SPEAKER–INDEPENDENT UPFRONT DIALECT ADAPTATION IN A LARGE VOCABULARY CONTINUOUS SPEECH RECOGNIZER

V. Fischer[1], *Y. Gao*[2], *E. Janke*[3]

[1]IBM Speech Systems, European Speech Research,
Vangerowstr. 18, D-69115 Heidelberg, F.R. of Germany
[2]IBM T. J. Watson Research Center,
P.O. Box 218, Yorktown Heights, NY 10598, USA
[3]IBM United Kingdom Laboratories Ltd., Hursley Park,
Winchester, Hants SO21 2JN, United Kingdom

ABSTRACT

Large vocabulary continuous speech recognition systems show a significant decrease in performance if a users pronunciation differs largely from those observed during system training. This can be considered as the main reason why most commercially available systems recommend — if not enforce — the individual end user to read an enrollment script for the speaker dependent reestimation of acoustic model parameters. Thus, the improvement of recognition rates for dialect speakers is an important issue both with respect to a broader acceptance and a more convenient or natural use of such systems.

This paper compares different techniques that aim on a better speaker independent recognition of dialect speech in a large vocabulary continuous speech recognizer. The methods discussed comprise Bayesian adaptation and speaker clustering techniques and deal with both the availability and absence of dialect training material. Results are given for a case study that aims on the improvement of a German speech recognizer for Austrian speakers.

1. INTRODUCTION

With the appearance of large vocabulary continuous speech recognition systems (LVCSRS) users are no longer forced to insert short pauses between words, but still have to face a significant loss in recognition accuracy, if their pronunciation differs largely from those observed during system training. Therefore, such systems may be unusable for dialect speakers, unless they are forced to speak in an inconvenient manner, or at least require a speaker dependent reestimation of acoustic model parameters.

The use of dialect affected speech for the training procedure of a Hidden Markov Model based speech recognizer is one way to overcome these limitations, but needs the collection of a substantial amount of data for a reliable estimation of the model parameters. In contrast, adding even a limited amount of dialect affected speech to a large portion of "clean" training data may result in a lower recognition rate for speakers that use standard pronunciation.

This paper compares a variety of techniques that can deal with the tradeoff outlined above, both in case of availability or absence of dialect training data. Results are given that were obtained in a study that aimed on the improvement of a German LVCSRS for Austrian speakers. Section 2 gives a brief outline of both the training procedure and the baseline recognition system. Section 3 applies pre-clustering of training speakers, which is appropriate if no additional dialect data is available for the training of the acoustic models. Section 4 compares different methods that can be applied if additional training data is available: the creation of an acoustic model for dialect speech, the training of a common recognizer for both dialect and clean speech, and the use of Austrian training data for the fast adaptation of a German recognizer. Finally, Section 5 gives a conclusion and an outlook on further work.

2. SYSTEM DESCRIPTION

The basic ideas underlying the LVCSRS used here are described in some detail in [2, 1]. The training of the system is a bootstrap procedure that assumes the availability of an initial speaker independent system. In a first step cepstral features and their first and second order derivatives are computed and viterbi aligned against the transcription of the training data.

For the training of context-dependent, subphonetic HMMs, phonetic contexts are extracted by passing the feature vectors through a decision tree. The data at each leaf of a tree is clustered, and described by a mixture of Gaussian densities components with diagonal covariance matrices. The so created models are refined by running a few iterations of the well known forward-backward training algorithm (see e.g. [9]). The total number of both context dependent HMMs and Gaussians is limited by the specification of an upper bound and depends on the amount and contents of the traininig data.

In the experiments described below we use a combination of a word based trigram language model and a class based trigram model that yields a significant improvement compared to the word based model alone in preliminary experiments (not described here). The class based model is automatically created by a combinatorial optimization procedure [5] that relies on the use of uni- and bigram information and a small set of regular expressions (approx. 50) for the encoding of rudimentary morphological information, like e.g. the endings of inflected words. The algorithm starts with a random assignment of words into a fixed number of classes, and maximizes the similarity between words and their possible classes.

