CONNECTIONIST
SPEAKER NORMALIZATION AND ADAPTATION

Victor Abrash, Horacio Franco, Ananth Sankar, and Michael Cohen

SRI International
Speech Technology and Research Laboratory
Menlo Park, CA, 94025, USA

ABSTRACT

In a speaker-independent, large-vocabulary continuous speech recognition systems, recognition accuracy varies considerably from speaker to speaker, and performance may be significantly degraded for outlier speakers such as nonnative talkers. In this paper, we explore supervised speaker adaptation and normalization in the MLP component of a hybrid hidden Markov model/ multilayer perceptron version of SRI's DECIPHER^{TM} speech recognition system. Normalization is implemented through an additional transformation network that preprocesses the cepstral input to the MLP. Adaptation is accomplished through incremental retraining of the MLP weights on adaptation data. Our approach combines both adaptation and normalization in a single, consistent manner, works with limited adaptation data, and is text-independent. We show significant improvement in recognition accuracy.

1. INTRODUCTION

In a speaker-independent (SI), large-vocabulary continuous speech recognition system, recognition accuracy varies considerably from speaker to speaker, and performance is significantly degraded for outlier speakers such as nonnative talkers. Techniques that have been used to improve the performance and robustness of speech recognition systems to these variations include adapting the speech models to the new speaker, or normalizing the input speech of new speakers to some canonical or prototype representation. In the hidden Markov model (HMM) speech recognition paradigm, adaptation typically means modifying the speaker-independent model parameters based on limited enrollment data from a new speaker.

Speaker *adaptation* can be performed in a Bayesian framework [1] with the prior information encapsulated in the speaker-independent models, or in maximum likelihood (ML) transform-based approaches [2][3] that map the new speaker's acoustic feature space to the training feature space. Mapping can be accomplished by transforming either the input speech features or the speech models. Transformation approaches are simple and amenable to fast adaptation with little enrollment data. The Bayesian approach usually has the desirable property of converging asymptotically to speaker-dependent recognition performance as the amount of adaptation data increases.

Speaker *normalization* [4] usually refers to mapping the new speaker's speech features to those of a training speaker, similar to some algorithms in the transform-based adaptation field. Many approaches are text-dependent, calling for the new speaker to record sentences with prespecified transcriptions so they can be time-warped and aligned with utterances from reference speakers.

At SRI, we are working with a hybrid HMM/ MLP version of DECIPHER^{TM}, which uses a single large multi-layer perceptron (MLP) network to estimate the state-dependent observation likelihoods of the HMM [5][6]. This design provides a convenient framework for speaker adaptation or normalization because transformations can be applied directly to input speech features, the speech model (MLP weights), or both. Many parameters are shared so less enrollment data are required to robustly adapt the weight values. This work explores supervised adaptation and normalization in the MLP component of our hybrid system, combining both approaches in a single, consistent manner. The approach works with limited adaptation data and is text-independent.

2. APPROACH

2.1. Normalization

Our approach to speech normalization is based on defining a "transformation network" (TN) that acts as a preprocessor to the main MLP network. The goal of the TN is to map the incoming speech into a "better" representation, one that enhances the ability of the main MLP classifier to compute the *a posterior* phonetic probabilities.

The architecture of the transformation network can be defined with different degrees of complexity, tying, and number of parameters, depending on the amount of adaptation data available. A characteristic of the TN is that the same transformation is shared among all the input frames so its parameters are tied across frames.

Advantages to this architecture include: a compact representation of multiple speakers; quick adaptation of TN parameters with small amounts of adaptation data; separate sub-networks for modeling speaker-dependent and speaker-independent characteristics; and more robust training of a small number of parameters with limited adaptation data.

In our initial experiments, we define our TN to compute $\hat{Y}_t$, the normalized speech feature, using a general linear transformation, $\hat{Y}_t = A \cdot Y_t + b$. The normalization parameters to learn are the elements of the matrix A and the vector b, a total of 702 parameters for each speaker.

The weights of our TN are obtained using a supervised adaptation paradigm, where the target signal was obtained from Viterbi forced alignments to the transcriptions of the adaptation sentences. Prior to training, the TN was initialized to the identity matrix, $A = I$ (i.e., $\hat{Y}_t = Y_t$). As in [4], the TN was trained by backpropagating phonetic classification errors through the main MLP without changing the parameters of the main network; only the TN weights were modified during training.

In Figure 1, we show the combined network architecture of the TN and the MLP. The TN lies between the speech input vectors and the speaker-independent MLP.

2.2. Adaptation

In our system, the speaker-independent model is formed by the connectivity of the MLP classifier, plus the values of its weights. For this experiment, we simply adapted the previously trained MLP, without any transformation network. Adaptation sentences are used to retrain the MLP classifier, with the speaker-independent weight values providing the initial training condition.

This alternative, which is in essence more straightforward than the previous normalization experiment, has the potential for greater performance gains since our MLP architecture provides a larger amount of modeling power than the simple TN used in these experiments. However, obtaining good generalization could be more difficult because of the large number of parameters (300,000 weights) in the MLP.

