INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Comparison of glottal source parameter values in emotional vowels

Yongwei Li[1], *Jianhua Tao*[1,2,3], *Bin Liu*[1], *Donna Erickson*[4,5], *Masato Akagi*[6]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[3]Center for Excellence in Brain Science and Intelligence, Chinese Academy of Sciences, China
[4]Haskins Laboratories, U.S.A.
[5]Federal University of Rio de Janeiro, Brazil
[6]Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Japan

yongwei.li@nlpr.ia.ac.cn, jhtao@nlpr.ia.ac.cn, liubin@nlpr.ia.ac.cn, ericksondonna2000@gmail.com, akagi@jaist.ac.jp

Abstract

Since glottal source plays an important role for expressing emotions in speech, it is crucial to compare a set of glottal source parameter values to find differences in these expressions of emotions for emotional speech recognition and synthesis. This paper focuses on comparing a set of glottal source parameter values among varieties of emotional vowels /a/ (joy, neutral, anger, and sadness) using an improved ARX-LF model algorithm. The set of glottal source parameters included in the comparison were T_p, T_e, T_a, E_e, and $F_0(1/T_0)$ in the LF model; parameter values were divided into 5 levels according to that of neutral vowel. Results showed that each emotion has its own levels for each set of the glottal source parameter value. These findings could be used for emotional speech recognition and synthesis.

Index Terms: Glottal source, emotional vowel analysis, ARX-LF model

1. Introduction

As the most natural means of human-human communication, speech consists of not only linguistic information but also non-linguistic information, such as emotion, gender, and age. In addition to linguistic information, non-linguistic information is also important for expressive speech synthesis and speech understanding. Especially, emotions in speech are extensively used by humans to express intentions and play an important role in speech communication [1]. Therefore, increased attention has been given to emotions in speech in recent years [2].

Emotional speech analysis provides an important basis for expressive speech recognition, synthesis and many other speech applications. Based on the source-filter theory, most studies on emotional speech analysis tried to find out the source-filter - related acoustic features that are closely related to the emotions in speech. Many studies suggested that F_0, energy, duration, speech rate, and spectral cues are the most important acoustic features for expressing emotions in speech [3, 4, 5]. A few studies showed that formant frequencies and voice quality contribute to emotions [6, 7]. Although the emotional speech synthesis and recognition systems have been built up based on these acoustic features, this research methodology is quite difficult to further improve the performance which is still far from that of a human being. It is well-known that glottal sources and vocal tract shapes play important roles in speech production. Studies have suggested that features of speech-production organs, such as glottal source waves and vocal tract shapes be directly investigated [8, 9].

Based on the source-filter model of speech production[10], several studies processed emotional speech to investigate glottal source waveform using inverse filtering, where glottal sources were considered residual signals. They suggest that glottal source waveform plays the most important role for emotional speech [11, 12]. Hence, further analysis of the glottal source parameters of emotional vowels, comparing the set of glottal source parameter values and finding differences of their values among emotions, is greatly helpful for improving speech emotion recognition and synthesis systems.

Several studies have investigated glottal source parameters for insight into different emotions, most of them by using inverse filtering where glottal sources were considered residual signals. However, it is difficult to handle the glottal source parameters without these parameters [13, 14]. Although some studies have attempted to extract and compare glottal source parameters using inverse filtering with glottal source models [12], it is difficult to find differences for the set of glottal source parameter values given the limitation of glottal source estimation methods.

Li *et al*. proposed a more precise algorithm for estimating glottal source wave and vocal tract shape parameters simultaneously based on an analysis-by-synthesis method that combines the Auto-Regressive eXogenous with the Lilijencrants-Fant (ARX-LF) model. They also investigated the contribution of glottal source and vocal tract cues to emotional vowels through perception experiments [15, 16]. However, glottal source parameters have not been compared extensively among emotions.

