Speaker Adaptive Modeling by Vocal Tract Normalization

Lutz Welling, Hermann Ney and Stephan Kanthak

Lehrstuhl für Informatik VI

Computer Science Department

RWTH Aachen – University of Technology

D-52056 Aachen, Germany

Abstract. This paper presents methods for speaker adaptive modeling using vocal tract normalization (VTN) along with experimental tests on three databases. We propose a new training method for VTN: By using single-density acoustic models per HMM state for selecting the scale factor of the frequency axis, we avoid the problem that a mixture-density tends to learn the scale factors of the training speakers and thus cannot be used for selecting the scale factor. We show that using single Gaussian densities for selecting the scale factor in training results in lower error rates than using mixture densities.

For the recognition phase, we propose an improvement of the well–known two–pass strategy: By using a non-normalized 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 two–pass strategy is an efficient method, but it is suboptimal because the scale factor 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.

In summary, on the German spontaneous speech task Verbmobil, the WSJ task and the German telephone digit string corpus SieTill, the proposed methods for VTN reduce the error rates significantly.

Index Terms: speech recognition, speaker adaptive recognition, vocal tract (length) normalization, speaker adaptive modeling and training.

EDICS: 1-RECO (speech recognition), 1-SPRD (speech production and perception).

1 Introduction

This paper deals with methods for vocal tract normalization (VTN) [7, 11, 12, 16, 18, 19]. The main idea of VTN is to normalize the frequency axis of the acoustic vectors for each speaker in the recognition process and thus to remove speaker dependent variability in the acoustic vectors. When using Gaussian mixtures as emission distributions in the HMMs, one of the main problems is that the various speaker-dependent scale factors tend to be modeled by the multimodality of the mixture distributions. As a result, it is difficult to reliably select the speaker-dependent scale factor of the frequency axis. This problem is addressed in a number of papers [11, 19, 21]. The method used in this paper for selecting the speaker-dependent scale factor was first presented in [21] and is based on using single Gaussian densities.

Using this single-density method for scale factor selection, this paper presents the following novel contributions for VTN:

- *Principles.* The principles of VTN-based recognition are outlined in the frame-work of the Bayes decision rule and speaker adaptive modeling.
- *Training method for VTN.* We present a training method for VTN where in contrast to the methods described in [11, 19] the scale factors for the training speakers are estimated only once using no cross validation. Our method uses single Gaussian densities per HMM

state for scale factor selection in training. We show that using single Gaussian densities for scale factor selection in training results in better recognition rates than using mixture densities.

- *Improved two–pass strategy.* An improved two–pass strategy [11] for VTN in recognition is presented: By using a non-normalized acoustic model for the first recognition pass instead of a normalized model, lower error rates are obtained.

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

- *Fast VTN.* A method for fast scale factor selection in recognition is described in detail. The method is similar to the method in [19], but the training of the mixture model for normalized acoustic vectors is simpler with our approach. In recognition tests, the fast VTN is compared with the improved two–pass strategy.

We will present recognition tests on three different databases, namely the German spontaneous speech task Verbmobil [8], the Wall Street Journal task and the German telephone digit string corpus SieTill. The results will show that the proposed method for VTN in training in combination with the improved two–pass strategy leads to consistent reductions in the word error rates.

2 Speaker Adaptive Acoustic Modeling

In order to use VTN appropriately, it is useful to study it in the framework of speaker adaptive modeling. Speaker adaptive modeling can be viewed as a special case of stochastic matching as presented in [17].

2.1 Principles

In speech recognition, we have a sequence of acoustic vectors or observations over time $t = 1, ..., T$:

$$X = x_1 ... x_t ... x_T \tag{1}$$

In conventional, i.e. non–adaptive, acoustic modeling, we assume a model distribution $p(X|W;\theta)$ with suitable reference parameters θ for each hypothesized word sequence W. Typically, an HMM framework is used:

$$p(X|W;\theta) = \sum_{s_1^T} \prod_{t=1}^{T} [p(s_t|s_{t-1}, W) \cdot p(x_t|s_t, W;\theta)] \tag{2}$$

with state sequence $s_1^T = s_1 ... s_t ... s_T$, transition probabilities $p(s_t|s_{t-1}, W)$ and emission probabilities $p(x_t|s_t, W;\theta)$. The typical model for the emission probabilities is that of multivariate Gaussian mixture distributions, where the model parameters θ stand for the mean vector and the covariance matrix.

In speaker-adaptive acoustic modeling, we assume that there is a dependence of the distribution on some speaker specific parameter α:

$$P(X|W;\theta, \alpha) \tag{3}$$

The typical example of such a speaker specific parameter is the scale factor of the frequency axis as used in VTN. As we will discuss later, also the widely used MLLR approach (MLLR = maximum likelihood linear regression) [13] fits into this framework, where the single parameter α has to be extended to a matrix of parameters. In addition to *speaker* adaptive modeling, the same method can be used to describe other types of adaptive modeling like microphone, acoustic channel and speaking rate adaptation.

In the HMM framework, the adaptation is typically concerned only with the emission distribution:

$$p(X|W;\theta,\alpha) = \sum_{s_1^T} \prod_{t=1}^{T} [p(s_t|s_{t-1},W) \cdot p(x_t|s_t,W;\theta,\alpha)] \tag{4}$$

To handle the unknown adaptive parameter α, we consider it to be a hidden variable (continuous-valued) that has to be integrated out:

$$\begin{aligned} p(X|W;\theta) &= \int d\alpha \; p(X,\alpha|W;\theta) && (5)\\ &= \int d\alpha \; p(\alpha|W) \cdot p(X|W;\theta,\alpha) && (6)\\ &\cong \max_{\alpha} \{p(\alpha|W) \cdot p(X|W;\theta,\alpha)\} && (7) \end{aligned}$$

where, in the final equation, we have used the well-known maximum approximation to obtain the final approximation to avoid the tremendous integration problem.

We have then the following Bayes decision rule for adaptive modeling:

$$\arg\max_{W} \left\{ p(W) \cdot \int d\alpha \; p(\alpha|W) \cdot p(X|W;\theta,\alpha) \right\} \tag{8}$$

with the language model $p(W)$. In the maximum approximation, we have:

$$\arg\max_{W} \left\{ p(W) \cdot \max_{\alpha} \{p(\alpha|W) \cdot p(X|W;\theta,\alpha)\} \right\} \tag{9}$$

2.2 Training

In training, for each speaker $r = 1, ..., R$, we have pairs (X_r, W_r) of training data X_r and word sequences W_r:

$$(X_r, W_r) \quad r = 1, ...R \tag{10}$$

From these training data, we estimate the unknown model parameters by maximum likelihood:

$$\arg\max_{\theta} \prod_{r=1}^{R} \int d\alpha_r \; p(\alpha_r|W_r) \cdot p(X_r|W_r;\theta,\alpha) \tag{11}$$

or in the maximum approximation:

$$\arg\max_{\theta} \prod_{r=1}^{R} \max_{\alpha_r} \{p(\alpha_r|W_r) \cdot p(X_r|W_r;\theta,\alpha_r)\} \tag{12}$$

As it stands, even in the maximum approximation, this joint estimation of the (speaker specific) adaptation parameter set $\{\alpha_r : r = 1, ..., R\}$ and the model parameter set θ is a complex optimization problem for which approximations will be considered later.

2.3 Realization by Transformations

To obtain a practical method, we have to specify the exact dependence of the model on the speaker specific parameter α. Although many types of suitable functional dependencies might be thought of, in virtually all approaches the functional form of the non-adaptive model is kept fixed, and the adaptive operations are implemented as transformations. Typically, we distinguish two variants of transformations:

- transformation of the model parameters θ:

  For each value of the speaker specific parameter α, we transform the non-normalized model parameter θ into normalized ones θ^α:

  $$\theta \rightarrow \theta^\alpha \tag{13}$$

  so that we have for the distribution:

  $$p(X|W;\theta,\alpha) \quad = \quad p(X|W;\theta^\alpha) \tag{14}$$

  The prototypical example of this approach is the so-called MLLR method (MLLR = maximum likelihood linear regression) [13, 9]. In MLLR, the parameter set α defines a transformation matrix $A(\alpha)$ and is applied to the mean vector θ of Gaussian distributions:

$$\theta \rightarrow \theta^{\alpha} \quad = \quad A(\alpha) \cdot \theta \tag{15}$$

maybe with an additive offset. In [2], the training problem of such matrix transformations as formulated by Eq.(12) is referred to as speaker-adaptive training (SAT).

- transformation of the observations X:

  This can be formulated as a mapping of the acoustic vectors:

$$X \rightarrow X^{\alpha} \tag{16}$$

  so that we have for the distribution:

$$p(X|W;\theta,\alpha) \quad = \quad p(X^{\alpha}|W;\theta) \tag{17}$$

  When changing the random variable from X to X^{α}, we must include the so-called Jacobian determinant for the transformation [14]. This is not expressed explicitly in our notation since we use the *same* symbols for the non-adaptive model $p(X|W;\theta)$ and the adaptive model $p(X^{\alpha}|W;\theta)$. Anyway, the effect as such tends to be small and is neglected in this paper.

