EXPERIMENTS USING DATA AUGMENTATION FOR SPEAKER ADAPTATION

Jerome R. Bellegarda, Peter V. de Souza, David Nahamoo, Mukund Padmanabhan, Michael A. Picheny,* and *Lalit R. Bahl*

IBM Research
T.J. Watson Research Center
Yorktown Heights, New York 10598

ABSTRACT

Speaker adaptation typically involves customizing some existing (reference) *models* in order to account for the characteristics of a new speaker. This work considers the slightly different paradigm of customizing some reference *data* for the purpose of populating the new speaker's space, and then using the resulting (augmented) data to derive the customized models. The data augmentation technique is based on the metamorphic algorithm first proposed in [1], assuming that a relatively modest amount of data (100 sentences) is available from each new speaker. This contraint requires that reference speakers be selected with some care. The performance of this method is illustrated on a portion of the Wall Street Journal task.

I. INTRODUCTION

In the recognition of typical large vocabulary natural language speech tasks, such as the U.S. Advanced Projects Agency's (ARPA) Wall Street Journal (WSJ) task, training the recognition system in a speaker-dependent fashion generally achieves a lower word error rate than training it speaker-independently [2]. This is because in speaker-dependent recognition the parameters of the system—such as acoustic prototypes, hidden Markov model (HMM) parameters [3], etc.—can be specifically tailored to each individual speaker. The main drawback of this strategy, however, is that a relatively large amount of training data is required from each new speaker before he or she can utilize the recognizer profitably.

For example, in [2] we evaluated the IBM large vocabulary speech recognition system on the WSJ task under both speaker-dependent and speaker-independent conditions, using a varying amount of training data in each case. We then compared the results obtained on the same test corpus under the two conditions. A typical conclusion was that speaker-dependent processing reduced the average error rate by about 15% when 600 training sentences were available, but by more than 30% when 2400 training sentences were available. This underscores the fact that speaker-dependent training may only be truly beneficial on a large amount of data.

Unfortunately, in most applications the user is not willing to provide that much training data. Furthermore, as speaker-independent systems become more and more representative, it makes sense to take advantage of some of their "universal" properties, e.g., for initialization or smoothing purposes. It therefore seems appropriate to switch to a speaker-adaptive paradigm, where a smaller amount of training data is necessary. In the case of the above experiments, for example, we might want to use a maximum of 100 sentences, as specified in recent ARPA evaluation proposals.

Speaker adaptation typically involves perturbing a set of existing (reference) *models* in the direction of the new data, to account for the idiosyncrasies of the new speaker. Alternatively, one can perturb some reference *data* to make it look like it came from the new speaker. This way, it is possible to populate the new speaker's space with the perturbed data, and then use the resulting (augmented) data to derive the customized models. The following implements this alternative approach using the metamorphic algorithm first proposed in [1]. Note that this does not result in incremental adaptation, but rather "block adaptation," where the customization is performed at specific intervals in time (such as after each day of use for instance).

The paper is organized as follows. In the next section we discuss our general strategy to perform data augmentation using the metamorphic framework. In Section III, we propose suitable solutions to two important implementation issues. Section IV gives a complete overview of the algorithm used to conduct the experiments. Finally, in Section IV we present experimental results obtained on the pilot portion (WSJ0) of the Wall Street Journal task.

*J.R. Bellegarda is now with the Advanced Technology Group, Apple Computer Inc., Cupertino, California 95014.

0-7803-2431-5/95 $4.00 © 1995 IEEE

II. GENERAL STRATEGY

Many techniques have been developed to carry out speaker adaptation, via codebook adaptation (e.g., see [4]), HMM parameter adaptation (cf., e.g., [5]), or both [6]. Broadly speaking, the underlying strategy is to find an appropriate mapping between a new speaker's codebook, or spectral space, or HMM parameters, and the corresponding entity from a reference speaker, or speakers [7]. The idea is to use this speaker-normalizing mapping to transform the previously acquired parameters of the reference speaker onto the space of the new speaker.

This transformation may be constructed globally through a linear mapping to make the two speakers look as alike as possible, or a non-linear mapping such as a multi-layer perceptron [8] to better handle the separation of phones. Alternatively, it may be constructed at the phone level as in [9], where a piecewise linear mapping was used to transform the feature vectors from the reference feature space(s) to the new feature space.

