Controlling “shout” expression in a Japanese POP singing performance: analysis and suppression study

Yuri Nishigaki1, Ken-Ichi Sakakibara2, Masanori Morise3, Ryuichi Nisimura,4 Toshio Irino4, Hideki Kawahara4
1Graduate School of Systems Engineering, Wakayama University, Wakayama, Japan
2Department of Communication Disorders, Health Science University of Hokkaido, Sapporo, Japan
3Faculty of Engineering, University of Yamanashi, Kofu, Japan
4Department of Design Information Sciences, Wakayama University, Wakayama, Japan
kis@hoku-iryo-u.ac.jp, mmorise@yamanashi.ac.jp, kawahara@sys.wakayama-u.ac.jp

Abstract

Degree of “shout” singing performance is effectively controlled by combining global spectral shape equalization, peak cancellation in frequency modulation spectrum of F0 trajectory, and synchronized shape-modulation of voice spectral envelope. This “shout-reduction” processing is based on a symmetry-based F0 extractor with fine temporal resolution, a temporally stable representation of instantaneous frequency of periodic signals, and the TANDEM-STRAIGHT, a speech analysis, modification and resynthesis framework. The proposed procedure successfully converted an expressive Japanese POP song performance with “shout” into a plain performance without damaging original naturalness. Possibility of adding artificial “shout” to plain performance is also discussed.

Index Terms: singing voice, shout, fundamental frequency, spectral modulation, frequency modulation

1. Introduction

Singing performance explores various aspects of voice quality [1, 2]. “Shout” is one such expression effectively used by professional singers and makes them sound impressive. Enabling flexible control of such expression in the post-production processing of recorded singing performances is a challenging topic and will become very useful. It also applicable to enhance expressions of artificial singers, such as a famous vocaloid [4]. “Hatune Miku.” [5, 6]

The next section briefly reviews research topics of analysis and synthesis of singing voices and describes their relations to this article. The following section introduces specifically adopted tools and procedures employed in this investigation. These tools and procedures are employed to analyze performances of a Japanese POP song and “shout” specific features are illustrated. Based on these observations, a systematic test on “shout” impression suppression is introduced to investigate physical correlates of the impression. The test stimuli were prepared by systematically removing “shout”-specific features using an extended TANDEM-STRAIGHT [7] framework with finer temporal resolution in both F0 and spectral envelope extraction. Finally, test results and discussions are presented.

2. Background

Production and physical attributes of singing voices have been studied mainly on western classical music and revealed physical correlates of specific features of singing voices, such as dynamics of F0 trajectories, vibrato, trill, voice register, singing formant, formant matching and so on [8, 9, 10, 11]. These findings have been introduced to implement various types of singing voice synthesis systems [12, 13, 14, 15] and improved quality related spectral aspects and dynamic aspects of F0 trajectories in synthesized singing voices. Also, by introducing those singing specific features, speaking voice to singing voice conversion was successfully demonstrated [16]. However, once the target singing is expanded to cover ethnic or popular music, detailed and new aspects of vocal expressions are needed to be studied. They involve non-linear interactions between source and filter parts, irregular excitation of voice source, and additional modulations by supra-laryngeal structures [1, 2, 17, 18]. These deviations from classical western singing result in frequency modulations of F0 as well as spectral shape modulations. These modulations are usually faster than the speed limit that conventional F0 extractors [19, 20, 21] can follow. Our temporally-fine processing tools were developed to investigate these aspects.

3. Tools and procedures

This section introduces test tools adopted for investigating “shout” specific features and procedures to analyze them. They are based on an F0 extractor with finer temporal resolution that is capable of analyzing fast modulations found in those wider vocal expressions mentioned in the previous section.

3.1. F0 extractor with finer temporal resolution

The introduced method is based on higher-order waveform symmetry of the fundamental component [22]. The F0 value is defined by the reciprocal of the zero-crossing interval of the fundamental component in this method, where higher-order waveform symmetry is used to select the best filter output from a set of specially designed low-pass filters [23, 22]. The filter nominal frequencies are log-linearly allocated and implemented using one of Nuttall’s windowing functions [24] to assure low sidelobe levels (< −90 dB) and reasonably steep sidelobe decay rate (−18 dB/dec).

