DEVELOPMENT OF DIALECT-SPECIFIC SPEECH RECOGNIZERS USING ADAPTATION METHODS

V. Diakoloukas V. Digalakis

Dept. of Electronics & Computer Engineering
Technical University of Crete
73100 Chania, Crete, GREECE
{vdiak,vas}@telecom.tuc.gr

L. Neumeyer

STAR Laboratory
SRI International
Menlo Park, CA, USA
leo@speech.sri.com

J. Kaja

Telia Research AB
S-13680 Haninge, SWEDEN
jaan.p.kaja@telia.se

ABSTRACT

Several adaptation approaches have been proposed in an effort to improve the speech recognition performance in mismatched conditions. However, the application of these approaches had been mostly constrained to the speaker or channel adaptation tasks. In this paper, we first investigate the effect of mismatched dialects between training and testing speakers in an Automatic Speech Recognition (ASR) system. We find that a mismatch in dialects significantly influences the recognition accuracy. Consequently, we apply several adaptation approaches to develop a dialect-specific recognition system using a dialect-dependent system trained on a different dialect and a small number of training sentences from the target dialect. We show that adaptation improves recognition performance dramatically with small amounts of training sentences. We further show that, although the recognition performance of traditionally trained systems highly degrades as we decrease the number of training speakers, the performance of adapted systems is not influenced so much.

1. INTRODUCTION

A wide variety of techniques have been proposed to perform large vocabulary, continuous speech recognition. However, the recognition accuracy of these systems has proven to be highly related to the correlation of the training and testing conditions. Performance degrades dramatically if a mismatch exists between these conditions, such as different channel, accent or speaker's voice characteristics.

Several speaker adaptation techniques have been recently proposed to improve the performance and robustness of speech recognition systems. These techniques include transformation based adaptation in the feature or the model space [1, 2, 3, 4], Bayesian adaptation [5, 6], or combined approaches [7].

In this paper, we consider the dialect issue on a speaker-independent (SI) speech recognition system. Based on the Swedish language corpus collected by Telia, we investigate the development of a Swedish multi-dialect SI speech recognition system which will require only a small amount of dialect-dependent data. This recognizer is part of a bidirectional speech translation system between English and Swedish that has been developed under the SRI-Telia Research Spoken Language Translator project [8]. We first investigate the effect of mismatched conditions in training and testing, and we find that the recognition performance of a speaker-independent system trained on a large amount of training data from the Stockholm dialect decreases dramatically when tested on speakers of another Swedish dialect, namely from the Scania region.

To improve the performance of the SI system for speakers of dialects for which minimal amounts of training data are available, we use *dialect adaptation* techniques. We apply both maximum likelihood (ML) transformation based approaches, as well as combined transformation-Bayesian approaches, in an effort to minimize the effect of different dialects.

2. DIALECT ADAPTATION METHODS

The SI speech recognition system for a specific dialect is modeled with continuous mixture-density hidden Markov models (HMM's) that use a large number of Gaussian mixtures [9]. The component mixtures of each Gaussian codebook (*genone*) are shared across clusters of HMM states, and hence the observation densities of the vector process y_t have the form:

$$P_{SI}(y_t|s_t) = \sum_{i=1}^{N_\omega} p(\omega_i|s_t) N(y_t; m_{ig}, S_{ig}),$$

where g is the genone index used by the HMM state s_t.

These models need large amounts of training data for robust estimation of their parameters. Since the amount of available training data for some dialects of our database is small, the development of dialect-specific SI models is not a robust solution. Alternatively, an initial SI recognition system trained on some *seed* dialects can be adapted to match a specific *target* dialect, in which case the adapted system utilizes knowledge obtained from the seed dialects. We choose to apply algorithms that we have previously developed and applied to the problem of speaker adaptation, since in our problem there are consistent differences in the pronunciation between the different dialects that we examine. The adaptation process is performed by jointly transforming all the Gaussians of each genone, and by combining transformation and Bayesian techniques. In the next two sections we describe the adaptation methods that we examined in this study.

2.1. Transformation based adaptation

In analogy with [1], we assume that for a given HMM state s_t the target-dialect SI vector process $[x_t]$ can be calculated by the corresponding process of the seed-dialect $[y_t]$ through the linear transformation

$$x_t = A_g y_t + b_g. \quad (1)$$

Therefore the observation densities of the dialect-adapted (DA) models can be written:

$$P_{DA}(x_t|s_t) = \sum_{i=1}^{N_\omega} p(\omega_i|s_t) N(x_t; A_g m_{ig} + b_g, A_g S_{ig} A_g^t). \quad (2)$$

In order to fully define our problem the parameters $A_g, b_g, g = 1, \ldots, N_g$ have to be estimated. N_g denotes the number of transformations for the whole set of genones. The parameter estimation process is performed using the EM algorithm [10]. In our experiments we consider two variations of the generic transformation above. In the first variation (method I), we assume the matrix A_g is diagonal [1], and is applied to both the means and covariances of the models, as in equation (2).

