EXHAUSTIVE SEARCH FOR LOWER-BOUND ERROR-RATES IN VOCAL TRACT LENGTH NORMALIZATION

J.G.A. Dolfing
Philips Research Laboratories
Weißhausstrasse 2
D-52066 Aachen, Germany
Email: Hans.Dolfing@philips.com

ABSTRACT

In the context of large-vocabulary, continuous speech recognition, we address the problem of speaker normalization. In particular, we address the main drawback of many vocal tract length normalization (VTLN) studies and explore the relation between achieved and potential error-rate reduction. In other words, we investigate the correlations between the estimated and optimal warping factors. In addition, we compare the merits of maximum-likelihood VTLN and fast VTLN with the best possible, optimal error-rate for every approach. The experimental results include achieved error-rate reductions of 13.5% and a potential error-rate reduction of about 20%. We show that maximum-likelihood VTLN achieves 90% of the potential, speaker-based, error-rate reduction.

1 INTRODUCTION

In this paper, we study the performance of vocal tract length normalization in the context of a large-vocabulary continuous speech recognition system for a dictation task. VTLN is a speaker-normalization technique and employed routinely in different speech recognition tasks [2, 10, 11, 12]. In contrast to most speaker adaptation techniques, VTLN changes the feature extraction stage instead of the hidden Markov models. The objective of the speaker normalization is to transform the features to better match the current speaker, to produce more compact hidden Markov models, and decrease the word error-rate (WER). It has been observed that the word error-rate reduction due to speaker adaptation and normalization is sometimes additive [7, 8, 10].

Main drawback of many studies on vocal tract length normalization is that the relation between vocal tract length normalization approach and implementation with respect to the error-rate reduction is unclear. Therefore, we investigate recognition without adaptation, compare the effect of two vocal tract length normalization approaches and provide experimental upper-bounds on the word error-rate reduction due to this one parameter speaker-normalization.

In Section 2, we briefly review the properties and alternatives to vocal tract length normalization. Section 3 describes the experimental setup and Section 4 discusses the experimental results. Finally, we draw conclusions in Section 5.

2 VOCAL TRACT LENGTH NORMALIZATION

Vocal tract length normalization [3, 4, 9] is an approach to frequency warping where we transform the frequency kernels with one speaker-specific, parameter \( \alpha \) commonly referred to as scale or warping factor. The objective is to compress or stretch the signal frequency axis to minimize the inter-speaker variations. Loosely speaking, we normalize differences between male and female voices. This approach should lead to better trained speaker-independent acoustic models because some of the speaker variations are explicitly normalized instead of modeled with more mixture densities. The approach in this paper follows [1, 3, 11] for both maximum-likelihood (ML) and fast vocal tract length normalization.

There are numerous alternative approaches to frequency normalization. Alternative estimations of the warping factor \( \alpha \) include [3, 5, 12, 13] and are based on formants, special models, and the Fisher criterion. The ‘all-pass transform’ [6] is a more recent approach to replace the one-parameter, speaker-normalization approach of VTLN with a multi-parameter approach. A linear transformation on the cepstral features is computed. The approach is a generalization of VTLN. The effectiveness on the approach is not compared with VTLN but [8] reports only minor differences compared to standard VTLN.

The allowed range of the warping factors \( \alpha \) depends on the chosen warping approach, i.e., warping of the center frequencies or warping of the original frequency axis. We follow the latter approach, arrange the frequency kernels equidistant on the Mel-scale up to \( f_{N\alpha}^\alpha \) where \( \alpha = 1 \) denotes the ‘average speaker’, and ignore the frequency components beyond the cut-off frequency. A more robust handling is piece-wise linear warping [2].

