INTERSPEECH 2014 Special Session: Phase Importance in Speech Processing Applications

Pejman Mowlaee[†], *Rahim Saeidi*[‡], Yannis Stylianou[*]

[†]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
[‡]Speech and Image Processing Unit, School of Computing, University of Eastern Finland, Finland
[*]Computer Science Dept. University of Crete, Crete, Greece

pejman.mowlaee@tugraz.at, rahim.saeidi@uef.fi, yannis@csd.uoc.gr

Abstract

In many speech processing applications, the spectral amplitude is the dominant information while the use of phase spectrum is not so widely spread. In this paper, we present an overview on why speech phase spectrum has been neglected in the conventional techniques used in different applications including: speech separation/enhancement, automatic speech and speaker recognition and speech synthesis. We proceed with giving highlights on the recent progress carried out in demonstrating the importance of phase in different applications and how it impacts on the overall performance. The paper is an introduction to the Interspeech 2014 special session *phase importance in speech processing applications*.

Index Terms: Phase spectrum, speech enhancement, speech recognition, speech analysis, speech synthesis.

1. Introduction

The importance of phase spectrum of speech signals has been a controversial topic and there has been disagreement on its role in different speech processing applications. While early studies reported the unimportance of phase spectrum in perception [1, 2], more recent studies elaborated the potential of using phase spectrum in different speech and audio processing applications: speech enhancement [3–5], (1) source separation [6–10], (2) speech recognition [11–13], (3) speaker recognition [14, 15], speech coding [16], formant extraction [17], waveform estimation [18], and speech analysis/synthesis [19]. These examples suggest that incorporating the phase information can push the limits of state-of-the-art phase-independent solutions employed for long by scientists in different aspects of audio and speech signal processing. The great potential of phase information in speech processing calls for a unified effort to get a better understanding of experts from several speech and audio processing communities.

The Interspeech 2014 special session on phase importance in speech processing applications organized by the authors in this paper aims to promote the phase-based speech signal processing and explore the recent advances and referring to the new methodologies proposed for different speech applications.

2. Why Phase Has Been Neglected?

In the following, we will concentrate on three speech applications: source separation and speech enhancement, automatic speech and speaker recognition, and speech synthesis. The goal here will be to exemplify how the conventional methods neglected the phase information (see this Section) and explain the more recent advances towards the recent phase-based approaches in Section 3.

2.1. Source Separation and Speech Enhancement

From an input-output system standpoint, both speech separation and speech enhancement methods fall into the category of analysis-modification-synthesis shown in Figure 1 (left panel). The key step is to select an analysis-synthesis signal representation which satisfies two criteria: signal reconstruction and being aliasing-free. As analysis-synthesis, short-time Fourier transform (STFT) is commonly chosen where the time and frequency resolution are restricted by the choice of the window length and type. In both separation and enhancement tasks, conventionally the noisy phase is selected for signal reconstruction leading to a limited quality. The choice of noisy phase for reconstruction is supported by the fact that the noisy phase has been shown in [20] to provide the minimum mean square error (MMSE) estimate for the clean speech phase. This is only true under the assumption that the Fourier spectral coefficients are independent (obviously not the case in practice).

In the modification stage, the separation and enhancement methods are different. In single-channel source separation, it is common to assume prior or side information about the underlying signals in the mixture. The source prior knowledge could be in the form of dictionaries trained on spectral amplitude of clean signals. Some examples for learning dictionary are non-negative matrix factorization (NMF) [21], hidden Markov models (HMM) [22], Gaussian mixture models (GMM) [23] and vector quantizer (VQ) [24, 25]. A sense of optimality is required to choose the states of the underlying sources. For this purpose, in single-channel source separation, MMSE estimators in log-domain (*logmax*) [26], in power spectrum domain [23], and in spectral amplitude domain (*Elliptic series*) [25] were previously proposed. All these MMSE estimators average out the phase information in their derivations. The systematic performance comparison of these estimators has been performed in [7], demonstrating that considerable improvement in parameter estimation is possible by taking the phase information into account. The optimal states selected from dictionaries are eventually used to separate the mixture either by applying a direct synthesis [24] or by applying a soft [25, 27] or a binary mask [28] onto the mixed observation. In speech enhancement, it is common to assume a circularly symmetric complex Gaussian distribution for complex speech spectrum and derive the MMSE estimation for the speech spectral amplitude. The aim

This project has received funding from the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement no FP7-ICT-2011-7-288121.

