STATISTICAL NORMALISATION OF PHASE-BASED FEATURE REPRESENTATION FOR ROBUST SPEECH RECOGNITION

Erfan Loweimi, Jon Barker and Thomas Hain

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

{eloweimi1, j.p.barker, t.hain}@sheffield.ac.uk

ABSTRACT

In earlier work we have proposed a source-filter decomposition of speech through phase-based processing. The decomposition leads to novel speech features that are extracted from the filter component of the phase spectrum. This paper analyses this spectrum and the proposed representation by evaluating statistical properties at various points along the parametrisation pipeline. We show that speech phase spectrum has a bell-shaped distribution which is in contrast to the uniform assumption that is usually made. It is demonstrated that the uniform density (which implies that the corresponding sequence is least-informative) is an artefact of the phase wrapping and not an original characteristic of this spectrum. In addition, we extend the idea of statistical normalisation usually applied for the magnitude-based features into the phase domain. Based on the statistical structure of the phase-based features, which is shown to be super-gaussian in the clean condition, three normalisation schemes, namely, Gaussianisation, Laplacianisation and table-based histogram equalisation have been applied for improving the robustness. Speech recognition experiments using Aurora-2 show that applying an optimal normalisation scheme at the right stage of the feature extraction process can produce average relative WER reductions of up to 18.6% across the 0-20 dB SNR conditions.

Index Terms: phase spectrum, robust speech recognition, phase distribution, statistical normalisation

1. INTRODUCTION

There has been a recent growth of interest in the phase-based speech processing. For example, there has been a dedicated special session in Interpeech 2014 [1], a tutorial session in Interspeech 2015 and a special issue in Speech Communication journal [2]. An expanding body of work is showing that the phase spectrum can be usefully employed in many speech processing applications, including in speech enhancement [3–6], speech reconstruction [7–11], speech recognition [12–18] and speaker recognition [19, 20]. However, integrating the phase spectrum into the speech processing pipeline is not as straightforward as for the magnitude spectrum. In contrast to the latter – which has a relatively simple structure that is easily related to speech perception – the former has a noise-like appearance with neither a clear trend nor meaningful extrema. As such interpreting the behaviour of phase and consequently modelling and extracting useful/compact representations from it remains a challenge.

One of the fundamental models in the magnitude-based speech signal processing has been the idea of source-filter deconvolution [21], for example, by cepstral liftering [22]. In [23], we proposed a novel framework for source-filter separation in the phase domain. The method was successful in segregating the vocal tract and excitation elements and its superiority in comparison with the magnitude-based approach was discussed and illustrated. The filter component of this spectrum was converted into a set of features for ASR and lead to better speech recognition performance than well-known features such as MFCC, PLP [24], MODGDF [18], PS [16] and CGDF [17].

In this paper, we aim to study the phase spectrum and its feature representations from a statistical standpoint at different stages within the parametrisation process. Such analysis will shed light on the behaviour of the phase spectrum and further clarify its structure. It will be shown that contrary to the general belief which considers phase to have a uniform distribution, it has a bell-shaped distribution in the clean condition. In addition, such findings help in devising an efficient normalisation scheme for the phase-based feature. As shown, an optimal statistical transformation can result in an absolute WER reduction of up to 4.7% on average (over 0–20 dB SNR).

This paper is structured as follows. In Section 2 the phase-based source-filter separation framework and the proposed parametrisation workflow are briefly reviewed. Section 3 is dedicated to studying the statistical attributes of the phase-based features and the utilised statistical normalisation schemes. Section 4 includes experimental results along with discussion and Section 5 concludes the paper.

2. PHASE-BASED SOURCE-FILTER SEPARATION

Speech is a mixed-phase signal as its complex cepstrum (CC) is neither causal nor anti-causal [7, 25]. As such it can be decomposed into *minimum-phase (MinPh)* $X_{MinPh}(\omega)$ and *all-pass (AllP)* $X_{AllP}(\omega)$ components

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

where $|X(\omega)|$ and $arg[X(\omega)]$ are the magnitude and unwrapped (continuous) phase spectra, respectively. For the MinPh signals the CC is causal, i.e., equals zero at negative quefrencies [25]. Based on this property, the Hilbert transform provides a one-to-one correspondence between the phase and magnitude spectra

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

where $*$ indicates convolution. By replacing the $log|X_{MinPh}|$ with $log|X(\omega)|$ based on (1), $arg[X_{MinPh}(\omega)]$ can be calculated. Equivalently, $arg[X_{MinPh}(\omega)]$ may be computed in the cepstrum domain by applying a causal lifter (Fig. 1) instead of (2).