This study aims to compare the set of glottal source parameter values among emotions in vowels. A high-quality analysis-by-synthesis method is used which can accurately estimate the glottal source and vocal tract parameters simultaneously based on an improved ARX-LF model [17]. We first estimate the set of glottal source parameter values of the vowels /a/ with four emotion states for four different speakers. Subsequently, these parameter values were divided into 5 levels according to the neutral vowel, and then these levels were compared among different emotions. Sets of different patterns (set of each parameter value levels) were found for each emotion. It is believed that the results in this paper will contribute to improving speech emotion synthesis and recognition systems.

http://dx.doi.org/10.21437/Interspeech.2020-1536

2. Methods

2.1. dataset

The sustained vowel /a/ with four emotional states (anger, joy, neutral, and sadness) uttered by two male and two female professional actors were recorded (44100 Hz sampling frequency) three times for each emotional state at experiment room in the Japan Advanced Institute of Science and Technology (JAIST), Japan. A listening experiment was further carried out to select four emotional states from each speaker. A total of 16 sustained vowels (4 speakers $\times$ 4 emotions) were investigated.

2.2. ARX-LF model

To estimate the glottal source parameter values of emotional vowels, a high-quality analysis tool of the source-filter model is required for separating glottal source and vocal tract. Among source-filter models, the ARX-LF model has been widely used for representing glottal source waves and vocal tract filter [18], and is especially good for estimating glottal source parameters of neutral vowels. Li *et al*. proposed an improved algorithm for analyzing emotional vowels based on the ARX-LF model, and showed higher estimation accuracy than the traditional inverse filter method [16, 17]. Thus, the improved algorithm of the ARX-LF model is used for estimating glottal source parameters of emotional vowels in this paper. For details on implementation of this improved approach, please refer to Li *et al*. [16, 17]. The ARX-LF model is briefly introduced in the following section.

The LF model is mainly characterized by six parameters to represent the derivative of the glottal flow, including five time-domain parameters T_p, T_e, T_a, T_c, T_0 and one amplitude parameter E_e. One period of glottal source waveform and its derivative waveform of the LF model is plotted in Fig. 1. T_0 is one period of glottal source waveform, T_p is the instant of the maximum glottal source waveform, T_e is the instant of the maximum negative differentiated glottal source waveform, T_a is the duration of the return phase, T_c is the instant at the complete glottal closure, and E_e is the amplitude at the glottal closure instant. Since T_c is often set to T_0 in a simple LF model version, five parameters are used for further discussing emotional vowels in this paper. The LF model in the time domain can be formulated as:

$$u(n) = \begin{cases} E_1 e^{an} \sin(wn) & 0 \le n \le T_e \\ -E_2[e^{-b(n-T_e)} - e^{-b(T_0-T_e)}] & T_e \le n \le T_c \\ 0 & T_c \le n \le T_0 \end{cases} \quad (1)$$

where E_1, E_2, a, b and w are the parameters related to T_p, T_e, T_a ,E_e and T_0 [10].

Given the LF glottal source waveform derivative, the speech signal $s(n)$ can be synthesized by means of an ARX model:

$$s(n) = -\sum_{i=1}^{p} a_i(n)s(n-i) + b_0 u(n) + e(n). \quad (2)$$

where $a_i(n)$ are the coefficients of the p-order ARX model characterizing the vocal tract filter, b_0 is related to the amplitude of LF glottal source waveform derivative and $e(n)$ is the residual signal.

3. Results and discussion

The glottal source parameters of emotional vowels were estimated by the improved method of the ARX-LF model [16, 17].

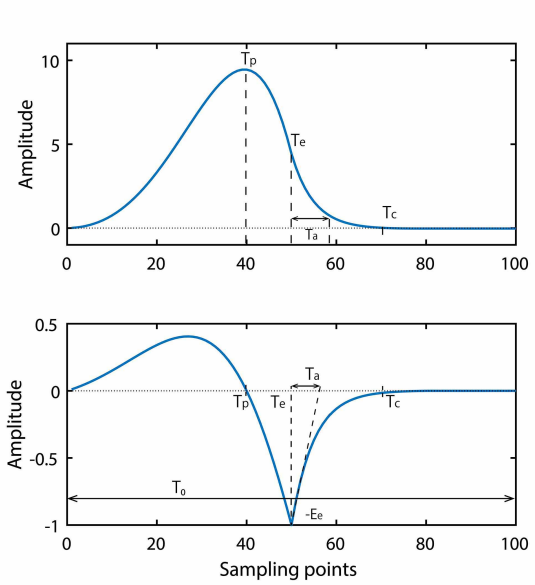

Figure 1: *One period of glottal source waveform (top) and its derivative waveform modeled by the LF model (bottom).*

Li *et al*. resynthesized vowels using averaged glottal source parameter values of the LF model, and the resynthesized vowels were presented to listeners for evaluating the perception of emotions. They validated that the perceived emotions of resynthesized vowels using averaged parameters of the LF model were similar to those of the original vowels [16]. Thus, in the following descriptions, the averaged parameter values of the LF model are used for discussion.

