An Investigation of Temporally Varying Weight Regression for Noise Robust Speech Recognition

Shilin LIU, Khe Chai SIM

School of Computing, National University of Singapore, Singapore
{shilin,simkc}@comp.nus.edu.sg

Abstract

In this paper, recently proposed Temporally Varying Weight Regression (TVWR) is investigated in two ways for noise robust speech recognition. Firstly, since typical model compensation approaches assume that the noise feature is independent and identically distributed, non-stationary noise environment can be poorly compensated using conventional model compensation approaches in the standard Hidden Markov Model (HMM) framework. TVWR, however, maintains both the basic HMM structure and additional time-varying property, therefore, model compensation for TVWR is proposed such that i.i.d. noise assumption can be relaxed. Secondly, although Noise Adaptive Training NAT has been proposed to optimize the "pseudo-clean" HMM model for a better performance by maximizing the likelihood of multi-condition data, NAT heavily depends on the simplicity of Vector Taylor Series (VTS) formulation. Hence, other advanced compensation approaches, such as Trajectory-based Parallel Model Combination (TPMC), have difficulties benefiting from this powerful training schema. This paper exploits the time-varying attribute of TVWR to approximate NAT such that any compensation technique can be applied during noise adaptive training. Experiments on the Aurora 4 corpus show that significant improvements over the standard HMM or NAT system can be obtained by compensating TVWR either trained using clean data or adaptively trained using multi-condition data.

Index Terms: trajectory modelling, noise robustness, adaptation, adaptive training

1. Introduction

Noise robust speech recognition has been studied for many years but not solved yet due to its unsatisfactory accuracy for various noise conditions. One of the most popular approaches to solve this problem is to learn the impact of the noise on the corrupted speech such that the underlying acoustic model can be adapted to various unknown noisy environments. In order to apply the model compensation approaches such as [1, 2, 3, 4, 5], the noise model representing the noise evolution along with the time instance is expected. However, this noise model is typically modelled as a single state evolution process with single Gaussian model, which assumes that the noise feature is independent and identically distributed. Although this assumption simplifies both the noise model estimation process and the model adaptation formulation, it makes the non-stationary noise poorly modelled. Many techniques [6, 7, 8, 9, 10] have been introduced to tackle the non-stationary noise problem, many of them have complicated formulations and require iterative estimation process. None of above approaches have solved this problem from the model compensation perspective: the adapted acoustic model works exactly the same as the clean model does in the clean environment.

On the other hand, Noise Adaptive Training NAT [11] has been proposed to optimize a "pseudo-clean" canonical HMM model by maximizing the likelihood of multi-condition training data. Although promising improvements have been obtained by NAT, this powerful training schema strictly depends on the formulation of VTS adaptation [1] in order to derive tractable update formulae. This means that other advanced model compensation approaches such as Data-driven Parallel Model Combination (DPMC) [5] and Trajectory-based Parallel Model Combination (TPMC) [4] have difficulties in benefiting from this training schema, since it is very difficult to evaluate the derivative of the corrupted statistics with respect to clean statistics or noise statistics for the optimization process.

In this paper, above two issues will be solved based on the recently proposed temporally varying weight regression (TVWR) [12] framework. Since the temporal correlation of the speech is partially modelled into the time-varying GMM weights, equivalent temporal correlation of the noise feature will be required to be modelled for consistent compensation of TVWR model such that non-stationary noise environment can be better characterized. In this paper, we will firstly introduce how TVWR model can be adapted for handling the non-stationary noise environment. Secondly, NAT is factorized with a time-varying factor together with the conventionally compensated component output probability, the time-varying attribute of TVWR is exploited to approximate such time-varying factor such that complex compensation techniques, such as TPMC [4], can benefit from this powerful training schema.

The remaining of this paper is organized as follows. Section 2 gives an overview of previously proposed TVWR. Approaches to make TVWR noise robustness are introduced in Section 3. Experimental evaluation are reported in Section 4.

