Analysis of voice source characteristics using a constrained polynomial model

Tokihiko Kaburagi [†‡] *and Koji Kawai* [†]

[†]Department of Acoustic Design
Kyushu Institute of Design, Fukuoka, Japan
[‡]CREST, Japan Science and Technology Corporation
kabu@kyushu-id.ac.jp

Abstract

This paper presents an analysis method of voice source characteristics from speech by simultaneously employing models of the vocal tract and voice source signal. The vocal tract is represented as a linear filter based on the conventional all-pole assumption. On the other hand, the voice source signal is represented by linearly overlapping multiple number of base signals obtained from a generalization of the Rosenberg model. The resulting voice source model is a polynomial function of time and has lesser degrees-of-freedom than the polynomial order. By virtue of the linearity of both models, the optimal values of their parameters can be jointly determined when the instants of the glottal opening and closing are given for each pitch period. We also present a temporal search method of these glottal events using the dynamic programming technique. Finally, experimental results are presented to reveal the applicability of the proposed method for several phonation conditions.

1. Introduction

Analysis method of the voice source characteristics would be useful to study the voice quality of speech in relation to the phonation types, laryngeal settings, or variation among different speakers [1, 2, 3]. Several analysis methods have been studied in which models of the voice source signal were used as well as that of the vocal-tract and their parameters were jointly estimated for input speech samples [4, 5, 6, 7]. However, due to the nonlinearity of model parameters on the configuration of the voice source signal, complicated estimation problems were required to be solved.

This paper presents a novel analysis method especially designed for the investigation of the voice source dynamics. The speech production process is represented using the conventional source-filter model where the vocal tract filter is driven by an overlapped voice source model. The base signal of this source model is constructed by expanding the definition of the Rosenberg model [1]. The resulting voice source model is a polynomial function of time where the degrees-of-freedom of the polynomial coefficients are constrained. By virtue of the linearity of both models, the optimal values of the model parameters are uniquely determined if the instants of the glottal opening and closing are given. Therefore, the problem to be solved results in the temporal estimation of these glottal events. Thus our method can partly avoid the complex nonlinearities existed in the methods studied thus far. We also show a computationally-efficient and effective procedure for searching the instants of the glottal events by employing the dynamic programming technique.

This paper is organized as follows. Section 2 describes the vocal-tract and voice-source models and how their parameter values are determined is shown in Section 3. In Sections 4 and 5, an analysis procedure is presented as well as the temporal search method of the glottal events. Section 6 shows experimental results and Section 7 summarizes this work.

2. Speech production model

To study the voice source characteristics, nonnasalized voiced speech is only considered as the object of the analysis. Then the speech production process can be modeled by linearly concatenating the driving signal in the form of the glottal volume velocity, linear filter representing the transfer function of the vocal tract, and radiation impedance at the mouth. Since the acoustic effect of the radiation impedance can be approximated as a high pass filter of about +6dB/oct. slope, the driving signal is treated as the differentiated air flow by combining it with the radiation impedance.

Based on the all-pole approximation of the vocal tract, speech signal is represented as

$$\tilde{s}[n] = -\sum_{i=1}^{p} a_i s[n-i] + g[n] \tag{1}$$

where $s[n]$ and $\tilde{s}[n]$ respectively represent the actual and predicted samples of speech. a_i $(1 \le i \le p)$ is the filter coefficient and p is the filter order. $g[n]$ represents the driving signal (differentiated glottal air flow) defined as the linear combination of piecewise base signals such that

$$g[n] = \sum_{j=1}^{r} \sum_{k=1}^{m} b_{jk} w_{jk}[n] \tag{2}$$

where r is the number of pitch periods within the analysis frame and m is that of overlapped base signals in each period.

The base signal $w_{jk}[n]$ $(1 \le k \le m)$ for the jth pitch period should be locally supported and take a nonzero value only in the time interval $n_{j1} \le n \le n_{j2}$ such that

$$w_{jk}[n] = 0 \quad (n < n_{j1}, n_{j2} < n)\,.$$

They are summed up with the weighting coefficients b_{jk} to form the driving signal. From the definition, n_{j1} and n_{j2} respectively correspond to the instants of the glottal opening and closing in the jth pitch period. In addition, we assume that each base signal can be shaped as a function of these two time points. Such a signal model could be the Rosenberg model [1] or the Milenkovic's polynomial model [8].

