
Proceedings Paper:
Loweimi, E., Barker, J. orcid.org/0000-0002-1684-5660 and Hain, T. orcid.org/0000-0003-0939-3464 (2015) Source-filter Separation of Speech Signal in the Phase Domain. In: 16TH ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION (INTERSPEECH 2015), VOLS 1-5. Interspeech 2015, 06-10 Sep 2016, Dresden, Germany. ISCA , pp. 598-602. ISBN 978-1-5108-1790-6

Source-filter Separation of Speech Signal in the Phase Domain

Erfan Loweimi, Jon Barker, Thomas Hain

Speech and Hearing Research Group (SPandH), University of Sheffield, Sheffield, UK

{e.loweimi, j.barker, t.hain}@dcs.shef.ac.uk

Abstract

Deconvolution of the speech excitation (source) and vocal tract (filter) components through log-magnitude spectral processing is well-established and has led to the well-known cepstral features used in a multitude of speech processing tasks. This paper presents a novel source-filter decomposition based on processing in the phase domain. We show that separation between source and filter in the log-magnitude spectra is far from perfect, leading to loss of vital vocal tract information. It is demonstrated that the same task can be better performed by trend and fluctuation analysis of the phase spectrum of the minimum-phase component of speech, which can be computed via the Hilbert transform. Trend and fluctuation can be separated through low-pass filtering of the phase, using additivity of vocal tract and source in the phase domain. This results in separated signals which have a clear relation to the vocal tract and excitation components. The effectiveness of the method is put to test in a speech recognition task. The vocal tract component extracted in this way is used as the basis of a feature extraction algorithm for speech recognition on the Aurora-2 database. The recognition results shows upto 8.5% absolute improvement in comparison with MFCC features on average (0-20dB).

Index Terms: Speech phase spectrum, Source-filter decomposition, Hilbert transform, phase wrapping, minimum-phase component, Trend/Fluctuation analysis

1. Introduction

Phase spectrum of the speech signal has a very complicated behaviour. Its shape is noise-like and lacking in any meaningful trend or extremum points making it hard to model or interpret. Further, there are several studies that suggest it is unimportant from a perceptual point of view [1–4]. Although there are notable works that have exploited the phase [5–20], it is still a common practice to discard this spectrum after the Fourier analysis. Experimental studies have lacked a fundamental mathematical foundation such that would allow measurement of the information encoded by the phase spectrum. Useful reviews about the phase spectrum may be found in [21], [22] and [23].

In the case of the magnitude spectrum, the source-filter model is commonly employed to decompose the speech signal into its fundamental elements, i.e. the vocal tract (filter) and the excitation (source). The theory of such deconvolution based on cepstral (log-magnitude) processing is well-established and most speech processing methods make use of it in some way [24, 25]. However, in the case of the phase spectrum there is no algorithm for direct phase modelling to facilitate such decomposition, despite the fact that convolution in the time domain is equivalent to addition in the phase domain, which should potentially be useful for carrying out deconvolution. So, developing a basic phase-based source-filter model, could pave the way for putting the speech phase spectrum to practical use and will lend support to the experimental results already reported.

In this paper, we present a novel method for source-filter modelling which enables the blind deconvolution of the speech using purely phase-based processing. This model provides insight into the way in which information is encoded in the phase spectrum and sheds light on the behaviour of the phase across the spectrum. The capability of the method to follow the temporal evolution of the vocal tract configuration and the excitation component will be demonstrated and discussed. In addition, we propose a parametrization method to demonstrate the efficacy of the suggested method in a speech recognition task.

The rest of this paper is structured as follows. Section 2 reviews the main issues with the phase spectrum-based analysis and modelling. In Section 3 the proposed method is introduced and investigated from different aspects. Section 4 presents the feature extraction process and noise-robust speech recognition results. Section 5 concludes the paper.

