final paper on

Speaker Adaptation Using Maximum Likelihood Linear Regression

submitted to fulfill the requirements for

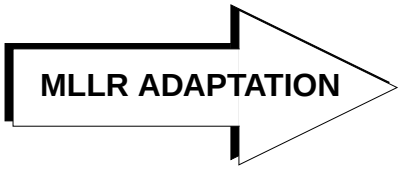
ECE 8993: Fundamentals of Speech Recognition

submitted to:

Dr. Joseph Picone
Department of Electrical and Computer Engineering
413 Simrall, Hardy Rd.
Mississippi State University
Mississippi State, Mississippi 39762

July 26, 1998

Speaker-Independent Recognition



Speaker-Dependent Recognition

submitted by:

Jonathan Hamaker
MS in Computer Engineering Candidate
Department of Electrical and Computer Engineering
Mississippi State University
429 Simrall, Hardy Rd.
Mississippi State, Mississippi 39762
Tel: 601-325-8335, Fax: 601-325-3149
Email: hamaker@isip.msstate.edu

ABSTRACT

In typical state-of-the-art large vocabulary continuous speech recognition (LVCSR) systems a single model is developed for all speakers. With this, we expect that our systems will generalize well to all speakers. However, from experience we know that there are speakers who are poorly modeled using this paradigm. Thus, it would be advantageous to adapt the models, during run-time, to the new speaker. Following this premise, many methods have been developed which use a small amount of a speaker's data to adapt the speaker-independent models to that speaker.

In this paper we describe one method which uses a maximum likelihood linear regression (MLLR) approach to speaker adaptation. MLLR builds a transform for the model means using linear regression so that the transformed mean of each model better represents the new speaker. Applying this approach to all of the models in an LVCSR system (particularly when using mixture models) would require an unreasonable number of additional parameters and a large amount of training data for full coverage. To attack this problem a small number of transforms are built and tying is used. MLLR has become a standard addition to basic LVCSR systems and has proven successful in every major speaker-independent speech recognition task to which it has been applied.

1. INTRODUCTION

Commercially available dictation systems have recently hit the speech products market. These have, for the most part received rave reviews from users. Most of these systems claim to work well out of the box but perform better as the user performs more dictation with it. This indicates that the systems used in these applications are somehow adjusting to the speaker — also that the speaker is adjusting to the subtleties of getting the application to work. It is this phenomena that we will discuss in this paper.

Speaker-independent recognition systems have been developed to the point that they perform very well for LVCSR in the general case. However, speaker-independent systems, in general, are known to have poorer performance than systems with speaker-dependent models [1, 2]. The main reason for this is that speaker-independent systems are discarding the knowledge that the same speaker is, in fact, speaking every utterance. In doing so, the system is negating the ability of the models to describe the peculiarities of each specific speaker (vocal tract shape and length, accent, etc.) in favor of a general model of any speaker.

On the other hand, there is a very large problem with developing such a speaker-dependent system: doing so would require a large amount of training data from every speaker involved which is impractical for most applications. There are vast amounts of training data available for speaker-independent tasks such as SWITCHBOARD [3]. This provides clear motivation for techniques which would allow us to adapt the speaker-independent models to a new speaker using a small amount of adaptation data. From this need, there have been many attempts to develop robust speaker adaptation techniques.

2. SPEAKER ADAPTATION

The basic idea of speaker adaptation can be seen in Figure 1. Essentially, we want to use a a small amount of adaptation data as possible to change our recognition system such that they model as much of the speaker-specific information as possible [4]. Many approaches have been developed which try to produce this effect.

