EXPLORING THE USE OF GROUP DELAY FOR GENERALISED VTS BASED NOISE COMPENSATION

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 studied the effect of statistical normalisation for phase-based features and observed it leads to a significant robustness improvement. This paper explores the extension of the generalised Vector Taylor Series (gVTS) noise compensation approach to the group delay (GD) domain. We discuss the problems it presents, propose some solutions and derive the corresponding formulae. Furthermore, the effects of additive and channel noise in the GD domain were studied. It was observed that the GD of the noisy observation is a convex combination of the GDs of the clean signal and the additive noise and also in the expected sense, channel GD tends to zero. Experiments on Aurora-4 showed that, despite training only on the clean speech, the proposed features provide average WER reductions of 0.8% absolute and 4.1% relative compared to an MFCC-based system trained on the multi-style data. Combining the gVTS with a bottleneck DNN-based system led to average absolute (relative) WER improvements of 6.0% (23.5%) when training on clean data and 2.5% (13.8%) when using multi-style training with additive noise.

Index Terms— Robust ASR, generalised VTS, phase spectrum, group delay, product spectrum

1. INTRODUCTION

The speech phase spectrum has recently received renewed attention. An expanding body of work propounds that phase can be employed in a multitude of applications [1], including in speech reconstruction [2,3], speech enhancement [4,5], robust speech recognition [6–10] and speaker recognition [11].

We recently developed a source-filter model in the phase domain [12, 13] which further sheds light on the phase structure, clarifies how it encodes the speech information and successfully separates the vocal tract and the excitation components through phase-based signal manipulation. Moreover, in [14], we scrutinised the statistical characteristics of the phase spectrum and its representations along the feature extraction pipeline in the clean condition. It was demonstrated that the unwrapped phase spectrum has a bell-shaped distribution. Also, the efficacy of statistical normalisation of the phase-based features was evaluated and lead to significant performance improvement in ASR.

Such gain in robustness motivates us to explore applying more advanced statistical techniques like VTS [15] and its generalised version (gVTS) [16, 17] for building robust phase-based features. In this paper, we investigate the problems encountered when extending the gVTS framework to the phase/group delay domain, propose some solutions and derive the corresponding formulae. Experimental results conducted on Aurora-4 [18] confirm the success of this approach in dealing with both additive noise and channel distortion.

The rest of this paper is organised as follows. Section 2 is dedicated to deriving the environment model in the GD domain and examining the effect of the additive noise. In Section 3, the problems of extending the (g)VTS formulae to the GD domain are investigated and some solutions are proposed. Section 4 derives the gVTS equations and Section 5 contains the experimental results as well as discussion. Finally, Section 6 concludes the paper.

2. ENVIRONMENT MODEL IN THE GROUP DELAY DOMAIN AND THE ADDITIVE NOISE EFFECT

In the (g)VTS approach to robust ASR, there is a need for an environment model which shows how the clean signal gets contaminated with the noise. The general model takes the form of $Y(\omega) = X(\omega)H(\omega) + W(\omega)$ where ω, Y, X, H and W are the radial frequency, (short-time) Fourier transforms (FT) of the noisy observation, clean signal, channel and additive noise, respectively. Assuming speech and noise are uncorrelated and using periodogram power spectrum estimation

$$|Y|^2 = |X|^2\,|H|^2 + |W|^2 \tag{1}$$

where $|.|^2$ denote the periodogram. With some algebraic manipulation, it can be shown that the group delay of the noisy observation, τ_Y, takes the following form

$$\tau_Y = \frac{|X|^2|H|^2}{|Y|^2}\,(\tau_X + \tau_H) \;+\; \frac{|W|^2}{|Y|^2}\tau_W. \tag{2}$$

Equation (2) shows the environment model in the group delay domain and underpins the relation between the group delay of the noisy observation with other variables.

