Commonalities of glottal sources and vocal tract shapes among speakers in emotional speech

Yongwei Li[1], Ken-Ichi Sakakibara[2], Daisuke Morikawa[3], Masato Akagi[1]

[1]*Japan Advanced Institute of Science and Technology*
[2]*Health Sciences University of Hokkaido*
[3]*Toyama Prefectural University*
{yongwei, akagi}@jaist.ac.jp, kis@hoku-iryo-u.ac.jp, dmorikawa@pu-toyama.ac.jp

Abstract

Discussion on commonalities of the glottal source waves and vocal tract shapes among speakers in emotional-speech production can help us to carry out further speech processing from the speech-production point of view. This requires accurate and independent measurement of glottal source waves and vocal tract shapes. Although magnetic resonance imaging and electromagnetic articulography can help achieve this requires, they are time consuming and costly. This paper explores the commonalities of the glottal source waves and vocal tract shapes among speakers in emotional speech based on a source-filter model. With the help of the proposed precise estimation algorithm, the glottal source waves and vocal tract shapes of emotional voices (vowel:/a/) were estimated, in which the emotional speech with four basic emotional speaking styles (neutral, joy, anger and sad) were uttered by four different speakers. The results are as follows. When compared with the spectra tilts of glottal source waves of the neutral emotion, (1) those of anger and joy increased, and that of sad decreased in the 200- to 700-Hz frequency range; (2) that of anger increased, but that of joy decreased, and that of sad was the same as that of neutral in 700- to 2000-Hz range; (3) all spectral tilts had the same tendency over 2000 Hz. For front vocal tract shapes, the area function of anger was the largest, that of sad was smallest, and those of joy and neutral were in the middle.

Keywords: emotional speech, ARX-LF model, glottal source waves, vocal tract shapes

1. Introduction

Emotional-speech recognition and synthesis are hot topics in speech-signal processing. Investigating the commonalities among speakers in emotional speech is important for emotional speech-signal processing. Although the commonalities of acoustic features of emotional speech have been investigated and used for emotional speech conversion and recognition (Schröder et al. 2001; Hamada, Elbarougy, R., and Akagi, M. 2014; Li and Akagi, M. 2016), it was also shown that it is difficult to model emotions by using only these acoustic features (Banse and Scherer, K. R. 1996). Thus, it has been suggested to investigate features of speech-production organs, such as glottal source waves and vocal tract shapes (Banse and Scherer, K. R. 1996). This requires accurate and independent measurement of glottal source waves and vocal tract shapes. However, the properties of production organs for emotional speech have not been extensively investigated for discussing commonalities. This is because (1) measurement of vocal-fold vibration and vocal tract shape simultaneously and directly when uttering emotional speech such as using magnetic resonance imaging (MRI) and electromagnetic articulograpgy (EMA), are precise but costly (Kitamura 2010; Erickson et al. 2016), and (2) independently and precisely estimating glottal source waves and vocal tract shapes from uttered emotional speech is still challenging.

The aim of this paper is to independently investigate the commonalities of glottal source waves and vocal tract shapes among speakers in emotional speech. To achieve this, three problems should be solved, (1)collecting emotional speech data, (2) separating glottal source waves and vocal tract shapes from emotional speech signals, and (3) discussing commonalities among speakers.

2. Database

Vowels /a/ were uttered by four actors three males and a female- with different speaking styles including eight emotion states; neutral, happy, sad, afraid, disgust, relax, surprise and anger. Neutral was uttered once and the other seven emotions were uttered three times with different degrees (weak, normal, and strong). Thus, there are a total of 88 (1+7*3=22 for each speaker) utterances.

2.1. Data selection

Even when listening to utterances of actors, listeners sometimes perceive the different emotions with the actors portrayed. Thus, a listening experiment was carried out to evaluate actors' utterances.

There are two types of emotion-evaluation approaches, categorical approach and dimensional approach (valence-arousal:V-A) (Hamada, Elbarougy, R., and Akagi, M. 2014). Using the dimensional approach, not only category but also degree of emotion can be described in the V-A space. Since the database has three different degrees of emotions, the dimensional approach was adopted to evaluate emotions.