As shown in [23], $arg[X_{MinPh}(\omega)]$, contrary to the wrapped phase ($ARG[X(\omega)]$), no longer exhibits a chaotic and trendless shape. In fact, it could be imagined as a superposition of two components, one changing slowly (*Trend*) and the other one oscillating quickly (*Fluctuation*). Accordingly, these components can be decomposed based on the difference in their rate of change. By

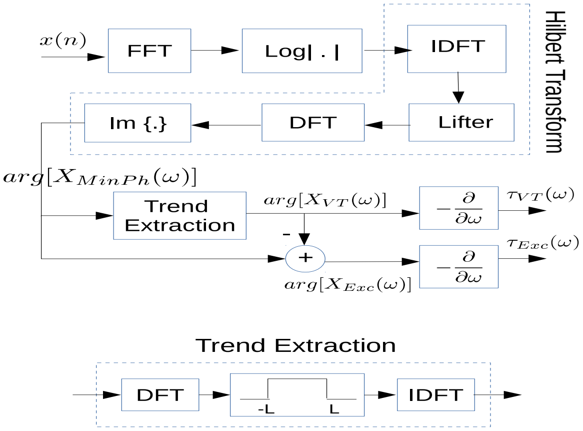

Fig. 1. Workflow of the proposed source-filter decomposition in the phase domain [23].

smoothing (low-pass filtering) the $arg[X_{MinPh}(\omega)]$, Trend can be extracted. As explained in [23], Trend and Fluctuation correspond to the vocal tract (VT) and excitation (Exc) components, respectively. Fig. 1 shows the proposed source-filter separation process.

For evaluating the efficacy of the suggested approach, a feature (named *BMFGDVT* [23]) was extracted from the vocal tract (filter) component of the phase spectrum and tested for its usefulness in ASR (Fig. 2). The performance was better than the listed features despite the relative simplicity of the parametrisation process. For a detailed discussion readers are referred to [23]. In the next section we study the statistical structure of this representation and try to improve its performance through statistical normalisation.

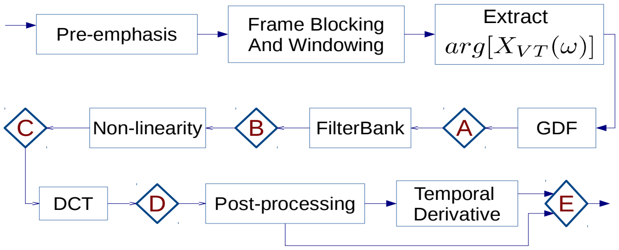

Fig. 2. The proposed phase-based front-end in [23].

3. STATISTICAL PROPERTIES OF THE PHASE-BASED REPRESENTATIONS

3.1. Statical Normalisation

Variability in the data representation due to nuisance factors is a significant issue in pattern recognition posing considerably more difficulties for the back-end in mapping the feature onto the correct class. This problem could be alleviated by developing either a more robust front-end or back-end. In the front-end, one sensible approach could be applying some knowledge about the properties of the clean data to mitigate the effect of unwanted disturbances. This involves evaluating the behaviour of the extracted pattern in a noise-free condition and embedding such knowledge into the pipeline, in a principled way, to attenuate the deviations induced by noise.

Prior knowledge about the clean data could have a deterministic or statistical basis. The latter is more effective in dealing with the variability problem and is added to the parametrisation process as a normalisation block aiming at giving a *desired* statistical property to the features. From the back-end standpoint, desired features are those which, among other things, are in harmony with the assumptions it makes about its input. Although any mismatch could be costly performance-wise, transforming the data by only considering the back-end could be problematic, too. In fact, there is the possibility of distorting the features through imposing some properties on them which do not comply with their original structure. So, an optimal normalisation should take both ends into account.