The additive noise emerges as an additive term in the periodogram domain whereas in the phase and group delay domain it has a different effect. To study its effect and for the

sake of argument, let us assume that there is no channel distortion ($H = 1$). In this case,

$$\tau_Y = \frac{\xi}{1+\xi}\,\tau_X \;+\; \frac{1}{1+\xi}\tau_W \;=\; c\,\tau_X + (1-c)\,\tau_W \quad (3)$$

where $\xi = \frac{|X|^2}{|W|^2}$ is *a priori* signal-to-noise ratio (SNR) and $c = \frac{\xi}{1+\xi}$. As seen, the noisy observation in the group delay domain is a *convex* combination of the clean part and the additive noise while in the periodogram domain it is just the sum of the corresponding power spectra of these two components.

3. DIFFICULTIES WITH (g)VTS IN GD DOMAIN

Having derived the environment model in the group delay domain, we now wish to extend the idea of (g)VTS to this domain. However, due to some properties of the environment model and the group delay, there are issues which should be addressed and resolved in advance.

3.1. Larger Number of Variables

For noise compensation using the (g)VTS framework, as well as the environment model in the target domain, the statistical distribution of all the involved variables is needed. While in the periodogram domain there are only four quantities (1), (2) shows that in the GD domain the environment model contains eight variables. Hence, eight probability distribution functions should be estimated. Considering eight variables instead of four, complicates the compensation process.

To decrease the number of variables, two factors can be considered: First, the variables that overlap in terms of the information they carry and are added/multiplied together can be re-expressed via one variable. Second, a term containing a variable that tends to zero in the expected sense, e.g. cross-correlation of speech and noise in (1), may be removed. In the work presented here we have used both of these points.

In this regard, let us multiply both sides of (2) by $|Y|^2$

$$|Y|^2\,\tau_Y = |X|^2|H|^2\,(\tau_X + \tau_H) \;+\; |W|^2\,\tau_W. \quad (4)$$

In general, $|Z|^2$ and τ_Z are not independent and actually, for many signals they are closely linked together. Therefore, it appears reasonable to encapsulate the multiplication of $|Z|^2\;\tau_Z$ into a single variable Q_Z to represent the information encoded in each one. This quantity was called group delay-power *product spectrum* (PS) in [19]. Accordingly,

$$Q_Y = Q_X\,|H|^2 + Q_H\,|X|^2 + \;Q_W \quad (5)$$

where Q_Z is the product spectrum of Z for $Z \in \{Y, X, W\}$. This decreases the number of variables from eight to six.

3.2. Dynamic Range Compression

The dynamic range of the product spectrum is comparable to the periodogram. So, it should be compressed using functions like log or power transformation (z^α) before statistical modelling. However, the admissible range for these functions is strictly restricted to the positive values. Although the power spectrum is always positive, the GD and subsequently the product spectrum may have negative value in some time-frequency bins. So, one needs to deal with the negative values before applying the compression function.

Taking the absolute value is not an appropriate solution as it makes some of the negative values larger than the small positive ones. This distorts the relative order/rank of the samples. The other possible solution which has been used for compressing the group delay in [7, 12] is to implement compression using $sign(x)\,|x|^\alpha$, inspired by [20]. Although this approach preserves the relative order, it poses two problems for a (g)VTS-based noise compensation process: first, the clean part can not be factored out

$$\begin{aligned} sign(Y)|Y|^\alpha &= sign(XH+W)\;|XH+W|^\alpha \\ &\neq \tilde{X}\;\breve{G}(\tilde{X},\tilde{H},\tilde{W}) \end{aligned} \quad (6)$$

where $\tilde{Z} = sign(Z)\,|Z|^\alpha$ for $Z \in \{X, H, W\}$, $sign$ indicates the signum function and $\breve{G}$ denotes the distortion function. Second, computing the Jacobians becomes complicated.

Another option which preserves the rank without complicating the factorisation and Jacobian computation is to add a constant, c, to the product spectrum to ensure it remains positive in all bins. However, finding the optimal c is problematic: setting it to the minus of the minimum value of the utterance causes inter-utterance variability whereas choosing a universal large enough value causes the compression function to operate in its saturation region, namely $(Q_Z + c)^\alpha \approx c^\alpha$.