  The prototypical example of this approach is the VTN, which is better referred to as *frequency axis normalization*: we determine a speaker dependent scale factor α of the frequency axis and use it to re-scale the frequency axis *before* computing the acoustic vectors for recognition. The details are given in the next section.

3 VTN Implementation

3.1 Vocal Tract Normalization and Front End

For VTN, a piecewise–linear frequency scaling is used [19]. Fig. 1 shows the acoustic front–end including VTN. As shown in the figure, the frequency scaling is applied to the magnitude spectrum before the mel frequency scaling is performed. The selection of the scale factor α is done using an exhaustive search for α in the range $0.88 \leq \alpha \leq 1.12$ with step size 0.02. A scale factor is estimated for each sentence in recognition and for each speaker in training.

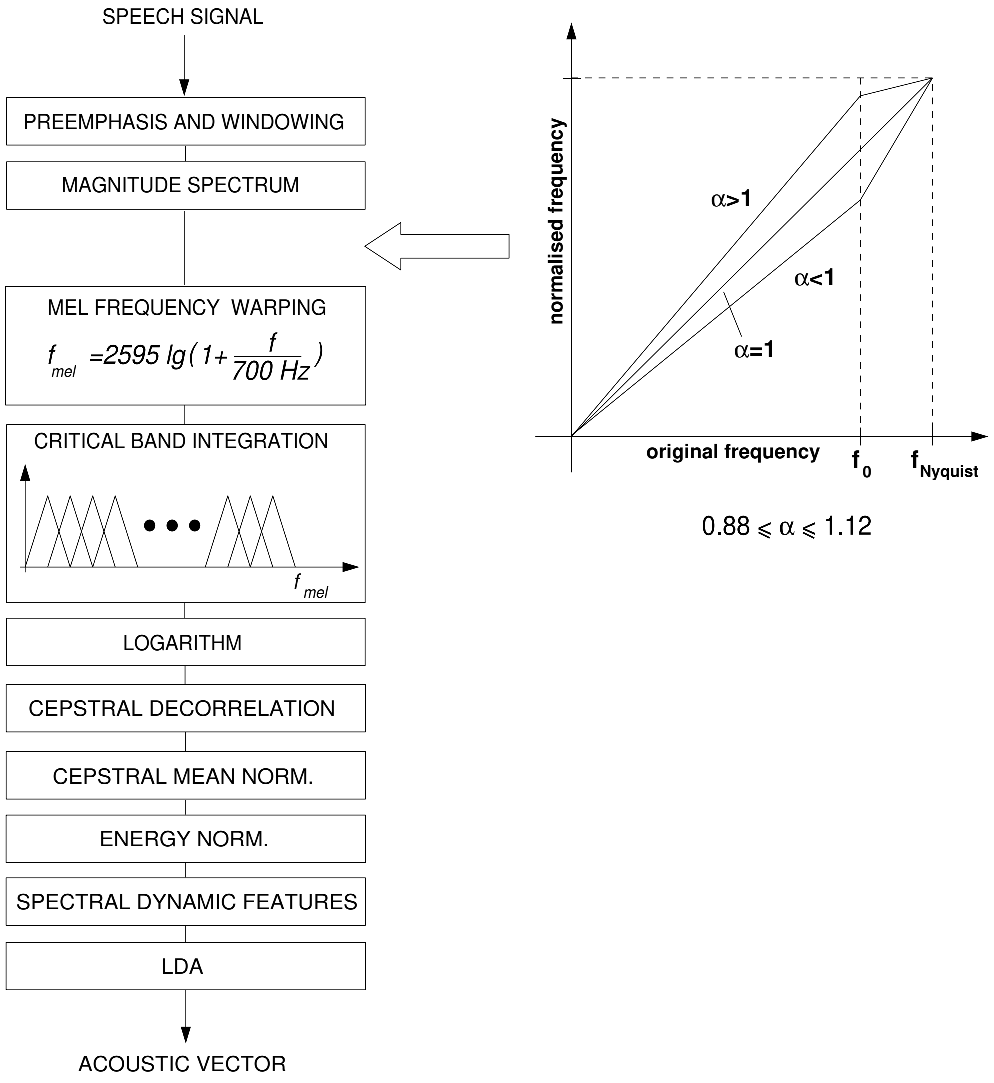


Figure 1: Acoustic front–end with VTN.

The term normalization will be used in two contexts:

- normalized observations X^{α}:

  the original observations are re-scaled using the scale factors α.

- normalized parameter $\tilde{\Lambda}$:

  they are computed in the training from *normalized* observations X^{α}.

As usual, the prior distribution $p(\alpha|W)$ is assumed to be uniform. Only recently, there has been some studies to make explicit use of the prior distribution [5, 6].

In the following, we will focus on methods for training and recognition with VTN. Since we will use different types of acoustic models with different model parameters, we have summarized the different model parameters in Table 1.

Table 1: Notation for acoustic models.

Adaptation	Symbol	Model structure
no	Λ	Mixture density per HMM state
	λ	Single density per HMM state
yes	$\tilde{\Lambda}$	Mixture density per HMM state
	μ	Mixture model (global)

In addition, we give a summary of often used symbols:

$X := \{x_1, \ldots, x_t, \ldots, x_T\}$	sequence of acoustic vectors
$W := \{w_1, \ldots, w_n, \ldots, w_N\}$	word sequence
$r = 1, ..., R$	training speaker index
X_r	acoustic vectors of training speaker r
W_r	word sequence of training speaker r
α_r	scale factor of training speaker r

3.2 Training

Without VTN, the acoustic vectors and therefore the acoustic model parameters Λ depend on the varying vocal tract lengths of the training speakers. The goal of applying VTN in the training phase is to train an acoustic model with parameters $\tilde{\Lambda}$ that are (more) independent of the vocal tract lengths of the training speakers. This leads to the following speaker–adaptive training criterion for the normalized model parameters $\tilde{\Lambda}$:

$$\tilde{\Lambda} \quad = \quad \arg\max_{\Lambda'} \prod_{r=1}^{R} \max_{\alpha} P(X_r^{\alpha}|W_r;\Lambda') \ . \tag{18}$$

The optimization of Eq. (18) is a difficult problem because the optimization over α has to be carried out separately for each speaker r and over the unknown model parameters Λ' at the same time. Two complex methods have been proposed so far [11, 19]. The common idea of both methods is that the scale factor for a training speaker r is selected using an acoustic model which has been trained without using data of speaker r. Also, both methods are based on an iterative process where each iteration consists of maximum likelihood estimation of scale factor and model parameter.

In this work, we propose a three–step procedure for the training of a normalized acoustic model $\tilde{\Lambda}$ with VTN. This method is conceptually simpler than the methods in [11, 19] and consists of the following steps:

1. An acoustic model λ consisting of a single Gaussian density per HMM state is estimated from the non-normalized acoustic vectors X_r of all training speakers r with $r = 1, \ldots, R$ by maximum likelihood training [15]:

$$\lambda \quad = \quad \arg\max_{\lambda'} \prod_{r=1}^{R} p(X_r|W_r;\lambda') \ . \tag{19}$$

2. For each training speaker r, a scale factor α_r is chosen as the scale factor which results in the maximum likelihood for the training data of this speaker $X_r^{\alpha_r}$ using the *single density* model λ:

$$\alpha_r \quad = \quad \arg\max_{\alpha} p(X_r^{\alpha}|W_r;\lambda) \quad \text{for} \quad r = 1,\ldots,R \; , \tag{20}$$

   where, for optimal results, we perform a forced time alignment for each training sentence.

3. A normalized model $\tilde{\Lambda}$ is trained on the normalized acoustic vectors $X_r^{\alpha_r}$ of all training speakers r with $r = 1,\ldots,R$ by maximum likelihood training:

$$\tilde{\Lambda} \quad = \quad \arg\max_{\Lambda'} \prod_{r=1}^{R} p(X_r^{\alpha_r}|W_r;\Lambda') \; . \tag{21}$$

In this method, the use of the single-density model for scale factor selection is crucial. We will study the effect of the choice of the acoustic model used for scale factor estimation in Section 4.3.

3.3 Recognition

This section is about recognition with VTN. Using the Bayes decision rule in the maximum approximation (see Eq. 9) and a uniform distribution of the scale parameter α, we obtain the criterion for the unknown word sequence W:

$$\arg\max_{W} \left\{ p(W) \cdot \max_{\alpha} p(X^{\alpha}|W;\tilde{\Lambda}) \right\} \; . \tag{22}$$

Three different methods for the optimization of Eq. (22) will be considered in this paper:

- **full optimization:**

  For a full optimization over the scale factor α and the word sequence W, we exchange the order of the optimizations over W and α. As a result, we obtain a a separate recognition pass for each value of the scale factor α:

$$\max_{\alpha} \left\{ \max_{W} \left\{ p(W) \cdot p(X^{\alpha}|W;\tilde{\Lambda}) \right\} \right\} \; . \tag{23}$$

The word sequence that corresponds to the maximizing value of α is then selected as recognized word sequence. Due to its computational cost, this method is used only if the recognition task is simple enough, such as digit string recognition.

