Use of Generalised Nonlinearity in Vector Taylor Series Noise Compensation 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

Designing good normalisation to counter the effect of environmental distortions is one of the major challenges for automatic speech recognition (ASR). The Vector Taylor series (VTS) method is a powerful and mathematically well principled technique that can be applied to both the feature and model domains to compensate for both additive and convolutional noises. One of the limitations of this approach, however, is that it is tied to MFCC (and log-filterbank) features and does not extend to other representations such as PLP, PNCC and phase-based front-ends that use power transformation rather than log compression. This paper aims at broadening the scope of the VTS method by deriving a new formulation that assumes a power transformation is used as the non-linearity during feature extraction. It is shown that the conventional VTS, in the log domain, is a special case of the new extended framework. In addition, the new formulation introduces one more degree of freedom which makes it possible to tune the algorithm to better fit the data to the statistical requirements of the ASR back-end. Compared with MFCC and conventional VTS, the proposed approach provides upto 12.2% and 2.0% absolute performance improvements on average, in Aurora-4 tasks, respectively.

Index Terms: robust speech recognition, feature extraction, noise compensation, Vector Taylor Series, power transformation

1. INTRODUCTION

Automatic speech recognition (ASR) in clean/matched acoustical environments has become a less challenging problem. However, once the signal gets corrupted by noise ASR performance starts to degrade with a rate that depends on the type and level of the disturbance. Both the recognition front-end and back-end can be modified in order to mitigate the sensitivity of the system to the environmental distortions. Three main strategies may be adopted in dealing with this issue [1, 2]. First, the signal may be passed through a pre-processing enhancement stage prior to being supplied to the parametrisation block. Second, the feature extraction techniques can be modified to produce features that are more robust against the distortions induced by noise [3–6]. A third solution is modification of the back-end to increase its ability to cope with noise, either through model compensation/adaptation or by using the noise data during the training phase (multi-style training). One of the most successful approaches to robust speech recognition is the vector Taylor series (VTS) method [7]. In this approach, Taylor series expansion is employed to linearise the non-linear relationship between the clean and noisy representations. This method has been studied from several different perspectives [8–11] and can produce state-of-the-art performance improvements.

However, the VTS method is not flexible in terms of the features it can enhance. Specifically, it can only be used for MFCC or log filterbank energies (FBE) and can not be directly applied to representations which – instead of log compression – apply a power transformation to the FBEs. Examples of such features are generalised-MFCC [12], PLP [3], PNCC [5] and phase-based features [4,13–18]. As will be illustrated in Section 3, substituting the log function with the power transformation has a significant effect on the statistical properties of these features and on their performance, particularly in noisy conditions. As such by combining VTS and power transformations a more flexible framework can be created.

In this paper, we develop a novel formulation for VTS assuming that a power transformation is employed instead of log compression. This expands the applicability of the VTS, increases the controllability of the compensation process and, performance-wise, it will be shown that it leads to a better speech recognition results.

The rest of this paper is organized as follows. In Section 2 the conventional VTS is reviewed. Section 3 contains the motivations and formulation of the new approach called generalised VTS (gVTS). Section 4 includes experimental results as well as discussion and section 5 concludes the paper.

2. REVIEW OF VTS

2.1. Signal contamination with noise

A typical acoustic environment can be modelled as follows

$$Y[k] = X[k]\,|H(k)|^2 + W[k], \tag{1}$$

where index k, $X[k]$, $|H[k]|$, $W[k]$ and $Y[k]$ denote the discrete frequency, power spectrum (PS) of the clean signal, frequency response of the linear channel, PS of noise and PS of the noisy observation, respectively. For simplification, it is assumed that there is no channel noise. Applying the logarithm to both sides of (1) yields

$$\begin{aligned} log\{Y[k]\} &= log\{X[k]\} + log\{1 + \frac{W[k]}{X[k]}\} \\ \tilde{Y}[k] &= \tilde{X}[k] + \underbrace{log\{1 + e^{\tilde{W}[k]-\tilde{X}[k]}\}}_{G(\tilde{W},\tilde{X})} \end{aligned} \tag{2}$$

where $log\{Z[k]\} = \tilde{Z}[k]$. Taking the discrete cosine transform (DCT) results in

$$\tilde{y}[q] = \tilde{x}[q] + \underbrace{C\ log\{1 + e^{C^{-1}\left(\tilde{w}[q]-\tilde{x}[q]\right)}\}}_{g(\tilde{w},\tilde{x})}. \tag{3}$$

where C and C^{-1} denote the DCT and inverse DCT matrices, respectively, and q, $\tilde{x}$, $\tilde{w}$, and $\tilde{y}$ are quefrency, (real) cepstrums of the clean data, noise and noisy observation, respectively. As seen in (2)

