Derivative of instantaneous frequency for voice activity detection using phase-based approach

NGUYEN Binh Thien*, Yukoh WAKABAYASHI†*, Takahiro FUKUMORI*, Takanobu NISHIURA*
* Ritsumeikan University, Shiga, Japan
E-mail: {gr0398xe@ed, wakayuko@fc, fukumori@fc, nishiura@is}.ritsumei.ac.jp
† Tokyo Metropolitan University, Tokyo, Japan

***Abstract*—In this paper, we consider the use of the phase spectrum in speech signal analysis. In particular, a phase-based voice activity detection (VAD) by using the derivative of instantaneous frequency is proposed. Preliminary experiments reveal that the distribution of this feature can indicate the presence or absence of speech. The performance of the proposed method is evaluated in comparison with the conventional amplitude-based method. In addition, we consider a combination of the amplitude-based and phase-based methods in a simple manner to demonstrate the complementarity of both spectra. The experimental results confirm that the phase information can be used to detect voice activity with at least 62% accuracy. The proposed method shows better performance compared to the conventional amplitude-based method in the case when a speech signal was corrupted by white noise at low signal-to-noise ratio (SNR). A combination of two methods achieves even higher performance than each of them separately, in limited conditions.**

I. Introduction

Speech processing in time-frequency domain has handled amplitude spectrum much more thoroughly than the phase spectrum. Previous studies argued that the most important information could be obtained from the amplitude spectrum, while very little information could be obtained from the phase spectrum. Wang and Lim [1] concluded from their experiments that a more accurate estimation of phase is unwarranted in speech enhancement. Vary [2] also showed that for the SNR above 6 dB, there is no degradation in the synthesized speech that could be perceived if the noisy phase is used as the estimation of clean phase spectrum. However, phase information is gaining more and more attention from the researchers. Paliwal *et al.* demonstrated the usefulness of the phase spectrum in speech signal processing [3] and human speech perception [4]. Gerkmann *et al.* [5] presented the review of phase processing for single-channel speech enhancement. Mowlaee *et al.* [6] reported the advances in phase-aware signal processing in speech communication. And recently, other studies demonstrated the importance of the phase in speech signal processing such as source separation [7]–[9], speech synthesis [10]–[12], and speech enhancement [13]–[17].

The phase spectrum itself contains valuable information about the structure of the signal, but it is hidden due to the phase wrapping issue, which causes a fuzzy pattern in the phase spectrogram as shown in Fig. 1(b). To extract this information, some other representations of the phase spectrum have been proposed. One of the most important representations

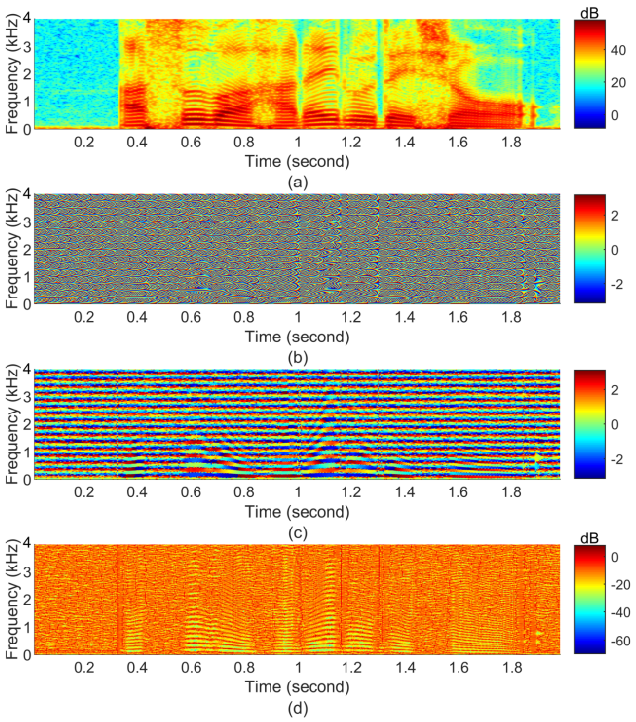

Fig. 1. (a) Amplitude spectrogram, (b) phase spectrogram, (c) instantaneous frequency, and (d) derivative of instantaneous frequency with respect to frequency in logarithmic scale.

is instantaneous frequency (IF) [18], shown in Fig. 1(c). The IF is derived from the phase spectrum by taking the derivative of the phase with respect to the time. By doing so, we can reduce the effect of the wrapping issue, thereby revealing the harmonic structure of the speech signal, as illustrated in Fig. 1(a). In speech signal processing, the IF can be used to detect the presence of vowels or to extract the harmonic frequencies, which is a very useful characteristic. To improve visualization, we use the derivative of the IF with respect to frequency (DIF). Fig. 1(d) shows the DIF spectrogram in a logarithmic scale. We can see that the structure of the speech signal is shown more clearly.