This paper builds on the framework described in [9] to perform speaker adaptation through data augmentation. The intuitive appeal of this approach is that the speaker-normalizing mapping is done prior to training. Hence, in principle, the derivation of the models is not restricted to the existing reference set up. On the other hand, any deficiency in the mapping may be amplified during later stages of processing. Thus, the trade-off, not surprisingly, is between flexibility and robustness.

In its original implementation, the metamorphic algorithm used about 10 times more reference data than new data. In some cases, however, the imperfect estimation of the transformation parameters sometimes led to an overwhelming of the original data by a large quantity of inadequately transformed data. It was found in [9] that transforming about the same amount of reference data as new data represented a reasonable compromise. Thus, in the example mentioned earlier, we would transform about 100 sentences of reference data. This immediately entails that, to be effective, this approach has to involve more than one reference speaker. Otherwise, the amount of training data after data augmentation would still be quite small.

From the point of view of data merging, it makes no difference whether the reference data is accumulated across several speakers, since the metamorphic paradigm maps everything onto the same place. Thus, it is in fact possible to use multiple reference speakers instead of a single one. The idea is to apply the metamorphic algorithm to transform the data of each of the reference speaker onto the space of the new speaker, thereby populating the space of the new speaker with enough data to reliably estimate the parameters of the (mixture Gaussian) HMM output distributions.

There are two potential problems with this framework, however. First, 100 sentences of data for each speaker may not be sufficient to estimate the (phone-based) piecewise linear mappings in the metamorphic approach. Least-squares estimation, in particular, is vulnerable to sparse data [9]. Second, if for some reference speakers the metamorphic algorithm performs less satisfactorily than for others, there is a danger of populating the space of the new speaker with transformed data of uneven quality. This in turn may introduce some bias in the modeling. These two potential difficulties are examined in detail in the next section.

III. IMPLEMENTATION

Let us first consider the issue of the reliability of the metamorphic transformations in the face of sparse data. This problem can be addressed by implementing various tying mechanisms in the mapping estimation. Recall from [1] or [9] that for a given phone the metamorphic framework involves three transformations. First, each reference feature vector X is mapped to the unit sphere, yielding:

$$\tilde{X} = S_r^{-1/2}(X - M_r), \tag{1}$$

where M_r and S_r are the mean vector and covariance matrix for this phone. Similarly, each training feature vector Y is mapped to the unit sphere, resulting in:

$$\tilde{Y} = S_t^{-1/2}(Y - M_t), \tag{2}$$

with similar definitions for M_t and S_t. Finally, there is an orthonormal transformation Γ which maps the reference vectors $\tilde{X}$ onto the training vectors $\tilde{Y}$ within the unit sphere. As a result, the metamorphically transformed data is obtained as:

$$\hat{X} \;=\; (S_t^{1/2}\Gamma S_r^{-1/2})X + (M_t - S_t^{1/2}\Gamma S_r^{-1/2}M_r), \tag{3}$$

where $\hat{X}$ represents the reference vector X in the space of the new speaker.

In preliminary experiments, we found that for the estimation of the covariance matrices S_r and S_t, tying at the phone level is sufficient. However, for the least squares estimation of Γ, tying across phones is necessary. In the implementation of [1], phones were tied together in groups of three or four, and the same orthonormal transformation Γ was shared within each group. Here, because the amount of data available is more constrained, a even stronger tying is called for. We found it necessary to share a single orthonormal transformation Γ across all phones.

Let us now consider the issue of the potentially uneven quality of the transformed data. This problem is related to the well-known phenomenon of "outlier" speakers, and can be addressed by implementing a reference speaker selection procedure. This procedure discards those speakers in the reference pool which are "far away" from the new speaker, in some appropriate metric. Thus, only speakers sufficiently "close" to the new speaker contribute to the data augmentation. We have found that the metamorphic algorithm has an easier time transforming the data for these speakers, thereby reducing the risk that low quality data overwhelm the original data in the new speaker's space.

