2014 RISP International Workshop on Nonlinear Circuits,
Communications and Signal Processing
NCSP'14, Honolulu, Hawaii, USA. February 28 - March 3, 2014

Glottal source analysis of emotional speech

Yongwei Li, Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, Japan
Phone/FAX:+81-080-4250-9518
E-mail: yongwei@jaist.ac.jp, akagi@jaist.ac.jp

Abstract

Emotional speech makes speech more expressive, emotional speech conversion is needed in many systerms. Analyzing of glottal source wave plays an important role in emotional speech conversion. The purpose of this paper is to analyze glottal source of emotional speech for emotional speech conversion based on the Auto-Regressive eXogenous (ARX) model combined with Liljencrant-fant (LF) model, in which the Glottal Clsure Instant (GCI) and Glottal Opening Instant (GOI) are two important parameters and greatly affect the accuracy of the ARX-LF model.

Therefore, a mean-based signal method is suggested to improve the estimation accuracy of GCI, and GOI is estimated from the Hilbert envelope of LP residual. The ARX-LF model with accurate GCI and GOI is applied for analysis of glottal source of emotional speech. The results show that the proposed approach improve the accuracy of glottal source wave of speech, and the different glottal source waves of different emotional speech can be obtained.

1. Introduction

Human speech consists of not only linguistic information but also non-linguistic information (i.e., emotion), which makes speech more expressive. Emotional speech conversion is useful for many systerms. Many approaches were mainly based on transformation from one acoustic feature set to another acoustic feature set. Those methods sometimes fail due to the fact the signal representation does not really fit with speech production mechanisms. Glottal source plays an important role in emotional speech production mechanisms. Thus, analysis of glottal source wave of emotional speech is important for emotional speech conversion.

Recently, many glottal source models for estimating the parameters from speech signals have been proposed such as Rosenberg-Klatt (RK model) and Liljencrant-Fant (LF) model [1, 2]. However, the parameters of these models were mainly estimated manually. The manual procedure causes a serious problem when analyzing a large number of speech data. To solve this problem, an Auto-Regressive eXogenous (ARX) model combined with the RK model was proposed by Ding [3]. It is well known that the information of frequency domain is important in speech signal processing. Michael *et al.* shows that the return phase of glottal source has strong relationship with frequency energy [4]. The shorter return phase is, the larger frequency energy is. However, the return phase cannot be represented by the RK model. The return phase can be obtained by LF model. Thus, an ARX model combined with the LF model was proposed by Vincent *et al.* [5]. The ARX-LF model is widely used for estimating glottal source parameters in recent years. For example, the glottal source parameters of singing signals is estimated by Motoda and Akagi based on the ARX-LF model [6].

In case of estimating glottal source wave of emotional speech. The parameters of glottal opening instant (GOI) are used as the start point of each period of LF model and glottal close instant (GCI) are used as the initial parameters of each period of LF model. Since the period of joy and anger speech is very short, enven small mistake may causes big error of analyzed results. However, the parameters of GOI and GCI are undervalued in the current ARX-LF model-based methods.

In this paper, the ARX-LF model with accurate GCI and GOI is used for estimating glottal source wave of emotional speech.

2. Speech production model

2.1 LF model

The LF model is used for representing the glottal flow derivative. The LF model has five parameters, including phase of maximum open of glottis (T_p), open phase of glottis (T_e), return phase (T_a), amplitude of GCI (E_e) and period (T_0). A typical LF waveform is depicted on Figure 1. The explicit expression of the model for one fundamental period is give by Eq.(1), in which the parameters a, b and w are implicitly related to T_p, T_e, T_a.