Previous works show the usefulness of the IF in such aspects of speech signal processing as speaker identification [19], source separation [20], and formant detection [21]. Responding to the success of recent phase-aware studies, we propose a

phase-based voice activity detection. VAD detects the presence or absence of human speech and plays an important role in speech processing, especially in speech coding [22] and speech recognition [23]. The conventional VAD algorithms [24]–[26] mostly use the amplitude information to recognize the presence or absence of speech. In our research, we use the phase information, specifically, the difference between the statistic distribution of the DIF in voiced/unvoiced segments to estimate voice activity.

The remainder of this paper is organized as follows. In Section II, we formulate and analyze the phase information in short-time Fourier transform (STFT) domain. In Section III, we describe the proposed phase-based VAD algorithm. Section IV reports the experiments and results. Finally, Section V concludes the paper.

II. Formulation and Analysis of Phase Features

A. Notation

Let t, ω, and T be the time index, frequency index, and frame length, respectively. The STFT of a continuous-time speech signal $x(t)$ is defined as:

$$X(\omega, t) = \int_0^T x(t+\tau)w(\tau)e^{-j\omega\tau}d\tau, \tag{1}$$

where $w(\tau)$ is the window function and j is the imaginary unit. Let $\angle$ denote the angle operator, then the phase spectrum at time t is denoted as $\angle[X(\omega, t)]$. The discrete time version of (1) for signal $x(n)$ can be given as follows:

$$X(k, l) = \sum_{n=0}^{N-1} w(n)x(n + lH)e^{-j2\pi kn/N}, \tag{2}$$

where $l = 0, \cdots, L-1$ is the frame index, and k, H, and N are the frequency bin index, hop size, and window length, respectively.

B. Analysis of derivative of instantaneous frequency

One of the most important phase-based features is IF, which is defined as a derivative of the phase with respect to time:

$$\phi(\omega, t) = \frac{\partial\angle[X(\omega, t)]}{\partial t}. \tag{3}$$

For discrete time signal processing, Kay [27] proposed a method to avoid the phase unwrapping problem for calculating the IF:

$$\phi(k, l) = \angle[X(k, l+1)X^*(k, l)], \tag{4}$$

where X^* is the complex conjugate of X. The time derivative can extract the temporal fluctuations of the phase information, especially when the signal moves from an unvoiced segment to a voiced segment and vice versa. Beyond those cases, the speech signal moves slowly comparing to the frame rate; therefore, the IF also changes slowly in the same frequency band. At the unvoiced segments, e.g., from the beginning to 0.3 second in Fig. 1, the IF depends linearly on the frequencies of the STFT, and the IF spectrum increases regularly along the frequency axis. However, due to the phase wrapping issue, the IF values can only increase from $-\pi$ to π, then drop to $-\pi$, and then continue to increase, repeatedly. This phenomenon causes horizontal stripes on the IF spectrogram. At the voiced segments, especially for the vowels, e.g., from 0.6 second to 0.8 second in Fig. 1, the signal has a harmonic structure, so the IF at positions near harmonic components will be affected by them. Along the frequency axis, the IF no longer increases steadily but is divided into many bands, where each band contains several STFT frequencies and where the center is the harmonic frequency. The width of each band is proportional to the width of the corresponding harmonic component in the amplitude spectrogram. In each harmonic band, the values of the IF are approximately the same and depend on the value of the dominant harmonic frequency.