- **two–pass strategy:**

  There are suboptimal, but efficient two–pass strategies to avoid the cost of the full optimization. In this paper, we will distinguish and use two variants, namely a baseline and an improved two–pass strategy. The **baseline two–pass strategy** has been proposed in [11, 12] and works as follows:

  1. A first recognition pass with non-normalized acoustic vectors X and a normalized acoustic model $\tilde{\Lambda}$ produces a preliminary transcription $\tilde{W}$:

     $$\tilde{W} = \arg\max_{W} \left\{ p(W) \cdot p(X|W;\tilde{\Lambda}) \right\} \ . \tag{24}$$

     Using the recognition result $\tilde{W}$, the scale factor α is then selected according to

     $$\tilde{\alpha} = \arg\max_{\alpha} p(X^{\alpha}|\tilde{W};\tilde{\Lambda}) \ . \tag{25}$$

  2. A second recognition pass with normalized acoustic vectors $X^{\tilde{\alpha}}$ and a normalized acoustic model $\tilde{\Lambda}$ produces the final recognition result:

     $$\arg\max_{W} \left\{ p(W) \cdot p(X^{\tilde{\alpha}}|W;\tilde{\Lambda}) \right\} \ . \tag{26}$$

  In addition to this baseline method, we will also consider an **improved two–pass strategy**, where a non-normalized reference model Λ is applied in this first pass:

  $$\tilde{W} = \arg\max_{W} \{ p(W) \cdot p(X|W;\Lambda) \} \ . \tag{27}$$

  As the experiments will show, this variant produces consistently better results.

- **fast VTN:**

  A significant reduction in computation time can be obtained by using a *text–independent* acoustic model $p(X^\alpha|\mu)$ with model parameter μ for scale factor selection [11, 12, 19, 21, 22]:

  $$\tilde{\alpha} = \arg\max_{\alpha} p(X^\alpha|\mu) \quad . \tag{28}$$

  The training of the text–independent model parameters μ will be addressed later.

  As in the two–pass strategy before, the recognition pass itself is then carried out using the obtained scale factor $\tilde{\alpha}$:

  $$\arg\max_{W} \left\{ p(W) \cdot p(X^{\tilde{\alpha}}|W;\tilde{\Lambda}) \right\} \ . \tag{29}$$

  In the following, this method will be referred to as *fast VTN*.

4 Experiments

This section is on our experimental results. We will first give the details of our experimental conditions. Then we will present experimental results for training and recognition with VTN.

When using mixture densities, we are always faced with the question of how many component densities to select. There is no perfect answer to that question because there might very well be an interaction between VTN and the number of densities. As a compromise, we varied the number of densities (using a certain stepsize) and typically selected that number of densities that resulted in optimal recognition accuracy.

4.1 Experimental Conditions

4.1.1 Recognition Tasks and Databases

Experimental tests were carried out on three different databases, namely the WSJ0, Verbmobil and SieTill corpora:

- WSJ0 task:

  The WSJ0 (WSJ= Wall Street Journal) is used by many research groups. For the WSJ0 experiments, recognition was done on the Nov.'92 development and evaluation sets (18 speakers, 740 sentences, 12132 words, 85 minutes) and training on the WSJ0 training corpus (84 speakers, 16 hours).

- Verbmobil task:

  Unlike the other two corpora, the Verbmobil corpus comprises recordings of spontaneous-speech dialogues. The domain of the dialogues is the scheduling of business meetings.

- SieTill task:

  The SieTill corpus is similar to the TI digit string corpus. The differences are that the language is German rather than English and that the corpus was recorded via telephone. The vocabulary consists of the German digits including both pronunciations of the digit 2, namely 'zwei' and 'zwo'.

More details of each of the three databases are given in Table 2.

Table 2: Statistics of speech corpora used for training and recognition experiments.

corpus		WSJ0	SieTill	Verbmobil
task		read newspaper articles	digit strings	spontaneous speech, dialogues
signal quality		clean speech	telephone speech	clean speech
sampling frequency		16 kHz	8 kHz	16 kHz
language		US English	German	German
vocabulary size		5000	11	7000
training	speakers	84	362	613
	record. hours	16	12	27
recognition	speakers	18	356	62
	sentences	740	13114	343
	words	12132	43086	6446
	record. hours	1.42	12	0.67
	language model	bigram/trigram	zerogram	trigram
	test-set perplexity	107/56	11	35

4.1.2 Recognition Systems

For the experiments on the large vocabulary corpora WSJ0 and Verbmobil, we used the recognizer described in [15]. This recognizer employs a cepstrum front–end that includes linear discriminant analysis (LDA) [20]. For acoustic modeling, continuous density hidden Markov models along with decision–tree based state tying are used. A time–synchronous left–to–right beam search strategy in combination with a tree-organized pronunciation lexicon is used for recognition.

For the tests on the SieTill database, we employed a small vocabulary recognition system based on whole–word hidden Markov models with continuous observation densities. The hidden Markov models are left–to–right models with forward, skip and loop transitions. The system has whole–word models for 11 German digits including the pronunciation variant '*zwo*' for '*zwei*'. The total number of states and thus emission distributions is 425 states including 1 state for silence. The emission probabilities are modeled as single Gaussian densities with a single variance vector pooled over all states.

In this work, all experiments were carried out using gender–independent acoustic models. For training, the adaptation was performed speaker by speaker. For recognition, the adaptation was performed sentence by sentence because this is the more practical method in a real application. For the two tasks Verbmobil and WSJ, both the normalized and the non-normalized emission distributions are Gaussian mixtures. Since their number may vary from test to test, we typically will report the number of Gaussian densities used for each experiment.

4.2 Effect of Silence

In areas of silence, the acoustic vectors do not depend on the vocal tract length. Therefore, it could be useful to exclude acoustic vectors assigned to the silence model from the scale factor selection. As Table 3 demonstrates, the recognition rates on the WSJ0 and SieTill data are indeed slightly better if the selection of the scale factor in training and recognition is performed on speech only, i.e. excluding silence portions. As usual, the word error rate (WER) between spoken word sequence and recognized word sequence is computed using the Levenshtein distance, resulting in the minimum number of deletion (DEL), substitution (SUB) and insertion

(INS) errors between the two word sequences. The experiments in Table 3 were done using the three–step procedure for training with VTN proposed in Section 3.2 and the two–pass strategy explained in Section 3.3.

In all experiments described in the remaining of this paper, the silence portions are *not* used for scale factor selection. This applies to both training and recognition.

Table 3: Effect of silence on scale factor selection for VTN.

Corpus	VTN	Silence vectors	#Densities	DEL – INS %	WER %
WSJ0	no	-	103k	1.4 – 0.6	6.8
	yes	yes	140k	1.0 – 0.6	6.1
		no	140k	1.2 – 0.5	5.9
SieTill	no	-	425	1.0 – 0.8	4.6
	yes	yes	425	1.0 – 0.8	4.2
		no	425	1.0 – 0.8	4.0

4.3 Training method

The training method for VTN as presented in Section 3.2 is conceptually simple. However, the choice of a suitable acoustic model for scale factor selection is crucial. Table 4 summarizes the effect of the acoustic resolution, i.e. the number of Gaussian densities used for scale factor selection, on VTN in recognition.

The table shows the number of Gaussian densities per HMM state for scale factor selection along with the corresponding recognition results on the WSJ0 data obtained with the baseline two–pass strategy as described in Section 3.3. As can be seen from Table 4, applying VTN only in recognition gives a word error rate of 6.8%, i.e. no improvement over the baseline result (see Table 3). Using VTN also in training with 32 Gaussians per HMM state for scale factor selection does not lead to any improvement either. The experimental analysis shows that a scale factor of 1.00 is selected for each training speaker. The reason is that the 32 Gaussians per HMM state are able to capture the scale factors of all training speakers. Using eight Gaussians per HMM state, the error rate is reduced down to 6.1%. The lowest word error rate of 5.9% is obtained with only a single Gaussian density per HMM state for the selection of the scale factors in training.

A remark about the number of densities is in order. As shown in Table 4, the number of densities had to be increased from 103k to 140k or 143k in order to obtain minimum error rate. The effect of the number of densities is shown in Fig. 2. This figure shows the word error rate as a function of the number of Gaussian component densities in the acoustic model with and without VTN in training and recognition. The minimum word error rates are obtained for 103k densities without VTN and for about 140k densities with VTN. As can be seen, the normalization by VTN results in "purer" models. For the same word error rate, the VTN-based system requires only about half the number of densities in comparison with the baseline recognition system without VTN. For the VTN-based system, the two–pass strategy described in Section 3.3 was used.

Fig. 3 shows the score $[-\log(p(X_r^\alpha|W_r;\lambda))]$ as a function of the scale factor α for the male speaker *20r* from the WSJ0 training data and for the female speaker KAE from the Verbmobil training data, respectively. Both examples shown in Fig. 3 represent the most typical case where the score is a smooth function of the scale factor with a single well-defined minimum.

