A STUDY ON SPEAKER NORMALIZATION USING VOCAL TRACT NORMALIZATION AND SPEAKER ADAPTIVE TRAINING

L. Welling [1], R. Haeb–Umbach [2], X. Aubert [2] and N. Haberland [1]

[1] RWTH Aachen – University of Technology, D-52056 Aachen, Germany
[2] Philips GmbH Forschungslaboratorien Aachen, D-52066 Aachen, Germany

ABSTRACT

Although speaker normalization is attempted in very different manners, vocal tract normalization (VTN) and speaker adaptive training (SAT) share many common properties. We show that both lead to more compact representations of the phonetically relevant variations of the training data and that both achieve improved error rate performance only if a complementary normalization or adaptation operation is conducted on the test data. Algorithms for fast test speaker enrollment are presented for both normalization methods: in the framework of SAT, a pre-transformation step is proposed, which alone, i.e. without subsequent unsupervised MLLR adaptation, reduces the error rate by almost 10% on the WSJ 5k test sets. For VTN, the use of a Gaussian mixture model makes obsolete a first recognition pass to obtain a preliminary transcription of the test utterance at hardly any loss in performance.

1. INTRODUCTION

Normalization techniques applied in the training of automatic speech recognizers aim at separating phonetically relevant variations from irrelevant variations caused e.g. by speaker particularities or the acoustic environment of the training data. The potential benefit of such an approach is twofold:

- If only the relevant variations of the training data are learnt, the resulting models are more compact, i.e. fewer parameters are required to capture the information relevant for decoding.
- If applied in training and recognition it should result in a reduction of the mismatch between training and testing conditions and therefore improve the error rate performance.

While simple and effective normalization techniques, such as cepstral mean subtraction, are widely in use, there is a growing interest in more elaborate techniques which incorporate the normalization in the maximum likelihood (ML) parameter estimation framework, e.g. [1, 2, 6, 7, 9].

In this paper, we concentrate on *speaker* normalization and investigate the commonalities and differences of two known approaches: speaker adaptive training (SAT) [2, 3, 4] and vocal tract normalization (VTN) [5, 8, 9, 10]. Although these methods rely on quite different adaptation techniques, namely, affine transformations of means for SAT and warping of the frequency axis for VTN, they share the same ML framework.

SAT integrates the adaptation technique of maximum likelihood linear regression (MLLR) in the HMM training and the parameters of both, per-speaker affine transformations and mixture densities, are jointly estimated. Piecewise linear transformations of the means have been shown to capture speaker characteristics reasonably well and when incorporated in the training process, they lead to "purer" models, as can be measured e.g. by their variance [3, 4].

However, as we will see, this only pays off in better recognition performance if a complementary action, in this case MLLR adaptation, is done during recognition. Moreover, the byproduct of SAT, an inventory of MLLR transformations related to the training speakers, can also be successfully used to speed up the enrollment of new test speakers as shown in [4]. This approach has been further refined and new results are presented.

VTN on the other hand, performs a normalization in the signal space: for each training speaker an optimal frequency warp scale is determined and the feature vectors are computed from the accordingly warped frequency axis. In contrast to SAT, a single scalar normalization parameter per speaker is applied, and its effect in the model space is highly nonlinear. But, similar to SAT, we will see that VTN models are more compact and that a complementary normalization is required on the recognition data to gain a performance benefit.

In principle, warp scale and model parameters should be estimated iteratively, just like the iterative estimation of transformation matrices and HMM parameters in SAT. This, however, is computationally very expensive. An approximation is to do just one or two iterations [8] or to split the training data in two sets, and alternately determine warp scale and model parameters on either of them [9].

For the determination of the best frequency warp scale in recognition a number of methods have been proposed [5, 9, 10]. One popular approach, the selection according to the largest likelihood of the test utterance given a hypothesized word sequence [8, 9], is again exactly the same approach as employed for fast enrollment of a test speaker in SAT [4].

Here, we propose a fast warp scale selection algorithm using a single Gaussian mixture model of speech in the normalized feature space. A similar approach based on a generic model of voiced speech has been introduced in [10] but our solution is simpler as it is applied on the whole (voiced and unvoiced) speech and the Gaussian mixture is trained *after* the optimal warping factors have been selected. Another related approach described in [9] used one mixture model for each warping factor.

The paper is organized as follows. In the next section, we investigate the normalization effect of both SAT and VTN. Section 3 is devoted to fast enrollment of test speakers, and we draw some conclusions in Section 4.