2. TVWR Overview

The basic idea of TVWR formulation is to approximate a standard HMM system using a long span of observations as input features such that the system can benefit from explicitly modelling the temporal correlation of successive observations and the system complexity is still relatively low. In TVWR, the output probability of state \( j \) is expressed as following mixture models:

\[
p(\omega_j, \tau_t | j) = \sum_{m=1}^{M} \frac{p(m | j)p(\tau_t | \omega_j, j, m)}{N(\omega_j | \mu_j, \Sigma_j)}
\]

(1)
where $\tau_t$ is a limited acoustic context of the clean observation $o_t$, denoted as $\tau_t = \{o_{t-\delta}, \ldots, o_{t-1}, o_{t+1}, \ldots, o_{t+\delta}\}$, where $\delta$ is the context expansion size, and $e_{j,m}$ is the temporally varying weight to be modelled. Approximations are made for a compact representation of $p(\tau_t|o_t, j, m)$:

$$p(\tau_t|o_t, j, m) \approx p(\tau_t|j, m)$$

$$= \sum_{i=1}^{N} p(\tau_t|i, j, m) P(i|j, m)$$

$$\approx Z_i \sum_{i=1}^{N} p(i|\tau_t) P(i|j, m)$$

where $Z_i = p(\tau_t)/P(i)$ is the component independent normalization term, which can be ignored during likelihood calculation, $i$ is a latent variable and uniform prior $P(i)$ is assumed during the above derivation. Typically, $i$ is defined as the mono-phone such that the posterior feature $p(i|\tau_t)$ can be easily predicted from a neural network or any other classifier. For convenience, $h_{1} = p(\tau_t|i), w_{j,m} = P(i|j, m)$, and $e_{j,m} = P(m|j)$ are defined for future reference.

As can be seen, TVWR system is simply a standard HMM system with additional time-varying GMM weights. Therefore, TVWR system estimated from the typical clean data will suffer from the same performance degradation if deployed in a noisy environment like the regular HMM system. Feature enhancement or multi-condition training data may be applied to make TVWR noise robust without any modification. However, this paper will focus on investigating in two ways to make TVWR noise robust: 1) model compensation for TVWR to handle the non-stationary noise environment; 2) approximating Noise Adaptive Training using TVWR to allow wider application of noise adaptive training.

### 3. Noise-robust TVWR

In this section, model compensation for TVWR will be introduced first, which will form the basis for the subsequent derivation of NAT approximation using TVWR. Before that, $x_t$, $y_t$, $n_t$ are defined as the clean, noisy speech and noise cepstral feature, respectively. (For convenience, only additive noise is considered in this paper.) Hence, the respective long span of each feature will be denoted as:

$$\chi_x = \{x_{t-\delta}, \ldots, x_{t-1}, x_{t+1}, \ldots, x_{t+\delta}\}$$

$$\psi_x = \{y_{t-\delta}, \ldots, y_{t-1}, y_{t+1}, \ldots, y_{t+\delta}\}$$

$$\eta_x = \{n_{t-\delta}, \ldots, n_{t-1}, n_{t+1}, \ldots, n_{t+\delta}\}$$

#### 3.1. Model Compensation for TVWR

In order to improve the noise robustness of TVWR, its formulation in Eq.1 needs to be compensated properly. When TVWR system is estimated using the maximum likelihood training criterion, the Gaussian parameters in $p(o_t|j, m)$ will represent the distribution of the speech data. In that case, any of the conventional compensation approaches [1, 2, 3, 4, 5] can be applied to adapt these Gaussian parameters. On the other hand, according to Eq.3, the posterior feature $p(i|\tau_t)$ also depends on the acoustic features, which are condition dependent. The condition dependent normalization term contributes uninformative knowledge to discriminate different components or states, so that it can be ignored even in a mismatch condition. As with the standard model compensation methods, $P(i|j, m)$, $P(m|j)$ are assumed to be unaffected by the channel and noise distortion.

Therefore, the remaining problem is how to generate condition dependent posterior feature. In the original TVWR [12] work, the posterior feature is generated using a neural network for better accuracy. However, it is difficult to directly adapt the neural network or any other discriminate models [13], as their parameters do not represent the statistics of the features. To circumvent this problem, generative models using long span features are introduced:

$$p(i|\chi_x) = \frac{N(\chi_x; \mu_i^k, \Sigma_i^k) P(i)}{\sum_i^N N(\chi_x; \mu_i^k, \Sigma_i^k) P(i)}$$

