

Efficient Vocal Tract Normalization in Automatic Speech Recognition

Sirko Molau, Stephan Kanthak, Hermann Ney

*Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen - University of Technology, D-52056 Aachen, Germany
{molau,kanthak,ney}@informatik.rwth-aachen.de*

Abstract

In this paper we study the effect of vocal tract normalization (VTN) on the word error rate (WER) in speaker independent large vocabulary speech recognition. Evaluation test results are reported for the German VerbMobil II (VM II) and the English Wall Street Journal (WSJ) corpus. In particular, we analyse:

- the effect of the type of warping function (linear vs. non-linear) on the WER;
- different methods for estimating the warping factor in recognition;
- incremental warping factor estimation for single-pass online recognition;
- phoneme dependence of the warping factors.

We find that a simple piecewise linear warping function performs better than non-linear frequency warping. In recognition, a two-pass approach performs as good as supervised VTN on the reference transcription even if the WER of the first recognition pass is of the order of 20..30%. Fast warping factor estimation with text independent models results in only a slight performance degradation but allows the system to run at the same speed as a single-pass recognizer without VTN. A minor improvement over baseline VTN is obtained with phoneme dependent warping factors.

1. Introduction

VTN is a normalization scheme which is typically applied in speaker independent recognition. The main idea is to eliminate variations in the speech signal caused by different lengths of the speakers' vocal tracts. This is achieved by warping the frequency axis of the speech spectrum during signal analysis.

The idea of VTN is not new. A number of authors reported on the successful application of vocal tract normalization on small and large vocabulary recognition tasks (i.e. Eide&Gish 96, Lee&Rose 96, Wegmann et al. 96, Welling et al. 98). Still, the implementation is not straightforward. The performance and computational costs of VTN significantly depend on algorithmic decisions and a number of different parameters which will be discussed in more detail here.

We will report recognition test results for the German VerbMobil II corpus (spontaneous speech, 10k vocabulary, scheduling task) and the English Wall Street Journal corpus (clean read speech, 20k vocabulary, newspaper texts). Their statistics are given in Table 1. We used the RWTH large vocabulary speech recognition system, which has been described in detail in Ney (98). The recognizer is a continuous Gaussian mixture density speech recognition system with a trigram Viterbi decoder. We use 3-state left-to-right HMM within-word models, a globally pooled variance vector, phonetically tied triphone models, and a linear discriminant analysis.

Recognition results are presented in terms of search space (average number of states, arcs, and trees) and recognition accuracy (deletions, insertions, and overall word error rate).

Table 1: Training and test corpora

Corpus	VerbMobil II		Wall Street Journal	
	Training	Test	Training	Test
Name	CD1-41	DEV-99	WSJ0+1	NAB 20k DEV-94 H1
Acoustic Data	61.5h	1.6h	81.4h	0.8h
Silence Portion	13%	11%	27%	19%
# Speakers	857	16	284	20
# Sentences	36,015	1,081	37,571	310
# Running Words	701,512	14,662	649,624	7,378
Perplexity (Trigram LM)	-	62.0	-	126.6

The remainder of the paper is organized as follows: In section 2 the baseline VTN training and two-pass recognition approach is described. In section 3 we examine the effect of the warping function on the performance of VTN. The following two sections discuss alternative approaches for estimating the warping factor in recognition, and in section 6 the effect of phoneme dependent warping factors is analysed. The paper is concluded in section 7.

2. Baseline VTN

In speaker-adaptive training, the warping factor for each training speaker is required. Suppose λ is an acoustic model, X_i^α is a set of acoustic vectors obtained by applying the warping factor α to all utterances of speaker i , and W_i is the corresponding transcription. Then, the optimal warping factor $\hat{\alpha}_i$ is obtained by maximizing the text dependent probability Pr :

$$\hat{\alpha}_i = \arg \max_{\alpha} Pr(X_i^\alpha | \lambda, W_i).$$

It was shown by Welling et al. (98) that λ is preferably a low resolution acoustic model (i.e. single densities), since mixture density models may have “learned” already different warping factors and therefore do not discriminate well between them anymore. Once the warping factors of all training speakers are determined, the training data are normalized accordingly, and normalized acoustic models $\bar{\Lambda}$ are obtained by the standard training algorithm.