2.1. Constrained polynomial model of the voice source signal

The base signal $w_{jk}[n]$ in Eq. (2) is constructed by sampling the following function of time defined for the interval $0 \leq t \leq 1$:

$$g(t) = (2+\alpha)t^{1+\alpha} - (3+\alpha)t^{2+\alpha} \tag{3}$$

where α is the parameter specifying the polynomial order. When α is zero, it corresponds to the Rosenberg model [1] except for the gain factor. It takes $g(0) = 0$ and $g(1) = -1$ at the both boundaries and satisfies a condition $\int_0^1 g(t)dt = 0$.

When the instants of the glottal opening and closing events (n_{j1} and n_{j2}) are given, $g(t)$ is sampled using the sampling interval of speech so that $t = 0$ and 1 respectively correspond to n_{j1} and n_{j2} to form the base signal. In addition, $g[n]$ is constructed as in Eq. (2) by adding multiple number of base signals having different polynomial order. If the maximum value of the order parameter in this summation is α_{max}, the resulting voice source signal ($g[n]$) is the sampled version of a polynomial function with the order of $\alpha_{max} + 2$. However, the number of the summation (m), i.e., degrees-of-freedom of the model, is generally smaller than that order. Our model then can be seen as a constrained polynomial representation of the voice source signal.

In the summation of the base signals, the weighting coefficients b_{jk} must satisfy the following inequality constraint

$$\sum_{k=1}^{m} b_{jk} \geq 0 \tag{4}$$

so that a negative peak of the voice source signal is represented when the glottis is closed. This constraint is directly derived from the property ($g(1) = -1$) of the generalized Rosenberg model described above. The voice source characteristics of speech under various phonation conditions could be represented in the proposed model by means of the weighted interpolation of base signals having different frequency characteristics with each other.

3. Joint estimation of the model parameters

This section explains how the parameters of the speech production model in Eqs. (1) and (2) are determined simultaneously when the instants of the glottal events are given. Suppose that speech samples are given as $s[n]$ $(0 \leq n \leq N-1)$ including several pitch periods where N represents the number of samples. Let $e[n]$ denotes the signal prediction error as

$$e[n] = s[n] - \tilde{s}[n] \ .$$

Then the vector representation of the prediction error can be written as

$$\mathbf{e} = \mathbf{s}_0 + \mathbf{q}G^T$$

where $\mathbf{e}$, $\mathbf{s}_k$, and $\mathbf{q}$ are vectors respectively representing the prediction errors

$$\mathbf{e} = (e[p], e[p+1], \cdots, e[N-1]) \ ,$$

speech samples

$$\mathbf{s}_k = (s[p-k], s[p-k+1], \cdots, s[N-k-1]) \ ,$$

and the model parameters

$$\mathbf{q} = (a_1, \cdots, a_p, -b_{11}, \cdots, -b_{1m}, \cdots, -b_{r1}, \cdots, -b_{rm}) \ .$$

T denotes the transposition. G is a matrix storing samples of input speech and base signals of the voice source model as

$$\begin{aligned} G \quad = \quad & [(\mathbf{s}_1)^T, \cdots, (\mathbf{s}_p)^T, (\mathbf{w}_{11})^T, \cdots, (\mathbf{w}_{1m})^T, \\ & (\mathbf{w}_{21})^T, \cdots, (\mathbf{w}_{2m})^T, \cdots, (\mathbf{w}_{r1})^T, \cdots, (\mathbf{w}_{rm})^T] \end{aligned}$$

where $\mathbf{w}_{jk}$ is the vector for the kth base signal in the jth pitch period as

$$\mathbf{w}_{jk} = (w_{jk}[p], w_{jk}[p+1], \cdots, w_{jk}[N-1]) \ .$$

The optimal values of the model parameters are determined so that the total squared error

$$E = \mathbf{e}\mathbf{e}^T$$

is minimized. This problem results in the simultaneous linear equations with respect to the unknown parameters as

$$\mathbf{q}G^TG = -\mathbf{s}_0G$$

and it can be solved explicitly to determine the properties of the vocal tract and voice source. It might be noted that the vocal tract is supposed to be stable within the analysis frame while period-to-period fluctuations of the voice source are considered both in the temporal and spatial domains to represent the jitter or shimmer during the phonation.