2. Properties of the phase spectrum

In speech processing, the magnitude spectrum plays the major role because of its harmony with our understanding of the physical characteristics of the speech production system and its benign behaviour from the mathematical standpoint. In particular, source (fundamental frequency) and filter (formant frequencies) elements, as the building blocks of the speech signal, can be obtained straightforwardly by decomposition in the cepstral domain. As in the cepstral domain, these two components are also additive in the phase domain. However, signal processing through phase manipulation is overwhelmingly difficult because of *phase wrapping*. This phenomenon is due to the $atan2(.,.)$ function and gives the phase spectrum a chaotic shape, lacking any meaningful trend or extremum points. As a result, direct interpretation and modelling of this spectrum becomes extremely complicated. To produce a more understandable representation of the phase, researchers resort to working with its derivative, i.e. the group delay function (GDF).

2.1. Group Delay Function

Most of the phase-related works in speech processing are based on the GDF, $\tau_X(\omega)$, which is defined as follows

$$\tau_X(\omega) = -\frac{d}{d\omega} arg[X(\omega)] = -Im\{\frac{d}{d\omega}\log(X(\omega))\} \quad (1)$$

where $arg[.]$ and $Im\{.\}$ denote the unwrapped (continuous) phase and imaginary part and ω is angular frequency. Phase unwrapping is not straightforward [26] but the GDF can be computed while avoiding this issue by utilizing real and imaginary parts,

$$\tau_X(\omega) = \frac{X_{Re}(\omega)Y_{Re}(\omega) + X_{Im}(\omega)Y_{Im}(\omega)}{|X(\omega)|^2} \quad (2)$$

where $Y(\omega)$ is the Fourier transform of $nx[n]$ and the Re subscript denotes the real part.

2.1.1. Pros and cons of working with GDF

There are several benefits of working with the GDF. First, under certain conditions, it has a relatively clear behaviour and resembles the magnitude spectrum. The second advantage is *additivity*, which means if two signals are convolved in the time domain their GDFs will be added. Third, it has a high frequency resolution, i.e. it is able to resolve closely located peaks in the frequency domain.

However, on the negative side, when the magnitude spectrum (denominator of Eq. (2)) gets close to zero, the GDF becomes very spiky and this will highly limit its usefulness. In the case of speech signals, the excitation component gives rise to such zeros and this, to a great extent, restricts the applicability of the GDF.

2.1.2. Proposed Solutions

Four methods have been proposed in the literature for dealing with the GDF spikiness, namely modified GDF (MODGDF) [27–29], chirp GDF (CGDF) [30], product spectrum (PS) [31], and model-based GDF [12, 14, 32, 33]. Despite the relative success of these techniques in tackling the aforementioned problem and returning an estimate of the vocal tract configuration, none of them provides insight into the way in which the source and filter components interact nor how such information is encoded in the phase spectrum. In fact, each one proposes a solution for alleviating the unfavourable effect of the denominator of Eq. (2), when $|X(\omega)|$ approaches zero: MODGDF, replaces the denominator with the cepstrally smoothed power spectrum; CGDF tries to move away from the unit circle by chirp processing; in the PS method the denominator is substituted with unity; and in the model-based case the autoregressive (AR) model is extracted and its group delay is computed.

2.1.3. Shortcoming of the Solutions

By suppressing the excitation element, these methods can not be successfully employed in applications where this component is of some significance. However, the more important issue with the GDF is that although the GDFs of the source and filter components are additive, decomposing the speech into these two parts through GDF-based processing is problematic in principle. This point will be explained in Section 3.3.2 but its direct implication is that the additive property of the GDF for doing deconvolution becomes practically ineffective.

3. Phase Spectrum Modelling

3.1. Preliminary

Speech is a *mixed-phase* signal [34] owing to its complex cepstrum being neither causal nor anticausal [24]; hence it can be decomposed into two complementary components, namely *minimum-phase (MinPh)*, $X_{MinPh}(\omega)$, and *all-pass (AllP)*, $X_{AllP}(\omega)$,

$$X(\omega) = |X(\omega)|.e^{j.arg[X(\omega)]} = X_{MinPh}(\omega).X_{AllP}(\omega) \quad (3)$$

$$\begin{cases} |X(\omega)| = |X_{MinPh}(\omega)| \\ arg[X(\omega)] = arg[X_{MinPh}(\omega)] + arg[X_{AllP}(\omega)] \end{cases} \quad (4)$$