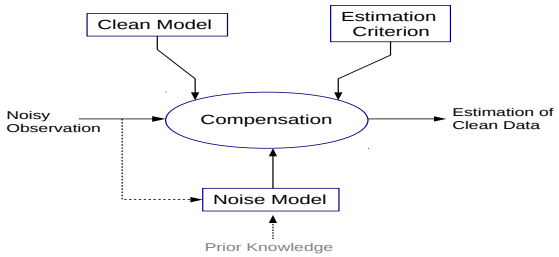

Fig. 1. Elements of the compensation workflow.

and (3), in both log and cepstral domains the noisy observation consists of two components, namely the clean part and an additive term. The latter is a non-linear function of the signal-to-noise ratio (SNR) and by SNR reduction, its influence increases. It may be thought of as a distortion function, $G(\tilde{W}, \tilde{X})$ or $g(\tilde{w}, \tilde{x})$, depending on the domain. The non-linearity of this function complicates estimation of the statistics of noisy observations given a model for clean speech.

2.2. Compensation

Fig. 1 depicts the elements of the compensation workflow. As seen, it is comprised of four parts, namely, the clean model, noise model, estimation criteria and the compensation process. For modelling the distributions of the clean (x) and noise (w) data, usually a GMM with M components and a single Gaussian are used, respectively,

$$\begin{cases} x \sim \sum_{m=1}^{M} \; p_x(m)\, \mathcal{N}(x; \mu_m^x, \Sigma_m^x) \\ w \sim \mathcal{N}(w; \mu^w, \Sigma^w), \end{cases} \tag{4}$$

where $p_x(m)$, μ_m^x and Σ_m^x denote the component weight, mean vector and (diagonal) covariance matrix of the m^{th} Gaussian of the clean feature model and μ^w and Σ^w are the mean vector and covariance matrix of the noise, respectively. The models of noise and clean speech data could be learned either in the frequency domain (from $log-FBEs$) or in the cepstrum domain (from the DCT of the log-FBEs). Since the covariance matrices are assumed to be diagonal, modelling in the cepstral domain is more suitable due to decorrelation effect of the DCT.

For estimating the clean features from the noisy observations, minimum mean square error (MMSE) is used as estimation criterion

$$\hat{x}_{MMSE} = \mathcal{E}[x|y] = \int x\, p(x|y) dx \tag{5}$$

where $\mathcal{E}$ denotes the expected value. In the log (frequency) domain

$$\begin{aligned} \hat{\tilde{X}}_{MMSE} &= \int \left(\tilde{Y} - G(\tilde{W}, \tilde{X})\right) p(\tilde{X}|\tilde{Y}) d\tilde{X} \\ &= \tilde{Y} - \sum_{m=1}^{M} p(m|\tilde{Y})\; G(\mu^{\tilde{W}}, \mu_m^{\tilde{X}}), \end{aligned} \tag{6}$$

and in the cepstrum (quefrency) domain

$$\begin{aligned} \hat{\tilde{x}}_{MMSE} &= \int \left(\tilde{y} - g(\tilde{w}, \tilde{x})\right) p(\tilde{x}|\tilde{y}) d\tilde{x} \\ &= \tilde{y} - \sum_{m=1}^{M} p(m|\tilde{y})\; g(\mu^{\tilde{w}}, \mu_m^{\tilde{x}}). \end{aligned} \tag{7}$$

