Estimating VTLN Warping Factors by Distribution Matching

Janne Pylkkönen

Adaptive Informatics Research Center, Helsinki University of Technology, Finland

janne.pylkkonen@tkk.fi

Abstract

Several methods exist for estimating the warping factors for vocal tract length normalization (VTLN), most of which rely on an exhaustive search over the warping factors to maximize the likelihood of the adaptation data. This paper presents a method for warping factor estimation that is based on matching Gaussian distributions by Kullback-Leibler divergence. It is computationally more efficient than most maximum likelihood methods, but above all it can be used to incorporate the speaker normalization very early in the training process. This can greatly simplify and speed up the training. The estimation method is compared to the baseline maximum likelihood method in three large vocabulary continuous speech recognition tasks. The results confirm that the method performs well in a variety of tasks and configurations.

Index Terms: speech recognition, speaker normalization, VTLN, warping factor estimation

1. Introduction

Vocal tract length normalization (VTLN) [1] is a widely used method for reducing the inter-speaker variability and the mismatch between the acoustic models and the new speakers. When used to normalize training set speakers in a manner of speaker adaptive training (SAT) it can help to produce more compact models and enable more efficient acoustic modeling [2]. As VTLN only needs one parameter to be estimated, it can be applied even with very limited amount of adaptation data.

Several methods for estimating the VTLN warping factors have been suggested. Traditionally a grid of warping factors is searched based on the maximum likelihood (ML) criterion over the adaptation data. This, however, requires evaluating the adaptation data several times using all possible warping factors. Computationally more efficient methods have been developed which use simpler acoustic models for VTLN estimation [1] or warping factor specific models [3] to select the optimal warping factor according to ML criterion. One popular alternative method not related to ML is to use formant estimation to predict the required frequency warping [4]. However, this can have problems with the robust estimation of the formant frequencies.

Along with the recent formulation of VTLN as a linear transformation of the cepstral coefficients [5, 6] new methods for VTLN estimation have become available. If the transformation is seen as a model adaptation rather than speaker normalization, the warping factor can be estimated by using an auxiliary function and the EM-algorithm [6], thus avoiding the direct optimization of the likelihood of the adaptation data. If the VTLN transformation is at the end of the feature processing, sufficient statistics can be used in speaker normalization case as well [7].

Usually the training procedure for VTLN adapted models is rather complicated. First an unadapted model is trained, after which the training data is iteratively adapted and models re-estimated [3]. This may require additional target models as complex acoustic models may model too much of the inter-speaker variability, thus hindering the proper estimation of the warping factors according to the ML criterion [6]. On the other hand, it has been noted that it is beneficial to use VTLN already before estimating discriminative feature transformations [8]. In such cases the VTLN factors should be known already prior to model and feature estimation. Therefore it would be desirable if the VTLN warping factors could be estimated in a model independent manner prior to actual training.

This paper presents a method for estimating VTLN warping factors based on simple distribution matching. The method operates on collected speech statistics and it can be applied prior to training as long as a preliminary phoneme segmentation of the training data is available.

2. Distribution matching by Kullback-Leibler divergence

The method for estimating VTLN warping factors presented here is based on matching speaker's speech distributions to the reference distributions according to the Kullback-Leibler divergence. The Kullback-Leibler divergence or relative entropy [9] is defined as

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}, \tag{1}$$

where p and q are some distributions and $\mathcal{X}$ is the set of possible values of x. The use of this measure as the criterion for speaker normalization can be argued by its connection to maximum likelihood, as presented below.

Let us consider a model selection problem where the intention is to find a distribution $q(x)$ that matches $p(x)$ as closely as possible. For this we can write (1) as

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} p(x) \log q(x) \tag{2}$$

and drop the first term that is independent of the $q(x)$. The model selection can therefore be formulated as

$$\tilde{q}(x) = \operatorname*{argmin}_{q(x)} D(p||q) = \operatorname*{argmax}_{q(x)} \sum_{x \in \mathcal{X}} p(x) \log q(x) \tag{3}$$

which equals maximizing the likelihood over the set of possible values distributed as $p(x)$. If we know the true distribution $p(x)$ minimizing K-L divergence therefore equals ML.

