Channel Compensation in the Generalised Vector Taylor Series Approach to Robust ASR

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

Vector Taylor Series (VTS) is a powerful technique for robust ASR but, in its standard form, it can only be applied to log-filter bank and MFCC features. In earlier work, we presented a generalised VTS (gVTS) that extends the applicability of VTS to front-ends which employ a power transformation non-linearity. gVTS was shown to provide performance improvements in both clean and additive noise conditions. This paper makes two novel contributions. Firstly, while the previous gVTS formulation assumed that noise was purely additive, we now derive gVTS formulas for the case of speech in the presence of both additive noise and channel distortion. Second, we propose a novel iterative method for estimating the channel distortion which utilises gVTS itself and converges after a few iterations. Since the new gVTS blindly assumes the existence of both additive noise and channel effects, it is important not to introduce extra distortion when either are absent. Experimental results conducted on LVCSR Aurora-4 database show that the new formulation passes this test. In the presence of channel noise only, it provides relative WER reductions of up to 30% and 26%, compared with previous gVTS and multi-style training with cepstral mean normalisation, respectively.

Index Terms: robust speech recognition, generalised Vector Taylor Series, Channel noise estimation

1. Introduction

Automatic speech recognition (ASR) performance in a noise-free condition can reach human parity [1, 2]. In noisy conditions, if sufficient data can be collected from conditions that match the test scenario then state-of-the-art performance is achievable using DNN-based techniques in either front-end [3, 4] or back-end [5–7]. However, in many situations matched training data is not available and purely data-driven approaches perform poorly. Therefore, it is worthwhile, from both a theoretical and practical standpoint, to consider how to build robust systems using only clean training data.

Vector Taylor Series (VTS) [8] is a well-established and powerful technique for robust ASR with formulations that allow it to be applied in either the feature [9] or model [10] domains. It has a well-principled foundation and rests on reasonable assumptions. Taylor series expansion is employed to linearise the nonlinear relationship between the clean and noisy representations. This allows the distribution of the noisy observations to be estimated and for the effects of the noise to be compensated. In its standard form, VTS is only applicable to features which use log for compressing the filter bank energies (FBE). In [11] we replaced the log with the generalised logarithmic function ($GenLog$) function [12] and called it generalised VTS (gVTS). This modification resulted in significant performance improvement in both clean and noisy conditions and extended the applicability of VTS to features which use a power transformation nonlinearity, e.g. PLP [13], generalised-MFCC [14], PNCC [15] and phase-based features [16–21].

The previous formulation of gVTS was based on the assumption that the signal is only contaminated with additive noise. For dealing with the channel noise, geometric mean normalization (GMN) was utilised which is equivalent to cepstral mean normalisation (CMN). It is deterministic in essence and hence, unlike statistical methods, is unable to model the variability induced by noise.

In this paper, we extend the formulation of gVTS assuming that both additive and channel noises are present. The new formulation requires an estimate of the channel noise (a challenging problem that is less studied than additive noise estimation). We present an iterative approach to this problem. Experimental results shows up to 30% and 26% relative WER reduction in dealing with channel noise compared with the previous GMN-based approach and multi-style training results, respectively.

The rest of this paper is organised as follows. In Section 2, the noise compensation process through gVTS is presented. Section 3 explains the proposed channel estimation approach. Section 4 contains experimental results along with discussion and Section 5 concludes the paper.

2. Noise Compensation through gVTS

2.1. Generalised Logarithmic Function

The main idea behind gVTS is to replace the log with $GenLog$

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

where α is its parameter and when α approaches zero, $GenLog$ converges to log. In the Statistics literature, this function is known as the Box-Cox transformation (BCT) [22]. It unifies the log and power transformation (z^{α}), and is claimed to be helpful in enhancing the linearity, Gaussianity and homoscedasticity [22]. Substituting the log operation with $GenLog$ in the MFCC framework yields generalised MFCC (gMFCC) [14]. As shown in [11] α has a substantial impact on the distribution of the filter bank energies (FBE) and can improve the WER in the noisy conditions. These properties foster using this transform.

