Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space

Yongwei Li,[1,a)] Junfeng Li,[2,b)] and Masato Akagi[1]
[1]*School of Information Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan*
[2]*Institute of Acoustics, Chinese Academy of Sciences, 21 North 4th Ring Road, Haidian District, Beijing 100190, People's Republic of China*

(Received 12 February 2018; revised 24 July 2018; accepted 6 August 2018; published online 22 August 2018)

Motivated by the source-filter model of speech production, analysis of emotional speech based on the inverse-filtering method has been extensively conducted. The relative contribution of the glottal source and vocal tract cues to perception of emotions in speech is still unclear, especially after removing the effects of the known dominant factors (e.g., F_0, intensity, and duration). In this present study, the glottal source and vocal tract parameters were estimated in a simultaneous manner, modified in a controlled way and then used for resynthesizing emotional Japanese vowels by applying a recently developed analysis-by-synthesis method. The resynthesized emotional vowels were presented to native Japanese listeners with normal hearing for perceptually rating emotions in valence and arousal dimensions. Results showed that glottal source information played a dominant role in perception of emotions in vowels, while vocal tract information contributed to valence and arousal perceptions after neutralizing the effects of F_0, intensity, and duration cues.
 https://doi.org/10.1121/1.5051323

[ZZ] Pages: 908–916

I. INTRODUCTION

Emotions in speech are extensively used by humans to express intentions and play an important role in speech understanding (Juslin and Scherer, 2005). The same textual message can be conveyed with different meanings by incorporating an appropriate emotion, and the intended message can be understood from the underlying emotions together with phonetic information and other linguistic factors. The perception of emotions in speech is usually described by means of a categorical or dimensional approach. The categorical approach, where emotions in speech are generally perceived as discrete states (e.g., *neutral*, *joy*, *anger*, and *sadness*), cannot capture the perceptual complexity of emotions (Scherer, 2003). In contrast, the dimensional approach provides a more flexible and continuous description of emotions in which the degree of emotions is usually represented in a two-dimensional space by arousal (the degree to which the listener perceives the emotion on a scale from calm to excited) and valence (the degree to which the listener perceives the emotion on a scale of negative emotion to a positive one) (Sauter *et al.*, 2010). The dimensional description of emotional speech is therefore widely used in analysis and recognition of emotions in speech in recent years.

The production process is often characterized by the glottal source and the acoustics of the vocal tract (the filter), which can be measured with instruments or estimated algorithmically. In recent years, some studies began to analyze emotions in speech using electromagnetic articulograph (EMA) (Erickson, 2004; Li *et al.*, 2010), magnetic resonance imaging (MRI) (Kim *et al.*, 2014; Lee *et al.*, 2006), and high-speed videoendoscopy (HSV) (Degottex *et al.*, 2008). However, these instrumental studies lacked flexibility for analyzing different degrees of emotional speech; moreover, it is also difficult and costly to collect sufficient recordings of emotional speech data. Therefore, many researchers attempted emotional speech analysis based on speech production models, e.g., the source-filter model (Fant *et al.*, 1985).

In many previous studies of emotional speech analysis, the glottal source (GS) and vocal tract (VT) parameters associated with the speech production models were usually estimated using the inverse filtering based approach (Iliev *et al.*, 2010; Yanushevskaya *et al.*, 2009; Yanushevskaya *et al.*, 2007). Based on the source-filter theory, the acoustic features of emotional speech have been found to be primarily source-related cues, including duration (e.g., utterance duration), intensity (e.g., loudness), F_0 (rate of vocal-fold vibration), and spectral cues (e.g., spectral flatness and harmonic-to-noise ratio) (Juslin and Laukka, 2001; Juslin and Scherer, 2005; Scherer, 2003). Many studies showed that the GS information plays a more important role in emotion than the VT ones. Juslin and Laukka (2001) and Scherer (2003) found that F_0 is the most important cue for emotional speech and that the higher mean F_0 usually correlates with the higher level of arousal, whereas the duration and spectral cues are generally used to predict the valence of emotion. The filter-related cues are mainly associated with the formant frequencies (e.g., F_1 and F_2). Mori and Kasuya (2007) suggested that F_1 and F_2 are important for perceiving the

[a)]Electronic mail: yongwei@jaist.ac.jp
[b)]Also at: School of Electronic Electrical and Communication Engineering, University of Chinese Academy of Sciences, 19(A) Yuquan Road Shijingshan District, Beijing 100049, People's Republic of China.

valence of emotional speech, and that the higher F_1 and F_2 correspond to the more positive valence perception of emotion. Scherer (1986) suggested that the key to vocal differentiation of discrete emotions seems to be *voice quality* and humans can perceive *voice quality* independent of F_0. Since expression of emotions in speech contains much redundance, perception of some emotional states is still possible even when many vocal cues are eliminated. To further examine the relative importance of the acoustic features on emotional speech, Erickson *et al*. (2008) modified the emotional speech by setting its F_0 at a constant value of 300 Hz and found that a decrease in the values of F_2, F_3, and F_4 often leads to the perception of *sad* emotion. Through keeping the overall sound pressure level constant and artificially setting the F_0 to 200 Hz, Laukkanen *et al*. (1997) showed that GS contribute to perception of arousal, and that GS and formants contribute to perception of valence.

To further analyze emotions in speech, a high-quality analysis-by-synthesis method has been recently developed which can accurately estimate the GS and VT information simultaneously based on the Auto-Regressive eXogenous with Lilijencrant-Fant (ARX-LF) model (Li *et al*., 2017). Since vowel perception can provide an important insight into speech perception (Airas and Alku, 2006; Lee *et al*., 2004; Leinonen *et al*., 1997; Ringeval and Chetouani, 2008), perception of emotions in Japanese vowels was examined in this present study. Using the analysis-by-synthesis method, the GS and VT parameters are first estimated, artificially modified and then used for resynthesizing the emotional vowels. Two psychoacoustic tests are conducted to examine the contributions of the GS and VT cues for perception of emotions in vowels. In the first experiment, the emotional vowels are resynthesized using the GS parameters from one emotional state and the VT parameters from another emotional state; while in the second experiment, the F_0, duration and intensity parameters are neutralized to those of neutral vowels before resynthesizing the emotional vowels. The synthesized emotional vowels are then presented to native Japanese listeners with normal hearing for perceptually rating emotions in terms of valence and arousal. Experimental results showed that the GS information plays a dominant role in perception of emotions in vowels relative to the VT cue, and that the VT cues contribute to emotion perception in valence and arousal after neutralizing the effects of F_0, intensity and duration cues.

