PARAMETRIC EMOTIONAL SINGING VOICE SYNTHESIS

Younsung Park, Sungrack Yun and Chang D. Yoo

Korea Advanced Institute of Science and Technology, Department of Electrical Engineering
2106, LG Semicon Hall, 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Republic of Korea
email: ys.park@kaist.ac.kr, yunsungrack@kaist.ac.kr, cdyoo@ee.kaist.ac.kr

ABSTRACT

This paper describes an algorithm to control the expressed emotion of a synthesized song. Based on the database of various melodies sung neutrally with restricted set of words, hidden semi-Markov models (HSMMs) of notes ranging from E3 to G5 are constructed for synthesizing singing voice. Three steps are taken in the synthesis: (1) Pitch and duration are determined according to the notes indicated by the musical score; (2) Features are sampled from appropriate HSMMs with the duration set to the maximum probability; (3) Singing voice is synthesized by the mel-log spectrum approximation (MLSA) filter using the sampled features as parameters of the filter. Emotion of a synthesized song is controlled by varying the duration and the vibrato parameters according to the Thayer's mood model. Perception test is performed to evaluate the synthesized song. The results show that the algorithm can control the expressed emotion of a singing voice given a neutral singing voice database.

Index Terms— Statistical singing voice synthesis, Emotion expression, Vibrato model

1. INTRODUCTION

There are considerable interest in human-machine interface (HMI) and affective computing [1] to enhance the capability of machines to express emotion. Until now, all efforts have been focused on expressing emotion by varying facial expression, action [2] and synthesized speech [3] of a robot, and although it is known that singing can be an effective way of expressing emotion [4], controlling the expressed emotion in singing has not yet been explored.

It is known that duration, pitch and *vibrato* of a music are important features for expressing emotion while singing [5], and this paper investigates the possibility of varying the expressed emotion while singing by varying the duration and *vibrato*. The *vibrato* represents the fluctuation of pitch frequency with time and embodies the characteristic of the singing voice and expressed emotion [5], [6].

There are two approaches to synthesizing voice: (1) sample-based voice synthesis and (2) statistical voice synthesis. The sample-based synthesis approach generates voice by concatenating various samples of the database. This approach requires a large database for natural sounding voice synthesis. The statistical voice-synthesis approach generates a statistical model such as the hidden semi-Markov model (HSMM) from which model parameter are used to synthesize sampled speech. Compared to the sample-based voice synthesis, the statistical voice-synthesis approach has an advantage that a voice characteristic is easily changed by modifying the model parameters [7].

In this paper, the statistical voice synthesis approach is taken, and the statistical characteristics of a note feature consisting of spectral parameters and vibrato parameters are modelled using a HSMM. The spectral parameters consist of mel-cepstral coefficient vectors and their dynamics, and the *vibrato* parameters consist of intonation, *vibrato* extent and *vibrato* rate. Singing voice is synthesized by the mel-log spectrum approximation (MLSA) filter [8] using the sampled feature as parameters of the filter.

To control the expressed emotion of synthesized song, the Thayer's model of mood is adopted [9]. The Thayer's model defines various emotions using only two characteristic of singing voice: energy and tense. In accordance, this paper investigates the expressed emotion in singing by varying two parameters that vary the duration, intonation, *vibrato* extent and *vibrato* rate. The two parameters influence the emotion as the two characteristic of the Thayer's model.

This paper is organized as follows. Section 2 provides background information. Section 3.1 describes the *vibrato* model. Section 3.2 introduces the Thayer's two-dimensional model of emotion. Section 3.3 describes an emotional parameter generation algorithm. Section 4 presents the results of the experiment. Section 5 concludes and provides future research directions.

2. THE OVERVIEW OF HSMM-BASED EMOTIONAL SINGING VOICE SYNTHESIS SYSTEM

The parameters of HSMM-based emotional singing voice synthesis system must be estimated before it can be used to synthesized song of various emotion. Fig. 1 describes the procedure for training and synthesizing singing voice through the learned parameters of HSMM.

In the training part, features consisting of *vibrato* and spectral parameters and duration parameters are extracted from the training data, and the HSMM is used to model the statistical properties of the features [10]. The *vibrato* parameters are obtained from F0 values of training data, and they are composed of three elements: intonation, *vibrato* extent and *vibrato* rate. Intonation is the averaged pitch height, *vibrato* extent is the difference between peak of F0 value and intonation value, and *vibrato* rate is the changing rate of F0 value. Using these features with their labels (notes), HSMM of each note is trained. The duration in the HSMM is modelled by a Gaussian.

