ON THE INTEGRATION OF DIALECT AND SPEAKER ADAPTATION IN A MULTI-DIALECT SPEECH RECOGNITION SYSTEM

V. Digalakis V. Doumpiotis S. Tsakalidis

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

ABSTRACT

The recognition accuracy in recent Automatic Speech Recognition (ASR) systems has proven to be highly related to the correlation of the training and testing conditions. Several adaptation approaches have been proposed in an effort to improve the speech recognition performance, and have typically been applied to the speaker- and channel-adaptation tasks. We have shown in the past that a mismatch in dialects between the training and testing speakers significantly influences the recognition accuracy, and we have used adaptation to compensate for this mismatch. The dialect of the speaker needs to be identified in a dialect-specific system, and in this paper we present results in this area. To achieve further improvement in recognition performance, we combine dialect- and speaker-adaptation.

1 INTRODUCTION

Performance in large-vocabulary continuous-speech recognition degrades dramatically if a mismatch exists between the training and testing 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 model space [1, 2, 3], Bayesian adaptation [4, 5], or combined approaches [6].

In this paper, we consider the dialect issue on a speaker-independent (SI) speech recognition system and the adaptation of a dialect-specific system to individual speakers. Based on the Swedish language corpus collected by Telia, we have developed a Swedish multi-dialect SI speech recognition system which requires 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 [7]. We have found in the past 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 first identify the dialect of the speaker, and then use a dialect-specific system which has been trained using adaptation techniques with a small amount of data from the target dialect. This dialect-specific system can be further adapted to the speaker, providing additional improvement in recognition, and we show the effect the seed system has in the final speaker-adapted recognition performance.

2 DIALECT AND SPEAKER 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 [8]. The component mixtures of each Gaussian codebook (*genone*) are shared across clusters of HMM states, and hence the observation densities of the vector process x_t have the form:

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

where s_t is the HMM state, x_t is the spectral feature obtained from the recognizer front-end at frame t, and 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.

Using the adaptation method proposed in [1], we assume that the dialect-adapted (DA) observation density of the HMM state s_t for dialect D can be obtained from the corresponding density of the seed-dialect system:

$$P_{DA}(x_t|s_t, D) = \sum_{i}^{N_\omega} p(\omega_i|s_t) N(x_t; m_{ig}(D), S_{ig}(D))$$

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

Adaptation is equivalent to estimation of the parameters $A_g(D), b_g(D), g = 1, \ldots, N_g$. N_g denotes the number of transformations for the whole set of genones. The parameter estimation process is performed using the EM algorithm [9].

In a similar manner, a speaker-adapted (SA) system to a particular speaker S can be obtained using the same transformation method and the dialect-adapted system as a seed model:

$$P_{SA}(x_t|s_t, S) = \sum_{i}^{N_\omega} p(\omega_i|s_t)$$

$$N(x_t; A_g(S)m_{ig}(D) + b_g(S), A_g(S)S_{ig}(D)(A_g(S))^t),$$

where D is the dialect of the speaker. In our experiments we assume the matrix A_g is diagonal [1].

3 DIALECT IDENTIFICATION

To use a the correct system for a particular speaker, his/her dialect must be first identified. One alternative is to run multiple dialect-specific systems in parallel and select the system which maximizes the a-posteriori probability of the dialect given the speaker data. Assuming that all dialects are equally likely, then the identified dialect D^* is simply the one that maximizes the likelihood of the data using the corresponding dialect-specific system. If a Viterbi recognizer is used, we can approximate the summation over all possible state sequences by maximizing the joint likelihood of the observed spectral features $X = [x_1, x_2, \ldots, x_T]$ and the most likely state sequence $S = [s_1, s_2, \ldots, s_T]$ over all possible dialects:

$$D^* = \arg\max_D \max_S p(X, S|D),$$

where

$$p(X, S|D) = \prod_{t=1}^{T} p(s_t|s_{t-1}) P_{DA}(x_t|s_t, D).$$

This maximization can be achieved by running in parallel one Viterbi decoder for each dialect, and at the end selecting the dialect of the recognizer with the highest likelihood as the identified one.

An alternative method that we examined is to base the dialect identification solely on the observation density likelihoods:

$$D^* = \arg\max_D \prod_{t=1}^{T} P_{DA}(x_t|s_t, D),$$

using again the states s_t of the most likely state sequence.

4 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 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.

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 [8]. 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 [8].