Frequency of the best of candidates in terms of symmetry is used to set the initial F0 value for the refinement procedure based on a temporally stable representation of the instantaneous
frequency of periodic signals [25]. The whole procedure of this F0 estimation runs faster than real-time when 1 ms frame rate and 40 Hz to 800 Hz F0 search range are applied for analyzing 44100 Hz sampled signals. (MacBookPro 2.6 GHz Intel Core i7, 16 GB memory, OS X 10.8.3, Matlab R2012b)

The modulation transfer function of the F0 extractor to frequency modulation of the fundamental component shows a smooth low-pass shape. The nominal cutoff frequency of this F0 extractor in terms of the FM modulation frequency is about 50 Hz for signals with 200 Hz average F0s [23, 22].

3.2. TANDEM-STRAIGHT with finer temporal resolution

In this research, a finer frame update rate (1 ms) is used for investigating the rapid and possibly synchronized spectral envelope modulation, found in the preliminary study [26], using the TANDEM-STRAIGHT spectral analysis. The TANDEM-STRAIGHT provides a time-frequency representation without trace of interferences due to periodicity of the analyzed periodic signals [7]. It is based on a temporally stable power spectral representation of periodic signals (TANDEM process) [27] followed by a periodicity suppression process on the frequency domain (STRAIGHT process) based on the consistent sampling theory [28]. While the whole procedures in TANDEM and STRAIGHT are heavily dependent on the estimated F0, they do not require any time-alignment to pitch marks. This flexibility of analysis frame allocation enabled finer frame update rate.

Since the TANDEM procedure uses a set of F0-adaptive time windows, the temporal resolution is also proportional to the length of the F0-adaptive time windows. The duration of the effective composite window in the TANDEM process is about one half of the nominal length of the constituent window [29]. Specifically, for example, the cutoff frequency in terms of amplitude modulation transfer function of spectral envelope calculation is about 70 Hz for signals with 200 Hz average F0.

3.3. Procedures for feature analysis

In addition to extracted STRAIGHT spectrograms and F0 trajectories, several statistical values are calculated to illustrate features of “shout” expression. These representations are intentionally kept low-level since the current research is an exploratory stage. In such stage, it is important to make interpretation and control straightforward by using conceptually and computationally transparent representations.

3.3.1. F0 frequency modulation (modulation power spectrum)

First, rapid F0 modulation is analyzed in terms of power spectrum of the modulation, \( P_0[k] \). Differentiated signals of the logarithmic F0 to base 2, \( l_0[n] \), are used in the analyses since our (musical) pitch perception is approximately log-linear. Differentiation is introduced to make the signals zero-mean.

\[
I_0[n] = \frac{\log_2(f_0[n+1]) - \log_2(f_0[n])}{t[n+1] - t[n]} \tag{1}
\]

\[
P_0[k] = \frac{\sum_{m=0}^{N_s-1} w[m] l_0[m + b_S] \exp(-j \frac{2\pi km}{M})^2}{N_s} \tag{2}
\]

where \( f_0[n] \) represents the fundamental frequency at \( n \)-th frame and \( t[n] \) represents the corresponding time of the \( n \)-th frame. The number \( N_s \) represents the length of the windowing function \( w[m] \) of the \( S \)-th voiced segment of the F0 trajectory. The initial frame index of the segment is represented by \( b_S \).

A discrete frequency index \( k \) corresponds to the modulation frequency \( f[k] = kf_s/M \), where \( f_s \) represents the analysis frame update rate (frequency in Hz) and \( M \) represents the buffer length of discrete Fourier transform. A set of contiguous \( P_0[k] \) are summed to yield \( P_0[k] \), which represents the modulation power spectrum of a larger musical unit “phrase” that consists of several bars. The power spectrum is calibrated to represent speed of logarithmic F0 change in terms of octaves per 1 Hz.