Histograms of the obtained scale factors using single Gaussian densities are shown in Fig. 4 for WSJ0 and SieTill corpora. In the histogram for the SieTill data, the range for the scale factor 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 Fig. 4, only approximately five percent of the training speakers are assigned a scale factor lower than 0.88 or higher than 1.12. Both histograms show that the distribution of the chosen scale factors is approximately bimodal with one mode for each gender.

The remaining of this section is about recognition with VTN. In the following, all reported experiments were carried out using the training method with VTN presented in this section.

Table 4: Effect of the number of densities per HMM state for scale factor selection in training on the word error rate (corpus: WSJ).

recognition method	#Dens./State for scale selection	#Dens. for recognition (cf. Fig. 2)	DEL – INS %	WER %
no VTN	-	103k	1.4 – 0.6	6.8
VTN in recognition only	-	103k	1.4 – 0.6	6.8
VTN in training + recognition	32	103k	1.4 – 0.6	6.8
	8	143k	1.2 – 0.6	6.1
	1	140k	1.2 – 0.6	5.9

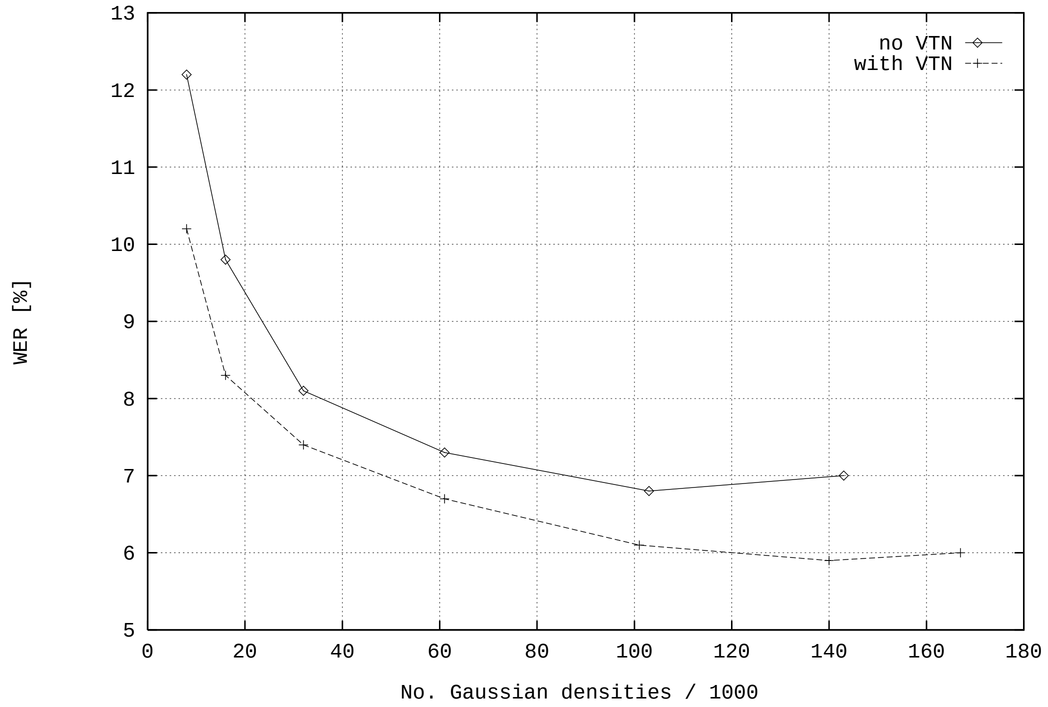


Figure 2: Word error rate as a function of the number of Gaussian component densities (corpus: WSJ0).

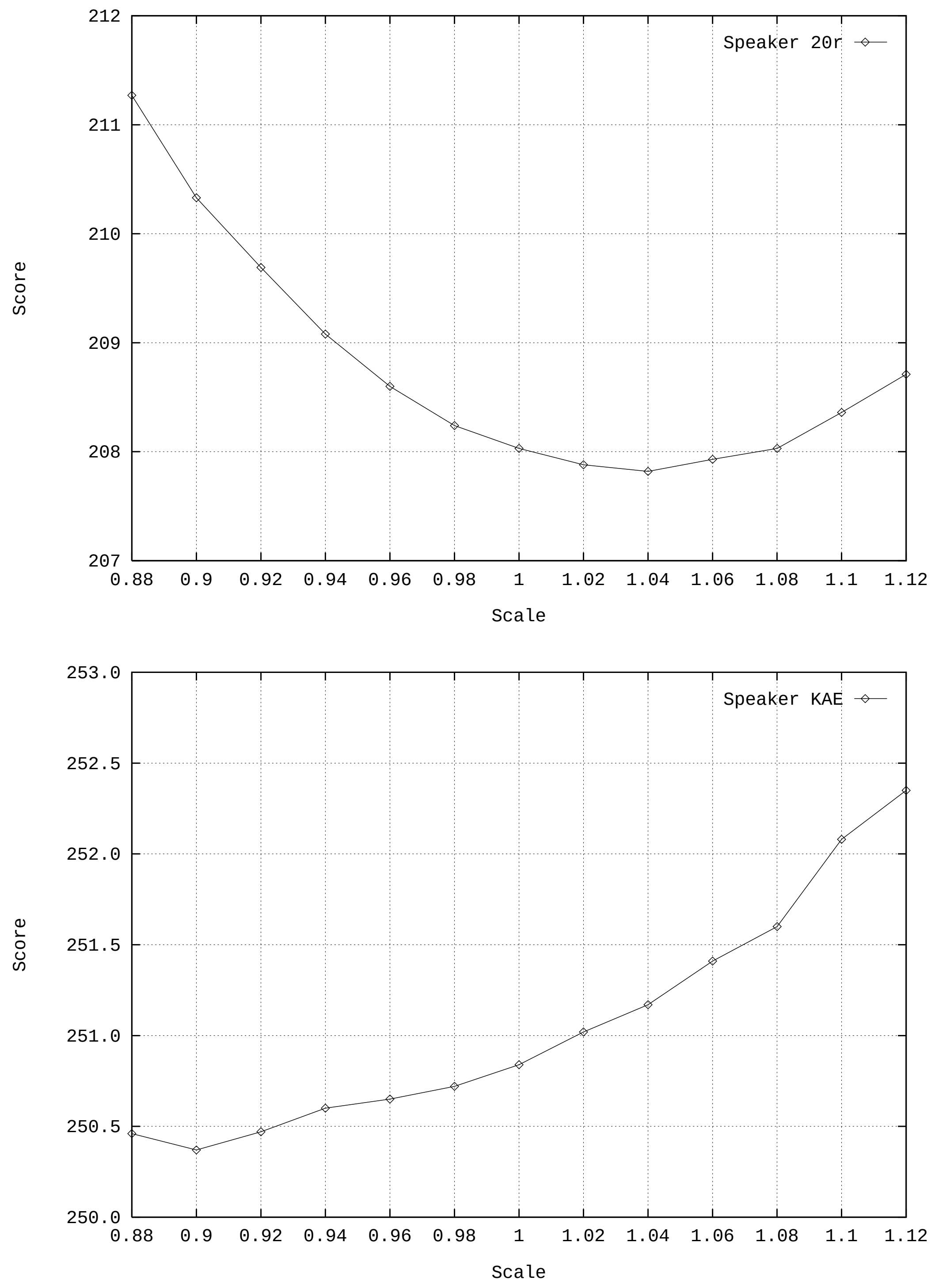


Figure 3: Score $[-\log(p(X_r^\alpha|W_r;\lambda))]$ as a function of the scale factor α (*top*: male speaker *20r* from WSJ0 training data, *bottom*: female speaker *kae* from Verbmobil training data).

