Voice and Emotional Expression Transformation based on Statistics of Vowel Parameters in an Emotional Speech Database

Toru Takahash, Takeshi Fujii, Masashi Nishi, Hideki Banno, Toshio Irino, Hideki Kawahara

Faculty of Systems Engineering
Wakayama University, Wakayama 640-8510, Japan
{tall, s045043, s035063, banno, irino, kawahara}@sys.wakayama-u.ac.jp

Abstract

We propose a simple method for modifying emotional speech sounds. The method aims at real-time implementation of an emotional expression transformation system based on STRAIGHT. We developed a mapping function of spectra, fundamental frequencies \( (F_0) \), and vowel durations from the statistical analysis of 1500 expressive speech sounds in an emotional speech database. The spectral mapping parameters are initially extracted at the centers of vowels and interpolated with bilinear functions. The spectral frequency warping functions are manually designed. The \( F_0 \) and duration mapping functions simply transform the average values in log frequency and linear time scales. We demonstrate that the spectral distortion is small enough when ‘Neutral’ speech sounds are transformed to expressive speech sounds (i.e. ‘Bright’, ‘Excited’, ‘Angry’, and ‘Raging’ speech sounds).

1. Introduction

There are several review papers on emotional speech [1, 2]. Recent research for voice transformation has focused on transforming ‘Neutral’ speech sounds to expressive speech sounds [3, 4, 5].

STRAIGHT[6] is known as a very high-quality speech analysis, modification, and synthesis procedure. Matsui et al [7] reported that morphing technique between expressive speech sounds based on STRAIGHT kept speech sounds high-quality and natural.

Toda et al [4] proposed an emotional expression transformation algorithm based on statistics and demonstrated that transformed speech sounds with other synthesis system were a natural. The transformation algorithm is based on the joint probability densities of source and target features. It is hard to apply the algorithm in real-time.

A simple and high-quality expressive speech synthesis procedure from only ‘Neutral’ speech must be developed. This paper describes an emotional expression mapping technique based on STRAIGHT. A simple emotional expression transformation rules are introduced. The rules are shown independent of input speech sounds. A key to design simple transformation rules is to focus on the statistics of vowel parameters. Conventional transformation rules are applied to all phonemes and the ‘Neutral’ spectra are changed. In contrast, the proposed transformation rules are applied to all of the vowel parts of the spectra. The vowel spectra are changed, however, not the consonant spectra.

2. Voice and emotional expression transformation system

A block diagram of a voice and emotional expression transformation system is shown in Figure 1. First, a ‘Neutral’ speech sound is input into the system and decomposed into a fundamental frequency \( (F_0) \) and vocal tract characteristics in a STRAIGHT analysis block. Then an average and a variance of \( F_0 \) are adapted to an average and a variance given by analyzing an emotional speech database in an \( F_0 \) mapping block, respectively. An adapted \( F_0 \) is warped to a target speech space in a time warping block. Some features given by analyzing an emotional speech database are added to the spectra in a spectral mapping block. Adapted spectra are warped to target speech spectral space in a time-frequency warping block. Finally, a target speech sound is comprised of \( F_0 \) and the spectra in a STRAIGHT synthesis block.

2.1. Emotional speech database

An emotional speech database is built and used in experiments. Normal and expressive speeches narrated by eight professional actors are sampled at 44.1 kHz with 16 bit linear PCM format. Normal speech is called ‘Neutral speech’ in this paper, and four types of expression called ‘Bright’, ‘Excited’, ‘Angry’, and
'Raging', are recorded.

2.2. $F_0$ transformation

A fundamental frequency transformation is based on statistics. The relation between a fundamental frequency trajectory, $F_0(t)$, $(t = 1, 2, \ldots, T)$ and a transformed fundamental frequency trajectory, $\hat{F}_0(t)$, is represented as

$$\log(F_0(t)) = \frac{\log(F_0(t)) - \mu_{m}}{\sigma_{m}}, \sigma_{m} + \mu_{m}, \quad (1)$$

where constant values $\mu_{m}, \mu_{m}, \sigma_{m}, \text{and} \sigma_{m}$ are derived from an emotional speech database. These parameters are defined as follows:

$$\mu_{m} = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} 1 \cdot N \cdot T_n}{1 \cdot N \cdot T_n} \log(F_0(n)), \quad (2)$$

$$\mu_{m} = \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} 1 \cdot T_n}{1 \cdot T_n} \log(F_0(n)), \quad (3)$$

$$\sigma_{m}^{2} = \frac{\sum_{t=1}^{T_n} \sum_{n=1}^{N} 1 \cdot N \cdot T_n (\log(F_0(t)) - \mu_{m})^2}{1 \cdot N \cdot T_n}, \quad (4)$$

$$\sigma_{m}^{2} = \frac{\sum_{t=1}^{T_n} \sum_{n=1}^{N} \frac{1}{M \cdot T_m} (\log(F_0(t)) - \mu_{m})^2}{1 \cdot M \cdot T_m}, \quad (5)$$

where $N$ and $M$ are the number of ‘Neutral’ and target expression utterances in an emotional speech database, respectively. Fundamental frequency trajectories corresponding to the $n$-th neutral utterance and the $m$-th target expression utterances are represented by $F_0(n)$ and $F_0(m)$, respectively.