where $\{\mu_i^k, \Sigma_i^k\}$ represents the clean generative model for latent variable $i$. In order to obtain the condition dependent posterior feature, model compensation for these clean generative models can be performed. Given the long span noise model $\{\mu_{n}^k, \Sigma_{n}^k\}$ for condition $k$, the noisy speech model $\{\mu_{\psi}^k, \Sigma_{\psi}^k\}$ for $i$ at condition $k$ can be obtained using extended VTS [3]:

$$\mu_{\psi}^k \approx \mu_{\psi} + Q \log \left[ 1 + \exp \left( Q^{-1} (\mu_{n,k} - \mu_{\psi}^k) \right) \right]$$

$$\Sigma_{\psi}^k = J \Sigma_{\psi} J^T + (I - J) \Sigma_{n} (I - J)^T$$

where

$$Q = I_{2\delta} \otimes C$$

$$J = Q \mathrm{diag} \left\{ 1 + \exp \left( Q^{-1} (\mu_{n,k} - \mu_{\psi}^k) \right) \right\} Q^{-1}$$

$I_{2\delta}$ is an identity matrix with dimension $2\delta$, $C$ and $C^{-1}$ are the DCT matrix and its pseudo-inverse matrix, $\mathrm{diag} \{ \}$ denotes a diagonal matrix using the input vector element as its diagonal element. Based on above equations, block-wise (a.k.a. frame-wise) covariance matrix at index $p$ and $q$ can be re-written as:

$$\Sigma_{\psi, pq}^k = J_p \Sigma_{\psi, pq}^k J_q + (I - J_p) \Sigma_{n, pq}^k (I - J_q)$$

where $J_p$, $J_q$ indicates the diagonal block of $J$ at index $p$ and $q$ respectively, and the block index ranges: $p, q \in [1, 2\delta]$. Conventional model compensation approaches in HMM framework assume that:

$$\Sigma_{n, pq}^k = 0 \quad \text{if } p \neq q \quad (14)$$

$$\Sigma_{n, pp}^k = \Sigma_{n, qq}^k \quad \forall p, q \quad (15)$$

which is poorly made for non-stationary noise environment. In order to compensate TVWR model, the correlation of successive noise data is explicitly modelled into $\Sigma_{n, pq}^k$ and compensated into the statistics $\Sigma_{\psi, pq}^k$ of the corrupted speech, the i.i.d noise assumption can be relaxed. However, the typical utterance level adaptation may not be able to provide sufficient noise data to estimate a full covariance noise model, which is particularly important for non-stationary noises, we will focus on model compensation at the condition level in this preliminary work.

#### 3.2. NAT Approximation using TVWR

In the recently proposed noise adaptive training NAT [11] technique, both the "pseudo-clean" speech and noise models are optimized to maximize the likelihood of the multi-condition data given the condition adapted noisy speech models. The state
emission probability for condition $k$ after NAT on HMM system can be expressed as:

$$p(y^k_j | j, k) = \sum_{m=1}^{M} c_{jmk}^{\text{NAT}} N(y^k_j ; \mu_{jmk}, \Sigma_{jmk})$$  \hspace{1cm} (16)$$

where $c_{jmk}^{\text{NAT}}$, $\mu_{jmk}$, and $\Sigma_{jmk}$ are NAT optimized weights and compensated speech model for condition $k$, $\mu_{jmk}$, $\Sigma_{jmk}$ are the conventionally compensated speech model without NAT, $z_{jmk}^{\text{NAT}}$ is the time-varying scaling factor to make the above equation hold. From the perspective of Eq-17, the existence of non-constant $z_{jmk}^{\text{NAT}}$ after NAT tells that the conventional compensation does not yield maximum likelihood of the condition specific noisy speeches. Since the auxiliary function of NAT is a highly nonlinear function of the variance variable, the variance update requires gradient or Newton’s method, which somehow complicates the optimization process [11].