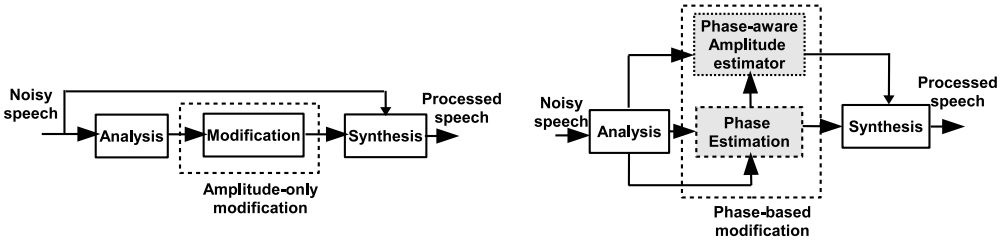

Figure 1: General system representation for typical speech enhancement/separation method as analysis-modification-synthesis. (Left) the conventional approach where the noisy speech phase is directly copied for signal reconstruction. (Right) the phase information is incorporated for signal modification (phase-aware amplitude estimation) as well as for signal reconstruction.

of the modification stage is to estimate a gain as a function of *a priori* and a *posteriori* SNRs often tabularized using a lookup table (see e.g. [29] for a list). The gain function is applied to the noisy spectral amplitude and the enhanced signal is synthesized using the enhanced spectral amplitude and noisy phase.

A typical way of including phase estimation and phase-aware processing for speech enhancement/separation is shown in Figure 1 (right panel). The phase processing of speech signal dates back to 1980s where several attempts were made to estimate the time-domain signal from a given modified spectral magnitude. This problem fits to several speech applications to name a few: speech enhancement, separation, time-scale modification and speech coding, where one is provided with a modified amplitude spectrum while there is no access to the original phase of the signal [30, 31]. Griffin and Lim proposed least square error estimation approach to estimate the time-domain signal from the given STFT spectral amplitude in an iterative way where STFT and inverse STFT steps are applied [32]. Several iterative-based techniques have been proposed to find an estimated phase from the spectral amplitude estimates of the underlying sources. Detailed overviews on performance comparison between different iterative techniques used for phase estimation in signal reconstruction are reported for speech enhancement [33] and source separation [8].

In both speech enhancement and source separation applications described above, for the synthesis stage, the observed noisy phase is directly used to reconstruct the enhanced signal. As the noisy phase has remaining contributions from the interfering source, both perceived quality and intelligibility are degraded, leading to limitation on the performance when noisy phase is used for signal modification or reconstruction.

2.2. Automatic Recognition Systems

Although the usefulness of speech phase in automatic speech and speaker recognition is not totally proven, phase spectrum has long been used for other applications like pitch and formant extraction [17, 34]. Most of the automatic speech and speaker recognition systems are built on short-term feature representation, typically calculated on the amplitude spectrum [35]. Amplitude spectrum can be calculated as the magnitude of complex Fourier transform or other parametric and non-parametric spectrum estimation methods including linear prediction, multitapering and their variants [36, 37]. The Mel-frequency Cepstral coefficients are among the most popular features derived by applying a perceptually weighted filter-bank on amplitude spectrum.

2.3. Speech Synthesis

In speech synthesis, the phase information is not used in an explicit way. Unit selection based text-to-speech synthesis systems try to select units using the magnitude spectrum as part of the concatenative cost during the selection of optimal units. The only phase information that is used is that of linear phase removals. This is in order to avoid linear phase mismatches which may result into audible clicks [38]. In general, linear phase mismatch is avoided by estimating a common reference point on the time-domain signal, like the glottal closure instants (GCIs), and then place analysis windows around these instances. However, in some cases, like creaky voice, voice offsets or expressive speech, these reference points are difficult to be defined and therefore need to be estimated.

Current HMM-based text-to-speech synthesis systems make use of minimum phase [39], since the cepstrum coefficients used in that systems are estimated by the magnitude spectrum only. The use of minimum phase for the generation of the synthetic speech signal artificially increases the correlation of speech in areas where naturally low correlation exists (i.e., fricatives). This is perceived as buzziness. To reduce this effect, researchers try to reconstruct the noise observed in the speech magnitude spectra by what is referred to as band aperiodicity [39]. This has the effect to introduce a mixed excitation (pulses plus noise at different frequency bands) in order to reduce the buzzy effect by reducing the high and unnatural correlation between consecutive speech sounds.

