
Proceedings Paper:
Loweimi, E., Barker, J. orcid.org/0000-0002-1684-5660 and Hain, T. orcid.org/0000-0003-0939-3464 (2018) On the usefulness of the speech phase spectrum for pitch extraction. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. Interspeech 2018, 02-06 Sep 2018, Hyderabad, India. ISCA , pp. 696-700.

10.21437/Interspeech.2018-1062

Interspeech 2018
2-6 September 2018, Hyderabad

On the Usefulness of the Speech Phase Spectrum for Pitch Extraction

Erfan Loweimi, Jon Barker and Thomas Hain

Speech and Hearing Research Group (SPandH), University of Sheffield
{e.loweimi, j.p.barker, t.hain}@sheffield.ac.uk

Abstract

Most frequency domain techniques for pitch extraction such as cepstrum, harmonic product spectrum (HPS) and summation residual harmonics (SRH) operate on the magnitude spectrum and turn it into a function in which the fundamental frequency emerges as *argmax*. In this paper, we investigate the extension of these three techniques to the phase and group delay (GD) domains. Our extensions exploit the observation that the bin at which $F(magnitude)$ becomes maximum, for some monotonically increasing function F, is equivalent to bin at which $F(phase)$ has maximum negative slope and $F(group\,delay)$ has the maximum value. To extract the pitch track from speech phase spectrum, these techniques were coupled with the source-filter model in the phase domain that we proposed in earlier publications and a novel voicing detection algorithm proposed here. The accuracy and robustness of the phase-based pitch extraction techniques are illustrated and compared with their magnitude-based counterparts using six pitch evaluation metrics. On average, it is observed that the phase spectrum can be successfully employed in pitch tracking with comparable accuracy and robustness to the speech magnitude spectrum.

Index Terms: pitch extraction, voicing detection, phase spectrum, group delay, source-filter separation

1. Introduction

The excitation component of the speech signal can either be a quasi-periodic train of laryngeal pulses or it may take a noise-like form [1]. Detecting the presence of the periodicity and quantifying the period value, are referred to as voicing detection and pitch extraction. These problems can be dealt with in the time/frequency domains by converting the wave/spectrum into a function where the fundamental frequency (F_0) emerges as *argmin* or *argmax*. Most frequency domain techniques utilise the speech magnitude spectrum, and turn it into a function at which the pitch[1] emerges as an argmax. Examples of these functions are cepstrum [2], harmonic product spectrum (HPS) [3] and summation residual harmonic (SRH) [4]. This paper aims to extend these functions to the phase domain.

The phase spectrum is not an appealing part of the Fourier transform for speech signal processing. However, it has recently received renewed attention. An expanding body of work propounds that phase can be employed in a multitude of applications [5,6], including in speech reconstruction [7,8], speech enhancement [9–14], robust speech recognition [15–21], speaker recognition [22, 23] and pitch/melody extraction for music signal [24, 25]. We recently developed a source-filter model in the phase domain [26, 27] which clarifies how the phase spectrum encodes speech information and allows the separation of the vocal tract and the excitation components through phase-based signal manipulation. The filter component was turned into a set of features and tested successfully in GMM/HMM and DNN-based ASR.

This paper aims at studying the usefulness of speech phase spectrum and its source component for pitch estimation. To this end, an extension of the cepstrum, HPS and SRH techniques to the phase and group delay (GD) domains is studied and the corresponding equations are derived. Moreover, we propose a voicing detection algorithm to facilitate F_0 tracking. Experimental results on the FDA [28] and Keele [29] databases, in both clean and noisy conditions, show that the phase spectrum can be successfully employed for robust and accurate pitch tracking.

The rest of this paper is organised as follows. In Section 2 the signal information distribution between phase and magnitude as well as the source-filter separation in the phase domain are briefly explained. Section 3 explores the extension of the cepstrum, HPS and SRH methods to the phase and GD domains. In Section 4 a novel algorithm is proposed for voicing detection to enhance the pitch tracking. Section 5 includes experimental results and discussion and Section 6 concludes the paper.