In recognition a similar procedure for the warping factor determination is applied. As the speaker identity is usually unknown, the optimal warping factor is computed utterance-wise. In addition, since also W_j of each spoken utterance j is not known, a preliminary transcription \hat{W}_j need to be obtained in a first recognition pass with unnormalized acoustic vectors X_j and model Λ . Then the factor $\hat{\alpha}_j$ is determined using the normalized full resolution acoustic model $\bar{\Lambda}$.

$$\hat{\alpha}_j = \arg \max_{\alpha} Pr(X_j^\alpha | \bar{\Lambda}, \hat{W}_j).$$

Finally, the acoustic vectors are normalized with $\hat{\alpha}_j$ and a second recognition pass is performed with the normalized acoustic model $\bar{\Lambda}$.

As shown in Table 2, replacing the actually spoken but unknown transcription W_j by \hat{W}_j does not degrade the recognition performance, even if the word error rate of \hat{W}_j is 20..30%. In fact, the WER on the VM II test corpus is almost the same under the following conditions:

- supervised VTN using the correct transcription W_i (WER=23.9%);
- $\hat{\alpha}_j$ is determined by exhaustive search (i.e. the utterance is recognized with all considered warping factors, and the one with the highest likelihood is chosen; WER=24.2%);
- two-pass VTN as described above (WER=23.9%).

In all cases, VTN gains about 8% relative in recognition accuracy.

Table 2: Comparison of two-pass and supervised VTN

Corpus	System	Search Space			Word Error Rate [%]		
		States	Arcs	Trees	Del	Ins	WER
VM II	Baseline (no VTN)	6710	3260	82	5.3	4.6	25.9
	Two-Pass VTN	6076	2959	75	5.0	3.9	23.9
	Exhaustive Search	7346	3581	96	5.1	3.9	24.2
	Supervised VTN	6064	2952	75	5.0	3.8	23.9

3. Warping Functions

The basic principle of VTN is to warp the frequency axis of the spectra during signal analysis with a warping function that is controlled by one or more parameters. In the literature, different warping function were suggested. Wegmann et al. (96) and Welling et al. (98) used a piecewise linear warping function $w_l(f)$, where f denotes the frequency. Up to a limiting frequency f_0 (7 kHz) the spectra are warped linearly with α , and between f_0 and the Nyquist frequency f_N (8 kHz) a different factor α' is applied to ensure that $w_l(f_N) = f_N$, and no frequency region is omitted. Acero & Stern (91) and McDonough (98) applied a bilinear warping function, and Eide & Gish (96) used a power function for spectral warping. Again, in both cases the equation $w(f_N) = f_N$ holds and, like in the piecewise linear case, all spectral lines are shifted either upwards or downwards. More flexible, but also more complex, is the all-pass transform of McDonough (98). This type of warping allows some parts of the spectrum to be shifted upwards, and others downwards at the same time.

We compared the piecewise linear warping $w_l(f)$ with power function warping $w_p(f)$, and a combination of both $w_c(f)$, i.e. the sequential application of both functions. The warping functions are shown in Figure 1:

$$w_l(f) = \begin{cases} \alpha \cdot f & f \leq f_0 \\ \alpha \cdot f + \frac{f_N - \alpha \cdot f_0}{f_N - f_0} \cdot (f - f_0) & f > f_0 \end{cases}$$

$$w_p(f) = \left(\frac{f}{f_N} \right)^\beta \cdot f_N$$

$$w_c(f) = \left(\frac{w_l(f)}{f_N} \right)^\beta \cdot f_N$$

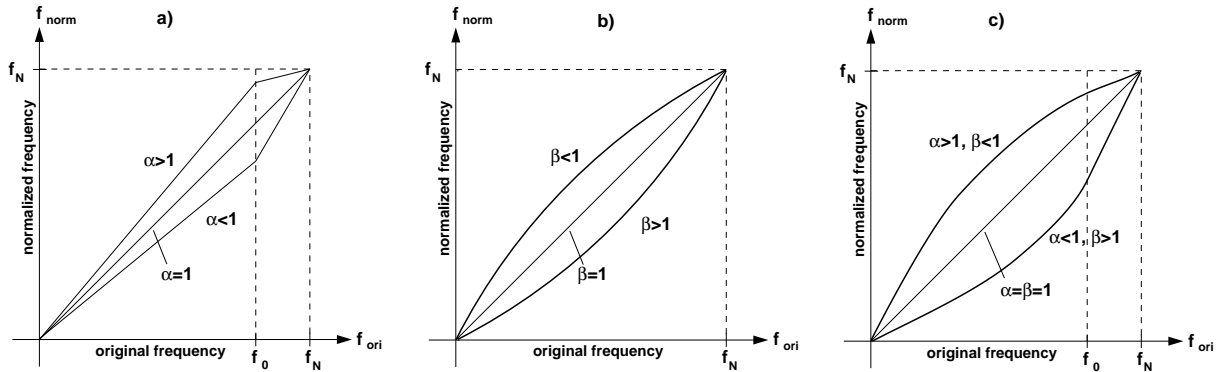


Figure 1: Schematic plot of different warping functions: a) piecewise-linear, b) power function, and c) a combination of both

We found that piecewise linear warping yields better results than the power function warping (WER 23.9% vs. 24.3%). In previous tests on the VM I EVAL-96 test set, the combination of both methods resulted in a minor performance improvement (18.3% vs. 18.0%), but not anymore on the more challenging VM II corpus (WER=24.0%), as can be seen in Table 3.