The second method (method II, [3, 2]) assumes that A_g is a block diagonal matrix which transforms only the means of the Gaussian distributions:

$$P_{DA}(x_t|s_t) = \sum_{i=1}^{N_\omega} p(\omega_i|s_t) N(x_t; A_g m_{ig} + b_g, S_{ig}). \quad (3)$$

Each of the three blocks of this matrix performs a separate transformation to every basic feature vector (cepstrum, and its first and second derivatives). Since the transformation is only applied to the means of the Gaussians, there is no underlying transformation of the form (1) in the feature space, as for method I. For the speaker adaptation problem, it was shown in [2] that method II with a block diagonal matrix significantly outperformed both method II with a full matrix and method I with a diagonal matrix.

2.2. Combined transformation and Bayesian approaches

Bayesian techniques use prior knowledge together with the small amount of training data to adapt the system. These techniques have several useful properties, such as asymptotic convergence and text independence. However, they suffer from slow adaptation rates. By combining the Bayesian with the transformation based approach, we expect to achieve faster adaptation as well as better convergence to the dialect-specific models as the number of training sentences increases. In order to implement the combined approach, we first adapt the SI models to match the new dialect using a transformation method. Then, these dialect adapted models serve as prior knowledge for the Bayesian adaptation step. For a more detailed description of how the combination is performed, the reader is referred to [7].

3. EXPERIMENTS

The adaptation experiments were carried out using a multi-dialect Swedish speech database collected by Telia. The core of the database was recorded in Stockholm using more than 100 speakers. Several other dialects are currently being recorded across Sweden. The corpus consists of subjects reading various prompts organized in sections. The sections include a set of phonetically balanced common sentences for all the speakers, a set of sentences translated from the English Air Travel Information System (ATIS) domain, and a set of newspaper sentences.

For our dialect adaptation experiments we used data from the Stockholm and Scanian dialects, that were, respectively, the seed and target dialects. The Scanian dialect was chosen for the initial experiments because it is one of three that are clearly different from the Stockholm dialect. The main differences between the dialects is that the long (tense) vowels become diphthongs in the Scanian dialect, and that the usual supra-dental /r/-sound becomes uvular. In the Stockholm dialect, a combination of /r/ with one of the dental consonants /n/, /d/, /t/, /s/ or /l/, results in supradentalization of these consonants and a deletion of the /r/. In the Scanian dialect, since the /r/-sound is different, this does not happen. There are also prosodic differences.

In addition, the Scanian dialect can be divided into 4 distinct areas (subdialects), namely Malmö, Helsingborg, Trelleborg and Kristianstad. In our experiments, the training and test sets consist of sentences chosen equally from the above subdialects in order to create a generic, subdialect-independent system. There is a total of 40 speakers of the Scanian dialect, both male and female, and each of them recorded more than 40 sentences. We selected 8 of the speakers (half of them male) to serve as testing data, and the rest composed the adaptation/training data with a total of 3814 sentences. Experiments were carried out using SRI's $DECIPHER^{TM}$ system [9]. The system's front-end was configured to output 12 cepstral coefficients, cepstral energy and their first and second derivatives. The cepstral features are computed with a fast Fourier transform (FFT) filterbank and subsequent cepstral-mean normalization on a sentence basis is performed. We used genonic HMM's with arbitrary degree of Gaussian sharing across different HMM states [9].

The SI continuous HMM system, which served as seed models for our adaptation scheme, was trained on approximately 21000 sentences of Stockholm dialect. The recognizer is configured so that it runs in real time on a Sun Sparc Ultra-1 workstation. The system's recognition performance on an air travel information task similar to the English ATIS one was benchmarked at a 8.9% word-error rate using a bigram language model when tested on Stockholm speakers. On the other hand, its performance degraded significantly when tested on the Scanian-dialect testing set, reaching a word-error rate of 25.08%. The degradation in performance was uniform across the various speakers in the test set (see Table 1), suggesting that there may be consistent differences across the two dialects. Hence, there is a great potential for improvement through dialect adaptation.