In VTLN the model selection problem is very restricted, as we are only trying to find one warping factor to define the actual form of the frequency warping function. Although no longer justified by the connection to the ML, for warping factor estimation we will be using the K-L divergence in "reverse direction", that is, selecting the warping factor as

$$\tilde{\alpha} = \underset{\alpha}{\operatorname{argmin}}\, D(p(\alpha)||q). \tag{4}$$

This is due to convenience of computing the K-L divergence when the distributions are modeled with Gaussians, as it obtains the following formulation:

$$\begin{aligned} D(p(\alpha)||q) =& \frac{1}{2}\Big(\log|\Sigma_q| - \log|\Sigma_{p(\alpha)}| + \operatorname{tr}(\Sigma_q^{-1}\Sigma_{p(\alpha)}) + \\ & (\mu_q - \mu_{p(\alpha)})^T \Sigma_q^{-1} (\mu_q - \mu_{p(\alpha)})\Big). \end{aligned} \tag{5}$$

In this form only the covariance of the distribution q has to be inverted. To ensure the inversion is possible we use q as the reference distribution so that we can be sure we have enough data to avoid the covariance matrix to be ill-conditioned. The covariances of the reference distributions can also be inverted in advance so that no matrix inversions are needed when new estimations are made.

3. Warping factor estimation procedure

The VTLN warping factor estimation procedure consists of two phases. First the speaker dependent speech statistics are collected from the data, after which the optimal warping factor is determined such that the warped statistics best match the reference distributions. During the training the statistics are collected once for all speakers and the reference distributions are formed from the speaker statistics by iteratively estimating the warping factors and computing the reference distributions as the averages of the warped speaker statistics. A few iterations of the warping factor re-estimation are needed so that no substantial changes to the warping factors occur. However, these iterations only require the manipulation of speaker statistics, not the actual data. The reference distributions are finally saved for estimating the VTLN warping factors for the new speakers.

3.1. Speech statistics

The statistics used in warping factor estimation are based on phoneme segmentation where each phoneme is represented by a three-state hidden Markov model (HMM). We used a segmentation generated by triphone models, after which context information was dropped and monophone statistics were collected. During evaluation the segmentation was obtained from the first pass recognition, during training we used aligned transcriptions.

Each HMM state is modeled by a single Gaussian so the statistics to be collected for each state are the mean and the covariance of the features. As we implement the VTLN as a linear transformation of the features (as in [6]), the effect of frequency warping can be simulated by applying the same linear transformation to these statistics. It is therefore enough to collect the statistics once, unwarped, and only apply the transformation to the statistics during the warping factor estimation.

The dimension of the Gaussians for the warping factor estimation is the dimension of the feature vector before VTLN transformation in the front-end. In our 16kHz system this was a 20-dimensional vector of mel-frequency cepstral coefficients (MFCC), in 8kHz case 15 coefficients were used to reflect the lack of higher frequency mel-filters. The mean values over a 1.25 second window had been removed from these MFCCs.

The benefit of using HMM states with single Gaussians over a Gaussian mixture model (GMM) is that as we know the phoneme segmentation prior to collecting the statistics, one pass through the data is enough for estimating the required distributions. With mixture models the data would have to be iterated several times in order to estimate the mixture models.

The Gaussian distribution of a certain HMM state was used in warping factor estimation if at least five samples were observed for that state. The silences were omitted in the processing.

3.2. Warping factor estimation by distribution matching

As mentioned earlier, the presented method is based on matching the speech distributions of a speaker and the reference model using the Kullback-Leibler divergence. As the distributions of the speech are collected conditional to phoneme HMM states, the matching is done between the Gaussian models of these HMM states. The divergence values of the states are weighted by the states' sample counts in the speaker data and then averaged over all states. The warping factor resulting the minimum of this averaged measure is then selected.

