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 normalisation (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 show 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

\[
\text{GenLog}(z; \alpha) = \begin{cases} 
\frac{1}{\alpha}(z^\alpha - 1), & z > 0 \\
\lim_{\alpha \to 0} \text{GenLog}(z; \alpha) = \log(z),
\end{cases}
\]

where \(\alpha\) is its parameter and when \(\alpha\) 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 (\(z^\alpha\)), 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] \(\alpha\) 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

\[
\hat{Y} = \hat{X}H + \left(1 + \frac{W}{HH}ight)^\alpha 
\]

where \(\hat{Z} = Z^\alpha\) for \(Z \in \{Y, X, H, W\}\). As seen, the clean representation (\(\hat{X}\)) is distorted by a distortion function, \(G\),
and the only missing part for evaluating (6) is $p(m|\tilde{Y})$.

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

$$p(m|\tilde{Y}) = \frac{p_{\tilde{Y}}(m)N(\mu_{\tilde{Y}}^m, \Sigma_{\tilde{Y}}^m)}{\sum_{m'=1}^{M} p_{\tilde{Y}}(m')N(\mu_{\tilde{Y}}^{m'}, \Sigma_{\tilde{Y}}^{m'})} \quad (7)$$

which, in turn, translates the problem of computing $p(m|\tilde{Y})$ into that of finding the distribution of $\tilde{Y}$, specifically, $p_{\tilde{Y}}(m)$, $\mu_{\tilde{Y}}^{m}$ and $\Sigma_{\tilde{Y}}^{m}$. Another assumption is that $\tilde{Y}$ and $\tilde{X}$ are jointly Gaussian within each mixture component and $p_{\tilde{Y}}(m) \approx p_2(m)$.

For computing $\mu_{\tilde{Y}}^{m}$ and $\Sigma_{\tilde{Y}}^{m}$, the statistics of $\tilde{Y}$ should be computed given those of $\tilde{X}$, $\tilde{H}$ and $\tilde{W}$. 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 $\tilde{Y}_m$ can be calculated. It runs as follows

$$\tilde{Y} \approx \tilde{Y}(X_0, W_0, H_0) + J \tilde{X}(X - X_0) + J W (\tilde{W} - W_0) + J H (\tilde{H} - H_0) \quad (8)$$

where $J^X$ is the Jacobian matrix of $\tilde{Y}$ with respect to $Z$ ($Z \in \{X, \tilde{H}, \tilde{W}\}$) and $(X_0, W_0, H_0)$ denotes the point around which $\tilde{Y}$ is linearised. Linearisation is performed around the mean values, namely $(\mu_{\tilde{X}}^X, \mu_{\tilde{H}}^H, \mu_{\tilde{W}}^W)$ which will be $M$ points altogether. Therefore, the Jacobians should be evaluated at each point. With some algebraic manipulation it can be shown that

$$J_{m}^{X} = \frac{\partial \tilde{Y}}{\partial \tilde{X}} \bigg|_{(\mu_{\tilde{X}}^X, \mu_{\tilde{H}}^H, \mu_{\tilde{W}}^W)} = \text{diag}(\mu_{\tilde{X}}^X (1 + \tilde{V}_m^a)^{-1}) \quad (9)$$

$$J_{m}^{H} = \frac{\partial \tilde{Y}}{\partial H} \bigg|_{(\mu_{\tilde{X}}^X, \mu_{\tilde{H}}^H, \mu_{\tilde{W}}^W)} = \text{diag}(\mu_{\tilde{X}}^X (1 + \tilde{V}_m^a)^{-1}) \quad (10)$$

$$J_{m}^{W} = \frac{\partial \tilde{Y}}{\partial W} \bigg|_{(\mu_{\tilde{X}}^X, \mu_{\tilde{H}}^H, \mu_{\tilde{W}}^W)} = \text{diag}(1 + \tilde{V}_m^a)^{-1} \quad (11)$$