2.2. Environment Model

Let's consider $Y = XH + W$ as the environment model where Y, X and W denote the power spectra of the noisy observation, clean speech and additive noise, respectively, and H is the squared magnitude spectrum of the channel. Taking the $GenLog$ from both sides yields

$$\breve{Y} = \breve{X}\,\breve{H}\,(1 + (\frac{\breve{W}}{\breve{X}\breve{H}})^{\frac{1}{\alpha}})^{\alpha} \tag{2}$$

where $\breve{Z} = Z^{\alpha}$ for $Z \in \{Y, X, H, W\}$. As seen, the clean representation ($\breve{X}$) is distorted by a distortion function, G,

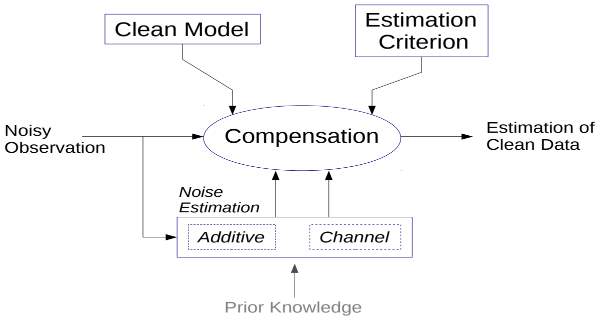

Figure 1: *Elements of the noise compensation process.*

$$G(\breve{X}, \breve{H}, \breve{W}) = \breve{H}\,(1 + (\frac{\breve{W}}{\breve{X}\breve{H}})^{\frac{1}{\alpha}})^{\alpha}, \tag{3}$$

and the higher the SNR, the closer the G to unity. The overarching goal of the noise compensation process is to counter this function and extract an estimate of $\breve{X}$ from $\breve{Y}$.

2.3. Noise Compensation

Fig. 1 depicts the elements of the noise compensation process which contains four parts: the statistical models of the clean data and noises, the estimation criterion and the compensation block. For modelling the distributions of the clean features and noises (both additive and channel), usually a GMM with M components and a single Gaussian are used, respectively,

$$\begin{cases} \breve{X} \sim \sum_{m=1}^{M} p_{\breve{x}}(m)\,\mathcal{N}(\mu_m^{\breve{X}}, \Sigma_m^{\breve{X}}) \\ \breve{W} \sim \mathcal{N}(\mu^{\breve{W}}, \Sigma^{\breve{W}}) \\ \breve{H} \sim \mathcal{N}(\mu^{\breve{H}}, \Sigma^{\breve{H}}) \end{cases} \tag{4}$$

where M, $p_{\breve{x}}(m)$, μ and Σ denote the number of components, weight, mean vector and covariance matrix, respectively. One advantage of using $GenLog$ over log is that α greatly affects the distribution of FBEs [11] such that adjusting α can make the feature distribution more Gaussian-like. This leads to improving the fit of the data to the Gaussian models.

The statistical models could be learned either in the frequency domain ($GenLog$ of FBEs) or in the cepstrum domain (DCT of the $GenLog$ of FBEs). Since the covariance matrices are assumed to be diagonal, modelling and compensation in the cepstral domain is more favourable. However, a GMM with a diagonal covariance matrix can still effectively model the distribution of correlated features at the cost of increasing M which, in turn, entails more training data to avoid the overfitting. As shown in [11] compensation in both domains returns almost equally good results.

Minimum mean square error (MMSE) is usually employed as the estimation criterion

$$\breve{X}_{MMSE} = \mathbb{E}[\breve{X}|\breve{Y}] = \int \breve{X}\,p(\breve{X}|\breve{Y})d\breve{X} \tag{5}$$

where $\mathbb{E}$ denotes the expected value. Rewriting (5) using (2), (3) and some algebraic manipulation yields

$$\begin{aligned} \breve{X}_{MMSE} &= \int \frac{\breve{Y}}{G(\breve{X}, \breve{H}, \breve{W})} \sum_{m=1}^{M} p(\breve{X}|m)\,p(m|\breve{Y})\,d\breve{X} \\ &= \breve{Y} \sum_{m=1}^{M} p(m|\breve{Y})\,\frac{1}{G(\mu_m^{\breve{X}}, \mu^{\breve{H}}, \mu^{\breve{W}})}, \end{aligned} \tag{6}$$

and the only missing part for evaluating (6) is $p(m|\breve{Y})$.

