INTERSPEECH 2009 BRIGHTON

Acoustic and Perceptual Effects of Vocal training in Amateur Male Singing

Takeshi SAITOU and Masataka GOTO

National Institute of Advanced Industrial Science and Technology (AIST)

{saitou-t,m.goto}[at]aist.go.jp

Abstract

This paper reports our investigation of the acoustic effects of vocal training for amateur singers and of the contribution of those effects to perceived vocal quality. Recording singing voices before and after vocal training and then analyzing changes in acoustic parameters with a focus on features unique to singing voices, we found that two different F0 fluctuations (vibrato and overshoot) and singing formant were improved by the training. The results of psychoacoustic experiments showed that perceived voice quality was influenced more by the changes of F0 characteristics than by the changes of spectral characteristics and that acoustic features unique to singing voices contribute to perceived voice quality in the following order: vibrato, singing formant, overshoot, and preparation.

Index Terms: singing voice, vocal training, psychoacoustic experiment

1. Introduction

The expression of singing voices varies greatly according to singer and singing style, and many amateur singers therefore take vocal training lessons in an effort to acquire a professional-level singing voice. Clarifying the acoustical and perceptual effects of vocal training will help us understand the mechanisms of singing voice production and perception, will help amateur singers acquire a professional-level singing voice, and will contribute to the development of high-quality singing voice synthesis systems. In this study we therefore investigate the ways in which the acoustic features and perceived quality of the singing voices of amateur singers were affected by vocal training.

Many studies have investigated the acoustic features unique to singing voices and some of them focused on relations between those features and the perceived voice quality. Bartholomew dealt with the acoustic features of a good singing voice [1], whereas Sundberg dealt with those affecting vocal ugliness [2]. Saitou, on the other hand, investigated acoustic features affecting perception of singing-ness [3]. Each of these investigators found that the perception of voice quality depends on acoustic features in fundamental frequency (F0) contour and spectrum, but the relations between those features and vocal training are not clear.

This paper therefore reports our investigation of the acoustic and perceptual effects of vocal training in amateur singers through acoustic analysis and two psychoacoustic experiments. Section 2 introduces singing voice data recorded before and after vocal training. Section 3 focuses on four different acoustic features unique to singing voices and shows the acoustic effects of vocal training. Section 4 shows experimental results demonstrating the effects of F0 and spectrum on perceived vocal quality, and Section 5 presents experimental results showing that the four different acoustic features in F0 and spectrum contribute to the perception of voice quality. Section 6 concludes the paper by summarizing the contributions of this research.

2. Singing voice data

We recorded amateur singers singing two Japanese popular songs before and after vocal training. The singers were three males who had never taken vocal training lessons, and the vocal training teacher was a professional tenor with a 20-year teaching career. Each singer received a total of nine hours of vocal training over three days. We also recorded the teacher singing the same songs. The singing voices were recorded in soundproof room at a sampling rate of 48 kHz with 24-bit resolution by using a microphone (SHURE SM87A) and a solid-state recorder (Marantz PMD671). The data was downsampled to 24 kHz and converted to 16-bit resolution. In this paper the singing voices that were recorded before and after vocal training are respectively referred to as untrained and trained voices.

To evaluate qualities of untrained and trained voices subjectively, we conducted brief psychoacoustic experiments in which the subjects were fifteen adults (nine male and six female) who had at least two years of vocal training. None of the subjects had hearing impairments. In the experiments the subjects listened to pairs of untrained and trained voices and judged which of the pair was better. The voices were presented at a comfortable loudness level through binaural earphones (Sennheizer HDA200). For both songs, all subjects judged the trained voices to be better than the untrained voices.

3. Acoustic effects of vocal training

To investigate acoustic effects of vocal training, we analyzed features in F0 contour and spectrum by using STRAIGHT [4] and then compared these features between untrained and trained voices.

3.1. F0 analysis

In the F0 contours of singing voices, there are several fluctuations unique to singing voices, and vibrato is the one usually related to professional voice quality [1, 2, 3]. Perceived quality of singing voices has also been reported to be influenced by other fluctuations [5]. We therefore analyzed three different F0 fluctuations, and compared each fluctuation between untrained and trained voices.

vibrato: a quasi-periodic frequency modulation (4–7 Hz) [6].

overshoot: a deflection exceeding the target note after a note change [5, 7].

preparation: a deflection in the direction opposite to a note change observed just before the note change [5].

