2018 RISP International Workshop on Nonlinear Circuits,
Communications and Signal Processing (NCSP2018)
Honolulu, Hawaii, USA, March 4-7, 2018

Synthesis of Expressive Singing Voice by F0, Amplitude Envelope and Spectral Feature Conversion.

Thi-Hao Nguyen, Masato Akagi

Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 932-1292, JAPAN
Email: {s1610058, akagi}@jaist.ac.jp

Abstract

This paper investigates correlates of acoustic features to emotional singing voices. By analyzing acoustic features that are strongly related to emotions, this research determines which feature is more significant to the emotional expressions in singing voices. We also propose a method to modify amplitude envelopes based on the entire F0 contour to have a higher naturalness as singing voice. The results show that the spectral feature is the most affecting acoustic feature to the emotion of singing voice. However, in order to obtain high naturalness and singing-ness for the synthesized voices, it is necessary to manipulate all three features that are F0 contour, amplitude envelope and spectral sequences.

1. Introduction

Singing voice analysis and synthesis are becoming an interesting topic recently. In this topic, expressiveness plays an important role in obtaining high quality of singing voices. Controlling of expression conducts a set of acoustic features that are related to emotion, style or singer individuality. Listeners, depending on their moods and situations, would like to hear a song with different emotional expressions. Therefore, it is necessary to have a computer-based application in singing voice performance for fulfilling such purpose.
Alonso [1] has tried to generate emotional singing voices using rule-based approach. The advantage in this approach is that they are relatively straightforward and completely deterministic. However, the rule-based method takes a long time for the analysis and synthesis phases to get the rules, and the system is yet difficult to control. Another singing voice synthesizer by Umbert has applied unit-selection methodology [2]. It constructs implicit rules based on units of a singing voices. Hence, it requires a huge database of units to process. Hidden Markov Model approaches are also taken into account for synthesizing expressive singing voices by using a statistical methodology to model the important features from database. This method is faced with the problem of overfitting of parameters. A novel speech-to-singing system proposed by Saitou et al. [3], which can produce a singing voices from simple resources: (i) a speaking voice reading a song's lyric, and (ii) its musical score. It succeeded in synthesizing a neutral singing voice. Nonetheless, the expressiveness was not taken into consideration.
The aim of this study is to investigate contributions of significant acoustic features to emotional singing voices. To do this, there are two sub-goals needed to be achieved: (i) analyzing the acoustic features that are strongly related to emotion, (ii) conducting the experimental examination to discuss importance of each acoustic feature in the emotional singing voice.

2. Acoustic feature analysis

In this section, we analyze the acoustic features in emotional singing voice.

2.1 Corpus

The Ryer Audio - Visual Database of Emotion Speech and Song (RAVDESS - 2015) is used for F0 and amplitude envelope analysis. This database is created by Livingstone and Russo in Department of Psychology, McMaster University, Canada and has been validated in a perceptual experiment involving 297 participants.
Regarding spectral features, since there exists the similarity between speaking voice and singing voice in spectral tendency and we want to remove the linguistic information in synthesizing voice, the emotional speech vowel database that was uttered by ten actors under different speaking styles including 8 emotions (Neutral, Afraid, Anger, Disgust, Joy, Relax, Sad, Surprise) is used.

2.2 Feature Selection

This study aims to investigate the correlates of acoustics feature to the emotional singing voice. Hence, a selection of distinctive features of emotional singing voice was considered. Specifically, we consider the analysis of the basic acoustic features of F0, amplitude envelope and spectral feature, which represent the differences between an emotional and neutral singing voice.
F0: F0 features generally have an important effect on the emotional expression of speech. Also, F0 fluctuations have the most influence on the singing-ness of synthesized voice

in Saitou's model [3]. This implies that F0 contours possibly contribute significantly to the emotion of a singing voice.

Amplitude envelope: This is a prosodic feature that contributes to the emotional expression [4]. Oncley [5] found that a singer-formant amplitude of a singing voice is modulated in the synchronization with the frequency modulation of each vibrato in the F0 contour. However, we realize that there exists a high correlation between the entire F0 contour and the amplitude envelope. We therefore propose a rule to re-synthesize such the amplitude envelope, for the purpose of synchronizing with the F0 contour, not only in the vibrato part but also in the overshoot and preparation ones

Spectral sequence: A spectral sequence contains two parameters, including spectral tilts and spectral balance, also bring substantial information of emotional expressions [6]. The preliminary results show that the re-synthesized voices generated using a typical spectral sequence of each emotion, have presented different impressions to listeners.