Ten people participated in the listening test. For arousal and valence evaluations, a 7-point scale from -3 to 3 (-3:very negative to 3:very positive for valence and -3:very calm to 3:very exited for arousal) was used, and the average evaluation values of the ten participants were calculated. Four basic emotion categories (neutral, joy, sad, and anger) with strong degree were selected from the database for further discussion of the commonalities on glottal source waves and vocal tract shapes. The averages of the evaluated values of the selected speech data in the V-A space are shown in Figure 1.

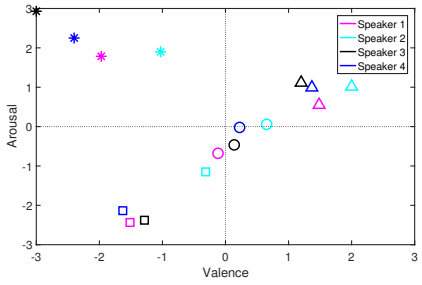

Figure 1: Selected speech data in V-A spaces. Emotional states; triangle: joy; asterisk: anger; circle: neutral; square: sad.

3. Separation of glottal source waves and vocal tract shapes

A source-filter model was proposed by Fant (Fant, Liljencrants, J., and Lin, Q. 1985), and is widely used to represent the speech-production process. Among the source-filter models, the ARX-LF model, which combines the auto-regressive exogenous (ARX) model with the Liljencrant-Fant (LF) model has been widely used for representing glottal source waves and vocal tract shapes with widely pronounced types, such as (Vincent, Rosec, O., and Chonavel, T. 2005; Kane and Gobl, C 2013). Thus, it is possible to use the ARX-LF model to estimate glottal source waves and vocal tract shapes for emotional speech

3.1. LF model

The LF model has six parameters to represent the derivative of the glottal flow; five parameters concerning time T_p, T_e, T_a,T_c, T_0 and one parameter concerning amplitude E_e as shown in Fig. 2. The glottal opening instant (GOI) is set to 0, and T_0 is the end of the period, T_p is the phase of the maximum opening of the glottis, T_e is the open phase of the glottis, T_a is the return phase, T_c is end of return phase and E_e is the amplitude at the glottal closure instant (GCI) point. The LF model in the time domain is formulated as Equation 1.

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

Parameters E_1, E_2, a, b and w are implicitly related to T_p, T_e, T_a ,E_e and T_0, respectively (Fant, Liljencrants, J., and Lin, Q. 1985).

3.2. ARX model

The ARX model simulates a vocal tract filter. The speech production process can be modeled as a time-varying system as follows:

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

where $s(n)$ is the observed speech signal, $u(n)$ is the derivative of the glottal waveform (LF waveform) at time n, $a_i(n)$ and $b_0(n)$ are coefficients of the IIR filter, p is filter order, and $e(n)$ is the residual.

The output signal of the LF model acts as an input signal $u(n)$ to the vocal tract filter. Equation(2) is called the ARX model, and the output signal $x(n)$ is a periodic component and $e(n)$ is a non-periodic component in speech.