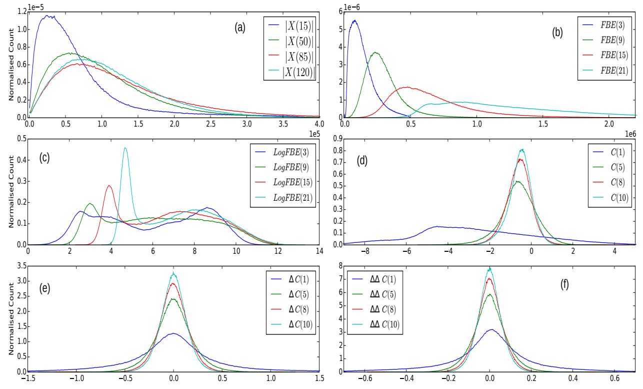

Fig. 3. Histograms at different stages for MFCC.

Here we aim at investigating the behaviour of the phase and its representations from statistical standpoint and extending the idea of statistical feature normalisation to the phase-based representations. So far, such techniques have been employed mainly for magnitude-based features. Estimation of the statistical structure of the phase in the clean condition provides a fresh perspective on the behaviour of this spectrum, helps in explaining the reason behind success/failure of each normalisation scheme and paves the way towards finding the optimal one.

3.2. Distribution of the phase-based representations

For evaluating the distribution of the phase spectrum and its representations, the histogram at various points along the proposed workflow (Fig. 2) was computed. In order to get statistically significant results, all the (clean) training data of Aurora 2 [26] has been employed which includes 8440 waves yielding more than 1.4 M frames. For comparison, the same process has been done for MFCC.

As shown in Fig. 3(a), the distribution of the magnitude spectrum is heavily right-skewed and may be thought of as a Rayleigh density, the assumption which is usually made in speech enhancement. Taking *Log* of the filter bank energies (FBE) results in a bimodal distribution. The left mode relates to the low-energy speech/silence and the right one is connected to the speech parts with a normal energy level. As seen, applying the *Log* has a significant statistical impact on the FBEs and pushes the distribution toward the Gaussian by decreasing both skewness and kurtosis. This allows the GMM-based back-end to obtain a much better fit.

Figure 4 demonstrates the histograms of the phase-based feature at different points across the parametrisation workflow. In a sharp contrast to the uniform assumption usually made about the phase spectrum, especially in the speech enhancement literature, $arg[X_{MinPh}(\omega)]$ has a bell-shaped distribution (Fig. 4(a)). On the other hand, as Fig. 5 shows, uniform density($U(-\pi,\pi)$) is a correct choice for the distribution of $ARG[X(\omega)]$ (principle phase) but this is the outcome of wrapping not an inherent property of phase.

In order to prove this point we first wrap the magnitude spectrum similar to the phase spectrum as follows

$$wrapped\,|X(\omega)| = (|X(\omega)| + \pi)\ mod\ 2\pi - \pi, \qquad (3)$$

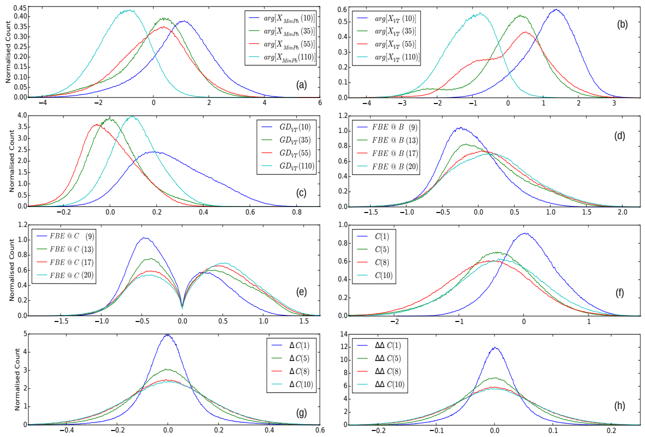

Fig. 4. Histograms of the phase spectrum and its representations along the workflow shown in Fig. 2.