As seen, the magnitude spectrum is only related to the MinPh part of the speech whereas the phase spectrum is linked to both the MinPh and AllP components. Since both vocal tract ($X_{VT}(\omega)$) and excitation ($|X_{Exc}(\omega)|$) elements manifest themselves in the magnitude spectrum, the minimum-phase component can be expressed as follows

$$\begin{cases} |X_{MinPh}(\omega)| = |X_{VT}(\omega)|.|X_{Exc}(\omega)| \\ arg[X_{MinPh}(\omega)] = arg[X_{VT}(\omega)] + arg[X_{Exc}(\omega)]. \end{cases} \quad (5)$$

Therefore, our goal, i.e. source-filter modelling in the phase domain, is to compute $arg[X_{VT}(\omega)]$ and $arg[X_{Exc}(\omega)]$. To do so, first we need to compute the minimum-phase component of the $X(\omega)$, namely $X_{MinPh}(\omega)$.

3.2. Computing the minimum-phase component

For recovering the minimum-phase component of a mixed-phase signal two approaches can be considered, parametric and non-parametric. In the parametric case, the z-transform of the sequence should be expressed in a rational form and all the poles and zeros which are located outside the unit circle are reflected inside. This method is not practical since the z-transform of an autoregressive-moving-average model is not available in practice. On the other hand, the non-parametric approach does not assume a particular form for the z-transform and takes advantage of the complex cepstrum properties. For the minimum-phase signals the complex cepstrum is causal, i.e. it equals zero at the negative quefrencies [24]. For such signals, the Hilbert transform provides a one-to-one relationship between the magnitude and phase spectra as follows

$$arg[X_{MinPh}(\omega)] = -\frac{1}{2\pi} log|X_{MinPh}(\omega)| * cot(\frac{\omega}{2}) \quad (6)$$

By replacing the $log|X_{MinPh}(\omega)|$ with $log|X(\omega)|$ based on Eq. (4), $arg[X_{MinPh}(\omega)]$ can be computed. Alternatively, the minimum-phase component can be calculated by putting a proper lifter on the cepstrum sequence. It should be noted that both of these approaches are very closely related, one operates in the frequency domain and one acts in the quefrency domain. If the independent variable was continuous both would return identical results, however, due to having discrete independent variables, the outcome of these two approaches would not be exactly the same. Fig 2(h) shows this slight difference which is due to the behaviour of the $cot(.)$ function as it tends to infinity at the edges. By increasing the FFT length such error reduces. We will use the cesptrum-based method henceforth because of its better numerical accuracy.

3.3. Source-Filter Modelling

Now, $arg[X_{MinPh}(\omega)]$ should be decomposed into excitation and vocal tract components. To this end, some prior knowledge is required. However, apart from the source/filter issue, by looking at the $arg[X_{MinPh}(\omega)]$ (Fig 2(h)), it can be imagined as a modulated signal with two major ingredients, namely carrier and message. The former varies fast and the later changes more slowly with respect to the independent variable, i.e. frequency. Based on this argument $arg[X_{MinPh}(\omega)]$ may be expressed as follows

$$arg[X_{MinPh}] = Trend + Fluctuation \quad (7)$$

where the $Trend$ element is the slowly-varying aspect and the $Fluctuation$ relates to the rapidly changing component [35].

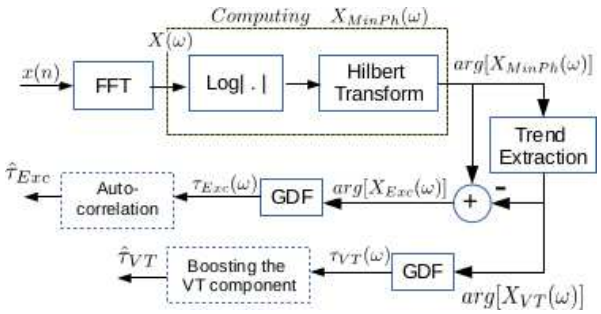

Figure 1: Workflow of the phase-based speech source/filter decomposition.

Comparing Eq. (7) with the way in which vocal tract and excitation components are combined in the log-magnitude spectrum, it is expected that the Trend is associated with the vocal tract and Fluctuation pertains to the excitation part of the phase.