2. Usefulness of the Speech Phase Spectrum for Pitch Extraction

2.1. Signal Information Distribution

Speech is a mixed-phase signal [8,30] because its complex cepstrum is neither causal nor anti-causal [31]. Therefore, it can be decomposed as follows

$$X(\omega) = X_{MinPh}(\omega)\, X_{AllP}(\omega) \quad (1)$$

$$|X(\omega)| = |X_{MinPh}(\omega)| \quad (2)$$

$$arg[X(\omega)] = arg[X_{MinPh}(\omega)] + arg[X_{AllP}(\omega)] \quad (3)$$

$$arg[X_{MinPh}(\omega)] = -\frac{1}{2\pi} log|X_{MinPh}(\omega)| * cot(\frac{\omega}{2}) \quad (4)$$

where $MinPh$, $AllP$, $|X(\omega)|$, $arg\{X(\omega)\}$ indicate the minimum-phase component, all-pass part, (short-time) magnitude spectrum and unwrapped (continuous) phase spectrum, respectively. For minimum-phase signals, the complex cepstrum is causal and based on that the Hilbert transform establishes a relationship between the phase and magnitude spectra such that the (unwrapped) phase can be computed from the magnitude spectrum and the magnitude may be calculated from the phase spectrum up to a scale error. Casting the information components into a Venn diagram, what is shared by both phase and magnitude spectra is the *minimum-phase-scale-excluded* part, the all-pass information is uniquely captured by the phase and the scale information resides in the magnitude spectrum.

The only piece of signal information which is missed by phase is the scale, which intra-framely it is merely a constant representing the signal intensity which has no bearing on the fundamental frequency and the formants. Therefore, information-wise, the phase spectrum has great potential to be used in pitch tracking.

[1] Pitch is the perception of F_0, but we use them interchnagely here.

10.21437/Interspeech.2018-1062

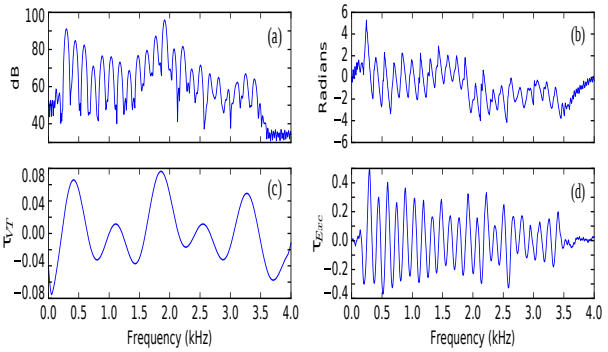

Figure 1: *Source-filter separation in phase domain [26, 27]. (a) magnitude spectrum($log|X|$), (b) phase of the minimum-phase part ($arg\{X_{MinPh}\}$), (c) group delay of the vocal tract (τ_{VT}), (d) group delay of the excitation of component (τ_{Exc}).*

2.2. Phase-based Source-Filter Separation

Having illustrated the phase information content qualitatively, one needs to know how the speech source and filter components are captured by the phase spectrum and how they can be separated in the phase domain. Since the vocal tract ($X_{VT}(\omega)$) and excitation ($X_{Exc}(\omega)$) components are convolved in the time domain, their log-magnitude and phase spectra are additive

$$log|X_{MinPh}(\omega)| = log|X_{VT}(\omega)| + log|X_{Exc}(\omega)| \quad (5)$$

$$arg\{X_{MinPh}(\omega)\} = arg\{X_{VT}(\omega)\} + arg\{X_{Exc}(\omega)\}. \quad (6)$$

As illustrated in [26], $arg\{X_{MinPh}(\omega)\}$ can be understood as a superposition of two components: a quickly oscillating *Fluctuation*, modulated by a slowly varying *Trend*. The former is associated with the excitation (Exc) part and the latter corresponds to the vocal tract (VT). Based on having different rates of change with respect to the independent variable (frequency) and using the additive property in Eq.(6), the source and filter were successfully separated in [26].

