Source-filter separation for nonstationary voiced speech based on sinusoidal representation

Masashi Ito^{1,*}, Keiji Ohara², Akinori Ito¹ and Masafumi Yano²

¹Department of Electrical and Communication Engineering, Tohoku University, 6-6-05 Aramaki-Aoba, Aoba-ku, Sendai, 980-8579 Japan

²Research Institute of Electrical Communication, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai, 980-8577 Japan

(Received 13 May 2009, Accepted for publication 8 July 2009)

Keywords: Speech analysis, Sinusoidal model, Vocal-tract transfer function

PACS number: 43.72.Ar [doi:10.1250/ast.31.181]

1. Introduction

Speech production is well explained by a source-filter model [1]. The source of voiced speech is a quasi-periodic signal produced by the vibration of vocal folds. This signal is then modified by a linear filter, the characteristics of which are governed by the physical shape of the vocal tract, to produce a resultant speech signal. In speech analysis, it is of prime importance to separate the source and filter characteristics appropriately. A cepstrum method estimates the fundamental frequency (F_0) and the amplitude response of the vocal-tract transfer function (VTTF) for input speech on the basis of the Fourier transform of the logarithmic-magnitude spectrum [2]. Linear predictive coding (LPC) is used to estimate the VTTF on the basis of the assumption that it is approximated by an all-pole model [3]. However, the analytical accuracy of these methods is not sufficient for natural utterances. A signal reconstructed from the estimated parameters is not perceptually equivalent to the original speech. This inaccuracy becomes crucial in nonstationary parts of the input speech because in the conventional methods, the source signal and/or VTTF is assumed to be stationary within a short analysis period.

To overcome this problem, the use of a sinusoidal model seems to be promising. The model approximates speech signals as the sum of sinusoidal components with instantaneous amplitudes and frequencies that continuously vary with time [4]. In the present study, we propose a novel analysis method to estimate VTTF and F_0 for nonstationary voiced speech on the basis of sinusoidal representation. The effectiveness of the proposed method is examined for synthesized and naturally uttered successive vowels.

2. Algorithm

2.1. Local vector transform

McAuley and Quatieri proposed a sinusoidal model to analyze speech signals. The model utilizes local peaks in the power spectrum to estimate sinusoidal parameters [4]. Since this model is not very effective for nonstationary parts of the input signal, the authors proposed an improved sinusoidal model, a local vector transform (LVT) [5]. The LVT explicitly assumes nonstationarity of the sinusoidal parameters

in the input signal $x(t)$, where instantaneous logarithmic amplitude $a(t)$ and phase $p(t)$ might vary with simple quadratic functions of time in an analysis frame centered at t_C , as follows.

$$a(t) = \text{Re}[\alpha(t - t_C)]|_{t \approx t_C} \quad (1)$$

$$p(t) = \text{Im}[\alpha(t - t_C)]|_{t \approx t_C} \quad (2)$$

$$\alpha(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 / 2 \quad (3)$$

The three complex coefficients α_0 , α_1 and α_2 can be determined from the short-term Fourier spectrum $X_0(\omega)$ and its first and second derivatives ($X_1(\omega) = dX_0(\omega)/d\omega$ and $X_2(\omega) = d^2X_0(\omega)/d\omega^2$) as follows. Here, γ represents a positive constant for Gaussian window $w(t)$.

$$X_0(\omega) = \int_{-T}^T x(t - t_C)w(t) \exp(-j\omega t) dt \quad (4)$$

$$w(t) = \exp(-\gamma \cdot t^2) / \sqrt{2\pi} \quad (5)$$

$$\alpha_0 = \log(2X_0^2 / \sqrt{X_1^2 - X_0X_2}) + X_1^2 / 2(X_1^2 - X_0X_2) \quad (6)$$

$$\alpha_1 = j \cdot (\omega + X_0X_1 / (X_1^2 - X_0X_2)) \quad (7)$$

$$\alpha_2 = \gamma - X_0^2 / (X_1^2 - X_0X_2) \quad (8)$$

Thus LVT can estimate not only instantaneous frequency $f(t)$ but also its velocity $df(t)/dt$ of the input signal. Figure 1 shows an example of the estimated parameters for a voiced speech signal. As reported in the previous study [5], the analytical accuracy of the LVT is equal to or higher than those of the examined conventional methods, at least for the lower harmonic components in voiced speech signals.