The principal difficulty in VTLN is: how to compute the best warping factor \( \alpha \) for a given speaker or sentence? We employ two different approaches to compute the best warping factor \( \alpha \) for a speaker \( i \) or sentence \( j \). Given \( N \) warping factors, a transcription \( W^\prime \), a gender-independent (GI) hidden Markov model \( \lambda \) trained with the original, unnormalized features \( X \), and a gender-independent hidden Markov model \( \lambda \) trained with the warped, normalized features \( X^\alpha \). Then we estimate the ‘best’ warping factor \( \hat{\alpha} \) as
● Standard maximum-likelihood VTLN:
\[ \hat{d} = \arg \max \alpha P(X^a | W, \lambda) \]

● Fast VTLN:
\[ \hat{d} = \arg \max \alpha P(X | \lambda_a) \]

Note that we assume that the transcription \( W' \) of the speech data is available to estimate \( \hat{d} \), e.g., from a development set. That is different from the approach in [11] where the transcription to determine the warping factor comes from an initial recognition pass. With VTLN applied in both recognition and training, the recognition then yields the recognized text \( \hat{W} = \arg \max \alpha P(W) | P(X^a | W, \hat{\lambda}) \).

While the advantage of the standard, maximum-likelihood VTLN is that it works reliably [3], its disadvantage is the N-fold feature extraction and an extra set of hidden Markov models for alignment [12]. The advantage of the fast VTLN is that only one feature stream is needed and the alignment with the N hidden Markov models \( \lambda_{a \min}, \ldots, \lambda_{a \max} \) is fast. However, the estimated warping factors produce worse error-rates so far [12, 3].

The achieved error-rate reduction of vocal tract length normalization depends heavily on many details, like the exclusion of the likelihood of silence frames or the correction of the likelihood with the Jacobian of the VTLN transformation. Although vocal tract length normalization employs only one free parameter for the speaker normalization, word error-rate reductions of more than 10% relative have been achieved [7].

3 EXPERIMENTAL SETUP

The investigated large-vocabulary recognition task uses an internal dictation database which contains eight speakers, four males and four females as testing material. This test set is clean office data, contains one hour of dictation material consisting of 99 sentences with an OOV rate of 1.7%. In addition, a development set contains 20 spoken sentences per speaker with known transcript. The main purpose of this development set is to have an independent data set to determine the warping factor for each speaker. One minute material of the development set is used to determine a warping factor for a speaker. We used 170 hours of training material which includes WSJ material and additional, disjunct dictation material.

The recognizer is a large-vocabulary continuous speech recognition (LVCSR) system based on hidden Markov models which employs gender-independent models with Laplacian densities, triphone models, and linear discriminant analysis (LDA). The recognizer is described in more detail in [1]. The statistical significance of the word error-rates at a 95% level is \( \approx 0.3\% \).

Given a speaker \( i \), a sentence \( j \), and a warping factor \( \alpha \) then \( e_{ij} \) denotes the number of errors, i.e., substitutions, insertions and deletions, while \( n_{ij} \) denotes the number of spoken words of this particular sentence. While \( \hat{\alpha} \) denotes the best, estimated warping factor with ML or fast VTLN, we use \( \alpha' \) to denote the optimal warping factor with lowest error-rate.

For speaker-based vocal tract length normalization, we estimate a warping factor \( \alpha \) and compute an error-rate for speaker \( i \) as

\[ \text{wer}_a = 100 \cdot \frac{\sum e_{ij} / \sum n_{ij}}{\text{wer}_a} \]

where \( \text{wer}_a \) is the word error-rate for speaker \( i \) at the best warping factor \( \alpha' \) and is \( \text{wer}_a = \min \alpha \text{wer}_a \) with \( \alpha'_a = \arg \min \alpha \text{wer}_a \).

For sentence-based vocal tract length normalization, we estimate a warping factor \( \alpha \) for every sentence \( j \) of speaker \( i \) and compute a best possible error-rate for speaker \( i \) as \( \text{wer}_a = 100 \cdot \frac{\sum (\min \alpha e_{ij}) / \sum n_{ij}}{\text{wer}_a} \).

Given eight speakers \( i \), 99 sentences \( j \), and 16 warping factors \( \alpha \). Then our approach is: \( \forall i \forall j \forall \alpha \) compute \( e_{ij} \). Therefore, we can compare the achieved and best possible error-rate reduction due to vocal tract length normalization. We use warping factors mostly equidistant in 0.02 steps between 0.78 to 1.20.

4 RESULTS

The following subsections discuss a series of experiments to compare several vocal tract length normalization techniques. While the initial experiments are speaker-based, we also investigate the benefits of a sentence-based approach. First, we investigate the merits of the standard, ML-based approach in Subsection 4.1. This is followed by a comparison with the lower bound on the word error-rate due to the single parameter for vocal tract length normalization in Subsection 4.2. Finally, we investigate the error-rate reductions due to sentence-based and fast vocal tract length normalization in Subsections 4.3 and 4.4.