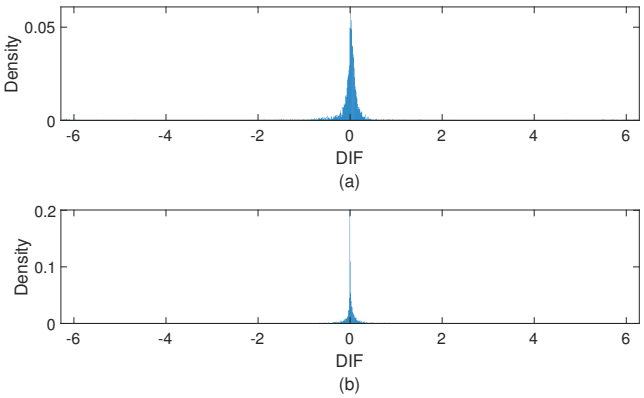

Fig. 2. Comparison between estimated probability densities of DIF in (a) unvoiced segment and (b) voiced segment.

To improve visualization, we define the DIF as the derivative of the IF with respect to frequency:

$$\psi(k, l) = \phi(k+1, l) - \phi(k, l). \tag{5}$$

As mentioned before, at the unvoiced segments, the IF spectrum increases regularly along the frequency; therefore, its frequency derivative is approximately a constant. We can see that the DIF spectrogram depicted in Fig. 1(d) has the same color at the unvoiced segments; however, there are still some thin horizontal lines due to the influence of the wrapping phenomenon. At the voiced segments, the IF in each harmonic band has the same value, hence the derivative of it is close to zero.

To clearly see the difference between the two segments, frequency bands larger than the cutoff frequency bin k_c are eliminated, because we can see from Fig. 1(d) that the DIFs at low frequencies contain much more information than at high frequencies. We analyze the DIFs at low frequencies by using their distributions. Fig. 2 illustrates the estimated distribution of the low-frequency DIF in voiced and unvoiced segments at five frames. We can clearly see the difference between the two segments: the DIF distribution concentrates near zero in the voiced segments, while it spreads out in the unvoiced segments. The difference between these two distributions expressed in their shapes and statistic measures such as variance, points at the presence or absence of speech

in the signal. In the next section, we propose a VAD method based on this difference.

III. VAD BASED ON THE DIFFERENCE OF DIF DISTRIBUTIONS

We propose a new VAD algorithm using the phase-based feature discussed in the previous section. We assume that the first small segment of the signal is unvoiced and estimate its distribution as a reference. To identify the segment of the signal as voiced or unvoiced, we compare it with the reference segment by calculating the distance between their distributions. If the two distributions have similar shapes, i.e., if the distance is small, the segment is unvoiced; otherwise, it is voiced.

The details of the algorithm are described as follows. After calculating the DIF, we eliminate all frequency bands larger than the cutoff frequency k_c:

$$\boldsymbol{\psi}(l) = \{\psi(0,l), \psi(1,l), \cdots, \psi(k_c,l)\}, \tag{6}$$

where $\boldsymbol{\psi}(l)$ is the DIF spectrum at frame l after removing high frequencies. The distribution $p(\boldsymbol{\Psi}_l)$ at frame l is estimated by the histogram of the DIF spectrogram of segment $\boldsymbol{\Psi}_l = \{\boldsymbol{\psi}(l), \cdots, \boldsymbol{\psi}(l + N_{\mathrm{H}} - 1)\}$, where N_{H} is the size of the segment. The unvoiced reference distribution $p(\boldsymbol{\Psi}_{\mathrm{ref}})$ is calculated as an average of the distributions of the first N_{R} frames, i.e., $\boldsymbol{\Psi}_{\mathrm{ref}} = \{\boldsymbol{\psi}(0), \cdots, \boldsymbol{\psi}(N_{\mathrm{R}} - 1)\}$. Next, we use the Euclidean distance to calculate the distances from all distributions to the reference distribution, i.e., the Euclidean distance between the histograms $\mathcal{L}\big(p(\boldsymbol{\Psi}_l), p(\boldsymbol{\Psi}_{\mathrm{ref}})\big)$. Then, the threshold η, determined in the experiment, is used to make decisions. If the distance is larger than the threshold, the frame is treated as voiced; otherwise, it is unvoiced. The binary mask $\mathcal{M}_l$ is generated from the distance vector and the threshold, where 1 corresponds to speech presence and 0 corresponds to speech absence. Finally, the mask is smoothed by eliminating the small-sized voiced/unvoiced segments (about 10 ms). The pseudo-code of this algorithm is given in Algorithm 1.