3. DIALECT ADAPTATION WITHOUT DIALECT DATA

Speaker clustering techniques have been applied successfully to create acoustic models for certain speaker types, like e.g. male/female or fast/slow speakers [7, 6]. If no additional dialect training data is available one might try to find a subset of the “clean” training data that is best suited to create an acoustic model for the dialect test speakers. The algorithm used here employs a preclustering of the training speakers that comprises the following steps [3]:

1. Partitioning of the set of training speakers into clusters of acoustically similar speakers,
2. training of an acoustic model for each cluster,
3. selection of a model that is best suited for the decoding of a test speaker's utterances, and
4. adaptation of the model to a test speaker's acoustic space.

In the first step the characteristics of a speaker is defined by the speaker dependent mean and variances of 186 allophonic HMMs. These are obtained from a viterbi alignment of each test speakers utterances against their transcription. The similarity between any two speakers is measured by the sum of Gaussian log likelihoods of corresponding allophones

$$\log P_i \quad = \quad -c_i \left[\frac{n}{2} \log(2\pi) + \frac{1}{2} |\underline{\underline{\Gamma}}_i| \right], \tag{1}$$

where c_i is the merged E-M count of the i-th allophone, $\underline{\underline{\Gamma}}_i$ is the variance of the i-th merged Gaussian, and n is the dimension. In a bottom up clustering procedure Eqn. 1 is employed to merge the speaker dependent allophonic Gaussians until the desired number of clusters is obtained.

The second step comprises the computation of an acoustic model from the training data of those speakers who belong to a given cluster. Since therefore only a subset of the complete training data is available for the estimation of the cluster dependent model parameters, a speaker indenpendent model is used for Bayesian smoothing according to Eqn. (3) – (5), see Section 4.

The third step computes the characteristics of each test speaker on a small amount of held out data and evaluates the Euclidean distance to each cluster. The cluster with the smallest distance is choosen for the decoding of the test speaker's data. For a further improvement of recognition results the cluster dependent HMM parameters can be moved towards the particular test speaker's acoustic model by use of MLLR adaptation [8]. However, since we do not want to intermix effects from speaker adaptation and dialect handling in this study, this step is ommited in the remainder.

		No. of clusters			
Test spkr	1	8	6	4	2
Gr	13.16	13.04	13.08	12.69	12.99
At	20.10	18.06	17.85	18.80	18.99

Table 1: Speaker independent error rates for dialect adaptation by speaker clustering.

Table 1 showes results for both the recognition of ”clean” and dialect affected speech for different numbers of speaker clusters. The first column (1 cluster) gives the error rates for the baseline recognizer, that are averaged over 20 German (Gr) and 20 Austrian (At) test speakers (10 female, 10 male) who read the same script. In the training procedure slightly more than 30000 Gaussian mixture components were estimated and approx. 2000 context dependent HMMs were trained from 90 hours of ”clean” speech that was read by 700 native German speakers.

For the Austrian test speakers the improvement is 11.19 percent, if 6 clusters are used. In contrast, for the German speakers the improvement is smaller (3.57 percent for 4 clusters), but for both groups of speakers the improvement does not merely rely on a gender-based splitting of the training corpus, which can be observed if the number of clusters is limited to two.

4. DIALECT ADAPTATION USING DIALECT TRAINING DATA

Whereas speaker clustering can be applied even in the absence of dialect training data, the availability of a limited amount of dialect affected speech allows for various methods that can improve the performance of a LVCSRS for dialect speakers:

- the training of an acoustic model from dialect data only, using the procedure outlined in Section 2,
- the incorporation of both dialect data and clean speech into the training procedure, and
- the fast adaptation of an already existing, “clean” acoustic model to the characteristics of dialect speakers.

