Speaker Adaptation from Limited Training in the BBN BYBLOS Speech Recognition System

Francis Kubala
Ming-Whei Feng, John Makhoul, Richard Schwartz

BBN Systems and Technologies Corporation
10 Moulton St., Cambridge, Ma. 02238

Abstract

The BBN BYBLOS continuous speech recognition system has been used to develop a method of speaker adaptation from limited training. The key step in the method is the estimation of a probabilistic spectral mapping between a prototype speaker, for whom there exists a well-trained speaker-dependent hidden Markov model (HMM), and a target speaker for whom there is only a small amount of training speech available. The mapping defines a set of transformation matrices which are used to modify the parameters of the prototype model. The resulting transformed model is then used as an approximation to a well-trained model for the target speaker. We review the techniques employed to accomplish this transformation and present experimental results conducted on the DARPA Resource Management database.

1. Introduction

Soon after a speech recognition system begins operation, small amounts of new speech data become available to the system as spoken utterances are successfully transcribed to text. This data is of potentially great value to the system because it contains detailed information on the current state of the speaker and the environment. The purpose of rapid speaker adaptation is to utilize such small samples of speech to improve the recognition performance of the system.

Speaker adaptation offers other benefits as well. For applications which cannot tolerate the initial training expense of high performance speaker-dependent models, adaptation can trade-off peak performance for rapid training of the system. For typical experimental systems being investigated today on a 1000-word continuous speech task domain, speaker-dependent training uses 30 minutes of speech (600 sentences), while the adaptation methods described here use only 2 minutes (40 sentences).

For applications in which an initial speaker-independent model fails to perform adequately due to a change in the environment or the task domain not represented in the training data, adaptation can utilize an economical initial model generated from the speaker-dependent training of a single prototype speaker. Again, looking at typical systems today, speaker-independent models train on 3 1/2 hours of speech (4200 sentences), while adaptation can use a speaker-dependent model trained from 30 minutes (600 sentences).

In this paper, we describe the speaker adaptive capabilities of the BBN BYBLOS continuous speech recognition system. Our basic approach to the problem is described first in section 2. Two methods for estimating the speaker transformation are described in section 3. In section 4 we present our latest results on a standard testbed database.

3. Methods For Computing the Transformation

In 1987 [5] we reported a new algorithm for estimating a probabilistic spectral mapping between two speakers. The transformation in this method is equivalent to expanding the HMM of the prototype, replacing each state by N states and connecting them in parallel by N transitions. The transition probabilities on each of these paths are then $p(k_i|s,)$, which are the original prototype probabilities for each spectrum, i, given the state, s. The pdf at each new state on these paths is $\hat{\mathbf{p}}(k'|k_i, \phi(s))$ which corresponds to one row of the transformation matrix, $\mathbf{T}_{\phi(\mathbf{s})}$.

Since the conditional probability, $\hat{p}(k'_j|s)$ in equation (3) is computed by the expanded HMM, the familiar forward-backward procedure can be used to estimate $\mathbf{T}_{\phi(\mathbf{s})}$. The target speech is first quantized by the prototype codebook and is automatically aligned against the prototype model. This method worked very well with low perplexity grammars but performance degraded unacceptably as the perplexity of the grammar increased.

We found that cross-speaker quantization was a significant factor in the performance degradation. Also, the transformed pdfs were excessively smooth. We think that the original models, which have been smoothed appropriately for the prototype by interpolating context-dependent phoneme models, may not be specific enough to preserve important detail under the transformation.

To overcome these problems, we investigated a text-dependent procedure which is described in [2]. In this method we constrain the prototype and target speaker to say a common set of training sentences. A class labeling, $\phi(s)$, is derived for each frame of prototype speech by using the prototype HMM to perform recognition while constrained to the correct word sequence. Matching pairs of utterances are time-aligned using a DTW procedure on the parameters of the training speech. This alignment of the speech frames defines a set of spectral co-occurrence triplets, $\{(k'_j, k_i, \phi(s))\}$, for all i, j, which can be counted to estimate the elements of each matrix $\mathbf{T}_{\phi(s)}$ directly.