Moreover, in [32] two modifications were made in the framework: the log was replaced with the generalised logarithmic function [33] and the derivative of the phase spectrum, i.e., GD was calculated using a regression filter [34]. It was demonstrated that the former is particularly useful when working with the filter part and the latter is helpful in processing the source element. Fig. 1 shows the results of source-filter separation in the phase domain which clearly highlights the potential of phase spectrum in pitch extraction.

Now, we need to mathematically underpin the phase-based pitch extraction process.

3. Phase-based Pitch Extraction

3.1. Preliminaries and Magnitude-based Techniques

Frequency domain algorithms for pitch extraction such as cepstrum, HPS and SRH, compute a magnitude-based function in which F_0 emerges as $argmax$ (in a pre-specified search range)

$$Cep[q] = \mathcal{F}^{-1}\{log|X[k]|\} \quad (7)$$

$$HPS[k] = \prod_{h=1}^{N_{harm}} |\, X[h\,k]\,|^2 \quad (8)$$

$$SRH[k] = E[k] + \sum_{m=2}^{N_{harm}} E[mk] - E[(m-\frac{1}{2})\,k] \quad (9)$$

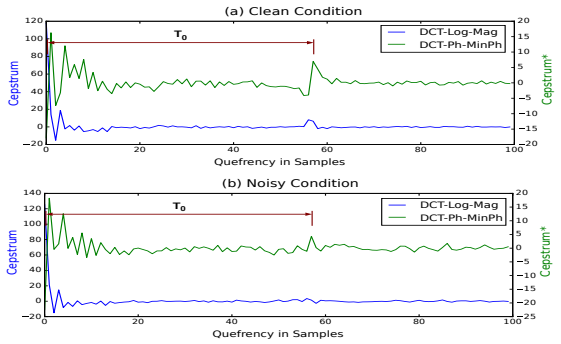

Figure 2: *Extension of the cepstrum method to the phase domain. (a) Clean condition, (b) noisy condition (Car, 10 dB).*

where $\mathcal{F}^{-1}$, q, k, N_{harm} and E denote the inverse Fourier transform, quefrency, discrete frequency, number of harmonics and the magnitude spectrum of the residual signal after linear prediction [35] (representing the excitation), respectively.

To extend these techniques to the phase domain, one first needs to find the equivalent of the $argmax$ of such magnitude-based functions in the phase and group delay domains.

3.2. Extension of the Cepstrum Method to Phase Domain

The cepstrum method is based on the fact that the periodicity in a domain gets translated into peaks (impulses) at the fundamental periodicity and its harmonics in the dual domain, namely after taking the Fourier transform (or discrete cosine transform (DCT)). Returning to Fig. 1 (a) and (b), both $log|X(\omega)|$ and $arg\{X_{MinPh}(\omega)\}$ have the same periodicity, although the pattern which is repeated is different. Therefore, taking the Fourier transform of both leads to impulses at the same bins, as demonstrated in Fig. 2. As such the cepstrum method can be extended to the phase domain as follows

$$\frac{1}{F_0} = \underset{q_{min}\le q\le q_{max}}{argmax} \underbrace{\mathcal{F}^{-1}\{arg[X_{MinPh}(\omega)]\}[q]}_{Cep^*} \quad (10)$$

where Cep^*, q_{min} and q_{max} denote the phase-based cepstrum, minimum and maximum quefrencies (for search), respectively. The same argument holds for the derivative of the phase, namely the GD, because differentiation is a linear transform.