In the synthesis part, emotional singing voice is synthesized from a musical score using trained HSMMs. From the pitch sequence of a musical score, a sequence of HSMM to model the musical score is determined. The total duration of the musical score

This research (paper) was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Knowledge Economy of Korea, and was in part supported (National Robotics Research Center for Robot Intelligence Technology, KAIST) by Ministry of Knowledge Economy under Human Resources Development Program for Convergence Robot Specialists.

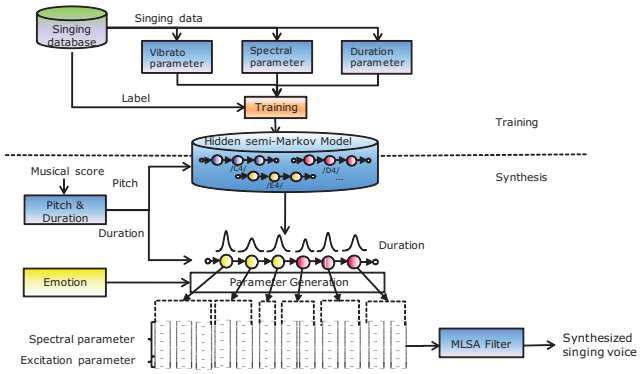

Fig. 1. Overview of HSMM-based emotional singing voice synthesis system.

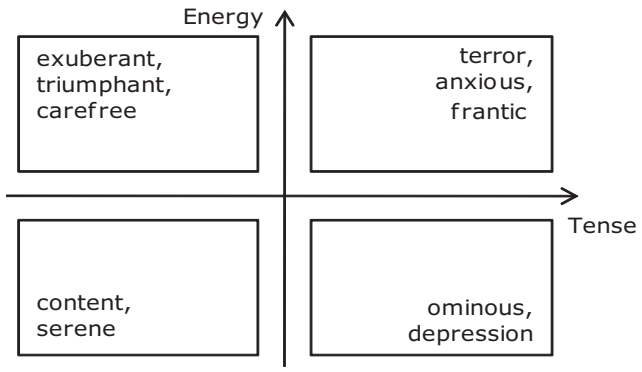

Fig. 2. Thayer's two dimensional model of emotion. Here, energy and tense are respectively defined as subjective sensation of energy, vigor, or peppiness and feeling of tension, anxiety, or fearfulness.

determines the duration of each HSMM in the sequence. The duration of each state of a HSMM is determined by a probabilistic duration model, and depending on the emotion, it is modified. According to the duration, each state generates parameter vectors consisting of mel-cepstrum and vibrato parameters using the parameter generation algorithm [10]. After changing the vibrato parameters by the algorithm in the next section, we synthesize a singing voice using the generated parameters and the MLSA filter [8].

3. HSMM-BASED EMOTIONAL SINGING VOICE SYNTHESIS SYSTEM

3.1. Vibrato model

There are many modes of vocalization in singing voice which includes *vibrato*, *marcato*, *accelerando*, *rubato* and *ritardando*, and these represent the different styles in singing. It has been proposed [5] that of the many modes, varying the degree of *vibrato* can lead to different expressed emotion.

The *vibrato* is the fluctuation of F0 values over time in singing. It is represented by three parameters: intonation, *vibrato* extent and *vibrato* rate. Pitch, otherwise known as F0 value, can be mathematically expressed as

$$\mathbf{f}[n] = \mathbf{m}[n] + \mathbf{e}[n]\cos(\theta[n]) \tag{1}$$

where $\mathbf{m}[n]$, $\mathbf{e}[n]$ and $\theta[n]$ represent the intonation of $\mathbf{f}[n]$ at time n, the *vibrato* extent and the periodic angle value which is related to the *vibrato* rate, respectively. The intonation $\mathbf{m}[n]$ is a running average of F0 values and is given by

$$\mathbf{m}[n] = \frac{\sum_{\tau=-L_1}^{L_2} \mathbf{f}[n+\tau]}{L_1 + L_2 + 1} \tag{2}$$

where $n - L_1$ and $n + L_2$ represent start point and end point of the moving average filter at time n, respectively. Let us define

$$\mathbf{d}[n] = \mathbf{f}[n] - \mathbf{m}[n] = \mathbf{e}[n]\cos(\theta[n]). \tag{3}$$

then the *vibrato* extent and *vibrato* rate parameters can be derived as

$$\mathbf{e}[n] = \sqrt{\mathbf{d}^2[n] + \hat{\mathbf{d}}^2[n]}, \tag{4}$$

$$\mathbf{r}[n] = \theta[n] - \theta[n-1] \tag{5}$$

where

$$\hat{\mathbf{d}}[n] = H[\mathbf{d}[n]] = \mathbf{e}[n]\sin(\theta[n]), \tag{6}$$