It is usually assumed that $\breve{Y}$ has a GMM distribution with M components, similar to $\breve{X}$. Using Bayes' rule

$$p(m|\breve{Y}) = \frac{p_{\breve{y}}(m)\,\mathcal{N}(\mu_m^{\breve{Y}}, \Sigma_m^{\breve{Y}})}{\sum_{m'=1}^{M} p_{\breve{y}}(m')\,\mathcal{N}(\mu_{m'}^{\breve{Y}}, \Sigma_{m'}^{\breve{Y}})} \tag{7}$$

which, in turn, translates the problem of computing $p(m|\breve{Y})$ into that of finding the distribution of $\breve{Y}$, specifically, $p_{\breve{y}}(m)$, $\mu_m^{\breve{Y}}$ and $\Sigma_m^{\breve{Y}}$. Another assumption is that $\breve{Y}$ and $\breve{X}$ are jointly Gaussian within each mixture component and $p_{\breve{y}}(m) \approx p_{\breve{x}}(m)$. For computing $\mu_m^{\breve{Y}}$ and $\Sigma_m^{\breve{Y}}$, the statistics of $\breve{Y}$ should be computed given those of $\breve{X}$, $\breve{H}$ and $\breve{W}$. However, due to the nonlinearity in (2) this can not be done analytically.

2.4. generalised VTS (gVTS)

Using the first-order Taylor series, the relationship in (2) can be linearised and consequently the statistics of $\breve{Y}_m$ can be calculated. It runs as follows

$$\begin{aligned} \breve{\mathbf{Y}} \;\approx\; & \breve{Y}(\breve{X}_0, \breve{W}_0, \breve{H}_0) \;+\; J^{\breve{X}}(\breve{X} - \breve{X}_0) \\ & +\; J^{\breve{W}}(\breve{W} - \breve{W}_0) \;+\; J^{\breve{H}}(\breve{H} - \breve{H}_0) \end{aligned} \tag{8}$$

where J^Z is the Jacobian matrix of $\breve{Y}$ with respect to Z ($Z \in \{\breve{X}, \breve{H}, \breve{W}\}$) and $(\breve{X}_0, \breve{W}_0, \breve{H}_0)$ denotes the point around which $\breve{Y}$ is linearised. Linearisation is performed around the mean values, namely $(\mu_m^{\breve{X}}, \mu^{\breve{H}}, \mu^{\breve{W}})$ which will be M points altogether. Therefore, the Jacobians should be evaluated at each point. With some algebraic manipulation it can be shown that

$$J_m^{\breve{X}} = \frac{\partial \breve{Y}}{\partial \breve{X}}\Big|_{(\mu_m^{\breve{X}}, \mu^{\breve{H}}, \mu^{\breve{W}})} = diag\{\mu^{\breve{H}}\,(1 + \breve{V}_m)^{\alpha-1}\} \tag{9}$$

$$J_m^{\breve{H}} = \frac{\partial \breve{Y}}{\partial \breve{H}}\Big|_{(\mu_m^{\breve{X}}, \mu^{\breve{H}}, \mu^{\breve{W}})} = diag\{\mu_m^{\breve{X}}\,(1 + \breve{V}_m)^{\alpha-1}\} \tag{10}$$

$$J_m^{\breve{W}} = \frac{\partial \breve{Y}}{\partial \breve{W}}\Big|_{(\mu_m^{\breve{X}}, \mu^{\breve{H}}, \mu^{\breve{W}})} = diag\{(\frac{1 + \breve{V}_m}{\breve{V}_m})^{\alpha-1}\} \tag{11}$$

where $diag[\,\boldsymbol{z}\,]$ turns vector $\boldsymbol{z}$ into a diagonal matrix and

$$\breve{V}_m = (\frac{\mu^{\breve{W}}}{\mu_m^{\breve{X}}\,\mu^{\breve{H}}})^{\frac{1}{\alpha}}. \tag{12}$$

