INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Robust Source-Filter Separation of Speech Signal in the Phase Domain

Erfan Loweimi, Jon Barker, Oscar Saz Torralba and Thomas Hain

Speech and Hearing Research Group (SPandH), University of Sheffield, Sheffield, UK
{eloweimi1, j.p.barker, o.saztorralba,t.hain}@sheffield.ac.uk

Abstract

In earlier work we proposed a framework for speech source-filter separation that employs phase-based signal processing. This paper presents a further theoretical investigation of the model and optimisations that make the filter and source representations less sensitive to the effects of noise and better matched to downstream processing. To this end, first, in computing the Hilbert transform, the log function is replaced by the generalised logarithmic function. This introduces a tuning parameter that adjusts both the dynamic range and distribution of the phase-based representation. Second, when computing the group delay, a more robust estimate for the derivative is formed by applying a regression filter instead of using sample differences. The effectiveness of these modifications is evaluated in clean and noisy conditions by considering the accuracy of the fundamental frequency extracted from the estimated source, and the performance of speech recognition features extracted from the estimated filter. In particular, the proposed filter-based front-end reduces Aurora-2 WERs by 6.3% (average 0-20 dB) compared with previously reported results. Furthermore, when tested in a LVCSR task (Aurora-4) the new features resulted in 5.8% absolute WER reduction compared to MFCCs without performance loss in the clean/matched condition.

Index Terms: phase spectrum, source-filter separation, group delay, generalised logarithmic function, regression filter

1. Introduction

Phase spectrum is not an obvious starting point for speech processing. In contrast to the magnitude spectrum whose fine and coarse structures have a clear relation to speech perception, the phase spectrum is difficult to interpret and manipulate. In fact, there is neither a meaningful trend nor extrema which may facilitate the modelling. Nevertheless, the speech phase spectrum has recently gained renewed attention. For example, it has been the focus of a special session in Interspeech 2014 [1], a tutorial session in Interspeech 2015 and a special issue in Speech Communication journal [2]. An increasing body of work is showing that the phase spectrum can be employed in a multitude of speech processing applications, including in speech enhancement [3–6], speech reconstruction [7–12], speech recognition [13–19] and speaker recognition [20, 21].

Now that the potential for phase-based speech processing has been established, there is a need for a fundamental model to help understand the way in which it encodes speech information. In this respect, we proposed a phase-domain source-filter model that allows for deconvolution of the vocal tract (filter) and excitation (source) components through phase manipulation [22]. This model shows how the excitation and vocal tract elements mix in the phase domain and provides a mathematical framework for segregating them.

In this paper we aim at further elaboration and optimisation of the proposed model to facilitate a more robust fundamental frequency (F_0) estimation from the source element while obtaining better performance in ASR from the filter part. In this regard, the computation of the phase spectrum by the Hilbert transform is modified: the generalised logarithmic function is used in place of the standard log function. This function provides one degree of freedom which can be tuned in order to achieve a better dynamic range (DR) and statistical distribution. Moreover, in computing the group delay (GD), a regression filter has been employed instead of the sample difference. By considering a wider context, estimation of the derivative becomes more accurate and further robust. It was observed that the first and second modifications are particularly useful in representing the vocal tract and excitation components, respectively.

The rest of this paper is organised as follows. Section 2 reviews the phase-based source-filter model. Section 3 describes the proposed modifications. Section 4 presents and discusses experimental results and Section 5 concludes the paper.

2. Phase-based Source-Filter Separation

Speech is a *mixed-phase* signal [11] as its complex cepstrum (CC) is neither causal nor anti-causal [23]. Therefore, it can be divided into *minimum-phase (MinPh)*, $X_{MinPh}(\omega)$, and *all-pass (AllP)*, $X_{AllP}(\omega)$, components

$$\begin{aligned} X(\omega) &= X_{MinPh}(\omega)\, X_{AllP}(\omega) \\ |X(\omega)| &= |X_{MinPh}(\omega)| \\ arg[X(\omega)] &= arg[X_{MinPh}(\omega)] \,+\, arg[X_{AllP}(\omega)] \end{aligned} \tag{1}$$

where $|X(\omega)|$ and $arg[X(\omega)]$ indicate the (short-time) magnitude and unwrapped (continuous) phase spectra, respectively.