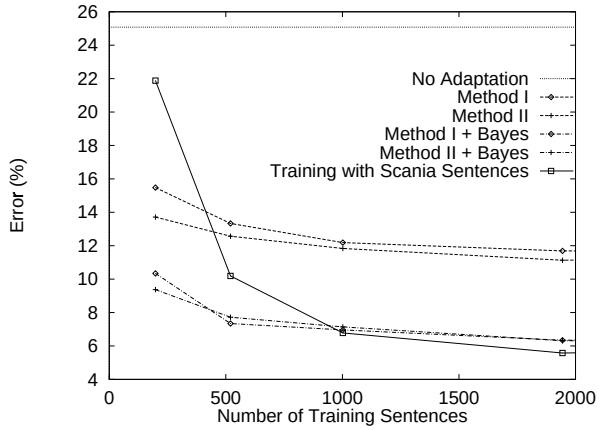

Figure 1. Dialect adaptation results for adaptation methods I, II, their combination with Bayes and standard ML training.

In the first set of experiments, we adapted the Stockholm-dialect system using various amounts of adaptation data from the training speakers of the Scanian dialect, and evaluated the performance of the adapted system to a separate set of testing speakers. This gives us a measure of the dialect-adapted, speaker-independent performance, since the adaptation and testing sets consist of different speakers. We also trained from scratch a Scania-dialect system using standard ML training based on the same adaptation data (ML-trained system), in order to estimate the adaptation benefits.

The results are summarized in Figure 1. We see that even with the first simplified algorithm, which does not take full advantage of large amounts of training data, we get a significant improvement in the performance. With as few as 198 sentences we get a 38% reduction and the word-error rate drops to almost 15%. Method II produces even better results, and the error rate for the same amount of training sentences falls to approximately 13%. However, when compared with the ML-trained system, we see that the transformation adaptation methods outperform the standard ML training only when a very small amount of training data is used (i.e. less than 400). For larger amounts of training data, the ML-trained system performs better, and this is due to the bad asymptotic properties of the transformation adaptation, as well as the relatively small vocabulary of the ATIS system.

In Figure 1, we also present the results of the combination of methods I and II with Bayesian adaptation. The combined schemes are proven to be far more efficient than the simple transformation methods I and II, and the adaptation takes better advantage of the amount of the training sentences. The error rate is reduced by 63%, 69% and 75%, with 198, 500 and 2000 adaptation sentences, respectively. Although no direct comparison can be made, using as few as 198 adaptation sentences, the error rate of 9.37% approaches the Stockholm dialect dependent performance. For more sentences the error rate drops even more, to 6.40%. In addition, the combined approach significantly outperforms the ML trained system when less than 1000 sentences are used, providing a solution that is more robust and easier to train.

Speaker	*Word Error Rate %*		
	Non adapted	Meth.II+Bayes 198 sent.	Meth.II+Bayes 3814 sent.
d09	24.94	8.53	8.31
d0b	27.05	12.32	9.90
d0k	21.92	8.49	5.42
d0j	28.64	9.24	6.70
d0r	29.85	13.93	6.71
d0v	19.72	7.66	5.10
d12	26.29	10.07	6.39
d13	22.88	5.26	2.75
total	25.08	9.37	6.40

Table 1. Word recognition performance across Scanian-dialect test speakers using non-adapted and combined-method adapted Stockholm dialect models

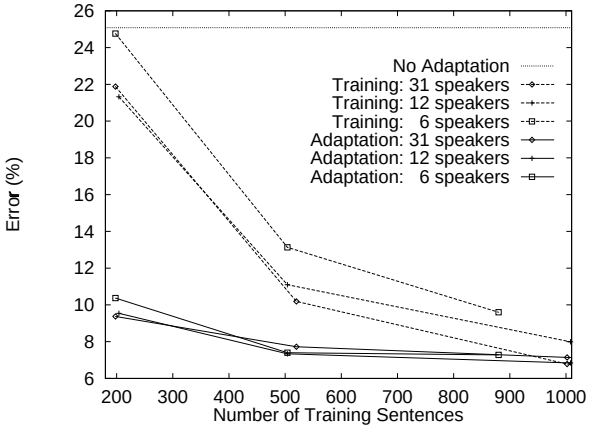

Figure 2. Comparison of dialect training and adaptation results for different number of speakers.

In Table 1, we present the word-error rate of the Stockholm dialect trained system for several Scanian-dialect test speakers. We can see that the improvement in terms of performance when the combined method is used for 198 and 3814 adaptation sentences is almost uniform across the speakers, which verifies the assumption that there is a consistent mismatch across speakers of these two different dialects.