Having evaluated the Jacobians, $\mu_m^{\breve{Y}}$ and $\Sigma_m^{\breve{Y}}$ can be calculated

$$\mu_m^{\breve{Y}} \approx \;\mu_m^{\breve{X}}\,\mu^{\breve{H}}(1 + (\frac{\mu^{\breve{W}}}{\mu_m^{\breve{X}}\mu^{\breve{H}}})^{\frac{1}{\alpha}})^{\alpha} \tag{13}$$

$$\Sigma_m^{\breve{Y}} \approx J_m^{\breve{X}}\Sigma_m^{\breve{X}}{J_m^{\breve{X}}}^T + J_m^{\breve{W}}\,\Sigma^{\breve{W}}\,{J_m^{\breve{W}}}^T + J_m^{\breve{H}}\,\Sigma^{\breve{H}}\,{J_m^{\breve{H}}}^T. \tag{14}$$

For mathematical convenience, $\Sigma_m^{\breve{Y}}$ is assumed to be diagonal. Extension of the modelling to the cepstrum domain can be easily carried out similar to the previous formulation of gVTS [11]. Since the overall performance does not differ, for saving space only the frequency-domain formulation is provided here.

Discarding the nonlinear terms in first-order VTS introduces some error. To reduce this error, in [23], second-order VTS was proposed. It can be shown that the magnitudes of the nonlinear terms are proportional to the eigenvalues of $\Sigma_m^{\breve{Y}}$ to the power of n, where n is the order of the nonlinear term. By increasing the number of Gaussian components, M, these values

– and consequently the contribution of nonlinear terms – become very small. So, first-order VTS using sufficient number of Gaussians is a reasonable approximation. gVTS has another advantage from this perspective: the nonlinear terms are inversely proportional to α. An easy way to verify this point is to set α to one in (2) which yields a linear relationship. So, by increasing this parameter the error associated with VTS linearisation decreases. Increasing α too much, however, does not have a constructive effect on the statistical modelling of the FBEs.

3. Channel Estimation

Channel noise estimation is a challenging problem and has been less studied than the estimation of the additive noise. The most commonly used algorithm is EM-based which was suggested in [8]. Here, we propose an alternative method, depicted in Fig. 2. It is an iterative technique and uses gVTS itself.

3.1. Workflow

If the characteristics of the microphone and its relative position to the speaker are considered to be fixed, channel distortion will not be a stochastic process. As such $\Sigma^{\breve{H}}$ may be set to zero and the channel can be characterised only by the mean ($\mu^{\breve{H}}$). Using a nonzero covariance matrix allows the uncertainty in the mean estimate to be taken into account. However, forming a reliable estimation for it is not straightforward.

We assume initially that the additive noise is absent, so

$$\breve{H}_t = \frac{\breve{Y}_t}{\breve{X}_t} \Rightarrow \mu^{\breve{H}} = \mathbb{E}\{\breve{H}\} = \mathbb{E}\{\frac{\breve{Y}_u}{\breve{X}_u}\} \tag{15}$$

where t and u denote frame index and the utterance, respectively. For mathematical simplification, we suboptimally assume that the random variables $\breve{Y}_u$ and $\frac{1}{\breve{X}_u}$ are uncorrelated

$$\mu^{\breve{H}} = \mathbb{E}\{\frac{\breve{Y}_u}{\breve{X}_u}\} = \mathbb{E}\{\breve{Y}_u\}\ \mathbb{E}\{\frac{1}{\breve{X}_u}\}. \tag{16}$$

$\mathbb{E}\{\breve{Y}_u\}$ can be approximated using sample mean as follows

$$\mathbb{E}\{\breve{Y}_u\} \approx \frac{1}{T}\sum_{t=1}^{T} \breve{Y}_t \tag{17}$$

where T indicates the number of frames of the utterance. Based on the law of large numbers, the larger the T, the better the estimate. Now, $\mathbb{E}\{\frac{1}{\breve{X}_u}\}$ should be calculated. In this manner, $\mathbb{E}\{\breve{X}_u\}$ may be estimated using the GMM of the clean model

$$\mathbb{E}\{\breve{X}_u\} \approx \sum_{m=1}^{M} p_{\breve{x}}(m)\ \mu_m^{\breve{X}}\,. \tag{18}$$

If the utterance is long enough with adequate phonetic diversity, the mean of the clean version would be close to the global mean of the clean speech. Using Jensen's inequality

$$\mathbb{E}\{\frac{1}{\breve{X}_u}\} \geq \frac{1}{\mathbb{E}\{\breve{X}_u\}}. \tag{19}$$

