SPEAKER AND GENDER NORMALIZATION FOR CONTINUOUS-DENSITY HIDDEN MARKOV MODELS

Alejandro Acero and Xuedong Huang

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052, USA

ABSTRACT

In this paper we describe a speaker-cluster normalization algorithm that we applied to both gender-normalization and speaker-normalization. To achieve parameter sharing the acoustic space is partitioned into classes. A maximum likelihood approach has been proposed under which the delta between the distribution mean and its corresponding acoustic class is mostly speaker-independent, whereas the means of the acoustic classes are mostly speaker-dependent. When applied to gender-normalization, the error rate reduction approaches that of a gender-dependent system but with half the number of parameters. For a speaker-normalized system, a 30% decrease in error rate was obtained in a batch recognition experiment in a context-dependent continuous-density HMM system.

1. INTRODUCTION

The error rate of state-of-the-art large-vocabulary speaker-independent continuous-speech recognition systems is still too high to be used in practical systems. Speaker-dependent systems have error rates that are typically 3 times lower than those of speaker-independent systems, but they require large amounts of speaker-dependent training data, which may not be available. Speaker adaptation techniques offer the benefit of speaker-independent accuracy and gradual convergence to speaker dependent accuracy. Many of the techniques used for speaker adaptation are based on *Maximum a Posteriori* (MAP) estimation [2]. While they offer excellent convergence properties, they cannot adapt parameters of the models which are not observed in the adaptation data, thus fairly large amounts of data are generally required.

Different modeling techniques have been proposed to deal with the general tradeoff between *trainability* and *specificity*. Parameter sharing has been proposed at the kernel level (Gaussian sharing/tying mostly) and the state level as an effective method for decreasing the total number of parameters, and therefore increase the reliability of their estimates for a fixed amount of training data. In this paper we will describe a modeling technique that achieves parameter sharing by combining and extending recent work done in speaker adaptation and cepstral normalization.

The use of correlation among parameters [1][6][9] has been an effective parameter tying approach for speaker adaptation. These model adaptation methods generally apply a transformation only to the speakers in the testing phase. In this paper, we propose to apply the transformation to the speakers in the training data as well, thereby accomplishing *speaker normalization*. The transformation could be applied to a single speaker or to clusters of speakers, for example grouped by gender.

Cepstral Mean Normalization (CMN) [7][8] has been proposed to normalize differences in acoustical environments. In order to eliminate the dependency of the cepstral mean on the amount of noise included in the calculation, researchers have extended CMN to the computation of two means: one for noise and one for speech [4][8]. In this paper we will extend this method to a larger number of acoustic classes.

The proper combination of context-dependent and context-independent information [5] on one hand, and speaker-dependent and speaker-independent on the other is an important area of research.

2. ALGORITHM DESCRIPTION

For this study we assume Continuous-Density sub-word Hidden Markov Models (CD-HMM). We will focus on adapting the probability density functions (*pdf*), though transition probabilities could also be updated by other techniques such as MAP estimation.

Let's define two partitions of the acoustic space:

1. *Acoustic Classes*. We partition the acoustic space into R classes, where each class r contains a set of HMM states Ω_r that either represents a phonetic entity such as a context-independent phone, or a data-driven cluster of output probabilities.
2. *Speaker Clusters*. We partition the universe of speakers into L speaker clusters. For *gender-normalization* we use one cluster for male speakers and another for female speakers. For *speaker normalization* we use one speaker per speaker cluster.

Under these assumptions, we model the *pdf* of a *p*-dimensional input feature vector *x* for speaker cluster *l* in a state *n* belonging to class *r*, as a mixture of *M* Gaussian random vectors:

$$f_{r,n}^{l}(x) = \sum_{m=1}^{M} \omega_{n,m} N(x, \mu_{n,m}^{l}, C_{n,m}) \quad (1)$$

where the weights $\omega_{n,m}$ and the covariance matrices $C_{n,m}$ are speaker-independent. The mean vector $\mu_{n,m}^{l}$ is a function of the speaker cluster *l*, and therefore speaker-dependent, is modeled as

$$\mu_{n,m}^{l} = \mu_{r}^{l} + \delta_{n,m} \quad (2)$$

where μ_{r}^{l} is shared among the set of states Ω_r in class *r*, and it assumed to be speaker-dependent and the delta parameters $\delta_{n,m}$ are considered speaker-independent.

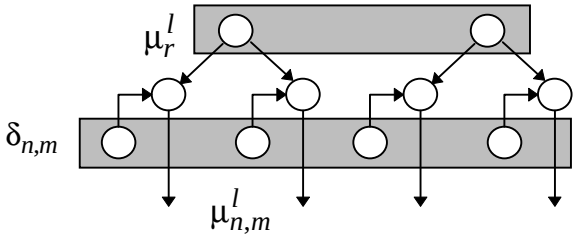

Figure 1. Hierarchical parameter tying of speaker-dependent and speaker-independent means.