II. THE ANALYSIS-BY-SYNTHESIS APPROACH FOR EMOTIONAL VOICE

In this section, the recently developed analysis-by-synthesis approach for emotional voice based on the Auto-Regressive eXogenous with Lilijencrant-Fant (ARX-LF) model (Li *et al*., 2017) is briefly reviewed.

A. ARX-LF model

According to the source-filter theory of speech production, the glottal source signal in the ARX-LF model is represented by the LF glottal flow derivative and the vocal tract transfer function by the ARX filter. More specifically, the glottal flow derivative is formulated in the LF model by six parameters, where five parameters are related to time and one parameter is related to amplitude. The definitions of these parameters are listed in Table I.

A typical LF glottal flow derivative is plotted in Fig. 1. The explicit expression of the LF glottal flow derivative for one fundamental period is given by

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

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 (Fant *et al*., 1985).

Given the LF glottal flow derivative, the speech signal $s(n)$ can be synthesized by means of an ARX model (Ohtsuka and Kasuya, 2002):

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

where $a_i(n)$ are the coefficients of the p-order ARX model characterizing the vocal tract, b_0 is the ARX filter gain and is assumed to be $b_0 = 1$, and $e(n)$ is the residual signal.

B. Analysis and synthesis of emotional voice based on ARX-LF model

In the *analysis* step of the analysis-by-synthesis approach, the GS and VT parameters of emotional voice are estimated using Eq. (2); in the *synthesis* step, these parameters can be modified and further used to resynthesize emotional voice using Eq. (2).

The ARX-LF model has been widely used for analyzing neutral voice in the past several decades (Fu and Murphy, 2006; Vincent *et al*., 2005), and it was extended for analyzing singing (Lu, 2002). Due to the difference in acoustic characteristics between neutral and emotional voice, an estimation approach for the parameters of the ARX-LF model was developed to improve the analysis-by-synthesis quality for emotional voice (Li *et al*., 2017). The parameter estimation of the ARX-LF model includes *parameter initialization* and *parameter update*. In the *initialization* process, the initial value of the glottal closure instant (GCI) parameter was suggested to be further shifted around the value determined by the SEDREAMS method (Drugman *et al*., 2012a), which took the properties of emotional voice into account and improved the GCI estimation accuracy; the parameters of

TABLE I. Definitions of the parameters describing the glottal flow derivative in the LF model.

T_0	One period of glottal flow
T_p	Instant of the maximum glottal flow
T_e	Instant of the maximum negative differentiated glottal flow
T_a	Duration of the return phase
T_c	Instant at the complete glottal closure
E_e	Amplitude at the glottal closure instant

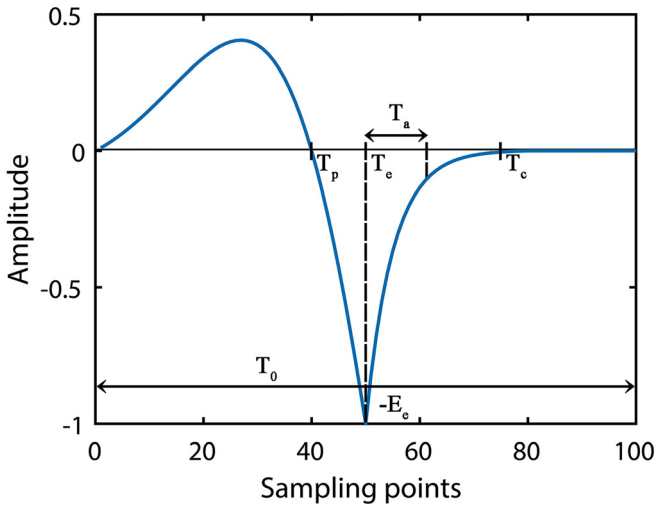

FIG. 1. (Color online) A typical one period of the LF glottal flow derivative.

the LF model were then empirically initialized to the values with the constraints suggested in Drugman *et al.* (2012b). In the *update* process, the parameters of the ARX-LF model were iteratively optimized in the sense of minimizing the mean square error for each period of the glottal source, with which the high-quality speech signal could be eventually resynthesized. For the detailed implementation of this analysis-by-synthesis approach, please refer to Li *et al.* (2017).

III. EXPERIMENT I: RELATIVE CONTRIBUTION OF GLOTTAL SOURCE AND VOCAL TRACT TO EMOTIONAL VOWEL PERCEPTION

A. Method

1. Stimuli

The Japanese voiced vowel (/a/) with four different emotions (i.e., neutral, joy, anger, and sadness) uttered by two professional male actors was used for the listening tests. These vowels are called original emotional vowels hereafter. The parameters of the ARX-LF model related to glottal source and vocal tract were extracted using the analysis-by-synthesis approach described in Sec. II. Since the estimation of these parameters was done for each glottal period, the estimated parameters (except T_0 and E_e) were first averaged across all periods, while T_0 and E_e were kept as those of the original signals. To examine the relative importance of GS and VT on emotional vowel perception, the emotional vowels were resynthesized by using the averaged GS parameters from one emotional state and the averaged VT parameters from another emotional state. As a result, 16 artificially synthesized emotional vowels were obtained for each speaker. To remove the effect of power difference on emotion perception, the energy of the synthesized emotional vowels was normalized in root mean square (RMS) prior to the listening tests.

2. Subjects

Ten normal-hearing listeners (four females and six males) participated in this experiment. All subjects were native Japanese-speaking listeners and were paid for their participation. The subjects were post-graduate students at the Japan Advanced Institute of Science and Technology, ranging in age from 23 to 30 yr old.

3. Procedure

For each speaker's vowel, there was a total of 16 resynthesized signals. All stimuli were presented to each subject at a constant sound pressure level of 65 dB through HDA-200 headphones in a soundproof room. In the test, each subject listened to a total of 32 tokens of resynthesized vowels (2 speakers × 4 sets of the glottal source parameters × 4 sets of the vocal tract parameters) and eight original emotional vowels (2 speakers × 4 emotional states). The presentation orders of the stimuli were randomized across each subject. Each subject was asked to give a score for each stimulus based on her/his perceptual impression of valence and arousal. Before the test, the basic information of emotion dimensions and the meanings of valence and arousal were introduced to subjects, as done in Elbarougy and Akagi (2014). The scores for the perceptual evaluation of emotions in the valence-arousal (V-A) space ranged from −2 to with a step of 0.1. For the evaluation of valence, the score −2 indicates very negative, with very positive; for the evaluation of arousal, the score −2 indicates very calm, with very excited.