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

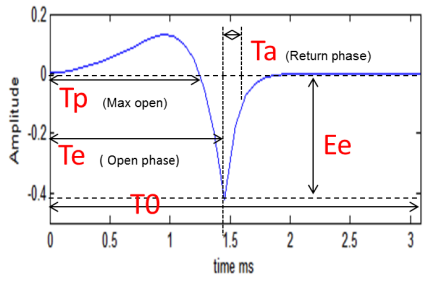

Figure 1: LF model that functionally mimics derivative of glottal flow

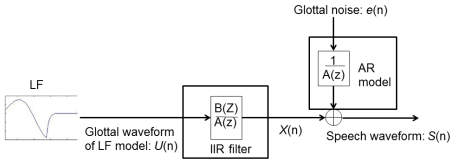

Figure 2: ARX model

2.2 ARX model

The ARX model can simulate vocal tract. Speech production process can be modeled as a time-varying IIR system as follows:

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

where $s(n)$ is observed speech signal and $u(n)$ is the glottal waveform (The LF waveform) at time n, a_i and b_0 are time-varying coefficients of filter, p is filter order, and $e(n)$ is the equation error.

The output signal of the LF model acts as an input signal $u(n)$ to the vocal tract filter. Eq.(2) is called ARX model, as illustrated in Figure. 2. Input signal $u(n)$ to the IIR filter is the glottal waveform, which is approximated by the LF model. We use the ARX model to represent the vocal tract filter. In this model, $e(n)$ as a glottal noise in the speech production and its power in the voiced sound is obtained from the equation error. The vocal tract transfer function is defined as follows:

$$H(z) = \frac{B(z)}{A(z)} = \frac{1}{1 + a_1 z^{-1} + ... + a_p z^{-p}} \qquad (3)$$

3. Estimation of GCI and GOI

GCI and GOI are important parameters for estimating the glottal source wave of emotional speech. Since the period of joy and anger speech is very short, enven small mistake may causes big error of analyzed results. The mean-based signal method is selected for estimating GCI parameter [7], and GOI parameter is estimated from the Hilbert envelop of LP residual [8].

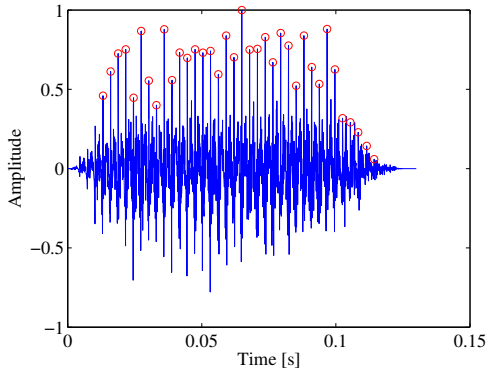

Figure 3: LP residual

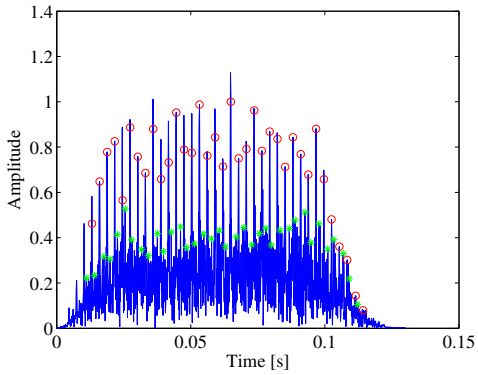

Figure 4: Hilbert envelope of LP residual

3.1 GCI detection

The procedure of estimation of GCI is divided into two successive steps. In the first step, a mean-based signal is calculated and the region of GCI is detected from it. The region of GCI is located between the minimum and the positive zero-crossing of the mean-based signal. The second step is to estimate the position of GCI. The GCI position corresponds to the strongest peak of the Linear Prediction (LP) residual within the region of GCI. GCI is estimated by picking peak of LP residual. The results is shown in Figure 3.

3.2 GOI detection

After the GCI position is determined, the GOI detection is getting to be easy and its region is located between two successive GCI. GOI are estimated by picking biggest peaks

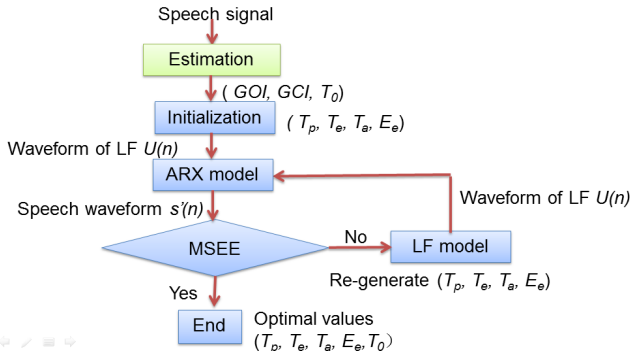

Figure 5: Flowchart of ARX-LF model with accurate GCI and GOI

form Hilbert envelope of LP residual. The results is plotted in Figure 4.