To compare the robustness and trainability of the standard ML training and adaptation algorithms, we performed training and adaptation experiments using fewer speakers in the training set, specifically 12 and 6 speakers. We use the term trainability above to refer to the ease with which a dialect-specific system can be developed. Clearly, the capability of developing a dialect-specific system with as few training speakers as possible is desirable, since it saves both time and money.

The smaller subsets of speakers were selected randomly out of the total number of 31 speakers available in the initial training set, and were equally divided across the two genders. We tried to select speakers from all 4 sub-dialects, so that the resulting system remains subdialect-independent.

The results are illustrated in Figure 2. We see that for standard ML training, the error rate is very large when fewer than 1000 sentences from 31 speakers are used. Moreover, the ML training error rate is getting even larger as the number of speakers in the training set decreases. For example, if we use roughly 500 training sentences, the 31-speaker error rate increases by 9% and 29% when sentences from 12 and 6 speakers are considered, respectively. On the other hand, for the dialect-adapted system, the error rate using 12 and 6 speakers in the adaptation data remains as small as when using the full set of 31 speakers. The small differences are within the statistical error.

The reason for the significantly better performance of the adaptation schemes over standard ML training for small number of speakers is that speaker variability in the systems developed using adaptation techniques is captured from the prior knowledge, which the systems trained using standard ML techniques lack. In general, when we compare adaptation and training results we can conclude that adaptation significantly outperforms training for small amounts of sentences, and small number of speakers. For example, when we perform training with 31 speakers and 520 sentences the results obtained are similar with the adaptation experiments with as few as 6 speakers and only 200 training sentences. Similarly, the performance of a system trained with 31 speakers and 1000 sentences is similar to that of a system trained with only 6 speakers and 500 sentences. Therefore, both the robustness and trainability of an adaptation-based system are highly increased, when compared to standard ML training.

4. CONCLUSIONS

In this paper we have discussed the issue of dialect mismatch in an ASR system. We found, for the pairs of dialects that we examined, that there is a consistent degradation in performance across speakers when there is a dialect mismatch. Hence, we selected to improve the performance of the system using adaptation methods. We tested transformation and combined transformation and Bayesian adaptation algorithms to adapt a Stockholm-trained system to the Scania dialect. The results showed that adaptation is capable of improving the robustness of our system, and that the performance of the adapted system improved dramatically over the mismatched condition with very small amounts of adaptation data. Moreover, we showed that the recognition performance of the adapted system does not degrade when we reduce the number of different speakers from which the training data was collected, something not true for standard ML training. Hence, in terms of robustness and trainability, adaptation is a much better alternative for the development of dialect-specific systems than standard ML training.

ACKNOWLEDGMENTS

The work we have described was accomplished under contract to Telia Research.

REFERENCES

[1] V. Digalakis, D. Rtischev and L. Neumeyer, "Speaker Adaptation Using Constrained Reestimation of Gaussian Mixtures," *IEEE Transactions Speech and Audio Processing,* pp. 357–366, September 1995.

[2] L.Neumeyer, A.Sankar and V.Digalakis, "A Comparative Study of Speaker Adaptation Techniques", *Proceedings of European Conference on Speech Communication and Technology*, pp. 1127–1130, Madrid, Spain, 1995.

[3] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language,* pp. 171–185, 1995.

[4] A. Sankar and C.-H. Lee, "A Maximum Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Transactions Speech and Audio Processing,* pp. 190–202, May 1996.

[5] C.-H. Lee, C.-H. Lin and B.-H. Juang, "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," *IEEE Trans. on Acoust., Speech and Signal Proc.,* Vol. ASSP-39(4), pp. 806–814, April 1991.

[6] J.-L. Gauvain and C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Observations of Markov Chains," *IEEE Transactions Speech and Audio Processing,* pp. 291–298, April 1994.

[7] V. Digalakis and L. Neumeyer, "Speaker Adaptation Using Combined Transformation and Bayesian Methods," *IEEE Transactions Speech and Audio Processing,* pp. 294-300, July 1996.

[8] M. Rayner,I. Bretan,D. Carter, M. Collins, V. Digalakis, B. Gambäck, J. Kaja, J. Karlgren, B. Lyberg, P. Price, S. Pulman and C. Samuelsson, "Spoken Language Translation with Mid-90's Technology: A Case Study," *Proc. Eurospeech '93*, Berlin, 1993.

[9] V. Digalakis, P. Monaco and H. Murveit, "Genones: Generalized Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognizers," *IEEE Transactions Speech and Audio Processing,* pp. 281-289, July 1996.

[10] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood Estimation from Incomplete Data," *Journal of the Royal Statistical Society (B)*, Vol. 39, No. 1, pp. 1–38, 1977.