The distance measure used in the speaker selection procedure is chosen to reflect the difficulty of reliably estimating the transformation Γ discussed earlier. As a result, it operates within the unit sphere. Taking (1) and (2) into account, the distance between two speakers R and T is defined to be:

$$D(R,T) = \sum_{(X \in R, Y \in T)} \|\tilde{X} - \tilde{Y}\|^2 , \qquad (4)$$

where the summation is over all pairs of matched anchor points for these two speakers (cf. [9]). Note that a distance of zero would mean that Γ evaluates to the identity matrix in the metamorphic processing associated with these two speakers. Thus, only speakers whose orthonormal transformation is "close" to the identity (i.e., whose data can be more reliably transformed) contribute to the data augmentation.

IV. DATA AUGMENTATION

This section provides an overview of the complete data augmentation algorithm used in the experiments reported below. To initialize the algorithm, we assume that some speech has been recorded for all speakers in the reference pool (reference data) as well as the new speaker (training data). We further assume that all recordings have been appropriately labelled and Viterbi aligned against suitable Markov word models [2]. Finally, we assume that for any given phone, suitable mean vectors M_t and covariance matrices S_t have been extracted for the new speaker T (cf. [1], [9]).

For each reference speaker R, we perform the following steps: (i) tie the reference data at the phone level for covariance estimation; (ii) extract mean vectors M_r and covariance matrices S_r; (iii) compute the distance $D(R,T)$ between this speaker and the new speaker, using the definition (4). After all reference speakers R have been ranked according to this distance, only those whose $D(R,T)$ is smaller than a pre-determined threshold are retained for further processing.

For each reference speaker so selected, we continue with the following: (iv) tie the reference data across all phones for least squares estimation; (v) estimate the orthonormal transformation Γ to map one normalized space onto the other; (vi) apply the resulting piecewise linear mapping to the reference data to map it to the space of the new speaker, cf. (3); (vii) merge the transformed reference data with the original training data, as the two are now extremely close.

This procedure effectively increases the amount of training data available from the new speaker. After that, training proceeds normally on the augmented data. In the IBM system, this results in a set of mixture Gaussian HMM output distributions, which are then smoothed back to the speaker-independent distributions for added robustness.

V. EXPERIMENTAL RESULTS

We have performed a series of experiments to illustrate some of the benefits of the above approach. We considered the 12-speaker speaker-dependent portion of the Wall Street Journal task released in the Fall of 1992. Without loss of generality, attention was restricted to 20K, non-verbalized punctuation data. All experiments were conducted with the standard MIT Lincoln Laboratory 20,000 word statistical trigram language models provided by Paul [10]. For the sake of completeness, we ran on both development and evaluation test sets. We ran development tests with the standard closed vocabulary trigram language model (as this is more amenable to analysis), and the evaluation tests with the standard open vocabulary trigram language model (as in official ARPA evaluations).

The training data was taken to be the first 100 sentences of the data provided by each speaker. The reference pool was taken to be the WSJ0 speaker-independent set (SI-84), which was also used to train the reference speaker-independent system. Of these 84 speakers, an average of about 20 were selected by the reference speaker selection procedure mentioned earlier. This resulted in about 1500 sentences transformed with the metamorphic algorithm onto the space of the new speaker. Thus, for each speaker, the 100 sentences of training data were supplemented by 1500 sentences of synthesized training data obtained through metamorphic acoustic mapping.

The error rates obtained using the speaker adaptation procedure described above were compared to the error rates obtained using the speaker-dependent system of [2] with 600 sentences of training data. (This speaker-dependent system is about 25% better than the SI-84 speaker-independent system on this data—cf. [2].) The results are summarized in Table I for the development test data and in Table II for the evaluation test data. Over the 12 speakers the error rate remains essentially identical even as the (original) training data is reduced 6-fold from 600 to 100 sentences. This shows that data augmentation through the metamorphic algorithm appears to be a viable approach to (non-incremental) speaker adaptation.

Speaker	Spkr–Dep 600 Sentences Training	Spkr–Adapt 100 Original +1500 Mapped
001	2.8 %	2.8 %
002	5.2 %	5.2 %
00a	11.8 %	12.1 %
00b	9.9 %	10.1 %
00c	11.8 %	12.2 %
00d	8.9 %	9.0 %
00f	12.3 %	12.8 %
203	7.7 %	7.8 %
400	9.4 %	9.7 %
430	5.5 %	5.5 %
431	7.7 %	7.8 %
432	3.8 %	3.8 %
Average	8.1 %	8.2 %

Table I. Word Error Rates – Development Test Data.