Flooring is another possible solution in which values below a preset threshold are clipped. A potential pitfall of this technique is that it can lead to information loss. However, this is tolerable as long as the discarded data plays an insignificant role. Plotting the product spectrum illustrates that the majority of the activity occurs on the positive side. Therefore, flooring can be safely performed with negligible information loss. The floor function takes the form of $floor(z;\theta_z) = max(z,\theta_z)$ where θ_z is a tunable threshold.

After filtering out the negative values, the compression function can be applied. Using the power transformation

$$\breve{Q}_Y = \breve{Q}_X\;\underbrace{\breve{H}\left(1 + (\frac{\breve{Q}_H\;\breve{X}}{\breve{Q}_X\;\breve{H}})^{\frac{1}{\alpha}} + (\frac{\breve{Q}_W}{\breve{Q}_X\breve{H}})^{\frac{1}{\alpha}}\right)^{\alpha}}_{\breve{G}(\breve{Q}_X,\breve{Q}_H,\breve{Q}_W,\breve{X},\breve{H})} \quad (7)$$

where $\breve{Q}_Z = (floor(Q_Z;0))^\alpha$, $\breve{Z} = (|Z|^2)^\alpha$ for $Z \in \{X, H, W\}$ and $\breve{G}$ indicates the distortion function.

The dynamic range compression issue is solved but there are still six variables whereas the environment model in the power spectrum domain includes only four. The two extra variables are related to the term $Q_H|X|^2 = \tau_H\,|H|^2\,|X|^2$ in (5). Without this term, the equation resembles that of the periodogram domain and this facilitates re-deriving the (g)VTS formulae in the product spectrum domain. In general, neither $|X|$ nor $|H|$ are zero. However, the spectral behaviour of the group delay of the channel, τ_H, is unclear.

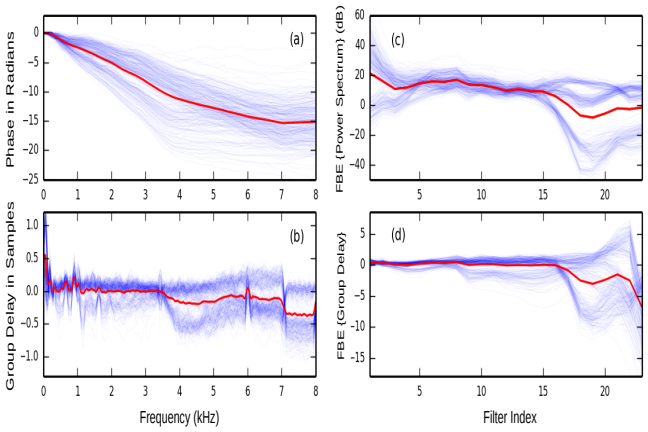

Fig. 1. Channel behaviour in the frequency domain before and after applying the filter bank, the red curve shows the average over all utterances. (a) unwrapped phase spectrum, (b) group delay, (c) $FBE\{|H|^2\}$, (d) $FBE\{\tau_H\}$.

3.3. Channel Phase Spectrum and Group Delay

To investigate the properties of $\tau_H(\omega)$, there is a need for a database of impulse responses of different channels. Here, we make use of the test sets A and C of the Aurora-4 database [18]. Both sets include 330 utterances with an average length of 7.3 seconds. Signals in the test set A were recorded using a close-talking microphone whereas in the test set C the same speech was simultaneously recorded by a different desktop microphone. Reportedly, 18 desktop microphones have been used in the recording process [18].

For the sake of argument let us assume that the microphone used in the test set A is ideal. This allows us to treat sets A and C as stereo data and facilitates channel estimation,

$$\begin{cases} \text{Test Set A} \;\Rightarrow\; Y^A = X \\ \text{Test Set C} \;\Rightarrow\; Y^C = X\ H \end{cases} \quad \Rightarrow \quad H_t = \frac{Y_t^C}{Y_t^A} \tag{8}$$

where Y^A and Y^C denote the short-time FT of the corresponding signals in the test sets A and C, H indicates FT of the channel and t is the frame index. Having computed H for each frame, the phase spectra as well as the group delay can be calculated. Averaging over the utterance frames produces an estimate of the channel behaviour.