Assuming a Gaussian model for the speech classes we can compute the K-L divergence between one state of a speaker's model and the reference model as in (5), by replacing p with the state dependent frequency warped Gaussian of the speaker and q with the Gaussian of the same state in the reference model. To smooth the covariances of the Gaussians in case of limited data we used a diagonal covariance version of that equation for those classes with fewer than 50 samples in the speaker data.

Although the minimization of the averaged K-L divergence is amenable to several numerical optimization algorithms, a simple grid search was used in the experiments to find the optimal warping factor. As the method only manipulates the statistics this was not an efficiency issue. It should be noted that this same distribution matching could be used to estimate more complex speaker normalizations as well, such as multiparameter SLAPT [6]. This was tested with SPEECON children database, but as it performed slightly worse than the usual VTLN this was not investigated further.

3.3. Jacobian compensation

It has been pointed out [5] that when optimizing a feature transformation under ML framework the Jacobian of the transformation should be taken into account as to avoid introducing any bias to the optimization process. If the transformation is a linear full-rank transformation, its contribution can be handled easily by computing the determinant of the transformation. Otherwise some workarounds are needed, such as modeling the nuisance dimensions (as with HLDA [10]), or using e.g. MMI criterion instead of ML [11].

The need for Jacobian compensation arises from the requirement that the acoustic models have to remain proper probability distributions after the transformation. In K-L based distribution matching compensation is not needed because the Gaussians remain proper probability distributions even after VTLN transformation is applied to the Gaussian statistics.

4. Evaluation

The presented warping factor estimation method was tested with three large vocabulary continuous speech recognition (LVCSR) tasks, two in Finnish and one in English. We used the HUT recognition system that is a HMM/GMM based LVCSR system. The acoustic models used triphones with decision tree tied states. The acoustic features were standard MFCC with delta and double delta features, followed by a global maximum

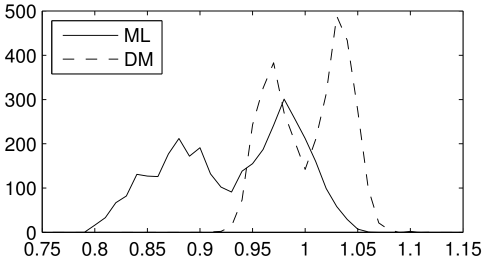

Figure 1: *Histograms of the training set warping factors in the English task for two different models.*

likelihood linear transformation [10]. Although the statistics for warping factor estimation were collected with higher dimensional features, only 12 mel frequency cepstral coefficients were used for the actual acoustic features. With power and deltas the dimension of the acoustic features was then 39.

The frequency warping was implemented as a linear transformation of the MFCC features as presented by McDonough *et al.* [6]. The baseline VTLN estimation was a ML-based grid search using VTLN adapted models. The baseline adapted models were trained by first training an unadapted model after which VTLN estimation and re-training were iterated 6-8 times. For the distribution matching method only the same number of iterations as for the unadapted models were used for training.

Although we used rather complex acoustic models (16-Gaussian mixtures), the VTLN training with ML estimation did succeed without separate target models. This was confirmed by looking at the warping factor histograms of the training set speakers. Figure 1 shows the histogram of the warping factors for the 3507-speaker training set of the English task for both ML and distribution matching (DM) estimation methods. Both methods show the distinctive bimodal distribution, but it is clear that ML has acquired a broader range for the warping factors. For ML the warping factors have also drifted towards the low values. One possible reason for this is the unequal number of males and females in the training set (1516 and 1991, respectively).

For the Finnish tasks the error rates were evaluated using letter error rate, as it suits well for measuring the recognition errors of a highly inflectional and compounding language. For the English task the errors were measured with word error rate.