In order to compute the $\hat{x}_{MMSE}$ based on (5) [1], the posterior probabilities ($p(m|y)$) should be estimated. In this case, it is assumed that the noisy features also follow a GMM model with the same number of Gaussians, namely M. Using Bayes' rule

$$p(m|y) = \frac{p_y(m)\; p(y|m)}{p(y)} = \frac{p_y(m)\; \mathcal{N}(y; \mu_m^y, \Sigma_m^y)}{\sum_{m'=1}^{M} p_y(m')\; \mathcal{N}(y; \mu_{m'}^y, \Sigma_{m'}^y)} \tag{8}$$

The problem of finding $p(m|y)$ is translated into that of finding $p_y(m)$, μ_m^y and Σ_m^y. Usually it is assumed that x and y are jointly Gaussian within each mixture component and that

$$p_y(m) \approx p_x(m). \tag{9}$$

For computing μ_m^y and Σ_m^y, a relationship (preferably linear) between x and y is required that allows the statistics of y to be computed given noise and clean models. At this point, first-order VTS is used for linearising the relation between vectors $\mathbf{x}$ and $\mathbf{y}$ around the point $(\mathbf{w_0}, \mathbf{x_0})$

$$\mathbf{y} \approx \mathbf{y}(\mathbf{w_0}, \mathbf{x_0}) + \mathbf{A}\,(\mathbf{x} - \mathbf{x_0}) + \mathbf{B}\,(\mathbf{w} - \mathbf{w_0}) \tag{10}$$

where A and B matrices are defined as follows

$$A_{i,j} = \left.\frac{\partial y_i}{\partial x_j}\right|_{(\mathbf{w_0}, \mathbf{x_0})}, \quad B_{i,j} = \left.\frac{\partial y_i}{\partial w_j}\right|_{(\mathbf{w_0}, \mathbf{x_0})}. \tag{11}$$

Depending on the domain chosen for modelling/compensation

$$A = \begin{cases} \frac{\partial \tilde{Y}_i}{\partial \tilde{X}_j} & = diag\left[\frac{1}{1+e^{(\mathbf{w}-\mathbf{x})}}\right] \\ \frac{\partial \tilde{y}_i}{\partial \tilde{x}_j} & = C\, diag\left[1 + exp(C^{-1}(\tilde{\mathbf{w}} - \tilde{\mathbf{x}}))\right] C^{-1} \end{cases} \tag{12}$$

where $diag[\,.\,]$ is the operation of generating a diagonal matrix from a vector. B in either domains equals

$$B = I - A. \tag{13}$$

As such, depending on the domain chosen for modelling (and compensation), μ_m^y and Σ_m^y can be calculated as follows

$$\text{log-FBE} \Rightarrow \begin{cases} \mu_m^{\tilde{Y}} \approx \mu_{\mathbf{m}}^{\tilde{\mathbf{X}}} + G(\mu^{\tilde{W}}, \mu_m^{\tilde{X}}) \\ \Sigma_m^{\tilde{Y}} \approx (A \Sigma_m^{\tilde{X}} A^T + B \Sigma^{\tilde{W}} B^T) \;\odot\; I \end{cases} \tag{14}$$

$$\text{Cepstrum} \Rightarrow \begin{cases} \mu_m^{\tilde{y}} \approx \mu_{\mathbf{m}}^{\tilde{\mathbf{x}}} + g(\mu^{\tilde{w}}, \mu_m^{\tilde{x}}) \\ \Sigma_m^{\tilde{y}} \approx (A \Sigma_m^{\tilde{x}} A^T + B \Sigma^{\tilde{w}} B^T) \;\odot\; I, \end{cases} \tag{15}$$

where I is the Identity matrix and $\odot$ denotes element-wise (Hadamard) multiplication aiming at diagonalising the covariance matrix for computational convenience.

3. GENERALISED VTS

3.1. Generalised Nonlinearity

Replacing the log function with the generalised logarithmic function was shown to have a significant effect on the WER [12]. To the best of our knowledge, the first use of this nonlinearity in speech processing dates back to 1984 [19] in which it was suggested as an extension to spectral root deconvolution system (SRDS) proposed in [20]. In parallel, in the statistics literature it was known from 1964 as

[1] When the domain is not explicitly mentioned, (e.g. x, instead of $\tilde{X}[k]$ or $\tilde{x}[q]$) it means that the argument holds for both domains.

the Box-Cox Transformation (BCT) [21]. BCT itself was proposed as a complementary approach to the Tukey Ladder of Powers [22] for enhancing the linearity, normality and homoscedasticity (variance stabilisation) of the data. In our work, we refer to the generalised logarithmic function or BCT as the generalised non-linearity (GN) and it is defined as follows

$$\begin{cases} GN(x;\gamma) = \frac{1}{\gamma}(x^{\gamma} - 1), & x > 0 \quad \gamma \neq 0 \\ \lim_{\gamma \to 0} GN(x;\gamma) = log(x) \end{cases} \tag{16}$$

where γ is the parameter of this transformation.