B. Results

The mean perceptual scores of the 16 synthesized emotional vowels and the original emotional vowels in the V-A space across 10 subjects for two speakers are plotted in Fig. 2. As shown in Fig. 2, for two speakers, since each emotional vowel synthesized with the averaged GS and VT parameters is close to its original emotional vowel, the averaged parameters of the ARX-LF model still contain much information in the V-A space. The emotional vowel synthesized with the neutral GS and neutral VT parameters was perceptually scored to almost (0, 0) in the V-A space, and the scores for the emotional vowels synthesized with the neutral GS and arbitrary VT parameters were also quite close to the center. Furthermore, it was found that the perceptual scores for the emotional vowels synthesized with the anger GS and arbitrary VT parameters were negative in valence and positive in arousal, while the emotional vowels synthesized with the joy GS and arbitrary VT parameters were perceptually scored as positive in both valence and arousal. But emotional vowels synthesized with the sad GS and arbitrary VT parameters were perceptually scored as negative in both valence and arousal, especially for speaker 1. More importantly, it was noted that the perceptual differences (i.e., the distances in the V-A space) for the emotional vowels synthesized with the given GS and arbitrary VT parameters were much smaller than those for the emotional vowels synthesized with the given VT and arbitrary GS parameters. In other words, in the V-A space, the vowels synthesized with the given GS and arbitrary VT parameters were centralized, whereas the vowels synthesized with the given VT and arbitrary GS parameters were more dispersedly distributed.

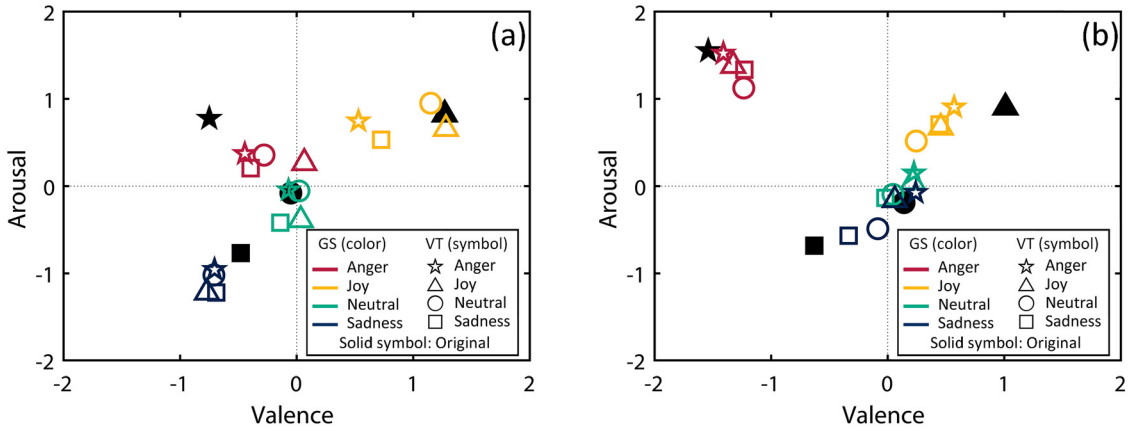

FIG. 2. (Color online) The perceptual results in the valence-arousal space of the emotional vowels synthesized with the arbitrary glottal source (GS) parameters and the vocal tract (VT) parameters for speaker 1 (a) and speaker 2 (b). The results of the original emotional vowels are also plotted in the V-A space with the solid symbols.

To examine the effects of GS parameters (GS-anger, GS-joy, GS-neutral, GS-sadness) and VT parameters (VT-anger, VT-joy, VT-neutral, VT-sadness), the perceptual scores (the scores in arousal and the scores in valence) of the resynthesized emotional vowels were subjected to statistical analysis using the scores as the dependent variable, and the GS and VT parameters as the two within subject factors. For the perceptual scores in *valence*, two-way analysis of variance (ANOVA) with repeated measures indicated the significant effects of GS [$F(3, 27) = 20.377$, $p = 0.0000$] and VT [$F(3, 27) = 7.378$, $p = 0.0009$] for speaker 1, and GS [$F(3, 27) = 30.534$, $p = 0.0000$] for speaker 2. There was significant interaction between GS and VT [$F(9, 81) = 3.298$, $p = 0.0018$] for speaker 1, and [$F(9,81) = 2.638$, $p = 0.0099$] for speaker 2. For the perceptual scores in *arousal*, two-way ANOVA with repeated measures indicated the significant effect of GS [$F(3, 27) = 33.729$, $p = 0.0000$] and VT [$F(3,27) = 8.340$, $p = 0.0004$] for speaker 1, and GS [$F(3, 27) = 82.417$, $p = 0.0000$] and VT [$F(3, 27) = 27.323$, $p = 0.0000$] for speaker 2. There was no significant interaction between GS and VT for speaker 1, but significant interaction between GS and VT [$F(9, 81) = 2.350$, $p = 0.0207$] for speaker 2.

C. Discussion

The emotional vowels synthesized with the averaged GS and VT parameters (rather than their real values that varied across glottal periods) was perceptually scored to (0, 0) in the V-A space. Additionally, the perceptual scores of the synthesized emotional vowels are close to that of the original one, which validated again the effectiveness of the ARX-LF model-based analysis-by-synthesis approach (Li *et al.*, 2017). The slight difference between the synthesized emotional vowels and the original ones may be caused by using the averaged parameters of the ARX-LF model. In comparison with emotional vowels synthesized with the given VT and arbitrary GS parameters, the emotional vowels synthesized with the given GS and arbitrary VT parameters have the much smaller differences in the V-A space. This indicates the dominant effect of the GS parameters (relative to the VT parameters) in perception of emotions in vowels, and is consistent with results from many previous studies (Sun *et al.*, 2009; Sundberg *et al.*, 2011; Waaramaa *et al.*, 2010). The positions in the V-A space of the synthesized vowels with four typical emotions obtained in this experiment are similar to the results shown in Rubin and Talarico (2009) and Gangamohan *et al.* (2016). Often the GS parameters are extracted from the glottal source waveform using the inverse filtering techniques (Alku, 2011; Rothenberg, 1973).