The use of μ_{r}^{l} introduces an extra degree of freedom, so in order to obtain these estimates we need to force some other constraint, namely that the $\delta_{n,m}$ parameters average to 0.

2.1 Initial Estimates of Mean Parameters

For this study, we will assume that the covariance matrices $C_{n,m}$ are diagonal. Similar expressions can also be obtained for the general full covariance case.

Given a set of speaker-independent estimates for $\mu_{n,m}$, $\omega_{n,m}$ and $\sigma_{n,m}^{2}$, and a set of observations $\mathbf{X}_l = \{x_1^l, \ldots, x_{T_l}^l\}$ from speaker cluster *l*, we can obtain initial values for μ_{r}^{l} and $\delta_{n,m}$, by computing μ_r for each class *r* as

$$\mu_r[i] = \frac{\sum_{n\in\Omega_r} \sum_{t=1}^{T_l} \sum_{m=1}^{M} \gamma_t^l(n,m) x_t^l[i] / \sigma_{n,m}^2[i]}{\sum_{n\in\Omega_r} \sum_{t=1}^{T_l} \sum_{m=1}^{M} \gamma_t^l(n,m) / \sigma_{n,m}^2[i]} \quad (3)$$

where $\mu_r^l = \mu_r$ for $l = 1, \cdots, L$ and $\gamma_t^l(n,m)$ is the *a posteriori* probability of state *n* and Gaussian *m* for frame *t* of speaker cluster *l*, that can be computed through the forward-backward algorithm or through the Viterbi approximation.

The corresponding initial $\delta_{n,m}$ are then simply

$$\delta_{n,m} = \mu_{n,m} - \mu_r \quad (4)$$

2.2 Iterative Estimation of Mean Parameters

Once we have some initial estimates, we can compute the maximum likelihood estimates through the EM algorithm:

1. *Estimate* the *a posteriori* probabilities $\gamma_t^l(n,m)$ given μ_r^l, $\delta_{n,m}$ and $\sigma_{n,m}^2$ and the training data.

2. *Maximize* the log-likelihood, by finding updated estimates $\tilde{\mu}_r^l$, $\tilde{\delta}_{n,m}$ and $\tilde{\sigma}_{n,m}^2$ given by:

$$\tilde{\mu}_r^l[i] = \frac{\sum_{n\in\Omega_r} \sum_{t=1}^{T_l} \sum_{m=1}^{M} \gamma_t^l(n,m)\left(x_t^l[i] - \delta_{n,m}[i]\right) / \sigma_{n,m}^2[i]}{\sum_{n\in\Omega_r} \sum_{t=1}^{T_l} \sum_{m=1}^{M} \gamma_t^l(n,m) / \sigma_{n,m}^2[i]}$$

$$\tilde{\delta}_{n,m}[i] = \frac{\sum_{l=1}^{L} \sum_{t=1}^{T_l} \gamma_t^l(n,m)\left(x_t^l[i] - \tilde{\mu}_r^l[i]\right)}{\sum_{l=1}^{L} \sum_{t=1}^{T_l} \gamma_t^l(n,m)} \quad (5)$$

$$\tilde{\sigma}_{n,m}^2[i] = \frac{\sum_{l=1}^{L} \sum_{t=1}^{T_l} \gamma_t^l(n,m)\left(x_t^l[i] - \tilde{\mu}_r^l[i] - \tilde{\delta}_{n,m}[i]\right)^2}{\sum_{l=1}^{L} \sum_{t=1}^{T_l} \gamma_t^l(n,m)}$$

3. Prepare for next iteration:

$$\tilde{\mu}_r^l = \mu_r^l, \quad \tilde{\delta}_{n,m} = \delta_{n,m} \quad \text{and} \quad \tilde{\sigma}_{n,m}^2 = \sigma_{n,m}^2 \quad (6)$$

4. Stop if reached convergence, otherwise go to step 1.

We have to note that the refined estimates in step 2 are the solution to a two-step maximization process: first compute $\tilde{\mu}_r^l$ that maximizes the log-likelihood and with these values compute $\tilde{\delta}_{n,m}$ and $\tilde{\sigma}_{n,m}^2$.

2.3 Relationships with other methods

If the number of classes *R* equals *1* and each utterance is considered to come from a different speaker cluster, the proposed method is equivalent to CMN but with a ML estimation approach, as proposed in [8] by Sankar

and Lee. In fact, these authors also applied to the case of two classes: one for noise and one for speech, similarly to [4].

The assumptions in Eq. (1) and (2) implicitly imposes a correlation among the means of all *pdfs* within a class *r*, but it is not a full correlation matrix. In [6], the authors mention that if enough adaptation data is available, a full correlation matrix provides some improvement over a diagonal correlation matrix. In our approach, we intend to use this technique with small amounts of speaker-dependent data, so the smaller the number of adaptation parameters the better.