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

$$\tilde{V}_m = (\mu_{\tilde{W}}^W \mu_{\tilde{X}}^X)^{\frac{1}{2}}. \quad (12)$$

Having evaluated the Jacobians, $\mu_{\tilde{Y}}^X$ and $\Sigma_{\tilde{Y}}^X$ can be calculated

$$\mu_{\tilde{Y}}^X \approx \mu_{\tilde{X}}^X \mu_{\tilde{H}}^H (1 + (\mu_{\tilde{W}}^W \mu_{\tilde{X}}^X)^\frac{1}{2})^a \quad (13)$$

$$\Sigma_{\tilde{Y}}^X \approx J_{m}^{X} \Sigma_{\tilde{X}} J_{m}^{X} + J_{m}^{W} \Sigma_{\tilde{W}} J_{m}^{W} + J_{m}^{H} \Sigma_{\tilde{H}} J_{m}^{H} \quad (14)$$

For mathematical convenience, $\Sigma_{\tilde{Y}}^X$ 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 $\Sigma_{\tilde{Y}}^X$, 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 $\alpha$. An easy way to verify this point is to set $\alpha$ to one in (2) which yields a linear relationship. So, by increasing this parameter the error associated with VTS linearisation decreases. Increasing $\alpha$ 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 $\Sigma_H$ may be set to zero and the channel can be characterised only by the mean $(\mu_H)$. 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

$$\hat{H}_t = \frac{\hat{Y}_t}{\hat{X}_t} \Rightarrow \mu_H = \mathbb{E}\{\tilde{H}\} = \mathbb{E}\{\frac{\tilde{Y}_u}{\tilde{X}_u}\}$$

(15)

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

$$\mu_H = \mathbb{E}\{\frac{\tilde{Y}_u}{\tilde{X}_u}\} = \mathbb{E}\{\tilde{Y}_u\} \mathbb{E}\{\frac{1}{\tilde{X}_u}\}.$$  

(16)

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

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

(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, $\mathbb{E}\{\frac{1}{\tilde{X}_u}\}$ should be calculated. In this manner, $\mathbb{E}\{\tilde{X}_u\}$ may be estimated using the GMM of the clean model

$$\mathbb{E}\{\tilde{X}_u\} \approx \frac{1}{M} \sum_{m=1}^{M} \mu_{\tilde{X}_m}.$$  

(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

$$\mathbb{E}\{\frac{1}{\tilde{X}_u}\} \geq \frac{1}{\mathbb{E}\{\tilde{X}_u\}}.$$  

(19)

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

$$\mu_H \approx \frac{1}{T} \sum_{t=1}^{T} \tilde{Y}_t \sum_{m=1}^{M} \mu_{\tilde{X}_m}.$$  

(20)

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

#### 3.2. Effect of Additive Noise

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

$$\mathbb{E}\{\tilde{Y}_u\} = \mathbb{E}\{\tilde{H} + \tilde{W}_u\} \approx \mu_H + \mathbb{E}\{\tilde{W}_u\}$$  

(21)

where $\tilde{H}$ and $\tilde{W}_u$ are assumed to be uncorrelated. This introduces an error term, $\mathbb{E}\{\tilde{W}_u\}$, 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 ($GM_{\tilde{X}_u}$).

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 $\tilde{Z} = \tilde{X}\tilde{H}$, which encapsulate $\tilde{X}$ and $\tilde{H}$ into one variable. Now $gVTS_1$ aims at alleviating the additive noise and finding $\tilde{Z}$. In this regard, $GM_{\tilde{X}_u}$ should be computed through adapting the $GM_{\tilde{X}_u}$ using $\tilde{H}$

$$\tilde{Z} \sim \sum_{m=1}^{M} p_{\tilde{X}_m} (m) N(\mu_{\tilde{X}_m}, \tilde{H}_d \mu_{\tilde{X}_m})$$  

(22)

where $\tilde{H}_d$ is $\text{diag}[\mu_H]$. 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, $\tilde{Z}$, would be an approximation for $\tilde{X}\tilde{H}$. 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.
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