3. Potential of Speech Phase Information

The structure in phase spectrum of audio signals has been demonstrated and found useful in *music processing* [40] e.g. in *onset detection* [41], *beat tracking* [42] or in *speech watermarking* in [43] where the idea was to embed the watermark data into the phase of unvoiced speech segments. The phase spectrum has also been shown to be useful in speech polarity determination [44] and detection of synthetic speech to avoid imposture in biometric system [45]. In *speech coding* and psychoacoustics, it was shown that the capacity of human perception due to phase is higher than expected, concluding that existing speech coders introduce certain distortions well perceived in particular for low-pitched voice [46]. The instantaneous higher order phase derivatives [47] and the phase information embedded in speech in [48] have been studied.

In the following, we consider the three aforementioned applications in Section 2 and describe the recent progress made towards estimation or incorporation of phase information.

3.1. Source Separation and Speech Enhancement

In single-channel speech enhancement or source separation, the issue of phase processing is an ill-conditioned problem to solve, even for two sources, and given the oracle spectral amplitude of the underlying sources in the mixed signal. This makes

the problem difficult and challenging, requiring additional constraints to solve. For example, recently, a phase estimation approach was proposed in [6] which relies on the geometry of interaction between the underlying signals and the property of group delay deviation to exhibit minimum at spectral peaks. Replacing the mixture phase with estimated phase in signal reconstruction of the separated signals led to improved perceived quality. In speech enhancement, it has been recently demonstrated that replacing the noisy phase with an estimated phase leads to improvement in the perceived quality [3, 5, 6].

As for the amplitude estimation part in modification stage, the phase importance in single-channel source separation was shown in [7]. The impact of phase in speech amplitude estimation has been recently investigated with positive outcomes [5, 49, 50]. The joint estimation of amplitude and phase spectrum in a closed-loop iterative configuration has been proposed in [3] and compared with the conventional methods using noisy phase or the open-loop configuration of [50] and upper-bound of amplitude estimation in [49] or signal reconstruction in [6].

In microphone array speech enhancement, the use of phase difference between microphones in dual-microphone was demonstrated to result in robust speech enhancement [51] and improved ASR [52]. The importance of phase information in speech enhancement has been studied extensively in [4, 53].

3.2. Automatic Recognition Systems

Extracting useful features from Fourier phase spectrum is not straightforward. This is due to the difficulties in phase wrapping, the dependency of phase spectrum on window starting sample and fast changes of phase spectrum when the zeros of complex spectrum $S(z)$ (calculated for a windowed speech $s(n)$ of limited support) lie near the unit circle in z-plane. In the literature, phase un-wrapping methods are studied and derivative of phase spectrum as group delay is employed which is less sensitive to phase wrapping issue [18, 54]. Several modifications on the group delay calculation are proposed to deal with the zeros of complex spectrum in preparing phase-based features for automatic speech recognition (ASR) [55]. Another approach to reduce the effect of zeros is to smooth the phase spectrum of *mixed-phase* speech signal before arriving at group delay function and next perform cesptral smoothing [56]. The application of instantaneous frequency (and it's deviation) along with delta-phase spectrum are considered for feature extraction in order to account for the large variability of phase caused by starting point of analysis window [57, 58].

As a bonus of utilizing phase-based features, there are several studies demonstrating the robustness of the group-delay based features against noise [59]. A common way to extract phase-derived features is by directly combining the features derived from amplitude and phase individually [60]. Features derived from Hilbert transform are considered as a way to utilize both amplitude and phase information in a unified way for speech [61] and speaker [62, 63] recognition.

3.3. Speech Synthesis

Currently, there are some attempts to introduce the phase spectrum (mixed phase) into the HMM-based text-to-speech synthesis systems, by suggesting the complex cepstrum approach [39]. In this case, both the magnitude and phase spectra are taken into account. Complex cepstra require phase unwrapping which implies a relatively high dimension Fourier transform. Phase unwrapping requires also to remove any linear phase component from the phase spectrum. For this reason, the estimation of complex cepstra is very sensitive to the position and type of analysis window. Especially for the position, an accurate estimation of the glottal closure instants is required [39]. If the paradigm of text-to-speech synthesis goes beyond HMMs (i.e., linear dynamical models, or deep neural networks), it may be possible to include the phase information without the current constraints put by the HMMs.