Figure 1 shows examples of the F0 contours of an untrained voice and a trained voice, and also shows examples of the three

Table 1: Characteristics of F0 fluctuations in untrained and trained voices of amateur singers and a professional singer (the vocal training teacher).

	Characteristics of F0 fluctuations (untrained → trained)					
	Vibrato		Overshoot		Preparation	
singer	rate (variance) [Hz]	extent (variance) [cent]	extent [%]	duration [ms]	extent [%]	duration [ms]
A	5.2(2.6) → 5.3(1.4)	41(10.2) → 52(6.6)	18 → 14	223 → 196	10.7 → 10.3	126 → 130
B	4.9(3.1) → 4.8(3.3)	43(23.4) → 49(16.1)	16 → 17	204 → 222	11.6 → 10.9	136 → 128
C	5.3(3.3) → 5.2(2.6)	48(20.8) → 57(14.3)	17 → 13	198 → 163	9.3 → 8.9	113 → 115
pro	6.3(1.3)	65(9.5)	11	126	9.8	119

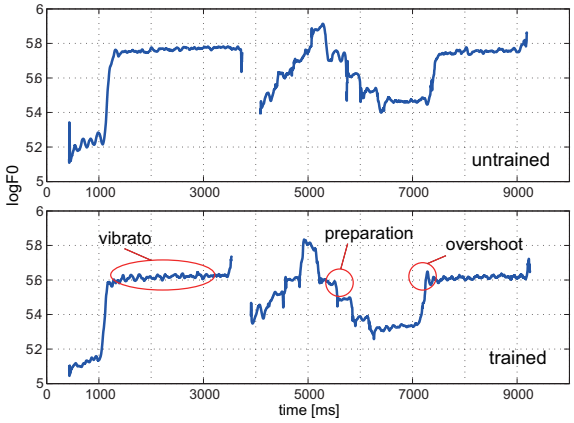

Figure 1: *F0 contours of singing voices of singer C.*

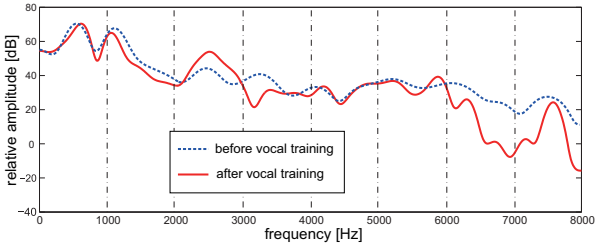

Figure 2: *Spectral envelopes of the vowel /a/ in the untrained and trained voices of singer A.*

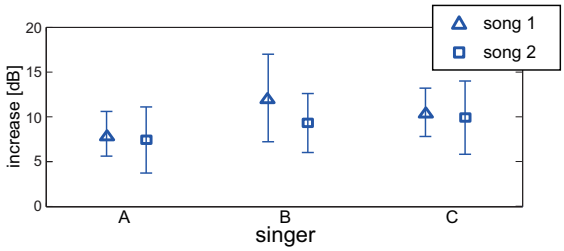

Figure 3: *Increased amplitude (mean ±SE) of spectral peaks near 3 kHz in two songs sung by the amateur singers.*

F0 fluctuations. Characteristics of the F0 fluctuation in the amateur and professional singing voices are listed in Table 1.

Vibrato was analyzed by measuring its rate (modulation frequency) and extent (amplitude). In each singer, the vibrato extent of the trained voice was larger than that of the untrained voice and approximated the vibrato extent of the professional singer's voice. And because the variance of vibrato extent in each of the trained voices was smaller than that in each of the corresponding untrained voice, the vibrato in the trained voices was steadier than that in the untrained voices. The vibrato rate and its variance, however, were not improved by vocal training. These results indicate that vibrato rate is harder to control than vibrato extent.