Assuming (again suboptimally) that the equality in (19) holds

$$\mu^{\breve{H}} \approx \frac{\frac{1}{T}\sum_{t=1}^{T} \breve{Y}_t}{\sum_{m=1}^{M} p_{\breve{x}}(m)\ \mu_m^{\breve{X}}}. \tag{20}$$

This runs the risk of underestimation of $\breve{H}$ but provides a practical framework for estimating the channel. It should also be noted that making error in scale and bias, namely $a\mu^{\breve{H}} + b$, is tolerable as they do not change the WER.

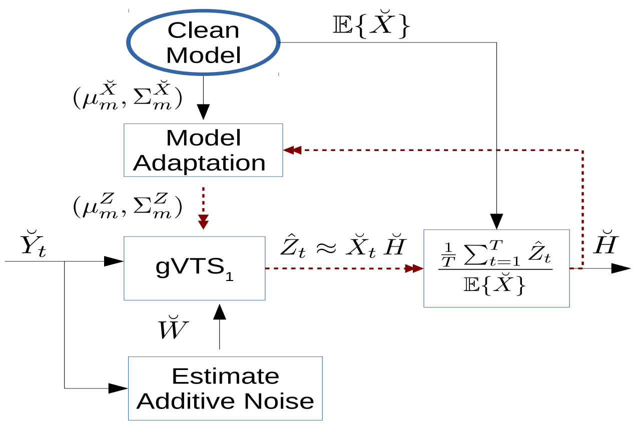

Figure 2: *Workflow of the proposed channel estimation method.*

3.2. Effect of Additive Noise

Now, let us extend (15) to the case where additive noise exists

$$\frac{\mathbb{E}\{\breve{Y}_u\}}{\mathbb{E}\{\breve{X}_u\}} = \mathbb{E}\{\big(H + \frac{W_u}{X_u}\big)^{\alpha}\} \approx \mu^{\breve{H}} + \mathbb{E}\{\frac{\breve{W}_u}{\breve{X}_u}\} \tag{21}$$

where H and $\frac{W}{X}$ are assumed to be uncorrelated. This introduces an error term, $\mathbb{E}\{\frac{\breve{W}_u}{\breve{X}_u}\}$, which is inversely proportional to SNR and leads to overestimation (since it is always positive). To deal with this term, the additive noise should be attenuated. Speech enhancement algorithms may seem useful but bring about the problem of distorting speech in the sense that the enhanced signal will no longer be consistent with the background statistical model of the clean speech ($GMM_{\breve{X}}$).

To this end, we suggest an iterative algorithm illustrated in Figure 2. First, the channel estimate is initialised using (20). Since at this stage the channel noise is not available, the older version of gVTS ($gVTS_1$) is used which only compensates for the additive noise. Let $Z = \breve{X}\breve{H}$, which encapsulate $\breve{X}$ and $\breve{H}$ into one variable. Now $gVTS_1$ aims at alleviating the additive noise and finding $\hat{Z}$. In this regard, GMM_Z should be computed through adapting the $GMM_{\breve{X}}$ using $\breve{H}$

$$Z \sim \sum_{m=1}^{M} p_{\breve{X}}(m)\ \mathcal{N}(z; \breve{H}_d\ \mu_m^{\breve{X}}, \breve{H}_d\ \Sigma_m^{\breve{X}}\ \breve{H}_d^{T}). \tag{22}$$

where $\breve{H}_d$ is $diag[\mu^{\breve{H}}]$. This process attenuates the additive noise and pushes the utterance closer to the background clean model in a statistical sense (i.e., likelihood is increased). It allows for a better channel noise estimation even in the clean condition because (18) will hold more closely. The gVTS1 output, $\hat{Z}$, would be an approximation for $\breve{X}\breve{H}$. As such an estimate for the channel frequency response can be formed using (20) for the the next iteration.

Fig. 3 illustrates the estimated frequency response versus the target (ground truth) values. As seen, the proposed approach shows a great potential for blindly capturing the trend and local shape of the channel. However, in some cases like Fig. 3 (b) and (c), despite capturing the overall trend, local estimates are inexact. In the next subsection we briefly review the causes of error for future optimisations.