3.3.1. *Trend/Fluctuation decomposition*

The underlying premise of Eq. (7) is that the two components have a different rate of change. If the phase sequence is assumed to be a time series for the sake of discussion, a slower element i.e. Trend will mainly occupy the low frequencies after computing the Fourier transform and can be recovered by low-pass filtering (Here, we have used a brickwall lowpass filter with 20 taps). Having extracted the Trend, by a simple subtraction, the Fluctuation component may be computed.

3.3.2. *Comparing additivity in the phase and GDF domains*

It should be noted that additivity is not a sufficient condition for performing such decomposition. In fact, such additivity holds for the GDFs of the vocal tract and excitation parts, too. The necessary condition for successful decomposition is the different variation rates of the so-called Trend and Fluctuation elements which, ideally, leads them to occupying non-overlapping frequency regions after Fourier analysis (recall that for the sake of discussion they are assumed to be time series). The greater the overlap between their supports in the frequency domain, the less effective the decomposition will be. When computing the derivative to get the GDF, the support of the Trend component expands toward high frequencies because its high frequency components, although weak, are not exactly zero. The derivative magnifies them linearly with frequency, i.e. the higher the frequency the greater the amplification,

$$\begin{cases} \tau_X(t) = -\frac{d}{dt} arg[X_{MinPh}(t)] = -\frac{d}{dt} Trend - \frac{d}{dt} Fluctuation \\ \mathcal{F}\{\tau_X(t)\} = -j\omega\mathcal{F}\{Trend\} - j\omega\mathcal{F}\{Fluctuation\} \end{cases} \quad (8)$$

where $\mathcal{F}\{.\}$ denotes the Fourier transform and t notifies time (as the phase sequence was assumed to be a time series for the sake of discussion). As a result, the overlap between the corresponding supports of the Trend and Fluctuation in the frequency domain increases and the efficacy of the decomposition after truncation substantially decreases. As such, the GDF of the vocal tract (Trend) would have high-frequency components with noticeable energy and blocking them with a low-pass filter will result in significant error. Therefore, the additive property of the GDF is not practically functional.

3.3.3. *Postprocessing*

To have a better representation of the data lying in the Filter and Source components of the phase we added a post-processing block to each branch. In the case of the excitation, in order to capture the periodicity more efficiently, autocorrelation of the $\tau_{Exc}(\omega)$ is computed. As seen in Fig 2(j), it allows for a better fundamental frequency estimation.

$$\hat{\tau}_{Exc}(\omega) = Autocorrelation\{\tau_{Exc}(\omega)\} \quad (9)$$

In the case of the vocal tract branch, the formant peaks were boosted using an approach similar to [28], i.e.

$$\begin{cases} \hat{\tau}_{VT}(\omega) = signum(\tau_{VT}(\omega)).|\tau_{VT}(\omega)|^{\alpha} \\ signum(\tau_{VT}(\omega)) = \frac{\tau_{VT}(\omega)}{|\tau_{VT}(\omega)|} \end{cases} \quad (10)$$

where α can be considered as a peak-boosting factor and should be less than 1. Figs. 2(i) and 3(e) depict the effect of this factor when it is set to 0.7. As seen, it adjusted the dynamic range and the bandwidth of the formants.

Fig. 1 shows the block diagram of the phase-based source/filter decomposition of a speech frame, $x[n]$. Fig. 2 illustrates different representations of a typical speech waveform and Fig. 3 depicts the spectrogram of the source and filter components computed through the proposed method. As seen, the suggested method succeeds in deconvolving the speech into source and filter elements, directly models the phase spectrum behaviour and clarifies the way in which information is encoded in the phase spectrum of the speech signal.

3.4. Comparison with the magnitude-based approach

The advantages of the proposed method compared with the magnitude-based approach are twofold: higher frequency res-

