ACOUSTIC ADAPTATION USING NONLINEAR TRANSFORMATIONS OF HMM PARAMETERS

Victor Abrash, Ananth Sankar, Horacio Franco, Michael Cohen

SRI International, Speech Research and Technology Laboratory, Menlo Park, CA, 94025
Email: {victor, ananth, hef, mcohen}@speech.sri.com

ABSTRACT

Speech recognition performance degrades significantly when there is a mismatch between testing and training conditions. Linear transformation-based maximum-likelihood (ML) techniques have been proposed recently to tackle this problem. In this paper, we extend this approach to use nonlinear transformations. These are implemented by multilayer perceptrons (MLPs) which transform the Gaussian means. We derive a generalized expectation-maximization (GEM) training algorithm to estimate the MLP weights. Some preliminary experimental results on nonnative speaker adaptation are presented.

1. INTRODUCTION

Automatic speech recognition (ASR) performance degrades significantly when there is a mismatch between the testing and training conditions. State-of-the-art speaker-independent (SI) recognizers are usually trained with a large number of clean (nonnoisy) utterances from many native speakers. However, such systems typically perform poorly in situations unlike their training environment, such as nonnative speech, speech in the presence of acoustic or channel noise, or speech recorded with different microphones. It is clear that ASR systems must be more robust to such variations in order to maintain reasonable performance in practical applications.

Recent work [1-6] has shown that under mismatched conditions, the performance and robustness of speech recognition systems can be improved by adapting the speech models to the speaker, the channel, or the microphone. Adaptation can be achieved using Bayesian [1,6] approaches or transformation-based maximum-likelihood (ML) approaches [1-5]. The goal is to make the model better match the test data, reducing the performance degradation caused by the mismatch. In this paper, we are interested in transformation-based ML approaches to adapt the speech models. The form of the transformation is decided upon, and its parameters are estimated by maximizing the likelihood of the adaptation data using the expectation-maximization (EM) algorithm [7].

Previous approaches to transformation-based ML adaptation [1-5] have used linear transformations in either the feature or the model space. In many real scenarios, the assumption of a linear transformation is inadequate. For example, it is known that additive noise in the spectral domain results in a nonlinear distortion in the cepstral domain. It also appears limiting to assume that the mismatch between native and nonnative speech can be modeled by a linear transformation.

In this paper, we extend the previous linear transformation based adaptation approaches [1-5] to use nonlinear transformations. The only requirement for the nonlinear transformations is that they be differentiable with respect to their parameters. In this paper, the nonlinear transformations are implemented by multilayer perceptrons (MLPs) that map the means of the original hidden Markov model (HMM) Gaussian observation densities. The parameters of these transformations are estimated using the EM algorithm so as to maximize the likelihood of the adaptation data. Closed-form reestimation formulae are not usually possible for nonlinear transformations. In order to solve this problem, we use a generalized expectation-maximization (GEM) learning algorithm which makes use of gradient ascent training at each M step of the EM algorithm.

In Section 2 we describe the nonlinear adaptation approach. Some preliminary experimental results on nonnative speaker adaptation are presented in Section 3. Discussion of the results are presented in Section 4.

2. ADAPTATION APPROACH

We assume that the means of the Gaussians of the HMMs are transformed by a nonlinear function:

$$\hat{\mu}_{mg} = f_g(\mu_{mg}) \tag{1}$$

In this paper, f_g is implemented by an MLP. The parameters of the MLP are estimated by maximizing the likelihood of the adaptation data. In order to solve the ML problem, we need to compute the state conditioned observation probability density functions (pdfs) of the HMMs. Using Equation 1, we can write this pdf as

$$p_{SA}(y_t|s_t) = \sum_m p(\omega_m|s_t) N(y_t; f_g(\mu_{mg}), \Sigma_{mg}) \quad (2)$$

where g is the index of the Gaussian codebook used by the HMM state s_t, m is the mixture index within this codebook, and $p(\omega_m|s_t)$ are the mixture weights. Only the parameters of $f_g(\mu_{mg}), g = 1, \ldots, N_g$ are estimated during adaptation, where N_g is the number of codebooks. Transformations can be shared by more than one codebook, in a fairly straightforward extension to the derivation presented here. Transformations are tied using the method described in [2].

To obtain ML estimates of the parameters of the transformation $f_g(\mu)$ we use the GEM algorithm [7]. This algorithm is a two step procedure. In the first step we compute the auxiliary function Q, which in the case of multiple Gaussian mixtures with diagonal covariance matrices is written as

$$Q(\theta;\theta_o) = \sum_{g \in G, m \in g, t} \frac{\gamma_t(g,m)}{2} \sum_{d=1}^{D} \left[\frac{y_t - f_g(\mu_{mg})}{\sigma_{gm}}\right]^2_{(d} \quad (3)$$

where θ_o and θ are the original and reestimated models respectively, the transformation f_g is tied over all mixtures m in all codebooks $g \in G$, G defines the transformation tying, $y_{(d)}$ refers to the d-th component of the vector y_t, and

$$\gamma_t(g,m) \equiv p(s_t \in g, \omega = m | y_t, \theta) \quad (4)$$

is the posterior joint probability that the state s_t is a member of genone g and the mixture is m, given the acoustic observation and the speech model.