The boxplot of fundamental frequency in a male is shown in Figure 2. The $F_0$ distribution is shown by boxes and lines. Lines over and below a box show frequencies at which cumulative distribution achieved 90% and 10%, respectively. Upper and bottom lines of a box show the frequencies that cumulative distribution achieved 75% and 25%, respectively. The center line of a box shows 50%. Vertical and horizontal axes give a fundamental frequency in Hz and types of expression, respectively.

These results show that the average $F_0$ of ‘Bright’, ‘Excited’, ‘Angry’, and ‘Raging’ are higher than the average $F_0$ of ‘Neutral’. Variance of expressive $F_0$ are wider than the variance of ‘Neutral’ $F_0$ except ‘Excited’. The variance of ‘Angry’ $F_0$ is narrower than the variance of ‘Neutral’.

The equation (1) shows a rule for adapting the average and variance of ‘Neutral’ $F_0$ to those of the target $F_0$. For each expression, averages and variances of $F_0$ are represented by $\mu_{m}$ and $\sigma_{m}$, respectively, where $m = 1, 2, 3, 4$. The $m$ corresponds to each expression (‘Bright’, ‘Excited’, ‘Angry’, ‘Raging’).

The parameters of the rules are simple and independent of input speech sounds. These features of the proposed procedure are suitable for real-time emotional expression transformation.

2.3. Spectrum transformation

Mapping feature $\delta(t) = \delta_1(t), \delta_2(t) \text{and} \delta_3(t)$ are investigated and defined as follows:

$$\delta_1(t) = \begin{cases} w_m(S_m(t)) - S_n(t), & \text{t is in the vowel segment}, \\ \{0, \ldots, 0\}, & \text{otherwise} \end{cases} \quad (7)$$

$$\delta_2(t) = \begin{cases} w_m(S_m^{\xi}(t)) - S_n^{\xi}(t), & \text{t is in the \xi segment}, \\ \{0, \ldots, 0\}, & \text{otherwise} \end{cases} \quad (8)$$

$$\delta_3(t) = \begin{cases} w_m(S_m^{\xi}(t)) - S_n^{\xi}(t), & \text{t is in the center position in the \xi segment}, \\ \{0, \ldots, 0\}, & \text{otherwise} \end{cases} \quad (9)$$

$$S_m(t) = \frac{1}{K_m^{\xi}} \sum_{\xi \in \{a,e,i,o,u\}} K_m^{\xi} \sum_{k=1}^{K_m^{\xi}} S_m^{\xi}(k) \quad (10)$$

$$S_n(t) = \frac{1}{K_n^{\xi}} \sum_{\xi \in \{a,e,i,o,u\}} K_n^{\xi} \sum_{k=1}^{K_n^{\xi}} S_n^{\xi}(k) \quad (11)$$

$$\delta_3(t) = \begin{cases} \alpha \delta_1(t_p) + (1 - \alpha) \delta_3(t_n), & \text{t is in the \xi segment except center position}, \\ \{0, \ldots, 0\}, & \text{otherwise} \end{cases} \quad (12)$$

$$\delta_3(t) = \begin{cases} \alpha \delta_1(t_p) + (1 - \alpha) \delta_3(t_n), & \text{t is in the \xi segment except center position}, \\ \{0, \ldots, 0\}, & \text{otherwise} \end{cases} \quad (13)$$

where the nearest vowel center positions at time $t$ are represented as $t_p$ and $t_n$, $t_p < t < t_n$, $t = 1, 2, \ldots$ and $\alpha = \frac{t - t_p}{t_n - t_p}$. The number of vowel $\xi$ in ‘Neutral’ and target expression utterances in an emotional speech database are represented by $K_m^{\xi}$ and $K_n^{\xi}$, respectively. The average spectrum corresponding to vowel $\xi$ in ‘Neutral’ and target expression utterances in an emotional speech database are $S_m^{\xi}(k)$ and $S_n^{\xi}(k)$, respectively. Spectral warping functions $w_m(\cdot)$ and $w_n(\cdot)$.
2.4. Frequency warping

