Visualization of phase characteristics of a vocal tract using phase-conversion method with electroglottograph signals

Masahiro Okada*

National Research Institute of Police Science, 6-3-1 Kashiwanoha, Kashiwa, 277-0882 Japan

(Received 11 December 2020, Accepted for publication 22 January 2021)

Keywords: Phase conversion, Short-time Fourier transform (STFT), Phase characteristics, Source-filter theory, Electroglottograph (EGG)

1. Introduction

In speech analyses, we typically use short-time Fourier transform (STFT) and visualize power spectrograms of speeches. Although STFT provides complex values in the form of magnitude and phase, phase information is rarely used because of its time sensitivity, indicating the lack of sufficient information. Phase has an important role especially in resonance analyses. In the frequency domain, phases reverse by about $-\pi$ in the vicinity of a resonance frequency. Despite the power at a certain frequency is small, its phase inversion confirms whether resonances have occurred.

Recently, the phase-conversion method that modifies the time sensitivity of phases was proposed [1,2]. In the present study, we applied it to speech analyses and verified its efficiency. For the first time, we attempted to visualize resonances for a vocal tract.

2. Theoretical idea

The STFT of a signal $x(t)$ with a real-valued symmetric time window $w(t) = w(-t)$ can be expressed as follows:

$$\begin{aligned} S_X(t, \omega) &= \int_{-\infty}^{\infty} x(\tau)w(\tau - t)e^{-i\omega(\tau - t)}d\tau \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\Omega)W(\Omega - \omega)e^{i\Omega t}d\Omega, \end{aligned} \quad (1)$$

where $X(\Omega)$ and $W(\Omega) = W(-\Omega)$ are the Fourier transforms of $x(t)$ and $w(t)$, respectively, and $i = \sqrt{-1}$. Note that $W(\Omega)$ is real-valued because $w(t)$ is symmetric. Here, let $X(\Omega)$ have a harmonic structure, i.e., $X(\Omega) \approx \sum_n X(\omega_n)\delta(\Omega - \omega_n)$, where $\delta(\Omega)$ is the Dirac delta function. Further, the bandwidth of the filter, $W(\Omega)$, is narrow enough to pass one of the harmonics only. Then, the phase of STFT can be represented as [2,3],

$$\angle S_X(t, \omega) \approx \omega_n t + \angle X(\omega_n) \quad \text{for } \omega \approx \omega_n. \quad (2)$$

According to the source-filter theory [4], human speeches in the frequency domain $X(\omega_n)$ can be represented as

$$X(\omega_n) = H(\omega_n)G(\omega_n), \quad (3)$$

where $H(\omega_n)$ and $G(\omega_n)$ are the resonance characteristics of a vocal tract and the glottal source characteristics at frequency ω_n , respectively. When we obtain the glottal source wave-

form, the phase of the vocal tract characteristics (i.e., $\angle H(\omega_n)$) can be calculated using Eq. (2) as

$$\angle H(\omega_n) = \angle X(\omega_n) - \angle G(\omega_n) \approx \angle S_X(t, \omega) - \angle S_G(t, \omega). \quad (4)$$

Therefore, the argument of the quotient $S_X(t, \omega)/S_G(t, \omega)$ indicates the phase characteristics of a vocal tract. We typically visualize power spectrograms only, but we can also visualize phase spectrograms based on the above idea.

3. Numerical simulation for time-varying speeches

In the previous section, we assumed steady speeches whose frequency components do not vary with time. We therefore perform a numerical simulation and verification for time-variant speeches in this section.

3.1. Time-varying pitch

First, we focused on the effects of fundamental frequency of vocal folds. According to Rosenberg [5], we modeled a glottal source waveform using the following equations:

$$g(t) = \begin{cases} \frac{1}{2} \left[1 - \cos\left(\frac{\pi}{T_{op}}t\right) \right] & \text{for } 0 \leq t < T_{op} \\ \cos\left(\frac{\pi}{2T_{cl}}(t - T_{op})\right) & \text{for } T_{op} \leq t < T_{op} + T_{cl} \\ 0 & \text{for } T_{op} + T_{cl} \leq t < T_0, \end{cases} \quad (5)$$

and

$$g(t) = g(t + mT_0). \quad (6)$$

The parameters, T_{op} , T_{cl} , and T_0 , indicate the opening period, closing period, and fundamental period, respectively, and m is an integer. In the simulation, they were set in such a way so that the following two conditions are satisfied: $(T_{op} + T_{cl})/T_0 = 0.8$ and $T_{op}/T_{cl} = 1.2$.

The fundamental frequency, also known as the pitch, $1/T_0$, was first varied upward and then downward in a linear manner in both the directions, from 100 Hz to 300 Hz and again back to 100 Hz.