3.1. Results of glottal source waveform of emotional vowels

The glottal source waveforms of vowels for one period in four emotional styles for two males and two females are firstly analyzed. For all speakers, compared to the neutral vowel, the glottal source waveform of vowels with sadness is relatively round, while those of the vowel with joy and anger are relatively steep. These results were consistent with those in [15].

3.2. Results of glottal source parameters of emotional vowels

In order to find the differences of the estimated glottal source parameter values in emotional vowels, the parameter values of the LF model were extracted from the vowels with different emotions. The averaged estimated glottal source parameters of the LF model (T_p, T_e, T_a, E_e, and F_0(1/T_0)) were further compared for two males and two females. The averaged glottal source parameter values (the ratio of T_p, T_e, and T_a to one period, E_e, and F_0) for the two males and two females are detailed in Tables 1 and 2, respectively.

The F_0 of the emotional vowels was firstly examined among all parameters, since it is well-known to be important in perceptual emotions [19]. As shown in Tables 1 and 2, the F_0 values of the emotional vowels varied greatly for all speakers. Table 1 and Table 2 demonstrated that the values of F_0 in anger and joy voices are higher than those in neutral and sadness voices; the values of F_0 in sadness voice are the smallest for both male and female speakers. These differences in F_0

Table 1: *Averaged estimated glottal source parameters of emotional vowels for two males*

	Male 1					**Male 2**				
	T_p	T_e	T_a	E_e	$F_0(Hz)$	T_p	T_e	T_a	E_e	$F_0(Hz)$
Anger	0.42	0.54	0.04	0.38	160	0.34	0.47	0.07	0.48	333
Joy	0.47	0.67	0.09	0.09	280	0.49	0.62	0.06	0.12	218
Neutral	0.57	0.75	0.05	0.16	123	0.40	0.53	0.04	0.06	197
Sadness	0.49	0.70	0.09	0.03	103	0.52	0.68	0.07	0.03	154

Table 2: *Averaged estimated glottal source parameters of emotional vowels for two females*

	Female 1					**Female 2**				
	T_p	T_e	T_a	E_e	$F_0(Hz)$	T_p	T_e	T_a	E_e	$F_0(Hz)$
Anger	0.36	0.49	0.09	0.38	353	0.28	0.38	0.05	0.07	414
Joy	0.32	0.45	0.07	0.08	500	0.33	0.47	0.08	0.05	414
Neutral	0.48	0.63	0.09	0.07	250	0.50	0.63	0.08	0.08	267
Sadness	0.53	0.78	0.10	0.05	266	0.53	0.74	0.09	0.02	245

of different emotions were also found in many previous studies [20, 21]. Moreover, it was observed that the values of F_0 of females with a given emotional style was much higher than that of males, which was also pointed out by many studies [21, 22].

As an important physiological parameter describing the glottal open instance and closure instance in one period, the open quotient (OQ) is proportional to the parameter T_e of the LF model as shown in Tables 1 and 2. It can be found that the values of T_e in joy and anger voices are smaller than those in the neutral voice, while the values of T_e in sadness voice were larger than those in neutral voices. Among them, the values of T_e in anger and joy voices are the smallest. It is because OQ may be strongly related to phonation styles. These results are consistent with those in [23]. Similar to parameter T_e, the values of parameter T_p in anger and joy voice were the smallest, while that of sadness voice was the largest. However, Tables 1 and 2 show no clear consistent relations of values of T_e and T_p between male and female.