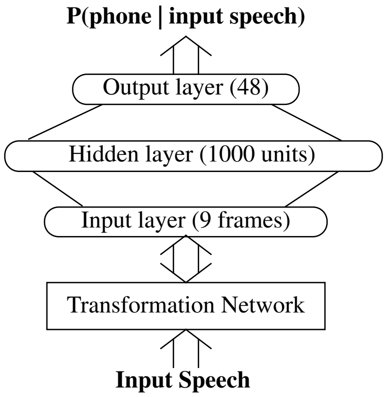

Figure 1: Full network formed by the cascade of a speech transformation network (TN) and an MLP phonetic classifier.

2.3. Joint Normalization and Adaptation

If improvements from normalization and adaptation are independent, we expect to see a gain from combining both methods. For this experiment, the MLP and the TN were jointly trained using the same scheme as before. The MLP was initialized with the speaker-independent weights, the TN was initialized with values corresponding to the identity transformation, and all weights in the combined system were updated using a small learning rate.

3. EXPERIMENTS

SRI's DECIPHER^{TM} speech recognition system was used for all experiments. Adaptation and normalization experiments were performed on the "spoke 3" and "spoke 4" development and test sets of the WSJ corpus [7], including read-speech from both native and nonnative speakers of American English. Recognition used the standard 5,000-word, closed-vocabulary bigram and trigram language models. The bigram language model was used to generate the 1,000 most likely N-best sentence hypotheses, which were then rescored with the trigram language model. Only the MLP was used to estimate acoustic observation likelihoods; the genonic HMMs were used only to bootstrap the neural network training and adaptation.

The bootstrap HMM models 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* 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 length-39 cepstral vector, consisting of 12 cepstral coefficients, cepstral energy, and their first- and second-differences. The cepstra were normalized with cepstral mean subtraction.

Viterbi alignments from these models were used to train an MLP with a 9-frame input window (234 inputs), 1,000 hidden units, and 48 outputs corresponding to the context-independent phonetic classes used by DECIPHER^{TM} (See [5] for more details). The MLP input vector for each frame was a length-26 vector of cepstral mean-subtracted coefficients, consisting of 12 cepstral coefficients, cepstral energy, and their first-differences. The MLP inputs were further normalized to be zero mean and unit variance.

The "spoke 3" data, which were spoken by nonnative talkers, were divided into 40 adaptation and 40 test sentences. The "spoke 4" data were spoken by native English talkers. The 100 sentences from each speaker were[7] arbitrarily divided into 50 adaptation and 50 test sentences. The adaptation/training sentences were further divided into three-quarters adaptation and one-quarter cross-validation sentences. Adaptation was halted when cross-validation performance ceased to improve.

3.1. Nonnative Speakers

Table 1 presents 1993 S3 development data results for two language models. The MLP baseline result is the word error of our hybrid HMM/MLP recognizer, a fully context-independent (CI) system. The MLP computed all observation likelihoods. The second and third columns show the word error rates obtained by rescoring the N-best sentence hypotheses with the bigram and trigram language models respectively. Normalization results (Section 2.1) are reported as "Adapt TN", adaptation results (Section 2.2) as "Adapt MLP", and joint normalization and adaptation (Section 2.3) as "Adapt TN & MLP".

On this outlier data, the speaker-independent MLP performed very badly. The TN alone improved performance 34% using the trigram language model, an encouraging result given the simplistic nature of the transformation used and the small number of additional parameters. Adapting the MLP weights (with no TN) yielded a 46% reduction in error, a somewhat surprising result; we initially expected that with the small amount of adaptation data available compared to the number of parameters, the system would be overtrained even using a cross-validation stopping criterion. Joint adaptation of both the TN and the MLP gave no further improvement.

Experiment:	Bigram	Trigram
SI HMM	24.55	21.52
SA HMM	11.42	8.70
Baseline MLP	33.34	28.47
Adapt TN	23.39	19.24
Adapt MLP	19.09	15.65
Adapt TN & MLP	19.04	15.59

Table 1: Word error rate for 5 male nonnative speakers from the 1993 WSJ S3 development set, for bigram and trigram language models.

The performance of the adapted MLP recognizer is better than the speaker-independent genone-based HMM system, but significantly worse than the speaker-adapted HMM [1][2], which is not surprising since we are comparing a triphone-based, context-dependent system to a simple, context-independent one.

Table 2 shows recognition performance by speaker, with the trigram language model (LM). By-speaker results are similar with the bigram LM. All speakers improved significantly; the worse their recognition performance was initially, the greater the effect of normalization or adaptation. Normalization is better than adaptation only for speaker 4na, who with acceptable SI recognition accuracy is less of an outlier speaker than the others.

Experiment:	Speaker					Avg
	4n0	4n3	4n5	4n9	4na	
Baseline	33.7	26.2	26.8	21.4	13.9	24.5
Adapt TN	25.9	21.4	24.6	14.8	9.1	19.2
Adapt MLP	19.5	14.2	21.5	11.7	11.2	15.7
Adapt TN & MLP	19.6	15.1	21.7	12.3	9.1	15.6

Table 2: Word error rate for 5 male nonnative speakers from 1993 WSJ S3 development set, by speaker, generated by the trigram language model.