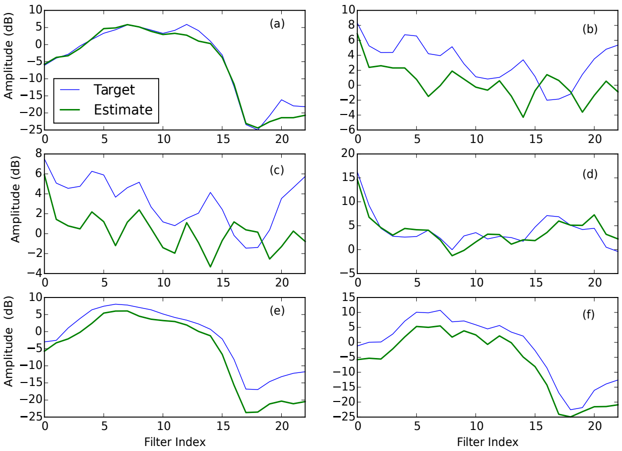

Figure 3: *Blind channel estimation based on the proposed method for 6 waves from the test set C of the Aurora-4 [24]. Target channel response was computed through comparing the noisy wave (Y) with its clean counterpart (X) from test set A. Underestimation is due to Jensen's inequality in (19).*

3.3. Difficulties with the Proposed Approach

In addition to earlier mentioned issues, (18) implicitly assumes that the channel used in recording the training data has a flat frequency response and this is not necessarily the case. Moreover, the frame length (25 ms) may not be longer than the effective length of the impulse response of the channel in the time domain. As such the frequency resolution will be insufficient for resolving the channel frequency response. In such case even target values in Fig. 3 are inaccurate as they are computed using short-term analysis (25 ms). Finally, averaging in (17) and (18) is performed across all frames, both speech and non-speech ones. At the non-speech segments where $X = 0$, channel contribution will be zero, too. As a result, such frames not only do not provide any useful information as far as channel estimation is concerned but also allow the additive noise to further overshadow the channel estimation process. So, ideally, for channel estimation only the speech segments should be considered.

4. Experimental Results

ASR experiments were carried out on Aurora-4 [24]. 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 [25]. Decoding was performed with standard 5k-word WSJ0 bigram language model. The evaluation set of Aurora-4 consists of 14 test sets, grouped into 4 subsets: clean, (additive) noisy, clean with channel distortion, noisy with channel distortion, referred to as A, B, C, and D, respectively. Aurora-4 has two extra training sets for multi-style training, namely $Multi1$ and $Multi2$. Training data in the former is contaminated with only additive noise and in the latter by both additive noise and channel distortion. Cepstral mean normalisation and GMN (for gVTS features) were applied. The feature vector is augmented by c_0, delta and acceleration coefficients. M was set to 256 and additive noise was estimated using the first and last 20 frames.

4.1. Discussion

It should be noted that the gVTS a priori assumes that the signal is corrupted by both additive and channel noises whereas this may not be the case. The ideal compensation process should not distort the signal and worsen the results when either of the noises is not present. Most of the robust methods (if not all) induce extra distortion in the clean condition and return a poorer results than the baseline. Test sets A, B and C are particularly useful to evaluate the parametrisation process from this viewpoint. For instance, for test set A, the gVTS should be as good as the baseline (MFCC) or for test set B, gVTS2 (this paper) should be as accurate as gVTS1 (older version). The results reported in Table 1 show that gVTS2 passes this test successfully. For example, in the case of test set A not only it does not worsen the results in comparison with MFCCs but also returns a lower WER. For test set B the results of gVTS2 are almost as good as gVTS1 unless the channel estimation process is over-iterated.

Performance-wise, the proposed channel estimation algorithm results in a remarkable improvement for Test Set C. The absolute WER is reduced from 21.1% to 14.4% which is equivalent to about 30% relative error reduction. At the same time, for test set B, the performance is almost kept to the same level. The optimum number of iterations (n in gVTS2-α-n at Table 1) for estimating the channel is empirically found to be 2 or 3.