$$x(n) = -\sum_{i=1}^{p} a_i(n)s(n-i) + b_0(n)u(n) \quad (3)$$

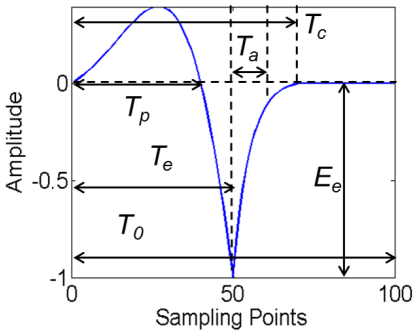

Figure 2: LF model

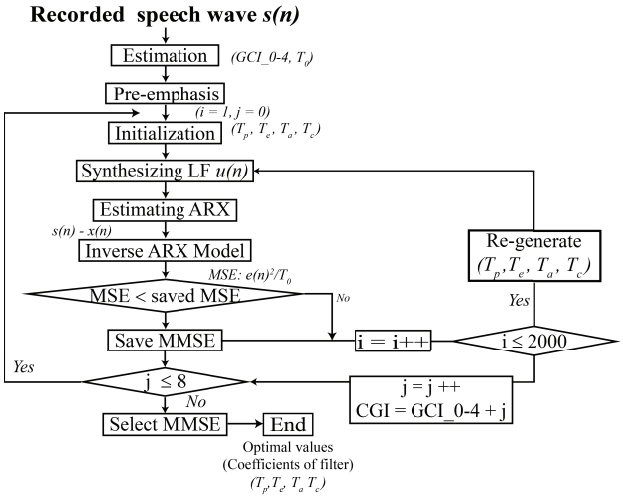

Figure 3: Estimation algorithm of glottal source wave and vocal tract shape

3.3. Scheme of analysis

The estimation procedure for a period of a glottal source wave is shown in Fig. 3, in which two main processes are included. In the first process, LF parameters and vocal tract coefficients can be obtained with a fixed GCI from a differential electro glottograph (EGG) signal as initial values. The initial values of the LF parameters are used for synthesizing $u(n)$. $u(n)$ is then exploited to re-synthesis $x(n)$ using the ARX model with the parameters of the vocal tract filter updated within each period in the mean square error (MSE) sense for $e(n)$ with the help of least square (LS) method (Ohtsuka and Kasuya, H. 2001). For the initial values, T_e is estimated from the signal by searching the GCI and GOI.