Fig. 1 depicts the phase spectrum, group delay and the filter bank energies (FBE) after passing $|H|^2$ and τ_H through the Mel filter bank, computed for all the 330 utterances along with the overall mean. As seen, in the expected sense, τ_H tends to zero before/after the filter bank, which permits the removal of $Q_H|X|^2$ from (5). Consequently,

$$Q_Y \approx Q_X\ |H|^2 + \ Q_W. \tag{9}$$

4. gVTS IN THE PRODUCT SPECTRUM DOMAIN

In the conventional VTS, the *log* function is used for compression whereas in gVTS, the power transformation (or generalised logarithmic function [21] also known as Box-Cox transformation [22]) is employed. Applying the power transformation (z^α) with parameter α, provides one more degree of freedom which is helpful in adjusting the statistical properties of the features. In ASR, it has been utilised for improving the robustness in features like PLP [23], Genenralised-MFCC (gMFCC) [21], PNCC [24] and the modified group delay [7].

For implementing the gVTS, one needs the statistical model of the involved variables, an estimate of the (additive/channel) noise and an estimation criterion. For modelling the clean features a GMM with M Gaussians is employed and each noise type is modelled through a single Gaussian

$$\breve{Q}_X \sim \sum_{m=1}^{M} p_m^{\breve{Q}_X}\ \mathcal{N}(\mu_m^{\breve{Q}_X}, \Sigma_m^{\breve{Q}_X}) \tag{10}$$
$$\breve{Q}_W \sim \mathcal{N}(\mu^{\breve{Q}_W}, \Sigma^{\breve{Q}_W}) \qquad \breve{H} \sim \mathcal{N}(\mu^{\breve{H}}, \Sigma^{\breve{H}})$$

where $p_m^{\breve{Q}_X}$, μ and Σ denote the weight, mean vector and (diagonal) covariance matrix, respectively. Using minimum mean square error (MMSE) as the estimation criterion

$$\hat{Q}_X^{MMSE} = \breve{Q}_Y \sum_{m=1}^{M} P(m|\breve{Q}_Y)\ \frac{1}{\breve{G}(\mu_m^{\breve{Q}_X}, \mu^{\breve{Q}_W}, \mu^{\breve{H}})} \tag{11}$$

where $\breve{G}$ denotes the distortion function, defined in (7). The only missing part in (11) is the $P(m|\breve{Q}_Y)$, and to compute it, the statistics of $\breve{Q}_Y$ should be estimated.

Similar to $\breve{Q}_X$, it is assumed that $\breve{Q}_Y$ follows a GMM distribution with M components. This recasts the problem into computing the GMM of $\breve{Q}_Y$, namely $\{p_m^{\breve{Q}_Y}, \mu_m^{\breve{Q}_Y}, \Sigma_m^{\breve{Q}_Y}\}$. The statistics of $\breve{Q}_Y$ should be computed given those of $\breve{Q}_X$, $\breve{Q}_W$, $\breve{Q}_H$ and the environment model in the target domain, namely (7). However, due to the non-linearity, this can not be done analytically. The first-order Taylor series is used to approximately linearise this non-linear relationship

$$\begin{aligned}\breve{Q}_Y \approx \breve{Q}_{Y_0} &+ J^{\breve{Q}_X}(\breve{Q}_X - \breve{Q}_{X_0}) \\ &+ J^{\breve{Q}_W}(\breve{Q}_W - \breve{Q}_{W_0}) + J^{\breve{H}}(\breve{H} - \breve{H}_0)\end{aligned} \tag{12}$$

where J^Z is the partial derivative (Jacobian) of $\breve{Q}_Y$ with respect to Z for $Z \in \{\breve{Q}_X, \breve{Q}_W, \breve{H}\}$ and $(\breve{Q}_{X_0}, \breve{Q}_{W_0}, \breve{H}_0)$ is the point about which (7) is linearised.