Comparing the results with multi-style training which usually constitutes the upper bound for the performance of the clean-trained systems in the noisy condition is important, too. As seen, on average the performance of the overall system is not far from this limit, especially if the Ave_2, in which all the test sets have the same weight, is taken into account. For test set C, however, the proposed method returns up to 26% relative higher performance compared with multi-style (Multi2) results which is a significant gain given that it has been achieved blindly, at low computational cost and without any stereo data.

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

Feature	A	B	C	D	Ave_1	Ave_2
MFCC-Clean	6.8	33.4	23.8	50.2	38.0	28.6
MFCC-Multi1	9.0	18.0	23.7	35.4	25.2	21.5
MFCC-Multi2	10.0	17.2	19.6	31.2	23.0	19.5
gVTS1-0.05	6.5	19.9	21.1	37.0	26.6	21.3
gVTS2-0.05-0	6.6	20.3	16.7	35.6	25.6	19.8
gVTS2-0.05-1	6.5	20.9	15.9	35.0	25.6	19.6
gVTS2-0.05-2	6.5	20.8	**14.4**	35.1	25.5	**19.2**
gVTS2-0.05-3	6.6	21.3	15.0	35.3	25.8	19.5
gVTS2-0.075-2	7.0	20.8	15.2	35.1	25.6	19.5
gVTS2-0.1-2	7.4	20.3	15.7	34.9	25.3	19.6

$Ave_1 = \frac{A+6B+C+6D}{14} \qquad Ave_2 = \frac{A+B+C+D}{4}$

5. Conclusion

In earlier work, we derived VTS equations assuming that the *log* nonlinearity is substituted by generalised logarithmic function ($GenLog$). We called this approach generalised VTS (gVTS). $GenLog$ has an extra parameter which affects the statistical distribution of the features and can improve the performance in both clean and noisy conditions. In the previous formulation of gVTS, it was assumed that the signal is only distorted by the additive noise. In this paper all the equations were re-derived assuming the presence of both additive and channel noises. In addition, a novel iterative approach for channel estimation was proposed. The experimental results in LVCSR task (Aurora-4) show significant gains in recognition accuracy without noticeable performance loss when either additive or channel noise does not exist. In future work we plan to extending the gVTS to other features which use power transformation and also to further improve the newly proposed channel estimation technique.

6. References

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving Human Parity in Conversational Speech Recognition," *ArXiv e-prints*, Oct. 2016.

[2] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall, "English Conversational Telephone Speech Recognition by Humans and Machines," *ArXiv e-prints*, Mar. 2017.

[3] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for lvcsr of meetings," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, April 2007, pp. IV–757–IV–760.

[4] D. Yu and M. L. Seltzer, "Improved bottleneck features using pre-trained deep neural networks." in *Interspeech*, vol. 237, 2011, p. 240.

[5] H. Bourlard and N. Morgan, "Hybrid hmm/ann systems for speech recognition: Overview and new research directions," in *Adaptive Processing of Sequences and Data Structures, International Summer School on Neural Networks, "E.R. Caianiello"-Tutorial Lectures*. London, UK, UK: Springer-Verlag, 1998, pp. 389–417.

[6] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[7] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Trans. Audio, Speech and Lang. Proc.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[8] P. J. Moreno, "Speech recognition in noisy environments," Ph. D. Dissertation, ECE Department, CMU, Tech. Rep., 1996.

[9] P. J. Moreno, B. Raj, and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2. IEEE, 1996, pp. 733–736.

[10] A. Acero, L. Deng, T. T. Kristjansson, and J. Zhang, "Hmm adaptation using vector taylor series for noisy speech recognition." in *INTERSPEECH*, 2000, pp. 869–872.

[11] E. Loweimi, J. P. Barker, and T. Hain, "Use of generalised nonlinearity in vector taylor series noise compensation for robust speech recognition," in *Interspeech*, 2016.

[12] 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.

[13] 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.

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

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

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

[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, Jan 2007.

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

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

[20] 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.

[21] ——, "Source-filter separation of speech signal in the phase domain." in *INTERSPEECH*. ISCA, 2015.

[22] 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.

[23] S. Bu, Y. Qian, K. C. Sim, Y. You, and K. Yu, "Second order vector taylor series based robust speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 1769–1773.

[24] 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, p. 94, 2002.

[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.