Table 3: Comparison of different VTN warping functions

Corpus	Warping Function	Search Space			Word Error Rate [%]		
		States	Arcs	Trees	Del	Ins	WER
VM II	Baseline (no VTN)	6710	3260	82	5.3	4.6	25.9
	Piecewise Linear	6076	2959	75	5.0	3.9	23.9
	Power Function	6071	2956	75	5.2	4.1	24.3
	Combined	6059	2950	74	5.1	4.0	24.0

4. Warping Factor Estimation in Recognition

The two-pass approach described so far results in the same improvement of recognition accuracy as supervised VTN. However, as a text dependent approach it requires two recognition passes, which makes real time operation more difficult. Welling (98) suggested a fast text independent algorithm based on Gaussian mixture models (GMMs). The idea is to train one GMM $\bar{\lambda}$ on all normalized training data that describes the distribution of normalized acoustic vectors in feature space.

During recognition, the utterance j is normalized with all considered warping factors, and the warping factor that maximizes the text independent probability Pr is chosen for recognition:

$$\hat{\alpha}_j = \arg \max_{\alpha} Pr(X_j^{\alpha} | \bar{\lambda}).$$

Silence frames do not contribute to the warping factor estimation, and the decision between speech and silence frames relies on a simple heuristic rule. With this approach Welling (99) obtained about half of the WER reduction of two-pass VTN on the 5k WSJ0 corpus.

We have improved this approach by introducing one GMM for each warping factor similar to the procedure suggested in Lee (96). After the initial warping factor determination on the training data, all data with the same warping factor is pooled, and one GMM is trained on these unnormalized acoustic vectors. Hence, each model now represents the distribution of unnormalized vectors of a specific warping factor in feature space. Since the training data is distributed over different models, there is much less data available to train each of the GMMs. To increase the robustness we tie the variance over all models. During recognition the unnormalized acoustic vectors are scored with all GMMs in order to find the best warping factor $\hat{\alpha}$. Speech frames are boosted by weighting the score of each frame with it's energy.

$$\hat{\alpha}_j = \arg \max_{\alpha} Pr(X_j | \lambda_{\alpha}).$$

Figure 2 shows the difference between the two GMM approaches. Warping factors are typically considered in the range $\alpha = 0.88 \dots 1.12$ with step size 0.02.

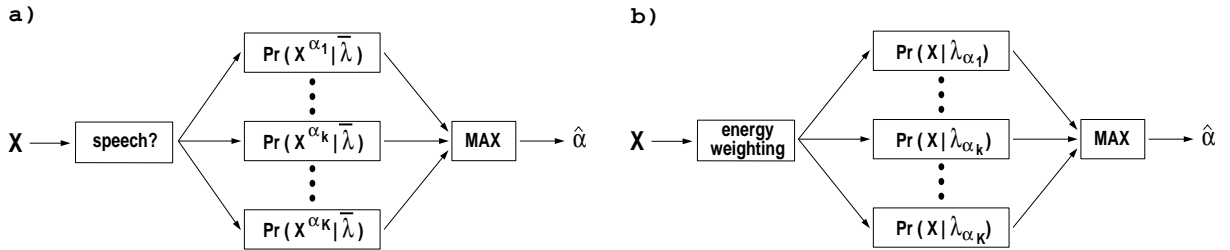


Figure 2: Different GMM-based warping factor determination algorithms: a) approach of Welling (98) with normalized acoustic vectors X_{α} and one GMM $\bar{\lambda}$ for all warping factors, b) new approach with unnormalized acoustic vectors X and one GMM λ_{α} for each warping factor

Recognition results for the VM II and WSJ corpus are given in Table 4. It turned out that due to the reduced number of signal analysis steps our new approach is faster than the one proposed by Welling. Additionally it gives also consistently better recognition results. In fact, it performs almost as good as the two-pass baseline VTN (24.1% vs. 23.9% on VM II, 12.3% in both cases on WSJ), but is faster than a recognition without VTN. The reason is that normalized acoustic models discriminate better and pruning with identical thresholds becomes therefore more efficient.