4. Temporal determination of the glottal events

It has been shown in the previous sections that each speech sample is predicted as the linear combination of the preceding samples and explicit driving signal. The base signal of the voice source model is constructed by specifying the opening and closing instants of the glottis, and then the values of the model parameters are determined so that the prediction error of the speech samples is minimized. Therefore, the problem to be solved clearly results in the optimal temporal assignment of the glottal events. To solve the problem, an iterative procedure is employed here in which the temporal estimation of the glottal events and the joint estimation of the model parameters are alternately performed.

Suppose that the instants n_{j1} and n_{j2} are initially guessed for each pitch period. Then it is convenient to define the analysis frame in synchronization with them. As shown in Fig. 1, both ends of the analysis frame are set at the instants of the glottal opening with the interval of L pitch periods and they are denoted as n_{01} and n_{L1}. In the temporal estimation of the glottal events, these end points are fixed and the positions of the remaining $2L-1$ time points are searched more precisely around the initial estimates. In the following, the instants $\{n_{01}, n_{02}, ..., n_{L1}\}$ are rewritten as $\{n_0, n_1, ..., n_{2L}\}$ for simplicity.

4.1. Definition of the segmental error

The instants $\{n_1, n_2, ..., n_{2L-1}\}$ are determined by fitting the voice source model to the residual signal obtained by inversely filtering the input speech. The squared error within the analysis frame can be defined as

$$E = \sum_{l=1}^{2L} E_l(n_{l-1}, n_l)$$

where E_l is the following segmental error:

$$E_l(n_{l-1}, n_l) = \sum_{n=n_{l-1}}^{n_l-1} (x[n] - g[n])^2 \ .$$

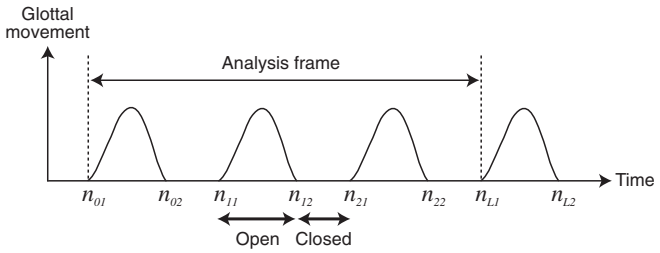

Figure 1: *Illustration of the pitch-synchronized analysis frame defined between the opening instants of the glottal movement.*

$x[n]$ is the residual signal computed as

$$x[n] = \sum_{i=0}^{p} a_i s[n-i] \quad (a_0 = 1)$$

where a_i is the coefficient of the vocal-tract filter obtained by solving the joint estimation problem. $g[n]$ is the voice source signal represented as

$$g[n] = \begin{cases} \sum_{k=1}^{m} b_k w_k[n] & \text{(glottis open)} \\ 0 & \text{(glottis closed)} \end{cases}$$

where b_k is the unknown weighting parameter. The segmental error can be explicitly written in terms of the residual signal as

$$E_l(n_{l-1}, n_l) = \mathbf{e}_l \mathbf{e}_l^T$$

when the glottis is open and

$$E_l(n_{l-1}, n_l) = \mathbf{x}_l \mathbf{x}_l^T$$

when the glottis is closed, where

$$\mathbf{x}_l = (x[n_{l-1}], x[n_{l-1}+1], ..., x[n_l - 1])$$

is the vector storing the samples of the residual signal,

$$\mathbf{e}_l = \mathbf{x}_l \{I - G(G^T G)^{-1} G^T\}$$

is the error vector when the residual signal is approximated by the constrained voice source model,

$$G = [(\mathbf{w}_1)^T, (\mathbf{w}_2)^T, \cdots, (\mathbf{w}_m)^T]$$

is the matrix storing the base signals of the voice source model, and I is the unit matrix.