3.3. Extension of HPS and SRH to Phase Domain

Let us revisit the Hilbert transform in Eq. (4). Using the Trapezoidal rule [36], it can be approximated as follows

$$arg\{X_{MinPh}(\omega)\} = -\frac{1}{2\pi}\mathcal{P}\int_{-\pi}^{\pi} log|X(\theta)|cot(\frac{\omega-\theta}{2})d\theta$$

$$arg\{X_{MinPh}[k]\} \approx c_1^k(\tilde{X}[k+1]-\tilde{X}[k-1]) \quad (11)$$

$$+ c_2^k(\tilde{X}[k+2]-\tilde{X}[k-2]) + c_3^k(\tilde{X}[k+3]-\tilde{X}[k-3]) + ...$$

where $\mathcal{P}$ indicates the Cauchy principle value of the integral, $\tilde{X}[k] = log|X(\omega_k)|$ and

$$c_n^k = \frac{\Delta\omega}{2\pi}cot(\omega_k - \omega_{k-n}) = \frac{\Delta\omega}{2\pi}cot(n\Delta\omega) \quad (12)$$

in which $\Delta\omega = \omega_k - \omega_{k-1}$, is the frequency resolution, and the lower the $\Delta\omega$ the higher the frequency resolution. Note that when $\Delta\omega \to 0 \Rightarrow cot(\Delta\omega) \to \infty$ and since the cotangent decays quickly around zero $cot(\Delta\omega) >> cot(n\Delta\omega)$, for $n > 1$. Considering these points and assuming the $\Delta\omega$ is small enough,

$$c_1^k >> c_2^k > c_3^k \quad ... \quad (13)$$

This allows the following approximation

$$arg\{X_{MinPh}[k]\} \approx c_1^k \ (\tilde{X}[k+1] - \tilde{X}[k-1]). \quad (14)$$

As a result, phase is proportional with the slope of the log-magnitude spectrum. To further clarify the Eq. (14), it is useful to compute the first order (central) derivative of phase, $\Delta\phi_X[k]$,

$$\begin{aligned}\Delta\phi_X[k] &= (arg\{X_{MinPh}[k+1]\} - arg\{X_{MinPh}[k-1]\})/2 \\ &\approx c_1^k \ (\tilde{X}[k+2] - 2\tilde{X}[k] + \tilde{X}[k-2])/2 \quad (15)\end{aligned}$$

which illustrates how phase varies, as a function of $\tilde{X}$. As seen, the rate of change of the phase is proportional to the second derivative of $\tilde{X}$. Around a maximum point, $\tilde{X}$ behaves like a concave function and its second derivative will be negative. Hence, the $\Delta\phi_X[k]$ would be negative and the phase will be a decreasing function of frequency in that neighbourhood.

To have a quantitative interpretation of $\Delta\phi_X$ or the second derivative of $\tilde{X}$, *curvature* [36] notion could be helpful. When the first derivative gets close to zero, which is the case at the local extrema of the $\tilde{X}$, curvature (almost) equals the second derivative. Intuitively, the larger the peak, the larger the curvature and subsequently, the larger the second derivative. Therefore, at the neighbourhood of the $\tilde{X}$'s local maxima, namely poles, the phase spectrum shows maximum negative (or downward) slope. By the same token, for the minima of $\tilde{X}$, i.e. zeros, $\Delta\phi_X[k]$ would be positive and maximum. As a result, zeros emerge as bins at which the phase spectrum has the maximum upward (positive) slope. Fig. 3 illustrates these points.

Putting these altogether and remembering that group delay is the negative derivative of the phase spectrum, one can write

$$\begin{aligned}argmax\{|X|\} &= argmax\{log|X|\} = argmax\{F(|X|)\} \\ &= argmin\{\Delta arg\{X\}\} \quad (16) \\ &= argmax\{-\Delta arg\{X\}\} = argmax\{\tau_X\}\end{aligned}$$

for some monotonically increasing function F. In particular, HPS in the group delay domain takes the following form