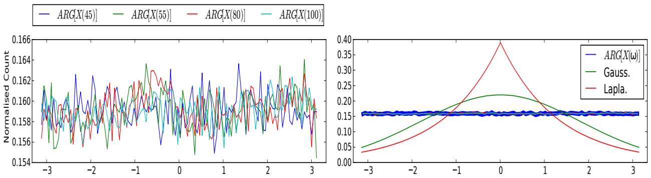

Fig. 5. Histogram of the principle phase spectrum, $ARG[X(\omega)]$.

and plot its histogram. Fig. 6 shows that the distribution of the *wrapped magnitude spectrum* is $U(-\pi, \pi)$, too. This corroborates the claim that the uniform density is only due to the wrapping. Second, a data sequence with uniform distribution over its support has the maximum-level of randomness (entropy) and is least-informative from an information theory viewpoint [27]. In our empirical study, it means that after making more than 1.4 M observations, still all the possible values of the phase are equiprobable and this implies that phase is a random information-less sequence.

In fact, making a uniform assumption for phase density creates two paradoxes: First, under some mild conditions, a signal is recoverable (up to a scale error) from its phase spectrum [8] and a non-informative uniformly-distributed sequence should not have such capability. Second, there is a one-to-one relationship between the magnitude and phase spectra of a MinPh signal. This implies that they carry the same amount of information and are just two mathematical realisations of the same information (knowledge). The apparent uniform distribution of the phase coupled with the one-to-one phase/magnitude relation imply that the magnitude spectrum is also devoid of information, which is obviously incorrect. Fig. 4(a), however, shows the true distribution and resolves these contradictions. Comparing it with Figs. 5 and 6 demonstrates that the uniform distribution is not a structural property of the phase spectrum and is merely a repercussion of phase wrapping.

After applying the non-linearity, the FBEs at point C (Fig. 2), $FBE @ C$, become bimodal (Fig. 4(e)) similar to the log of FBEs in MFCC, although, this time the underlying reason is different. Contrary to the magnitude spectrum, phase and its representations are scale-blind and energy level has no role to play. The reason here stems from the fact that zero is a fixed-point of the power transformation used as nonlinearity ($sign(x)|x|^a$, where x is $FBE @ B$). As such, points with a value very close to zero remain identical whereas others move away and this gives rise to bimodality.

Finally, we need to investigate where the bell-shaped distribution of the phase-based features (Fig. 4(f)) stands in comparison with the Gaussian density. As seen in Fig. 7, although the Gaussian assumption seems to be a reasonable approximation, the distribution of this feature has a higher kurtosis and is Leptokurtic or super-Gaussian. The Laplace distribution is an example of such densities. As juxtaposed in Fig. 7, it forms the upper bound from this perspective and could be regarded as another approximation for the feature's distribution. This implies that both Gaussianisation and Laplacianisation might be helpful in pushing the noisy phase-based features toward their clean counterpart and hence be useful in improving the robustness. This hypothesis will be validated in Section 4 but before that we briefly review how the normalisations are implemented.

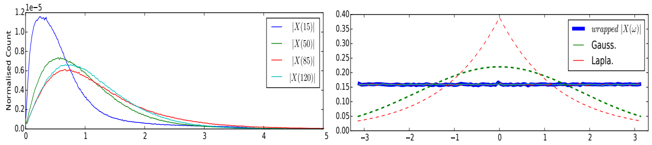

Fig. 6. Effect of wrapping on the density of the magnitude spectrum.

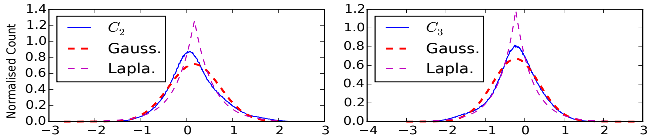

Fig. 7. Comparison of histograms of the proposed phase-based feature (C) with Gaussian and Laplacian densities.

3.3. Implementing the Statistical Transformation

This subsection briefly overviews the statistical normalisation schemes which are used in the experiments, namely, Gaussianisation [28], Laplacianisation and histogram equalisation (HEQ) [29]. These techniques are computationally low-cost as neither need noise estimation nor stereo data and no explicit expression for how the features get contaminated by noise is required. The equation which underpins them mathematically is as follows

$$CDF_Y(y) = CDF_X(x) \Rightarrow x = CDF_X^{-1}\left(CDF_Y(y)\right), \qquad (4)$$

where X and Y are random variables (rv) associated with the clean and noisy observations, respectively, and x and y are their realisations. Implementing (4) involves finding the quantile function of X, i.e., $CDF_X^{-1}(x)$, and cumulative distribution function of Y.