The first approach requires a large amount of dialect data for the proper training of the HMMs, and is expected to result in a specialized recognizer for dialect speech. In contrast, the other two options may be used if less dialect training data is available, because parameters are estimated from a larger, common set of data. Moreover, these methods seem suitable

	Recognizer		
Test spkr	GrGr	GrAt	AtAt
German	13.16	13.72	22.27
Austrian	20.10	15.61	12.24

Table 2: Error rates for incorporation of dialect data into the training procedure.

		smoothing factor k		
Test spkr	GrGr	1	50	500
German	13.16	17.10	16.54	15.62
Austrian	20.10	14.18	14.08	14.36

Table 3: Error rates for dialect adaptation by Bayesian smoothing.

to achieve a good compromise, i.e. a single recognizer that works well for the different groups of speakers.

Table 2 compares the baseline system (GrGr) and two recognizers that were trained with additional data from the austrian speakers (GrAt) as well as with dialect data alone (AtAt). In all experiments we used approx. 15 hours of speech that was collected from 100 Austrian training speakers (50 female and 50 male). By the specification of an upper bound for the number of context dependent HMMs and the number of Gaussian mixture components, it was assured that each recognizer has approximately the same dimension.

Whereas the training of an acoustic system with dialect data alone results in a specialized recognizer for the Austrian dialect, the acoustic system trained with both sets of data shows no substantial degradation for the German speakers (4.2 percent), but results in a 22.3 percent relative improvement for the Austrian speakers.

The fast adaptation approach bears some interest, because it avoids the time consuming training procedure. For that purpose, the forward-backward algorithm is used to create EM-counts

$$c_i = \sum_t c_i(t), \tag{2}$$

where $c_i(t)$ is the a posteriori probability of the i-th Gaussian at time t, computed from all observed dialect data $\underline{x}_t$. The means $\underline{\mu}_i^{si}$, variances $\underline{\underline{\Gamma}}_i^{si}$, and mixture component weights ω_i^{si} of the baseline system are used to reestimate the parameters $\underline{\mu}_i^d$, $\underline{\underline{\Gamma}}_i^d$, ω_i^d, $i = 1, \ldots, N$, of the adapted system by Bayesian smoothing and tying [4] according to the following equations:

$$\underline{\mu}_i^d = \frac{\sum_t c_i(t)\underline{x}_t + \alpha_i \underline{\mu}_i^{si}}{c_i + \alpha_i} \tag{3}$$

$$\begin{aligned} \underline{\underline{\Gamma}}_i^d &= \frac{\underline{\underline{\Upsilon}}_i + \alpha_i(\underline{\underline{\Gamma}}_i^{si} + \underline{\mu}_i^{si}\underline{\mu}_i^{si,T})}{c_i + \alpha_i} - \underline{\mu}_i^d \underline{\mu}_i^{d,T}, \\ \underline{\underline{\Upsilon}}_i &= \sum_t c_i(t)\underline{x}_t\underline{x}_t^T \end{aligned} \tag{4}$$

$$\omega_i^d = \frac{c_i + \alpha_i}{\sum_{l \in L}(c_l + \alpha_l)}, \; \alpha_j = k \cdot \omega_j^{si} \tag{5}$$

Here, N denotes the total number of mixture components, L denotes the set of Gaussians that belong to the same leaf as the i-th Gaussian, and k is a constant, the *smoothing factor*.

Table 3 gives results for different values of the smoothing factor. The error rate for the German speakers increases significantly if little weight is given to the seed HMM parameters (k is small), and approaches the baseline error rate for a more rigid smoothing. In contrast, the error rate for Austrian speakers is lowered by 30.0 percent for $k = 50$, but does not improve further for smaller values of k. However, given the 20.4 percent degradation for the German speakers, it becomes evident, that this approach is well suited for the fast creation of a acoustic system for dialect speakers, but is not feasible, if a single acoustic system for both groups of speakers is needed.