Overshoot was analyzed by measuring its extent and duration. The extent indicates an F0 variance exceeding the target note and the duration means a stabilization time. As shown in Table 1, for two of the amateur singers (A and C) the vocal training reduced the extent and duration of the overshoot to amounts approximating those in the voice of the professional singer. Since Krom [8] had reported that controlling overshoot was one of the expression ways in professional singing, the analysis results indicate that the vocal technique of singer A and C was improved by training.

Preparation was also analyzed by measuring its extent and duration. Although vocal training changed both of these characteristics, the changes were smaller than the training-induced changes in the extent and duration of overshoot.

3.2. Spetcrum analysis

Sundberg showed that the spectral envelope of a male operatic singing had a remarkable peak called the "singing formant" near 3 kHz and that the peak strongly affected the perceived quality of the singing voice [9]. Because a singing formant has also been reported to be important in some traditional Japanese singing [10], we analyzed the effects of vocal training on spectral peaks near 3 kHz.

Examples of spectral envelopes of the vowel /a/ in the singing voices of singer A are shown in Figure 2, where one sees that a spectral peak near 2.6 kHz is higher in the trained voice. Increases in the amplitude of the spectral peak near 3 kHz after vocal training are shown in Fig. 3. This figure shows that the spectral peaks, in both songs of all singers, increased by vocal training and the increases in the peaks was 13 dB on the average. On the other hand, the spectral peaks in professional voices was 18 dB higher on average than that in untrained voices. These results suggest the possibility that a singing formant is generated by vocal training.

4. Psychoacoustic experiment 1

Psychoacoustic experiment 1 was conducted to investigate how perceived voice quality is affected by the F0 contour and spec-

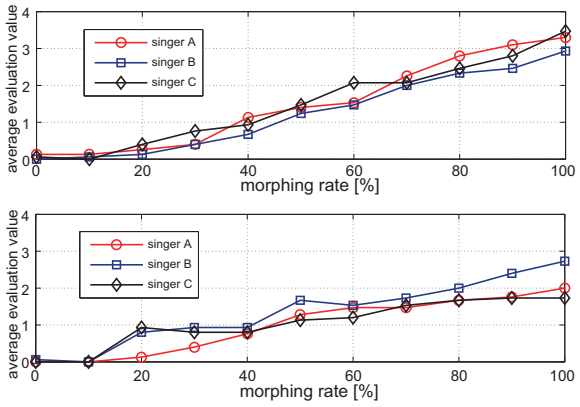

Figure 4: *Result of the psychoacoustic experiment 1: the evaluation value of voice quality for F0 (upper fig.) and spectral (lower fig.) morphing.*

trum changes resulting from vocal training.

4.1. Stimuli

Two kinds of stimulus sets were made by using STRAIGHT-morphing [11]. The first set consisted of singing voices made by morphing F0 contour from the untrained voice to the trained voice. Stimuli of the second set were made by morphing spectral envelopes. In making each stimulus set, morphing rate was changed from 0 to 100% in 10% steps. The morphing rate "0%" corresponds to the untrained voice and "100%" corresponds to the trained voice. There were 66 stimuli in each set (11 morphing rates $\times$ 3 singers $\times$ 2 songs).

4.2. Methods

The experiment was conducted in accordance with the paired comparison methodology. The subjects were sixteen adults (eleven male and five female) with normal hearing ability. All had taken vocal training lesson for over a year.

The subjects listened to pairs of stimuli presented in random order at intervals of 500 ms. The first stimulus of a pair was an untrained voice, and the second was a morphed singing voice of the stimulus sets. The stimuli were presented at a comfortable loudness level through binaural earphones (Sennheizer HDA200). Since the subjects listened to each pair only once, the number of trials was 132.

The subjects were asked to evaluate a vocal quality of the second stimulus relative to that of the first on a five-grade scale: 0 for worse, 1 for almost the same, 2 for slightly better, 3 for clearly better, and 4 for much better. To confirm the criterion of judgment, the subjects had a rehearsal session with ten pairs prior to each actual experiment. The experiment was conducted in a sound-proof room.

4.3. Results

Figure 4 shows experimental results. The horizontal and vertical axes respectively show a morphing rate and an evaluation value. As shown in each figure, the evaluation value of the stimulus was increased steadily when the morphing rate was increased. Moreover, in the case of over 60% morphing, the perceptual effect of F0 morphing is obviously larger than that of spectral morphing. These results indicate that perceived voice quality is affected more by the changes of F0 characteristics than by the changes of spectral characteristics.