4. The scheme of analysis

GCI and GOI are estimated as described in the previous section. The period (T_0) is the time duration between two consecutive GCIs. The scheme of analysis is shown in Fig. 5. We set some initial values for glottal sources value in one period, and waveform of LF model is constructed which can be used in the ARX process. The Kalman filter is used to estimate the time-varying coefficients of vocal tract filter. The re-synthesized speech waveform is estimated after the ARX process. The next step is to go to LF model again and a small random glottal source values instead of initial value in LF process, then go to ARX processed again. The optimal glottal source parameters is estimated by searching the smallest values of mean square equation error (MSEE) between original speech waveform and re-synthesized speech waveform.

5. Evaluation

5.1 Database

Speech sentence is selected from the voice database of Fujitsu Lab. A vowel /a/ is selected for analysis. The sampling frequency is 12 kHz.

5.2 Results and discussion

The results are plotted in Figure 6. The Figure 6 (a) is the waveform of LF model, in which we estimated the optimal values of T_p, T_e, T_a and E_e. The Figure 6 (b) is the vocal tract information, in which the peaks are formant frequency.

In order to evaluate our method, the LP spectrum envelope method to estimate the formant frequency is used, as shown in the Figure 6 (c), in which the peaks are formant frequency.

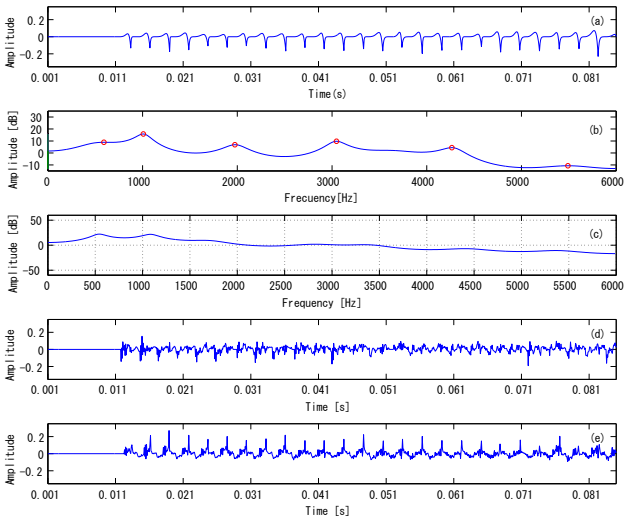

Figure 6: Example of netural speech: (a) the waveform of LF model, (b) the vocal tract imformation, (c) the LP spectrum envelope, (d) the error of our result, (e) the error of Motoda's result.

The results show that the formant frequency of LP spectrum envelope method similar with our result.

The error between the estimated signal and the orginal signal is calculated, and the error was compared with the method of estimation of GCI parameter and GOI parameter in different approach such as Motoda's approach. The error of our approach and the error of previous approach in neutral speech were plotted in the Figure 6 (d) and the Figure 6 (e) respectively, the results show that the error of our approach is smaller than previous method. Thus, our approach is more suitable for estimating glottal source wave of speech.

The mean-based signal method estimating GCI and the Hilbert envelope of LP residual method for estimating GOI can improve the accuracy of output of ARX-LF model. The next section is to use ARX-LF model with accurate GCI and GOI to analysis glottal source wave of emotional speech.

6. Emotional speech analysis

6.1 Database

Emotional speech signals were selected from the voice database produced by Fujitsu Lab. The utterance is:

- I ra n/a/ i me-ru ga a ta ra su te te ku da sa i

in which the phoneme /a/ is selected from different emotional speech sentence including neutral, joy, anger and sad. The sampling frequency is 12 kHz.