and

$$\theta[n] = \arctan(\mathbf{d}[n], \hat{\mathbf{d}}[n]). \tag{7}$$

Here, $H[\mathbf{d}[n]]$ denotes the Hilbert transform of $\mathbf{d}[n]$. The Hilbert transform causes the phase shift with the $\pi/2$ or $-\pi/2$. Thus, we have $\hat{\mathbf{d}}[n] = \mathbf{e}[n]sin(\theta[n])$ and $\mathbf{d}[n] + j\hat{\mathbf{d}}[n] = \mathbf{e}[n]e^{j\theta[n]}$. From the $\mathbf{d}[n]$ and $\hat{\mathbf{d}}[n]$, *vibrato* extent $\mathbf{e}[n]$ and periodic angle $\theta[n]$ and *vibrato* rate $\mathbf{r}[n]$ can be obtained.

3.2. Emotion model for expression

In music psychology, mood taxonomy is a key descriptor in describing a music. In a traditional approach, many adjectives are used to describe the mood of music such as happy, sad, pathetic, hopeful, gloomy and so on [11]. In Hevner's classification for mood of music [12], there are approximately 67 adjectives. However, there are no relationship between the adjectives; thus, each adjective has its own numerical value to describe the mood of a music.

Many adjectives are needed to be defined in a systematic fashion for describing emotions. Russell finds the interrelationship between adjectives in a two dimensional model for emotion taxonomy[13]. Later, Thayer adopts the Russell's model to music using two-dimensional energy-tense emotion model [9] as shown in Fig. 2. In the model, the emotion of a music is represented by a point in the energy-tense plane. Here, energy is defined as subjective sensation of energy or vigor, and tense is defined as feeling of tension or anxiety [9]. Song sung exuberantly or in terror has high positive energy, and song sung depressingly or serenely has high negative energy. The tense represents the degree of pleasure: song sung exuberantly or in content has high negative tense value, and song sung in terror or depressingly has high positive tense value.

3.3. Emotional parameter generation algorithm

For synthesizing singing voice of various emotion using the MLSA filter, appropriate parameters of the filter must be generated. The parameters consist of duration, *vibrato* and mel-cepstrum. The mel-cepstrum parameter is generated using the parameter generation al-

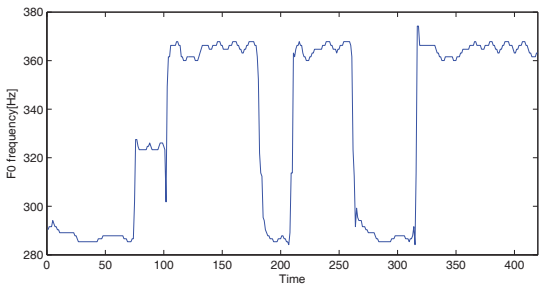

Fig. 3. Example of F0 contour of a synthesized singing.

gorithm in [10]. The duration ratio of each state in a HSMM is determined by the algorithm in [14]:

$$\mathbf{D}^* = \arg\max_{\mathbf{D}} \log P(\mathbf{D}|\lambda) = [\xi_1, ..., \xi_K] \tag{8}$$

where $\mathbf{D}$, λ, ξ_n and K are, respectively, the vector of duration of each state, HSMM parameters, the duration of n-th state and the number of states. Since the entire duration of a HSMM is fixed by the musical score to T, the state duration is scaled as

$$\mathbf{D}^{'} = \mathbf{D}^* \cdot T/t = [\xi^{'}{}_1, ...\xi^{'}{}_K] \tag{9}$$

where

$$t = \sum_{n=1}^{K} \xi_n, \tag{10}$$

$\mathbf{D}^{'}$ is the scaled duration vector and $\xi^{'}$ is the scaled duration of n-th state, respectively.

The *vibrato* parameters from HSMM can be obtained as

$$\begin{aligned} \mathbf{O}^* &= \arg\max_{O} P(\mathbf{O}|q, \lambda, \hat{\mathbf{D}}), \\ &= \{[\mu_{m_1}\ \mu_{e_1}\ \mu_{r_1}]^T, ..., [\mu_{m_T}\ \mu_{e_T}\ \mu_{r_T}]^T\} \end{aligned} \tag{11}$$

where $\mathbf{O}$, q and $\hat{\mathbf{D}}$ represent observation vector sequence, state sequence and modified duration following the emotion as will be described by equation (12). The mean vector of the intonation, *vibrato* extent and *vibrato* rate of the state corresponding to the n-th frame are denoted by μ_{m_n}, μ_{e_n} and μ_{r_n}.