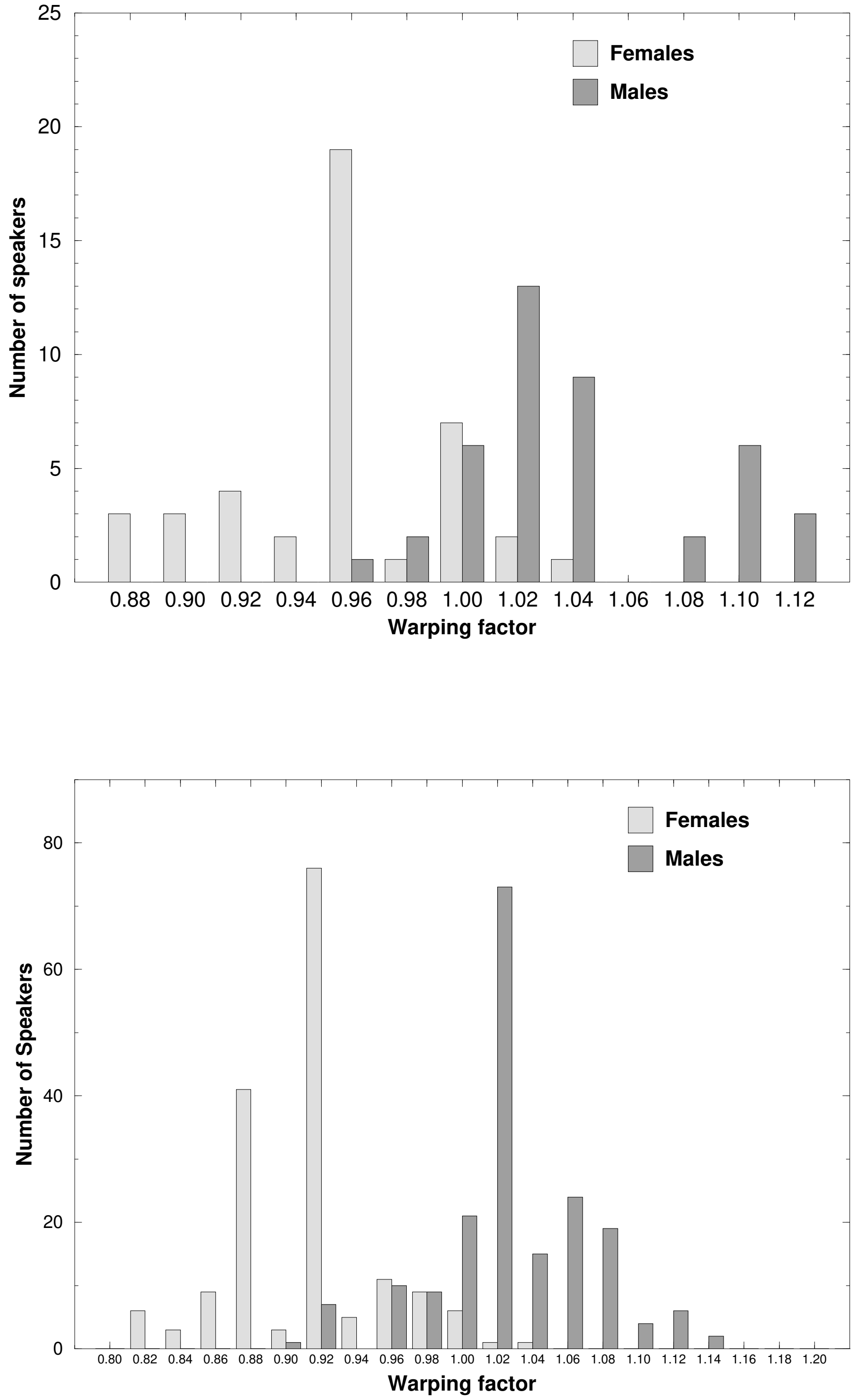


Figure 4: Histograms over scale factors for training speakers: *top* WSJ0; *bottom*: SieTill.

4.4 Improved Two–Pass Strategy

The baseline two–pass strategy [11, 12] as described in Section 3.3 uses a normalized acoustic model in both recognition passes. However, the number of word errors in the preliminary transcription $\tilde{W}$ can be reduced if a non-normalized model is used in the first recognition pass. As Table 5 shows, the error rates on the WSJ0, Verbmobil and SieTill databases using non-normalized acoustic vectors are consistently better with a non-normalized acoustic model Λ than with a normalized model $\tilde{\Lambda}$.

Therefore, we propose to improve the two–pass strategy by using a non-normalized model Λ instead of an normalized model $\tilde{\Lambda}$ in the first recognition pass. An experimental comparison of the baseline and the improved two–pass method is given in Table 6. Without VTN in training and recognition, a word error rate of 6.8% is obtained on the WSJ0 testing data. The two–pass strategy described in Section 3.3 gives an error rate of 5.9%. The improved two–pass strategy 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 (see Table 6). The baseline two–pass strategy reduces the error rate down to 4.0%. Again, the improved two–pass strategy outperforms the baseline method and gives a word error rate of 3.9%.

Table 5: Effect of acoustic model normalization in first–pass recognition.

Corpus	Normalized Acoustic Model	#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	425	1.0 – 0.8	4.6
	yes	425	1.0 – 0.9	5.3

Table 6: Comparison of baseline and improved two–pass strategy for recognition with VTN.

Corpus	type of VTN	#Dens.	DEL – INS %	WER %
WSJ0	no	103k	1.4 – 0.6	6.8
	baseline two–pass	140k	1.2 – 0.5	5.9
	improved two–pass	103k/140k	1.2 – 0.5	5.7
SieTill	no	425	1.0 – 0.8	4.6
	baseline two–pass	425	1.0 – 0.8	4.0
	improved two–pass	425	1.0 – 0.8	3.9

4.5 Suboptimality of Two–Pass Strategy

The previously described two–pass strategies are suboptimal methods for maximizing Eq. (22) since the scale factor and the word sequence are determined sequentially and not in combination.

In this section, we show that for telephone digit string recognition this suboptimality of the two–pass reduces the gain in recognition performance due to VTN significantly.

Table 7 compares the error rates on WSJ0 and SieTill for the improved two–pass strategy and for scale factor selection based on the correct transcription of the test sentences instead of the preliminary transcription $\tilde{W}$. As Table 7 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 two–pass strategy compared to 3.6% with the correct transcription.

A full optimization over the scale factor and the word sequence can be implemented by a separate recognition pass for each possible value of the scale factor. The combination of word sequence and scale factor which maximizes Eq. (23) is selected. This method leads to an error rate of 3.6% on the SieTill data, as shown in Table 7, instead of 3.9% with the improved two–pass strategy and 4.6% with no VTN. Thus we observed that for telephone digit string recognition the suboptimal two–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 according to Table 8 and the more difficult acoustic conditions for the telephone–line recorded SieTill data.

Another question is concerned with the difference between sentence-wise and speaker-wise adaptation. Table 8 shows some statistics for this question. In addition to the average duration of the test sentences and the average number of words per test sentence, Table 8 also gives the absolute deviation of the sentence-wise estimation of the scale factor α, i.e. for each speaker r, we compute the average absolute difference between the scale factor α_{rs} estimated from a single sentence s and the scale factor $\bar{\alpha}_r$ estimated from all sentences. Then these speaker-dependent absolute deviations are averaged over all speakers so that we have for the absolute deviation d_α:

$$d_\alpha \quad = \quad << |\alpha_{rs} - \bar{\alpha}_r| >_s>_r \tag{30}$$

where the brackets $< >$ denote the averaging operation over the sentences s and speakers r. As can be seen from Table 8, this absolute deviation is 0.0105 for on the WSJ0 test data and 0.0292 for the SieTill test data. Therefore, for a given speaker the changes in scale factors among different sentences are significantly smaller on the WSJ0 than on the SieTill data. However, compared to the possible range of scale factors from 0.88 to 1.12, these differences are small on both databases.

Table 7: Effect of the suboptimality of the two–pass strategy for VTN in recognition on the word error rate.

Corpus	type of VTN	#Dens.	DEL – INS %	WER %
WSJ0	no	103k	1.4 – 0.6	6.8
	improved two–pass	103k/140k	1.2 – 0.5	5.7
	correct transcription	140k	1.1 – 0.5	5.7
SieTill	no	425	1.0 – 0.8	4.6
	improved two–pass	425	1.0 – 0.8	3.9
	correct transcription	425	1.0 – 0.7	3.6
	full optimization	425	1.0 – 0.7	3.6

Table 8: Effect of sentence-wise and speaker-wise estimation of scale factors α.

Corpus	#Speakers	#Sentences	Average Sentence		abs. deviation
			Length [sec]	#Words	of α
WSJ0	18	740	7.0	16.4	0.0105
SieTill	356	13114	3.2	3.3	0.0292

4.6 Fast VTN

In the following, we first describe the implementation details of the fast method for scale factor selection in recognition which was introduced in Section 3.3. Then, we report an experimental comparison of the fast VTN with the improved two–pass method.

Similar to the approach in [19], the method presented in this work is based on a Gaussian mixture model that represents the distribution of normalized acoustic vectors. After the training phase described in Section 3.2, a *text–independent* Gaussian mixture model μ is trained on the normalized acoustic vectors. During recognition, the scale factor is selected using the Gaussian mixture model μ according to Eq. (28).

In the experiments, the Gaussian mixture model had a single diagonal covariance matrix and 64 component densities. The fast selection of the scale factor was performed on speech excluding silence portions: For each sentence, we compute the component density μ_{sil} of the mixture model μ which is selected most often using the maximum approximation. Acoustic vectors assigned to μ_{sil} are not used for scale factor selection. Fig. 5 illustrates this fast method of scale factor selection.

Table 9 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 improved two–pass strategy described in Section 4.4 clearly outperforms the fast method and gives an error rate of 5.7%. The fast VTN has not been tested on the SieTill data since the fast method is not necessary for digit string recognition.