Here, a vocal tract filter was modeled as a fourth-order auto-regression (linear time-invariant) system, with the two formant frequencies being set to $F_1 = 800$ Hz and $F_2 = 1,200$ Hz, corresponding to the Japanese vowel /a/.

In the STFT analysis after the speech synthesis, a Blackman window, whose dynamic range between its main lobe and its side lobe is high [6], was used, and its length was set to 25 msec. The time shift of the window was one sixteenth of the window length. The calculations were performed using

*e-mail: m.okada@nrips.go.jp
[doi:10.1250/ast.42.150]

Fig. 1 Power spectrumgram of the synthesized speech (upper) and phase spectrumgram of the filter (lower). The frequency region is restricted up to 3,000 Hz to focus on formants, although the sampling frequency was set to 8,000 Hz.

Python 3.7, and the obtained phase spectrumgram was represented using the twilight cyclic colormaps in Matplotlib to avoid phase unwrapping.

Figure 1 shows the power spectrumgram of the synthesized speech and the phase spectrumgram of the filter estimated by our proposed method. It can be observed that near the two poles ($F_1 = 800$ Hz and $F_2 = 1,200$ Hz), phases are reversed by about $-\pi$, and the result is in good agreement with our theory.

3.2. Time-varying formants

Next, we considered time-variant formant frequencies. The first formant frequency, F_1 , was first set to 800 Hz, then 200 Hz, and again back to 800 Hz. Similarly, the second formant frequency, F_2 , was first set to 1,200 Hz, then 2,000 Hz, and again back to 1,200 Hz. This is in correspondence with the time-variant Japanese vowel, that is, /a/, /i/, and back to /a/. In the transition, formant frequencies were varied linearly, whereas the pitch was fixed to 100 Hz. The other parameters had the same values as in the time-variant pitch simulation.

In this case, the assumption of the source-filter theory (Eq. (3)) is discarded because the vocal tract system is time variant. We therefore introduced an approximation that only the impulse response of the filter varies with time; in other words, a speech $x(t)$ can be calculated by the following formula:

$$x(t) = \int_{-\infty}^{\infty} g(\tau)h_{\tau}(t - \tau)d\tau, \quad (7)$$

where $h_{\tau}(t - \tau)$ is the impulse response varying with time delay τ .

Figure 2 shows the power spectrumgram and the phase spectrumgram, as in Fig. 1. It is observed that near the two time-varying poles, phases are reversed by about $-\pi$ and thus, the result is in good agreement with our theory. In addition, we simulated a situation in which both the formants and pitch vary, which is shown in Fig. 3. Here, the pitch varies upward only. According to the figure, the obtained phase spectrumgram remains viewable despite time-variant pitches.

Fig. 2 Power spectrumgram of the synthesized speech (upper) and phase spectrumgram of the filter (lower) in the situation that only formant frequencies vary. As in Fig. 1, the frequency region is restricted up to 3,000 Hz.

Fig. 3 Power spectrumgram of the synthesized speech (upper) and phase spectrumgram of the filter (lower) in the situation that both formant frequencies and the pitch vary. As in Fig. 1, the frequency region is restricted up to 3,000 Hz.

4. Real data application

To investigate the effectiveness of our idea in real data, we finally attempted to apply it to natural speeches. Here, we used electroglottograph (EGG) signals as glottal source waveforms because EGG measurements are non-invasive and hardly interfere with the utterances of a speaker. Typically, we analyzed some reading-style speeches from MULTEXT-J (Japanese MULTEXT) corpus [7,8], which contains both Japanese speeches and simultaneously recorded EGG signals of three males and three females. These signals were used after being down sampled to 8,000 Hz.

A speech is physically generated from a glottal source signal through a vocal tract. Hence, there is a time delay between the two signals. There are some correction methods, for example, time shifting which corresponds to the vocal tract length [9]. In this study, however, we corrected the delay by time shifting while computing a cross-correlation between the envelopes of a speech and an EGG. The envelopes were calculated using Hilbert transform. The cross-correlation was computed by utilizing the entire length of an utterance to

Fig. 4 Power spectrogram of the speech, /bezitariaN/, (upper) and phase spectrogram of the filter (lower) for a male.

Fig. 6 Power spectrogram of the speech, /bezitariaN/, (upper) and phase spectrogram of the filter (lower) for the other male.

Fig. 5 Power spectrogram of the speech, /bezitariaN/, (upper) and phase spectrogram of the filter (lower) for another male.

Fig. 7 Power spectrogram of the speech, /bezitariaN/, (upper) and phase spectrogram of the filter (lower) for a female.

contain more voiced segments because EGG signals often have noise or meaningless values at unvoiced segments.