Though the similarities of the distribution of the perceptual scores on the V-A space of four typical emotions were clearly observed for two speakers, the relative distances of emotional vowels (i.e., sadness, joy, and anger) to neutral vowels (i.e., the center of the V-A space) differed greatly. To explore possible reasons for this difference, the F_0 contour and amplitude E_e of the emotional vowels of the two speakers were examined, since these two cues are well-known to be important in emotional speech perception (Banziger and Scherer, 2005). The F_0 contours and intensity amplitudes (E_e) of the emotional vowels of the two speakers are plotted in Fig. 3. As shown by (a) and (b) in Fig. 3, the F_0 patterns (e.g., the mean F_0, F_0 range, and F_0 shape) of the emotional vowels differed greatly for the two speakers, especially for anger and joy. Compared to the neutral vowel, the largest differences in the mean F_0 and F_0 range were for joy for speaker 1 and anger for speaker 2, which correspond to the largest distances relative to the neutral vowel in the V-A space. This indicates that the mean F_0 and F_0 range can substantially account for perception of emotions in vowels, consistent with results reported by Audibert *et al.* (2005). For example, anger is characterized by a higher value of mean F_0, which might be partially caused by the heightened subglottal pressure during vowels in speech (Sundberg *et al.*, 2011). As shown by (c) and (d) in Fig. 3, the amplitude shapes of emotional vowels were different for two speakers. Relative to the neutral vowel, the largest differences in amplitude shape were for anger for two speakers. This indicated the important contribution of the intensity-related E_e parameter for emotion perception, also observed by Auberge and Cathiard (2003); Tao *et al.* (2009). Moreover, Fig. 3 indicates large differences in duration of the emotional

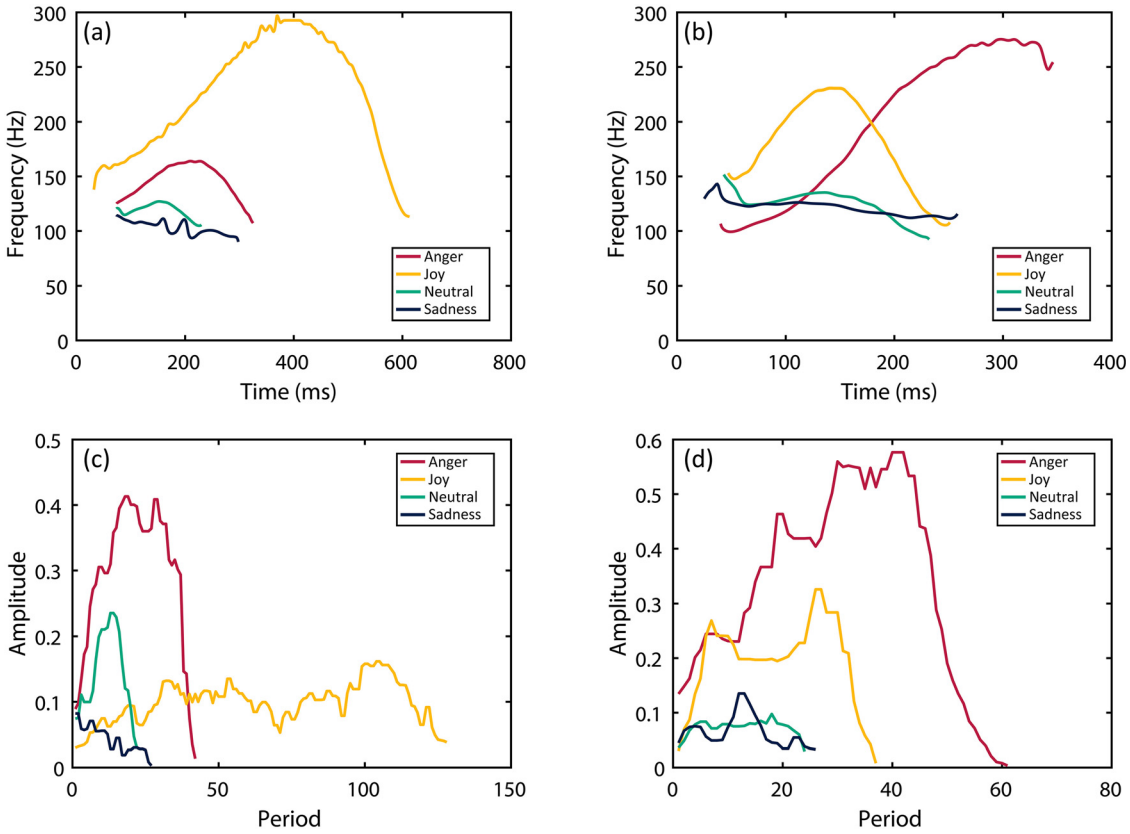

FIG. 3. (Color online) The F_0 contours of the synthesized emotional vowels for speaker 1 (a) and speaker 2 (b); the amplitude parameter E_e of the synthesized emotional vowels for speaker 1 (c) and speaker 2 (d).

vowels in this experiment, which might also affect perception of emotions in speech as suggested by Audibert *et al.* (2005) and Audibert *et al.* (2006).

IV. EXPERIMENT II: CONTRIBUTION OF GLOTTAL SOURCE AND VOCAL TRACT TO EMOTIONAL VOWEL PERCEPTION AFTER NEUTRALIZING THE F_0, INTENSITY AND DURATION CUES

The results of experiment I showed that the GS cues play more important roles in emotion perception than the VT cues. In the present study, six parameters were used to describe the glottal source wave by the LF model. As experiment I showed, F_0, E_e, and duration are important cues for emotion (especially arousal) perception, which was also reported in Auberge and Cathiard (2003), Audibert *et al.* (2005), and Audibert *et al.* (2006). A follow-up question is whether the other GS and VT cues contribute to emotion perception. And how do these cues contribute to emotion perception in the dimensions of valence and arousal. This is the task of experiment II.

A. Method

The same voiced vowel (/a/) with four different emotions (i.e., neutral, joy, anger and sadness) in experiment I was used again. The GS and VT parameters in the ARX-LF model were estimated and then averaged across all periods as in experiment I. To examine the importance of the GS parameters (except F_0 and E_e) on emotion perception in the V-A space, the T_0 (i.e., $1/F_0$) and E_e parameters in the LF model of the emotional vowels (i.e., joy, sadness, and anger) were first replaced by those of the neutral vowels, but other parameters for the LF model were preserved for each emotional state. In the synthesis process, the emotional vowels were resynthesized using the GS parameters (referred to as F_0 and E_e-neutralized GS parameters) from one emotional state and the VT parameters of another emotional state. As a result, 32 tokens of emotional vowels were resynthesized for listening tests. Replacement of the T_0 and E_e parameters removed their effects on perception of emotion, which is equivalent to normalizing duration of the synthesized emotional vowels, and this helped to isolate the contribution of the other GS cues to emotion perception.

The 10 normal-hearing listeners who participated in experiment I also participated in this experiment, and were paid for their participation. During the test, each subject listened to a total of 32 emotional vowel tokens (2 speakers $\times$ 4 sets of the glottal source parameters $\times$ 4 sets of the vocal tract parameters). Each subject was asked to give a score for each stimulus based on her/his perceptual impression in valence and arousal. The testing procedure was the same as that in experiment I.