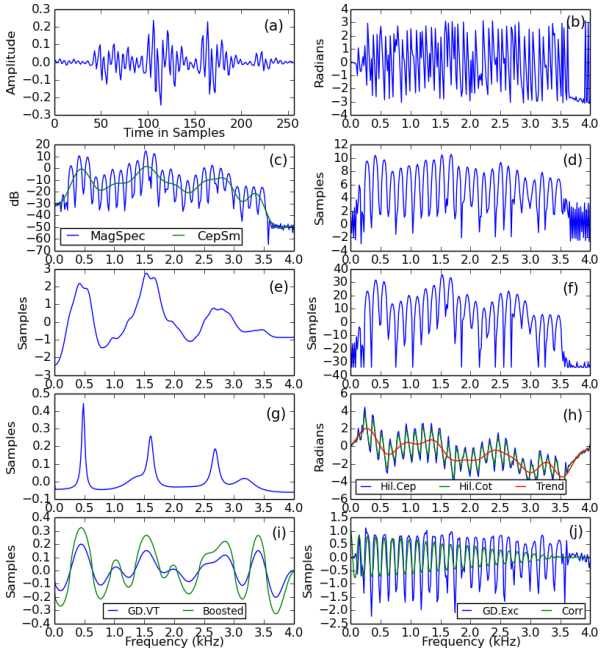

Figure 2: Different representations of a speech signal. (a) waveform, (b) wrapped phase spectrum ($ARG[X(\omega)]$), (c) magnitude spectrum and its cepstrally smoothed version, (d) MODGDF, (e) CGDF, (f) PS, (g) GDF of AR model (order 13), (h) $arg[X_{MinPh}(\omega)]$, (i) $\tau_{VT}(\omega)$ (Filter), (j) $\tau_{Exc}(\omega)$ (Source).

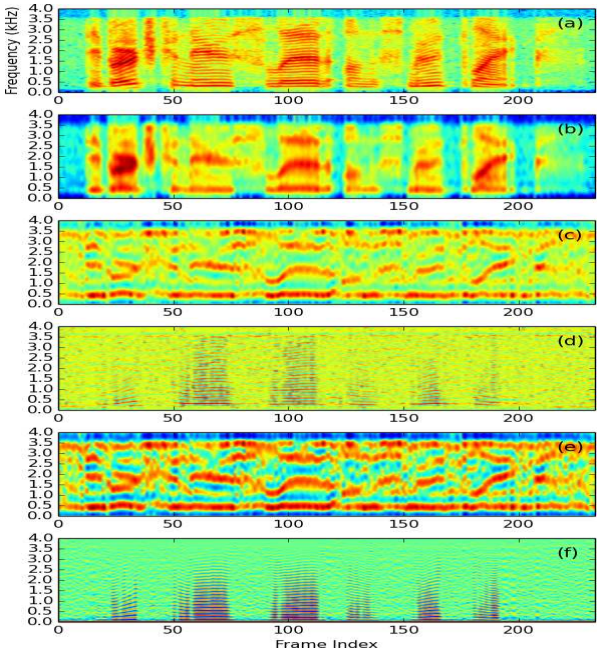

Figure 3: Spectrograms of the Source and Filter components based on the proposed method for $sp10$ from NOIZEUS corpus [36]. (a) magnitude spectrum, (b) its cepstrally smoothed version, (c) $\tau_{VT}(\omega)$, (d) $\tau_{Exc}(\omega)$, (e) $\hat{\tau}_{VT}(\omega)$, (f) $\hat{\tau}_{Exc}(\omega)$.

olution and less frequency leakage. As seen in Fig. 2 (c) and (i), even without boosting, it has a higher capability in resolving the local peaks of the envelope. Comparing the spectrograms of Fig. 3 (b) with (c) and (e), clearly shows that the formants are more distinct in the filter component estimated via $arg[X_{VT}(\omega)]$. Such better distinction of the formants potentially implies that this representation affords better phoneme discrimination. It should be noted that in the magnitude-based processing there is a trade-off between the resolution and leakage and both of them can not be improved simultaneously.

4. Experimental Results

4.1. Feature Extraction

In order to investigate the effectiveness of the proposed method, we used it as the basis of a feature extractor for speech recognition and compared its performance with other well-known features. Four simple approaches were taken for turning the filter component into a feature for ASR

i) $arg[X_{VT}] \rightarrow DCT \Rightarrow PHVT$

ii) $\tau_{VT} \rightarrow DCT \Rightarrow GDVT$

iii) $\hat{\tau}_{VT} \rightarrow MelFilterbank \rightarrow DCT \Rightarrow MFGDVT$

iv) $\hat{\tau}_{VT} \rightarrow Mel\ Filterbank \rightarrow Boost \rightarrow DCT \Rightarrow BMFGDVT$

where the rightmost name is the name assigned to each feature derived from the filter component of the phase spectrum.