In practice, the linearisation is performed around the means of the Gaussians, namely $(\mu_m^{\breve{Q}_X}, \mu^{\breve{Q}_W}, \mu^{\breve{H}})$ i.e., M points. With some algebraic manipulation

$$J_m^{\breve{Q}_X} = \frac{\partial \breve{Q}_Y}{\partial \breve{Q}_X} = diag\{\frac{\mu^{\breve{H}}}{(1+\breve{V}_m)^{1-\alpha}}\} \tag{13}$$

$$J_m^{\breve{Q}_W} = \frac{\partial \breve{Q}_Y}{\partial \breve{Q}_W} = diag\{(\frac{\breve{V}_m}{1+\breve{V}_m})^{1-\alpha}\} \tag{14}$$

$$J_m^{\breve{H}} = \frac{\partial \breve{Q}_Y}{\partial \breve{H}} = diag\{\frac{\mu_m^{\breve{Q}_X}}{(1+\breve{V}_m)^{1-\alpha}}\} \tag{15}$$

where $\breve{V}_m = (\frac{\mu^{\breve{Q}_W}}{\mu_m^{\breve{Q}_X}\mu^{\breve{H}}})^{\frac{1}{\alpha}}$. Now, the GMM of $\breve{Q}_Y$ can be estimated: linear relationship implies $p_m^{\breve{Q}_Y} \approx p_m^{\breve{Q}_X}$ and

$$\mu_m^{\breve{Q}_Y} \approx \mu_m^{\breve{Q}_X}\mu^{\breve{H}}(1+(\frac{\mu^{\breve{Q}_W}}{\mu_m^{\breve{Q}_X}\mu^{\breve{H}}})^{\frac{1}{\alpha}})^{\alpha} \quad (16)$$

$$\Sigma_m^{\breve{Q}_Y} \approx J_m^{\breve{Q}_X}\Sigma_m^{\breve{Q}_X}{J_m^{\breve{Q}_X}}^T + J_m^{\breve{Q}_W}\Sigma^{\breve{Q}_W}{J_m^{\breve{Q}_W}}^T + J_m^{\breve{H}}\Sigma^{\breve{H}}{J_m^{\breve{H}}}^T.$$

Extension of the modelling to the cepstrum domain can be easily carried out similarly to [16]. Since the overall performance does not differ noticeably, to save space only the frequency-domain formulation is provided here.

5. EXPERIMENTAL RESULTS

5.1. Set-up and Parametrisation

ASR experiments were conducted on the Aurora-4 [18] database. HMMs were trained with 16 components per mixture and all the acoustic models were standard phonetically state-clustered triphones trained from scratch using a standard HTK regime [25]. The test set consists of 4 subsets: clean, (additive) noisy, clean with channel mismatch and noisy with channel mismatch, referred to as A, B, C and D, respectively. As well as the clean (*CL*) training data, Aurora-4 has two extra sets for multi-style training, namely Multi1 (*M1*) and Multi2 (*M2*). Training data in the former is contaminated with only the additive noise whereas in the latter both additive and channel noise are present. For the DNN part, the network consists of four hidden layers with 1300 nodes, followed by a bottleneck (BN) [26] layer containing 26 nodes placed before the output layer. The network was trained using TNET [27].

The feature vector is augmented by c_0, delta and acceleration coefficients. M was set to 512 and the mean vector of the additive noise was estimated via the median of the first/last 50 frames. The channel noise was estimated using the method we proposed in [17] using three iterations. The product spectrum (PS) was parametrised in an MFCC-like framework through replacing the periodogram with the product spectrum [19]. A generalised PS (gPS) feature was also calculated by replacing the log with the power transformation.

5.2. Discussion

Table 1 shows the word error rate (WER) for different test sets. It provides a remarkable accuracy improvement in the noisy condition (test sets B-D) along with some WER reduction in the clean-matched condition (test set A). This means it enhances both robustness and discriminability of the features.

The optimal value for the parameter α depends on SNR and distortion type. In general, $0.05 - 0.1$ appears to be an optimal range and the higher the α the better the performance in the noisy condition and the lower the accuracy in the clean condition. Note also that, on average, the system trained on only clean data based on the proposed approach outperforms the one trained on multi-style training data (both M1 and M2) using MFCCs.