If rv Z is defined as $Z = CDF_Y(Y)$, Probability Integral Transform (PIT) [30] shows that it follows $U(0, 1)$. However, computing the quantile function of the clean (reference) features, is not straightforward and a closed-form expression is only available for a few density functions. In practice mostly numerical techniques are employed. Table-based HEQ is an example of such methods in which this function is estimated from the training data. HEQ does not make any assumption about the target distribution, contrary to the Gaussianisation and Laplacianisation, and this turns it into a more flexible approach. For the Gaussian and Laplace distributions, the closed-form expression for the quantile function exists

$$\begin{cases} Gaussianisation \;\rightarrow\; x_i = \sqrt{2}\, erf^{-1}(2z_i - 1) \\ Laplacianisation \;\rightarrow\; x_i = \begin{cases} ln(2z_i), & z_i < 0.5 \\ -ln(2 - 2z_i), & z_i \geq 0.5, \end{cases} \\ z_i = \frac{r_i - \beta}{N} \;\;,\; i = 1,\; 2,\; \dots,\; N \end{cases} \qquad (5)$$

where erf^{-1}, ln, z, N and r_i denote inverse error function, natural logarithm, realisation of rv Z, number of observations and the rank of y_i after ascending sort, respectively. β is used to avoid extreme values and usually set to 0.5 [28]. Note that for mathematical convenience, normalisation is (sub-optimally) carried out for each dimension independently. The difference of these techniques to mean-variance-normalisation should also be noted: The former affects all the moments while the latter only touches the first- and second-order statistics. As such they have a deeper statistical impact.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1. Parametrisation Process

Parameter setting is identical to [23]. Aurora-2 [26] is employed as the database and HMMs were trained from the clean data using HTK [31] based on the Aurora-2 standard (simple) recipe. The effect of Gaussinisation (Gaus), Laplacianisation (Lap) and HEQ at different stages, signified by A to E at Fig. 2, is investigated. Recognition results are reported in Tables 2-4 (average 0-20 dB) and Fig. 8 (versus SNR). BMFGDVT [23] is considered as the baseline.

4.2. Discussion

Comparing Table 1 with Tables 2-4 shows that normalisation, in most cases, enhances recognition performance, although the amount of improvement depends on the type of normalisation and the stage at which it is performed. As seen, point E, namely just before the back-end, appears to be the best place for carrying out the normalisation. Gaussianisation at this point decreases relative word error rate reduction (RER) by up to 18.6% which is a noteworthy gain, considering the low computational overhead involved.

Comparison of Table 2 and 3 show that Gaussainisation is a more effective normalisation scheme than the Laplacianisation. Two points can help explain this. First, as depicted in Fig. 7, although the true distribution of the phase-based feature is super-Gaussian, it is not as Leptokurtic as Laplacian and is closer to Gaussian. As a result, Gaussianisation leads to less distortion than Laplacianisation because it is more consistent with the original statistical structure of the features. The second reason is that the GMM-based back-end better fits data with a Gaussian distribution.

Comparing Tables 2 and 3 with 4 shows that both Gaussianisation and Laplacianisation return better results than HEQ. Three points should be noted: First, HEQ assumes that the noise-corruption process is a monotonic transform and does not cause any information loss. This demand is not met here due to the random effect of the noise. Second, the simple Table-based HEQ approach utilised here is not state-of-the-art; more advanced HEQ techniques might lead to better results. Third, as shown in Figure 7, the true distribution of the phase-based feature is relatively close to both Gaussian and Laplacian distributions. Thus both can approximate it to a reasonable extent and the flexibility of HEQ is unnecessary.

Figure 8 demonstrates the performance of these techniques (after applying each one at the corresponding optimal point) versus SNR (averaged over all test sets). As seen, these methods are especially useful in SNRs below 10 dB and return absolute accuracy improvement of over 7% and 10% in SNRs of 5 and 0 dB, respectively.