As shown in Tables 1 and 2, for male and female speakers, the values of E_e in anger voice were the largest, those of sadness voice were the smallest, and those of neutral and joy voice were in the middle. These differences were also pointed out by Li *et al*. [16]. Similar to parameters T_e and T_p, there were no clear consistent relations of values of E_e between male and female.

Tables 1 and 2 show that the values of T_a in sadness voices were the smallest for all speakers. However, there were no clear consistent relations between other emotional styles for males and females.

3.3. Set of glottal source parameters related to neutral vowel

The glottal source parameter values along with a fuzzy comparison of these parameters values among different emotions (high/low, big/small) were mentioned in section 3.2. How the patterns of glottal source parameter values differ among different emotions is the task of 3.3.

To further compare the glottal source parameters of emotional vowels based on the estimated glottal source parameters values in section 3.2, the percentage difference of the parameter values in joy, anger, and sadness compared to neutral voice were calculated. Let the parameter values be vector $\theta \in \{T_p, T_e, T_a, E_e, F_0\}$. The percentage difference of each parameter in emotional vowels ($d_j, j = joy, anger, sadness, neutral$) relative to neutral can be calculated as

$$d = \frac{\theta_j - \theta_n}{\theta_n} \times 100\%, \quad j = 1, 2, 3, 4. \tag{3}$$

where θ_n are five glottal source parameter values in the neutral voice. Note that the percentage difference of neutral voice for five glottal source parameters was equal to 0 by Eq. 3.

According to the percentage values d, they were divided into five levels: very low ($d < -25\%$), low ($-25\% \leq d < -5\%$), comparable ($-5\% \leq d < +5\%$), high ($+5\% \leq d < +25\%$), very high ($d \geq +25\%$). Different levels of glottal source parameters for anger, joy, sadness, and its averaged levels of four speakers are plotted in Figure 2. The values to represent different levels of emotions in Figure 2 ranged from -2 to +2 with a step of 1. Value -2 indicates very low level, -1 indicates low level, 0 indicates comparable level, +1 indicates high level, and +2 indicates very high level. Note that the five parameter levels of the neutral voice were 0.

As shown in Figure 2, for each emotion of the four speakers, the levels of the set of parameter values in each emotion were different. More specifically, it was found that each emotion has its own pattern of sets of glottal source parameter values.

For anger and joy voice in most speakers, most of the levels of glottal source parameter values were shared in joy and anger voice when compared with those of neutral voice. Joy and anger voice shows a very high level or high level for parameters F_0 and E_e, while a very low level or low level for parameters T_p and T_e. In contrast, the relative various levels for parameter T_a performed differently depending on speakers. These levels of parameter values in joy and anger voice may have a strong glottal source waveform(very high E_e) and a shallower spectral tilt, which is in line with the findings in [16]. When joy compares to anger voice, levels of parameters values of E_e and T_a were varied. More specifically, E_e with anger voice has a higher level than that of joy voice for three speakers, T_a with joy voice in male speakers has a higher lever than that of anger voice in female speakers. It is noted that the shared pattern of sets of glottal source parameter values may cause difficultly to distinguish anger and joy emotion. In fact, evaluation of valence (anger