Table 1: Average error γ

	T_p	T_e	T_a	T_c	$F1$	$F2$
/γ/(%)	5.75	3.41	34.6	5.65	2.05	0.59

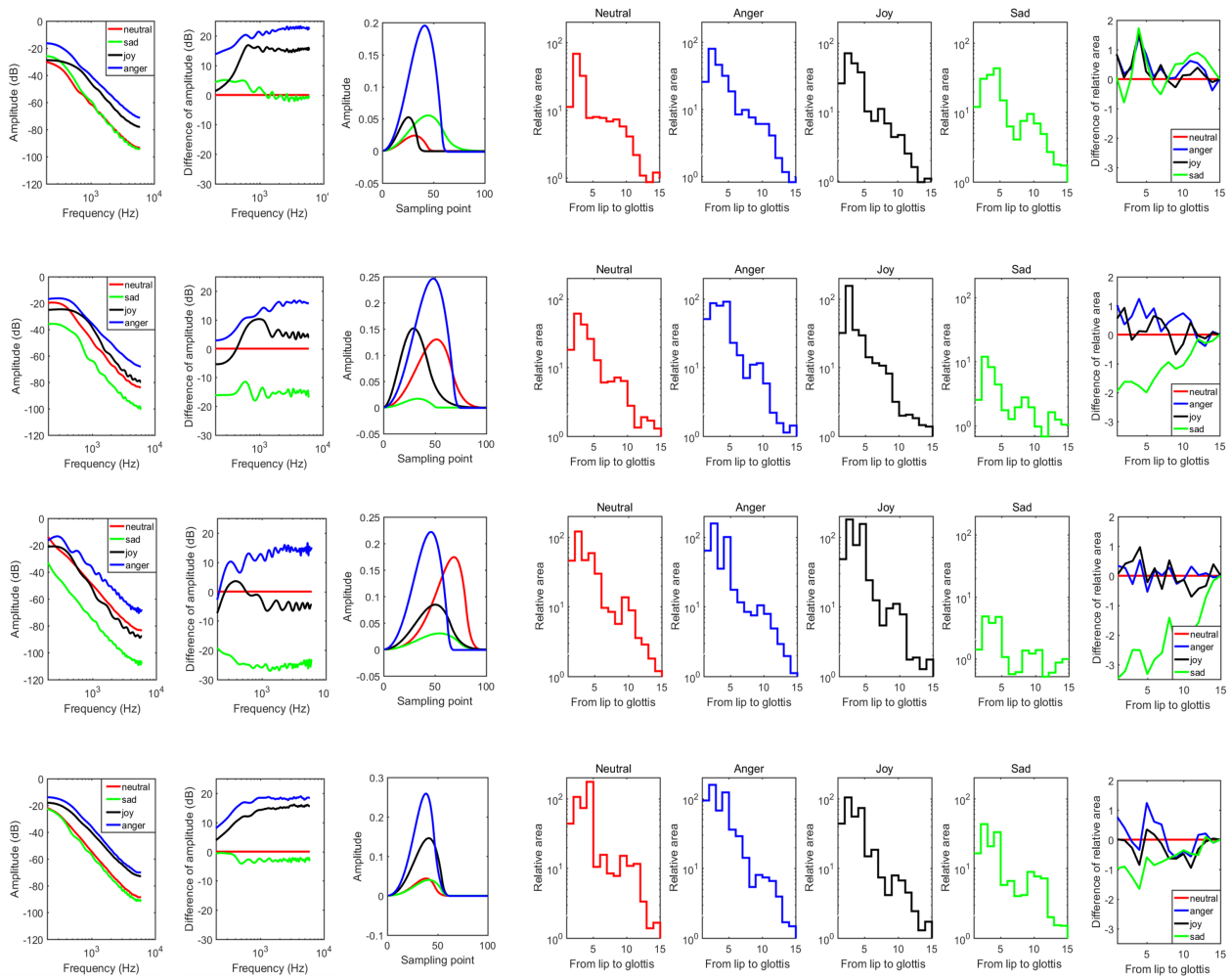

Figure 4: Results of four speakers (one speaker per row): (a) Spectra of glottal source wave (first coulumn); (b) difference in spectra between neutral and other emotions (second column); (c) glottal source waves (third column); (d) vocal tract area functions and their difference between neutral and other emotions (other columns).

In the second process, we want to obtain more accurate LF parameters and vocal tract coefficients. The GCI parameters shift around the initial GCI, and the first process is updated again for the shifted GCI. For the given GCI, the iteration process in the minimization of mean-square error (MMSE) optimization is set to 2000. The optimal glottal source parameters and vocal tract coefficients are finally estimated by MMSE.

3.4. Evaluation of proposed estimated schemes

A frequently used method for evaluating the estimation algorithm is to estimate synthetic vowels, given the parameter values of synthesized voices. The accuracy of the estimation algorithm can be evaluated by comparing estimated parameter values with referenced values.

3.4.1. Synthesized data

Synthesized voices are produced using Kawahara's method (Kawahara et al. 2015). For the LF parameters, T_e: 0.3 to 0.9 with steps of 0.05; T_p/T_e: 0.65 to 0.85 with steps of 0.05 steps; T_a: 0.03, 0.08; T_0s are estimated from actual voices. Two types of typical vowels are considered (/a/, /i/) for the filter. The total number of synthesized voices is 14040 period.

3.4.2. Results and discussion

The values of (T_p, T_e, T_a, T_c) as glottal source wave parameters and first and second formant frequencies (F_1 and F_2) as vocal tract shape parameters were evaluated based on the reference values. The errors (γ) between the reference parameters $\beta \in \{T_p, T_e, T_a, T_c, F_1, F_2\}$ and estimated parameters $\hat{\beta}$ were calculated as follows:

$$\gamma = \frac{|\hat{\beta} - \beta|}{\beta} \times 100\% \tag{4}$$

The average errors γ are listed in Table 1, which shows that most of the parameters of the ARX-LF model can be correctly estimated, and errors are less than 6% except for T_a. Since T_a was the smallest, error of T_a was (34.6%), which is the largest compared with the other parameters.

Table 2: First formant frequency [Hz] values among speakers

	anger	joy	neutral	sad
Speaker 1	861	902	762	650
Speaker 2	920	1020	885	961
Speaker 3	727	791	809	691
Speaker 4	973	861	703	709

3.5. Estimation of glottal source wave and vocal tract shape from actual emotional speech

To discuss the commonalities of glottal source waves and vocal tract shapes among speakers in emotional speech, glottal source waves and vocal tract shapes were estimated from emotional speech uttered by four different speakers. The estimated results are illustrated in Figure 4.

4. Commonalities of glottal sources and vocal tract shapes

Since the spectral tilts are frequently used to describe the characteristics of glottal source waves, the spectral tilt was adopted to discuss the commonalities of glottal source waves properties. For vocal tract shapes, the most obvious characteristic is vocal tract area functions. Thus, area functions normalized with the glottis side was adopted to discuss the commonalities and vocal tract shapes properties.

The commonalities of glottal source waves properties among speakers are summarized in Figure 4. When compared with the spectral tilts of the glottal source waves of neutral emotion (1) those of anger and joy increased, and that of sad decreased in the 200- to 700-Hz frequency range; (2) that of anger increased, but that of joy decreased, and that of sad was the same as that of neutral in the 700- to 2000-Hz range; (3) all spectral tilts had the same tendency over 2000 Hz.