B. Results

The mean perceptual scores of the emotional vowels synthesized with the F_0 and E_e-neutralized GS parameters and VT parameters in the V-A space across 10 subjects for two speakers are plotted in Fig. 4.

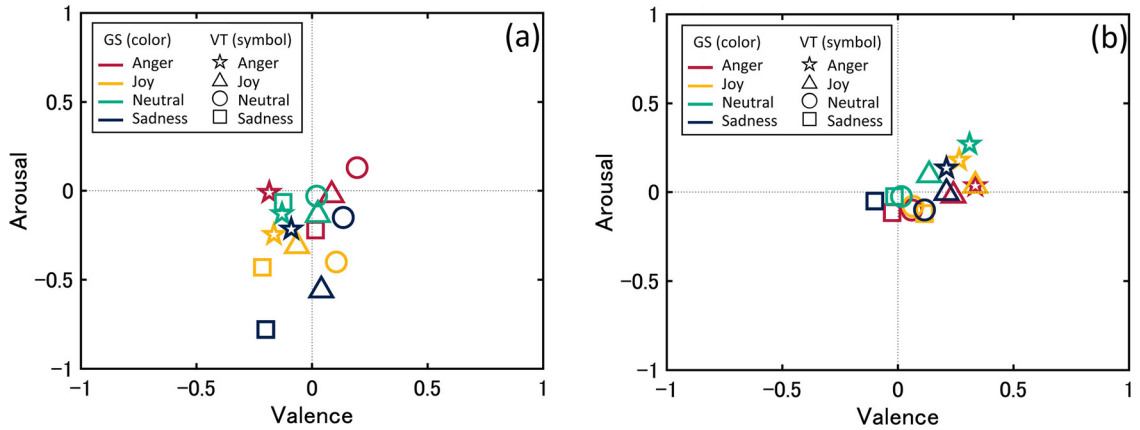

FIG. 4. (Color online) The perceptual results in the valence-arousal space of the emotional vowels synthesized with the different F_0 and E_e-neutralized GS and VT parameters for speaker 1 (a) and speaker 2 (b).

As shown in Fig. 4, for both speakers, the perceptual scores of the synthesized emotional vowels (i.e., joy, anger and sadness) moved towards the position of the synthesized neutral vowel (neutral GS and neutral VT) at the center (0, 0) in the V-A space. For speaker 1, the perceptual scores of most of the synthesized emotional vowels were negative in arousal in the V-A space. Though the perceptual scores in valence of the emotional vowels synthesized with different GS and VT parameters were centralized, those in arousal were relatively scattered. More specifically, for the emotional vowels synthesized with the given VT parameters, the perceptual scores in arousal were often the lowest for vowels with the sadness GS, and were often the highest for vowels with the anger GS. For speaker 2, it was found that the perceptual scores in valence of the synthesized emotional vowels were relatively scattered. For the emotional vowels synthesized with the given GS parameters, the perceptual scores in valence were the lowest for vowels with sadness VT, followed by those for the emotional vowels synthesized with the neutral and joy VT parameters, and the scores for the emotional vowels synthesized with the anger VT parameters were the highest in valence.

To examine the effects of GS parameter (GS-anger, GS-joy, GS-neutral, GS-sadness) and VT parameter (VT-anger, VT-joy, VT-neutral, VT-sadness), the perceptual scores (the scores in arousal and the scores in valence) of the F_0 and E_e-neutralized emotional vowels were subjected to statistical analysis using the scores as the dependent variable, and the GS and VT parameters as the two within subject factors. For the perceptual scores in *valence*, two-way ANOVA with repeated measures indicated the significant effects of VT [$F(3, 27) = 6.267$, $p = 0.0023$] for speaker 1, and GS [$F(3, 27) = 3.800$, $p = 0.0215$], and VT [$F(3, 27) = 6.005$, $p = 0.0028$] for speaker 2. There were no significant interactions between GS and VT for both speakers. For the perceptual scores in *arousal*, two-way ANOVA with repeated measures indicated the significant effects of GS [$F(3, 27) = 24.988$, $p = 0.0000$] and VT [$F(3, 27) = 8.132$, $p = 0.0005$] for speaker 1, and VT [$F(3, 27) = 15.138$, $p = 0.0000$] for speaker 2. There was a significant interaction between GS and VT [$F(9, 81) = 4.208$, $p = 0.0002$] for speaker 1, and there was no significant interaction between GS and VT [$F(9, 81) = 0.486$, $p = 0.8799$] for speaker 2.

C. Discussion

As seen in Figs. 2 and 4, the comparison of the perceptual scores in the V-A space of the synthesized emotional vowels shows movements towards those of the synthesized neutral vowels, which further indicates the important contributions of the F_0, E_e, and duration cues on perception of emotions in vowels, similar to that reported in Audibert *et al.* (2005) and Auberge and Cathiard (2003).

After removing the effects of the F_0, E_e, and duration cues, the VT information significantly contributed to valence and arousal perception for both speakers, which was consistent with the findings by Laukkanen *et al.* (1997). In contrast, the F_0 and E_e-neutralized GS cues performed differently for valence and arousal perception for the two speakers. More specifically, speaker 2 related the F_0 and E_e-neutralized GS cues to the valence of the perceived emotion, whereas speaker 1 related the F_0 and E_e-neutralized GS cues to the arousal of the perceived emotion. The different effect of the F_0-modified GS cues was also observed in Laukkanen *et al.* (1997).

To further analyze the effects of the F_0 and E_e-neutralized GS and VT cues on the perception of emotions in the V-A space, the spectral tilt of GS waveform and the first formant frequency (F_1) of VT were used to characterize the F_0 and E_e-neutralized GS and VT cues, respectively. In this present study, the spectral tilts of GS were calculated by STRAIGHT (Kawahara *et al.*, 1999) from 180- to 700-Hz, and the first formant frequencies (F_1) of VT were calculated from the ARX model.

