IMPROVED METHODS FOR VOCAL TRACT NORMALIZATION

L. Welling, S. Kanthak and H. Ney

RWTH Aachen – University of Technology, D-52056 Aachen, Germany
{welling, kanthak, ney}@informatik.rwth-aachen.de

ABSTRACT

This paper presents improved methods for vocal tract normalization (VTN) along with experimental tests on three databases.

We propose a new method for VTN in training: By using acoustic models with single Gaussian densities per state for selecting the normalization scales it is avoided that the models learn the normalization scales of the training speakers. We show that using single Gaussian densities for selecting the normalization scales in training results in lower error rates than using mixture densities.

For VTN in recognition, we propose an improvement of the well–known multiple–pass strategy: By using an unnormalized acoustic model for the first recognition pass instead of a normalized model lower error rates are obtained. In recognition tests, this method is compared with a fast variant of VTN.

The multiple–pass strategy is an efficient method but it is suboptimal because the normalization scale and the word sequence are determined sequentially. We found that for telephone digit string recognition this suboptimality reduces the VTN gain in recognition performance by 30% relative.

On the German spontaneous scheduling task Verbmobil, the WSJ task and the German telephone digit string corpus SieTill the proposed methods for VTN reduce the error rates significantly.

1. INTRODUCTION

This paper deals with improved methods for vocal tract normalization (VTN) [2, 3, 5, 6]. In [8], we presented a new method for VTN in training. Similar to the methods described in [2, 6], our approach avoids the problem that the acoustic models used for selecting the normalization scales in training learn the normalization scales (in the following called scales) of the training speakers. However, compared to the methods in [2, 6] our approach is conceptually simpler: The intermediate acoustic models that we use for scale selection are trained on the full training corpus and consist of only a small number of Gaussian densities per state. Thus the intermediate models do not learn the scales of the training speakers. This paper extends our work in [8] and presents the following novel contributions:

- *VTN in training.* The effect of the acoustic model for scale selection in training on the resulting word error rates is demonstrated. We show that using single Gaussian densities per generalized triphone state for scale selection in training results in better recognition rates than using mixture densities for scale selection.

- *Improved multiple–pass strategy.* An improved multiple–pass strategy [2] for VTN in recognition is presented: By using an unnormalized acoustic model for the first recognition pass instead of a normalized model lower error rates are obtained.

- *Fast VTN.* A method for fast scale selection in recognition is described in detail. The method is similar to the method in [6] but the training of the mixture model for normalized acoustic vectors is simpler with our approach.

- *Experimental comparison.* The improved multiple–pass strategy and the fast VTN are compared in recognition tests.

- *Suboptimality of multiple–pass strategy.* The multiple–pass strategy is an efficient method but it is suboptimal because the scale and the word sequence are determined sequentially. We found that for telephone digit string recognition this suboptimality reduces the VTN gain in recognition performance by 30% relative.

- *Results on WSJ, Verbmobil and SieTill.* We present recognition tests on three different databases, namely the German spontaneous scheduling task Verbmobil [1], the Wall Street Journal task and the German telephone digit string corpus SieTill. The results show that the proposed method for VTN in training in combination with the improved multiple–pass strategy lead to consistent reductions in the word error rates.

This work was partly funded by the German Ministry of Science and Technology (BMBF) in the framework of the VERBMOBIL project under grant 01 IV 701 T4. The responsibility for the contents of this study lies with the authors.

2. EXPERIMENTAL CONDITIONS

For all experiments in this work, we used the recognizer described in [4, 7] and the following set–up:

A piecewise–linear frequency normalization is used [6]. The selection of the scale α is done on speech excluding silence using an exhaustive line search for the α in the range $0.88 \leq \alpha \leq 1.12$ with step size 0.02. A scale is estimated for each sentence in recognition and for each speaker in training. We used gender–independent acoustic models.