Previous studies have shown that "happy" emotion is expressed by shorter duration and higher intonation, *vibrato* extent, *vibrato* rate while "sad" emotion is expressed by longer duration and lower intonation, *vibrato* extent, *vibrato* rate [5][15]. Thus, "happy" emotion is related to high negative tense and high positive energy, and "sad" emotion is related to high positive tense and high negative energy. In accordance, the relationship between features to synthesize singing voice and the Thayer's model shows the following when energy is increased and tense is decreased, the duration is shorter while intonation, *vibrato* extent and *vibrato* rate are increased. When energy is decreased and tense is increased, the duration is longer while intonation, *vibrato* extent and *vibrato* rate are decreased. In the proposed algorithm, an emotion in a singing voice is expressed by modifying the parameters $\mathbf{D}^{'}$, μ_{m_n}, μ_{e_n} and μ_{r_n} as follows

$$\hat{\mathbf{D}} = \mathbf{D}^{'}(1 + (1+\alpha)\beta k_d) = [\hat{\xi}_1, ...\hat{\xi}_K], \tag{12}$$

$$\hat{\mathbf{m}}[n] = \mu_{m_n}(1 - (1+\alpha)\beta k_m), \tag{13}$$

$$\hat{\mathbf{e}}[n] = \mu_{e_n}(1 - (1+\alpha)\beta k_e), \tag{14}$$

$$\hat{\mathbf{r}}[n] = \mu_{r_n}(1 - (1+\alpha)\beta k_r) \tag{15}$$

Table 1. Experimental result for happy emotion: with various combinations of duration (D), intonation (I), *vibrato* extent (E) and *vibrato* rate (R).

Category	Average rank
D	1.85
D+E+R	2.02
D+I+E+R	2.14
E+R	4.37
I+E+R	4.57

Table 2. Experimental result for sad emotion: with various combinations of duration (D), intonation (I), *vibrato* extent (E) and *vibrato* rate (R).

Category	Average rank
D	2.45
D+E+R	2.34
D+I+E+R	1.82
E+R	4.45
I+E+R	3.97

where k_d, k_m, k_e and k_r are the scaling factor for duration, intonation parameter, *vibrato* extent parameter and *vibrato* rate parameter, respectively. The emotion parameters α, β can be respectively considered as energy parameter ranging from -1 (less energetic) to 1 (most energetic) and tense parameter ranging from -1 (most happy) to 1 (most sad). We restrict the dynamic range of the $\hat{\mathbf{D}}$, $\hat{\mathbf{m}}$, $\hat{\mathbf{e}}$, and $\hat{\mathbf{r}}$ by

$$0 < \hat{\mathbf{D}} < 2\mathbf{D}^{'}, \tag{16}$$

$$0.5\mu_{m_n} \leq \hat{\mathbf{m}}[n] \leq 1.5\mu_{m_n}, \tag{17}$$

$$0.5\mu_{e_n} \leq \hat{\mathbf{e}}[n] \leq 1.5\mu_{e_n}, \tag{18}$$

$$0.5\mu_{r_n} \leq \hat{\mathbf{r}}[n] \leq 1.5\mu_{r_n}. \tag{19}$$

Thus, k_d, k_m, k_e, and k_r should be set in the range:

$$0 < k_d < \frac{1}{(1+\alpha)|\beta|}, \tag{20}$$

$$\frac{0.5}{(1+\alpha)|\beta|} \leq k_m \leq \frac{1.5}{(1+\alpha)|\beta|}, \tag{21}$$

$$\frac{0.5}{(1+\alpha)|\beta|} \leq k_e \leq \frac{1.5}{(1+\alpha)|\beta|}, \tag{22}$$

$$\frac{0.5}{(1+\alpha)|\beta|} \leq k_r \leq \frac{1.5}{(1+\alpha)|\beta|}. \tag{23}$$

Using the *vibrato* parameters, the F0 value used in the synthesizer is generated following the equation (1) where

$$\hat{\theta}[n] = \sum_{m=1}^{n} \hat{\mathbf{r}}[m]. \tag{24}$$

4. EXPERIMENT

In contrast to the availability of many speech database, there is no known database for neutral singing. A database of 56-minutes recording of singing by a professional female singer is collected. Singing sung using restricted word 'Ra' for each note ranging from E3 to G5 was recorded without any emotion at 44.1kHz sampling with 16 bit/sample.

A HSMM of 5 states with left-to-right model with no skip is used. Feature vectors consist of 25 mel cepstral coefficients, corresponding delta and acceleration coefficients [10], and three *vibrato* parameters. The singing data was labelled manually by segmenting according to the F0 values.