2. ON THE EFFECT OF SPEAKER NORMALIZATION

The basic idea underlying SAT is that the characteristics of each training speaker can be expressed by a set of linear transformations mapping the speaker-independent means on the speaker-specific acoustic domain, and the estimation of these transformation parameters is embedded in the mixture density HMM training [2]. This leads to a ML formulation for jointly estimating three sets of parameters:

- First, a set of MLLR transformations is estimated for adapting the speaker-independent (SI) means to each of the training speakers.
- Next, the SI means are reestimated as a weighted average of the inverse affine transformations applied to the speaker-dependent (SD) means. As shown in [4], the weights are the potential matrices of the corresponding SD distributions when transformed back to the SI acoustic space.
- Third, the SI covariances are reestimated in a similar manner as the means.

In [4], we observed that the overall variance of the SAT models (averaged over all densities and dimensions) is indeed significantly smaller than with standard SI models. Further, the log-likelihood of the training data is increased, both supporting the expectation that the SAT models should be "purer". However, this is not an advantage per se, as can be seen from Table 1. This table presents word error rates obtained over four WSJ 5k test sets, resp. the dev and eval sets of Nov'92 and '93, comprising 38 speakers and 24,630 spoken words. For the SAT experiments, the signal front-end consists of a 30 channel filter bank followed by a gender-independent LDA and training was done on the 84 speakers (m+f) of WSJ0. When MLLR adaptation is performed on the test data, it is done *unsupervised* and incrementally after each sentence (spoken by the same speaker). Starting with standard SI models (no SAT), unsupervised MLLR improves the accuracy by 16% relative, from 9.4% to 7.9%. In contrast, SAT models "as such" are about 10% worse but MLLR adaptation reduces the errors by 26%, leading to a final score of 7.7% which represents a small 2.5% improvement. It is clear that SAT provides a rather poor starting point that penalizes the unsupervised adaptation process. When combined with *supervised* MLLR adaptation, SAT has been shown to bring substantially larger benefits [2, 3, 4].

Table 1: Effects of SAT and unsupervised MLLR on word error rate (WER) for gender-independent (GI) models. WSJ 5k 92/93 dev/eval test sets, bigram lm.

MLLR applied in		#dens	del – ins	WER
train (SAT)	recog	(GI)	[%]	[%]
no	no	106k	2.2 – 0.8	9.4
no	yes	”	2.1 – 0.6	7.9
yes	no	”	2.0 – 1.2	10.4
yes	yes	”	1.8 – 0.8	7.7

VTN techniques perform a normalization in the signal space by, typically linearly, warping the frequency axis by a speaker-specific warping factor [8, 9, 10]. While, very similar to SAT, the per-speaker warping factors and the model parameters can be estimated iteratively, we employed a simplified training scheme with just one iteration:

1. An intermediate model λ with a small number of densities per state is estimated from the unwarped features of all training speakers by maximum likelihood (ML) training.
2. For each training speaker r, a warp scale α_r is chosen as the scale for which the training data of this speaker, $X_r(\alpha)$, achieve the greatest likelihood, given the transcriptions W_r and the intermediate model λ:

$$\alpha_r = \arg\max_{\alpha} Pr(X_r(\alpha)|W_r, \lambda).$$

   We used an exhaustive line search for α in the range $0.88 \leq \alpha \leq 1.12$ with step size 0.02.
3. A model Λ is trained on the warped utterances by ML training.

Table 2 again presents results on the four WSJ 5k test sets, now for a gender-independent (GI) and a gender-dependent (GD) setup. In the GD case we trained separate models on the 42 male and 42 female WSJ0 speakers, respectively, while in the GI case we trained one model set on the whole si84 training data. The baseline word error rate of 9.0% is better than in Table 1, because we employed a slightly different setup with a MFCC+LDA front-end here.

Similar to SAT, speaker normalization in training only results in worse error rate performance compared to the baseline system without VTN, in particular in the GI case. Only if VTN is also applied in recognition, a reduction in error rate can be achieved. For the results in this table, we used a preliminary transcription of the test sentence for the warping factor selection, see next section. Although the baseline error rate for a SD setup is slightly better, the results for VTN in training and recognition tend to be better

Table 2: Effects of VTN on the word error rate for gender-dependent (GD) and gender-independent (GI) models. WSJ 5k 92/93 dev/eval test sets, bigram lm.