5. Psychoacoustic experiment 2

Psychoacoustic experiment 2 was conducted to examine the contributions of three different F0 fluctuations (vibrato, overshoot, and preparation) and singing formant to the perception of voice quality.

5.1. Stimuli

Stimuli were made by using our recently developed system that can synthesize a singing voice when given the musical score of a song and a speaking voice reading the lyrics of that song [12]. As shown in Fig. 5, the system is based on *STRAIGHT* and comprises three models controlling three acoustic parameters: the F0, phoneme duration, and spectrum. The three F0 fluctuations and singing formant can be controlled by the F0 and spectral control models. In this experiment we put the untrained singing voices of the three amateur singers into the system and made six different singing voices by replacing characteristics of the four acoustic features with professional characteristics.

UNTRAINED: singing voice for which all acoustic features are controlled with untrained characteristics.

PRO-VB: singing voice for which only the **UNTRAINED** vibrato is controlled with professional characteristics.

PRO-OS: singing voice for which only the **UNTRAINED** overshoot is controlled with professional characteristics.

PRO-PR: singing voice for which only the **UNTRAINED** preparation is controlled with professional characteristics.

PRO-SF: singing voice for which only the **UNTRAINED** singing formant is controlled with professional characteristics.

PRO-ALL: singing voice for which all acoustic features are controlled with professional characteristics.

In synthesizing each stimulus, the F0 contour was generated first by adding the three fluctuations to a melody contour which is an input of the F0 control model, and then the generated F0 contour was replaced with that of the input singing voice. The untrained and professional characteristics for each F0 fluctuation were set to the values shown in Table. 1. Spectral envelopes, on the other hand, were generated by emphasizing the spectral peak near 3 kHz by 18 dB during vowel portion of the input singing voice. By using the generated F0 contour and modified spectral envelopes, all stimuli were synthesized.

5.2. Methods

The experiment was conducted in accordance with the Scheffe's paired-comparison method (Ura's modified method) [13]. The subjects were the same as the experiment 1.

The subjects listened to pairs of stimuli presented in random order at intervals of 500 ms. Each pair consists of two of the stimuli, the total number of the pairs was 180 ($=_6P_2 \times 3$ singers $\times$ 2 songs). The subjects were asked to evaluate the vocal quality of stimuli on a seven-step scale ranging from "-3" (The first stimulus is very good singing in comparison with the second) to "+3" (The second stimulus is very good singing in comparison with the first). Experimental environments were the same as in experiment 1.

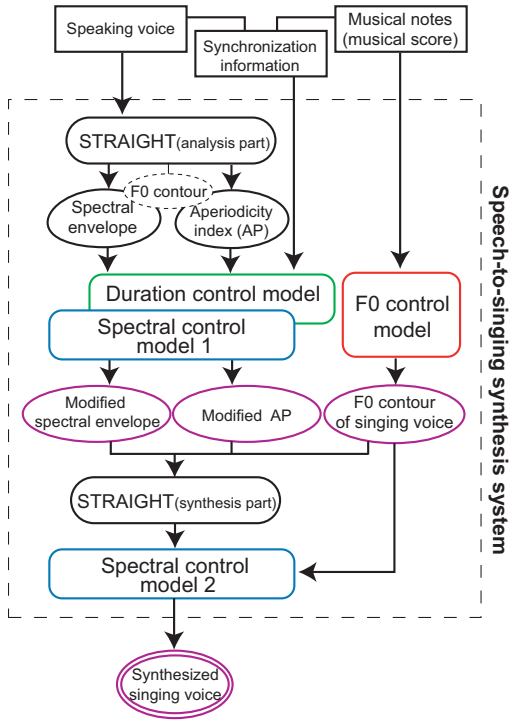

Figure 5: *Block diagram of the speech-to-singing synthesis system [12].*

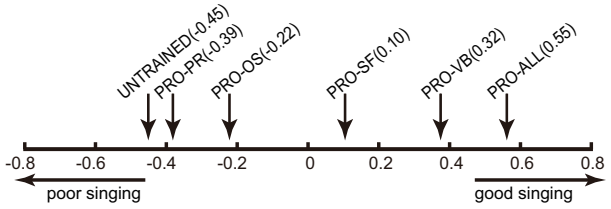

Figure 6: *Results of the psychoacoustic experiment 2: subjectively evaluated degree of the vocal quality for 6 stimuli.*

5.3. Results

Figure 6 shows the experimental results. The numbers under the horizontal axis indicates the degree of the vocal quality of the stimuli. The result of the F-test confirmed that there are significant differences amongst all stimuli at the 5% critical rate.