Since vocal tract ($X_{VT}(\omega)$) and excitation ($X_{Exc}(\omega)$) components are convolved in the time domain, the magnitude spectrum ($|X(\omega)|$) is the product of the corresponding magnitude spectra. Given $|X(\omega)|$ is only linked to the MinPh part

$$|X(\omega)| = |X_{VT}(\omega)|\, |X_{Exc}(\omega)| = |X_{MinPh}(\omega)|. \tag{2}$$

Based on causality of the CC for MinPh signals, the Hilbert transform provides a mapping between magnitude and phase

$$arg[X_{MinPh}(\omega)] = -\frac{1}{2\pi}\, log|X_{MinPh}(\omega)| \,*\, cot(\frac{\omega}{2}) \tag{3}$$

where $*$ denotes convolution. By replacing the $log|X_{MinPh}(\omega)|$ with $log|X(\omega)|$, $arg[X_{MinPh}(\omega)]$ can be calculated. Equivalently, the computation may be performed in the cepstrum domain by applying a causal lifter (Fig. 1) [23]. Substituting (2) into (3) yields

$$\begin{aligned} arg[X_{MinPh}(\omega)] &= -\frac{1}{2\pi} log\big(|X_{VT}(\omega)|\, |X_{Exc}(\omega)|\big) * cot(\frac{\omega}{2}) \\ &= arg[X_{VT}(\omega)] + arg[X_{Exc}(\omega)], \end{aligned} \tag{4}$$

which shows that the source and filter are additive in the (unwrapped) phase domain. As illustrated in [22], $arg[X_{MinPh}(\omega)]$, in contrast to the wrapped phase

http://dx.doi.org/10.21437/Interspeech.2017-210

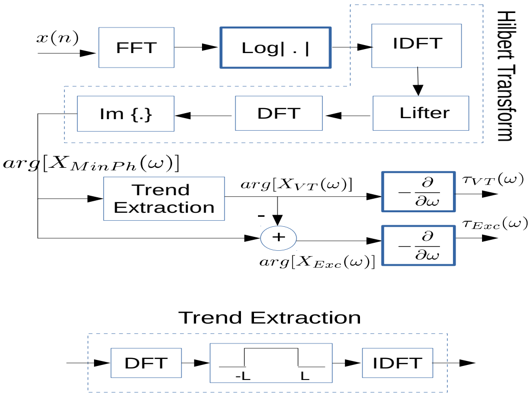

Figure 1: *Phase-based source-filter decomposition [22].*

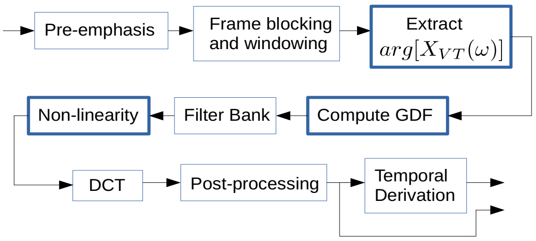

Figure 2: *Phase filter-based feature extraction proposed in [22].*

($ARG[X(\omega)]$), is no longer chaotic and can be understood as a superposition of two components: a quickly oscillating *Fluctuation*, modulated by a slowly varying *Trend*. As shown in [22], Trend and Fluctuation correspond to the vocal tract (VT) and excitation (Exc) parts, respectively.

In addition to the source and filter elements (embedded in $X_{MinPh}(\omega)$), the speech signal also includes timing information which captures the corresponding temporal evolution. This aspect is encoded in the AllP part and resides uniquely in the phase spectrum. Since speech is processed in short frames (in which stationarity holds), the frame index can be taken as a proxy for the timing information. However, for mid/long-term processing, the importance of such information and usefulness of the AllP part increases.

For evaluating the effectiveness of the proposed method in ASR, a simple feature (named *BMFGDVT* [22]) was extracted from the filter component of the phase (Fig. 2) and tested in an ASR system. On average, it showed better performance than conventional well-known features on Aurora-2 task (Table 1).