Setup	VTN in		#dens	del – ins	WER
	train	recog	(m+f)	[%]	[%]
GD	no	no	95k+95k	1.7 – 0.9	8.9
	no	yes	”	1.7 – 1.0	8.7
	yes	no	”	1.7 – 1.1	9.1
	yes	yes	”	1.6 – 0.9	8.5
GI	no	no	150k	1.7 – 0.9	9.0
	no	yes	”	1.6 – 0.9	8.5
	yes	no	”	1.7 – 1.4	10.9
	yes	yes	”	1.5 – 0.9	8.0

in the GI case. Obviously, VTN was able to discard gender-specific variations from the training data and could beneficially exploit the larger training database. This is consistent with the experience of other researchers, e.g. [10].

The normalization by VTN results in “purer” models, as can be seen from Table 3. The same recognition performance can now be obtained with a factor of two to three fewer model parameters, compared to a baseline system without VTN.

In order to verify the error rate reduction on a different corpus, we conducted experiments on the German SIETILL telephone digit string database. This corpus consists of 362 training speakers (42860 digits) and 356 testing speakers (43095 digits) which represent a large variety of line and speaker characteristics. We used a recognizer with continuous HMMs, *single* Gaussian emission probabilities, whole-word models, MFCC front-end and cepstral mean subtraction. Table 4 shows the drastic performance improvement if VTN is used in training and recognition. Our interpretation is that such a large improvement was obtained since the recognizer used *single* Gaussian emission probabilities. In our view, the similarity of the error rates for GD and GI models is due to the large size of the training corpus for each gender. This experiment again supports our conclusion that VTN provides a means to overcome the need for gender-dependent acoustic models.

Table 3: Word error rates as a function of the number of model parameters (densities). WSJ 5k 92/93 dev/eval sets, GI models, bigram lm.

#dens	no VTN		VTN in train+recog	
	del – ins [%]	WER [%]	del – ins [%]	WER [%]
150k	1.7 – 1.0	9.0	1.5 – 0.9	8.0
95k	1.8 – 1.0	9.3	1.7 – 0.9	8.3
55k	2.1 – 1.0	9.9	1.8 – 0.9	8.7
30k	2.3 – 1.0	10.9	2.0 – 0.9	9.4
16k	2.7 – 1.0	12.1	2.3 – 0.8	10.4
8k	3.2 – 1.1	14.0	2.9 – 1.0	12.3

Table 4: Effect of VTN on word error rates on SIETILL telephone digit string corpus for GD and GI models.

Models	VTN in		del – ins	WER
	train	recog	[%]	[%]
GD, single dens.	no	no	0.6 – 0.7	5.6
	no	yes	0.6 – 0.7	5.0
	yes	yes	0.4 – 0.5	2.9
GI, single dens.	no	no	0.6 – 1.1	7.5
	no	yes	0.5 – 1.0	5.9
	yes	yes	0.4 – 0.6	3.0

3. FAST SPEAKER ENROLLMENT

When adaptation proceeds *unsupervised*, MLLR is affected by the recognition errors especially when multiple regressions are considered and this constitutes an handicap to SAT. In contrast, VTN appears quite robust against script errors, presumably due to its single parameter. Therefore, in [4] we proposed to utilize the MLLR transformation matrices estimated during SAT to obtain better initial models for a new test speaker. This results in an effective pre-transformation step taking advantage of the training speakers who appear similar to the new one. This approach has been further improved by allowing for linear combinations of transformation matrices among the selected training speakers. Given the first unknown utterance of a new speaker, the following algorithm has been used:

1. Decode using the non-adapted SAT models.
2. For each speaker considered during training:
   - Transform the SI means to the speaker-specific means.
   - Compute the likelihood of the hypothesized word sequence.
3. Select the N transformation matrices yielding the highest likelihoods.
4. Form a new *global* transformation matrix as a linear combination of the N selected and transform the SI means using this new transformation.
5. Carry out the final decoding of the test utterance.

To be more robust against decoding errors, a single matrix is worked out based on one utterance of the new speaker. Results are summarised in Table 5.

Combined with SAT models, the pre-transformation step alone reduces the error rate by almost 20% for $N = 7$ and achieves a 10% improvement versus standard SI models. When further combined with unsupervised incremental MLLR, a final score of 7.4% is obtained, representing a relative gain of 6% with respect to the figure of 7.9% achieved by adapting the standard SI models with MLLR (Table 1).

This 2-pass decoding scheme is conceptually very similar to standard VTN implementations [8, 9] which we also applied to obtain the VTN results of the last section.