Speaker adaptation techniques for HMM-based recognition systems fall into two

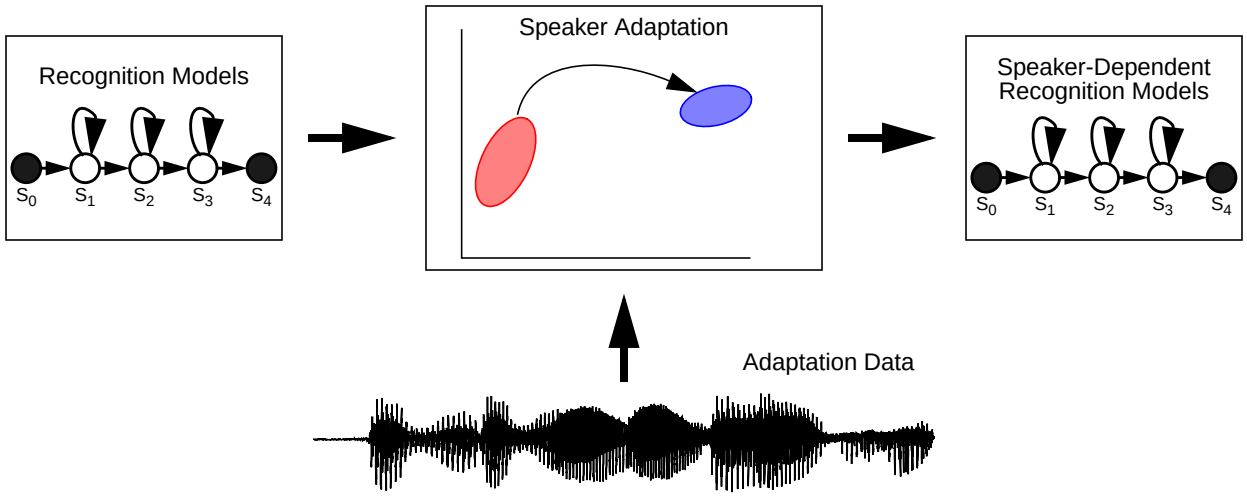


Figure 1. A high-level representation of the speaker adaptation process using HMMs. The speaker adaptation process uses the adaptation data to affect the modeling process such that the models are a closer match to the adaptation data. In the adaptation depicted, both the mean and variance of the data is affected by the transform.

basic categories. The first of these employs methods which transform the input speech of the new speaker to the match the characteristics of the training speech. These are known as **spectral mapping techniques**. Second are methods which transform the model parameters to better match the characteristics of the adaptation data. These techniques are known as **model mapping approaches**. The following sections describe each of these, in brief.

2.1. Spectral Mapping Approach

The spectral mapping approach is based on the belief that a recognition system can be improved by matching the new speaker's features vectors to the vectors of the training data [5]. The mapping is designed so that the difference between the reference vector and the mapped vector is minimized. These differences are due to the spectral differences of the speakers' speech production systems.

Initial attempts at spectral mapping adaptation were used in the spectral template matching systems [6, 7, 8]. These consider the template to be from the reference speaker and automatically generate a transformation to minimize the difference between the new speaker and the reference speaker [5]. Other approaches [9] have mapped both the reference data and the new speaker's data into a common vector which is said to maximally correlate the two. A variation on these methods which is similar to speaker normalization used a transform to map each speaker in the speaker-independent training set onto a reference speaker [10, 11]. Thus, the models generated act as speaker-dependent models. This approach is illustrated in Figure 2.

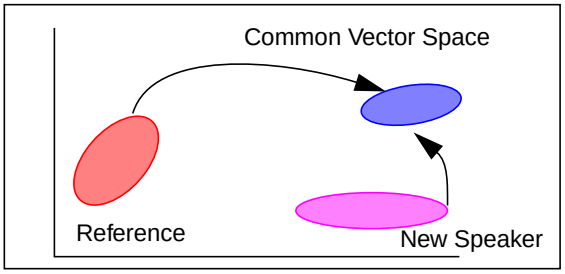


Figure 2. Spectral Mapping approach where both the reference and new speaker feature vectors are mapped to a common space which maximizes the correlation between the two.

2.2. Model Mapping Approach

The aim of spectral mapping is to improve the match between the reference speaker and new speaker. This goal does not explicitly try to increase the accuracy of the models for the new speaker. This is where the model mapping approach attempts to make its improvements. Rather than trying to map all speakers to one space, the model mapping approach adjusts the model parameters to best represent the new speaker.