In this method the target speech is quantized by a codebook derived from the target's own training data thereby eliminating the cross-speaker quantization problem. The smoothing problem is overcome by using the prototype speech itself as the prototype model while estimating the transformation.

We found that the second method outperformed the first using 30 seconds of training speech and an artificial grammar of perplexity 60. This remained true even after controlling for the quantization problem of the first method by adapting the prototype codebook after the manner of [6].

Several enhancements have been made to the DTW-based method. As described in [3], we introduced an iterative normalization procedure which modifies the speech parameters of one speaker by shifting them toward the other speaker. A VQ codebook partitions the speech of one speaker into groups of spectra which quantize to a common VQ codeword. The DTW alignment maps the partition onto corresponding groups of spectra for the other speaker. The shift is then determined by the difference vector between the means of these corresponding groups of spectra. Each iteration of aligning and shifting reduces the mean-squared error of the aligned speech parameters until convergence.

More recently, we have used additional features in the DTW to improve the alignment between utterances, and additional codebooks in the HMM to improve the prototype model.

2. Basic Approach to Speaker Adaption

We view the problem of speaker adaptation as one of modeling the difference between two speakers. One of the speakers, who we call the prototype, is represented by a speaker-dependent HMM trained from large amounts of speech data. The other speaker, called the target, is represented by only a small sample of speech. If the difference between the speakers can be successfully modeled, then one strategy for speaker adaptation is to make the prototype speaker look like the target speaker. This can be accomplished by finding a transformation which can be applied to the prototype HMM that makes it into a good model of the target speech.

The difference between speakers is a complex one, involving the interaction of spectral, articulatory, phonological, and dialectal influences. A non-parametric probabilistic mapping between the VQ spectra of the two speakers has appropriate properties for such a problem. A probabilistic transformation can capture the many-to-many mapping typical of the differences between speakers and it can be made robust even when estimated from sparse data. Non-parametricity makes few constraining assumptions about the data under transformation. Mapping VQ spectra between speakers constrains the transformation to dimensionswhich can be estimated reasonably from the limited training data.

We begin with high performance speaker-dependent phonetic models which have been trained from a large sample of speech from the prototype speaker. The speaker-dependent training procedure in the BYBLOS system has been described in [1]. For each state of the prototype HMM, we have a discrete probability density function (pdf) represented here as a row vector:

$$\mathbf{p}(s) = [p(k_1|s), p(k_2|s), ..., p(k_N|s)] \tag{1}$$

where $p(k_i|s)$ is the probability of the VQ label k_i at state s of the prototype HMM model, and N is the size of the VQ codebook.

The elements of the desired transformed pdf, $\mathbf{p}'(s)$, can be computed from:

$$p(k'_j|s) = \sum_{i=1}^{N} p(k_i|s)p(k'_j|k_i, s) \tag{2}$$

Since we have insufficient data to estimate a separate transformation for each state we approximate $\mathbf{p}'(s)$ by:

$$\hat{p}(k'_j|s) = \sum_{i=1}^{N} p(k_i|s)p(k'_j|k_i, \phi(s)) \tag{3}$$

where $\phi(s)$ specifies an equivalence class defined on the states s.

For each of the classes, $\phi(s)$, the set of conditional probabilities, $\{p(k'_j|k_i, \phi(s))\}$, for all i and j form an $N \times N$ matrix, $\mathbf{T}_{\phi(s)}$, which can be interpreted as a probabilistic transformation matrix from one speaker's spectral space to another's. We can then rewrite the computation of the transformed pdf, $\hat{\mathbf{p}}'(s)$, as the product of the prototype row vector, $\mathbf{p}(s)$, and the matrix, $\mathbf{T}_{\phi(s)}$:

$$\hat{\mathbf{p}}'(s) = \mathbf{p}(s) \times \mathbf{T}_{\phi(s)}; \quad \mathbf{T}_{ij\phi(s)} = p(k'_j|k_i, \phi(s)) \tag{4}$$

There are many ways to estimate $\mathbf{T}_{\phi(s)}$. We describe next two procedures that we have investigated.

