
<td>0.0041</td>
<td>11</td>
<td>0.0008</td>
</tr>
<tr>
<td>Spectral Slope</td>
<td>0.1217</td>
<td>11</td>
<td>0.0005</td>
</tr>
</tbody>
</table>

Figure 2: Comparison of harmonics intensities between pre and post training for individual subject.

Figure 3: Comparison of spectral dynamic range between pre and post training for individual subject.
5. Acoustic Correlate Analysis

Previous section indicated that four acoustic parameters reflected voice quality improvements. This section analyzes the correlation between the selected acoustic parameters and subjective evaluation on four voice training items, respiration, phonation, articulation, and facial expression.

5.1. Data 2 for acoustic correlate analysis

Another data set was collected for analyzing the correlation between subjective evaluation and acoustic feature parameter values. The subjects were three males and two females from 26 to 56 years old. The speech data were sampled at 44.1 kHz and recorded in WMA format. The data were down sampled into 11025 Hz, because the frequency range under around 6 kHz was appropriate for the analysis of this paper. The spectra were analyzed every 10 ms with 23.2ms Hanning window (256 samples). The frame width was different from that used in the previous section. The number of poles for LPC analysis was 10. Total number of training sessions was 9 for pre and post training, respectively. Utterances were isolated phrase sequences or continuous short sentences. Only the voiced frames were used for analysis. Total number of frames analyzed for pre and post training were 3352 and 3184, respectively.

5.2. Results

Table 2 analyzes the correlation between acoustic features and subjective evaluations by a voice training teacher for the Data 2. The choice of subjective evaluation was three levels, “very good”, “good”, and “no improvement”. T-test showed that harmonics intensity, spectral dynamic range, and formant intensity indicated significant difference between pre and post training. Table 2 shows that the quality of phonation is correlated with spectral dynamic range and harmonics intensity. Articulation is correlated with harmonics intensity. The subjective evaluation on respiration was “very good” for all subjects. Therefore, the correlation coefficient could not be calculated. This is the reason why respiration is not shown in Table 2. This fact indicated that harmonics intensity, spectral dynamic range, and formant intensity were strongly correlated with respiration quality.

Table 2. Correlation between acoustic features and subjective evaluations; P: phonation, A: articulation, and F: facial expression.

<table>
<thead>
<tr>
<th>Acoustic Feature</th>
<th>P</th>
<th>A</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harmonics Intensity</td>
<td>0.74</td>
<td>0.76</td>
<td>0.45</td>
</tr>
<tr>
<td>Spectral Dynamic Range</td>
<td>0.81</td>
<td>0.61</td>
<td>0.45</td>
</tr>
<tr>
<td>Formant Intensity</td>
<td>0.45</td>
<td>0.69</td>
<td>0.40</td>
</tr>
<tr>
<td>Spectral Slope</td>
<td>-0.29</td>
<td>-0.28</td>
<td>0.01</td>
</tr>
</tbody>
</table>

6. Conclusions

This paper revealed that four acoustic parameters well represented voice quality improvement by voice training. The four parameters included harmonic structure intensity, spectral dynamic range, formant intensity, and spectral slope. Experimental results indicated that these four parameters were significantly improved by voice training. Experimental results suggested that voice training mainly improves the harmonic structure, the intensities of high frequency harmonic components and vocal tract resonance. Correlation analysis between subjective evaluation and acoustic parameters clarified that the training of abdominal respiration was most important for improving the above four acoustic parameters.

7. Acknowledgements

This work was supported by the Open Research Center Fund of MEXT Japan.

8. References