$$\begin{aligned}F_0 &= \underset{k_{min}\le k\le k_{max}}{argmax} \{HPS[k]\} = argmax\{\prod_{h=1}^{N_{harm}} |X[hk]|^2\} \\ &= argmax\{\sum_h \tilde{X}[hk]\} = argmax\{\sum_h \tau_X[hk]\} \quad (17)\end{aligned}$$

and the SRH's equivalent in the group delay domain would be

$$\begin{aligned}F_0 &= \underset{k_{min}\le k\le k_{max}}{argmax} \{SRH\ [k]\} \quad (18) \\ &= argmax\{\tau_{Exc}[k] + \sum_{m=2}^{N_{harm}} \tau_{Exc}[mk] - \tau_{Exc}[(m-\frac{1}{2})k]\}.\end{aligned}$$

To distinguish the phase-based approaches from their magnitude-based counterparts, from now onwards, the phase-based ones are denoted by *, namely Cep*, HPS* and SRH*.

4. Proposed Method for Voicing Determination

To track the pitch, as well as the F_0 value, one requires estimating the voiced/unvoiced state of the frame. One solution is to threshold the value of the cepstrum, HPS or SRH at its argmax, namely the putative pitch. However, finding an optimal value for the threshold is not straightforward and it varies with signal-to-noise ratio (SNR), noise type, etc. Furthermore, it turns out

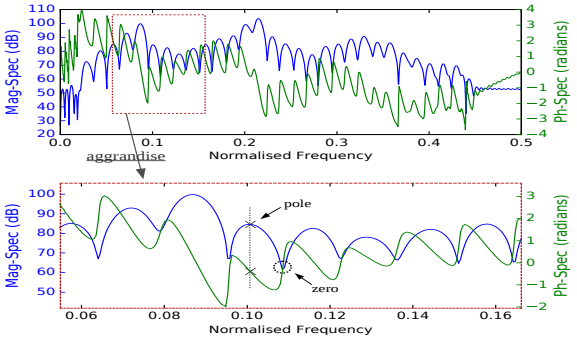

Figure 3: *Relation of the phase spectrum (Ph-Spec) slope with the local minima/maxima of magnitude spectrum (Mag-Spec).*

that even if the optimal threshold is found, some misses and false alarms still occur (Fig. 4(e)). In this section, we propose a simple yet effective method for voicing detection.

The red curve at Fig. 4(a) show the unprocessed (raw) pitch track along with the ground truth track and Fig. 4(b) shows the value of the SRH^* function at the F_0. Based on the properties of the raw pitch track and the value of the function at the fundamental frequency we proposed the following algorithm:

Step 1 The raw F_0 track varies smoothly at voiced frames and is spiky at the unvoiced ones. Initially, assume frames at which $|F_0[n] - F_0[n-1]| < threshold$ (n is frame index) are voiced. It was observed that the $threshold$ is not a critical parameter and 20 to 50 Hz is fine. It was set to 30 in the experiments. Fig. 4(c) shows the resulting track which is still spiky at unvoiced frames.

Step 2 Remove the spikes through accepting a frame as voiced, only if it is in the center of a sequence of at least c consecutive voiced frames. c was set to 5 here. As shown in Fig. 4(d), it returns a reasonable estimate for the voiced/unvoiced flags.

Step 3 Using the unvoiced flags, one can compute the threshold for voicing detection: set the threshold to the mean plus standard deviation of the function at the unvoiced frames. Large threshold gives rise to higher miss rate and low threshold brings about more false alarms. As seen in Fig. 4(b) this returns a reasonable value for the threshold. However, as Fig. 4(e) illustrates, relying only on the threshold gives rise to false alarms.

Step 4 To deal with the false alarms which the threshold causes, compute the logical AND of the voiced flags of the Step 2 and Step 3. Actually, it was observed that the Step 3 can be totally skipped, although keeping it along with applying Step 4 can slightly improve the performance. The results are almost identical to Fig. 4(d).