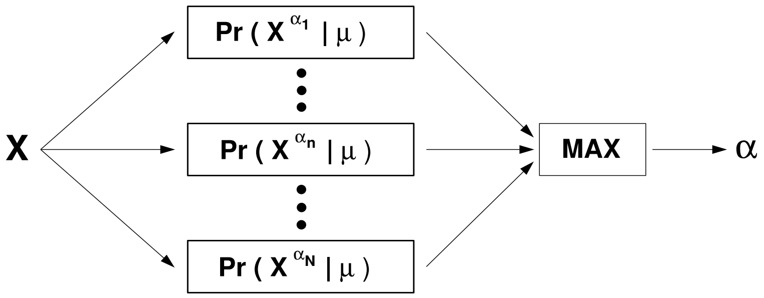


Figure 5: Fast selection of scale factor in recognition.

Table 9: Comparison of fast VTN and improved two–pass strategy for recognition with VTN (corpus: WSJ).

VTN	#Dens.	DEL – INS %	WER %
no	103k	1.4 – 0.6	6.8
fast	143k	1.3 – 0.6	6.2
improved two–pass	103k/140k	1.2 – 0.5	5.7

4.7 Summary of 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 3.2. For VTN in recognition, we employed the improved two–pass strategy as explained in Section 4.4 on the WSJ0 and Verbmobil tasks and the full optimization as explained in Section 4.5 on the SieTill data. As Table 10 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 10: Summary: effect of VTN in training and recognition on the word error rate.

Corpus	PP	VTN	#Dens.	DEL – INS %	WER %
WSJ0	56	no	103k	0.8 – 0.5	4.9
		yes	103k/140k	0.7 – 0.5	4.1
Verbmobil	35	no	195k	3.2 – 3.1	16.7
		yes	195k/185k	3.1 – 3.1	15.9
SieTill	11	no	425	1.0 – 0.8	4.6
		yes	425	1.0 – 0.7	3.6

5 Conclusions

The main contributions of this paper are:

- The principles of VTN-based speech recognition have been outlined in the frame-work of the Bayes decision rule and of speaker adaptive modeling. This has resulted in a straightforward interpretation of the training problem in speaker adaptive modeling.

- Using single Gaussian densities for scale factor selection in training leads to distributions of scale factors that reflect typical variations of vocal tract lengths among speakers. Based on this method, we have presented a three–step procedure for the training of a normalized acoustic model.

- We have reduced the error rates obtained with the two–pass strategy by using a non-normalized instead of a normalized model in the first recognition pass.
- We have compared a fast method for VTN with the improved two–pass strategy. The fast method has reduced the error rate on the WSJ0 task from 6.8% to 6.2% compared to 5.7% obtained with the improved two–pass strategy.
- We have found that for digit string recognition the suboptimality of the two–pass strategy has reduced the possible VTN improvement in recognition performance by 30% relative.

By using the presented methods for VTN, the word error rates has been 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.

Acknowledgment

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.

The authors would like to thank Sirko Molau (RWTH Aachen) for software support.

The authors would like to thank the anonymous reviewers for helpful comments to improve the quality of the paper.

References

[1] A. Acero, X. Huang: "Speaker and gender normalization for continuous-density hidden Markov models", *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 342-345, Atlanta, GA, USA, May 1996.

[2] T. Anastasakos, J. McDonough, R. Schwarz, J. Makhoul: "A compact model for speaker-adaptive training," *Proc. Int. Conf. on Spoken Language Processing*, Vol. 2, pp. 1137-1140, Philadelphia, PA, Oct. 1996.

[3] T. Anastasakos, J. McDonough, J. Makhoul: "Speaker adaptive training: A maximum likelihood approach to speaker normalization," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 1043-1046, Munich, Germany, Apr. 1997.

[4] X. Aubert, E. Thelen: "Speaker adaptive training applied to continuous mixture density modeling," *Proc. Europ. Conf. on Speech Communication and Technology*, Vol. 4, pp. 1851-1854, Rhodes, Greece, Sep. 1997.

[5] C. Chesta, O. Siohan, C.-H. Lee: "Maximum a posteriori linear regression for hidden Markov model adaptation", *Proc. Europ. Conf. on Speech Communication and Technology*, Vol. 1, pp. 211-214, Budapest, Hungary, Sep. 1999.

[6] W. Chou: "Maximum a posterior linear regression with elliptically symmetric variate priors", *Proc. Europ. Conf. on Speech Communication and Technology*, Vol. 1, pp. 1-4, Budapest, Hungary, Sep. 1999.

[7] E. Eide, H. Gish: "A parametric approach to vocal tract length normalization," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 346-349, Atlanta, GA, May 1996.

[8] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, M. Westphal: "The Karlsruhe Verbmobil speech recognition engine," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 83-86, Munich, Germany, Apr. 1996.

[9] M. J. F. Gales: "Maximum likelihood linear transformations for HMM-based speech recognition," Computer Speech and Language, Vol. 12, pp. 75-98, 1998.

[10] Y. Gong: "Source normalization training for HMM applied to noisy telephone speech recognition", *Proc. Europ. Conf. on Speech Communication and Technology*, pp. 1555-1558, Rhodes, Greece, Sep. 1997.

[11] L. Lee, R. Rose: "Speaker normalization using efficient frequency warping procedures," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing* Vol. 1, pp. 353-356, Atlanta, GA, May 1996.

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

[13] C. J. Leggetter, P. C. Woodland: "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer, Speech and Language*, Vol. 9, No. 2, pp. 171–185, April 1995.

[14] J. McDonough, W. Byrne: "Speaker adaptation with all-pass transforms", *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 757-760, Phoenix, AZ, March 1999.

[15] H. Ney, L. Welling, S. Ortmanns, K. Beulen, F. Wessel: "The RWTH large vocabulary continuous speech recognition system," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 853-856, Seattle, WA, May 1998.

[16] D. Pye, P. C. Woodland: "Experiments in speaker normalization and adaptation for large vocabulary speech recognition," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 1047-1050, Munich, Germany, Apr. 1996.

[17] A. Sankar, C.-H. Lee: "A maximum likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 3, pp. 190-202, May 1996.

[18] H. Wakita: Normalization of vowels by vocal tract length and its application to vowel identification. IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-25, No. 2, pp. 183-192, April 1977.

[19] S. Wegmann, D. McAllaster, J. Orloff, B. Peskin: "Speaker normalization on conversational telephone speech," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 339-341, Atlanta, GA, May 1996.

[20] L. Welling, N. Haberland, H. Ney: "Acoustic front-end optimization for large vocabulary speech recognition", *Proc. Europ. Conf. on Speech Communication and Technology*, Vol. 4, pp. 2099-2102, Rhodes, Greece, Sep. 1997.

[21] L. Welling, R. Haeb–Umbach, X. Aubert, N. Haberland: "A study on speaker normalization using vocal tract normalization and speaker adaptive training," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 797-800, Seattle, WA, May 1998.

[22] L. Welling, S. Kanthak, H. Ney: "Improved methods for vocal tract normalization," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 761–764, Phoenix, AZ, April 1999.

Seperate List of Figures/Tables

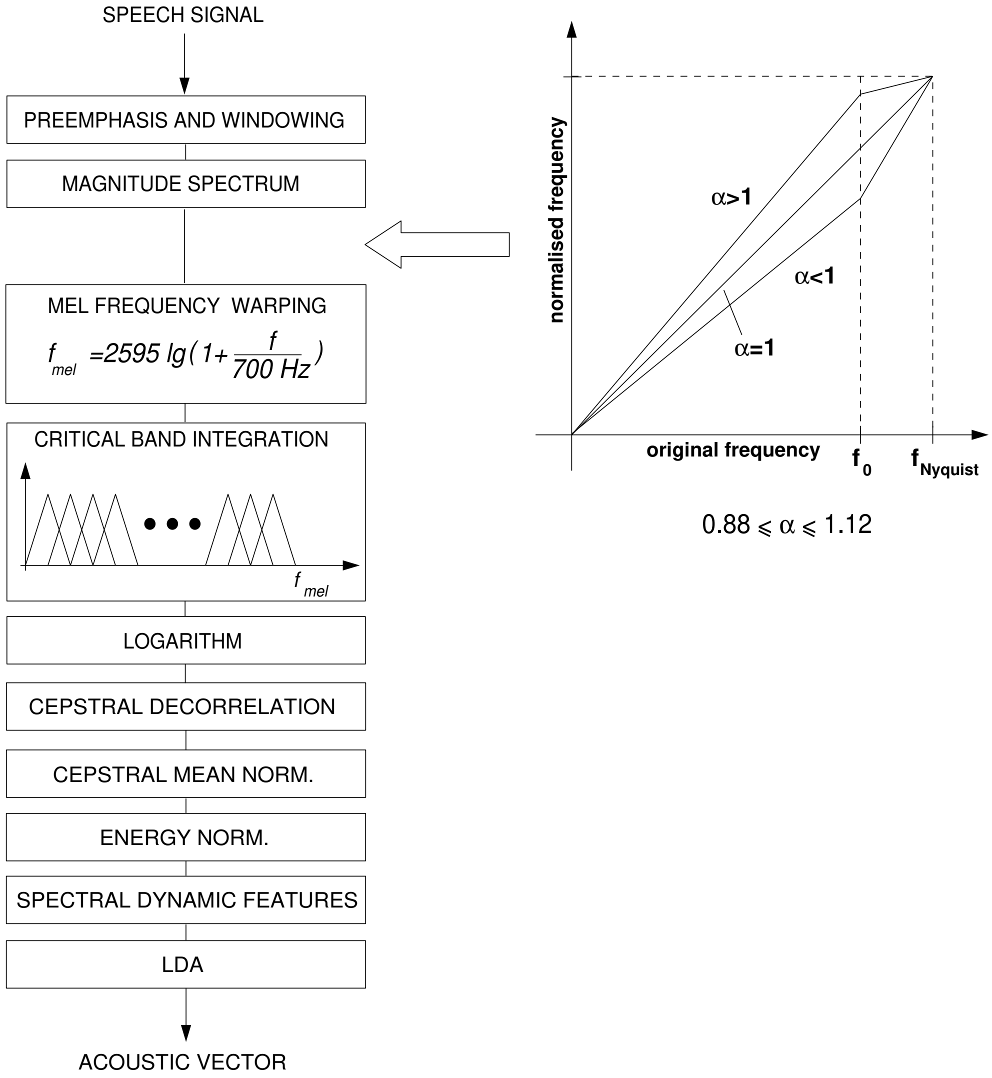


Figure 1: Acoustic front–end with VTN.