4. Experimental Results

The DARPA Resource Management database [4] defines a protocol for evaluating speaker adaptive recognition systems which is constrained to use 12 sentences common to all speakers in the database. To avoid problems due to unobserved spectra, we have chosen to develop our speaker adaptation methods on a larger training set, which restricts us to the speaker-dependent portion of the database for performance evaluation.

This segment of the database includes training and test data for 12 speakers sampled from representative dialects of the United States. We have used the first 40 utterances (2 minutes of speech) of the designated training material for our limited training sample. Two development test sets have been defined by the National Institute of Standards and Technology (NIST). These test sets consist of 25 utterances for each speaker. Each test set is drawn from different sentence texts and includes about 200 word tokens.

For all of our experiments, we have used one male prototype speaker originally from the New York area. 30 minutes of speech (600 sentences) were recorded at BBN in a normal office environment and used to train the prototype HMM. The speech is sampled at 20 kHz and analysed into 14 mel-frequency cepstral coefficients at a frame rate of 10 ms. 14 cepstral derivatives, computed as a linear regression over 5 adjacent frames, are derived from the original coefficients. The transformation matrices are made to be phoneme-dependent by defining the equivalence classes, $\phi(s)$, over the 61 phonemes in the lexicon.

Experiment	Features	Normalized	Codebooks	% Word Error
1	14	NO	1	17.8
2	14	YES	1	14.7
3	28	NO	1	15.3
4	28	YES	1	13.2
5	28	NO	2	10.8
6	28	YES	2	9.8

Table 1: Comparison of speaker adaptation results averaged over 8 speakers for the Word-Pair grammar and the Oct. '87 test set.

We have performed our development work on 8 speakers using the test set designated by NIST as Oct. '87 test. The results of this work, using the standard word-pair grammar, are summarized in Table 1, where:

% Word Error = 100 × [(*substitutions* + *deletions* + *insertions*) / *number of word tokens*]

For each experiment we show the number of features used in the DTW alignment, whether the iterative normalization procedure was used, and the number of codebooks used in recognition.

Using experiment (1) as a baseline, the table shows a 45% decrease overall in word error rate for using all three improvements together. Comparing experiments using 14 features with their counterparts using 28 features shows that the contribution due to the differential features is roughly a 10% - 14% reduction in error rate. A similar comparison for using/not-using the normalization reveals a 9% - 17% reduction. Finally, using the second codebook reduces the error rate by 26% - 29%.

It should be mentioned that the 40 sentences used for training in these experiments are drawn equally from 6 recording sessions separated by several days. Furthermore, the test data is from another session altogether. For the adaptation methods described here, it is reasonable to assume that the training data would be recorded in a single session and only a few minutes before the transformed models were ready for use. This means that the adaptation training and test data should realistically come from the same recording session. From earlier published experiments using single-session training and test, we believe the multi-session material accounts for about 1/5 of the total word error for the experiments reported here.

				Word	Word	Sentence
Speaker	Substitutions	Deletions	Insertions	Correct	Error	Error
DAS (F)	2.0	1.5	0.0	96.6	3.5	16.0
DMS (F)	2.2	2.8	0.0	95.0	5.0	20.0
DTD (F)	4.3	0.4	0.4	95.3	5.1	32.0
TAB	2.2	3.4	0.0	94.4	5.6	32.0
PGH	3.9	2.0	0.0	94.1	5.9	32.0
CMR (F)	2.6	1.7	1.7	95.7	6.0	36.0
HXS (F)	3.2	0.5	3.2	96.4	6.9	32.0
DTB	7.5	2.6	0.0	89.9	10.1	48.0
ERS	8.5	1.4	1.9	90.1	11.8	52.0
RKM	7.2	1.9	2.9	90.9	12.0	48.0
JWS	9.0	4.5	0.0	86.5	13.5	52.0
BEF	8.4	5.8	0.9	85.8	15.1	56.0
AVG	5.1	2.4	0.9	92.6	8.4	38.0

Table 2: Recognition performance by speaker for the Word-Pair grammar and the May '88 test set.