Table 1 shows the effect of substituting the *log* in MFCC with GN for different values of γ. It is called γ-MFCC and as seen, using GN instead of *log* has a noticeable effect on the performance. If the γ parameter is tuned correctly (0.075), it provides notably better results without increasing the computational complexity of the parametrisation process. This improvement can be attributed to the impact of the transformation on the distribution of the FBEs.

To demonstrate this point, FBEs of 2000 signals from Aurora-2 clean data were pooled and the histogram of each sub-band was plotted. Fig. 2 shows the average of the histograms. As can be observed, γ has a clear effect on the distribution of the features fed into the DCT block in the MFCC parametrisation process. This, in turn, affects how well the back-end model (HMM/GMM) fits the data. It should be noted that -1 and γ may be removed from the numerator and denominator of Eq. (16) without loss of generality since they do not affect the discriminability of the features, as they are identical for all classes. As such the GN is equivalent to a power transformation.

Table 1. *Average (0-20 dB) accuracy (%) for Aurora-2. Feature vector size is 39: static, log-energy, augmented by Δ and Δ-Δ.*

Feature	γ	TestSet A	TestSet B	TestSet C
MFCC	$log \leftrightarrow 0$	66.2	71.4	64.9
γ-MFCC	0.01	68.0	72.2	69.7
γ-MFCC	0.05	74.5	76.7	76.0
γ-MFCC	0.075	**75.4**	**76.2**	**76.9**
γ-MFCC	0.1	73.3	74.3	74.5
γ-MFCC	0.15	70.0	71.4	68.8
γ-MFCC	0.2	67.2	69.3	63.2

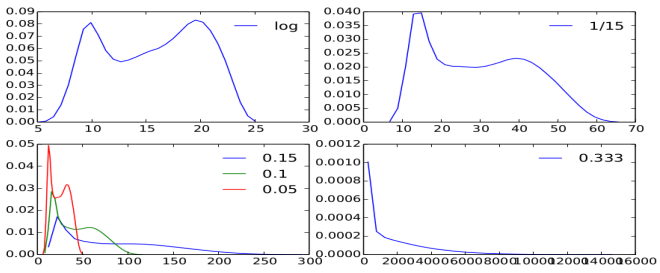

Fig. 2. Effect of γ on the histogram (distribution) of the FBEs.

3.2. Formulation of gVTS

For simplicity, the linear channel effect is removed (we return to this point later). Applying GN to both sides of (1) yields

$$Y^{\gamma}[k] = X^{\gamma}[k]\big(1 + (\frac{W^{\gamma}[k]}{X^{\gamma}[k]})^{\frac{1}{\gamma}}\big)^{\gamma} \tag{17}$$

Let

$$\begin{cases} \breve{X}[k] = X^{\gamma}[k] \\ \breve{W}[k] = W^{\gamma}[k] \\ \breve{Y}[k] = Y^{\gamma}[k] \end{cases} \quad \begin{cases} \breve{V}[k] = (\frac{\breve{W}[k]}{\breve{X}[k]})^{\frac{1}{\gamma}} \\ G(\breve{W}, \breve{X}) = (1 + \breve{V})^{\gamma} \\ Y[k] = \breve{X}[k] G(\breve{W}[k], \breve{X}[k]) \end{cases} \quad \begin{cases} \breve{\mathbf{x}} = \mathbf{C}\,\breve{\mathbf{X}} \\ \breve{\mathbf{w}} = \mathbf{C}\,\breve{\mathbf{W}} \\ \breve{\mathbf{y}} = \mathbf{C}\,\breve{\mathbf{Y}} \end{cases} \tag{18}$$

where bold letters (both lower-case and upper-case) denote vectors.