Table 1. *Average (0-20 dB) recognition rates for Aurora-2 [23].*

Feature	TestSet A	TestSet B	TestSet C	Ave. All
MFCC	66.2	71.4	64.9	67.5
PLP	67.3	70.6	66.2	68.0
MODGDF	64.3	66.4	59.5	63.4
CGDF	67.0	73.0	59.4	66.5
PS	66.0	71.2	64.6	67.3
Baseline	**73.2**	**77.4**	**73.4**	**74.7**

Table 2. *Average accuracy after Gaussianisation at points $A - E$.*

Feature	A	B	C	Ave. All	RER(%)
Gaus-A	74.1	78.3	74.4	75.6	3.6
Gaus-B	73.0	76.0	74.1	74.4	-1.9
Gaus-C	74.0	76.7	74.9	75.2	2.0
Gaus-D	78.6	80.2	77.0	78.6	15.4
Gaus-E	79.3	81.0	77.8	**79.4**	18.6

Table 3. *Average accuracy after Laplacianisation at points $A - E$.*

Feature	A	B	C	Ave. All	RER(%)
Lap-A	74.4	78.5	74.8	75.9	4.7
Lap-B	73.9	76.7	74.8	75.1	1.6
Lap-C	74.0	76.7	75.2	75.3	2.4
Lap-D	75.5	77.5	74.0	75.7	4.0
Lap-E	77.5	79.3	75.9	**77.6**	11.5

Table 4. *Average accuracy after HEQ at points $A - E$.*

Feature	A	B	C	Ave. All	RER(%)
HEQ-A	74.0	78.0	74.9	75.6	3.5
HEQ-B	74.2	78.0	75.2	75.8	4.3
HEQ-C	74.5	78.4	75.4	76.1	5.5
HEQ-D	76.5	78.2	73.5	76.1	5.5
HEQ-E	77.0	78.7	74.9	**76.9**	8.7

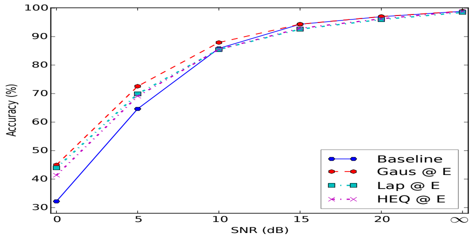

Fig. 8. Accuracy versus SNR for different normalisation schemes.

5. CONCLUSION

After developing a framework for source-filter separation through phase spectrum manipulation, we proposed a feature extraction algorithm from the vocal tract component of phase. In this paper, the statistical properties of this spectrum and its representations at different points along the parametrisation process was studied. It was observed that contrary to the general assumption made about the phase spectrum, its density has a bell-shaped structure and is not uniform. It is also argued that the uniform distribution is an artefact of phase wrapping and is not an intrinsic characteristic of this spectrum. We tried three statistical normalisation schemes to improve the performance of the proposed feature in noisy condition. Analysis of the statistical structure of the phase-based representations provided an understanding of the reason behind success/failure of each approach and helped in selecting the best normalisation scheme. It was observed that optimal statistical normalisation at the right stage can have a remarkable impact on the performance of the phase-based features. This suggests that taking the statistical properties of the phase spectrum into account has significant potential to improve the performance of the phase-based techniques in speech processing and is an avenue for future research.

6. REFERENCES

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

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

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

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

[8] 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.

[9] 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.

[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] 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.

[12] 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*, 2001, vol. 1, pp. 133–136 vol.1.

[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] 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.

[15] 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.

[16] D. 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.

[17] B. Bozkurt, L. Couvreur, and T. 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] R. Padmanabhan, S. H. K. Parthasarathi, and H. A. Murthy, "Robustness of phase based features for speaker recognition.," in *INTERSPEECH*. 2009, pp. 2355–2358, ISCA.

[20] S. R. Madikeri, A. Talambedu, and H. 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] G. Fant, *Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations*, Description and Analysis of Contemporary Standard Russian. De Gruyter, 1971.

[22] D. G. Childers, D. P. Skinner, and R. C. Kemerait, "The cepstrum: A guide to processing," *Proceedings of the IEEE*, vol. 65, no. 10, pp. 1428–1443, Oct 1977.

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

[24] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

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

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

[27] David J. C. MacKay, *Information Theory, Inference & Learning Algorithms*, Cambridge University Press, New York, NY, USA, 2002.

[28] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *2001: A Speaker Odyssey - The Speaker Recognition Workshop*, Crete, Greece, 2001, pp. 213–218, International Speech Communication Association (ISCA).

[29] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, May 2005.

[30] P. G. Hoel, S. C. Port, and C. J. Stone, *Introduction to Probability Theory*, Houghton Mifflin series in statistics. Houghton Mifflin, 1971.

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