The relationships of F_1 and the perceptual scores of the emotional vowels in valence and arousal are plotted in Fig. 5, in which the four perceptual scores corresponding to the given GS parameters and arbitrary VT parameters are linearly regressed in the sense of minimum mean square error (MMSE). As shown in Fig. 5, the great scattering of F_1 of the synthesized emotional vowels with arbitrary GS and arbitrary VT was clearly observed, as well as the relative

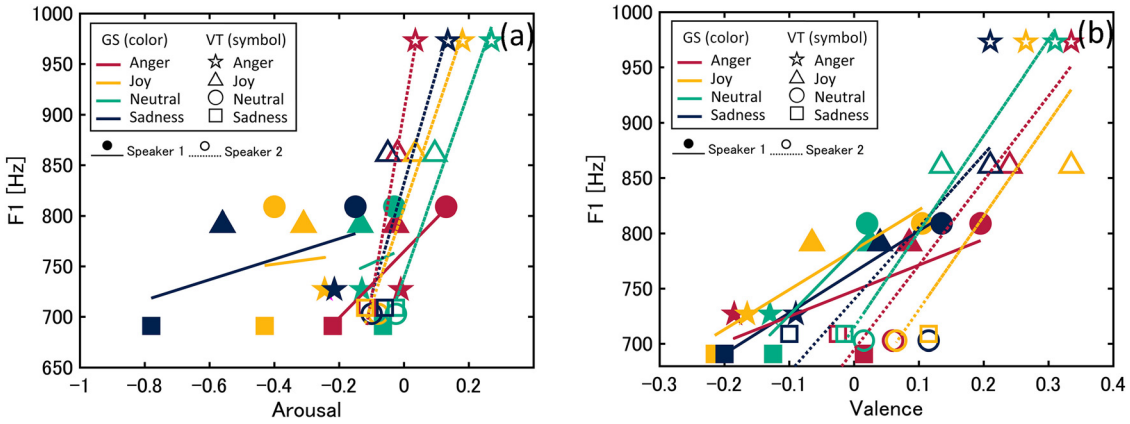

FIG. 5. (Color online) The relationship of the first formants (F_1) and the perceptual scores in arousal (a) and in valence (b), where the perceptual scores for four emotional vowels synthesized with the given GS and arbitrary VT parameters were linearly regressed in the sense of minimum mean square error (MMSE).

differences in F_1 of the emotional vowels compared to those of the neutral vowels for the two speakers. The differences in F_1 of different emotional vowels were also found in many previous studies (Erickson *et al.*, 2006; Goudbeek *et al.*, 2009). After linearly regressing the perceptual scores in the V-A space of the emotional vowels with the given GS parameters, it was found that the varying range in F_1 of the emotional vowels for speaker 1 was much smaller than that for speaker 2. This was attributed to the individualized characteristics of F_1 in emotional vowels (Erickson *et al.*, 2016). To further examine the reliability of the linear regressions in Fig. 5, their coefficients of determination (R^2) were calculated and listed in Table II. The numbers in bold, which are larger than 0.6, denote a higher coefficient of determination. From Fig. 5 and Table II, it was noted that F_1 varied proportionally to the perceptual scores in arousal for speaker 2 and in valence for two speakers. This implied that emotions with higher mean values of F_1 were usually positive in valence, which is in line with the findings in Laukkanen *et al.* (1997); Mori and Kasuya (2007). Moreover, it was observed that the slopes of the regression lines of the perceptual scores for speaker 1 were much lower than those for speaker 2, which indicates the significant differences in the degree of F_1 variation against the perceptual scores in valence and arousal for different speakers. This difference might be due to the different degree of contributions of the VT cues in the production process of the emotional speech for different speakers.

TABLE II. The coefficients determination (R^2) of the linear regressions for the relationships between the perceptual scores in the V-A space and F_1 or the spectral tilt of glottal source waveform.

	Speaker 1				Speaker 2			
	Anger	Joy	Neutral	Sadness	Anger	Joy	Neutral	Sadness
F_1-arousal	**0.74**	0.01	0.02	0.3	**0.99**	**0.97**	**0.97**	**0.79**
F_1-valence	0.46	**0.85**	**0.89**	**0.97**	**0.93**	**0.68**	**0.97**	0.54
Spectral tilt-arousal	**0.87**	**0.74**	**0.80**	**0.63**	**0.61**	**0.91**	**0.67**	0.03
Spectral tilt-valence	0.18	0.50	0.01	**0.80**	0.11	0.01	**0.63**	0.50

The relationships of the spectral tilt of the GS waveforms and the perceptual scores of the emotional vowels in arousal and valence are plotted in Fig. 6 in which the four perceptual scores corresponding to the given VT parameters and arbitrary GS parameters are linearly regressed in the MMSE sense. Figure 6 shows that the varying range of the spectral tilts of the GS waveforms for speaker 1 were largely smaller than those for speaker 2. The difference in the range of the spectral tilt of the GS waveforms could be from the different styles (e.g., individuality and degree of emotions) in the emotional vowel productions for each speaker. That individuality affects emotional speech was also pointed out by Bulut and Narayanan (2008). It is further noted that the spectral tilt of the GS waveforms varied proportionally to the perceptual scores in arousal for both speakers, as shown in Fig. 6(a). Noted also is that the slopes of the regression lines for speaker 1 are similar to those for speaker 2, which indicates that for a given amount of spectral tilt change of the GS waveform, the change of emotion perception in arousal is roughly similar. Figure 6(b) shows that no clear consistent patterns of GS waveform spectral tilts with valence perceptual scores, as is evidenced also by the low coefficients of determination in Table II. This result is similar to the findings in Laukkanen *et al.* (1997) which stated no significant relation of the GS waveform with valence perception.

Figures 5 and 6 clearly show large differences in the range of F_1 and spectral tilt of the GS waveform for the two speakers. Specifically, the ranges of F_1 for speaker 1 were much smaller than those for speaker 2, while the ranges of spectral tilt for speaker 1 were much larger than those for speaker 2. These differences implied that the speaker factor might have a significant effect on emotion perception in vowels.

V. CONCLUSION

In this paper, two experiments were carried out to examine the effects of the glottal source and vocal tract cues on emotion perception in terms of valence and arousal. Using the recently developed ARX-LF model-based analysis-by-synthesis approach, the glottal source and vocal tract

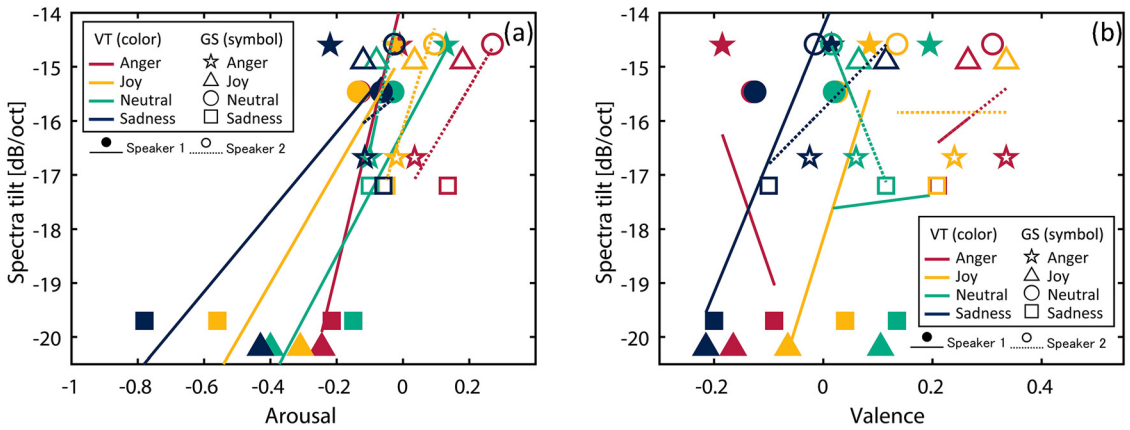

FIG. 6. (Color online) The relationship of the spectral tilt of glottal source waveform and the perceptual scores in arousal (a) and in valence (b), where the perceptual scores for four emotional vowels synthesized with the given GS and arbitrary VT parameters were linearly regressed in the sense of minimum mean square error (MMSE).