The following databases were used: For the WSJ0 experiments, recognition was done on the Nov.'92 development and evaluation sets (18 speakers, 740 sentences) and training on the WSJ0 84–speaker corpus. For the Verbmobil task, testing was done on the 1996 evaluation data (62 speakers, 343 sentences) and training on the 1996 training corpus (568 speakers). The SieTill corpus is a German telephone digit string database. It consists of 362 training speakers (42860 digits) and 356 testing speakers (43095 digits) representing a large variety of line and speaker characteristics.

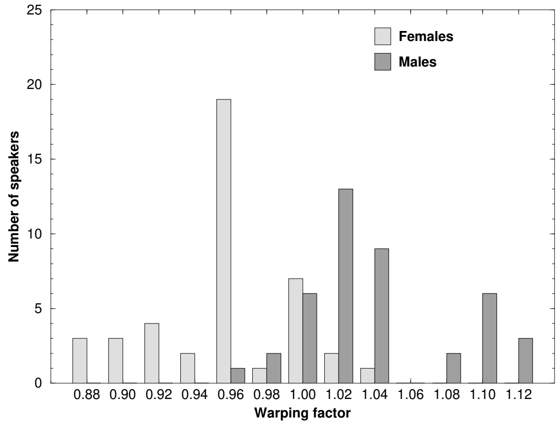

Figure 1: Histogram of scales for WSJ0 84–speaker corpus.

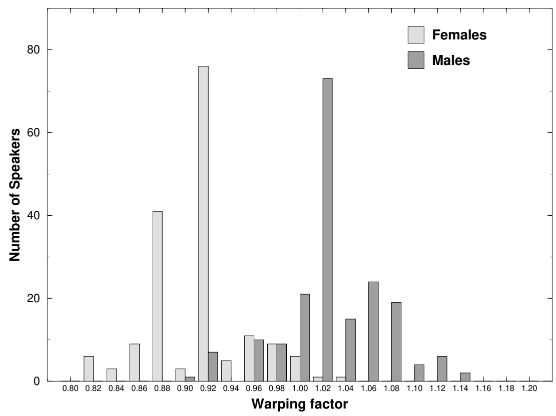

Figure 2: Histogram of scales for SieTill training speakers.

3. VTN IN TRAINING

For the training of a normalized acoustic model using VTN, we propose the following three–step procedure:

1. An intermediate acoustic model λ consisting of a single Gaussian density per generalized triphone state is estimated from the unnormalized acoustic vectors of all training speakers by maximum likelihood training [4].
2. For each training speaker r, a scale α_r is chosen as the scale for which the training data of this speaker, X_r^α, achieve the greatest likelihood, given the transcriptions W_r and the single density model λ:
$$\alpha_r = \arg\max_\alpha Pr(X_r^\alpha | W_r, \lambda).$$
3. A normalized model $\overline{\Lambda}$ is trained on the normalized acoustic vectors by maximum likelihood training.

This method is conceptually simple. However, the choice of the intermediate acoustic model for scale selection is critical: As can be seen from Table 1, an intermediate acoustic model with too high resolution can learn the scales of the training speakers. The table contains the number of Gaussian densities per state for scale selection along with the corresponding recognition results on the WSJ0 data obtained with the baseline multiple–pass strategy as explained in Section 4. Table 1 shows that applying VTN only in recognition gives a word error rate of 6.8%. Using VTN also in training with 32 Gaussians per state for scale selection does not lead to an improvement since scales of 1.00 are produced for most training speakers. Using eight Gaussians per state, the error rate is reduced down to 6.1%. The lowest word error rate of 5.9% is obtained with only one single Gaussian density per state for the selection of the scales in training.

Table 1: Effect of the number of densities per state for scale selection in training on the word error rate (WSJ0, bigram language model (lm) with $PP = 107$).

VTN	#Dens./State scale selection	#Dens. recognition	DEL – INS [%]	WER [%]
rec. only	-	103k	1.4 – 0.6	6.8
train.+rec.	32	$\alpha_r \approx 1.00$		
	8	143k	1.2 – 0.6	6.1
	1	140k	1.2 – 0.6	5.9

Histograms of the obtained scales using single Gaussian densities are shown in Figure 1 for the WSJ0 and in Figure 2 for the SieTill corpus. In Figure 2, the range for the scale α was extended to $0.80 \leq \alpha \leq 1.20$ in order to verify that the baseline set–up with $0.88 \leq \alpha \leq 1.12$ is adequate. According to Figure 2, only approx. five percent of the training speakers get a scale lower than 0.88 or higher than 1.12. Both histograms show that the distribution of the chosen scales is approximately bimodal with one mode per gender.