Algorithm 1 Phase-based VAD algorithm

Require: Speech signal $x(n)$
Ensure: VAD binary mask $\mathcal{M}_l$
  $X(k,l) = \mathrm{STFT}[x(n)]$
  $\phi(k,l) = \angle[X(k,l+1)X^*(k,l)]$
  $\psi(k,l) = \phi(k+1,l) - \phi(k,l)$
  $\boldsymbol{\psi}(l) = \{\psi(0,l), \cdots, \psi(k_c,l)\}$
  $\boldsymbol{\Psi}_l = \{\boldsymbol{\psi}(l), \cdots, \boldsymbol{\psi}(l + N_{\mathrm{H}} - 1)\}$
  $\boldsymbol{\Psi}_{\mathrm{ref}} = \{\boldsymbol{\psi}(0), \cdots, \boldsymbol{\psi}(N_{\mathrm{R}} - 1)\}$
  estimate distributions $p(\boldsymbol{\Psi}_l)$ and $p(\boldsymbol{\Psi}_{\mathrm{ref}})$
  if $\mathcal{L}\big(p(\boldsymbol{\Psi}_l), p(\boldsymbol{\Psi}_{\mathrm{ref}})\big) > \eta$ **then**
    $\mathcal{M}_l = 1$
  else
    $\mathcal{M}_l = 0$
  end if
  return $\mathcal{M}_l$ after hang-over

IV. EXPERIMENTS AND RESULTS

A. Experimental setup

The experimentation involves parameter tuning and accuracy testing. In the first experiment, we determine the threshold for the phase-based VAD algorithm by performing the algorithm with varying thresholds and then choosing the best one. The second experiment compares the performance of the proposed algorithm with the amplitude-based VAD algorithm. We also combine both methods (i.e., combine the binary masks) using the AND and OR operators and then compare all the results to investigate the relationships between the phase and the amplitude spectra. For the amplitude-based method, we use Sohn's algorithm [24], which is implemented as a function in MATLAB voicebox library [28].

To evaluate the results, we use precision, recall, F-measure, and accuracy [29]. Precision is defined as the ratio of correct voiced decisions to the total voiced samples of the estimated mask, while recall is the ratio of correct voiced decisions to the total voiced samples of the reference mask. The F-measure can be derived from precision and recall:

$$\text{F-measure} = \frac{2}{1/\text{precision} + 1/\text{recall}}. \tag{7}$$

The F-measure considers both precision and recall. The higher the F-measure, the better the result. The ideal value of the F-measure is 1, corresponding to the perfect precision and recall. Accuracy is defined as the ratio of the correct decisions to the total length of the speech signal. In the first experiment, the F-measure is used to determine the thresholding parameter. Then in the second experiment, the performance of the VAD algorithm using that parameter is evaluated by accuracy.

The tests are performed on the Japanese newspaper article sentences (JNAS) database [30], containing speech recordings of the Japanese-speakers reading excerpts from the Mainichi Newspaper. The sampling frequency is 16 kHz. The sound samples are impaired by adding white noise, babble noise, and traffic noise with varying SNR of 5 dB, 15 dB, and 25 dB. The reference decisions for the clean speech materials are made by labeling manually as in Fig. 4(a). In our implementation, the Hann window is used with 32 ms duration and 4 ms frame shift, and the number of FFT points is 4,096 with zero padding. We also choose the segment size N_{H} of five frames and the cutoff frequency k_c of 2 kHz. We assume that the first 100 ms segment is unvoiced, corresponding to N_{R} is 25 frames.

B. Threshold tuning

Distance threshold, which is used for making decisions, is one of the most important parameters in the phase-based VAD algorithm. In this experiment, we perform the algorithm on 100 clean sound samples from JNAS database with varying thresholds, compute the average precision, recall, and F-measure for each threshold, and finally choose the threshold with the highest F-measure.