parameters were first obtained by analyzing the emotional Japanese vowels, followed by the controllable modifications, and then exploited for resynthesizing emotional vowels. The resynthsized emotional vowels were presented to native Japanese listeners with normal hearing for perceptually rating emotions according to valence and arousal. Based on these results, the following findings were obtained:

(1) The ARX-LF model-based analysis-by-synthesis approach thus helped to analyze the perceptual contributions of the GS and VT information on emotion perception. The effect of the GS information was dominant on perception of emotional vowels relative to the VT information, which was in line with previous findings (Sun *et al.*, 2009; Sundberg *et al.*, 2011; Waaramaa *et al.*, 2010). Moreover, emotions in vowels were found to be highly speaker-dependent, which was attributed in part to the large variation in emotional vowel production, such as interspeaker differences in F_0.
(2) The VT cues contributed to an emotion perception in valence and arousal, after neutralizing the effects of the F_0, E_e, and duration cues. F_1 varied proportionally to the perceptual scores in valence and arousal, and the scattering of F_1 of the emotional vowels was sizeable for each speaker, which was attributed to the individualized VT characteristics during production of emotional vowels. The positive proportionality of the spectral tilt of the GS waveform to arousal perception was observed, with no clear relation to valence perception. Moreover, the range of spectral tilt of the GS waveform was different for each speaker, most likely due to different styles of emotional speech production.
(3) Comparison of results of experiments I and II showed that the GS cues play an important role in perception of emotions in vowels, and that the emotions in the synthesized vowels vary greatly across speakers, regardless of the neutralization of the F_0, E_e, and duration parameters.

One limitation of this study is the small number of speakers (only two speakers). Though many differences in glottal source waveforms and vocal tract cues were observed for the two speakers, some common findings were also found; for example, F_0 has an important role in perception of emotions in vowels, which was consistent with the previous study reported by Erickson *et al.* (2006).

The use of voiced vowel /a/ in the present study minimized the effects of other factors (e.g., prosodic and linguistic content) in order to examine the relative contributions of the GS and VT cues to emotion perception in vowels. It is believed that this present study is useful to provide insight into emotion perception in speech. However, perception of emotions in speech is based on the combined effects of prosodic (both glottal and vocal tract information) and linguistic factors. All these factors should be taken into account in future work, allowing the acoustic parameters (e.g., F_0, duration, and energy) to be modulated in a natural fashion and to be more easily generalizable to real-life utterances.

Also, this study focused on acoustic and perceptual characteristics of emotional vowels in Japanese. Since culture and language play important roles in emotional expression, future work is necessary to examine different language scenarios with respect to production and perception of emotional speech.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Donna Erickson and Dr. Aijun Li for their valuable comments on our study. This study was supported by a Grant-in-Aid for Scientific Research (A) (No. 25240026) and China Scholarship Council (CSC).

Airas, M., and Alku, P. (**2006**). "Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalised amplitude quotient," Phonetica **63**(1), 26–46.
Alku, P. (**2011**). "Glottal inverse filtering analysis of human voice production-a review of estimation and parameterization methods of the glottal excitation and their applications," Sadhana **36**(5), 623–650.
Auberge, V., and Cathiard, M. (**2003**). "Can we hear the prosody of smile?," Speech Commun. **40**(1), 87–97.
Audibert, N., Auberge, V., and Rilliard, A. (**2005**). "The prosodic dimensions of emotion in speech: The relative weights of parameters," in *INTERSPEECH*, Lisbon, Portugal, pp. 525–528.
Audibert, N., Vincent, D., Auberge, V., and Rosec, O. (**2006**). "Expressive speech synthesis: Evaluation of a voice quality centered coder on the different acoustic dimensions," in *Proc. Speech Prosody*, Dresden, Germany, Vol. 2006, pp. 525–528.