For all features, we have used the default parameters reported in their respective publications. Frame length, frame shift and number of filters were set to 25 ms and 10 ms and 23, respectively. Feature vectors have been augmented by log-energy as well as delta and delta-delta coefficients. Cepstral mean normalization (CMN) is performed. In the proposed method, α coefficient (Eq. (10)) was set to 0.7 and brickwall lowpass filter with 20 taps was used for extracting the Trend. Hamming window were applied for all features except for the proposed methods where Chebyshev (30 dB) window has been employed due to [15]. Aurora-2 [37] has been used as the database and the HMMs were trained from the clean data using HTK [38] based on the Aurora-2 standard recipe. Recognition results (average 0-20 dB) of the different magnitude-based and phase-based features are reported in Table 1.

4.2. Discussion

As seen, the proposed method outperforms the other magnitude and phase-based features despite its relative simplicity. In particular, the fact that it provides better performance than power normalized cepstral coefficients (PNCC) [39], is remarkable given the complexity of the PNCC. The relative robustness of the proposed method can be explained in part by the greater distinction of the formants arising from there being higher frequency resolution and less leakage which, in turn, decreases the confusion occurring after SNR reduction in the spectrum. Another important factor should be noted: As mentioned, the vocal tract component corresponds with the trend of the MinPh parts phase spectrum. Each point of the Trend may be considered as a pseudo-mean of the neighbouring points in the vicinity. This, in turn, increases the inertia of each of the Trend's individual points. As a result, the tendency for a point to preserve its initial condition and resist disturbance is higher. This reduces the sensitivity to the noise and increases the feature's robustness.

Table 1: *Average (0-20dB) recognition rates Aurora-2 [37].*

Feature	TestSet A	TestSet B	TestSet C
MFCC	66.2	71.4	64.9
PLP	67.3	70.6	66.2
PNCC	71.2	72.8	71.5
MODGDF	64.3	66.4	59.5
CGDF	67.0	73.0	59.4
PS	66.0	71.2	64.6
i) PHVT	69.0	74.8	67.1
ii) GDVT	70.5	75.9	69.1
iii) MFGDVT	72.8	77.3	72.8
iv) BMFGDVT	**73.2**	**77.4**	**73.4**

5. Conclusions

In this paper we proposed a novel phase-based method for the source-filter decomposition of speech. The minimum-phase part of the signal was computed through a Hilbert transform. Vocal tract and excitation components were then successfully deconvolved by Trend/Fluctuation analysis of the corresponding phase spectrum. This analysis clarifies the behaviour of the the phase spectrum and the way in which it encodes information. The efficiency of the method in comparison with performing the same process in the log-magnitude (cepstrum) and GDF domains was discussed and illustrated. In addition, the vocal tract component was transformed into a feature vector and its effectiveness was evaluated in speech recognition. Recognition results demonstrate the efficacy of the proposed method. Notably, despite its simplicity, it outperforms recent robust feature extraction techniques such as PNCC. Given the centrality of the source-filter modelling, the proposed method paves the way for putting the phase spectrum to practical use by providing a foundation for speech signal processing through phase manipulation. Further optimization and studying the statistical behaviour of the proposed phase-based representations is a broad avenue for future research.

6. References

[1] A. Duff, E. Lewis, C. Mendenhall, A. Carman, R. McClung, and W. Hallock, *A Text-book of Physics*, ser. Blakiston's science series. P. Blakiston's son & Company, 1912.

[2] H. v. Helmholtz and A. J. Ellis, *On the sensations of tone as a physiological basis for the theory of music / by Herman L.F. Helmholtz*, 2nd ed. Longmans, Green London, 1885.

[3] M. R. Schroeder, "Models of hearing," *Proceedings of the IEEE*, vol. 63, no. 9, pp. 1332–1350, Sept 1975.

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

[5] K. K. Paliwal and L. D. Alsteris, "Usefulness of phase spectrum in human speech perception." in *INTERSPEECH*. ISCA, 2003.

[6] K. Paliwal and L. Alsteris, "On the usefulness of stft phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 2, pp. 153–170, 2005, cited By (since 1996)39.

[7] K. Paliwal, K. Wjcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465 – 494, 2011.