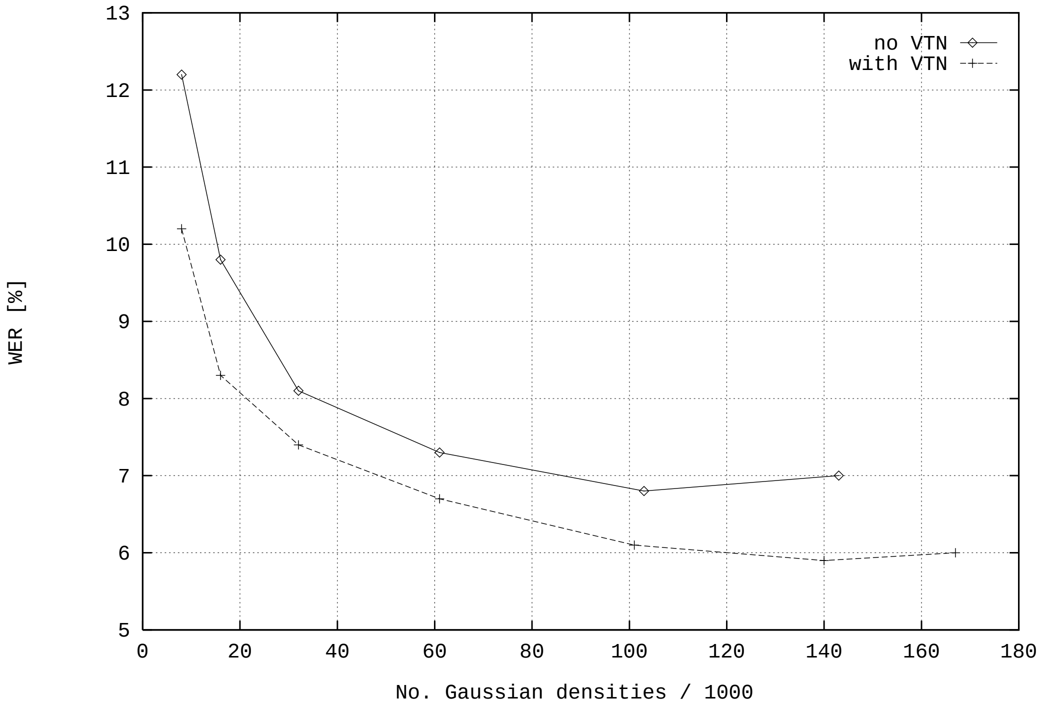
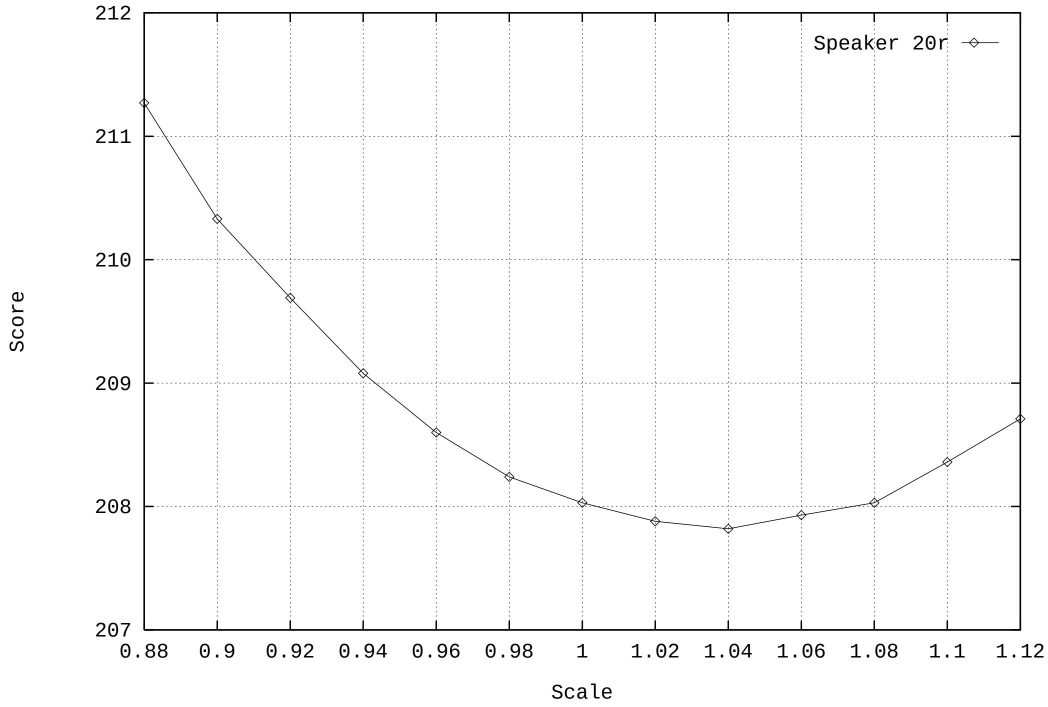


Figure 2: Word error rate as a function of the number of Gaussian component densities (corpus: WSJ0).



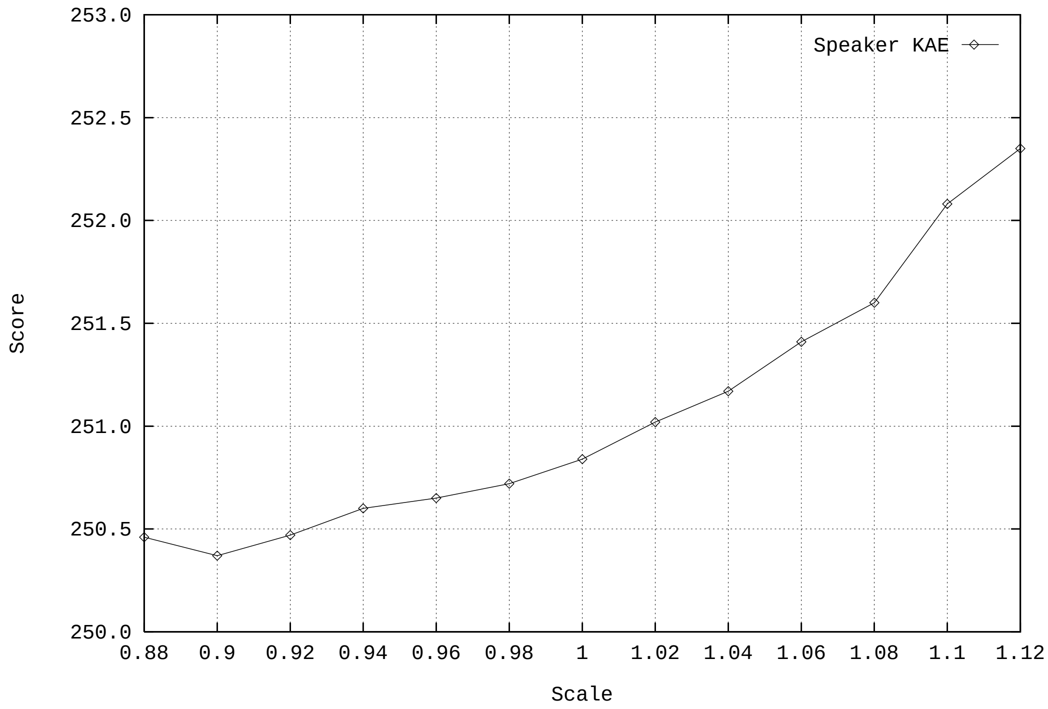


Figure 3: Score $[-\log(p(X_r^\alpha|W_r;\lambda))]$ as a function of the scale factor α (*top*: male speaker *20r* from WSJ0 training data, *bottom*: female speaker *kae* from Verbmobil training data).