Banziger, T., and Scherer, K. R. (**2005**). "The role of intonation in emotional expressions," Speech Commun. **46**(3), 252–267.
Bulut, M., and Narayanan, S. (**2008**). "On the robustness of overall F_0-only modifications to the perception of emotions in speech," J. Acoust. Soc. Am. **123**(6), 4547–4558.
Degottex, G., Bianco, E., and Rodet, X. (**2008**). "Usual to particular phonatory situations studied with high-speed videoendoscopy," in *International Conference on Voice Physiology and Biomechanics*, Tampere, Finland, pp. 19–26.
Drugman, T., Bozkurt, B., and Dutoit, T. (**2012b**). "A comparative study of glottal source estimation techniques," Comput. Speech Lang. **26**(1), 20–34.
Drugman, T., Thomas, M., Gudnason, J., Naylor, P., and Dutoit, T. (**2012a**). "Detection of glottal closure instants from speech signals: A quantitative review," IEEE Trans. Audio, Speech, Lang. Process. **20**(3), 994–1006.
Elbarougy, R., and Akagi, M. (**2014**). "Improving speech emotion dimensions estimation using a three-layer model of human perception," Acoust. Sci. Technol. **35**(2), 86–98.
Erickson, D. (**2004**). "Acoustic and articulator analysis of sad Japanese speech," in *Proc. Fall Meet. Phonet. Soc. Jpn.*, 2004, Tokyo, Japan.
Erickson, D., Shochi, T., Menezes, C., Kawahara, H., and Sakakibara, K.-I. (**2008**). "Some non-F_0 cues to emotional speech: An experiment with morphing," in *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil, pp. 677–680.
Erickson, D., Yoshida, K., Menezes, C., Fujino, A., Mochida, T., and Shibuya, Y. (**2006**). "Exploratory study of some acoustic and articulatory characteristics of sad speech," Phonetica **63**(1), 1–25.
Erickson, D., Zhu, C., Kawahara, S., and Suemitsu, A. (**2016**). "Articulation, acoustics and perception of mandarin Chinese emotional speech," Open Linguist. **2**(1), 620–635.
Fant, G., Liljencrants, J., and Lin, Q.-g. (**1985**). "A four-parameter model of glottal flow," STL-QPSR **26**(4), 1–13.
Fu, Q., and Murphy, P. (**2006**). "Robust glottal source estimation based on joint source-filter model optimization," IEEE Trans. Audio, Speech, Lang. Process. **14**(2), 492–501.
Gangamohan, P., Kadiri, S. R., and Yegnanarayana, B. (**2016**). "Analysis of emotional speech-a review," in *Toward Robotic Socially Believable Behaving Systems-Volume I* (Springer, Cham), pp. 205–238.
Goudbeek, M., Goldman, J. P., and Scherer, K. R. (**2009**). "Emotion dimensions and formant position," in *Tenth Annual Conference of the International Speech Communication Association*, Brighton, UK, 6–10 September.
Iliev, A. I., Scordilis, M. S., Papa, J. P., and Falcao, A. X. (**2010**). "Spoken emotion recognition through optimum-path forest classification using glottal features," Comput. Speech Lang. **24**(3), 445–460.
Juslin, P. N., and Laukka, P. (**2001**). "Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion," Emotion **1**(4), 381–412.
Juslin, P. N., and Scherer, K. R. (**2005**). "Vocal expression of affect," in *The New Handbook of Methods in Nonverbal Behavior Research* (Oxford University Press, Oxford, UK), pp. 65–135.
Kawahara, H., Masuda-Katsuse, I., and De Cheveigne, A. (**1999**). "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based *F*0 extraction: Possible role of a repetitive structure in sounds," Speech Commun. **27**(3), 187–207.
Kim, J., Toutios, A., Kim, Y.-C., Zhu, Y., Lee, S., and Narayanan, S. (**2014**). "Usc-emo-mri corpus: An emotional speech production database recorded by real-time magnetic resonance imaging," in *International Seminar on Speech Production (ISSP)*, Cologne, Germany.
Laukkanen, A.-M., Vilkman, E., Alku, P., and Oksanen, H. (**1997**). "On the perception of emotions in speech: The role of voice quality," Logoped. Phoniatr. Vocol. **22**(4), 157–168.
Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., and Narayanan, S. (**2004**). "Emotion recognition based on phoneme classes," in *Eighth International Conference on Spoken Language Processing*, Jeju Island, South Korea, 4–8 October.
Lee, S., Bresch, E., Adams, J., Kazemzadeh, A., and Narayanan, S. (**2006**). "A study of emotional speech articulation using a fast magnetic resonance imaging technique," in *INTERSPEECH*, Pittsburgh, PA, pp. 1792–1795.
Leinonen, L., Hiltunen, T., Linnankoski, I., and Laakso, M.-L. (**1997**). "Expression of emotional–motivational connotations with a one-word utterance," J. Acoust. Soc. Am. **102**(3), 1853–1863.
Li, A., Fang, Q., Hu, F., Zheng, L., Wang, H., and Dang, J. (**2010**). "Acoustic and articulatory analysis on mandarin Chinese vowels in emotional speech," in *2010 7th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, Tainan, Taiwan, pp. 38–43.
Li, Y., Sakakibara, K.-I., Morikawa, D., and Akagi, M. (**2017**). "Commonalities of glottal sources and vocal tract shapes among speakers in emotional speech," in *International Seminar on Speech Production (ISSP)*, Tianjing, China, pp. 79–81.
Lu, H.-L. (**2002**). "Toward a high-quality singing synthesizer with vocal texture control," Ph.D. dissertation, Stanford University, Stanford, CA.
Mori, H., and Kasuya, H. (**2007**). "Voice source and vocal tract variations as cues to emotional states perceived from expressive conversational speech," in *INTERSPEECH*, Antwerp, Belgium, pp. 27–31.
Ohtsuka, T., and Kasuya, H. (**2002**). "Robust ARX-based speech analysis method taking voicing source pulse train into account," Acoustical Soc. Jpn. **58**(7), 386–397.
Ringeval, F., and Chetouani, M. (**2008**). "A vowel based approach for acted emotion recognition," in *Ninth Annual Conference of the International Speech Communication Association*, Brisbane, Australia, 22–26 September.
Rothenberg, M. (**1973**). "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," J. Acoust. Soc. Am. **53**(6), 1632–1645.
Rubin, D. C., and Talarico, J. M. (**2009**). "A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words," Memory **17**(8), 802–808.
Sauter, D. A., Eisner, F., Calder, A. J., and Scott, S. K. (**2010**). "Perceptual cues in nonverbal vocal expressions of emotion," Q. J. Exp. Psychol. **63**(11), 2251–2272.
Scherer, K. R. (**1986**). "Vocal affect expression: A review and a model for future research," Psychol. Bull. **99**(2), 143–165.
Scherer, K. R. (**2003**). "Vocal communication of emotion: A review of research paradigms," Speech Commun. **40**(1), 227–256.
Sun, R., Moore, E., and Torres, J. F. (**2009**). "Investigating glottal parameters for differentiating emotional categories with similar prosodics," in *ICASSP*, IEEE, Taibei, Taiwan, pp. 4509–4512.
Sundberg, J., Patel, S., Bjorkner, E., and Scherer, K. R. (**2011**). "Interdependencies among voice source parameters in emotional speech," IEEE Trans. Affective Comput. **2**(3), 162–174.
Tao, J., Li, Y., and Pan, S. (**2009**). "A multiple perception model on emotional speech," in *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009, ACII 2009* (IEEE, New York), pp. 1–6.
Vincent, D., Rosec, O., and Chonavel, T. (**2005**). "Estimation of lf glottal source parameters based on an ARX model," in *INTERSPEECH*, Lisbon, Portugal, pp. 333–336.
Waaramaa, T., Laukkanen, A.-M., Airas, M., and Alku, P. (**2010**). "Perception of emotional valences and activity levels from vowel segments of continuous speech," J. Voice **24**(1), 30–38.
Yanushevskaya, I., Ni Chasaide, A., and Gobl, C. (**2009**). "Voice parameter dynamics in portrayed emotions," in *Models and analysis of vocal emissions for biomedical applications: 6th International Workshop* (Firenze University Press, Firenze, Italy), pp. 1000–1004.
Yanushevskaya, I., Tooher, M., Gobl, C., and Chasaide, A. N. C. (**2007**). "Time- and amplitude-based voice source correlates of emotional portrayals," in *International Conference on Affective Computing and Intelligent Interaction, ACII2007: Lecture Notes in Computer Science* (Springer, Berlin), pp. 159–170.