Table 4: Comparison of Two-Pass and GMM-based fast VTN

Corpus	System	Search Space			Word Error Rate [%]			RTF
		States	Arcs	Trees	Del	Ins	WER	
VM II	Baseline (no VTN)	6710	3260	82	5.3	4.6	25.9	10.7
	Two-Pass VTN	6076	2959	75	5.0	3.9	23.9	21.7
	Fast VTN	6161	2999	75	5.1	4.0	24.1	8.9
WSJ	Baseline (no VTN)	16250	4798	98	1.7	2.2	13.2	15.2
	Two-Pass VTN	12738	3808	78	1.5	2.3	12.3	31.2
	Fast VTN	12964	3876	80	1.5	2.2	12.3	13.6

4. Incremental Warping Factor Estimation

In online recognition tasks, it is desirable to have only little delay between speech recording and the output of the recognition result. This does not only require a fast recognizer, but also algorithms with minimum signal analysis delay between recording and the start of search. So far, the warping factor was always estimated on the whole utterance before the recognition started. The fast VTN approach presented in the previous section, however, allows also incremental warping factor estimation without delay.

As shown in Figure 3, the signal analysis following FFT is carried out twice. The acoustic vector normalized with the currently best warping factor is used for recognition. The unnormalized acoustic vector is immediately evaluated with the GMMs and the probabilities $Pr(X|\lambda_\alpha)$ are accumulated over time. As long as not enough time frames have been collected (200 time frames, i.e. 2 seconds of speech), the warping factor for recognition is set to 1.0. After a few seconds, the best warping factor does not change significantly anymore and recognition takes place with optimally normalized acoustic vectors.

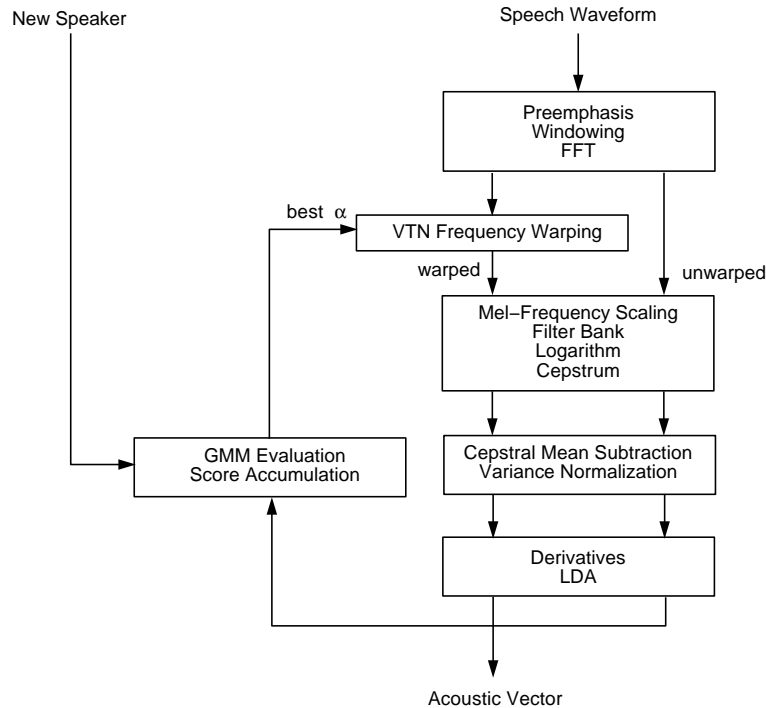


Figure 3: Signal analysis with fast VTN and incremental warping factor estimation.

Tests have shown that incremental warping factor estimation without any delay causes only little performance degradation in the first few seconds of a new speakers' speech (WER 24.4% vs. 24.1%). Thus, also in recognition systems that are accelerated to near real-time (Table 5) we observe a significantly improved WER of 24.8% from VTN compared to 26.1% for the baseline system. The real time factors (RTF) in Table 5 have been measured on a 600 MHz Pentium III machine. Both systems were independently optimized, the details how the recognizer was sped up were presented in Sixtus et al. (00).