3.2. Native Speakers

Table 3 and Table 4 show results for equivalent adaptation and normalization experiments conducted with *native* speakers of American English. For the 1993 data and trigram LM, the word error rate decreased by 20.7%, 15.3%, and 22.3% respectively for the normalization, adaptation, and joint cases. Recognition errors were reduced 15.2%, 15.3%, and 18.7% for the 1994 data. The improvement is smaller for the latter dataset because the acoustic and language model weights were optimized on the 1993 dataset, and then applied to the 1994 speakers.

These are significant improvements for native speakers, especially compared to Maximum Likelihood (ML) techniques reported in [1] and [2]. The error reduction is less than the outlier nonnative speakers because the SI models were already a fairly good match to the natives.

Experiment:	Bigram	Trigram
Baseline	25.36	21.77
Adapt TN	20.27	17.26
Adapt MLP	22.09	18.44
Adapt TN & MLP	20.35	16.91

Table 3: Word error rate for 4 male native speakers from the 1993 WSJ S4 development and test sets.

Experiment:	Bigram	Trigram
Baseline	30.54	27.10
Adapt TN	26.02	22.99
Adapt MLP	26.44	22.96
Adapt TN & MLP	25.00	22.04

Table 4: Word error rate for 4 male native speakers from the 1994 WSJ S4 development and test sets.

The results indicate that speaker normalization alone, using a simple transformation network, is as good as either adapting the MLP or jointly adapting the MLP and training a TN; the difference between these three cases is not statistically significant. We achieved a 15%-20% improvement in accuracy with only 706 more weights, or an additional 0.25% parameters. This improvement can be compared to the nonnative experiments, where applying the TNs did not help as much as adapting all the MLP weights.

4. DISCUSSION

From these experiments we see that a simple linear transformation with a small number of parameters can capture a significant amount of speaker-dependent characteristics to obtain a meaningful increase in performance, especially for non-outlier speakers. Recognition error with a trigram language model was reduced between 35% and 45% for nonnative speakers, and between 15% and 20% for natives. In both sets of experiments, joint adaptation of the MLP and TN produced a small additional gain which suggests that the type of transformation that the TN imposes may be useful even when we have a large number of adaptation parameters available.

For nonnative speakers, speaker adaptation (incrementally adapting the MLP weights) was most effective, while for natives, speaker normalization with a transformation network proved the most effective, lowest cost option.

We hypothesize that the difference between these results occurs because the speaker-independent MLP weights provide a poor model for the nonnatives, which can best be improved by *speaker adaptation* of the entire weight matrix, and any effects of applying a TN are lost because of poor modeling in the much larger MLP. For native, non-outlier speakers, on the other hand, the SI weights provide a reasonable model that is hard to improve with a small amount of adaptation data. In this case, we can apply our TN-based *speaker normalization* method, in combination with an acceptable SI model, to model speaker-dependent correlations in the cepstral features and achieve an inexpensive performance gain.

With a more sophisticated transformation we would likely be able to achieve a greater gain. In previous transformation-based adaptation algorithms [2][3], a relatively small number of adaptation parameters were estimated; in our approach, we can successfully adapt the whole set of model parameters, although applying multiple transformations meaningfully is more difficult. This approach could potentially lead to better performance than standard systems achieve where only a subset of the parameters are reestimated.

Future research potentially includes combining speaker-adapted MLP and genone likelihoods for a combined performance gain, or using more complex normalization architectures with multiple-frame inputs or acoustic-region-dependent transformation weights.

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*. ICASSP, 1995.
2. V. Digalakis, D. Rtischev and L. Neumeyer. *Fast Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures*. IEEE Trans. on Speech and Audio Processing; submitted April 1994 — accepted January 1995.
3. A. Sankar and C.H. Lee. *Stochastic Matching for Robust Speech Recognition*. IEEE Signal Processing Letters, Vol. 1, No. 8, August 1994.
4. R. Watrous. *Speaker Normalization and Adaptation Using Second-Order Connectionist Networks*. IEEE Transactions on Neural Networks, Vol. 4, No. 1, Jan. 1993, pp. 21-30.
5. M. Cohen, H. Franco, N. Morgan, D. Rumelhart, and V. Abrash. *Hybrid Neural Network/Hidden Markov Model Continuous Speech Recognition*. ICSLP, 1992.
6. S. Renals, N. Morgan, M. Cohen, and H. Franco. *Connectionist Probability Estimation in the DECIPHER Speech Recognition System*. ICASSP, 1992, pp. 601-604.
7. F. Kubala et al. *The Hub and Spoke Paradigm for CSR Evaluation*. Proceedings of the ARPA Human Language Technology Workshop, 1994, pp. 37-42.
8. V. Digalakis and H. Murveit. *GENONES: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer*. ICASSP, 1994, pp. I537-I540.