UNSUPERVISED NEURAL-NETWORK BASED VOCAL TRACT LENGTH NORMALISATION

Philip Harding[1] *and Matthew Gibson*

Nuance Communications Ltd., Cambridge, UK
[1]*philip.harding@nuance.com*

Abstract: In this paper an efficient, unsupervised, method of warp factor estimation for vocal tract length normalisation (VTLN) is proposed. VTLN is a method of feature-based speaker normalisation where the frequency spectrum is warped to produce speaker-independent spectral features that are invariant to vocal tract length. The degree to which the spectrum is warped is determined by the warping factor, and it is the estimation of this warping factor that is the focus of this paper. The warping factor is typically obtained using a maximum likelihood-based technique that requires a state alignment for each utterance and a GMM acoustic model trained on warped features. The warp factor is typically quantised, with one of N warp factors selected for each utterance. The proposed method of warp factor estimation makes use of a small neural network, trained on un-warped features, to directly estimate the quantised warp factor. Experimental results are presented where, unlike previously published methods of unsupervised warp factor estimation [1, 2], the proposed method is shown to give equivalent performance to the typical supervised GMM-based method in terms of ASR accuracy at a significantly lower computational cost.

1 Introduction

Performance of automatic speech recognition (ASR) is influenced by acoustic mismatches between training and evaluation data. These mismatches include environmental factors such as noise and reverberation, as well as varying speaker characteristics. A range of techniques have been developed to mitigate against these mismatches, and these can be broadly classified as being either model-based or feature-based [3]. Model-based approaches include multi-condition training and model adaptation, whilst feature-based approaches include CMLLR, i-vectors and vocal tract length normalisation (VTLN), and it is VTLN on which this paper is focused.

VTLN warps acoustic features in the spectral domain such that the resulting features are mapped onto the feature space of a canonical speaker [4]. The degree to which the spectrum is warped is determined by the warping factor and this must be estimated for every utterance, unless the speaker is known in advance [5]. Both supervised and unsupervised methods of estimation have been proposed in the past, with best performance observed using computationally expensive, supervised, methods [1, 2, 6, 7]. This paper proposes a computationally inexpensive method of estimation which is shown to give equivalent performance to the current best method.

Existing methods of warp factor estimation (WFE) are summarised in Section 2, with a description of the proposed method given in Section 2.3. Experimental results are then presented in Section 3, which includes a justification of the use of VTLN with HMM-DNN acoustic models in Section 3.2. Final results are presented in Section 3.4.

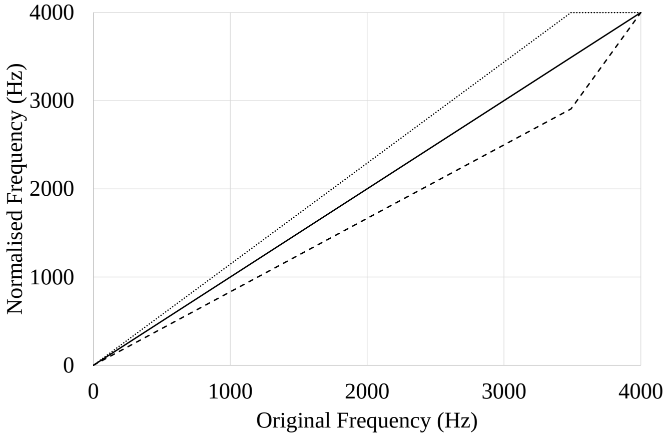

Figure 1 – Piecewise function used to warp spectral features in this work. Dashed and dotted lines illustrate maximum degrees of warping (warping factor range of 0.85 - 1.15).

2 Vocal-Tract Length Normalisation (VTLN)

There are two key components to VTLN: i.) the method of warping the spectrum, and ii.) the method of estimating the degree to which the spectrum is warped. Section 2.1 gives details of the warping function used in this work whilst Sections 2.2 and 2.3 describe existing methods of WFE and the proposed method of estimation, respectively.