Table 1: Results of experiment

	Tp	Te	Ta	Ee	T0
Neutral	1.39 ms	1.64 ms	0.084 ms	0.1322	3 ms
Joy	1.16 ms	1.4 ms	0.0375 ms	0.28	2.16 ms
Sad	2.66 ms	3.16 ms	0.229 ms	0.069	4.58 ms
Anger	0.804 ms	0.947 ms	0.0738 ms	0.2563	2.58 ms

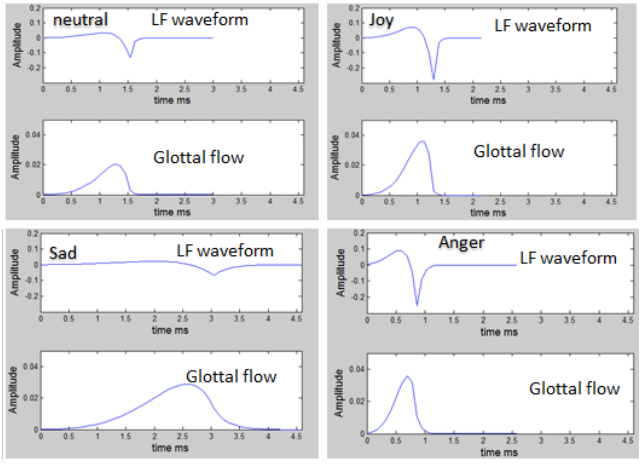

Figure 7: Glottal flow and its derivative of emotional speech

6.2 Results and discussion

Analyzed results are plotted in Figure 7.

The figure 7 contains four panels; neutral, joy, anger and sad. Each panel includes two part: the top one is glottal flow derivative, and the bottom one is glottal flow.

For the neutral speech, the result shows that the shape of glottal flow is similar to a sine wave. This is consistent with human speech production. When the airflow from lungs arrives at glottal fold, the glottal fold opens with a normal speed. Therefore, the glottal flow looks like a sine shape. For joy speech and anger speech, the shape of glottal flow looks like triangular wave. This result fits with human speech production mechanism. As human is in the excited state when producing joy and anager speech, the position under glottis has high pressure, the glottal folds vibrate more quickly. Therefore, the shape is like triangular wave. For sad speech the shape of glottal flow is more smooth because air flow arrives at glottal folds with a relatively slow speed which makes the glottal folds open slowly when human is in sad state.

The detailed values of parameter are shown in table 1. The obtained values for these parameter are of different emotional speech. A special parameter T_a is related to frequency energy. The smaller T_a is, the higher frequency energy is. The results show that the high frequency energy of joy and anger speech is more than neutral speech, and much more than sad speech.

7. Conclusion

In this paper, the glottal source wave of emotional speech is analyzed by using ARX-LF model with accurate GCI and GOI approach. The different glottal source waves of different emotional speech can be obtained with the approach.

References

[1] D. Klatt and L. Klatt, " Analysis, synthesis, and perception of voice quality varations among female and male talkers," J. Acoust. Soc. Am, vol. 87. pp. 820-857, 1990.

[2] G. Fan, J. Liljencrants, and Q. Lin, "A four-parameter model model of glottal flow," STL-QPSR4, pp. 1-13, 1986.

[3] D. Weng, H. Kasuya and S. Adachi, "Simultaneous estimation of vocal tract and voice source parameters based on an ARX model," IEICE Trans.Inf. Syst., Vol. E78-D, NO.6, pp. 738-743, June 1995.

[4] P. Michael D., Q. Thomas F. and R. Douglas A. " Modeling of the glottal flow derivative waveform with application to speaker identification," IEEE Transactions on speech and audio processing, Vol. 7, No.5, September 1999.

[5] D. Vincent, O. Rosec and T. Chonavel, "Estimation of LF model glottal source parameters based on arx model," INTERSPEECH, pp. 333-336, 2005

[6] H. Motoda and M. Akagi, "A singing voice synthesis system to characterize vocal registers using ARX-LF model," Proceedings of NCSP 2013, USA, pp. 93-96, 2013.

[7] T. Drugman, T. Dutoit , " Glottal closure and opening instant detection from speech signals," in Proc. INTERSPEECH, 2009.

[8] K. Ramesh, S. R. M. Prasana, D. Govind, " Detection of glottal opening instants using Hilbert envelope," INTERSPEECH, pp. 25-29, 2013.