3.3.2. F0 frequency modulation (AR model analysis)

Since resonant-like spectral peaks were found in modulation power spectra of “shout” expression, autoregressive (AR) model [30, 31] parameters \( \{a[k]\}_k \) are estimated from the auto-covariance coefficients \( r[n] \) calculated by inverse Fourier transform of \( P_0[k] \).

\[
r[n] = \frac{1}{M} \sum_{k=0}^{M-1} P_0[k] \exp\left( \frac{j 2\pi kn}{M} \right) \tag{3}
\]

\[
a = R^{-1}r \tag{4}
\]

where vector notations of the AR coefficients \( \{a[k]\}_k \), auto-covariance coefficients \( \{r[k]\}_k \), and a Toeplitz matrix \( \{R_{k,l}\}_k \) are introduced. Frequencies and bandwidths of the resonant peaks are calculated from the roots \( \{z_k\}_k \) of the following polynomial equation.

\[
1 - \sum_{k=1}^{K} a[k] z^{-k} = 0, \tag{5}
\]

where the imaginary part and the real part of the complex number \( z_k \) provide the frequency and the bandwidth respectively.

3.3.3. Octave-band level difference

Phrase-wise long-term spectra \( I_{S_0}[k] \) are calculated from STRAIGHT spectra \( P_{ST}(f,t) \) to illustrate the global spectral shape difference between “shout” and “plain” performances.

\[
L_p[k] = 10\log_{10} \left[ \frac{1}{N_s} \sum_{m \in S_0} P_{ST}(\lambda, l[m]) d\lambda \right], \tag{6}
\]

where \( S_0 \) represents a set of frame indices \( m \) of the voiced segments in the \( p \)-th musical “phrase” and \( N_s \) represents the total number of voiced frames in the \( p \)-th “phrase.” The frequencies \( f_S[k] \) and \( f_H[k] \) represent the lower and the higher frequency boundary of the \( k \)-th frequency band, respectively. In the current analysis, one-third octave bands are used.

3.3.4. Amplitude modulation of octave-band level

Differentiated logarithmic octave band level modulation in the voiced segments are analyzed in terms of amplitude modulation power spectra. The results are calibrated to represent the speed of modulation in terms of dB/s.

4. Analysis results

4.1. Test materials

The test material was selected from a database that consists of performances of the original J-POP song “Ride”, composed for research use, sang in several different styles. It was performed by a Japanese professional male singer. The selected performances were sung in his usual expressive singing (full
of “shout”) and in plain singing style. They were recorded at 44100 Hz with 16 bit. The song is 157 s long, and only voiced segments were analyzed using a semi-automatic voice activity detection procedure based on the segmental power and the symmetry measure of the F0 extractor [22]. Two performances were subdivided into 32 musical “phrase” each. One “phrase” consisted of 2 to 5 voiced segments.

4.2. F0 trajectories and modulation power spectrum

Figure 1 shows waveforms and F0 trajectories of selected performances of the 17-th “phrase” of the J-POP song. The 17-th “phrase” is one of the most expressive part (the corresponding lyrics is “zare goto mo tsurai” in Japanese). The trajectories were extracted by the temporally fine-resolution F0 extractor. Only voiced segments are displayed in the F0 plot. Fast vibrations visible on the “shout” F0 trajectory are mainly due to the modulation spectral peak described in the next paragraph.

Figure 2 shows the power spectra of the differentiated logarithmic F0 trajectories with “shout” (blue line) and “plain” (red line) expressions.

Figure 3: Waveforms of the 17-th “phrase” with “shout” (blue line) and “plain” (red line) expressions in the upper plot. The lower plot shows F0 trajectories of voiced segments with “shout” (blue line) and “plain” (red line) expressions.

4.3. Spectral envelope variations

Figure 3 shows STRAIGHT spectrograms of the same “phrase” performed in different singing styles. The “shout” spectrogram shows a specific texture consisting of vertical lines especially around 3 kHz. The amplitude modulation power spectrum of the differentiated logarithmic octave band level around this frequency region also has the similar peak around 70 Hz.

4.4. Octave-band level difference

Figure 4 shows the one-third octave band level difference between “shout” and “plain” performances.

Figure 4: Difference of the one-third octave band levels between “shout” and “plain” performances.
“shout” expression in Fig. 3.

4.5. Summary of the test results
All 32 “phrases” were analyzed and three common features of “shout” expression were observed. First feature (feature “Q”) is the peak around 70 Hz in the power spectrum of the differentiated log frequency F0 trajectories. Second feature (feature “F”) is the peak around 70 Hz in the power spectrum of the differentiated dB octave band levels. Third feature (feature “E”) is the positive level region around 2 kHz in the band level difference.