We evaluated the three improvements to the system by testing on new data designated as the May 88 test set which is defined for 12 speakers. For this experiment, we added 2 features, normalized energy and differential energy, and an additional codebook for the energy features. All parameters for this experiment were fixed prior to testing. The results shown in Table 2 were obtained from the first run of the system on the May '88 test data. All entries in the table are percentages, where:

% Word Correct = 100 × [1 − (*substitutions* + *deletions*) / *number of word tokens*]

% Sentence Error = 100 × [*number of sentences with any error* / *number of sentences*]

and % Word Error is defined as in Table 1.

The speakers in Table 2 are ordered from the top by increasing word error rate. It is evident from the table that the speakers cluster into two distinct performance groups. It is remarkable that all 5 female speakers are included in the higher performance group despite the fact that the prototype is male. The ordering of speakers shown here is not predicted by their speaker-dependent performance or by subjective listening.

The average word error rate of 8.4% for this test set is comparable to previously reported results from speaker-independent systems on this identical test set. Using training from 105 speakers (4200 sentences), the word error rates for the Sphinx system of CMU was 8.9% and for the Lincoln Labs system; 10.1%. New results from these systems, on different test data but from the same 12 speakers, are reported elsewhere in these proceedings.

5. Conclusion

Three improvements to the DTW-based speaker adaptation method have been combined to achieve a 45% overall reduction in recognition word error rate on development test data. The largest single improvement was due to the addition of a codebook derived from a set of cepstral derivative features. This improvement does not affect the estimation of the between-speaker transformation. This suggests that further improvements to the speaker-dependent prototype model can lead to significant improvements in the adapted model's performance.

The performance of the system on new evaluation test data was 8.4% word error averaged over 12 speakers, using the standard word-pair grammar. The system used a total of 600 sentences from a single prototype speaker and and a training sample of 40 sentences from each of the 12 test speakers. The performance of the system is comparable to several speaker-independent systems trained on 4200 sentences from 105 speakers, and tested on the same data. This result suggests that speaker adaptation may be the most cost-effective solution for applications which must be brought up quickly and must accommodate changing task domains or test conditions.

Acknowledgement

This work was supported by the Defense Advanced Research Projects Agency and monitored by the Space and Naval Warfare Systems Command under Contract No. N00039-85-C-0423.

References

[1] Chow, Y., M. Dunham, O. Kimball, M. Krasner, F. Kubala, J. Makhoul, P. Price, S. Roucos, and R. Schwartz (1987) "BYBLOS: The BBN Continuous Speech Recognition System," *IEEE ICASSP-87*, paper 3.7.1.

[2] Feng, M., F. Kubala, R. Schwartz, J. Makhoul (1988) "Improved Speaker Adaptation Using Text Dependent Spectral Mappings," *IEEE ICASSP-88*, paper S3.9.

[3] Feng, M., R. Schwartz, F. Kubala, J. Makhoul (1989) "Iterative Normalization for Speaker-Adaptive Training in Continuous Speech Recognition," *IEEE ICASSP-89*, To be published.

[4] Price, P., W. Fisher, J. Bernstein, and D. Pallett (1988) "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," *IEEE ICASSP-88*, paper S13.21.

[5] Schwartz, R., Y. Chow, F. Kubala (1987) "Rapid Speaker Adaptation using a Probabilistic Spectral Mapping," *IEEE ICASSP-87*, paper 15.3.1.

[6] Shikano, K., K. Lee, R. Reddy (1986) "Speaker Adaptation Through Vector Quantization," *IEEE ICASSP-86*, paper 49.5.