4. VTN IN RECOGNITION

This section deals with different strategies for VTN in recognition. First, we shortly summarize the baseline multiple–pass strategy [2]. Then, we propose an improved multiple–pass approach and a method for fast VTN. These three recognition strategies will then be compared in recognition tests.

In [2], the multiple–pass strategy has been introduced:

1. A recognition pass with unnormalized acoustic vectors X and a normalized acoustic model $\overline{\Lambda}$ produces a preliminary transcription W':
$$W' = \arg\max_W Pr(W) Pr(X | W, \overline{\Lambda})$$
2. The scale $\hat{\alpha}$ is selected according to
$$\hat{\alpha} = \arg\max_\alpha Pr(X^\alpha | W', \overline{\Lambda}) \quad .$$
3. A second recognition pass with normalized features $X^{\hat{\alpha}}$ and a normalized acoustic model $\overline{\Lambda}$ gives the final transcription $\hat{W}$:
$$\hat{W} = \arg\max_W Pr(W) Pr(X^{\hat{\alpha}} | W, \overline{\Lambda})$$

Table 2: Comparison of error rates for recognition using an unnormalized Λ versus a normalized acoustic model $\overline{\Lambda}$. WSJ0: bigram lm with $PP = 107$; Verbmobil: trigram lm with $PP = 35$.

Corpus	Model normalization	#Dens.	DEL – INS [%]	WER [%]
WSJ0	no	103k	1.4 – 0.6	6.8
	yes	140k	1.3 – 0.8	7.5
Verbmobil	no	195k	3.2 – 3.1	16.7
	yes	185k	3.7 – 2.9	18.7
SieTill	no	358	1.0 – 0.8	4.6
	yes	358	1.0 – 0.9	5.3

4.1. IMPROVED MULTIPLE–PASS STRATEGY

The multiple–pass strategy as described above uses a normalized model in both recognition passes. However, the number of word errors in the preliminary transcription W' can be reduced, if an unnormalized model is used in the first pass. As Table 2 shows, the error rates on the WSJ0, Verbmobil and SieTill databases using unnormalized acoustic vectors are consistently better with an unnormalized acoustic model Λ than with a normalized model $\overline{\Lambda}$.

Therefore, we propose to improve the multiple–pass strategy by using an unnormalized model Λ instead of a normalized model $\overline{\Lambda}$ in the first recognition pass. Experimental results will be given in Section 4.3.

4.2. FAST VTN

In the following, we describe a fast method for scale selection in recognition which does not require a preliminary transcription [8]. Similar to the approach in [6], the method is based on a Gaussian mixture model that represents the distribution of the normalized feature vectors.

After the training data have been normalized as explained in Section 3, a Gaussian mixture model M is trained on the normalized acoustic vectors by employing the LBG algorithm and the maximum likelihood criterion. During recognition, the scale is selected using the Gaussian mixture model M:

1. The scale $\hat{\alpha}$ is selected according to

$$\hat{\alpha} = \arg\max_{\alpha} Pr(X^{\alpha}|M) \quad .$$

   A maximum approximation is used to compute $Pr(X^{\alpha}|M)$: For each acoustic vector $X^{\alpha}(t)$ at time t, the sum over the component densities of M is replaced by the maximum.

2. A recognition pass with normalized features $X^{\hat{\alpha}}$ and the normalized acoustic model $\overline{\Lambda}$ gives the transcription $\hat{W}$:

$$\hat{W} = \arg\max_{W} Pr(W)Pr(X^{\hat{\alpha}}|W, \overline{\Lambda})$$

In our tests, the Gaussian mixture model had a single diagonal covariance matrix and 64 component densities. The fast scale selection was done on speech excluding silence: For each sentence, we compute the component density m_{sil} of the mixture model M which is selected most often according to the maximum approximation. Acoustic vectors $X(t)$ attributed to m_{sil} are not used for scale selection. Figure 3 illustrates this fast scale selection method.