The singing voice is synthesized using the HMM-based speech synthesis system (HTS) which is used for speech synthesis [14]. The F0 value obtained from the synthesized neutral singing voice using the *vibrato* effect is shown in Fig. 3.

We evaluated emotions of the synthesized song by conducting a subjective listening test. We synthesized 5 well-known Korean children's songs with various emotions: happy ($\beta = -0.5$) and sad ($\beta = 0.5$) and $\alpha = 0.5$ and scaling factor $k_d = 0.5$, $k_m = 1.0$, $k_e = 2.0$, $k_r = 0.5$. Results indicate that *vibrato* extent and *vibrato* rate have little influence of perceived emotion. To investigate the effect of the parameters on synthesized singing voice, we combined various parameters into the following categories:

1. duration (D)
2. duration, *vibrato* extent, *vibrato* rate (D+E+R)
3. duration, intonation, *vibrato* extent, *vibrato* rate (D+I+E+R)
4. intonation, *vibrato* extent, *vibrato* rate (I+E+R)
5. *vibrato* extent, *vibrato* rate (E+R).

By changing the parameters defined in each category, we synthesized five different emotional singing voices for expressing sadness and happiness. Perceptual evaluation is conducted to rank from 1 (most acceptable) to 5 (least acceptable) for each category. Subjects were 7 persons, and each person listened the neutral singing voice first and ranked the emotional singing voice.

The average rank for each emotion is shown in Table 1 and 2. In these results, duration change is the more effective to represent an emotion in a singing voice than *vibrato* parameters since duration can be most perceptible characteristic in a song among the emotion parameters. The effect of emotion parameters are different for each emotion. Varying the *vibrato* parameters did not have noticeable effect on producing happy and sad; however, varying the *vibrato* parameters with duration change had a noticeable effect on sad.

5. CONCLUSION

In this paper, HSMM-based emotional singing voice synthesis system was introduced. The database which is recorded neutrally with restricted word 'Ra' for each note ranging from E3 to G5 is used. For each note, HSMM is constructed for synthesizing singing voice. The procedure to synthesize singing voice is as follows: (1) Given the musical score, the sequence of pitch and duration is determined; (2) Duration, spectral parameters and vibrato parameters are sampled from appropriate HSMMs; (3) Singing voice is synthesized by MLSA filter using the sampled features. In the synthesis procedure, The duration and vibrato parameters are controlled to express emotion in a synthesized song following the Thayer's mood model. To evaluate the algorithm, perception test is performed, and the results show the rank of the combination of parameters for each emotion.

6. REFERENCES

[1] R. W. Picard, *Affective Computing*, MIT Press, 1997.

[2] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 119–155, 2003.

[3] M. Schröder, "Emotional speech synthesis: A review," in *Proc. ISCA Int. Seventh European Conference on Speech Communication and Technology*. ISCA, 2001.

[4] C. Bartneck, "How convincing is mr. data's smile: Affective expressions of machines," in *Proc. of IEEE Int. Conf. Acoust., Speech, Signal Process*. User Modeling and User-Adapted Interaction, 2001, vol. 11, pp. 279–295.

[5] CE Seashore, "Measurements on the Expression of Emotion in Music," *In Proc. of the National Academy of Sciences*, vol. 9, no. 9, pp. 323–325, 1923.

[6] D. Myers and J. Michel, "Vibrato and pitch transitions," *J. Voice*, vol. 1, no. 2, pp. 157–161, 1987.

[7] AW Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Proc. IEEE Int. International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007*, 2007, vol. 4.

[8] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, 1983.

[9] R.E. Thayer, *The biopsychology of mood and arousal*, Oxford University Press, USA, 1989.

[10] N. Miyazaki K. Tokuda, T. Masuko and T. Kobayashi, "Speech parameter generation algorithms for hmm-based speech synthesis," in *Proc. of IEEE Int. Conf. Acoust., Speech, Signal Process*, 2000, pp. 1315–1318.

[11] L. Lu, D. Liu, and H.J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 1, pp. 5–18, 2006.

[12] K. Hevner, "Expression in music: A discussion of experimental studies and theories," *Psychological Review*, vol. 42, no. 2, pp. 186–204, 1935.

[13] J.A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[14] H. Zen, K. Tokuda, T. Masuko, T. Kobayasih, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Transactions on Information and Systems*, vol. 90, no. 5, pp. 825, 2007.

[15] Y. Feng, Y. Zhuang, and Y. Pan, "Music information retrieval by detecting mood via computational media aesthetics," in *Proc. IEEE/WIC International Conference on Web Intelligence*, 2003.