2.1 Warping functions

The warping function parameterises how the spectrum is warped. Such functions typically take a single parameter, the warp factor. Three typical warp functions are illustrated in [6], namely: piecewise-linear, power function and combined piecewise-linear power function. In this work the piecewise-linear warping function is used, as illustrated in Figure 1.

2.2 Existing methods of warp factor estimation

Most existing methods of WFE utilise GMMs. Best results are observed when a text-dependent maximum-likelihood method of estimation is used where acoustic feature vectors, X are warped with each warp factor, α, and then the warp factor that maximises the probability in Equation 1, $\hat{\alpha}$, is chosen where θ represents the acoustic model [6].

$$\hat{\alpha} = \arg\max_{\alpha} \Pr\left(X^{\alpha} | \theta, W\right) \tag{1}$$

This method is computationally expensive as an initial recognition pass must be run to obtain the word sequence, W. Several unsupervised methods have been proposed that remove the requirement of W, but these methods do not usually perform as well as the supervised method [1, 2, 6]. In this work the method described in Equation 1 is used as a baseline.

2.3 Neural-network based warp factor estimation (NN-WFE)

The method of WFE proposed in this work uses a simple feed-forward neural network to estimate warp factor index (WPX) posteriors. A range of warping factors are quantised to N equally spaced warp factor indices. The neural-network warp factor estimation (NN-WFE) model is then trained with $N+1$ output states, where the additional state models non-speech. This differs from previous approaches where either the warp factor is directly estimated [2], or where, instead of warping the acoustic features with the estimated warp factor, WPX posteriors are augmented with MFCCs for acoustic modelling [8].

Training follows the standard stochastic gradient descent algorithm using the cross-entropy criterion. This method requires every frame to be associated with a reference label indicating either the target warp factor or the absence of speech. In this work, reference warp factors are obtained using the existing supervised GMM-based method. Once the model has been trained WPX estimates can be obtained by running a forward pass of the model at each frame. This level of granularity is not typically required; usually we are most interested in estimating the WPX at the utterance level. There are several methods through which these frame-level estimates can be converted to utterance-level estimates, two of which are described below:

Voting The output state with the highest posterior (either one of N warp factor indices or non-speech) is selected at each frame and counted as a single vote. After all frames of the utterance have been processed, the WPX with the highest number of votes is selected.

Summed posteriors Posteriors are summed for each state across all frames of the utterance. Once all frames have been processed, the WPX corresponding to the state with the highest summed posterior is selected (excluding the non-speech state).

These techniques are evaluated in Section 3.3.1.

3 Experimental Results

Single-speaker telephony data, namely voicemail recordings, are used to evaluate the proposed method; Section 3.1 describes the experimental setup in more detail. Next, the motivation behind the use of VTLN with neural-network acoustic models is justified in Section 3.2, before warp selection methods and optimal network configurations for NN-WFE are explored in Sections 3.3.1 and 3.3.2 before final results are presented in Section 3.4.

Unless otherwise stated, reported word error rate reductions (WERR) and runtime factor reductions (RTFR) are relative to an equivalent model trained on unwarped features.

3.1 Experimental setup

Training and test data is sampled from two voicemail-to-text deployments: English as spoken in the USA (eng-USA) and Castilian Spanish (spa-ESP). In both cases, training sets comprise 2000 hours of audio, with a further 10-20 hours split between development and evaluation sets.

MFCCs are used as the front-end for all models. The filterbank has 24 channels, and subsequently all 24 DCT coefficients are retained and then cepstral mean normalised at the utterance level. Warping factors are quantised to 16 warp factor indices where an index of 7 is equivalent to no warping. The maximum range of warping is illustrated in Figure 1.

For the results in this paper, feed-forward neural-network acoustic models are evaluated (HMM-DNN). Acoustic models are trained using the standard cross-entropy criterion, and then further refined by sequence training. n-gram and RNN language models, trained on voicemail messages of the target language, are used for search and rescoring, respectively.

