LINEAR DISCRIMINANT – A NEW CRITERION FOR SPEAKER NORMALIZATION

Martin Westphal, Tanja Schultz, Alex Waibel

Interactive Systems Laboratories
University of Karlsruhe (Germany), Carnegie Mellon University (USA)
{westphal,tanja,waibel}@ira.uka.de

ABSTRACT

In Vocal Tract Length Normalization (VTLN) a linear or nonlinear frequency transformation compensates for different vocal tract lengths. Finding good estimates for the speaker specific warp parameters is a critical issue. Despite good results using the Maximum Likelihood criterion to find parameters for a linear warping, there are concerns using this method. We searched for a new criterion that enhances the inter-class separability in addition to optimizing the distribution of each phonetic class. Using such a criterion, Linear Discriminant Analysis determines a linear transformation in a lower dimensional space. For VTLN, we keep the dimension constant and warp the training samples of each speaker such that the Linear Discriminant is optimized. Although that criterion depends on all training samples of all speakers it can iteratively provide speaker specific warp factors. We discuss how this approach can be applied in speech recognition and present first results on two different recognition tasks.

1 Speaker Normalization using VTLN

Vocal **T**ract **L**ength **N**ormalization (VTLN) has proven to decrease the word error rate of a speech recognition system, compared to systems not using such an approach to reduce the variability introduced by different speakers. The main effect addressed here is a shift of the formant frequencies of the speakers caused by their different vocal tract lengths. Two issues have been investigated. The first is how to map one speaker's spectrum on that of a "standard" or average speaker, depending on a warp parameter which is correlated with the vocal tract length. The other issue is how to find an appropriate warp parameter for each speaker. Most studies assume that the same algorithm is used for training and test, but this is not always necessary.

[Acero (1990)] has used a bilinear transform with one speaker dependent parameter. In a first attempt he observed that the algorithm chose a degenerate case where all input frames are transformed into a constant. Therefore, he enforced a constant average warping parameter over all speakers. Modeling the vocal tract as a uniform tube of length L, the formant frequencies are proportional to 1/L. Therefore, some approaches use a linear warp of the frequency scale to normalize speakers. The warp can be performed in the time or spectral domain. In the latter case, a new spectrum is derived by interpolation or by modifying the Mel frequency filter bank. When the warp is applied in the spectral domain, the problem of mismatching frequency ranges occurs. [Wegmann *et al* (1996)] used a piecewise linear spectral mapping to avoid this problem. They estimated the slope of the transformation function based on a maximum likelihood criterion. [Eide and Gish (1996)] proposed a compromise of different vowel models, namely the uniform tube model and the Helmholtz resonator. They warped the frequency axis *f* of a speaker according to

$$f' = k_S^{\frac{3f}{8000Hz}} f$$

The single warping parameter k_s was estimated using the speaker's formant values and the average formant values of all speakers. [Gouvêa and Stern (1997)] used the first three formants to estimate a linear transformation.

In a previous study [Zhan and Westphal (1997)], we compared the **M**aximum **L**ikelihood method (ML) with the formant based approach and considered different warping functions. The ML method outperformed the formant based approach and was used successfully on a number of speech recognition tasks with the **J**anus **S**peech **R**ecognition **T**ool**k**it (JRTk) [Finke *et al* (1997)]. We use a piecewise linear warping function to interpolate the spectral values as in [Wegmann *et al* (1996)]. Similar to their experiments, it turned out to be important to use only voiced speech samples to calculate the likelihood score. An experiment with different feature streams (warped and not warped) for voiced and unvoiced models showed that the performance is better when using a warped spectrum for all models. To obtain good warp factor estimates with only a very limited amount of test speaker data, we do not use a generic voiced model to calculate the likelihood for the different warps, but the acoustic model of the recognizer. On a German spontaneous speech recognition task (GSST), we achieved similar results for estimating the VTLN parameter on a single utterance (average duration: 7s) versus using all utterances of a speaker.

2 VTLN based on the ML Criterion

This section describes how we use the ML criterion in our system to derive warp factors for each speaker, and motivates a new criterion that will be introduced in this paper.

To obtain a speaker normalized system, we keep a list with one warp factor for each training speaker. The factors are initialized with 1.0, which means no warp. Starting with a speech recognition system without VTLN, we try different warps for each speaker and select the one with the best likelihood on voiced speech samples. These factors are based on a broad distribution of unwarped speech data and can only be a first approximation. After estimating warp factors for each speaker, we perform an EM-update of the acoustic model using the new factors. Thus the model can be iteratively improved.