Table 3: Effect of different strategies for VTN in recognition on the error rate on WSJ0 (bigram lm with $PP = 107$) and SieTill.

Corpus	VTN	#Dens.	DEL – INS [%]	WER [%]
WSJ0	no	103k	1.4 – 0.6	6.8
	fast	143k	1.3 – 0.6	6.2
	multiple–pass:			
	baseline	140k	1.2 – 0.5	5.9
	improved	103k/140k	1.2 – 0.5	5.7
SieTill	no	358	1.0 – 0.8	4.6
	multiple–pass:			
	baseline	358	1.0 – 0.8	4.0
	improved	358	1.0 – 0.8	3.9

4.3. EXPERIMENTAL COMPARISON

To compare the different strategies for VTN in recognition, we carried out recognition tests on the WSJ0 and SieTill databases.

Table 3 summarizes the recognition results. Without VTN in training and recognition, a word error rate of 6.8% is obtained on the WSJ0 testing data. The fast VTN leads to a significant reduction from 6.8% to 6.2%. However, the baseline multiple–pass strategy described in Section 4 clearly outperforms the fast method and gives an error rate of 5.9%. The improved multiple–pass strategy presented in Section 4.1 leads to a further reduction in error rate down to 5.7%.

The results on the SieTill database are similar: Using no VTN, a word error rate of 4.6% is obtained. The baseline multiple–pass strategy reduces the error rate down to 4.0%. Again, the improved multiple–pass strategy outperforms the baseline strategy and gives a word error rate of 3.9%. The fast VTN has not been tested on the SieTill data since the fast method is not necessary for digit string recognition.

4.4. SUPOPTIMALITY OF MULTIPLE–PASS STRATEGY

From the viewpoint of the Bayes' decision rule, the unknown word sequence and the scale should be determined according to the following criterion:

$$\hat{W} = \arg\max_{W,\, \alpha} Pr(W)Pr(X^{\alpha}|W, \overline{\Lambda}) \; . \tag{1}$$

The previously described multiple–pass strategies are suboptimal methods for maximizing Equation 1 since the scale and the

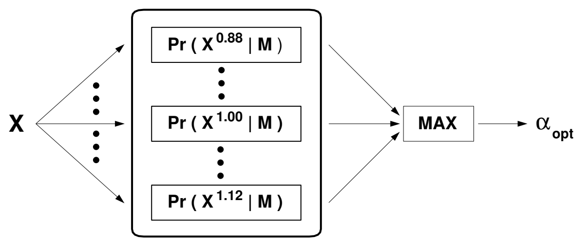

Figure 3: Fast scale selection in recognition.

Table 4: Effect of approximations for VTN in recognition on the word error rate on WSJ0 (bigram lm with $PP = 107$) and SieTill.

Corpus	VTN	#Dens.	DEL – INS [%]	WER [%]
WSJ0	improved mult.-pass	103k/140k	1.2 – 0.5	5.7
	correct transcription	140k	1.1 – 0.5	5.7
SieTill	improved mult.-pass	358	1.0 – 0.8	3.9
	correct transcription	358	1.0 – 0.7	3.6
	full optimization	358	1.0 – 0.7	3.6

word sequence are determined sequentially and not in combination. In this section, we show that for telephone digit string recognition this suboptimality of the multiple–pass reduces the gain in recognition performance due to VTN significantly.

Table 4 compares the error rates on WSJ0 and SieTill for the improved multiple–pass strategy and for scale selection based on the correct transcription of the test sentences instead of the preliminary transcription W'. As Table 4 shows, an error rate of 5.7% results for both methods on the WSJ0 data. However, on the SieTill data the error rates differ significantly: An error rate of 3.9% is obtained with the improved multiple–pass strategy compared to 3.6% with the correct transcription.