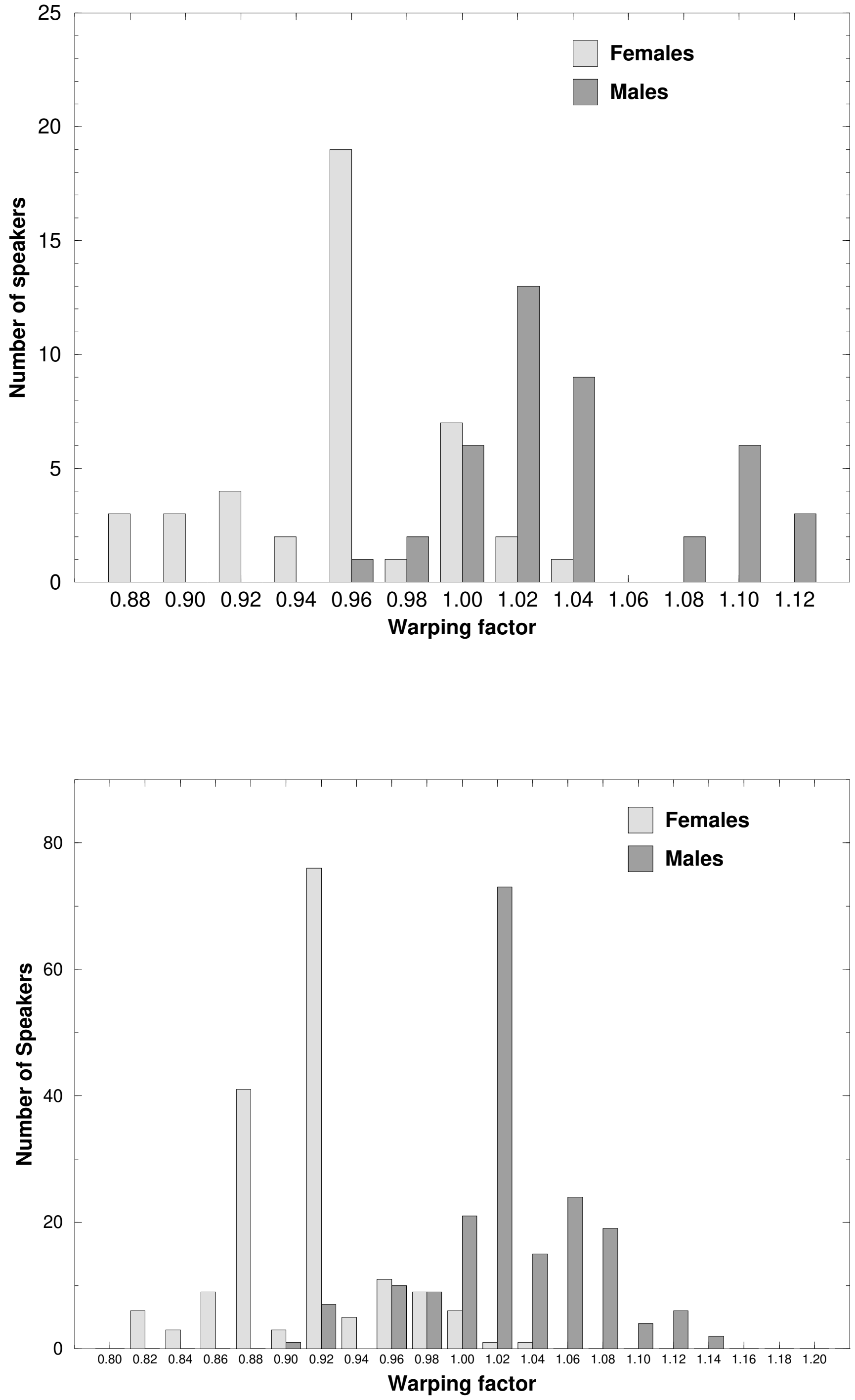


Figure 4: Histograms over scale factors for training speakers: *top* WSJ0; *bottom*: SieTill.

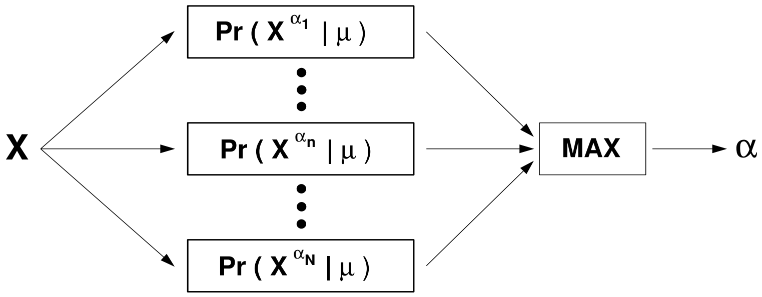


Figure 5: Fast selection of scale factor in recognition.

Table 1: Notation for acoustic models.

Adaptation	Symbol	Model structure
no	Λ	Mixture density per HMM state
	λ	Single density per HMM state
yes	$\tilde{\Lambda}$	Mixture density per HMM state
	μ	Mixture model (global)

Table 2: Statistics of speech corpora used for training and recognition experiments.

corpus		WSJ0	SieTill	Verbmobil
task		read newspaper articles	digit strings	spontaneous speech, dialogues
signal quality		clean speech	telephone speech	clean speech
sampling frequency		16 kHz	8 kHz	16 kHz
language		US English	German	German
vocabulary size		5000	11	7000
training	speakers	84	362	613
	record. hours	16	12	27
recognition	speakers	18	356	62
	sentences	740	13114	343
	words	12132	43086	6446
	record. hours	1.42	12	0.67
	language model	bigram/trigram	zerogram	trigram
	test-set perplexity	107/56	11	35

Table 3: Effect of silence on scale factor selection for VTN.

Corpus	VTN	Silence vectors	#Densities	DEL – INS %	WER %
WSJ0	no	-	103k	1.4 – 0.6	6.8
	yes	yes	140k	1.0 – 0.6	6.1
		no	140k	1.2 – 0.5	5.9
SieTill	no	-	425	1.0 – 0.8	4.6
	yes	yes	425	1.0 – 0.8	4.2
		no	425	1.0 – 0.8	4.0

Table 4: Effect of the number of densities per HMM state for scale factor selection in training on the word error rate (corpus: WSJ).

recognition method	#Dens./State for scale selection	#Dens. for recognition (cf. Fig. 2)	DEL – INS %	WER %
no VTN	-	103k	1.4 – 0.6	6.8
VTN in recognition only	-	103k	1.4 – 0.6	6.8
VTN in training + recognition	32	103k	1.4 – 0.6	6.8
	8	143k	1.2 – 0.6	6.1
	1	140k	1.2 – 0.6	5.9

Table 5: Effect of acoustic model normalization in first–pass recognition.

Corpus	Normalized Acoustic Model	#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	425	1.0 – 0.8	4.6
	yes	425	1.0 – 0.9	5.3

Table 6: Comparison of baseline and improved two–pass strategy for recognition with VTN.

Corpus	type of VTN	#Dens.	DEL – INS %	WER %
WSJ0	no	103k	1.4 – 0.6	6.8
	baseline two–pass	140k	1.2 – 0.5	5.9
	improved two–pass	103k/140k	1.2 – 0.5	5.7
SieTill	no	425	1.0 – 0.8	4.6
	baseline two–pass	425	1.0 – 0.8	4.0
	improved two–pass	425	1.0 – 0.8	3.9

Table 7: Effect of the suboptimality of the two–pass strategy for VTN in recognition on the word error rate.

Corpus	type of VTN	#Dens.	DEL – INS %	WER %
WSJ0	no	103k	1.4 – 0.6	6.8
	improved two–pass	103k/140k	1.2 – 0.5	5.7
	correct transcription	140k	1.1 – 0.5	5.7
SieTill	no	425	1.0 – 0.8	4.6
	improved two–pass	425	1.0 – 0.8	3.9
	correct transcription	425	1.0 – 0.7	3.6
	full optimization	425	1.0 – 0.7	3.6

Table 8: Effect of sentence-wise and speaker-wise estimation of scale factors α.

Corpus	#Speakers	#Sentences	Average Sentence Length [sec]	Average Sentence #Words	abs. deviation of α
WSJ0	18	740	7.0	16.4	0.0105
SieTill	356	13114	3.2	3.3	0.0292

Table 9: Comparison of fast VTN and improved two–pass strategy for recognition with VTN (corpus: WSJ).

VTN	#Dens.	DEL – INS %	WER %
no	103k	1.4 – 0.6	6.8
fast	143k	1.3 – 0.6	6.2
improved two–pass	103k/140k	1.2 – 0.5	5.7

Table 10: Summary: effect of VTN in training and recognition on the word error rate.

Corpus	PP	VTN	#Dens.	DEL – INS %	WER %
WSJ0	56	no	103k	0.8 – 0.5	4.9
		yes	103k/140k	0.7 – 0.5	4.1
Verbmobil	35	no	195k	3.2 – 3.1	16.7
		yes	195k/185k	3.1 – 3.1	15.9
SieTill	11	no	425	1.0 – 0.8	4.6
		yes	425	1.0 – 0.7	3.6