4.1 Maximum-likelihood vocal tract length normalization

The maximum-likelihood vocal tract length normalization computes a warping factor \( \alpha \) for every speaker [1, 11]. The warping factor is found by searching the range of warping factors as indicated in Section 3. We use a two step approach. First, the search for the correct warping factor is based on a subset of one minute speech from the additional data set where we align the known transcript with the speech data, and select the warping factor with best likelihood. The alignment uses single-density, triphone models trained on the original features [12]. The likelihood of silence frames is excluded from the total likelihood. The search for the best warping factor is done with a linear scan over all factors. Alternatively, we employ a binary search for the best warping factor which exploits the parabolic shape of the likelihoods over the warping factors.

Second, we employ the determined warping factors for every speaker and conduct an experiment with combinations of vocal tract length normalization in training and recognition. The pruning threshold and all other recognition parameters are fixed. The baseline word error-rate is 23.43%. The comparison of the effect of vocal tract length normalization in training and recognition is given in Table 1.

As expected, the combination of vocal tract length normalization in both recognition and training is most beneficial. The use of vocal tract length normalization only in training is counterproductive as observed in [11]. To put the word error-rate reduction of 13.5% in perspective, remember that we use a GL model with few densities.
Table 1: Comparison of error-rates in [%] of vocal tract length normalization in training and recognition with the baseline without vocal tract length normalization. In brackets are the relative reductions in [%] with respect to the baseline error-rate.

<table>
<thead>
<tr>
<th>VTLN in training</th>
<th>VTLN in recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>23.43 (-)</td>
<td>22.06 (- 5.8%)</td>
</tr>
<tr>
<td>25.27 (+7.8%)</td>
<td>20.26 (-13.5%)</td>
</tr>
</tbody>
</table>

4.2 Word error-rate lower bounds for vocal tract length normalization

After the computation of the baseline with and without vocal tract length normalization, we determine a lower-bound on the word error-rate due to vocal tract length normalization. Because we have determined for every speaker the warping factor $\alpha^*$ which generates the lowest word error-rate, it is easy to judge the performance of the implemented vocal tract length normalization. In Table 2, we see that vocal tract length normalization in training and recognition with the best possible warping factor gives a total reduction of 14.3% in word error-rate. Naturally, the optimal reduction is more than the achieved reduction of 13.5% in Table 1.

Table 2: Comparison of vocal tract length normalization error-rates in [%]. We compare the empirical lower bound of ‘Optimal’ VTLN to the error-rate based on the estimated warping factors with ML VTLN. In brackets are the relative reductions in [%] with respect to the baseline error-rate.

<table>
<thead>
<tr>
<th>VTLN in training</th>
<th>ML VTLN $\alpha$</th>
<th>Optimal VTLN $\alpha^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>22.06 (-5.8%)</td>
<td>21.89 (-6.6%)</td>
</tr>
<tr>
<td>yes</td>
<td>20.26 (-13.5%)</td>
<td>20.07 (-14.3%)</td>
</tr>
</tbody>
</table>

In contrast to most other studies on vocal tract length normalization, we can now compare the benefits of the estimated and the optimal vocal tract length normalization warping factors. This leads to the most interesting observation in Table 2 where we show that the implemented, speaker-based vocal tract length normalization ($\alpha$) gives us 88% and 93% of the best, possible WER reduction (with $\alpha^*$) for vocal tract length normalization in recognition only and in training and recognition, respectively.

Additionally, we computed the correlations between the estimated and optimal warping factors. For vocal tract length normalization only in recognition, the correlation is $r = 0.96$ while vocal tract length normalization in both training and recognition yields $r = 0.98$. This corresponds with the fact that the latter case yields an error-rate reduction closer to the lower-bound error rate.

4.3 Sentence-based vocal tract length normalization

In this experiment, we determined for every sentence in the test set the warping factor $\alpha^*_j$ which yields the lowest error-rate. We compare this sentence-based, lower bound error-rate in Table 3 with the speaker-based, lower bound in the previous subsection. Note that this approach increases the degrees of freedom for vocal tract length normalization which should result in a better normalization and higher error-rate reduction, provided we can estimate the warping factors $\alpha^*_j$ correctly.