Fig. 3(a) depicts the precision and recall curve, showing the trade-off characteristics of precision and recall. Fig. 3(b) shows the value of the F-measure for varying thresholds. We can see

TABLE I
ACCURACY OF VAD ALGORITHMS FOR VARIOUS ENVIRONMENTAL CONDITIONS

Noise type	SNR (dB)	Accuracy (%)			
		Phase-based	Amplitude-based	Combine phase and amplitude	
				AND	OR
White noise	5	74.83	72.28	65.99	**81.12**
	15	83.40	89.01	82.03	**90.37**
	25	87.04	**94.68**	88.87	92.85
Babble noise	5	62.04	**81.25**	63.29	80.00
	15	78.74	**85.73**	81.72	82.76
	25	84.54	87.87	**88.89**	83.51
Traffic noise	5	69.13	**81.99**	76.80	74.32
	15	75.22	84.05	**84.17**	75.10
	25	76.49	87.20	**88.06**	75.63

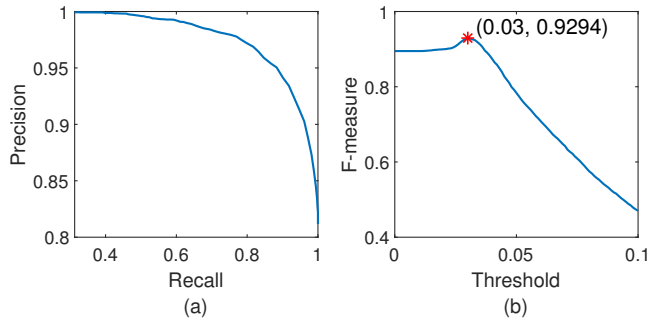

Fig. 3. Evaluation results of threshold tuning for phase-based VAD method: (a) precision and recall curve and (b) F-measure according to threshold.

that the best threshold for the phase-based VAD algorithm is 0.03, corresponding to the highest F-measure of 0.9294.

C. Evaluation of proposed method

In this section, we compare the accuracies of the proposed phase-based method, the amplitude-based method [24], and the combination method. We perform the algorithms on 200 sound samples from JNAS database. White noise, babble noise, and traffic noise are added to these signals with varying SNR.

The results of the experiment are summarized in Table I. In most testing conditions, the amplitude-based algorithm performs better than the phase-based algorithm. However, in the case of white noise at SNR of 5 dB, the phase-based algorithm yields better results. For the babble noise and the traffic noise at low SNR, the phase-based method gives the worst results. These observations can be interpreted by the characteristics of the DIF. The phase contains the information about the frequency of the signal; therefore, it is sensitive to periodic noises like babble noise (containing human voice) or traffic noise (containing vehicle horn sound), while white noise is a random signal having equal intensity for all frequencies which does not affect the phase much. We can also see from Table I that the method combining the phase-based algorithm

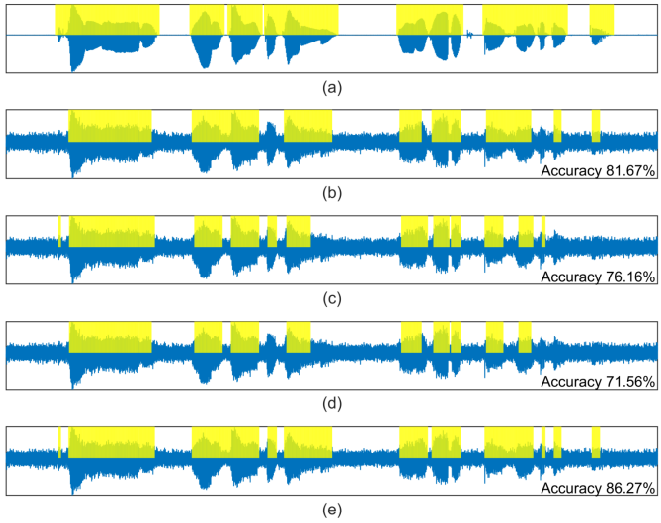

Fig. 4. Example of improved VAD algorithm using combination method with white noise added to the speech signal at 5 dB SNR: (a) clean speech signal with manual mask; (b) phase-based VAD algorithm; (c) amplitude-based VAD algorithm; (d) combination using AND operator, and (e) combination using OR operator.

and the amplitude-based algorithm using OR operator can improve the performance of the VAD algorithm for speech signals corrupted by white noise at SNR of 5 dB and 15 dB thanks to the aforementioned characteristic of DIF. Fig. 4 illustrates an example of the improvement. As shown in (b), the phase-based method estimates voice activity at the end of the second utterance and last two utterances, which are not detected by (c) the amplitude-based method. In contrast, the amplitude-based VAD can point at some voiced segments, which cannot be indicated by the phase-based algorithm. Consequently, combining two methods with the OR operator can yield better results. The combination method with AND operator can also increase the accuracy for babble noise at

SNR of 25 dB and for traffic noise at SNR of 15 dB and 25 dB, but the improvement is not significant. Overall, we can see that the algorithm using only phase information can detect voice activity with 62% accuracy at the least, and the phase information can complement the amplitude information.