Similarly, the state emission probability for TVWR after adaptation to condition $k$ can be expressed as:

$$p(y^k_j, \psi^k_j | j, k) = \sum_{m=1}^{M} c_{jkm} N(y^k_j ; \mu_{jmk}, \Sigma_{jmk})$$  \hspace{1cm} (18)$$

where $c_{jkm}$ is the time-varying weight adapted to condition $k$, $\psi^k_j$ is the long span noisy context from condition $k$. When comparing Eq-17 with Eq-18, both NAT and TVWR share some similar structure but differ in the underlying formulation and training process. Since the superiority of NAT over the conventional compensation approach during the above training process. Since Gaussian parameters in the above training process are not optimized to be “pseudo-clean”, the possible accuracy loss by using the initial Gaussian parameters will be compensated by adaptive training regression parameters. Note that the power of above training schema heavily benefits from the large number of regression parameters, which is about half of that of Gaussian parameters. Compared to the complicated update formulae of Gaussian parameters in the standard NAT formulation [11], update formulae for TVWR regression parameters are much simpler and easier. Although above training process assumes that the condition independent Gaussian parameters is unchanged, these Gaussian parameters may be updated in step 5 just using the standard HMM update formulae with a better forward-backward alignment. On the other hand, if VTS compensation is applied, the “pseudo-clean” speech and regular sized noise model parameters may also be updated just following the standard NAT procedure [11]. However, as this paper will focus on approximating NAT using TVWR formulation and any compensation approach, noise adaptive training of TVWR system (including generative models for the posterior feature and long span noise model with full covariance matrix) will be our future work.

4. Experimental Results

In this section, experiments are conducted on the Aurora 4 [14] corpus, which contains 3 datasets: clean dataset, multi-condition dataset, multi-noise dataset. Each dataset includes about 14 hours speech utterances, or 7138 utterances. Since only additive noise is considered in this paper, the multi-condition dataset with channel distortion will not be used. In
multi-noise training dataset, 6 noises are artificially added at 15 dB average SNR. For evaluation, only the first 7 test sets from the same microphone as the training data are used. For convenient presentation, Set-A and Set-B are introduced: Set-A contains 1 clean test set while Set-B contains 6 noisy test sets with the same noises as multi-noise training data but different 10 dB average SNR. Each test set has the same 330 utterances for 5k closed vocabulary recognition task (NIST Nov’92 WSJ0).

The clean baseline HMM system is a decision tree state-clustered triphone system with 3226 distinct states and 16 components per state. The acoustic features are 39 dimensional MFCC [15], including 12 static coefficients, zero-th coefficient and the first two derivatives. Given the clean baseline system, multi-noise HMM system was estimated by one iteration of single pass retraining. Monophone posterior features, $h_{11}$, were generated from multiple GMM models: each monophone has 8 full-covariance mixture components, which may be estimated from the clean or multi-noise data. The context expanded features used here include a sequence of MFCC features spanning a window of 8 static features, i.e. 13*8 dimensions.

The model compensation approaches investigated here include VTS [1] and Trajectory PMC (TPMC) [4], where TPMC has been shown to be better than VTS. To perform the noise adaptive training, only multi-noise training data are used. While the conventional NAT re-estimated both the “pseudo-clean” speech and noise models using VTS, NAT approximation using TVWR (NA-TVWR starting from multi-noise HMM) only updated the regression parameters and static weights (NA-TVWR* updated condition independent Gaussian parameters using the standard HMM update formulae). In the decoding, noise models are initially estimated by the head and tail (20) observations from each condition. To evaluate model compensation on the clean model, these initial noise models are always used without re-estimation for a fair comparison to TPMC. When evaluating VTS-based NA-TVWR and NAT-HMM, these noise models are re-estimated iteratively using multiple (2) iteration recognized hypotheses. The recognition results below includes a bigram full decoding followed by a trigram lattice-rescoring using HTK [16].

The recognition results of various approaches using different training data are shown in Table 1 and Table 2, where the “Avg” column indicates the average performance of all mentioned 7 test sets. When all systems are estimated using the clean data, both the standard HMM and TVWR systems without compensation suffered from dramatic performance degradation in the noisy test conditions. After performing model compensation, all the systems significantly outperformed the baseline systems. Compared to the standard HMM system using either VTS or TPMC, both adapted TVWR systems show about 2% absolute average improvements. Since most testing speeches are corrupted by non-stationary noise, modelling the temporal correlation of the noise feature makes TVWR gain more than the standard HMM system. On the other hand, although more accurate corrupted dynamic parameters (delta, delta-delta) were estimated by TPMC for a better representation of the impact of non-stationary noise on the speech, TVWR can still obtain significant improvement. This tells that some characteristics of non-stationary noise corrupted speech cannot be modelled well in the standard HMM framework.