The spectral frequency warping functions, \( w_m \) and \( w_m^\ell \), are manually designed to match spectral peaks corresponding to formants. Although it is probably hard to design for all functions, it is not hard because the number of functions is 20 (Four types of expression and five vowels are treated in this case). All functions are represented by partly bilinear function \( w(\cdot) \) formed as

\[
\tilde{S}(n) = \begin{cases} 
S(w(n)), & w(n) \text{ is an integer.} \\
\alpha S\left([w(n)]\right) + (1 - \alpha)\tilde{S}\left([w(n)]\right), & \text{otherwise.}
\end{cases}
\]

\[
\alpha = 1 - n + \lfloor n \rfloor, \\
n = 1, 2, \ldots, N,
\]

where \( w(1) = 1 \) and \( w(N) = N \). \([n] \) and \( \lfloor n \rfloor \) round the elements of \( n \) to the nearest integers less or greater than or equal to \( n \), respectively.

2.5. Speech speed

Speech speed is changed at constant speeds for each target expression. Duration distributions of all vowels in source (‘Neutral’) and target (‘Excited’) utterances are shown in Figure 4 as dotted and solid lines, respectively. It shows a distribution peak of ‘Neutral’ around 70 ms and a distribution peak of ‘Excited’ around 90 ms. To adapt ‘Neutral’ duration to ‘Excited’ distribution, ‘Neutral’ average duration \( \bar{d}_n \) is adapted to ‘Excited’ average duration \( \bar{d}_m \). Durations \( d'_n \) adapted from ‘Neutral’ to ‘Excited’ are as follows:

\[
d'_n = d_n \frac{\bar{d}_m}{\bar{d}_n}, \\
\frac{\bar{d}_n}{\bar{d}_m} = \frac{1}{N} \sum_{m=1}^{M} d_m, \\
\frac{\bar{d}_m}{\bar{d}_m} = \frac{1}{M} \sum_{m=1}^{M} d_m,
\]

where \( N \) and \( M \) are the number of vowels in ‘Neutral’ and ‘Excited’ utterances in an emotional speech database, respectively. \( d_n \) and \( d_m \) are the durations of vowels in ‘Neutral’ and ‘Excited’.

2.6. Transformation example

The spectrogram corresponding to input ‘Neutral’ speech sounds is shown in Figure 5, and the results of spectral transformation, spectral warping, and changing duration are shown in Figure 6 for ‘Angry’. Vertical and horizontal axes give time in ms and frequency in kHz, respectively. Spectral peaks are represented by dark color.

Figures 5 and 6 show that the speech speed changes faster than ‘Neutral’ speech speed by adding emotional expression. It is shown that the spectral peaks in Figure 5 at 0.5 kHz in 150 to 200 ms move to 0.4 kHz in 150 to 200 ms in Figure 6 by spectral warping and mapping features.

3. Experiments

The experiments were conducted to evaluate effectiveness of three proposed transformation. The spectra transformed by the proposed method were compared with the non-transformed spectra. Although target expressive speech sounds are not required in the proposed transformation system, virtual target expressive speech sounds are prepared for the evaluation by means of actually making the actor pronounce the desired expressive
speeches. The sounds have the same phonetic labels as those of ‘Neutral’ speech sound for the source. The spectral distortion between transformed speech and virtual target speech corresponding to the part of vowels is shown in Figure 7. Vertical and horizontal axes show spectral distortion and target expression. Spectral distortions between ‘Neutral’ speech and virtual target speech are shown as black bars. Because non-transformed speech is the same as input speech (‘Neutral’ speech), the black bars are considered baseline distortion. Spectral distortion between transformed with δ₁, δ₂, and δ₃ (expression(7),(8), and(7)) are shown as dark gray, light gray, and white bars, respectively. Spectral distortions decrease after transformation with δ₃ for all expression.

4. Conclusions

Voice and emotional expression transformation based on STRAIGHT is described in this paper. A simple transformation is introduced, such as transforming average and variance of F₀ to statistics in an emotional speech database, and spectral transformation rules are made from average spectrum for each vowel and each type of expression in the database. Once an emotional speech database is given, such rules independent of input speech can be provided. Experimental results show that spectral distortions decrease after transformation for all emotional expression. Informal listening tests suggested that the proposed method generally yielded emotional expression modifications toward the desired expression. An incremental algorithm of the proposed method is under development for implementing a STRAIGHT-based real-time speech modification system and will be reported elsewhere.

5. Acknowledgements

This work was partly supported by the Ministry Education, Culture, Sports, Science and Technology e-Society leading project.

6. References