A three stage decoding workflow is used, consisting of: i.) search, ii.) language model rescoring and iii.) confusion network decoding. For the conventional method of WFE a small HMM-GMM acoustic model is used to generate the word sequence, W, required for estimation.

3.2 VTLN with neural-network acoustic models

In the past, significant improvements in accuracy were observed when VTLN features were used in the front-end of HMM-GMM based systems [3]. Since then, more complex neural network

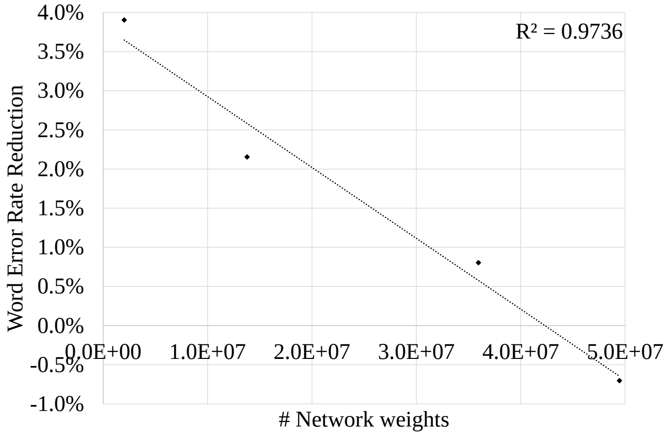

Figure 2 – Relative improvement observed by applying VTLN on spa-ESP voicemail data as a function of HMM-DNN acoustic model size. Relative improvements are measured against models trained on unwarped data whilst warped systems use existing GMM-based supervised method of WFE.

based models have become prominent, and there is an argument that these more complex models can be trained to be robust to changes in speaker characteristics [9].

In order for a neural network to become invariant to individual speaker characteristics there must exist sufficient modelling capacity within the network. There is therefore the possibility that smaller networks may still benefit from feature engineering. To evaluate this, a range of HMM-DNN systems were trained with a variety of network topologies, with and without applying VTLN. Figure 2 plots the improvement from VTLN as a function of the total number of weights in the acoustic model. Almost 3.5% WERR was measured with the smallest model, with improvements from VTLN observed to be inversely proportional to network size to the point where the largest network, with almost 50M weights, experiences a slight degradation from the use of VTLN features. Whilst best absolute WER is observed with the largest of the networks, small networks are useful in applications where very high throughput (and therefore very low computational cost) is demanded, and it is in these scenarios that the conventional approach to WFE has been prohibitively expensive. The proposed method of WFE has very low computational cost and therefore allows the use of VTLN in such scenarios.

3.3 Neural-network warp factor estimation

3.3.1 Warp selection method

The two utterance-level warp selection methods described in Section 2.3 are now compared. A NN-WFE model was trained on eng-USA data with varying context windows. Figure 3 compares the summed posterior against the voting scheme in terms of final WER as the context window of the neural network is increased; no significant difference was observed and so the summed posterior method is selected for its theoretically superior properties.

3.3.2 Network configuration

The size of the output layer is fixed by our existing configuration where the range of warp factors is quantised in to sixteen warp indices. In terms of the size of the network this leaves two properties to optimise: i.) the breadth of the context window at the input layer, and ii.) the number and size of the hidden layers.

The effect of altering the size of the input context window is first evaluated by training a range of networks with a single hidden layer of 256 nodes and then sweeping across a range of context window sizes. Symmetric context windows are used where *n* frames either side of the current frame are presented at the input layer. Context is reported in terms of *n*, and so the total