When noise adaptive training on the multi-noise data was performed, NAT-HMM+VTS obtained more than 2% absolute improvement over the multi-noise HMM. NAT-HMM+VTS achieved better recognition accuracy on the clean test Set-A, probably because its “pseudo-clean” speech model has been optimized while TVWR still use the multi-noise estimated Gaussian parameters. However, our approximation approach, NA-TVWR+VTS achieved 0.9% absolute more average improvement over NAT-HMM+VTS. These results show that adaptively trained time-varying weight is able to approximate some property of NAT, while the additional gain from TVWR over NAT is probably due to the useful temporal correlation of noise features. When NA-TVWR was performed with TPMC, NA-TVWR+TPMC averagely outperformed other VTS-based systems. Although NA-TVWR+TPMC still is not as good as NAT-HMM+VTS at test Set-A due to using the multi-noise data estimated Gaussian parameters, NA-TVWR+TPMC performs quite well averagely and comparable to TVWR+TPMC trained by the clean data. After the condition independent Gaussian parameters were optimized just using the standard HMM update formulae, NA-TVWR+TPMC performed best and also slightly better than the TPMC-based clean TVWR system. According to these results, any other compensation approach can potentially perform quite well given the multi-condition data in the TVWR framework.

Table 1: Word Error Rate(%) for different approaches using clean training data.

<table>
<thead>
<tr>
<th>System</th>
<th>Adaptation</th>
<th>Set-A</th>
<th>Set-B</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td></td>
<td>6.4</td>
<td>51.0</td>
<td>44.7</td>
</tr>
<tr>
<td>TVWR</td>
<td>-</td>
<td>5.8</td>
<td>56.5</td>
<td>49.3</td>
</tr>
<tr>
<td>HMM</td>
<td>TVWR</td>
<td>7.8</td>
<td>18.3</td>
<td>16.7</td>
</tr>
<tr>
<td>HMM</td>
<td>VTS</td>
<td>6.8</td>
<td>16.0</td>
<td>14.7</td>
</tr>
<tr>
<td>HMM</td>
<td>TPMC</td>
<td>6.9</td>
<td>14.8</td>
<td>13.6</td>
</tr>
<tr>
<td>TVWR</td>
<td>TPMC</td>
<td>6.2</td>
<td>12.3</td>
<td>11.4</td>
</tr>
</tbody>
</table>

Table 2: Word Error Rate(%) for different approaches using multi-noise training data.

<table>
<thead>
<tr>
<th>Training</th>
<th>Adaptation</th>
<th>Set-A</th>
<th>Set-B</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-HMM</td>
<td></td>
<td>8.5</td>
<td>18.7</td>
<td>17.2</td>
</tr>
<tr>
<td>NAT-HMM</td>
<td>VTS</td>
<td>7.3</td>
<td>16.0</td>
<td>14.8</td>
</tr>
<tr>
<td>NA-TVWR</td>
<td>VTS</td>
<td>8.1</td>
<td>14.9</td>
<td>13.9</td>
</tr>
<tr>
<td>NA-TVWR</td>
<td>TPMC</td>
<td>7.9</td>
<td>12.3</td>
<td>11.8</td>
</tr>
<tr>
<td>NA-TVWR+</td>
<td>TPMC</td>
<td>7.2</td>
<td>11.8</td>
<td>11.1</td>
</tr>
</tbody>
</table>

5. Conclusions

In this paper, the recently proposed Temporally Varying Weight Regression (TVWR) model is investigated in two ways to perform noisy speech recognition. Firstly, model compensation for TVWR is proposed such that i.i.d. noise assumption can be relaxed, which can potentially improve the accuracy in the non-stationary noise environment. Secondly, instead of directly updating the model and noise parameters in the standard NAT procedure, the time-varying attribute of TVWR is used to approximate the NAT algorithm such that any compensation approach can benefit from the multi-condition data training. Experimental results show that promising improvements can be obtained by performing model-based compensation for TVWR. Besides, the proposed NAT approximation technique achieved significant improvement over the traditional NAT using VTS compensation. This approximation can benefit more complicated compensation technique, such as Trajectory-based PMC, where efficient NAT algorithm cannot be derived.
6. References