In the second step of the GEM algorithm, we estimate the parameters of the transformation function $f_g(\mu)$ in order to increase the value of the auxiliary function. These parameters are then used to recompute the value of the auxiliary function, and the algorithm proceeds iteratively.

It is shown in [7] that this iterative procedure is guaranteed to increase the likelihood of the adaptation data. The difference between the "regular" EM algorithm and the GEM algorithm is that in the former the auxiliary function is maximized at each step, rather than just being increased.

The value of Q is increased by gradient ascent techniques. By taking the gradient of $Q(\theta;\theta_o)$ with respect to the transformation parameters θ, we define the following learning rule:

$$\theta_g(k+1) = \theta_g(k) + \eta(\nabla_{\theta_g} Q(\theta)), \quad (5)$$

where θ_g is some parameter in f_g, k is the gradient iteration, and η is the learning rate. The gradient is computed using the chain rule,

$$\nabla_{\theta_g} Q(\theta) = \frac{\partial Q}{\partial f_{\theta_g}} \cdot \frac{\partial f_{\theta_g}}{\partial \theta_g} \quad , \quad (6)$$

which can be expanded as

$$\nabla_{\theta_g} Q(\theta) = \sum_{g \in G, m \in g, t} \frac{\gamma_t(g,m)}{\sigma^2_{gm}} \sum_{d=1}^{D} [y_t - f_g(\mu_{mg})]_{(d)} \frac{\partial f_{g(d)}}{\partial \theta_g}. \quad (7)$$

In this work, the nonlinear function f_g is implemented with an MLP. The factor $\frac{\partial f_{g(d)}}{\partial \theta_g}$ is computed using error back-propagation. An MLP is not required; any function that is differentiable with respect to its parameters can be used.

The complete GEM algorithm for training the transformation parameter values is summarized as follows:

1. Initialize all transformations as close as possible to the identity transform, $f_g(0) \approx I, g = (1, \ldots, N)$.
2. **E-step:** Perform one iteration of the forward-backward algorithm on the speech data, using Gaussians transformed with the current value of the transformations. For all component Gaussians and all mixtures g, collect the sufficient statistics $\gamma_t(g,m)$ and $\gamma_t(g,m)y_t$, and compute Q.
3. **M-step:** Repeat the following gradient ascent loop until Q is approximately maximized in step (d):

   (a) Compute $\nabla^k_{\theta_g} Q(\theta)$.

   (b) Compute $\theta_g(k+1) = \theta_g(k) + \eta(\nabla_{\theta_g} Q(\theta))$.

(c) Set $\hat{\mu}_{mg} = f_g(\mu_{mg})$.

(d) Compute $Q(\theta_{(k+1)};\theta_{(k)})$.

4. Repeat steps 2-3 until the overall likelihood of the adaptation data converges, or for a predetermined number of iterations.

3. EXPERIMENTS

3.1. Experimental Setup

SRI's DECIPHER speech recognition system was used for all experiments. The HMMs consisted of gender-dependent genonic models (with an arbitrary degree of Gaussian sharing across different HMM states [8]). These models were trained with about 140 speakers and 17,000 Wall Street Journal (WSJ) sentences for each gender, and consisted of 12,000 context-dependent phonetic models sharing 500 Gaussian mixture codebooks with 32 Gaussians in each mixture. The input to each Gaussian was a 39 dimensional cepstral vector, consisting of 12 cepstral coefficients, cepstral energy, and their first and second differences. The cepstra were normalized with cepstral mean subtraction.

Adaptation was performed on the "spoke 3" development and test sets of the WSJ corpus [9], consisting of read speech from nonnative speakers of American English. For these experiments, the data was divided into 20 adaptation and 20 test sentences. Recognition used the standard 5,000-word, closed-vocabulary bigram language model. Results are reported using the five male speakers in the 1994 S3 development set.

3.2. Transformation Architecture

The nonlinear transformation was implemented by a MLP in parallel with the standard linear transformation, as shown in Figure 1. This is equivalent to an overall MLP with direct connections between its input and output layers. The overall transformation has 39 linear input units, a varying number of sigmoidal hidden units, and 39 linear output units.