Despite significant improvements, ML based VTLN has the following drawbacks. First, when applying an iterative warp factor search as described above, we sometimes observed a drift of the average warp factor. Without any cross validation, the feature space keeps shrinking. The samples are mapped such that all coefficients are equal which might optimize the likelihood but gives bad recognition results. A second concern results if using **L**inear **D**iscriminant **A**nalysis (LDA) as the last preprocessing step to create sample vectors with a reduced number of coefficients. LDA selects a sub space that facilitates discrimination of given classes (phonemes or parts of it). Variance within a class, for example caused by different speakers, is minimized. The optimal sub space will certainly be different as soon as a speaker normalization scheme such as ML-VTLN is introduced. When we search for the warp factors, we either do it without LDA or end up with a suboptimal LDA transform. In any case, we have to calculate a new LDA transformation matrix with the new factors and have to train the system again.

The idea underlying VTLN is to normalize the speech signals of different speakers such that it is similar to the speech of a "standard" or average speaker. ML-warp factors can not guarantee such standardization because most recognizers model speech units as Gaussian mixtures. They contain clusters (e.g. male and female speakers), and when a speaker is warped the likelihood might by highest when the samples are warped to the nearest cluster.

We performed an experiment where we used only one Gaussian per class. Thus the warp factors are forced to map all speakers into a single cluster. Another intention was to speed up the system by reducing the computational cost for calculating a number of Gaussians for each class. On the **G**erman **S**pontaneous **S**cheduling **T**ask (GSST), we trained a small context-independent system with ML-VTLN. It had one Gaussian per class and used Mel frequency spectral coefficients without LDA. The drift effect was very strong and the training resulted in degenerated warp parameters which had a good likelihood, but were essentially useless for speech recognition. Based on this experiment, we wanted a method that reduces the variance of the phonetic classes, but does not destroy the structure of the feature space, such that a recognizer is still able to discriminate between classes.

3 VTLN based on the LD Criterion

3.1 The Linear Discriminant Criterion

The **L**inear **D**iscriminant Criterion (LD) is based on the covariance matrices of a given sample set. It is assumed that each sample is assigned to a certain class. For classification purposes it is desirable that all samples of a class build a small scatter around the center of the class. The class centers should be widely spread in the feature space. This can mathematically be expressed by the following equation:

$$LD = \frac{|T|}{|W|}$$

where T is the total covariance matrix of all samples and W is the average within covariance matrix of samples belonging to the same class c_i :

$$W = \sum_i p(c_i) \cdot W_i$$

In **L**inear **D**iscriminant **A**nalysis (LDA) [Fukunaga 1972], this criterion is maximized in a subspace of the original feature space defined by a linear transformation. It is used to derive a $m \times n$ matrix to reduce the n dimensional feature vectors to a dimension $m \leq n$.

3.2 LD Warp Factor Estimation

For speaker normalization we want to find a parameter for each speaker such that the samples of a phonetic class have a smaller variance, under the constraint that different classes can still be discriminated. This is exactly what is measured by the LD criterion. Since we can not optimize the warp parameters of all speakers simultaneously, we have chosen an iterative approach just like in the ML based VTLN method. A set of new warp factors is tried for each speaker separately, while the parameters for the other speakers are kept constant. The warp factor with the best LD value is chosen for the next iteration. Note that this value depends on all other speakers' samples which are warped according to their currently best warping factor. To avoid recalculating the two covariance matrices using all samples of the whole data base, we use the scheme depicted in **Figure 1**.

Our experiments show that the new criterion is a u-shaped function over the warping factor. When using the same simple preprocessing as for the single Gaussian experiment with the same number of classes, the algorithm was able to find good warping parameters which settle after a small number of iterations.

To compare with our standard ML-VTLN approach, we used the same preprocessing and polyphone classes as the recognizer. **Figure 2** shows the average warp factor change between iterations for LD and ML-VTLN. In the first iteration, starting with all factors equal 1, LD-VTLN distributes the warp factors more but then does less changes than ML-VTLN. **Figure 3** shows the LD value for all speakers over the iterations. Since this value depends on the warp factors only, we could also determine it for the ML-VTLN. The value for iteration 0 stands for the system without VTLN which means all warp factors are set to 1. With only one iteration this value could be increased

Given: Samples of all speakers and their phonetic class c_i.