2.2. Time warping

For higher harmonic components, it is difficult, in many analysis methods including the LVT, to estimate accurate sinusoidal parameters within the period where F_0 of the input signal greatly varies with time. This is because spectral energy leaks and interferes with each other in these components. Figure 2 shows an example of the interference effect. The upper panel represents a power spectrum of voiced speech uttered by a female speaker. Since F_0 of the signal (233 Hz) greatly increases in the analysis period, there are no prominent peaks corresponding to the higher components in the spectrum.

*e-mail: itojin@makino.ecei.tohoku.ac.jp

Fig. 1 Local vector transform for speech signals.

Fig. 2 Power spectra of input and time-warped speech signals.

Fig. 3 Input and estimated VTTFs for synthesized vowels.

A time-warping technique is known to be effective for reducing this spectral interference [6]. The lower panel in Fig. 2 shows a power spectrum for the time-warped signal, in which the time axis of the input signal is converted to cancel the F_0 increase on the basis of $df(t)/dt$ for the first harmonic component estimated by the LVT. When each component becomes separable in the spectrum, its sinusoidal amplitude and phase can be simply estimated by the Hilbert transform with the use of an appropriate band-pass filter. Then, the input signal is expressed by the estimated sinusoidal parameters as follows.

$$x(t) = \sum_{k=1}^K \exp(A_k(t)) \cos(P_k(t)) \quad (9)$$

Here, $A_k(t)$ and $P_k(t)$ represent the instantaneous logarithmic amplitude and phase for the k -th component, respectively.

2.3. Source-filter separation

Next, the VTTF and source information are extracted from the sinusoidal representation of Eq. (9). The main feature of the source signal is F_0 . The fundamental angular frequency $\omega_0 (= 2\pi F_0)$ can be determined as follows.

$$\omega_0(t) = \frac{1}{K} \sum_{k=1}^K \frac{1}{k} \frac{dP_k(t)}{dt} \quad (10)$$

The VTTF consists of the logarithmic amplitude $M(t, \omega)$ and phase $\theta(t, \omega)$. It should be noted that they are not time-invariant functions. We define the VTTF by which a single-component sinusoidal signal $y(t) = \exp(a(t)) \cos(p(t))$ is modified as follows.

$$y'(t) = \exp \left[a(t) + M \left(t, \frac{dp(t)}{dt} \right) \right] \cos \left[p(t) + \theta \left(t, \frac{dp(t)}{dt} \right) \right] \quad (11)$$

From the estimated sinusoidal parameters and $\omega_0(t)$, $M(t, \omega)$ and $\theta(t, \omega)$ are determined at angular frequency $\omega(t) = k \cdot \omega_0(t)$ as follows.

$$M(t, k\omega_0(t)) = A_k(t) \quad (12)$$

$$\theta(t, k\omega_0(t)) = P_k(t) - k \left(\int_0^t \omega_0(\tau) d\tau + \varphi_0 \right) \quad (13)$$

$$\varepsilon = \sum_{k=1}^K \int \theta(t, k\omega_0(t))^2 dt \quad (14)$$

Here, φ_0 represents the fundamental phase at $t = 0$ and is determined by minimizing ε in Eq. (14). $M(t, \omega)$ and $\theta(t, \omega)$ at angular frequency $\omega \neq k \cdot \omega_0$ are determined by a simple interpolation of the available neighboring values.