Step 5 Apply a median filter to smooth the F_0 track (Fig. 4(f)). The filter length was set to 7 here. It is not a critical choice (3, 5 and 9 may be used) but it should be an odd number, otherwise it could give rise to octave error at the voiced/unvoiced borders.

5. Experimental Results

5.1. Setup

For evaluating the proposed phase-based approach, the FDA [28] and Keele [29] databases have been used. The sampling frequency for both databases is 20 kHz. Frame length, frame shift, FFT size and N_{harm} were set to 64 ms, 10 ms, 4096 and

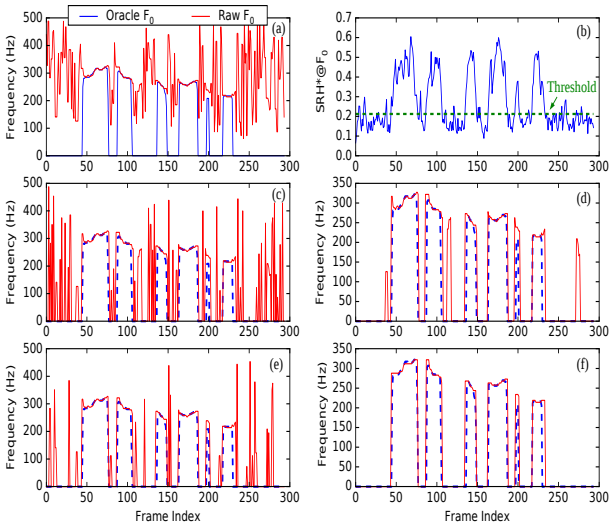

Figure 4: *Steps of the proposed method for voiced-unvoiced detection. (a) raw (unprocessed) F_0 track vs the oracle track, (b) track of the SRH* at F_0 ($SRH^*@F_0$) vs. the threshold computed in Step 3, (c) Step 1, (d) Step 2, (e) Step 3, (f) Step 5.*

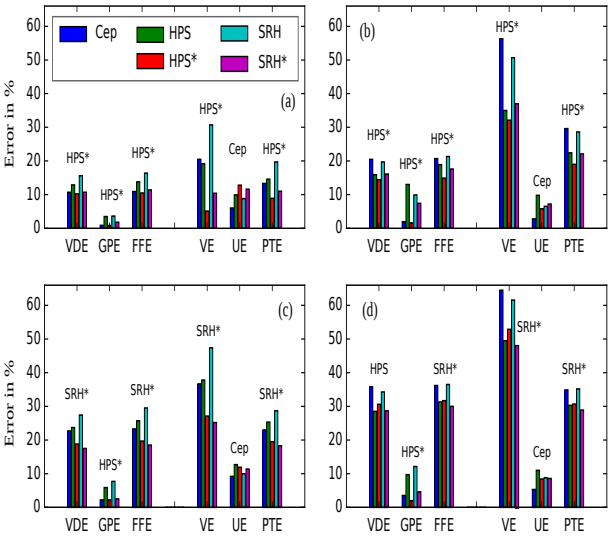

Figure 5: *Pitch performance measures in clean and noisy (Babble, 10 dB) conditions. (a) FDA-clean, (b) Keele-clean, (c) FDA-noisy, (d) Keele-noisy. The best technique for each metric is mentioned on top of the corresponding bars.*

5, respectively. The proposed voicing detection algorithm was employed with the aforementioned parameters for both phase and magnitude-based approaches. Pitch was computed in a single pass without extra search space limiting other than the minimum/maximum F_0 values, i.e., 50/500 Hz.

The FaNT tool [37] was used to add noise to clean signals. It deploys speech detection before the SNR calculation which leads to a more meaningful definition of SNR. Note that skipping speech detection and computing the power using all the speech/non-speech frames leads to a smaller power for the clean signal, especially when the silence part is large. Consequently, for a given SNR, the power of the noise would be lower, too. This, in turn, leads to a spurious measure of noise robustness.