4.1. Finnish dictation task

The distribution matching method was compared against the maximum likelihood baseline for different amounts of adaptation data in a Finnish dictation task. The task was based on the Finnish SPEECON database [12]. For the training we used 21 hours of clean speech (sample rate 16kHz) from 207 speakers. The acoustic models had 1853 tied states, each modeled with a 16-Gaussian mixture model.

The test set consisted of read sentences from 30 speakers not present in the training set, about 1.5 hours of speech in total in 755 sentences. The adaptation was done using 1, 3 or 5 enrollment sentences with unknown transcriptions, or with all the test data as in two-pass recognition.

The results are shown in Table 1. It can be seen that both the ML and DM methods achieve the best performance after 3 enrollment sentences. The results for one adaptation sentence is slightly worse with the distribution matching method. Using ML estimation with the DM model in two-pass recognition gave the same result as using DM in the test set as well.

Table 1: *Speech recognition results for different amounts of adaptation data, measured in letter error rate. 1-pass is unadapted recognition, Enr(N) is adaptation with N separate enrollment sentences, and 2-pass is adaptation with all the test sentences.*

VTLN	1-pass	Enr(1)	Enr(3)	Enr(5)	2-pass
ML	5.59%	3.52%	3.42%	3.42%	3.41%
DM	5.52%	3.63%	3.42%	3.42%	3.43%

Table 2: *Speech recognition results for children speech, measured in letter error rate.*

VTLN	1-Pass	2-Pass	3-Pass
ML	26.8%	7.82%	7.22%
DM	32.2%	8.74%	7.74%
DM (ML in test)	32.2%	8.24%	7.59%

It should be noted that better unadapted results would be obtained if an unadapted model was used, but in this test all the models were trained using VTLN in the training set. The unadapted error rate here reflects the quality of the segmentation obtained from the first-pass decoding that was used for adaptation.

4.2. Children speech

For more radical adaptation the child speaker portion of SPEECON was used with the models trained from adult speech, i.e. the same models as in the previous task.

The test set consisted of 25 children, aged 9–12 years, each reading 21–35 sentences. The total amount of test set sentences was 725, consisting of about 1 hour of speech.

All the test material was used for adaptation in an unsupervised manner. As the first (unadapted) recognition pass gave rather poor results, the adaptation was re-run using the better segmentations from the second pass. The results of the three passes of recognition are given in Table 2.

The results show that the distribution matching method did not quite achieve the performance of the baseline models. Using ML estimation in the test set with the DM models gave slightly better results, but still 5% worse than the baseline.

4.3. English conversational telephone speech task

Warping factor estimation methods were also compared in English conversational telephone speech task. The models were trained from a 200-hour portion of the Fisher corpus. The speech data had been sampled at 8kHz rate and it was much more noisy than the speech in the Finnish tasks. For the test set, all utterances lasting over 2 seconds from 34 new speakers were used, totaling about 2 hours of speech. The acoustic models had 6094 tied states, again modeled with 16-Gaussian mixture models. The language model for the recognition was trained from the part 2 transcripts of the Fisher corpus, excluding the material in the test set.

Table 3 shows results for several different configurations. Now also the unadapted model was tested to see the amount of improvement from VTLN overall. Adaptation was again unsupervised. In addition to ML and DM models two combinations

Table 3: *Speech recognition results for English CTS task, measured in word error rate. Field "Estimation" refers to VTLN warping factor estimation in the test set.*

Model	Estimation	WER
Unadapted	-	46.7%
ML	ML	44.6%
DM	DM	44.4%
DM	ML	44.4%
DM+ML	ML	44.1%

were tested: DM model with ML estimation in the test set (as in Finnish tasks) and a model where one iteration of ML based VTLN estimation and retraining over the training set was run over the DM model (denoted as DM+ML).