1. Accumulate all samples x_{ij} of a class c_i in a mean accumulator m_i, and all samples in a scatter accumulator *S*. The samples are warped according to the current warp factor of the speaker they belong to.

$$m_i = \sum_j x_{ij}$$

$$S = \sum_{ij} x_{ij} \cdot x_{ij}^T$$

Note that with these two accumulators and the counts for each class, *W* and *T* and therefore *LD* can be calculated.

2. For each speaker:

❍ Warp the samples of the speaker according to the current warp factor and remove their contribution from the accumulators. Keep them as *m_i(speaker)* and *S(speaker).*

For each warp of a set of warp factors within a grid window around the current one:

❍ Warp the samples of the speaker and accumulate it to *m_i(speaker)* and *S(speaker*).

❍ Use these accumulators to calculate *LD(T,W)* for the considered warp factor and speaker.

❍ Pick the warp factor with the best *LD* for that speaker.

3. Proceed with 1 until the average warp factor change falls below a threshold or a maximum number of iteration is reached.

Figure 1: LD warp factor estimation scheme

by a factor of 2.3 by the LD-VTLN training scheme. A similar value was also reached by the ML-VTLN in the 4th iteration.

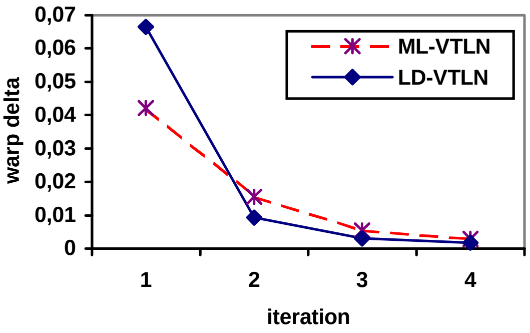

Figure 2: Average warp factor delta between iterations.

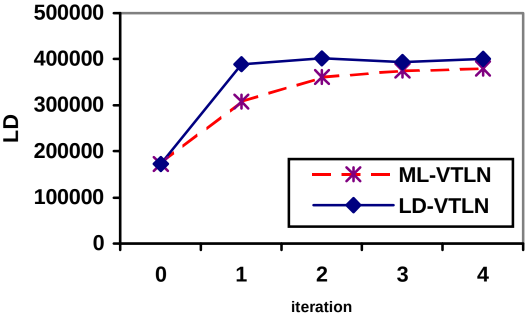

Figure 3: LD values over 4 iterations.

4 Comparison with LDA

Since for the newly proposed LD-VTLN, we use the same criterion as for LDA, we want to discuss the differences and possibilities to combine them.

For LDA the samples are static in a given feature space. It will pick the best “view” in a linear sub space such that the coefficients will be decorrelated and discriminative features will be preserved. When using LD-VTLN the dimension and the feature space are kept constant but the samples of each speaker can be warped until they eventually build easier to discriminate clusters. The matrices *T* and *W* to calculate the LDA transformation matrix are a byproduct of the LD-VTLN and so LDA can be put on top of it at any time. Since the acoustic model of the recognizer is not involved to find the warp parameters as for the ML-VTLN, we could use the feature space before the dimension reduction. The criterion can also be measured in the reduced space for any given dimension, but this requires an additional step to perform the LDA for each speaker and warp factor. For our experiments we therefore used the LD criterion in the original space.

5 Experiments and Results

In this section, we present results using LD-VTLN on two very different speech recognition tasks and compare it with the ML-VTLN. The first database consists of conversational German speech from scheduling dialogs [Finke *et al* (1997)]. The second is a Chinese dictation task from the GlobalPhone project [Schultz and Waibel (1998)]. They provide not only different speaking styles, but also very different language characteristics.

The German Spontaneous Scheduling database (GSST) consists of 1671 speakers with 14000 utterances for training. The compared systems are context dependent and use 2500 clustered polyphone models. The preprocessing is based on 13 Mel cepstral coefficients with first and second order derivatives. After cepstral mean subtraction, LDA is used to reduce the input to 32 dimensional feature vectors. Speaker adapted Viterbi alignments to initialize the recognizers and to assign each sample to a phonetic class as well as the search parameters were taken from a previous ML-VTLN system. A new standard ML-VTLN system was trained over four combined warp/EM iterations with fixed Viterbi alignments (see **Figure 4**). The performance was very similar to previous VTLN systems.