For both databases, the provided pitch references were used as a ground truth. To evaluate F_0 tracks objectively, the following measures were employed: Voicing decision error (VDE) [38], gross pitch error (GPE) [38], fine frame error (FFE) [39], voiced error (VE) [40], unvoiced error (UE) [40] and pitch tracking error (PTE) [40]. For details of each metric, please refer to the respective references.

5.2. Results and Discussion

Experimental results in clean and noisy (Babble, 10 dB) conditions are illustrated in Fig. 5 (average of male and female speakers). As seen, for both FDA and Keele, and in clean and noisy conditions, the phase-based approach for most measures and test scenarios outperforms its magnitude-based counterpart.

For the cepstrum method, it was observed that the phase and magnitude-based approaches return the same results. This stems from the fact that the Hilbert transform which links the log-magnitude and the phase spectra and the DCT/FFT which takes these spectra to the cepstrum domain are both linear and under a linear transform the argmax does not change. That is why both Cep and Cep* return identical estimate for pitch. In Fig. 5 and for saving space only one of them is shown. As the results indicate, the cepstrum method is less robust than the other techniques, although it is accurate at high SNRs.

In case of the HPS, the proposed approach returns better results than its magnitude-based counterpart in many test cases. This could be expected considering the higher frequency resolution and lower frequency leakage properties of the group delay which paves the way for pinpointing the pitch bin with higher accuracy. As seen in Fig. 5, HPS, provides a good level of accuracy/robustness and works well in the phase domain.

For the SRH method, the phase-based framework outperforms its magnitude-based counterpart for almost all measures. This may be explained by the high frequency resolution of the GD (similar to HPS) but also there is one more contributing factor: Note that, in the magnitude-based SRH, the residual signal is computed using linear prediction whereas in the phase-based one the cepstral smoothing is utilised to extract the source component. Since in pitch tracking, the frame length is usually longer than 20–40 ms (to achieve higher frequency resolution, especially for low-pitch speakers), the stationarity assumption may be violated to some extent. Therefore, considering a parametric autoregressive model for a frame with longer length (64 ms here) introduces some error. Cepstral smoothing requires the stationarity assumption, too, but since it is nonparametric it can better handle this issue.

6. Conclusion

In this paper the usefulness of the speech phase spectrum for pitch tracking was studied. After illustrating that the phase includes source information, the extension of three magnitude-based techniques, namely cepstrum, HPS and SRH to the phase domain was investigated. It was shown than the $argmax$ of an increasing monotonic function of the magnitude spectrum is equivalent to the $argmin$ of the phase change w.r.t. the frequency and the $argmax$ of the group delay. Using these points, the equivalents of the aforementioned techniques in the phase and GD domains were derived. In addition, a novel method was proposed for voicing detection to facilitate pitch tracking. The performance of the magnitude-based and phase-based techniques was studied and compared in clean and noisy conditions using six pitch evaluation metrics. It was shown that the phase-based approach along with the proposed method for voicing detection can be successfully employed in pitch tracking. Testing the proposed approach in classification/recognition tasks such as ASR and emotion recognition is suggested for future work.

7. References

[1] G. Fant, *Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations*, ser. Description and Analysis of Contemporary Standard Russian. De Gruyter, 1971.

[2] A. Noll, *Short-time Spectrum and "cepstrum" Techniques for Vocal-pitch Detection*, ser. Bell telephone system technical publications. Bell Telephone Laboratories, 1964.

[3] ——, *Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and maximum likelihood estimate*, ser. Proceedings of the Symposium on Computer Processing in Communications. Polytechnic Press: Brooklyn, New York, April 1969, vol. XIX.

[4] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics." in *INTERSPEECH*. ISCA, 2011, pp. 1973–1976.

[5] E. Loweimi, "Robust phase-based speech signal processing; from source-filter separation to model-based robust asr," Ph.D. dissertation, Sheffield, UK, Feb 2018. [Online]. Available: http://etheses.whiterose.ac.uk/19409/

[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] 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.