This shows that the vocal quality of the stimuli can be increased by converting characteristics of acoustic features to professional ones. Moreover, the value of **PRO-VB** was better than that of **PRO-OS**, **PRO-PR**, and **PRO-SF**, indicating that the perceptual effect of vibrato was greater than that of any of the other three acoustic features. These results also show the following ranking (in decreasing order of importance) of the acoustic features with regard to their effects on perceived vocal quality: vibrato, singing formant, overshoot, and preparation. However, overshoot and preparation contributed to perceived vocal quality less than the other features.

6. Conclusion

This paper described our investigation of the acoustic and perceptual effects of vocal training in amateur male singing. Comparing acoustic features unique to singing voices in recordings made before and after vocal training, we found that characteristics of two kinds of F0 fluctuations (vibrato and overshoot) and singing formant were changed by vocal training. Then results of psychoacoustic experiments showed that changes of F0 characteristics affected voice quality more than changes of spectral characteristics did and that acoustic features affecting the perception of the quality of singing voices can be ranked in the following descending order of importance: vibrato, singing formant, overshoot, and preparation. In future work we will investigate the relations between acoustic features and vocal training in more detail and use the results of those investigations to develop a high-quality singing voice synthesis system.

7. Acknowledgements

We thank Ken-Ichi Sakakibara for many useful comments and much good advice. This work was supported by CrestMuse, CREST, JST.

8. References

[1] W. T. Bartholomew, “A Physical Definition of ”Good Voice-Quality” in the Male Voice,” J. Acoust. Soc. Am., Vol.55, 838-844, 1934.

[2] J. Sundberg, “The KTH synthesis of singing,” Advances in Cognitive Psychology. Special issue on Music Performance, 2(2-3), 131-143, 2006.

[3] T. Saitou *et al.*, “Analysis of acoustic features affecting “singingnes” and its application to singing voice synthesis from speaking voice,” Proc. ICSLP2004, Vol. III, pp. 1929-1932, 2004.

[4] H. Kawahara, *et al*., “Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency based on F0 extraction: Possible role of a repetitive structure in sounds,” Speech Commun., Vol. 27, pp. 187-207, 1999.

[5] T. Saitou, *et al*., “Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis,” Speech Commun., Vol. 46, pp. 405-417, 2005.

[6] C. E. Seashore, “The Vibrato,” University of Iowa Studies in the Psychology of Music, Vol. I, 1932.

[7] H. Mori, *et al*., “F0 dynamics in singing: Evidence from the data of a baritone singer,” IEICE Trans. Inf. & Syst., Vol. E87-D, No. 5, pp. 1086-1092, 2004.

[8] G. de Krom, *et al*., “Timing and accuracy of fundamental frequency changes in singing,” Proc. ICPhS 95, Vol. I, pp. 206-209, 1995.

[9] J. Sundberg, “The Science of Singing Voice,” Northern Illinois University Press, 1987.

[10] I. Nakayama, “Comparative studies on vocal expression in Japanese traditional and western classical-style singing, using a common verse,” Proc. ICA2004, 1295-1296, 2004.

[11] H. Kawahara, and H. Matsui, “Auditory Morphing based on an Elastic Perceptual Distance Metric in an Interference-free Time-frequency Representation,” Proc. 2003 IEEE ICASSP, I, pp. 256-259 (2003).

[12] T. Saitou, *et al*., “Speech-To-Singing Synthesis: Converting Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices,” Proc. 2007 Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 215-218 (2007).

[13] S. Ura, *et al*., “Sensory Evaluation Handbook (in Japanese),” JUSE Press Ltd., pp. 366-384, 1973.