4.2. Dynamic programming solution

The instants $\{n_1, n_2, ..., n_{2L-1}\}$ are determined so that the overall error E described above is minimized. Here we apply the dynamic programming technique to reduce the computational cost in searching the optimal combination of the temporal alignments of the glottal events. The accumulated error along the search path is first computed as

$$\begin{aligned} D_1(n_1) &= E_1(n_0, n_1) \\ D_l(n_l) &= \min_{n_{l-1}} D_{l-1}(n_{l-1}) + E_l(n_{l-1}, n_l) \\ &\qquad \text{for} \quad l = 2, 3, ..., 2L-2 \\ D_{2L-1}(n_{2L-1}) &= \min_{n_{2L-2}} D_{2L-2}(n_{2L-2}) + \\ &\quad E_{2L-1}(n_{2L-2}, n_{2L-1}) + E_{2L}(n_{2L-1}, n_{2L}) , \end{aligned}$$

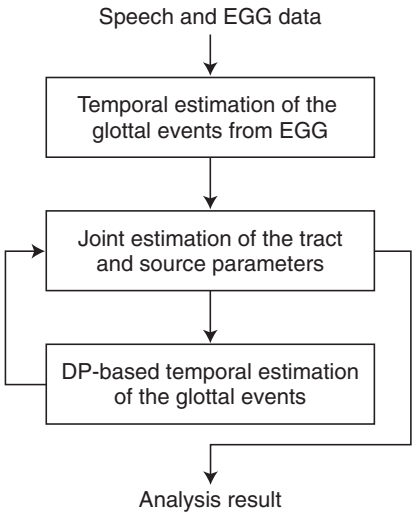

Figure 2: *The procedure for jointly determining the optimal values of the vocal-tract and voice-source model parameters for simultaneously recorded speech and EGG signals.*

here, the search range of each time point is set as

$$n_l = n_l^0, n_l^0 \pm 1, ..., n_l^0 \pm \delta n$$

for the initial position (n_l^0) and the range parameter (δn). Then the solution can be obtained as

$$\begin{aligned} n_{2L-1}^* &= \arg \min_{n_{2L-1}} D_{2L-1}(n_{2L-1}) \\ n_{l-1}^* &= \arg \min_{n_{l-1}} D_l(n_l^*) \\ &\qquad \text{for} \quad l = 2L-1, ..., 2 \end{aligned}$$

where n_l^* represents the optimal time alignment.

5. Analysis procedure

Figure 2 shows the analysis procedure for determining the optimal values of the vocal-tract and voice-source parameters. To derive an accurate solution, EGG (Electroglottgraph) signal is recorded simultaneously with speech. It is first high-pass filtered to eliminate the bias components, and the zero-crossing points in time are selected as the initial instants of the glottal events. Then the joint estimation procedure and the temporal estimation procedure are alternately performed to obtain the analysis result. The analysis frame is time-shifted so that the overlap between the adjacent frames is permitted. Then the temporal determination of the glottal events is performed more accurately and it leads to a better solution. The constraint given in Eq. (4) can be incorporated in the temporal estimation procedure by adding a penalty term to the segmental error when the constraint is not satisfied.

6. Experiment

In Fig. 3, waveform of the base signal of the voice source model is shown in the left and its magnitude spectrum is plotted in the right for the order parameter (α) of zero, one, and two. The fundamental frequency and open quotient are set at 100Hz and 0.5 respectively for the sampling frequency of 8kHz. It is clear that

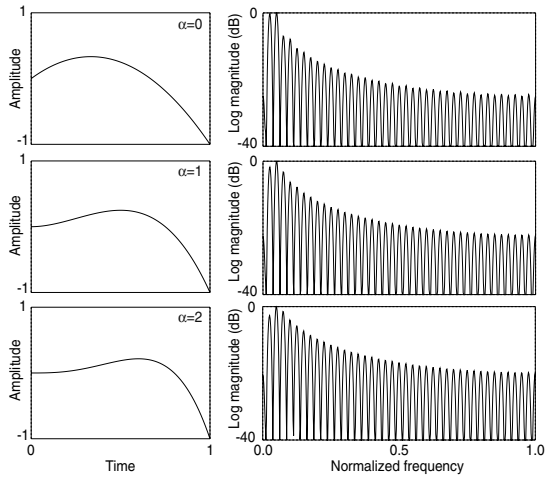

Figure 3: *Waveform and frequency characteristics of base signals of the voice source model.*

the order parameter can control the sharpness of the waveform and the envelope of the magnitude spectrum.