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

[9] K. Wojcicki, M. Milacic, A. Stark, J. Lyons, and K. Paliwal, "Exploiting conjugate symmetry of the short-time fourier spectrum for speech enhancement," *Signal Processing Letters, IEEE*, vol. 15, pp. 461–464, 2008.

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

[11] 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.

[12] M. Krawczyk-Becker and T. Gerkmann, "An evaluation of the perceptual quality of phase-aware single-channel speech enhancement," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 364–369, 2016.

[13] M. Pirolt, J. Stahl, and P. Mowlaee, "Phase estimation in single-channel speech enhancement using phase invariance constraints," in *ICASSP*, 2017, pp. 5585–5589.

[14] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice*. Wiley, 2016.

[15] E. Loweimi, S. Ahadi, T. Drugman, and S. Loveymi, "On the importance of pre-emphasis and window shape in phase-based speech recognition," in *Lecture Notes in Computer Science, Advances in Non-Linear Speech Processing (NOLISP)*, vol. 7911 LNAI, 2013, pp. 160–167.

[16] E. Loweimi, S. Ahadi, and T. Drugman, "A new phase-based feature representation for robust speech recognition," in *ICASSP*, May 2013, pp. 7155–7159.

[17] 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, 2007.

[18] E. Loweimi and S. Ahadi, "A new group delay-based feature for robust speech recognition," in *ICME*, July 2011, pp. 1–5.

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

[20] E. Loweimi, J. Barker, and T. Hain, "Statistical normalisation of phase-based feature representation for robust speech recognition," in *ICASSP*, 2017, pp. 5310–5314.

[21] ——, "Exploring the use of group delay for generalised vts based noise compensation," in *ICASSP*, 2018.

[22] R. Padmanabhan, S. Parthasarathi, and H. Murthy, "Robustness of phase based features for speaker recognition," in *INTERSPEECH*. ISCA, 2009, pp. 2355–2358.

[23] K. Vijayan, R. Pappagari, and K. Sri Rama Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Communication*, vol. 81, no. C, pp. 54–71, Jul. 2016.

[24] R. Rajan and H. Murthy, "Group delay based melody monopitch extraction from music," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 186–190.

[25] ——, "Two-pitch tracking in co-channel speech using modified group delay functions," *Speech Communication*, vol. 89, pp. 37 – 46, 2017.

[26] E. Loweimi, J. Barker, and T. Hain, "Source-filter separation of speech signal in the phase domain." in *INTERSPEECH*. ISCA, 2015, pp. 598–602.

[27] E. Loweimi, J. Barker, O. Saz Torralba, and T. Hain, "Robust source-filter separation of speech signal in the phase domain," in *INTERSPEECH*, 2017, pp. 414–418.

[28] P. Bagshaw, S. Hiller, and M. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching," in *EUROSPEECH*, 1993.

[29] F. Plante, G. Meyer, and W. Ainsworth, "A pitch extraction reference database," in *EUROSPEECH*, 1995, p. 837840.

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

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

[32] E. Loweimi, J. Barker, and T. Hain, "Channel compensation in the generalised vector taylor series approach to robust ASR," in *INTERSPEECH*, 2017, pp. 2466–2470.

[33] T. Kobayashi and S. Imai, "Spectral analysis using generalized cepstrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 5, pp. 1087–1089, Oct 1984.

[34] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[35] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[36] S. C. Chapra and R. Canale, *Numerical Methods for Engineers*, 5th ed. New York, NY, USA: McGraw-Hill, Inc., 2006.

[37] G. Hirsch, "Fant - filtering and noise adding tool," http://dnt.kr.hsnr.de/download.html, 2005.

[38] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 399–418, Oct 1976.

[39] W. Chu and A. Alwan, "Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 3969–3972.

[40] B. Lee and D. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in *INTERSPEECH*. ISCA, 2012, pp. 598–602.