All the VTLN estimation variants gave similar results. The best method was to use DM estimated warping factors from the beginning of the training process and run one iteration of ML-VTLN estimation at the end of the training. Due to rather broad range of VTLN warping factors in the training set the ML model gave rather poor performance for the first recognition pass (58.7% as opposed to 48.4% of the DM model), but using the segmentation from the unadapted model didn't help increase its final performance.

5. Discussion

For the Finnish adult task the DM models showed similar performance as with ML estimated VTLN models. Only the case with a single enrollment sentence showed slight degradation in the performance, but even that was not statistically significant according to the Wilcoxon signed rank test (as implemented in the NIST SCTK). For the children speech task the DM models didn't work quite as well as ML, but again the differences were rather small. The performance difference between the ML and DM models in the 3-pass recognition of the children speech was statistically significant, but the test showed no statistical significance for the performance difference between the DM and ML models when ML criterion was used with the DM models to estimate the test set warping factors.

In the English task, all VTLN models were statistically significantly better then the unadapted model, but between the different VTLN estimation methods the Wilcoxon signed rank test did not show statistically significant differences.

The biggest advantage of the presented method is that it greatly simplifies the training procedure as the VTLN warping factors are available from the very beginning of the training. Thus there is no need to separately train the unadapted models and retrain them with VTLN. For the models used in this paper, this saved about one third of the iterations needed for training the adapted models with ML VTLN estimation. The use of ML VTLN estimation along with DM trained models was also showed to work well, which can be useful if a recognition system with an optimized ML-based VTLN estimation method already exists.

One possible improvement for the distribution matching method is to increase the complexity of the models used for warping factor estimation. The easiest way to do this would be to use context dependent models instead of monophone HMM state models for the speech statistics. This, however, would require a more elaborate solution to the data sparsity problem, such as some kind of clustering.

6. Conclusions

This paper presented a method for estimating warping factors for vocal tract length normalization based on matching the HMM state dependent Gaussian distributions with the reference model using Kullback-Leibler divergence. The method is computationally efficient and it can provide estimates of the VTLN warping factors already early in the training process. This simplifies the training procedure and can e.g. enable the use of VTLN in the estimation of discriminative feature transformations prior to actual model training.

The evaluations reported here show that the method works in a variety of tasks with equal performance as the baseline maximum likelihood estimation method. Also the different combinations of the distribution matching and maximum likelihood methods showed good performance.

7. Acknowledgments

The author is funded by the Graduate School of Language Technology in Finland.

8. References

[1] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proc. ICASSP*, 1996, pp. 339–341.

[2] L. Welling, R. Haeb-Umbach, X. Aubert, and N. Haberland, "A study on speaker normalization using vocal tract normalization and speaker adaptive training," in *Proc. ICASSP*, 1998, pp. 797–800.

[3] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP*, 1996, pp. 353–356.

[4] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. ICASSP*, 1996, pp. 346–348.

[5] M. Pitz, S. Molau, R. Schlüter, and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," in *Proc. Eurospeech*, 2001, pp. 2653–2656.

[6] J. McDonough, T. Schaaf, and A. Waibel, "Speaker adaptation with all-pass transforms," *Speech Communication*, vol. 42, pp. 75–91, 2004.

[7] J. Lööf, H. Ney, and S. Umesh, "VTLN warping factor estimation using accumulation of sufficient statistics," in *Proc. ICASSP*, 2006, pp. 1201–1204.

[8] G. Saon, M. Padmanabhan, and R. Gopinath, "Eliminating inter-speaker variability prior to discriminant transforms," in *Proc. ASRU*, 2001, pp. 73–76.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 2nd edition, 2005.

[10] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on speech and audio processing*, vol. 7, no. 3, pp. 272–281, 1999.

[11] K. Visweswariah and R. Gopinath, "Adaptation of front end parameters in a speech recognizer," in *Proc. Interspeech*, 2004, pp. 1977–1980.

[12] D. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling, "SPEECON - speech databases for consumer devices: Database specification and validation," in *Proc. LREC*, 2002, pp. 329–333.