Table 5: Comparison of accelerated speech recognition systems

Corpus	System	Search Space			Word Error Rate [%]			RTF
		States	Arcs	Trees	Del	Ins	WER	
VM II	Baseline (no VTN)	1942	969	36	5.0	4.6	26.1	1.2
	Incr. Fast VTN	1801	779	18	5.5	4.1	24.8	1.2

5. Phoneme dependent VTN

One of the advantages of VTN is the low number of free parameters – usually one warping factor per speaker. This parameter can be estimated robustly on very little speech data, whereas adaptation schemes like maximum likelihood linear regression (MLLR) with larger parameter sets require more adaptation data.

On the WSJ corpus we tested whether VTN can be improved by slightly increasing the number of free parameters. In the baseline two-pass recognition approach, a preliminary phonetic transcription of the utterance is available after the first pass. This can be used to assign a phoneme label to each acoustic vector, and determine phoneme dependent warping factors for the utterance.

The results are presented in Table 6. When the number of parameter was increased in recognition only, the word error rate did not change (WER=12.3%). When speaker and phoneme dependent warping factor were used in training only, the recognition accuracy improved slightly to 12.1%. Using this approach both in training and recognition deteriorated the recognition performance a little, however (WER=12.4%). A closer inspection revealed that utterances by male speakers were modeled best with one warping factor, whereas female speakers’ utterances gained significantly from phoneme dependent modeling. Interpolating the frame-wise scores of phoneme dependent VTN with the best warping factor for the whole utterance yielded an optimal interpolation factor of 1.0 for male speaker (i.e. only one warping factor per utterance), and of 0.0 for female speakers (i.e. full weight for the phoneme dependent warping factors). The WER decreased to 11.9%.

Table 6: Speaker vs. speaker and phoneme dependent VTN

Corpus	Training	Test	male	female	Word Error Rate [%]		
			WER [%]	WER [%]	Del	Ins	WER
WSJ	no VTN	no VTN	15.5	11.0	1.7	2.2	13.2
	Speaker Dep. α	One α per Utt.	14.1	10.5	1.5	2.2	12.3
		Pho. Dep. α	14.3	10.4	1.5	2.2	12.3
	Speaker & Phoneme Dep. α	One α per Utt.	13.7	10.6	1.5	2.3	12.1
		Pho. Dep. α	14.7	10.1	1.5	2.2	12.4
		Interpolated	13.7	10.1	1.4	2.2	11.9

6. Conclusions

VTN is normalization scheme widely used in automatic speech recognition. We have shown that a piecewise linear warping function performs at least as good as a non-linear function. The two-pass VTN results in the same word error rate as supervised VTN with given reference transcription.

We presented an improved text independent approach for fast warping factor determination that yields almost the same WER improvements as two-pass VTN, but causes no increase in real time. It could be modified for the use in online systems without introducing time delay.

The recognition performance gain of VTN could be increased using phoneme dependent warping factors. However, male and female speaker behaved different in our tests on the WSJ corpus.

References

- [Acero & Stern 91] Acero A. and Stern R. M.: “Robust speech recognition by normalization of the acoustic space”, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Toronto, pp. 893–896, May 1991.
- [Eide & Gish 96] Eide E. and Gish H.: “A parametric approach to vocal tract length normalization”, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Atlanta, GA, pp. 346–349, May 1996.
- [Lee & Rose 96] Lee L. and Rose R.: “Speaker normalization using efficient frequency warping procedures”, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Atlanta, GA, pp. 353–356, May 1996.
- [McDonough 98] McDonough J. W.: “Speaker Normalisation with All-Pass Transforms”, *Research Notes No. 28, Johns Hopkins University*, Baltimore, MD, September 1998.
- [Ney *et al.* 98] Ney H., Welling L., Ortmanns S., Beulen K., and Wessel F.: “The RWTH Large Vocabulary Continuous Speech Recognition System”, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, WA, pp. 853–856, May 1998.
- [Sixtus *et al.* 00] Sixtus A., Molau S., Kanthak S., Schlüter R., and Ney H.: “Recent Improvements of the RWTH Large Vocabulary Speech Recognition System on Spontaneous Speech”, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Istanbul, pp. 1671–1674, June 2000.
- [Wegmann *et al.* 96] Wegmann S., McAllaster D., Orloff J., and Peskin B.: “Speaker normalization on conversational telephone speech”, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Atlanta, GA, pp. 339–342, May 1996.
- [Welling *et al.* 98] Welling L., Haeb–Umbach R., Aubert X., and Haberland N.: “A study on speaker normalisation using vocal tract normalisation and speaker adaptive training”, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, WA, pp. 797–800, May 1998.
- [Welling *et al.* 99] Welling L., Kanthak S., and Ney H.: “Improved methods for vocal tract normalization”, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Phoenix, AZ, pp. 761–764, March 1999.