Table 5: Effects on word error rate of SAT combined with pre-transformation step and unsupervised MLLR. Results on WSJ 5k 92/93 dev/eval test sets, GI models, bigram lm.

MLLR applied in			del – ins	WER
train	pre-transf.	recog	[%]	[%]
no	no	no	2.2 – 0.8	9.4
yes	no	no	2.0 – 1.2	10.4
yes	yes ($N = 1$)	no	1.9 – 0.8	8.8
yes	yes ($N = 7$)	no	1.9 – 0.8	8.5
yes	yes ($N = 7$)	yes	1.8 – 0.7	7.4

In the following, we present a fast selection algorithm for the warping factor in VTN which does not require a first recognition pass to obtain a preliminary transcription. The method is based on a Gaussian mixture model that represents the distribution of the normalized feature vectors.
After the training data have been warped as explained in the previous section, a Gaussian mixture model M is trained on the warped data by employing the LBG algorithm and the ML criterion.
During recognition, the warp scale is selected using the Gaussian model M as follows:

- Generate warped features $X(\alpha)$ for all warp scales α.
- Select warp scale $\hat{\alpha}$: $\hat{\alpha} = \arg\max_{\alpha} Pr(X(\alpha)|M)$.
- Decode the sentence using the features $X(\hat{\alpha})$.

In our tests, the Gaussian mixture model had a single diagonal covariance matrix and 64 component densities. The selection of the warping factor was done on speech excluding silence in training and recognition.
We compared this fast selection method with the standard 2-pass VTN on the WSJ0 5k Nov'92 dev and eval test sets, comprising 18 speakers and 12132 spoken words. By using the fast selection method, the decoding time is reduced by a factor of 2 with only an increase of 2% relative in the word error rate, see Table 6.

Table 6: Effect of fast warping factor selection on error rate and relative real time factor (RTF). Results on WSJ 5k 92 dev/eval test sets, GI models, bigram lm.

VTN in		#dens	del – ins	WER	rel.
train.	recog.	(GI)	[%]	[%]	RTF
no	no	122k	1.4 – 0.7	7.0	1.00
yes	2-pass	143k	1.2 – 0.6	6.1	2.05
yes	fast select.	"	1.3 – 0.6	6.2	1.05

4. SUMMARY

The effects of speaker normalization by using VTN and SAT were studied. Both methods lead to more compact acoustic models: we showed that by using VTN the same error rates are obtained with significantly less acoustic model parameters compared to a system with no VTN, and that VTN provides a means to overcome the need for gender-dependent acoustic models. We showed for both VTN and SAT that if normalized acoustic models are used, a complementary normalization step has to be carried out during recognition.

Furthermore, methods for fast test speaker enrollment were presented for both SAT and VTN: for SAT, a pre-transformation step based on a linear combination of MLLR matrices obtained in the training phase was investigated. An error rate reduction of 10% relative versus standard SI models was obtained. Using unsupervised incremental MLLR, the improvement with respect to adapting the standard SI models was 6% relative. For VTN, we selected the warp scale using a Gaussian mixture model. On a WSJ 5k task, this method was shown to give similar error rates as the common VTN method which requires 2 recognition passes.

5. REFERENCES

[1] A. Acero, X. Huang, "Speaker and Gender Normalization for Continuous-Density Hidden Markov Models", in *Proc. ICASSP*, Vol. 1, pp 342-345, Atlanta, GA, USA, 1996.

[2] T. Anastasakos, J. McDonough, R. Schwarz, J. Makhoul, "A Compact Model for Speaker-Adaptive Training," in *Proc. ICASSP*, 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," in *Proc. ICASSP*, pp.1043-1046, Munich, Germany, Apr. 1997.

[4] X. Aubert, E. Thelen, "Speaker Adaptive Training Applied to Continuous Mixture Density Modelling," in *Proc. EUROSPEECH* Vol. 4, pp. 1851-1854, Rhodes, Greece, Sep. 1997.

[5] E. Eiden, H. Gish, "A Parametric Approach to Vocal Tract Length Normalization," in *Proc. ICASSP*, Vol. 1, pp. 346-349, Atlanta, GA, May 1996.

[6] Y. Gong, "Source Normalization Training for HMM Applied to Noisy Telephone Speech Recognition", in *Proc. EUROSPEECH*, pp 1555-1558, Rhodes, Greece, 1997.

[7] A. Sankar, C.-H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition," in *IEEE T-SAP*, Vol. 4, No. 3, pp. 190-202, May 1996.

[8] 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. 1996.

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

[10] 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.