5. DISCUSSION AND FUTURE WORK

This paper compared different methods for an improved recognition of dialect affected speech. As a case study, we dealt with the recognition of Austrian speakers by a German LVCSRS.

If no dialect affected speech data is available for the training of the acoustic system, preclustering of the training speakers turned out to be appropriate. For this method the speaker independent error rate for the Austrian test speakers decreased by 11.2 percent, without lowering the error rate for the German speakers.

Since — in the case of 8 clusters — for 9 out of 40 test speakers the selected cluster did not produce the best decoding results we performed an additional cross-decoding experiment. The results for an ideal cluster identifaction algorithm are given in Table 4 and suggest that more work in this direction is neccessary. Further experiments (not reported here) show that an additional benefit can be obtained from the use of gender based seed models for Bayesian smoothing in Eqn. (3) – (5).

Clearly, the system performance for Austrian speakers benefits from the availability of additional training material. Here, the training of a common acoustic system yielded a 22.4 percent improvement for the Austrian speakers, and resulted in an acceptable small degradation for the German speakers. Thus, we consider this method as a good compromise and have recently applied speaker clustering to this

		No. of clusters			
Test spkr	1	8	6	4	2
German	13.16	11.96	12.10	12.24	12.99
Austrian	20.10	16.93	17.03	18.24	18.99

Table 4: Speaker independent error rates for an ideal cluster selection mechanism.

acoustic system to achieve a further improvement. Preliminary experiments indicate that a reliable selection of a dialect dependend cluster can be achieved by a hierarchical voting mechanism.

Finally, one might think of the introduction of additional, dialect specific baseforms both for the training of the acoustic models and the recognition procedure.

Acknowledgement We would like to thank our colleagues in the IBM European Speech Research Group (located in Seville, Rome, Paris, Hursley, and Heidelberg) and in the IBM Human Language Technology Research Group (Thomas J. Watson Research Center, Yorktown Heights) for many valuable suggestions and the continuous exchange of ideas.

6. REFERENCES

1. L. Bahl, S. Balakrishnan-Aiyer, J. Bellegarda, M. Franz, P. Gopalakrishnan, D. Nahamoo, M. Novak, M. Padmanabhan, M. Picheny, and S. Roukos. Performance of the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task. In *Proc. of the IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, pages 41–44, Detroit, 1995.
2. L. Bahl, P. de Souza, P. Gopalakrishnan, D. Nahamoo, and M. Picheny. Context-dependent vector quantization for continuous speech recognition. In *Proc. of the IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, 1993.
3. Y. Gao, M. Padmanabhan, and M. Picheny. Speaker adaptation based on pre-clustering training speakers. In *Proc. of the 5th European Conference on Speech Communication and Technology*, pages 2091–2094, Rhodes, Greece., 1997.
4. J. Gauvain and C. Lee. Maximum a posteriori estimation of multivariate gaussian mixture observations of markov chains. *IEEE Trans. on Speech and Audio Processing*, 2(2):291–298, 1994.
5. S. Geissler. Improving a speech recognition system by unsupervised classification of large vocabularies. Technical report, IBM Deutschland Informationssysteme GmbH, Institute for Logics and Linguistics, 1997.
6. T. Hazen and J. Glass. A Comparison of novel techniques for instantaneous speaker adaptation. In *Proc. of the 5th European Conference on Speech Communication and Technology*, pages 2047–2050, Rhodes, Greece, 1997.
7. T. Kosaka, S. Matsunaga, and S. Sagayama. Tree-structured speaker clustering for speaker-independent continuous speech recognition. In *Proc. of the 3rd Int. Conf. on Spoken Language Processing*, pages 1375–1378, Yokohama, 1994.
8. C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9:171–185, 1995.
9. L. Rabiner. *Fundamentals of Speech Recognition.* Signal Processing Series. Prentice Hall, Inc., Englewood Cliffs, NJ, 1993.