Bayesian MAP (maximum a posteriori) approaches are the most commonly used techniques for model adaptation of HMMs. In a MAP approach, the transformation is chosen such that the new model parameters chosen maximize

$$F(\lambda|O) = \frac{F(O|\lambda)F(\lambda)}{F(O)} \quad (1)$$

where O is the adaptation observation sequence and λ is the parameter set defining the distribution. Different methods have been used to estimate the value of λ including a segmental K-means approach [12] and an EM-based approach [13]. Most of these MAP approaches are limited in that they only adapt the parameters that are directly observed in the adaptation data.

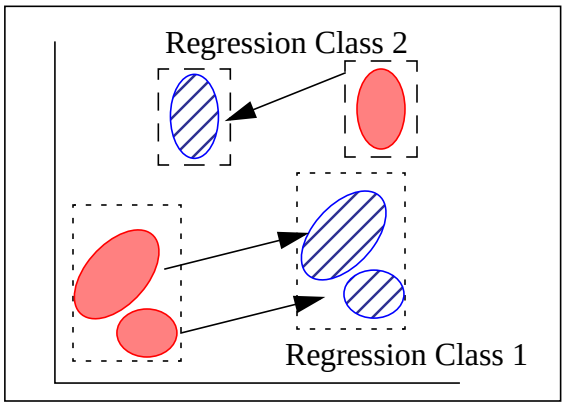


Figure 3. Representation of MLLR as it transforms the means of the mixtures. Notice that the shapes of the mixtures remains the same. Also note that the two distributions in Regression Class 1 are transformed by the same matrix.

3. MLLR

Maximum likelihood linear regression (MLLR) — (developed by Leggetter [1, 5, 14]) was designed to overcome the disadvantages of both the spectral mapping and model mapping techniques. MLLR is a transform-based method which adapts the model parameters like the MAP-based adaptation but is robust enough to produce effects from a small amount of training data. This approach was developed from work by Hewitt [15] which applied a least squares regression to adapt templates in dynamic time warping. MLLR extends this idea to the continuous density HMMs and uses ML to optimize the regression.

3.1. MLLR Basics

Ideally, all parameters of the system should be adapted to the new speaker but, in practice, this would require too much adaptation data to accurately estimate the adapted models. For this reason, MLLR only adapts the means of the models. This is also justified by the assumption that the primary difference between speakers is in the average formant positions for phones rather than the distribution of the intra-speaker variation [14]. This is the same reasoning given in many VQ/HMM adaptation schemes [16, 17].

MLLR uses a set of regression-based transforms to adapt all of the HMM means to the new speaker. The number of transforms in this set can be as small as one — a global transform where all means are adapted by the same regression — or large enough so that each HMM mean had a unique regression transform. An example of the effects of MLLR is shown in Figure 3. A method analogous to state-tying is used to find the

optimal number of regression classes given the adaptation data [5].

For a HMM mixture component s with mean μ_s, the adapted mean is given by

$$\hat{\mu}_s = W_s \xi_s \quad (2)$$

where W_s is an $n \times (n+1)$ transformation matrix and ξ_s is the extended mean vector for mixture component s given by

$$\xi = [w, \mu_{s1}, \ldots, \mu_{sn}]'. \quad (3)$$

w is an offset term that may represent when the feature vectors differ from the mean vector by an additive term such as in a different recording environment [4]. With this transformation, the mixture density function becomes

. (4)

$$b_j(O) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(o - W_s\mu_s)'\Sigma_s^{-1}(o - W_s\mu_s)}$$

3.2. MLLR Derivation

The derivation of MLLR produces a maximum likelihood estimate of all W_s transformation matrices to maximize the likelihood of the adaptation data given the adapted models. Rather than replicate the lengthy derivation here, we point the reader to [14, 4] which give the extensive details of the derivation.

One point to note from the derivation is that, in the MLLR framework, finding a closed-form solution to the transformation matrices is not possible. Thus, the MLLR approach will only work with diagonal covariance matrices. Also, the computational cost of the method given is very high but can be reduced if the transformation matrix is restricted to a diagonal form. Ignoring the offsets, all matrices can be reduced to diagonal making matrix inversion trivial.