3. Experiment

3.1. Synthesized vowels

The effectiveness of the above-mentioned algorithm applied to synthesized speech signals, of which F_0 and VTTF are clearly known, is examined. The signal is successive vowels /ai/ synthesized by a cascade-Klatt synthesizer [7]. The lower four formant frequencies are smoothly varied from /a/ ($F_1 = 0.70$; $F_2 = 1.14$; $F_3 = 2.70$; $F_4 = 3.66$ kHz) to /i/ ($F_1 = 0.21$; $F_2 = 2.52$; $F_3 = 3.52$; $F_4 = 3.94$ kHz) with a sigmoid function. F_0 increases linearly with time as $F_0(t) = 125 + 625t$ (Hz). Figures 3(a) and 3(b) show amplitude and phase responses of the VTTF for the synthesized vowels, respectively.

For this input signal, F_0 and VTTF are estimated using the proposed algorithm. The estimated F_0 was in good agreement with that of the input signal. Figures 3(c) and 3(d) shows the estimated $M(t, \omega)$ and $\theta(t, \omega)$ for the input signal, respectively. As shown in the figure, both responses were very close to those of the input VTTF (Figs. 3(a) and 3(b)). The small differences between the input and estimated responses reflects a limitation of the interpolation process, in which $M(t, \omega)$ and $\theta(t, \omega)$ at $\omega \neq k \cdot \omega_0(t)$ must be estimated from neighboring sinusoidal parameters $A_k(t)$ and $P_k(t)$ that were obtained at only frequency $\omega = k \cdot \omega_0(t)$.

3.2. Naturally uttered vowels

The proposed algorithm is also examined for its application to natural utterance. The input speech is successive vowels /ai/ actually spoken by a male speaker. Figure 4(a) shows a spectrogram of the speech. Figure 4(b) shows sinusoidal parameters for the speech, estimated using the algorithm described in Sect. 2.1 and 2.2. Since the actual

Fig. 4 Sinusoidal representation for vowels.

sinusoidal parameters are unknown for natural utterances, a signal was reconstructed from the estimated sinusoidal parameters and compared with the input to evaluate the accuracy of the analysis. Figure 4(c) shows the input, reconstructed, and residual signals. The ratio of the signal to residual power was 32.1 dB. Furthermore, the reconstructed signal was perceptually indistinguishable from the original speech. The results confirmed the accuracy of the sinusoidal parameter estimation to be sufficient.

Figures 5(a) and 5(b) show estimated $M(t, \omega)$ and $\theta(t, \omega)$ for input speech, respectively. As shown in the figure, neither of the responses differed greatly from those of the phonetically identical synthesized vowels (Figs. 3(a) and 3(b)). This implies that the VTTFs might be estimated well although

Fig. 5 Estimated VTTF for naturally uttered vowels.

there were some discontinuities observed, particularly in the phase responses ($t = 100$ ms and $f > 1.5$ kHz; $f > 4.5$ kHz; the initial and final parts).

4. Conclusion

On the basis of a sinusoidal representation, a method of analyzing F_0 and VTTF for nonstationary voiced speech signals is proposed. The experimental results indicated that the proposed method is effective, at least for the examined synthesized and naturally uttered speech signals. Further research is needed to clarify the effectiveness of the method by using a large dataset and performing quantitative comparison with the conventional methods.

Acknowledgements

This study was supported by KAKENHI (21700282 and 17075003).

References

- [1] G. Fant, *Acoustic Theory of Speech Production* (Mouton, The Hague, 1960).
- [2] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, **41**, 293–309 (1967).
- [3] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, **50**, 637–655 (1971).
- [4] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust. Speech Signal Process.*, **34**, 744–754 (1986).
- [5] M. Ito and M. Yano, "Sinusoidal modeling for nonstationary speech based on a local vector transform," *J. Acoust. Soc. Am.*, **121**, 1717–1727 (2007).
- [6] T. Abe and M. Honda, "Sinusoidal model based on instantaneous frequency attractors," *IEEE Trans. Audio Speech Lang. Process.*, **14**, 1292–1300 (2006).
- [7] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, **67**, 971–995 (1980).