Speaker	Spkr–Dep 600 Sentences Training	Spkr–Adapt 100 Original +1500 Mapped
001	9.2 %	9.2 %
002	8.6 %	8.6 %
00a	12.2 %	13.2 %
00b	7.3 %	7.2 %
00c	13.9 %	14.7 %
00d	10.3 %	10.4 %
00f	19.2 %	20.8 %
203	12.2 %	12.5 %
400	9.8 %	9.8 %
430	14.0 %	14.9 %
431	10.4 %	10.4 %
432	6.9 %	6.8 %
Average	11.2 %	11.5 %

Table II. Word Error Rates – Evaluation Test Data.

VI. CONCLUSION

We have proposed to use the metamorphic framework presented in [9] to perform speaker adaptation through data augmentation. The intuitive appeal of this approach is that the speaker-normalizing mapping is done prior to training, meaning that, in principle, the derivation of the models is not restricted to the existing reference set up. Since the amount of data provided by the new speaker is relatively modest, we use multiple reference speakers instead of a single one. To reduce the risk that low quality data overwhelm the original data in the new speaker's space, we implement a reference speaker selection procedure which discards those speakers in the reference pool deemed "too far away" from the new speaker. In addition, for the sake of robustness we implement various tying mechanisms in the estimation of the metamorphic transformations. Experiments performed on the Wall Street Journal corpus shows that this approach appears to be viable for block adaptation with 100 sentences of data.

REFERENCES

[1] J.R. Bellegarda *et al.*, "Robust Speaker Adaptation Using a Piecewise Linear Acoustic Mapping," in *Proc.* 1992 *ICASSP*, San Francisco, CA, pp. I445–I448, March 1992.

[2] L.R. Bahl *et al.*, "Performance of the IBM Large Vocabulary Continuous Speech Recognition System on the ARPA Wall Street Journal Task," in *Proc.* 1995 *ICASSP*, Detroit, MI, elsewhere in these Proceedings.

[3] L.R. Bahl *et al.*, "Multonic Markov Word Models for Large Vocabulary Continuous Speech Recognition," *IEEE Trans. Speech Audio Proc.*, Vol. SAP–1, No. 3, pp. 334–344, July 1993.

[4] K. Shikano, K. Lee, and R. Reddy, "Speaker Adaptation Through Vector Quantization," in *Proc.* 1986 *ICASSP*, Tokyo, Japan, pp. 2643–2646, April 1986.

[5] G. Rigoll, "Speaker Adaptation for Large Vocabulary Speech Recognition Systems Using 'Speaker Markov Models'," in *Proc.* 1989 *ICASSP*, Glasgow, UK, pp. 5–8, May 1989.

[6] Y. Zhao, "An Acoustic–Phonetic–Based Speaker Adaptation Technique for Improving Speaker–Independent Continuous Speech Recognition," *IEEE Trans. Speech Audio Proc.*, Vol. SAP–2, No. 3, pp. 380–394, July 1994.

[7] F. Kubala, R. Schwartz, and C. Barry, "Speaker Adaptation From a Speaker–Independent Training Corpus," in *Proc.* 1990 *ICASSP*, Albuquerque, NM, pp. 137–140, April 1990.

[8] J.P. Tubach *et al.*, "Speaker Adaptation for Speech Recognition Systems: Multiple Linear Regression and Multilayer Perceptrons," *Trait. Signal*, Vol. 7, No. 4, pp. 285–292, 1990.

[9] J.R. Bellegarda *et al.*, "The Metamorphic Algorithm: A Speaker Mapping Approach to Data Augmentation," *IEEE Trans. Speech Audio Proc.*, Vol. SAP–2, No. 3, pp. 413–420, July 1994.

[10] D.B. Paul and J.M. Baker, "The Design for the Wall Street Journal–based CSR Corpus", in *Proc.* DARPA *Speech Nat. Lang. Work.*, Morgan–Kaufmann, pp. 357–362, Feb. 1992.