2.3 Feature Extraction and Modification

The acoustic features should be manipulated to have emotional singing voice.

F0: F0 contour is extracted by using STRAIGHT [7]. After that, crucial properties of four components of F0 contour that well reflect the singingness including overshoot, vibrato, preparation, fine fluctuation are thoroughly studied among different emotions. In [3], Saitou et al. had proposed a model to synthesize proper F0 contours using for speech-to-singing synthesis system shown in the Figure. 1. As in this figure, the transfer functions of the second-order system are used to represent the F0 fluctuations, the responses of these functions with step function can be obtained as:

$$h(t) = \begin{cases} \frac{k}{(2\sqrt{\zeta^2-1})}(exp(\lambda_1\omega t) - exp(\lambda_2\omega t))), & |\zeta| > 1 \\ \frac{k}{\sqrt{1-\zeta^2}}exp(-\zeta\omega t)sin(\sqrt{1-\zeta^2}\omega t), & 0 < |\zeta| < 1 \\ ktexp(-\omega t), & |\zeta| = 1 \\ \frac{k}{\omega}sin(\omega t), & |\zeta| = 0 \end{cases} \quad (1)$$

where $\lambda_1 = -\zeta + \sqrt{\zeta^2 - 1}$, $\lambda_2 = -\zeta - \sqrt{\zeta^2 - 1}$, ω is the natural frequency and ζ is the damping coefficient and k is the proportional gain of the system. Each fluctuation represents as follows:

1. Overshoot: the second-order damping model ($0 < |\zeta| < 1$)
2. Vibrato: the second-order oscillation model ($|\zeta| = 0$)
3. Preparation: the second-order damping model ($0 < |\zeta| < 1$)

After using the model in Figure. 1 to generate F0 contour, the nonlinear least-square-error method [8] will be used to minimize the errors between the generated F0 and actual ones. Hence, we have the set of parameters value ofω, ζ and k for different emotion. The extracted parameter values show that there exists the variation between the values of parameter among emotions, but the variation is not so high due to F0 contour in singing voice have to follow the melody of the song.

From the set of parameter values, we constructed F0 contour using the model 1.

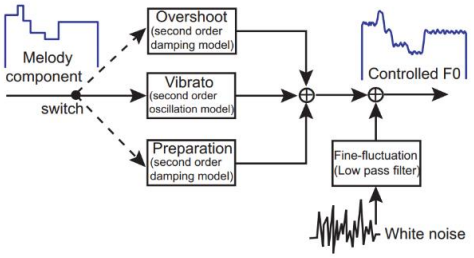

Figure 1: Block diagram of the F0 control model for singing voices [2]

Amplitude envelope: We modify the amplitude envelope by using the F0 fluctuation as Equation 2

$$ampEnv = [\mathrm{k} \cdot F_0] \quad (2)$$

With each specified part, we use the different coefficients. The modifications of overshoot and preparation follow Equation 3.

$$ampEnv = interpolate[\mathrm{k} \cdot peakPicking(F_0)] \quad (3)$$

Regarding vibrato part, we use Saitou's results as Equation 4

$$ampEnv = (1 + k_{am}sin(2\pi f_{am}t))F_0(t) \quad (4)$$

where f_{am} is the rate of amplitude modulation (AM) and k_{am} is the extend of AM. All the set of k, k_{am} and f_{am} are different among emotions.

Spectral sequence: The spectral sequences used for synthesizing singing voices are extracted from the emotional vowel speaking voices using STRAIGHT. After obtaining the spectral sequences representing the emotions, we carefully lengthen it to the desired duration. To preserve the fine fluctuation in the spectral sequences and to keep naturalness of the synthesized voices, we repeat each frame of the vowel sounds with the same number of repeating times.