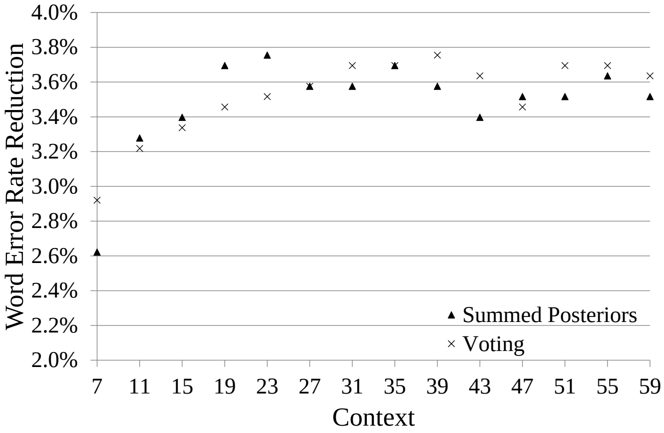

Figure 3 – WERR from proposed method over unwarped model comparing summed posterior and voting utterance level WFE methods in terms of network input context

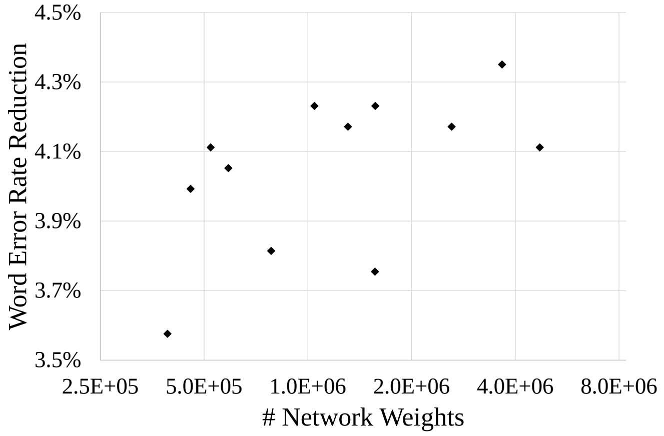

Figure 4 – WERR from proposed method over unwarped model as a function of network size

number of input frames is $2n+1$. Performance of the estimation network is shown in terms of WERR relative to a non-VTLN baseline in Figure 3.

Performance is shown to improve until $n = 31$, at which point performance plateaus. The input context is therefore fixed at $n = 31$ in all future experiments. Given a frame analysis window of 25ms and a shift of 10ms this corresponds to a temporal window of 645ms.

Next, the configuration of the hidden layers is considered. A range of network topologies with one to four hidden layers, each with 256 to 1024 nodes, were evaluated. The effect of varying the complexity of the network is illustrated in Figure 4.

Best performance is observed using a network comprising of 3.7M weights (three hidden layers with 1024 nodes each), however similar performance (within 0.12% WERR) is observed with a much smaller network, comprising 1.0M weights (2 hidden layers with 512 nodes each). In the interests of computational complexity, the smaller network is selected.

In Section 3.2 it was noted that increasing the size of the acoustic model improves WER. For this method of WFE to be useful, the use of a model with 1.0M weights must give an improvement greater than just increasing the size of the acoustic model by a similar number of weights. Linear regression was used to map between acoustic model size and WER, with and without VTLN; WER was then estimated before and after adding 1.0M weights to a range of acoustic model topologies. In all cases, gains from VTLN were significantly greater.

Figure 5 compares the warp factors estimated by the supervised GMM-based method and the unsupervised DNN-based method. The distribution of warp factors typically observed for male speakers (lower warping factors) is broadly similar, whilst for warp factors typically observed for female speakers the variance of the distribution of estimated warp factors is smaller

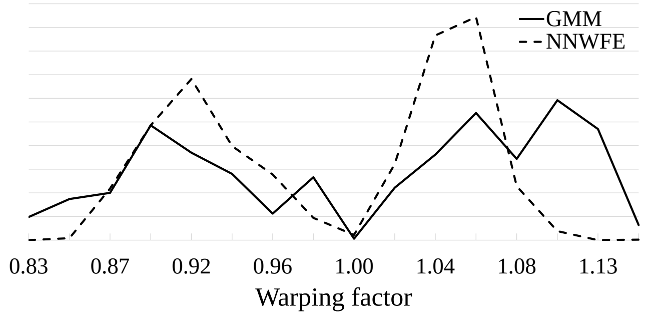

Figure 5 – Distributions of estimated warp factors on development set, comparing supervised GMM-based method and unsupervised neural network based method

than that of the GMM-based method.