3. Improved Source Filter Separation

3.1. Generalised Logarithmic Function

In the classic definition of the Hilbert transform, $arg[X_{MinPh}]$ is computed through (3). Here, we modify this and instead of log, utilise the generalised logarithmic function ($GenLog$) [24]

$$\begin{cases} GenLog(x;\alpha) = \frac{1}{\alpha}(x^{\alpha}-1), \quad x>0 \quad \alpha \neq 0 \\ \lim_{\alpha\to 0} GenLog(x;\alpha) = log(x), \end{cases} \tag{5}$$

where α is its parameter. In the Statistics literature, this function is known as the Box-Cox transform [25]. It unifies the power and log transforms and is helpful in improving the Gaussianity.

$GenLog(x;\alpha)$ provides one degree of freedom that allows two main properties of the representation to be adjusted, namely its DR and statistical distribution. Fig. 3 shows that by increasing α, the DR of the representation gets larger. Note that for the magnitude-based features, the power spectrum which has a high DR is fed into the filter bank (FB) and then the compression is carried out through power transformation (log is its special case). If the order of compression and FB is swapped in the pipeline, the performance degrades. However, in the case of the proposed phase-based feature, $\tau_{VT}(\omega)$ which has a limited DR (comparable to $log|X(\omega)|$), enters the FB. Similar to the magnitude-based features it could be costly performance-wise.

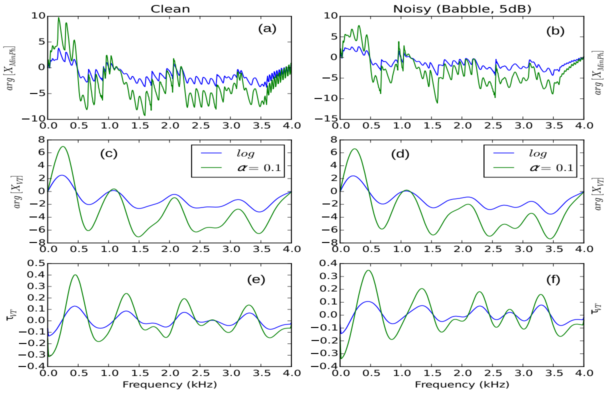

Figure 3: *Effect of using $GenLog(|X(\omega)|)$ in the Hilbert Transform at the clean and noisy (Babble, 5 dB) conditions. (a) $arg[X_{MinPh}]$-clean, (b) $arg[X_{MinPh}]$-noisy, (c) $arg[X_{VT}]$-clean, (d) $arg[X_{VT}]$-noisy (e), τ_{VT}-clean, (f) τ_{VT}-noisy.*

Contrary to the magnitude spectrum, DR of the GD is not related to the signal energy level at different bins. It depends on the relative location of the poles/zeros with respects to the unit circle. Zeros located next to the unit circle (primarily associated with the excitation component) increase the DR of GD and make it too spiky if left uncontrolled. By removing the source part, the spikiness issue is greatly alleviated but the DR of the GD is significantly reduced, too (Fig. 3). Tuning α allows the DR of $\tau_{VT}(\omega)$ to be adjusted without increasing the spikiness. Another advantage of using the $GenLog$ relates to the noisy condition where contamination with noise results in DR reduction. Increasing α counters this effect of the noise and consequently improves the robustness (Fig. 3).

Rewriting (4) using the $GenLog$ function yields

$$arg[X_{MinPh}(\omega);\alpha] = -\frac{1}{2\pi\alpha}|X_{VT}(\omega)|^{\alpha}\ |X_{Exc}(\omega)|^{\alpha} * cot(\frac{\omega}{2}). \tag{6}$$

Although the $GenLog$ function adds flexibility to the framework, based on (6), it poses a substantial problem: the useful additive relationship between the source and filter resulting from the log function (eq. (4)) is replaced with multiplication (eq. (6)). This could hinder source and filter separation because the Trend-*plus*-Fluctuation premise is undermined. However, as seen in Fig. 3(a), as long as α is set to a sufficiently small value, e.g. 0.1, the Trend and Fluctuation remain *quasi-additive*. While given (6) this may seem counterintuitive, Maclaurin series expansion of function $f(\alpha) = z^{\alpha}$, where $z = |X_{VT}(\omega)|\ |X_{Exc}(\omega)|$, shows the reason

$$\begin{aligned} f(\alpha) &= 1 + \alpha\, logz + \alpha^2(log\ z)^2 + \alpha^3(log\ z)^3 + ... \\ &\approx 1 + \alpha\, logz = 1 + \alpha\,(log|X_{VT}(\omega)| + log|X_{Exc}(\omega)|). \end{aligned} \tag{7}$$

As far as $\alpha \ll 1$, nonlinear terms in (7) remain negligible and the Trend-*plus*-Fluctuation assumption stays reasonable. So, α should be set large enough to supply a sufficient DR but small enough to avoid the violation of the quasi-additive combination.