3.4. Useful Phase Representations

In the following, we present a list of phase representations derived from instantaneous phase spectrum, found useful in different speech applications. Taking the Fourier transformation from a segment of speech signal $s(n)$, the instantaneous phase at time index n and frequency bin k is indicated by $\theta_k(n)$.

Relative phase shift (RPS): The RPS relates the instantaneous phase of harmonic multiples with respect to the instantaneous phase of the fundamental frequency, and is given by

$$\text{RPS}_k(n) = \theta_k(n) - k\theta_1(n), \tag{1}$$

where θ_k and θ_1 refer to the instantaneous phase of the kth harmonic and fundamental frequency both calculated at time instant n, respectively, with k as the harmonic index. Recently, in [64], the RPS representation has been proposed for analysis, modification and synthesis of phase spectrum.

Time-frequency derivatives of phase: Group delay at time n and frequency bin k is defined as the frequency derivative of instantaneous phase

$$\tau_k(n) = \theta_{k+1}(n) - \theta_k(n). \tag{2}$$

The instantaneous frequency is defined as the time derivative of the instantaneous phase

$$\text{IF}_k(n) = \theta_k(n+1) - \theta_k(n). \tag{3}$$

Temporal derivative of phase was also used to derive instantaneous frequency deviation as a useful representation for displaying the pitch information in the form of fine harmonic detail [57]. IF deviation spectrum was shown to represent both pitch and formant structure similar to the magnitude spectrum. Similar representation was proposed in [65] termed as group delay deviation well observed to exhibit minima at spectral peaks [66]. The group delay deviation was used to solve the ambiguity problem in phase estimation in single-channel source separation [6] and speech enhancement [3, 50].

Phase dispersion: Considering a harmonic model with pitch-synchronous analysis [67], a segment of speech $s(n)$ is approximated by its harmonic representation given by:

$$s_h(n) = \sum_{k=1}^{K} A_k \cos(k\omega_0 n + \phi_k), \tag{4}$$

where A_k and ϕ_k are the amplitude and phase of the kth harmonic, K is the number of harmonics, and ω_0 is the fundamental frequency in radians. The phase argument inside the cosine term is denoted by $\theta_k(n)$ and can be decomposed to three parts [68]:

$$\theta_k(n) = \Phi[n, \Omega_k(n)] + k\omega_0\tau + \psi_k \tag{5}$$

where $\Phi[n, \Omega_k(n)]$ is the phase Fourier spectrum of the filter that is *minimum phase* with $\Omega_k(n)$ as the time-varying frequency, $k\omega_0\tau = \int_0^\tau \Omega_0(\sigma)d\sigma$ is the linear phase term with Ω_0 as time-varying fundamental frequency and ψ_k denotes the dispersion term. The linear phase accounts for the translation of

the center of the analysis window τ samples with respect to the time analysis instant. The dispersion phase ψ_k was shown to exhibit a structured pattern for voiced frames whose probability density function can be described using wrapped Gaussian mixture model [68]. Phase dispersion term was used to further define a phase distortion pitch cycle in [16] to quantify the level of quantization error occurred in speech coding.

Phase distortion: The phase distortion is obtained by removing the contributions of minimum phase and linear phase parts from the instantaneous phase, representing the shape of the glottal signal [69] and glottal model parameters [70].

4. Conclusion

This paper gives an overview of the recent scientific progress toward phase-based signal processing in different speech processing applications. The authors hope that the current review provides a quick start for new researchers interested to continue research in different speech applications using phase spectrum.

The Interspeech 2014 Special Session "Phase Importance in Speech Processing Applications" is organized by the authors in this paper aiming to fully consider the new progress in phase-based signal processing made in different speech applications. By gathering researchers focused on phase processing from different fields, the goal is to organize the required first steps to establish a new community of researchers for closer collaboration on the topic.

5. References

[1] A. Oppenheim and J. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529 – 541, May 1981.

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

[3] P. Mowlaee and R. Saeidi, "Iterative closed-loop phase-aware single-channel speech enhancement," *IEEE Signal Processing Letters*, vol. 20, no. 12, pp. 1235–1239, Dec. 2013.

[4] K. K. Paliwal, K. K. Wojcicki, and B. J. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[5] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Processing Letters*, vol. 20, no. 2, pp. 129 –132, Feb 2013.

[6] P. Mowlaee, R. Saiedi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," in *Proc. Interspeech*, 2012, pp. 1–4.