4. SUMMARY

MLLR is one of many transform-based methods used for speaker adaptation. MLLR has the distinction of being the research industry standard method. MLLR has been shown to be an effective method for speaker adaptation on every task to which it has been applied including SWITCHBOARD, Broadcast News, and smaller vocabulary tasks so there it is easy to understand the reason it has become a default setting in most research recognition systems. The drawback is that the computations required to build a robust set of adaptation transformations eliminates its usefulness for consumer-level applications.

5. ACKNOWLEDGEMENTS

I'd like to express my gratitude to all that have helped me in so many ways over the past three years. Of particular importance have been Dr. Joe Picone, Aravind Ganapathiraju, and Neeraj Deshmukh. I would also like to commend the patience of all those that had to put up with me saying "not now!" during the writing of this paper.

REFERENCES

[1] C. J. Leggetter, and P. C. Woodland, "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression," *Proceedings of the ARPA Spoken Language Technology Workshop, Barton Creek, 1995.*

[2] X. D. Huang and K. F. Lee, "On Speaker-Independent, Speaker-Dependent, and Speaker-Adaptive Speech Recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 877-880, Toronto, Canada, 1991.

[3] J. Godfrey, E. Holliman and J. McDaniel, "Telephone Speech Corpus for Research and Development," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 517-520, San Francisco, California, USA, March 1992.

[4] H. Christensen, "Speaker Adaptation of Hidden Markov Models using Maximum Likelihood Linear Regression," Ph.D. Thesis, Institute of Electronic Systems, Department of Communication Technology, Aalborg University, 1996.

[5] C. J. Leggetter and P. C. Woodland, "Speaker Adaptation Using Linear Regression, Technical Report CUED/F-INFENG/TR.181", Cambridge University Engineering Department, June 1994.

[6] Y. Grenier, "Speaker Adaptation through Canonical Correlation Analysis," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, pp. 888-891, Denver, Colorado, USA, 1980.

[7] Y. Grenier, L. Miclet, J. C. Maurin, and H. Michel, "Speaker Adaptation for Phoneme Recognition, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, pp. 1273-1275, Atlanta, Georgia, USA, 1981.

[8] F. S. Gurgen and H. C. Choi, "On the Frame-Based and Segment-Based Non-linear Spectral Transformation for Speaker Adaptation," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 534-539, Perth, Australia, December 1994.

[9] K. Choukri, G. Chollet, and Y. Grenier, "Spectral Transformations through Canonical Correlation Analysis for Speaker Adaptation in ASR," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4, pp. 2659-2662, Tokyo, Japan, 1986.

[10] F. Kubala, R. Schwartz, and C. Barry, "Speaker Adaptation from a Speaker Independent Training Corpus," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 137-140, Albuquerque, NM, USA, 1990.

[11] H. C. Choi and R. W. King, "A Two-Stage Spectral Transformation Approach to Fast Speaker Adaptation," *Proceedings of Speech Science and Technology*, Vol. 2, pp. 540-545, Perth, Australia, December 1994.

[12] C. H. Lee, C. H. Lin, and B. H. Juang , "A Study on Speaker Adaptation of Continuous Density HMM Parameters," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 145-148, Toronto, Canada, 1991.

[13] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, 39:1-38 Series B, 1977.

[14] C. J. Leggetter, "Improved Acoustic Modeling for HMMs using Linear Transformations," Ph. D. Thesis, Cambridge University, 1996.

[15] A. J. Hewett, "Training and Speaker Adaptation in Template-based Speech Recognition," Ph.D. Thesis, Cambridge University, 1989.

[16] Y. Hao, and D. Fang, "Speech Recognition Using Speaker Adaptation by System Parameter Transformation," *IEEE*

Transactions on Speech and Audio Processing, Vol. 2, No. 1, Part 1, pp. 63-68, January 1994.

[17] S. Nakamura and K. Shikano, "Speaker Adaptation Applied to HMM and Neural Networks," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 89-92, Glasgow, Scotland, 1989.