3. Experimental Examination

3.1 Listening test

The aim of this test is to investigate importance of each acoustic feature in the emotional singing voice. It is also to confirm whether expressiveness of singing voice can be perceived or not with vowel /a/.

Human emotional states can be represented categorically such as happy, angry, sad or as a point in n-dimensional space, such as Valence-Activation space [9]. Emotion categories can be represented as regions in the V-A space, where the neutral state locates in the center, and the other emotion locates in a specific region as shown in Figure. 2. Using the dimensional approach not only gives us the result about category but also the result about the degree of the emotion. Therefore, to investigate the effect of these acoustic features to emotion in singing voice, we calculate the distance and direction of the synthesized voices and neutral one. The synthesized voices include:

- **All:** the three acoustic features are modified.
- **F0 and Spec:** F0 contour and Spectra are modified.
- **F0 and Amp:** F0 contour and Amplitude Envelope are modified.
- **Spec and Amp:** Spectra and Amplitude Envelope are modified.
- **F0:** Only F0 contour is modified.
- **Amp:** Only Amplitude Envelope is modified.
- **Spec:** Only Spectra is modified.

Ten listeners participated in the test. The listeners were required to listen to the synthesized singing voices and evaluate degrees of Valence and Activation of these voices. Each dimension is evaluated using 21 scales (Valence: from Very Negative to Very Positive; Activation: from Very Calm to Very Excited; range from -2 to 2 by 0.2 step).

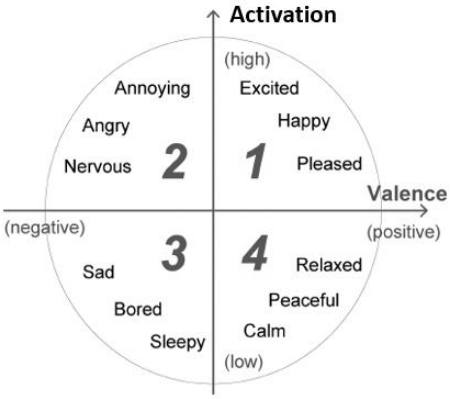

Figure 2: Position of emotions in V-A domain

3.2 Results

Averages of evaluation positions in V-A space of all the listeners are shown in Figure. 3 for the female voice and Figure. 4 for the male voice. As we can see in the figures, the angry and happy stimuli are correctly distributed in its own region in V-A space while sad voices are mostly selected as the neutral voice.

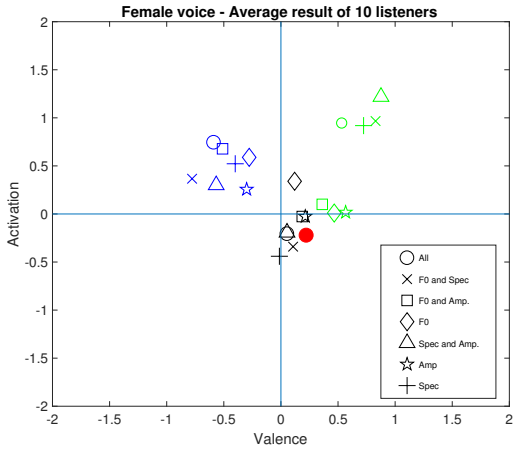

Figure 3: Position in V-A domain of stimuli of female voice. Emotion states: red-neutral; green-happy; blue-angry; black-sad

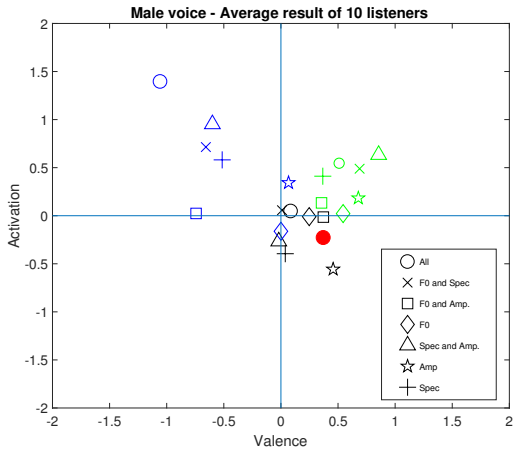

Figure 4: Position in V-A domain of stimuli of male voice. Emotion states: red-neutral; green-happy; blue-angry; black-sad

4. Discussion

From the results of the experimental examination we can see some significant findings as below:

- Listener can distinguish the emotions among stimuli, especially Anger voice and Happy voice.
- The combination having 'spectral' modification give the better result than others.
  This can be stated that the spectral feature is the most important acoustic feature for giving emotion expression in the singing voice.
- The combination having highest degree of emotion is different among emotions (Spectral and Amplitude Envelope for Happy; Spectral for Sad and All of three acoustic feature for Angry).