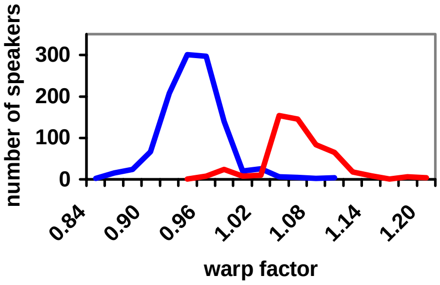

Figure 4: Warp factor distribution for ML-VTLN (GSST, left: males, right: females) after 4 iterations

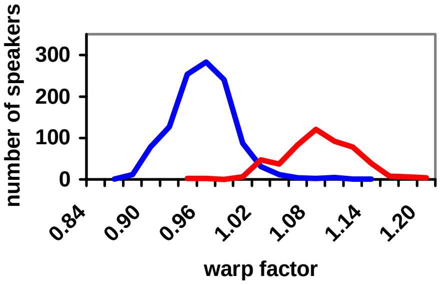

Figure 5: Warp factor distribution for LD-VTLN (GSST, left: males, right: females) after 1 iteration

To train the LD system we took 20 seconds of every speaker to estimate the warp factors. After the first warp iteration (see **Figure 5**) we trained a new system over four iterations with the given Viterbi alignments, keeping the warping factors constant. Both systems were tested using 343 utterances of 70 speakers. The ML-system achieved a word error rate of 15.4%, whereas the LD-system was slightly worse with 15.6%. The performance could not be increased by additional LD-warp iterations.

The Chinese database consists of 77 training speakers with 5124 utterances (150,000 spoken units). For the experiments, we used a context dependent system with 1500 clustered polyphone models. The preprocessing is similar to the German system except for 3 additional coefficients (e.g. zero crossing rate) and a reduction to 24 instead of 32 dimensions. Tested on 149 utterances from 6 different speakers we found that the LD-VTLN results in slightly better error rates in terms of pinyin units.

Table 1 compares the systems' performance for both tasks with and without speaker normalization. It shows that the relative error reduction using VTLN is between 8% and 11%.

Task	No VTLN	ML-VTLN	LD-VTLN
German SST	**16.8%**	**15.4%**	**15.6%**
Chinese Dictation	**20.3%**	**18.4%**	**18.0%**

Table 1: error rates on two speech recognition tasks

6 Conclusion and Future Work

In this paper, we proposed a new criterion for vocal tract length normalization. We showed how it can be applied to estimate a new set of warping parameters without training an acoustic model based on Gaussian mixtures. The derived normalization parameters can be found within only a few iterations and are as good as the one we get from our standard ML-VTLN. Memory requirements for this approach are low since only one matrix and one vector per class are needed as accumulators. The new criterion harmonize better with LDA and is more stable than the ML approach. We think that we could further benefit by using only certain classes for the evaluation of the LD-criterion. As for ML-VTLN it might be better to use only phonetic classes that are affected by different vocal tract lengths.

REFERENCES

1. Fukunaga, K. (1972) “Introduction to Statistical Pattern Recognition”, Academic Press, New York and London.
2. Acero, A. (1990) “Acoustical and Environmental Robustness in Automatic Speech Recognition”, Ph.D. thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh.
3. Wegmann, S.; McAllaster, D.; Orloff, J.; Peskin, B. (1996) “Speaker Normalization on Conversational Telephone Speech”, Proc. ICASSP-96, Vol. 1, pp. 339-341, Atlanta
4. Eide, E.; Gish, H. (1996) “A Parametric Approach to Vocal Tract Length Normalization”, Proc. ICASSP-96, Vol. 1, pp.346-348, Atlanta
5. Lee, L.; Rose R. (1996) “Speaker Normalization using Efficient Frequency Warping Procedures”, Proc. ICASSP-96, Vol. 1, pp. 353-356, Atlanta
6. Finke, M.; Geutner, P.; Hild, H.; Kemp, T.; Ries, K.; Westphal, M. (1997) “The Karlsruhe-Verbmobil Speech Recognition Engine”, Proc. ICASSP-97, Munich
7. Zhan, P.; Westphal, M. (1997) “Speaker Normalization based on Frequency Warping”, Proc. ICASSP-97, Vol. 1, pp.1039-1042, Munich
8. Schultz, T.; Waibel, A. (1998) “Language Independent and Language Adaptive Large Vocabulary Speech Recognition”, Proc. ICSLP-98, Sydney
9. Gouvêa, E.; Stern, R. (1997) “Speaker Normalization through Formant-Based Warping of the Frequency Scale”, Eurospeech-97, Vol. 3, pp. 1139-1142, Rhodes