The first difference from VTS is that the distortion function $g(.,.)$ cannot be expressed explicitly in the quefrency domain as was done in (3)

$$\breve{y}[i] = \sum_{k=1}^{d} C_{ik}\,\breve{Y}[k] = \sum_{k=1}^{d} C_{ik}\breve{X}[k]\,G(\breve{W}[k], \breve{X}[k]), \tag{19}$$

where d indicates the dimension of vectors in the frequency domain [2]. It is due to the loss of the additive property that the *log* function provides and the fact that the DCT of the multiplication of two sequences is not equal to the convolution of the DCTs.

This could be costly performance-wise because as explained in Section 2.2, modelling in the cepstrum domain better satisfies the assumptions made by the model. On the plus side, one of the effects of the GN is enhancing the Normality of the features. Therefore, the compatibility between the transformed data and such models improves. This allows for reaching a high level of fit with a fewer Gaussians in the mixture.

The next step is estimation of the clean feature using MMSE. In the frequency domain

$$\begin{aligned} \breve{X}_{MMSE} &= \int \breve{X}\, p(\breve{X}|\breve{Y})\, d\breve{X} = \int \frac{\breve{Y}}{G(\breve{W}, \breve{X})}\, p(\breve{X}|\breve{Y}) dx \\ &= \breve{Y} \sum_{m=1}^{M} p(m|\breve{Y}) \frac{1}{G(\mu^{\breve{W}}, \mu_m^{\breve{X}})}. \end{aligned} \tag{20}$$

As with VTS, it is assumed that the distribution of $\breve{y}$ is a GMM with M Gaussians and diagonal covariance matrix and components weight are computed by (9). Following the same line, we arrive at

$$\begin{cases} \mu_m^{\breve{Y}} \approx \mu_m^{\breve{X}}\, G(\mu^{\breve{W}}, \mu_m^{\breve{X}}) \\ \Sigma_m^{\breve{Y}} \approx (A \Sigma_m^{\breve{X}} A^T + B \Sigma_m^{\breve{W}} B^T) \odot I. \end{cases} \tag{21}$$

By some algebraic manipulation, it can be shown that

$$A_{i,j} = \frac{\partial \breve{Y}[i]}{\partial \breve{X}[j]} = \begin{cases} (1 + \breve{V}[i])^{\gamma - 1} & , i = j \\ 0 & , i \neq j \end{cases} \tag{22}$$

which in matrix form would be

$$A = diag\big[(1 + \breve{\mathbf{V}})^{\gamma - 1}\big]. \tag{23}$$

Similarly B equals

$$B = diag\big[(\frac{1 + \breve{\mathbf{V}}}{\breve{\mathbf{V}}})^{\gamma - 1}\big]. \tag{24}$$

As mentioned, modelling in the cepstral domain results in a better fit and potentially higher recognition rates. So, it is advantageous to develop the compensation formula also in the quefrency domain

[2] In fact, number of filters of the filterbank, e.g. 23.

[3]. As mentioned earlier, the distortion function cannot be directly expressed in the cesptrum domain but (20) can be rewritten:

$$\breve{x}_{MMSE} = (C^{-1}\breve{\mathbf{y}}) \sum_{m=1}^{M} p(m|C^{-1}\breve{\mathbf{y}}) \frac{1}{G(C^{-1}\mu^{\breve{w}}, C^{-1}\mu_m^{\breve{x}})}. \tag{25}$$

The only missing element in Equation (25) is $p(m|C^{-1}\breve{\mathbf{y}})$. In this regard, we first compute $p(m|\breve{\mathbf{y}})$ using modelling in the cepstral domain, and then find its relationship to $p(m|\breve{\mathbf{Y}})$. To this end, A and B can be computed as follows

$$\begin{aligned} A_{i,j} &= \frac{\partial \breve{y}[i]}{\partial \breve{x}[j]} = \frac{\partial}{\partial \breve{x}[j]} \left(\sum_{k=0}^{d-1} C_{ik} \underbrace{\sum_{p=0}^{d-1} C_{kp}^{-1} \breve{x}_p}_{\breve{X}[k]} (1+\breve{V}_k)^{\gamma} \right) \\ &= \sum_{k=0}^{d-1} C_{ik}(1+\breve{V}_k)^{\gamma} C_{kj}^{-1} C_{ik} - \gamma(1+\breve{V}_k)^{\gamma-1} \frac{\breve{V}_k}{\gamma\, \breve{x}_k} C_{kj}^{-1}\, \breve{x}_k \\ &= \sum_{k=0}^{d-1} C_{ik}\, (1+\breve{V}[k])^{\gamma-1}\, C_{kj}^{-1}. \end{aligned} \tag{26}$$