The commonalities of vocal tract shapes properties among speakers are summarized in Figure 4. The area function of anger was the largest, that of sad was smallest, and those of joy and neutral were in the middle.

Moreover, first formant frequency (F1) values were calculated among speakers and emotions and are listed in Table 2. Table 2 and Figure 4 show that, for four speakers, the mouth opens more, resulting in higher F1 for anger and joy (see Erickson et al. 2016 who reported a similar finding using EMA).

5. Conclusion

The commonalities of glottal source waves and vocal tract shapes among speakers in emotional speech were discussed in the following three steps. (1) For collecting emotional speech data, the emotions and degree of speech of the database were evaluated using the dimensional approach, and four basic emotions with strong degree (neutral, joy, anger and sad) were selected from the V-A emotional spaces. (2) For separating glottal source waves and vocal tract shapes from emotional speech signals, they were estimated from selected emotional speech data using a proposed precise estimation algorithm based on the ARX-LF model. (3) For discussion on the commonalities of glottal source waves and vocal tract shapes among speakers, the spectra of glottal source waves and width of the front area function were investigated. For glottal source waves, those of anger and joy increased, and that of sad decreased in the 200- to 700-Hz frequency range; that of anger increased, but that of joy decreased, and that of sad was the same as that of neutral in the 700- to 2000-Hz; all spectral tilts had the same tendency over 2000 Hz. For front vocal tract shapes, the area function of anger was the largest, that of sad was smallest, and those of joy and neutral were in the middle. These results are similar to the previous finding using EMA.

The results are expected to be used for further discussion of the commonalities of glottal source waves and vocal tract shapes of emotional speech among speakers and applications to emotional speech processing from the speech production points of view.

The commonalities of four basic emotions with strong degree were discussed. For future work , we will discuss the commonalities of these emotions with different degrees and attempt to find contributions of glottal source waves and vocal tract shapes on the perception of emotions for further understand emotional speech production mechanisms.

6. Acknowledgements

This study was supported by a Grant-in-Aid for Scientific Research (A) (No. 25240026) and China Scholarship Council (CSC).

7. References

Banse, R. and Scherer, K. R. (1996). "Acoustic profiles in vocal emotion expression." In: *Journal of personality and social psychology* 70.3, p. 614.

Erickson, D., Zhu, C., Kawahara, S., and Suemitsu, A. (2016). "Articulation, Acoustics and Perception of Mandarin Chinese Emotional Speech". In: *Open Linguistics* 2.1.

Fant, G., Liljencrants, J., and Lin, Q. (1985). "A four-parameter model of glottal flow". In: *STL-QPSR* 4.1985, pp. 1–13.

Hamada, Y., Elbarougy, R., and Akagi, M. (2014). "A method for emotional speech synthesis based on the position of emotional state in Valence-Activation space". In: *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*. IEEE, pp. 1–7.

Kane, J. and Gobl, C (2013). "Evaluation of Automatic Glottal Source Analysis". In: *Advances in Nonlinear Speech Processing*, pp. 1–8.

Kawahara, H., Sakakibara, K. I., Banno, H., Morise, M., Toda, T., and Irino, T. (2015). "Aliasing-free implementation of discrete-time glottal source models and their applications to speech synthesis and F0 extractor evaluation". In: *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*. IEEE, pp. 520–529.

Kitamura, T. (2010). "Similarity of effects of emotions on the speech organ configuration with and without speaking". In: *Eleventh Annual Conference of the International Speech Communication Association*.

Li, X. and Akagi, M. (2016). "Multilingual Speech Emotion Recognition System Based on a Three-Layer Model." In: *INTERSPEECH*, pp. 3608–3612.

Ohtsuka, T. and Kasuya, H. (2001). "Aperiodicity control in ARX-based speech analysis-synthesis method". In: *Seventh European Conference on Speech Communication and Technology*.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., and Gielen, S. (2001). "Acoustic correlates of emotion dimensions in view of speech synthesis". In: *Seventh European Conference on Speech Communication and Technology*.

Vincent, D., Rosec, O., and Chonavel, T. (2005). "Estimation of LF glottal source parameters based on an ARX model." In: *Interspeech*, pp. 333–336.