Table 1. *WER for Aurora-4 (HMMs trained on clean data).*

Feature	α	A	B	C	D	Ave
MFCC-CL	log	7.0	33.7	23.6	49.9	28.6
MFCC-M1	log	9.1	18.4	23.4	35.9	21.7
MFCC-M2	log	10.7	17.0	19.1	31.3	19.5
PS	log	7.1	33.7	23.7	49.9	28.6
gPS	0.05	7.0	25.3	23.2	42.9	24.6
gPS	0.1	8.1	22.1	25.6	40.8	24.1
gVTS	0.05	**6.5**	20.2	**13.9**	34.3	**18.7**
gVTS	0.075	7.1	19.8	15.0	34.0	19.0
gVTS	0.1	7.4	**19.6**	15.4	**33.9**	19.1

Table 2. *WER for BN trained on clean and multi-style data.*

Feature	α	A	B	C	D	Ave
BN{gPS}-CL	0.1	5.5	24.2	26.8	45.4	25.5
BN{gVTS}-CL	0.1	**4.6**	**20.6**	**16.0**	**36.7**	**19.5**
BN{gPS}-M1	0.1	5.5	11.1	23.5	32.3	18.1
BN{gVTS}-M1	0.1	**5.3**	12.4	**14.3**	**30.6**	**15.6**
BN{gPS}-M2	0.1	5.7	10.8	13.0	24.7	13.6
BN{gVTS}-M2	0.1	**5.6**	11.9	**12.3**	26.5	14.1

Table 2 shows the results of a combined gVTS/DNN (*BN{gVTS}*) system in the clean and multi-style conditions. When only clean data is available for training, DNNs on their own cannot deal with the variability induced by noise. However, when combined with gVTS, mismatch condition performance approaches that of a conventional GMM-HMM system in mismatch condition while benefiting from using DNN in the matched condition. In multi-style training, when only additive noise is available (M1), although DNN (on its own) leads to a significant performance improvement in dealing with additive noise, it fails in coping with channel mismatch. In this case, the gVTS can play a complementary role. Finally, if the DNN is trained on both additive and channel noise (M2), although its combination with gVTS could still be useful in the test sets A and C, on average, the DNN-only system outperforms the gVTS/DNN system.

6. CONCLUSION

This paper extended the method of additive and channel noise compensation with generalised Vector Taylor Series (gVTS) to the group delay-power product spectrum domain. The problems which this presents were discussed, some solutions were proposed and the corresponding gVTS formulae were derived. Experimental results on Aurora-4 showed that a system trained only on the clean data using the proposed feature, on average, outperforms an MFCC-based system trained using multi-style data. Combination of the gVTS features with the bottleneck feature in clean training mode resulted in remarkable WER reductions in the clean-match condition with minor performance loss in the unmatched condition. This potentially allows robust systems to be built using DNNs even when only clean training data is available.

7. REFERENCES

[1] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1 – 29, 2016.

[2] E. Loveimi and S.M. Ahadi, "Objective evaluation of magnitude and phase only spectrum-based reconstruction of the speech signal," in *ISCCSP*, March 2010.

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

[4] 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.

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

[6] E. Loweimi, S.M. 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)*, 2013, vol. 7911 LNAI, pp. 160–167.

[7] 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.

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

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

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

[11] K. Vijayan, R.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, July 2016.

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

[13] 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.

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

[15] P.J. Moreno, B. Raj, and R.M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *ICASSP*. IEEE, 1996, vol. 2, pp. 733–736.

[16] E. Loweimi, J. Barker, and T. Hain, "Use of generalised nonlinearity in vector taylor series noise compensation for robust speech recognition," in *INTERSPEECH*, 2016, pp. 3798–3802.

[17] 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.

[18] 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.

[19] D. Zhu and K.K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *ICASSP*, May 2004, vol. 1, pp. I–125–8 vol.1.

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

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

[22] G.E.P. Box and D.R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 26, no. 2, pp. 211–252, 1964.

[23] 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.

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

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

[26] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *ICASSP*, April 2007, vol. 4, pp. IV–757–IV–760.

[27] K. Vesely, L. Burget, and F. Grezl, "Parallel training of neural networks for speech recognition," in *INTERSPEECH*, 2010, pp. 2934–2937.