3.4 Final results

The final method is evaluated on both eng-USA and spa-ESP systems trained on voicemail data. Performance is evaluated relative to a non-VTLN baseline and compared against the supervised GMM-based WFE in Table 1. Acoustic scoring and search parameters of the recognition pass were tuned to best WER at the RTF of the chosen speed mode ($\ll$ real-time) for each system. In the case of the system using the baseline GMM-based WFE, the parameters of the HMM-GMM recognition pass used to obtain W were tuned independently; however, when added to the cost of the main recognition pass these systems are found to exceed the RTF target by a large margin. Separate attempts to jointly tune both passes to the target RTF failed to yield a working point within the RTF budget.

Table 1 – Final results, presented on eng-USA and spa-ESP voicemail systems tuned to the same real-time factor, comparing GMM-based WFE and proposed NN-based WFE

Language	**GMM**		**NN-WFE**	
	WERR	**RTFR**	**WERR**	**RTFR**
eng-USA	4.65	-211.84	5.20	0.00
spa-ESP	2.77	-79.60	3.22	0.37

Due to the very low computational cost of estimating the warp factors using NN-WFE, both NN-WFE based systems meet the RTF target. Word error rate reductions of 3.2 - 5.2% exceed the improvements observed using the GMM-based method to give final systems that are both faster and more accurate than the previous approach.

4 Summary

The proposed method of unsupervised neural network-based warp factor estimation has been shown to meet, and in some cases exceed, performance of the traditional supervised GMM-based method at a significantly lower cost at runtime. This enables the use of VTLN in applications where very low computational complexity is required, with small HMM-DNN acoustic models shown to benefit most.

References

[1] CERVA, P., K. PALECEK, J. SILOVSKY, and J. NOUZA: *Using Unsupervised Feature-Based Speaker Adaptation for Improved Transcription of Spoken Archives.* In *Twelfth Annual Conference of the International Speech Communication Association*. 2011.

[2] PACZOLAY, D., A. KOCSOR, and L. TÓTH: *Real-time Vocal Tract Length Normalization in a Phonological Awareness Teaching System.* In *Text, Speech and Dialogue*, pp. 309–314. Springer, 2003.

[3] WOODLAND, P. C.: *Speaker Adaptation for Continuous Density HMMs: A Review.* In *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*. 2001.

[4] ACERO, A. and R. M. STERN: *Robust Speech Recognition by Normalization of the Acoustic Space.* In *Acoustics, Speech, and Signal Processing, International Conference on*, pp. 893–896. IEEE, 1991.

[5] ZHAN, P. and A. WAIBEL: *Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition.* Tech. Rep., Carnegie-Mellon University Pittsburgh, PA, School of Computer Science, 1997.

[6] MOLAU, S., S. KANTHAK, and H. NEY: *Efficient Vocal Tract Normalization in Automatic Speech Recognition.* In *Proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*. Citeseer, 2000.

[7] EIDE, E. and H. GISH: *A Parametric Approach to Vocal Tract Length Normalization.* In *Acoustics, Speech, and Signal Processing, International Conference on*, vol. 1, pp. 346–348. IEEE, 1996.

[8] SERIZEL, R. and D. GIULIANI: *Vocal Tract Length Normalisation Approaches to DNN-based Children's and Adults' Speech Recognition.* In *Spoken Language Technology Workshop (SLT)*, pp. 135–140. IEEE, 2014.

[9] SEIDE, F., G. LI, X. CHEN, and D. YU: *Feature Engineering in Context-Dependent Deep Neural Networks for Conversational Speech Transcription.* In *Automatic Speech Recognition and Understanding, IEEE Workshop on*, pp. 24–29. IEEE, 2011.