In the analysis, the window lengths were reset to 25 msec and 12.5 msec for male speeches and female speeches, respectively. The time shift of the window and the other parameters were the same as those used in the numerical simulations.

Figures 4–8 show spectrograms as in the previous sections. Here, cross-correlation computations failed in one female; thus, the female was excluded from the results. As predicted by our theory, the time sensitivity of phases seems to have disappeared. The obtained phase spectrograms show a similar structure and phases drastically vary around formant frequencies, especially in male speeches, although speeches and EGG signals of all speakers differ from each other. Moreover, the phase structure is observable even if the powers are small (i.e., in frequency regions far from formants).

Compared with male speeches, female speeches are less viewable. The proposed method depends on the density of the harmonics in an analyzed speech, because it utilizes the discretization of the filter characteristics by the harmonics of the glottal sources in the frequency domain (see Eq. (3)). This is the reason the phase spectrograms of male speeches have finer frequency resolutions.

Fig. 8 Power spectrogram of the speech, /bezitariaN/, (upper) and phase spectrogram of the filter (lower) for the other female.

5. Conclusions and limitations

In this study, the author proposed visualizing the phase spectrogram of a vocal tract and verified its effectiveness. However, this method has several limitations as mentioned below:

- (a) Time shifting by cross-correlation method tended to be incorrect for female speeches. This might be because their EGG signals were weaker than those of males, indicating this method would depend on firm vibrations of vocal folds.
- (b) In natural speech spectrograms, the value of phase, especially that in the higher-frequency domain, would be affected by errors in the time delay correction. This is because the time shifting operation is expressed by multiplication with $\exp(-i\omega t_s)$ in the Fourier domain, where t_s denotes the shift time; the phase depends highly on t_s at higher frequencies. Therefore, the value of phase at higher frequencies would be less reliable than that at lower frequencies.
- (c) EGG signals were used in this work, owing to their non-invasiveness. However, they do not represent glottal sources, which are considered as acoustic volume velocities [10], but the degree of contact of vocal folds [11]. From this perspective, EGG signals are not appropriate for our theoretical model, and our results cannot be considered as vocal tract characteristics.
- (d) The source-filter theory (Eq. (3)), in reality, contains radiation characteristics, and we excluded them in this paper. Although influences on phases by radiations might be smaller than resonances, they should be considered.

The proposed method still has the above problems, and they should be improved in the future. We however hope that this method or idea helps in some speech analyses.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP20K23357, JP18H01671.

References

- [1] İ. Bayram and M. E. Kamasak, "A simple prior for audio signals," *IEEE Trans. Audio Speech Lang. Process.*, **21**, 1190–1200 (2013).
- [2] K. Yatabe, Y. Masuyama, T. Kusano and Y. Oikawa, "Representation of complex spectrogram via phase conversion," *Acoust. Sci. & Tech.*, **40**, 170–177 (2019).
- [3] F. Auger, E. Chassande-Mottin and P. Flandrin, "Making reassignment adjustable: The Levenberg-Marquardt approach," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 3889–3892 (2012).
- [4] G. Fant, *Acoustic Theory of Speech Production: With Calculations Based on X-ray Studies of Russian Articulations*, 2nd printing (Mouton, Hague, 1970), pp. 15–26.
- [5] P. J. Murphy, "Spectral characterization of jitter, shimmer, and additive noise in synthetically generated voice signals," *J. Acoust. Soc. Am.*, **107**, 978–988 (2000).
- [6] Y. Wakabayashi, "Speech enhancement using harmonic-structure-based phase reconstruction," *Acoust. Sci. & Tech.*, **40**, 162–169 (2019).
- [7] S. Kitazawa, T. Kitamura, K. Mochiduki and T. Itoh, "Preliminary study of Japanese MULTEXT: A prosodic corpus," *Proc. Int. Conf. Speech Processing*, pp. 825–828 (2001).
- [8] S. Kitazawa, S. Kiriya, T. Itoh and N. Campbell, "Japanese MULTEXT: A prosodic corpus," *Proc. Int. Conf. Language Resources and Evaluation*, pp. 2167–2170 (2004).
- [9] P. Alku, J. Pohjalainen, M. Vainio, Anne-Maria Laukkanen and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction," *J. Acoust. Soc. Am.*, **134**, 1295–1313 (2013).
- [10] I. Steinecke and H. Herzel, "Bifurcations in an asymmetric vocal-fold model," *J. Acoust. Soc. Am.*, **97**, 1874–1884 (1995).
- [11] N. Henrich, C. d'Alessandro, B. Doval and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *J. Acoust. Soc. Am.*, **115**, 1321–1332 (2004).