3.2. Group Delay

Group delay (GD), $\tau_X(\omega)$, is defined as the negative spectral derivative of $arg[X(\omega)]$. High spectral resolution is an important advantage of GD [18] but spikiness is a major problem. Cepstral smoothing [18], chirp processing [17] and signal modelling [15, 26] have been used for addressing this issue. However, what makes the GD a special representation for the phase spectrum is that (if its spikiness issue is resolved) it resembles the magnitude spectrum. As a result, one of the key problems with phase, i.e. ambiguous shape, will be solved allowing a wide range of magnitude-based methods to be employed.

3.2.1. *Usefulness of the Group Delay*

While the bulk of GD-related research is concerned about circumventing the spikiness, an important question is overlooked: why does GD bear a resemblance to the magnitude spectrum? In other words, among all the possible mathematical representations for the (unwrapped) phase spectrum, what is special about its derivative which renders the foregoing useful similarity?

The answer stems from the way in which information is encoded in the phase spectrum. Contrary to the magnitude spectrum where information is distributed in the amplitude values, in the phase domain it resides in the level-crossing structure. For the sake of argument let's consider a simple single-pole ($re^{j\theta}$) function where information means r and θ. For the magnitude spectrum, the bin in which the maximum occurs gives θ and the corresponding amplitude value determines r. In the phase domain, however, the bin at which zero-crossing takes place yields θ and the slope at that point gives r.

Loosely speaking, in the magnitude spectrum the information appears in an amplitude modulation (AM) format whereas for the phase spectrum it looks like frequency modulation (FM) where information gets encoded in the slopes rather than amplitude values. By computing the derivative, similar to FM demodulation through *discriminator* (aka *slope detector*) [27], the information would be demodulated and moved into the amplitude domain. This pushes the overall structure of the GD toward an AM signal, similar to the magnitude spectrum, and consequently facilitates the interpretation/processing.

3.2.2. *Computing the Group Delay*

Since the phase of the DFT is a discrete sequence, numerical differentiation is typically approximated by a finite difference ($diff$). The 1st-order diff, as is typically used, is intrinsically noisy. We propose to fit a line (regression filter [28]) to a short spectral interval around each bin and take the slope as the GD

$$\tau_X[k] = -\frac{\sum_{m=-k_0}^{k_0} m \; arg[X[k+m]]}{\sum_{j=-k_0}^{k_0} j^2}, \tag{8}$$

where k and τ_X denote the discrete frequency and GD, respectively, and $2k_0 + 1$ is the length of the context in frames. The regression filter has a bandpass frequency response, contrary to the sample difference which acts like a high-pass filter. Increasing k_0 lowers the high cut-off frequency and smooths the τ_X.

Note that effect of the regression filter on the GD of VT is limited as this component is the output of a low-pass filter (Trend Extraction), and its high-frequency content is already weak. However, it is especially effective for applications that employ the excitation component. As illustrated in Fig. 4, this approach allows accurate fundamental frequency extraction from the speech phase spectrum.

To further clarify this point, F_0 was estimated by computing $argmax$ of the summation of residual harmonics (SRH) [29]

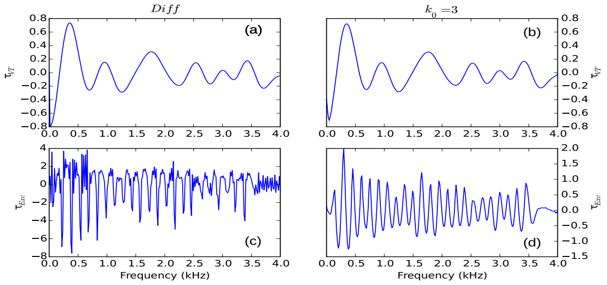

Figure 4: *Sample difference (Diff) vs regression filter for computing the GD of filter and source ($\alpha = 0.1$). (a) $\tau_{VT}[Diff]$, (b) $\tau_{VT}[k_0 = 3]$, (c) $\tau_{Exc}[Diff]$, (d) $\tau_{Exc}[k_0 = 3]$.*