[7] P. Mowlaee and R. Martin, "On phase importance in parameter estimation for single-channel source separation," in *The International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[8] P. Mowlaee and M. Watanabe, "Iterative sinusoidal-based partial phase reconstruction in single-channel source separation," in *Proc. Interspeech*, 2013, pp. 832–836.

[9] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE signal processing letters*, vol. 20, no. 3, pp. 217 – 220, 2013.

[10] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 421 –424, May 2010.

[11] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, pp. 133–136.

[12] T. Kleinschmidt, S. Sridharan, and M. Mason, "The use of phase in complex spectrum subtraction for robust speech recognition," *Computer Speech and Language*, vol. 25, no. 3, pp. 585 – 600, 2011.

[13] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the Modified Group Delay Feature in Speech Recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 190–202, Jan. 2007.

[14] I. Hernaez, I. Saratxaga, J. Sanchez, E. Navas, and I. Luengo, "Use of the harmonic phase in speaker recognition." in *Proc. Interspeech*, 2011, pp. 2757–2760.

[15] P. Rajan, T. Kinnunen, C. Hanilci, J. Pohjalainen, and P. Alku, "Using group delay functions from all-pole models for speaker recognition," in *INTERSPEECH*, 2013.

[16] H. Pobloth and W. B. Kleijn, "Squared error as a measure of perceived phase distortion," *J. Acoust. Soc. Am.*, vol. 114, no. 2, pp. 1081–1094, 2003.

[17] H. A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function," *speech communication*, vol. 10, no. 3, pp. 209 – 221, 1991.

[18] B. Yegnanarayana, J. Sreekanth, and A. Rangarajan, "Waveform estimation using group delay processing," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 33, no. 4, pp. 832 – 836, Aug 1985.

[19] G. Degottex, A. Roebel, and X. Rodet, "Phase minimization for glottal model estimation," *IEEE Trans. on Acoustics, Speech and Language Processing*, vol. 19, no. 5, pp. 1080–1090, July 2011.

[20] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[21] G. J. Virtanen, T., B. Raj, and P. Smaragdis, "Compositional models for audio processing," *accepted to IEEE Signal Processing Magazine*, 2014.

[22] S. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems (NIPS)*, 2000, pp. 793–799.

[23] A. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1766–1776, Aug 2007.

[24] P. Mowlaee, M. Christensen, and S. Jensen, "New results on single-channel speech separation using sinusoidal modeling," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 5, pp. 1265 – 1277, 2011.

[25] P. Mowlaee, R. Saeidi, M. Christensen, Z.-H. Tan, T. Kinnunen, P. Franti, and S. Jensen, "A joint approach for single-channel speaker identification and speech separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2586 –2601, Nov. 2012.

[26] M. Radfar, A. H. Banihashemi, R. Dansereau, and A. Sayadiyan, "Nonlinear minimum mean square error estimator for mixture-maximisation approximation," *Electronics Letters*, vol. 42, no. 12, pp. 724–725, June 2006.

[27] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 45–66, Jan 2010.

[28] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[29] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2008, pp. 4037–4040.

[30] M. Hayes, J. Lim, and A. Oppenheim, "Signal reconstruction from phase or magnitude," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 28, no. 6, pp. 672 – 680, 1980.

[31] T. Quatieri and A. Oppenheim, "Iterative techniques for minimum phase signal reconstruction from phase or magnitude," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29, no. 6, pp. 1187 – 1193, Dec. 1981.

[32] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236 – 243, apr 1984.

[33] L. D. Alsteris and K. K. Paliwal, "Iterative reconstruction of speech from short-time Fourier transform phase and magnitude spectra," *Computer Speech & Language*, vol. 21, no. 1, pp. 174–186, 2007.

[34] F. Charpentier, "Pitch detection using the short-term phase spectrum," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11, 1986, pp. 113–116.

[35] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[36] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 599–602, 2010.

[37] T. Kinnunen, R. Saeidi, F. Sedlak, K. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li, "Low-Variance Multitaper MFCC Features: a Case Study in Robust Speaker Verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1990–2001, 2012.

[38] Y. Stylianou, "Removing linear phase mismatches in concatenative speech synthesis," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 3, pp. 232–239, Mar 2001.

[39] R. Maia, M. Akamine, and M. J. F. Gales, "Complex cepstrum as phase information in statistical parametric speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4581–4584.