Table 3: Comparison of optimal vocal tract length normalization error-rates in [%]. We compare the lower-bound error-rates for speaker-based and sentence-based vocal tract length normalization.

<table>
<thead>
<tr>
<th>VTLN in training</th>
<th>VTLN WER lower-bounds.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker-based</td>
<td>Sentence-based</td>
</tr>
<tr>
<td>no</td>
<td>21.89 (-6.6%)</td>
</tr>
<tr>
<td>yes</td>
<td>20.07 (-14.3%)</td>
</tr>
</tbody>
</table>

In Table 3, we show that the error-rate reductions due to vocal tract length normalization are potentially much larger if we would compute a warping factor for every sentence of every speaker instead of only once per speaker. For vocal tract length normalization only in recognition, we see in Table 3 that the potential of sentence-based vocal tract length normalization is more than twice the error-rate reduction than speaker-based VTLN. With vocal tract length normalization in both training and recognition, the sentence-based approach yields another significant error-rate reduction from $-14.0\%$ to $-19.7\%$.

In addition, we computed the correlations of all optimal, sentence-based warping factors $\alpha^*_j$, with the optimal, speaker-based warping factors $\alpha^*$. If we apply vocal tract length normalization only in recognition then the correlation is $r = 0.74$. With vocal tract length normalization in both training and recognition, we find a correlation of $r = 0.92$.

4.4 Fast vocal tract length normalization

In this subsection, we investigated whether the fast, speaker-based vocal tract length normalization [3, 12] is a viable alternative to the standard, maximum-likelihood vocal tract length normalization. Loosely speaking, do we get a similar error-rate reduction without one feature extraction per warping factor? In the experiments by [3, 12], the error-rate after fast vocal tract length normalization was approximately 30% relative worse compared to maximum-likelihood vocal tract length normalization.

We concentrate on an experiment with vocal tract length normalization only in recognition and train the fast models with the sentences of the test data based on the optimal labels $\alpha^*_j$ from Subsection 4.3. At the cost of the strict separation between training and test data, this will give us insight whether we can reach the error-rate reduction of ML vocal tract length normalization at all. Other experimental parameters include LDA, whether or not to exclude the likelihood from the silence frames, and the topology of the fast model. The two topologies are a hidden Markov model with one state (Fast1) and a hidden Markov model with six states (Fast2), both with a maximum of 32 densities per state. The Fast2 model is an attempt to catch some temporal information in the fast VTLN models. The results are summarized in Table 4.
We learn from Table 4 that the exclusion of the likelihood of silence frames plus the use of LDA significantly reduces the error-rate for model Fast1 but not significant for Fast2. In general, the correspondence of the estimated warping factors with the optimal warping factors is better for Fast1.

It was observed that the parabolic shape of likelihoods plotted over warping factors is more noisy compared to Subsection 4.1. However, parabolic smoothing with a generalized linear model did not improve the estimations.

The best obtained error-rate with fast vocal tract length normalization is 22.03% which is almost the same as 22.06% which we obtained with maximum-likelihood vocal tract length normalization. Because of the way the fast models were trained, this does not allow further conclusions before the experiment is repeated with more, separate training data. We can only observe that it is possible in principle to estimate warping factors as well as with maximum-likelihood vocal tract length normalization.

5 CONCLUSION

In this paper, we showed that the speaker-based error-rate reduction due to standard, maximum-likelihood vocal tract length normalization is about 10% relative smaller than the best possible error-rate reduction. Second, we investigated sentence-based vocal tract length normalization, computed lower bounds on the word error-rate, and showed that there is 30% more potential error-rate reduction in a system where vocal tract length normalization is applied in both training and recognition. We conclude that the potential error-rate reductions of vocal tract length normalization are not fully exploited in both sentence-based and speaker-based setups. Finally, we showed that it is possible, in principle, for fast vocal tract length normalization to achieve the same error-rate reduction as with maximum-likelihood vocal tract length normalization. However, more experiments are needed for firm conclusions.

6 REFERENCES