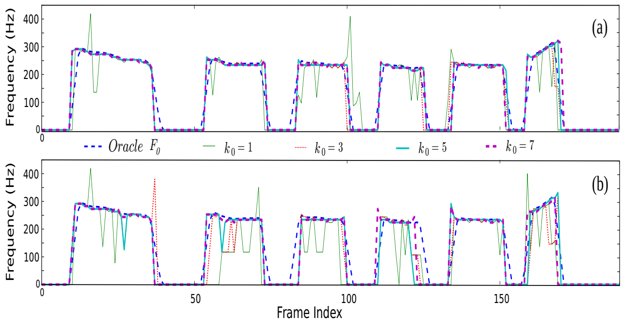

Figure 5: *Effect of k_0 on the accuracy/robustness of phase-based F_0 estimation using (9) for $sb003.sig$ from [30] ($\alpha = 0.1$). (a) clean, (b) noisy (Gaussian white, 5 dB).*

$$SRH(k) = \tau_{Exc}(k) + \sum_{m=2}^{N_{harm}} [\tau_{Exc}(m\,k) - \tau_{Exc}((m-0.5)\,k)] \tag{9}$$

where N_{harm} denotes the number of harmonics (set to 5 here). Figure 5 depicts the pitch estimated using (9) versus ground truth values taken from [30] in clean and noisy (5 dB) conditions. As seen, k_0 has a substantial impact on both accuracy and robustness of the phase-based pitch extraction process and can lead to a reliable phase-based F_0 estimation.

3.3. Non-linearity

In general the nonlinear compressive function applied to the filter bank energies (FBE) mimics the human auditory system's conversion of sound pressure into loudness and is usually implemented through the power transformation. From a machine standpoint, it is important for reshaping the distribution of the features. For phase-based features, a power transformation cannot be used directly since the admissible range is restricted to positive values whereas the FBEs may become negative if the filter bank is fed with GD.

Bickel and Doksum [31] modified the power transform ($GenLog$) such that it could also operate on negative values

$$y = \frac{sign(x)|x|^{\gamma} - 1}{\gamma}, \tag{10}$$

where $sign(\ .\ )$, $|\ .\ |$ and γ are the signum function, absolute value and the parameter of the transform, respectively. -1 from the numerator and γ from the denominator may be removed without loss of generality as they do not change the class discriminability of the features. That is why in [18, 22] only $sign(x)|x|^{\gamma}$ has been used.

Comparing (5) and (10) shows both α and γ have similar statistical effect. As such keeping both is redundant and one of them may be eliminated. Based on the argument propounded in Subsection 3.1. regarding the DR, we omit the non-linearity block placed after the FB (or integrate it into the $GenLog$).

4. Experimental Results and Discussion

4.1. Parametrisation

For evaluating the effectiveness of the proposed modifications on the robustness of the filter component a number of ASR experiments were carried out on Aurora-2 [32] and Aurora-4 [33]. For Aurora-4 HMMs were trained with 16 components per mixture and all acoustic models were standard phonetically state-clustered triphones trained from scratch using a standard HTK regime [34]. Decoding was performed with the standard 5k-word WSJ0 bigram language model. The evaluation set of Aurora-2 consists of 10 test sets, grouped into A, B and C where A and B only contain additive noises and C includes both additive and channel distortions. Aurora-4 has 14 test sets, grouped into 4 subsets: clean, (additive) noisy, clean with channel distortion and noisy with channel distortion, that will be referred to as A, B, C and D, respectively. For the DNN part, the network consists of four hidden layers with 1300 nodes, followed by a bottleneck (BN) [35] layer containing 26 nodes placed before the output layer. The network was trained using TNet [36] and standard HMM-GMM models were trained on the BN features. For Trend Extraction, L (Fig. 1) was set to $\frac{f_s}{400}$, where f_s is the sampling rate in Hz. Features are mean normalised and augmented by the log-energy (E) along with delta (D) and acceleration (A) coefficients.