[40] A. Rad and T. Virtanen, "Phase spectrum prediction of audio signals," in *International Symposium on Communications Control and Signal Processing (ISCCSP)*, May 2012, pp. 1 –5.

[41] A. Holzapfel and Y. Stylianou, "Beat tracking using group delay based onset detection." in *ISMIR*, 2008, pp. 653–658.

[42] M. E. P. Davies and M. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1009–1020, March 2007.

[43] K. Hofbauer, G. Kubin, and W. Kleijn, "Speech watermarking for analog flat-fading bandpass channels," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1624–1637, Nov. 2009.

[44] I. Saratxaga, D. Erro, I. Hernez, I. Sainz, and E. Navas, "Use of harmonic phase information for polarity detection in speech signals." in *INTERSPEECH*, 2009.

[45] P. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 4844–4847.

[46] H. Pobloth and W. Kleijn, "On phase perception in speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1999, pp. 29–32.

[47] D. J. Nelson, "Instantaneous higher order phase derivatives," *Digital Signal Processing*, vol. 12, no. 23, pp. 416 – 428, 2002.

[48] S. A. Fulop, *Speech Spectrum Analysis*. Springer, 2011.

[49] P. Mowlaee and R. Saeidi, "On phase importance in parameter estimation in single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2013, pp. 7462–7466.

[50] P. Mowlaee, M. Watanabe, and R. Saeidi, "Show & tell: Phase-aware single-channel speech enhancement," in *Proc. Interspeech*, 2013, pp. 1–4.

[51] G. Shi, P. Aarabi, and J. Hui, "Phase-based dual-microphone speech enhancement using a prior speech model," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 109–118, Jan 2007.

[52] S. Guangji, M. Shanechi, and P. Aarabi, "On the importance of phase in human speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1867 –1874, Sept. 2006.

[53] L. D. Alsteris and K. K. Paliwal, "Further intelligibility results from human listening tests using the short-time phase spectrum," *Speech Communication*, vol. 48, no. 6, pp. 727 – 736, 2006.

[54] G. Nico and J. Fortuny, "Using the matrix pencil method to solve phase unwrapping," *IEEE Trans. on Signal Processing*, vol. 51, no. 3, pp. 886–888, March 2003.

[55] B. Bozkurt and L. Couvreur, "On the use of phase information for speech recognition," in *Proc. EUSIPCO*, 2005, pp. 2–5.

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

[57] A. P. Stark and K. K. Paliwal, "Speech analysis using instantaneous frequency deviation," in *Proc. Interspeech*, 2008, pp. 22–26.

[58] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, "The delta-phase spectrum with application to voice activity detection and speaker recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2026–2038, 2011.

[59] S. H. Parthasarathi, R. Padmanabhan, and H. A. Murthy, "Robustness of group delay representations for noisy speech signals," *International Journal of Speech Technology*, vol. 14, no. 4, pp. 361–368, 2011.

[60] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 133–136.

[61] E. Loweimi, S. Ahadi, and T. Drugman, "A new phase-based feature representation for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 7155–7159.

[62] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, 2006.

[63] S. Sadjadi and J. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5448–5451.

[64] I. Saratxaga, I. Hernaez, D. Erro, E. Navas, and J. Sanchez, "Simple representation of signal phase for harmonic speech models," *Electronics Letters*, vol. 45, no. 7, pp. 381 –383, 2009.

[65] A. P. Stark and K. K. Paliwal, "Group-delay-deviation based spectral analysis of speech," in *Proc. Interspeech*, 2009, pp. 1083–1086.

[66] B. Yegnanarayana and H. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Trans. on Signal Processing*, vol. 40, no. 9, pp. 2281–2289, Sep 1992.

[67] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive am-fm signal decomposition with application to speech analysis," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 290–300, Feb 2011.

[68] Y. Agiomyrgiannakis and Y. Stylianou, "Wrapped gaussian mixture models for modeling and high-rate quantization of phase data of speech," *IEEE Trans. on Audio, Speech & Language Processing*, vol. 17, no. 4, pp. 775–786, 2009.

[69] G. Degottex, E. Godoy, and Y. Stylianou, "Identifying tenseness of lombard speech using phase distortion," in *Proc. The Listening Talker: An interdisciplinary workshop on natural and synthetic modification of speech in response to listening conditions*, Edinburgh, UK, May 2012, p. 60.

[70] G. Degottex, A. Roebel, and X. Rodet, "Function of phase-distortion for glottal model estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2011, pp. 4608–4611.