V. Conclusions

We investigated the use of the phase spectrum in speech processing and proposed a new phase-based VAD algorithm using the derivative of the instantaneous frequency. We compared the performance of the proposed phase-based method with the conventional amplitude-based method in various noisy environments. Experimental results showed that voice activity can be estimated by using only the phase information. While the amplitude-based algorithm showed better accuracy in most cases, the phase-based algorithm yielded better results in the case of white noise at low SNR. We also demonstrated the possibility of combining the phase and amplitude information for better speech signal analysis.

In our future work, we will try to improve the performance of the phase-based VAD algorithm by using other distances. We will also research more effective methods for combining the phase and the amplitude information.

VI. Acknowledgment

This work is supported by JSPS KAKENHI Grants Number JP19K21546, JP18K19829, and JP19H04142.

References

[1] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[2] P. Vary and M. Eurasip, "Noise suppression by spectral magnitude estimationmechanism and theoretical limits," *Signal Processing*, vol. 8, no. 4, pp. 387–400, 1985.

[3] L. D. Alsteris and K. K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital signal processing*, vol. 17, no. 3, pp. 578–616, 2007.

[4] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[5] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, 2015.

[6] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.

[7] J. Le Roux, G. Wichern, S. Watanabe, A. Sarroff, and J. R. Hershey, "The phasebook: Building complex masks via discrete representations for source separation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 66–70.

[8] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.

[9] Z.-Q. Wang, K. Tan, and D. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 71–75.

[10] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[11] S. Takamichi, Y. Saito, N. Takamune, D. Kitamura, and H. Saruwatari, "Phase reconstruction from amplitude spectrograms based on von-Mises-distribution deep neural network," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 286–290.

[12] S. Takaki, T. Nakashika, X. Wang, and J. Yamagishi, "STFT spectral loss for training a neural speech waveform model," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7065–7069.

[13] Y. Wakabayashi, T. Fukumori, M. Nakayama, T. Nishiura, and Y. Yamashita, "Phase reconstruction method based on time-frequency domain harmonic structure for speech enhancement," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5560–5564.

[14] Y. Wakabayashi and N. Ono, "Maximum a posteriori estimation of spectral gain with harmonic-structure-based phase reconstruction for phase-aware speech enhancement," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1649–1652.

[15] Y. Wakabayashi, T. Fukumori, M. Nakayama, T. Nishiura, and Y. Yamashita, "Single-channel speech enhancement with phase reconstruction based on phase distortion averaging," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 9, pp. 1559–1569, 2018.

[16] P. Mowlaee and J. Kulmer, "Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1521–1532, 2015.

[17] M. Krawczyk-Becker and T. Gerkmann, "On MMSE-based estimation of amplitude and complex speech spectral coefficients under phase-uncertainty," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2251–2262, 2016.

[18] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. I. Fundamentals," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 520–538, 1992.

[19] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1097–1111, 2008.

[20] L. Gu, "Single-channel speech separation based on instantaneous frequency," Ph.D. dissertation, Columbia University, 2010.

[21] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1912–1924, 1999.

[22] A. ITU, "Silence compression scheme for g. 729 optimized for terminals conforming to recommendation v. 70," *ITU-T Recommendation G*, vol. 729, 1996.

[23] J. Ramírez, J. C. Segura, C. Benítez, A. De la Torre, and A. Rubio, "An effective subband osf-based vad with noise reduction for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1119–1129, 2005.

[24] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE signal processing letters*, vol. 6, no. 1, pp. 1–3, 1999.

[25] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 412–424, 2006.

[26] L.-S. Huang and C.-H. Yang, "A novel approach to robust speech endpoint detection in car environments," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, vol. 3. IEEE, 2000, pp. 1751–1754.

[27] S. Kay, "A fast and accurate single frequency estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 1987–1990, 1989.

[28] M. Brookes *et al.*, "Voicebox: Speech processing toolbox for matlab," *Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html*, vol. 47, 1997.

[29] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[30] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.