4.2. Discussion

4.2.1. *Connected-digit task: Aurora-2*

For conducting a fair comparison, the effect of replacing log with $GenLog$ in the MFCC pipeline (generalised-MFCC [37]), was also evaluated. It is denoted by γ-MFCC in Table 1 and results in a significant increase in performance in noisy conditions, although in the clean condition log is a better option. Choosing an appropriate value for γ plays a key role and on average, 0.075 turned out to be an optimal choice. In the proposed feature, α and k_0 should to be tuned. Table 1 shows that the optimum value for α is around 0.1 (α-BMFGDVT-0.1) and it provides a significant WER reduction and robustness improvement compared with the previous version which was applying the log function (BMFGDVT).

The effect of using the regression filter for computing GD is shown in the last part of Table 1. Here, α is fixed to 0.1. Setting k_0 to 2 or 3 provides optimal results although, due to the reasons presented in Section 3.2, it has a limited influence on the filter components. In general, compared with other phase-based features, the proposed filter-based representation shows a superior performance. Fig. 6 illustrates the WER versus SNR and shows that the advantage of the proposed modifications is greatest in SNRs below 15 dB.

4.2.2. *LVCSR+DNN*

In order to further investigate the capabilities of the proposed parametrisation scheme, the Aurora-4 database was also used. First, HMMs were trained using only clean data. As seen in Table 2, despite returning notably better results in the noisy condition (5.8% absolute WER reduction, on average), the proposed feature (α-BMFGDVT-0.1-2) in the clean condition performs as well as MFCCs. Finally, BN features were extracted from the proposed phase-based representation and MFCCs in multi-style trained mode. Although in such circumstance MFCC work quite well, the proposed feature slightly outperforms it.

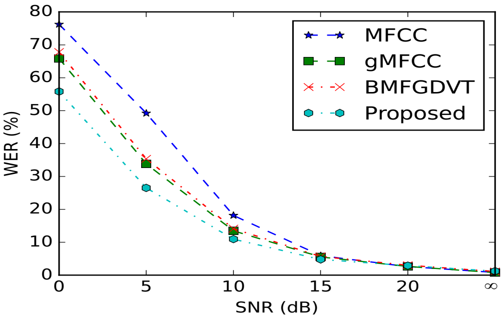

Figure 6: *Performance of different features vs SNR for Aurora-2 task (averaged over A, B and C testsets).*

Table 1: *WER (average 0-20 dB) for Aurora-2 [32].*

Feature	TestSet A	TestSet B	TestSet C
MFCC-E-D-A	32.7	27.5	34.0
γ-MFCC-0.075	24.6	23.8	23.1
PLP	32.7	29.4	33.8
MODGDF [18]	35.7	33.6	40.5
CGDF [17]	33.0	27.0	40.6
PS [14]	34.0	28.8	35.4
ARGDMF [15]	24.6	21.0	24.0
BMFGDVT [22]	**26.8**	**22.6**	**26.6**
α-BMFGDVT-0.12	22.8	20.8	20.7
α-BMFGDVT-0.1	**22.1**	**19.5**	**20.5**
α-BMFGDVT-0.08	22.3	19.3	21.0
α-BMFGDVT-0.1-1	21.7	19.2	20.3
α-BMFGDVT-0.1-2	**21.5**	**18.9**	**20.3**
α-BMFGDVT-0.1-3	21.5	18.8	20.5

Table 2: *WER for Aurora-4 [33].*

Feature	A	B	C	D	*Ave*
MFCC [Clean]	7.4	34.5	29.4	50.3	39.0
Proposed [Clean]	7.1	26.5	28.1	45.1	33.2
BN-MFCC [Multi]	7.2	12.9	14.3	27.0	18.7
BN-Proposed [Multi]	7.0	12.7	14.3	26.6	18.4

5. Conclusion

In earlier work we developed a framework for source-filter separation through phase-based speech processing. This paper aimed at further clarification of the theoretical aspects of the proposed framework and also improving the efficacy and robustness of the extracted source and filter components. In this regard, the formula of the Hilbert transform was altered by substituting the log with $GenLog$, group delay was computed through a regression filter instead of sample difference and non-linearity block was integrated into $GenLog$. The proposed modifications enable the tuning of the dynamic range and statistical properties of the excitation and vocal tract representations and improve the robustness. The filter-based feature provided better performance than MFCCs in both Aurora-2 connected-digit and Aurora-4 LVCSR tasks. Combining the phase and magnitude-based features in a DNN-based setup is an avenue for future research. The pilot F_0 extraction tests from the phase source component in clean/noisy conditions also appeared to be promising and establishes another direction for future works.