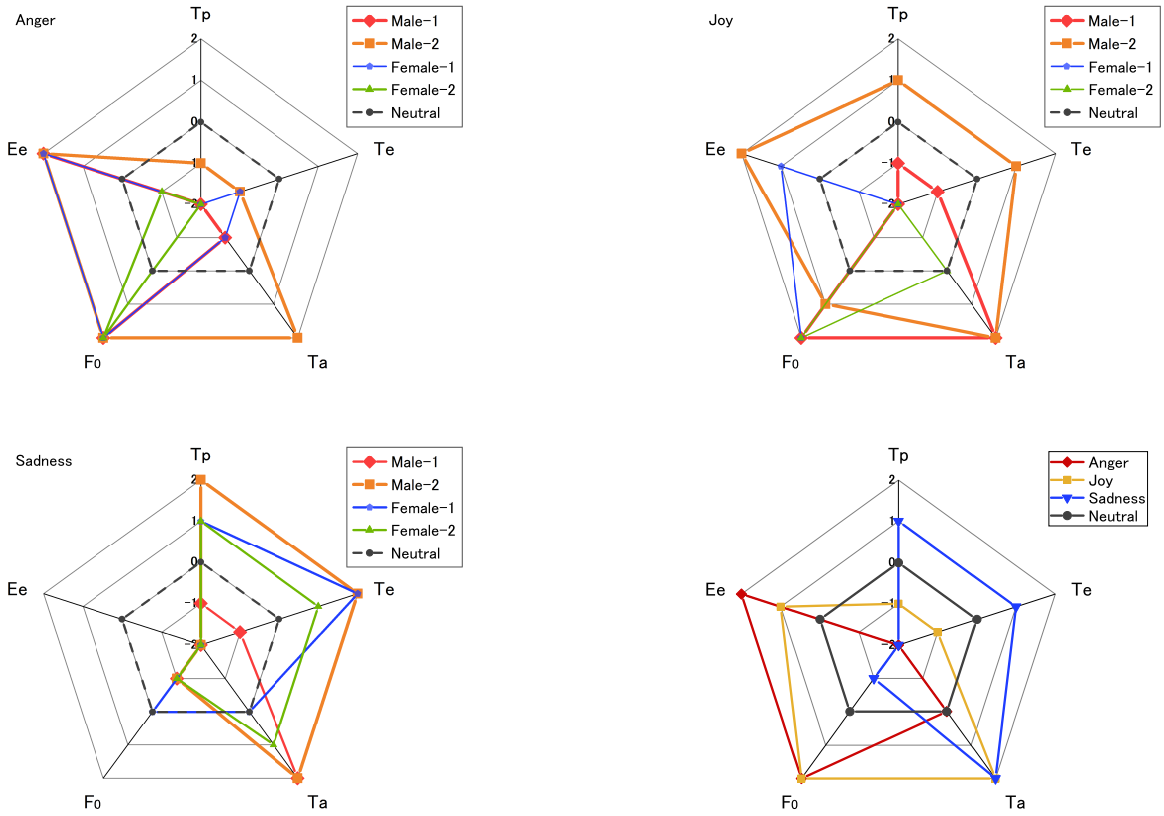

Figure 2: *Levels of glottal source parameters in two male and two female speakers for each emotion: -2 (very low level), -1 (low level), 0 (comparable: neutral level), +1 (high level), +2(very high level).*

and joy) is more difficult than that of arousal by speech emotion recognition systems, as reported in many studies [5, 24].

For sadness voice in four speakers, when compared with the levels of glottal source parameter values of neutral voice, sadness voice shows a high level and very high level for parameters T_p, T_e, and T_a, low level and very low level for parameters F_0 and E_e. These parameter values in sad voice may have a weak glottal source waveform (very low E_e and low F_0) and a sharper spectral tilt (higher T_e/OQ) since sadness is usually referred to as breathy voice, which is in line with the findings in [15, 25].

Average levels of glottal source parameter values in each emotion are shown in Figure 2, When compared to the neutral voice, anger voice has a very high level E_e and F_0, a very low level T_p and T_e. Joy voice has a very level F_0 and T_a, high level E_e, and a low level T_p and T_e. Sadness voice has a very high level T_a, high level T_p and T_e, low level F_0, and a very low level E_e.

The above results demonstrate that each emotion has its own pattern for the set of glottal source parameter values. The pattern on the set of glottal source parameter values in each emotion is slightly different for different speakers, which may be caused by the individual personalities and the different degrees of emotion.

4. Conclusion

In this paper, sets of glottal source parameters in emotional vowels were compared. Using the recently improved ARX-LF model-based analysis method, the glottal source parameter (T_p, T_e, T_a, E_e, and F_0) values were estimated from emotional vowels with four speakers (two males and two females). By comparing the estimated glottal source parameter values for each emotion, the results indicated that different emotion voices have their own patterns on glottal source parameter values.

The findings of this present study are useful to provide insight into emotional speech. It is believed that the results of this research contribute to speech emotion recognition and synthesis systems, and would be useful for improving performance of these applications.

Limitations of this study are the small number of speakers and only the vowels /a/ were discussed. Other factors (e.g., linguistic content, degree of emotion, and more vowels) should be taken in to account in future work.