Rewriting (26) in matrix form yields

$$A = C\; diag\big[(1+\breve{\mathbf{V}})^{\gamma-1}\big]\; C^{-1} \tag{27}$$

and by some algebraic manipulation

$$B = C\; diag\big[(\frac{1+\breve{\mathbf{V}}}{\breve{\mathbf{V}}})^{\gamma-1}\big]\; C^{-1}. \tag{28}$$

Therefore,

$$\begin{cases} p_{\breve{y}}(m) \approx p_{\breve{x}}(m) \\ \mu_m^{\breve{y}} \approx \sum_{k=0}^{d-1} C_{ik}\; \left(C_k^{-1} \cdot \mu_m^{\breve{x}}\right) G(C_k^{-1} \cdot \mu^{\breve{w}}, C_k^{-1} \cdot \mu_m^{\breve{x}}) \\ \Sigma_m^{\breve{y}} \approx (A\Sigma_m^{\breve{x}} A^T + B\Sigma_m^{\breve{w}} B^T) \odot I, \end{cases} \tag{29}$$

where C_k^{-1} indicates the k^{th} row of the C^{-1} matrix and $'\cdot'$ denotes inner product. Returning to (25), the issue was computing $p(m|\breve{\mathbf{Y}})$ and so far only $p(m|\breve{\mathbf{y}})$ is available. Since the $\breve{\mathbf{Y}}$ and $\breve{\mathbf{y}}$ are linear transforms of each other, and y is modelled by a GMM, it can be shown the components weights, likelihoods and consequently the posterior probabilities do not change

$$\breve{\mathbf{Y}} = C^{-1}\, \breve{\mathbf{y}} \Rightarrow \begin{cases} p_{\breve{\mathbf{Y}}}(m) = p_{\breve{\mathbf{y}}}(m) \\ p(\breve{\mathbf{Y}}|m) = p(\breve{\mathbf{y}}|m) \end{cases} \Rightarrow\; p(m|\breve{\mathbf{Y}}) = p(m|\breve{\mathbf{y}}). \tag{30}$$

3.3. Linear Channel

If a linear channel had been considered in Equation (17), then the foregoing algorithm would approximately return $\breve{X}[n,k]\,|H[k]|^{2\gamma}$, where n indicates the frame index. Assuming the stationarity of the channel, one can normalise the channel through geometric mean normalisation (GMN) as follows

$$\breve{x}[n,k] = \frac{\breve{X}[n,k]\;|H[k]|^{2\gamma}}{\sqrt[N]{\prod_{n=1}^{N} \breve{X}[n,k]\;|H[k]|^{2\gamma}}}. \tag{31}$$

[3]We refer to the $DCT(GN(X[k]))$ as cepstrum, too.

4. EXPERIMENTAL RESULTS

4.1. Parametrisation

The feature vector is 39-dimensional and includes static, delta and delta-delta temporal derivatives and cepstral mean normalization is applied. The experiments are carried out on Aurora-4 [24] which is a medium vocabulary task based on Wall Street Journal (WSJ0) corpus. HMMs were trained from clean data ($\approx$ 14 hours) with 16 components per mixture using maximum likelihood estimation. All acoustic models were standard phonetically state-clustered triphones ($\approx$ 2100 states) which were trained from scratch using a standard HTK regime [25]. Decoding was performed with standard 5k-word WSJ0 bigram language model. Clean model for VTS were trained using standard EM with 6 iterations. Noise model was estimated using the first and last 20 frames. The evaluation set of Aurora-4 consists of 14 test sets which can be grouped into 4 subsets: clean, noisy, clean with channel distortion, noisy with channel distortion, which will be referred to as A, B, C, and D, respectively.