5. Suppression study
Simple resynthesis of the “shout” performance using the analysis results by the extended temporally-fine TANDEM-STRAIGHT procedure yielded perceptually equivalent reproduction of the “shout” impression. This makes it possible to evaluate contributions of the above mentioned three features by testing resynthesized performance using all possible suppression patterns of these features.

5.1. Procedures
5.1.1. F0 frequency modulation: feature “Q”
The anti-resonance filter \( Q(z) \) is designed to cancel the complex conjugate poles \( z_k, z_k^* \) associated with the peak around 70 Hz, that is estimated by the AR model. It is cascaded to the 2nd order critical damping low-pass filter \( T(z) \) consisted of complex conjugate poles \( z_q, z_q^* \) with the same nominal frequency to yield a notch filter. In addition to this notch filter, the following bilinear equalizer \( M(z) \) is cascaded to adjust level differences in the higher modulation frequency region (\( > 20 \) Hz).
\[
M(z) = \frac{1 - r_L z^{-1}}{1 - r_H z^{-1}}, \tag{7}
\]
where \( r_L \) and \( r_H \) are determined to adjust the lower and higher corner frequencies.

5.1.2. Spectral shape modulation: feature “F”
The same modulation suppression filter \( Q(z)T(z)M(z) \) is used to suppress the spectral shape modulations of the dB STRAIGHT spectrogram.

5.1.3. Octave-band level difference: feature “E”
The piecewise linearly interpolated negative dB difference, for example, shown in Fig. 4 was used to design a spectral level equalizing FIR filter. A 10 ms size Hann window is applied to truncate the response.

5.2. Subjective evaluation
A subjective evaluation test was conducted using resynthesized stimuli listed in Table 1. Each stimulus label represents the suppression condition. For example, “Q0F0E0” indicates that features “Q”, “F” and “E” are all suppressed (\( = 0 \)). Three male and three female subjects were participated in the test. All possible combinations excluding the same stimulus pairs are counter-balanced and randomized. An interactive test program implementing 2AFC paradigm was used in the test. The subject was instructed to select the stimulus with stronger impression of “shout” from sequentially presented two stimuli.
Table 2 shows the test results. The numbers in the table represents the count of the selection pattern. For example, the number 13 in the “Q1F1E0” and “010” box represents that the stimulus with the label “Q1F1E0” is selected 13 times when compared with the stimulus “Q0F1E0.” The average preference differences due to feature “Q”, “F” and “E” are 12.5%, 47.6% and 16.7% respectively. A generalized linear model (GLM) was also applied to analyze the results. The logit model of GLM [32, 33] was applied using the statistical package R [34] and revealed that all three factors are highly significant. It yielded
\[
y = 1.231Q + 4.58F + 1.61E, \tag{8}
\]
where each feature has value “1” when it exists and “0” when it doesn’t. \( y \) is the logit converted odds. These suggest that all these features are contributing to “shout” impression. It also suggests that it may be possible to convert a “plain” performance to a “shout” performance by manipulating these features. Test stimuli used in this article are stored in this conference proceedings. Other samples are linked in our web page [35].

6. Conclusion
Introduction of an F0 extractor with fine temporal resolution into the TANDEM-STRAIGHT, speech analysis, modification and resynthesis framework, enabled reconstruction of a professional performance of a Japanese POP song with “shout” expression. It also revealed that there are three contributing factors of this “shout” impression. The most important factor is high modulation frequency (around 70 Hz) peak in the power spectrum of shape-modulation of voice spectral envelope. The secondary factor is global spectral peak around 2 kHz. The third factor is high modulation frequency (around 70 Hz) peak in the frequency modulation power spectrum of F0 trajectory. Removing all these contributing factors effectively converted an expressive “shout” performance into a plain performance without introducing severe artifacts. Development of an algorithm for inverse conversion from a plain expression to an expressive “shout” expression is currently underway.

7. Acknowledgements
This work is supported by the Aid for Scientific Research No.24300073, 23700221 and 24650085 from JSPS.
8. References


[23] ——, “Deviation measure of waveform symmetry and its application to high-speed and temporally-fine F0 extraction for vocal sound texture manipulation,” in Interspeech2012, 2012, session: O2d.05.