5. Acknowledgements

This work is supported by the National Key Research & Development Plan of China (No.2017YFC0820602), the National Natural Science Foundation of China (NSFC) (No.61831022, No.61771472, No.61773379, No.61901473) and the Key Program of the Natural Science Foundation of Tianjin (Grant No. 18JCZDJC36300).

6. References

[1] P. N. Juslin and K. R. Scherer, "Vocal expression of affect," *The new handbook of methods in nonverbal behavior research*, pp. 65–135, 2005.

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

[3] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–577.

[4] J. Tao, Y. Li, and S. Pan, "A multiple perception model on emotional speech," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–6.

[5] X. Li and M. Akagi, "Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model," *Speech Communication*, vol. 110, pp. 1–12, 2019.

[6] D. Erickson, A. Rilliard, T. Shochi, J. Han, H. Kawahara, and K. Sakakibara, "A cross-linguistic comparison of perception to formant frequency cues in emotional speech," *COCOSDA, Kyoto, Japan*, pp. 163–167, 2008.

[7] K. R. Scherer, "Vocal affect expression: A review and a model for future research." *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[8] M. Airas and P. Alku, "Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalised amplitude quotient," *Phonetica*, vol. 63, no. 1, pp. 26–46, 2006.

[9] A. N. Chasaide and C. Gobl, "Voice quality and f0 in prosody: Towards a holistic account," *Scientific Programming*, 2004.

[10] G. Fant, J. Liljencrants, and Q.-g. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[11] S. R. Kadiri, P. Alku, and B. Yegnanarayana, "Comparison of glottal closure instants detection algorithms for emotional speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7379–7383.

[12] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of excitation source features of speech for emotion recognition," in *INTERSPEECH 2015*, 2015.

[13] R. Sun, E. Moore, and J. F. Torres, "Investigating glottal parameters for differentiating emotional categories with similar prosodics," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4509–4512.

[14] Z. Xiao, Y. Chen, and Z. Tao, "Contribution of glottal waveform in speech emotion: A comparative pairwise investigation," in *2018 IEEE International Conference on Progress in Informatics and Computing (PIC)*. IEEE, 2018, pp. 185–190.

[15] Y. Li, K.-I. Sakakibara, D. Morikawa, and M. Akagi, "Commonalities of glottal sources and vocal tract shapes among speakers in emotional speech," in *International Seminar on Speech Production*. Springer, 2017, pp. 24–34.

[16] Y. Li, J. Li, and M. Akagi, "Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space," *The Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. 908–916, 2018.

[17] Y. Li, K.-I. Sakakibara, and M. Akagi, "Simultaneous estimation of glottal source waveforms and vocal tract shapes from speech signals based on arx-lf model," *Journal of Signal Processing Systems*, vol. 92, pp. 831–838, 2020.

[18] D. Vincent, O. Rosec, and T. Chonavel, "Estimation of lf glottal source parameters based on an arx model," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[19] Y. Xue, Y. Hamada, and M. Akagi, "Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space," *Speech Communication*, vol. 102, pp. 54–67, 2018.

[20] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE transactions on Audio, Speech, and Language processing*, vol. 14, no. 4, pp. 1145–1154, 2006.

[21] D. Erickson, C. Zhu, S. Kawahara, and A. Suemitsu, "Articulation, acoustics and perception of mandarin chinese emotional speech," *Open Linguistics*, vol. 1, no. open-issue, 2016.

[22] E. Ramdinmawii and V. K. Mittal, "Emotional speech discrimination using sub-segmental acoustic features," in *2017 2nd International Conference on Telecommunication and Networks (TELNET)*. IEEE, 2017, pp. 1–7.

[23] S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, and J. Schließer, *Methods in empirical prosody research*. Walter de Gruyter, 2012, vol. 3.

[24] R. Elbarougy and M. Akagi, "Improving speech emotion dimensions estimation using a three-layer model of human perception," *Acoustical science and technology*, vol. 35, no. 2, pp. 86–98, 2014.

[25] I. Yanushevskaya, C. Gobl, and A. N. Chasaide, "Voice parameter dynamics in portrayed emotions," in *MAVEBA*, 2009.