6. References

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

[2] Pejman Mowlaee, Rahim Saeidi, and Yannis Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1 – 29, 2016, Phase-Aware Signal Processing in Speech Communication.

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

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

[5] E. Loweimi, S.M. 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.

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

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

[8] Leigh D. Alsteris and Kuldip 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.

[9] E. Loveimi and S.M. Ahadi, "Objective evaluation of magnitude and phase only spectrum-based reconstruction of the speech signal," in *Communications, Control and Signal Processing (ISCCSP), 2010 4th International Symposium on*, March 2010, pp. 1–4.

[10] E. Loveimi and S.M. 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.

[11] Erfan Loweimi, Seyed Mohammad Ahadi, and Hamid Sheikhzadeh, "Phase-only speech reconstruction using very short frames.," in *INTERSPEECH*. 2011, pp. 2501–2504, ISCA.

[12] K. Vijayan and K. S. R. Murty, "Analysis of phase spectrum of speech signals using allpass modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2371–2383, Dec 2015.

[13] E. Loweimi and S.M. 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.

[14] Donglai Zhu and K.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*, May 2004, vol. 1, pp. I–125–8 vol.1.

[15] E. Loweimi, S.M. 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.

[16] E. Loweimi, S.M. Ahadi, T. Drugman, and S. Loveymi, "On the importance of pre-emphasis and window shape in phase-based speech recognition," *Lecture Notes in Computer Science*, vol. 7911 LNAI, pp. 160–167, 2013.

[17] Baris Bozkurt, Laurent Couvreur, and Thierry Dutoit, "Chirp group delay analysis of speech signals," *Speech Communication*, vol. 49, no. 3, pp. 159 – 176, 2007.

[18] R.M. Hegde, H.A. Murthy, and V.R.R. 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.

[19] E Loweimi, J Barker, and T Hain, "Compression of model-based group delay function for robust speech recognition," *The University of Sheffield Engineering Symposium Conference Proceedings Vol. 1*, vol. 1, 2014.

[20] Srikanth R. Madikeri, Asha Talambedu, and Hema A. Murthy, "Modified group delay feature based total variability space modelling for speaker recognition," *I. J. Speech Technology*, vol. 18, no. 1, pp. 17–23, 2015.

[21] Karthika Vijayan, Pappagari Raghavendra Reddy, and K. Sri Rama Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Commun.*, vol. 81, no. C, pp. 54–71, July 2016.

[22] Erfan Loweimi, Jon Barker, and Thomas Hain, "Source-filter separation of speech signal in the phase domain.," in *INTERSPEECH*. 2015, ISCA.

[23] Alan V. Oppenheim and Ronald W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2009.

[24] 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.

[25] G. E. P. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society. Series B (Methodological*, pp. 211–252, 1964.

[26] Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Julien Epps, "Group delay features for emotion detection.," in *INTERSPEECH*. 2007, pp. 2273–2276, ISCA.

[27] S.S. Haykin and M. Moher, *Communication Systems*, Wiley, 2010.

[28] 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.

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

[30] Paul C. Bagshaw, Steven M. Hiller, and Mervyn A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching," in *EUROSPEECH*, 1993.

[31] Kjell A. Doksum Peter J. Bickel, "An analysis of transformations revisited," *Journal of the American Statistical Association*, vol. 76, no. 374, pp. 296–311, 1981.

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

[33] N Parihar and J Picone, "Aurora working group: Dsr front end lvcsr evaluation au/384/02," *Inst. for Signal and Information Process, Mississippi State University, Tech. Rep*, vol. 40, pp. 94, 2002.

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

[35] F. Grezl and P. Fousek, "Optimizing bottle-neck features for lvcsr," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2008, pp. 4729–4732.

[36] Karel Veselý, Lukás Burget, and Frantisek Grézl, "Parallel training of neural networks for speech recognition," in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2010, pp. 2934–2937.

[37] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Unified approach to mel-generalized cepstral analysis," in *Proc. ICSLP-94*, 1994, pp. 1043–1046.