Table 2. *Word error rates (WER) for Aurora-4. Ave=* $\frac{A+6B+C+6D}{14}$

Feature	γ	A	B	C	D	$Ave.^*$
MFCC	$log \leftrightarrow 0$	6.8	33.4	23.8	50.2	38.0
γ-MFCC	0.05	7.3	25.4	23.9	42.9	31.5
γ-MFCC	0.075	7.6	23.7	24.8	41.6	30.3
γ-MFCC	0.10	8.3	22.3	25.3	40.1	29.1
VTS-log	log	6.8	21.9	22.1	38.9	28.1
VTS-cep	log	6.7	21.6	21.7	37.5	27.4
gVTS-log	0.075	6.9	19.4	23.7	37.7	26.6
gVTS-cep	0.075	7.1	19.6	24.9	37.1	26.6
gVTS-log-GMN	0.075	7.0	19.9	21.6	36.5	26.2
gVTS-cep-GMN	0.075	6.8	19.3	21.4	36.0	25.7
gVTS-log-GMN	0.05	6.5	19.8	20.4	36.0	25.8
gVTS-cep-GMN	0.05	6.5	19.9	20.6	36.1	25.9
gVTS-log-GMN	0.1	7.3	19.2	21.3	36.6	26.0
gVTS-cep-GMN	0.1	7.4	18.9	21.3	35.8	25.5

4.2. Discussion

Table 2 shows that the proposed generalised VTS yields higher performance than VTS on average. It should be noted that VTS is itself a powerful method and its recognition performance is difficult to improve without using a more complex modelling in the back-end. Applying GMN is helpful in removing the effect of channel noise as explain in Section 3.3. As seen, it is specially influential in case of C and D test sets in which the signals are contaminated with channel noise. Another point that may be deducted from Table 2 is that the optimum value for γ depends on the SNR. For clean data, the lower the γ, the better the WER. On the other hand, by SNR reduction larger values for this parameter returns better results. Based on Table 1 and Table 2, 0.05-0.1 is an optimum range for this parameter.

5. CONCLUSIONS

A novel formulation for VTS has been presented that introduces the use of power transformation (or so-called generalised logarithmic function) instead of the logarithm in the feature extraction process. The advantages of this modification was discussed and demonstrated from both statistical and ASR performance points of view. Moreover, the proposed formulation, expands the potential of VTS approach by enabling it to be applied to a wider range of features, including PLP, PNCC and phase-based representations. This opens up an ample opportunities for future research.

6. REFERENCES

[1] Jinyu Li, Li Deng, R. Haeb-Umbach, and Y. gong, *Robust Automatic Speech Recognition — A Bridge to Practical Applications (1st Edition), 306 pages*, Elsevier, October 2015.

[2] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745 – 777, April 2014.

[3] 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.

[4] 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.

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

[6] S. Ganapathy, "Robust speech processing using arma spectrogram models," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 5029–5033.

[7] 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*, May 1996, vol. 2, pp. 733–736 vol. 2.

[8] Alex Acero, Li Deng, T. Kristjansson, and J. Zhang, "Hmm adaptation using vector taylor series for noisy speech recognition," in *Proc. Int. Conf. on Spoken Language Processing*, October 2000.

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

[10] R. C. van Dalen and M. J. F. Gales, "Extended vts for noise-robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 733–743, May 2011.

[11] Jinyu Li, Michael L. Seltzer, and Yifan Gong, "Improvements to vts feature enhancement," in *Proc. ICASSP*, 2012.

[12] Keiichi Tokuda, Takao Kobayashi, Takashi Masuko, and Satoshi Imai, "Mel-generalized cepstral analysis a unified approach to speech spectral estimation," in *Proc. ICSLP-94*, 1994, pp. 1043–1046.

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

[14] H.A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International conference on*, April 2003, vol. 1, pp. I–68–71 vol.1.

[15] 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.

[16] Erfan Loweimi and Seyed Mohammad 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.

[17] 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.

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

[19] 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.

[20] Jae Lim, "Spectral root homomorphic deconvolution system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 3, pp. 223–233, Jun 1979.

[21] 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.

[22] John W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, 1977.

[23] 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.

[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, pp. 94, 2002.

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