A full optimization over the scale and the word sequence can be implemented by a separate recognition pass for each scale. The combination of word sequence and scale which maximizes Equation 1 is selected. This method leads to an error rate of 3.6% on the SieTill data, as shown in Table 4, instead of 3.9% with the improved multiple–pass strategy and 4.6% with no VTN (see Table 3). Thus we observed that for telephone digit string recognition the suboptimal multiple–pass strategy reduces the VTN gain in recognition performance by 30% relative. In our view, this effect results from the short average length of the SieTill testing utterances of only 3.2 seconds (3.3 digits) compared to 7.0 seconds for the WSJ0 task.

5. RESULTS ON WSJ, VERBMOBIL AND SIETILL

In this section, we summarize the improvements in word error rate due to the presented methods for VTN on three different corpora. For VTN in training, we used the method presented in Section 3. For VTN in recognition, we employed the improved multiple–pass strategie as explained in Section 4.1 on the WSJ0 and Verbmobil tasks and the full optimization as explained in Section 4.4 on SieTill. As Table 5 shows, we get a significant reduction in word error rate by 16% relative on the WSJ0 data, 5% relative on Verbmobil and 22% relative on SieTill.

Table 5: Effect of VTN in training and recognition on the word error rate (WSJ0 and Verbmobil: trigram lm).

Corpus	VTN	PP	#Dens.	DEL – INS [%]	WER [%]
WSJ0	no	56	103k	0.8 – 0.5	4.9
	yes		103k/140k	0.7 – 0.5	4.1
Verb–mobil	no	35	195k	3.2 – 3.1	16.7
	yes		195k/185k	3.1 – 3.1	15.9
SieTill	no	11	358	1.0 – 0.8	4.6
	yes		358	1.0 – 0.7	3.6

6. SUMMARY

The main contributions of this paper are:

- Using single Gaussian densities for scale selection in training leads to distributions of scales that reflect typical variations of vocal tract lengths among speakers. Based on this method, we presented a three–step procedure for the training of a normalized acoustic model.
- We reduced the error rates obtained with the multiple–pass strategy by using an unnormalized instead of a normalized model in the first recognition pass.
- We compared a fast method for VTN with the improved multiple–pass strategy. The fast method reduces the error rate on the WSJ0 task from 6.8% to 6.2% compared to 5.7% obtained with the improved multiple–pass strategy.
- We found that for telephone digit string recognition the suboptimality of the multiple–pass strategy reduces the VTN gain in recognition performance by 30% relative.

By using the presented methods for VTN the word error rates were reduced from 4.9% to 4.1% on the WSJ0 task, from 16.7% to 15.9% on the Verbmobil task and from 4.6% to 3.6% on the SieTill corpus.

7. REFERENCES

[1] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, M. Westphal, "The Karlsruhe–Verbmobil speech recognition engine," in *Proc. ICASSP*, Vol. 1, pp. 83-86, Munich, Germany, Apr. 1996.

[2] L. Lee, R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP* Vol. 1, pp. 353-356, Atlanta, GA, May 1996.

[3] L. Lee, R. C. Rose, "A frequency warping approach to speaker normalization," in *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 1, pp. 49-60, Jan. 1998.

[4] H. Ney, L. Welling, S. Ortmanns, K. Beulen, F. Wessel, "The RWTH large vocabulary continuous speech recognition system," in *Proc. ICASSP*, Vol. 2, pp. 853-856, Seattle, WA, May 1998.

[5] D. Pye, P. C. Woodland, "Experiments in speaker normalisation and adaptation for large vocabulary speech recognition," in *Proc. ICASSP*, Vol. 2, pp. 1047-1051, Munich, Germany, Apr. 1997.

[6] S. Wegmann, D. McAllaster, J. Orloff, B. Peskin, "Speaker normalization on conversational telephone speech," *Proc. ICASSP*, Vol. 1, pp. 339-341, Atlanta, GA, May 1996.

[7] L. Welling, N. Haberland, H. Ney, "Acoustic front-end optimization for large vocabulary speech recognition", *Proc. EUROSPEECH*, Vol. 4, pp. 2099-2102, Rhodes, Greece, Sep. 1997.

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