Figure 4 shows analysis results of vowel /a/ uttered by a male subject using three phonation conditions, i.e., modal, falsetto, and pressed. The traces in the left represent the waveform of the original speech in (A) and that of the estimated voice source signal in (B). Magnitude spectrum of the estimated voice source signal is shown in the right. In this experiment, four base signals with the order parameter of zero, two, four, and six were used. The order of the vocal-tract filter and the frame length (number of pitch periods) were 14 and two for the modal and pressed voices and ten and three for the falsetto. One pitch period was overlapping in the shift of the analysis frame. The results would indicate that our method can extract features of the voice source characteristics, such as the envelope of the magnitude spectrum and the shape and open quotient of the waveform, depending on the type of the phonation. As the measure of the objective evaluation, signal-to-noise ratio (SNR) of the signal prediction error was 10 to 15dB.

7. Summary

A speech analysis method is proposed based on the joint estimation of the vocal-tract and voice-source parameters. The voice source is modeled by overlapping piecewise base signals defined for the open phase of the glottal movements. Our method is capable of flexibly representing the voice source characteristics, including the jitter and shimmer, during the phonation. Experimental results indicate that the proposed method has the potential for effectively capturing the characteristics of the glottal source signals among different phonation types and for investigating the nature of speech from the viewpoint of the voice source dynamics.

This research was partly supported by the Grant-in-Aid for Scientific Research from the JSPS (Grant No.14101001).

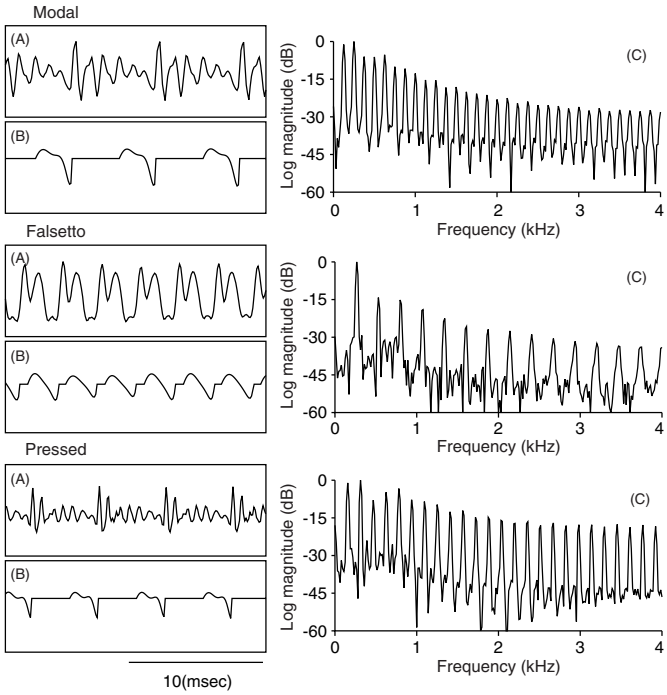

Figure 4: *Analysis results of vowel utterances using three phonation types.*

8. References

[1] Rosenberg, A. E., "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am., **49**, 583-590, 1971.

[2] Childers, D. G. and Lee, C. K., "Vocal quality factors: Analysis, synthesis, and perception," J. Acoust. Soc. Am. **90**, 2394-2410, 1991.

[3] Klatt, D. H. and Klatt, L. C., "Analysis,synthesis,and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am., **87**, 820-857, 1990.

[4] Fujisaki, K. and Ljungqvist, M., "Estimation of the voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform," Proc. Int. Conf. Acoust., Speech, Signal Processing, 637-640, 1987.

[5] Milenkovic, P., "Glottal inverse filtering by joint estimation of an AR system with a linear input model," IEEE Trans. Acoust., Speech, Signal Processing, ASSP-34, 28-42, 1986.

[6] Funaki, K., Miyanaga, Y., and Tochinai, K., "A time varying ARMAX speech modeling with phase compensation using glottal source model," Proc. Int. Conf. Acoust., Speech, Signal Processing, 1299-1302, 1997.

[7] Kasuya, H., Maekawa, K., and Kiritani, S., "Joint estimation of voice source and vocal tract parameters as applied to the study of voice source dynamics," Proc. ICPhS99, 2505-2512, 1999.

[8] Milenkovic, P., "Voice source model for continuous control of pitch period," J. Acoust. Soc. Am., **93**, 1087-1096, 1993.