[8] L. Alsteris and K. Paliwal, "Importance of window shape for phase-only reconstruction of speech," vol. 1, 2004, pp. I573–I576, cited By (since 1996)12.

[9] ——, "Further intelligibility results from human listening tests using the short-time phase spectrum," *Speech Communication*, vol. 48, no. 6, pp. 727–736, 2006.

[10] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Communication*, vol. 22, no. 4, pp. 403–417, 1997.

[11] T. Gerkmann, M. Krawczyk, and R. Rehr, "Phase estimation in speech enhancement – unimportant, important, or impossible?" in *Electrical Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of*, Nov 2012, pp. 1–5.

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

[13] E. Loweimi, S. Ahadi, and S. Loveymi, "On the importance of phase and magnitude spectra in speech enhancement," in *Electrical Engineering (ICEE), 2011 19th Iranian conference on*, May 2011, pp. 1–1.

[14] E. Loweimi and S. Ahadi, "A new group delay-based feature for robust speech recognition," in *Multimedia and Expo (ICME), 2011 IEEE International conference on*, July 2011, pp. 1–5.

[15] E. Loweimi, S. M. Ahadi, and H. Sheikhzadeh, "Phase-only speech reconstruction using very short frames." in *INTERSPEECH*. ISCA, 2011, pp. 2501–2504.

[16] E. Loveimi and S. Ahadi, "Objective evaluation of phase and magnitude only reconstructed speech: New considerations," in *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International conference on*, May 2010, pp. 117–120.

[17] H. Pobloth and W. Kleijn, "On phase perception in speech," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International conference on*, vol. 1, Mar 1999, pp. 29–32 vol.1.

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

[19] P. Mowlaee and R. Saeidi, "On phase importance in parameter estimation in single-channel speech enhancement," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7462–7466.

[20] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, vol. 1, 2001, pp. 133–136 vol.1.

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

[22] 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.

[23] *INTERSPEECH 2014 Special Session on Phase Importance in Speech Processing Applications*, 2014.

[24] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2009.

[25] J. Deller, J. Hansen, and J. Proakis, *Discrete-Time Processing of Speech Signals*, ser. An IEEE Press classic reissue. Wiley, 2000. [Online]. Available: http://books.google.co.uk/books?id=lKUeAQAAIAAJ

[26] A. Baldi, F. Bertolino, and F. Ginesu, "Phase unwrapping algorithms: A comparison," in *Interferometry in Speckle Light*, P. Jacquot and J.-M. Fournier, Eds. Springer Berlin Heidelberg, 2000, pp. 483–490.

[27] B. Yegnanarayana and H. A. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Transactions on Signal Processing*, vol. 40, no. 9, pp. 2281–2289, 1992, cited By (since 1996)52.

[28] H. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International conference on*, vol. 1, April 2003, pp. I–68–71 vol.1.

[29] R. Hegde, H. Murthy, and V. Gadde, "Significance of the modified group delay feature in speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 190–202, Jan 2007.

[30] B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp group delay analysis of speech signals," *Speech Communication*, vol. 49, no. 3, pp. 159 – 176, 2007.

[31] D. Zhu and K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International conference on*, vol. 1, May 2004, pp. I–125–8 vol.1.

[32] Y. B., "Formant extraction from linear prediction phase spectra," *Journal of the Acoustical Society of America (JASA)*, vol. 63, no. 5, pp. 1638 – 1640, May 1978.

[33] V. Sethu, E. Ambikairajah, and J. Epps, "Group delay features for emotion detection." in *INTERSPEECH*. ISCA, 2007, pp. 2273–2276.

[34] T. Drugman, B. Bozkurt, and T. Dutoit, "Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Commun.*, vol. 53, no. 6, pp. 855–866, Jul. 2011.

[35] G. Box, G. Jenkins, and G. Reinsel, *Time Series Analysis: Forecasting and Control*, ser. Wiley Series in Probability and Statistics. Wiley, 2013.

[36] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7-8, pp. 588–601, Jul. 2007.

[37] D. Pearce and H.-G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions." in *INTERSPEECH*. ISCA, 2000, pp. 29–32.

[38] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge University Press, 2006.

[39] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition." in *ICASSP*. IEEE, 2012, pp. 4101–4104.