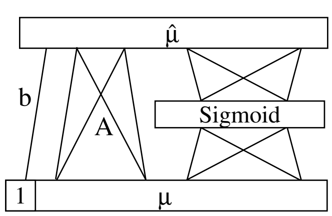

Figure 1: Linear and nonlinear transformation components.

By setting certain connections appropriately, several different transformations can be computed with this architecture:

1. **Linear**:

$$\hat{\mu}_{mg} = A\mu_{mg} + b$$

2. **Linear + Nonlinear**:

$$\hat{\mu}_{mg} = A\mu_{mg} + b + MLP(\mu_{mg})$$

In all cases, the transformation is initialized to compute the identity function, that is, $A = I, b = 0$, and the weights to and from the sigmoid units are set to small random values.

Twenty five, fifty, seventy five, and one hundred hidden units were tried, and it was found that fifty hidden units gave the best response. Therefore, results with other numbers of hidden units are not reported.

3.3. Preliminary Results

Transformation Type	Number of Transforms		
	1	5	40
SI	24.9	24.9	24.9
Linear (R)	22.5	18.9	16.4
Linear (G)	22.0	19.0	16.6
Linear+Nonlinear (G)	22.0	18.8	16.5

Table 1. Word Error Rates (percent) for before and after adaptation with different numbers of transformations.

The first row of Table 1 shows the speaker independent performance test set. In the second row, the (R) indicates that linear transformation parameters were estimated with the EM algorithm as described in [3, 4] for 1, 5, and 40 transformations. In the third row, linear transformation parameters were estimated using the GEM algorithm, as indicated by the (G). Finally, in the last row, linear and nonlinear transformation parameters were learned simultaneously.

From the table, we can see that all three approaches performed comparably. We therefore decided to try to initialize the GEM training presented in this paper from the final conditions obtained from the EM training (row 2). The motivation for this was to see if we could improve performance based on a good starting point.

In Table 2, linear transformation parameters are estimated with the EM algorithm (row 2), saving the parameters for later use. In the last row, these values are read in, a nonlinear component is added, and both the linear and nonlinear parameters are updated using GEM. From the table, we notice a modest improvement from adding the nonlinear component.

Transformation Type	Number of Transforms		
	1	**5**	**40**
SI	24.9	24.9	24.9
Linear (R)	22.5	18.9	16.4
Linear (R) Linear+Nonlinear (G)	22.4	18.3	16.0

Table 2. Word Error Rates (percent) for before and after adaptation with different numbers of adaptation. In the bottom two rows, the linear transformation component was initialized with the output of the second row.

4. DISCUSSION

In this paper we have presented a novel nonlinear transformation based adaptation algorithm. The nonlinearity was implemented by a multilayer perceptron, whose weights were estimated using a generalized EM algorithm. This technique made use of gradient ascent training embedded within each maximization step of the EM algorithm.

Experimental results were presented for nonnative speaker adaptation. So far, our results show a modest improvement using the nonlinear technique as compared to previous linear approaches.

We are continuing to explore this technique to see if we can further improve our performance. We also intend to investigate the application of nonlinear adaptation of HMM parameters for other types of acoustic mismatches, such as noisy speech.

ACKNOWLEDGMENTS

We gratefully acknowledge support for this work from ARPA through Office of Naval Research Contract N00014-94-C-0181. The Government has certain rights in this material. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Government funding agencies.

REFERENCES

1. V. Digalakis and L. Neumeyer, “Speaker Adaptation Using Combined Transformation and Bayesian Methods,” 1995 IEEE ICASSP, pp. I-680 - I-683.
2. V. Digalakis, D. Rtischev, and L. Neumeyer, “Fast Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures,” *IEEE Trans. on Speech and Audio Processing;* Vol. 3, No. 5, pp. 357—366, 1995.
3. C.J. Leggetter and P.C. Woodland, “Flexible Speaker Adaptation using Maximum Likelihood Linear Regression,” ARPA SLT Workshop, pp. 110-115, January 1995.
4. L. Neumeyer, A. Sankar, and V. Digalakis, “A Comparative Study of Speaker Adaptation Techniques,” Proceedings EUROSPEECH, 1995.
5. A. Sankar and C. H. Lee, “A Maximum Likelihood Approach to Stochastic Matching for Robust Speech Recognition”, Accepted for *IEEE Trans. on Speech and Audio Processing*.
6. C.-H. Lee and J.-L. Gauvain, “Speaker Adaptation Based on MAP Estimation of HMM Parameters,” 1993 IEEE ICASSP, pp. II-558 — II-561.
7. 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.
8. V. Digalakis and H. Murveit, “GENONES: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer,” 1994 IEEE ICASSP, pp. I537-I540.
9. F. Kubala et al., “The Hub and Spoke Paradigm for CSR Evaluation,” Proceedings of the ARPA Human Language Technology Workshop, 1994, pp. 37-42.