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

System Description	Test set		
	Stockholm	Scanian	Total
Stockholm Trained	10.3	25.1	18.0
Scania Adapted	13.2	6.9	10.0
Known Dialect	10.3	6.9	8.3
Dialect Identification			
Recognizer likelihood	10.3	7.3	8.7
Observation Probability	11.2	7.8	9.4

Table 1: Word-error rates (%) for Dialect-dependent, Cross-dialect, Known-dialect and automatic dialect identification methods.

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, reaching a word-error rate of 25.08% when tested on the Scanian-dialect testing set. The degradation in performance was uniform across the various speakers in the test set, suggesting that there are consistent differences across the two dialects. In our previous work [10], we adapted the Stockholm system to the Scania dialect, and we achieved a word-error rate of 6.9% using only a few hundred sentences from six speakers.

When the multi-dialect system is used in dialect-independent mode, the dialect of the speaker must be identified. In Table 1 we present the recognition results of the seed Stockholm and Scania adapted systems on test sets from the Stockholm and Scanian dialects. When the dialect of the speaker is known, then the matched system to the testing conditions performs well and achieves a word-error rate of 8.3%. In multi-dialect mode, however, the dialect must be identified automatically. Using the dialect identification methods that we mentioned above, we were able to identify the correct dialect 96.75% of the time using the recognizer likelihood, and only 79.25% using the observation-density likelihood. As a result, an automatic configuration where the two recognizers were running in parallel and the one with the highest likelihood was selected and its hypothesis was adopted, achieved a word-error rate of 8.7%, very close to the lower bound performance of the known-dialect case, which is 8.3%.

Once the dialect of the speaker is known, then the system can be further adapted to the speaker. In Figure 1 we present speaker-adaptation results on speakers of the Scanian dialect for different amounts of speaker-adaptation data and for different seed systems. Specifically, we adapt to speakers of the Scanian dialect three systems: the original Stockholm speaker-independent system, and two systems that were first adapted to the

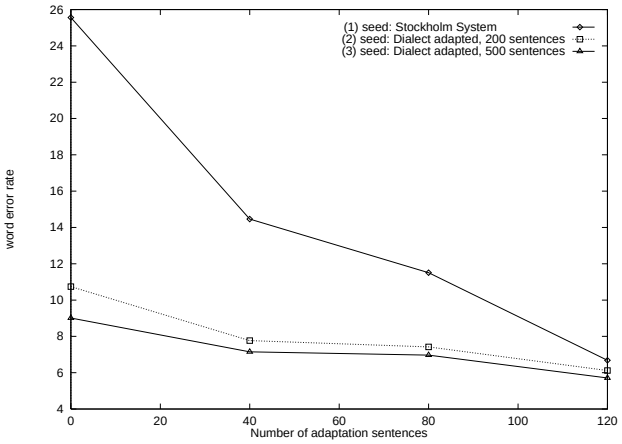

Figure 1: Speaker adaptation results using different seed systems.

Scanian dialect using two hundred and five hundred sentences from other Scanian speakers. We see that even using such a small amount of dialect-dependent data accelerates the convergence of the speaker-adaptation process significantly. For example, the dialect-adapted systems using 40 sentences of the speaker achieve an adapted performance of around 7.0%, whereas the original Stockholm system has an adapted word-error rate of 14.0% when the same 40 adaptation sentences are used.

5 CONCLUSIONS

In this paper we have shown that highly-accurate dialect identification is possible using the observation-density likelihood of hidden Markov models. Identifying the dialect of the speaker is important in multi-dialect applications, where there are significant differences between the dialects and the recognizer can benefit by knowing the dialect of the speaker. In addition, knowledge of the dialect is useful in order to bootstrap the speaker-adaptation process using seed models matched to the speaker's dialect. We have shown that the adaptation process can be accelerated significantly when dialect-specific models are used.

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 Reestima-

tion 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] 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.

[5] 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.

[6] V. Digalakis and L. Neumeyer, "Speaker Adaptation Using Combined Transformation and Bayesian Methods," *IEEE Transactions Speech and Audio Processing,* June 1996.

[7] 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.

[8] 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,* June 1996.

[9] 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.

[10] V. Diakoloukas, V. Digalakis, L. Neumeyer and J. Kaja, "Development of Dialect-Specific Speech Recognizers Using Adaptation Methods," *Proc. Int'l. Conf. on Acoust., Speech and Signal Processing,* Munich, 1997.