Table 1: Average distance and direction between the emotional and neutral voice on V-A domain (Female)

	Emotion	F0	Spectral	Amplitude Envelope	F0 and Spectral	F0 and Amplitude Envelope	Spectral and Amplitude Envelop	F0, Spectral and Amplitude Envelope
Distance	Happy	0.34	1.25	0.42	1.33	0.35	**1.58**	1.11
	Sad	0.57	**0.32**	0.2	0.17	0.2	0.17	0.17
	Angry	0.95	0.97	0.7	1.16	1.16	0.95	**1.26**
Direction	Happy	43	66	35	63	67	66	73
	Sad	100	223	93	225	100	170	174
	Angry	122	130	138	150	129	146	130

Table 2: Average distance and direction between the emotional and neutral voice on V-A domain (Male)

	Emotion	F0	Spectral	Amplitude Envelope	F0 and Spectral	F0 and Amplitude Envelope	Spectral and Amplitude Envelop	F0, Spectral and Amplitude Envelope
Distance	Happy	0.3	0.64	0.51	0.78	0.36	**1**	0.79
	Sad	0.25	0.37	0.34	**0.46**	0.21	0.4	0.4
	Angry	0.38	1.2	0.65	1.4	1.14	1.53	**2.2**
Direction	Happy	55	90	53	66	93	61	80
	Sad	119	206	285	141	90	186	136
	Angry	169	137	118	137	167	129	131

- The results of Sad voice is still not obtained the good result. The Sad voices are almost evaluated as a neutral voice.

5. Conclusions

This research successes in extracting the three appropriate acoustic features and manipulating them in order to obtaining the singing voices with emotions. The synthesized singing voices, even without the linguistic information, still express the emotions and the listeners can distinguish them adequately. In addition, by carrying out the subjective test and analyzing the results, the spectral feature is determined as the most affecting acoustic feature to the emotion of singing voices. However, the analysis results also show that it is needed to modify all the three acoustic features to obtain high naturalness and singing-ness.

Acknowledgments

This study was supported by a Grant-in-Aid for Scientific Research (A) (No. 25240026)

References

[1] M. Alonso. Expressive performance model for a singing voice synthesizer. Master's thesis, Universitat Pompeu Fabra, 2005.

[2] M. Umbert, J. Bonada, and M. Blaauw, "Generating singing voice expression contours based on unit selection," Proc. Stockholm Music Acoustics Conference (SMAC), 315-320, 2013.

[3] T. Saitou, M. Unoki and M. Akagi, "Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis," Speech Communication 46, 405-417, 2005.

[4] Y. Xue, Y. Hamada and M. Akagi., "Emotional speech synthesis system based on a three-layered model using a dimensional approach", Proc. APSIPA, pp. 505-514, 2015.

[5] P. B. Oncley, "Frequency, Amplitude, and Waveform Modulation in the Vocal Vibrato," JASA., vol. 49, iss. 1A, pp. 136, 1971.

[6] R. Elbarougy and M. Akagi, "Speech Emotion Recognition System Based on a Dimensional Approach Using a Three-Layered Model," Proc. Int. Conf. APSIPA ASC, 2012

[7] H. Kawahara, I. Masuda-Katsuke and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time frequencey smoothing and an instantaneous-frequency-based F0 extractionL Possible role of a repetitive structure on sounds," Speech Communication 27, 187-207, 1999.

[8] W. H. Press, B. P. Flannery, S. Teukolsky, W. T. Vetterling. "Numerical Recipes in C". Cambridge University Press, Cambridge.

[9] J. A. Russell. "A circumplex model of affect". Journal of Personality and Social Psychology, 39(6), pp. 1161-1178, 1980