3. ML GENDER NORMALIZATION

Given the differences existing between male and female voices, many systems currently estimate two sets of models, one for male speakers and another for female speakers. Gender selection is done at run-time by selecting the model with highest probability. This approach [3], results in a modest error rate at the expense of doubling the size of the acoustic models and fragmenting the training data.

Given a set of gender-independent HMM models and given a partition with *R* classes, the iterative maximum likelihood procedure described in Sec. 2.2 is applied for $L = 2$ to estimate the *R* gender-dependent means μ_r^l for both male and female clusters, as well the gender-independent parameters $\delta_{n,m}$, $\omega_{n,m}$ and $\sigma_{n,m}^2$.

Since generally *R* << *N*, (*i.e.* the number of acoustic classes is much smaller than the total number of clustered states), this ML Gender Normalization algorithm increases only slightly the total number of parameters in the system over the case of gender-independent model. As a comparison, for gender-dependent models the number of parameters is doubled. Therefore, this method has the potential to offer a lower error rate, either with a better trained set of models for a given number of senones, or with a more detailed acoustic model (*i.e.* larger number of senones) since the training data is not fragmented.

In recognition, we have to do *gender selection* for each utterance. This could be achieved by running both male and female models in parallel and selecting the one with highest likelihood. In practice other less expensive methods based on VQ [3] can be used to select the male/female cluster.

4. ML SPEAKER NORMALIZATION

A similar approach can be followed to normalize speaker differences. In this case each speaker has its own cluster. Since the amount of data available from each speaker will be smaller than in the gender normalization case, the number of classes *R* has to be chosen to assure trainability. The training phase is the same as described in Sec. 2.2, where *L* equals the total number of speakers in the training database.

In recognition, we can follow two different approaches:

1. *Speaker selection*. Similarly to the approach taken in gender normalization, we can select one of the speakers in the training phase as the speaker that best matches the incoming speech. This has the advantage that it could require very little speech, but it could computationally expensive if the number of speakers R is large.
2. *Speaker means training*. The speaker-dependent means are trained from speech from the target speaker.

In the latter case, the speaker means could be obtained in two different ways:

1. *Adaptation mode*. In this case, the speaker provides some adaptation data used to train the speaker means, which are then used on new test data.
2. *Batch recognition mode*. In many typical dictation applications, the user does not need to see the results of the recognition after speaking each sentence, and can wait until he or she has finished speaking to see all the sentence transcriptions. In this case, unsupervised training of the speaker means can be performed on this data to refine the recognition results. One iteration on the data is generally sufficient.

5. EXPERIMENTAL RESULTS

For the experimental evaluation we used a context-dependent continuous-density HMM [5] with 54 phonemes and 3 states per phoneme. The system was trained using the Wall Street Journal corpus, and evaluated on the 5000-word closed-vocabulary task with a bigram language model. We evaluated the algorithm on the *si_dev5* data set, the November 1992 development set used in ARPA evaluations, which contains 410 utterances from 4 female and 6 male speakers. We used *R=162* classes, *i.e.* one class per context-independent state.

5.1 Gender Normalization Experiments

For the sake of faster turnaround, we used only 1000 senones, or clustered states, with 2 Gaussians per state. Training was done using only 2000 utterances of the Wall Street Journal corpus. The results can be seen in Table 1. The error rate of this baseline system was 14.9% on half of the *si_dev5* data set. The use of gender-dependent models decreased the error rate to 13.6%. Finally, using the method of gender normalization described in Section 3, the error rate was 13.8%.

With the proposed gender-normalization algorithm, we attain most of the error rate decrease of the gender-

dependent system but without doubling the number of parameters.

Method	BASE	GENDER DEP.	GENDER NORM.	SPKR. NORM
ERROR RATE	14.9%	13.6%	13.8%	10.6%

Table 1. Error rates of a baseline CD-HMM system with gender-independent model, gender-dependent models, gender-normalized models and speaker-normalized models.

5.2 Speaker Normalization Experiments

For this experiment we used the same reduced configuration described in Sec. 5.1. The speaker selection procedure described in Section 4 was not used, because it would be computationally expensive unless a small set of speaker clusters is chosen. Instead, we trained the speaker means.

To evaluate the batch mode described in Section 4, we ran the *si_dev5* data set through the speaker-independent system. With the transcriptions generated by this recognizer, we trained the speaker means μ_r^l for the *R=162* classes. Finally we re-ran the recognizer with the speaker-normalized models. The resulting error rate, showed in Table 1, dropped to 10.6%, which is an error rate reduction of 30% with respect to the baseline.

In the adaptation mode described in Section 4, the speaker means were trained from the other half of *si_dev5* not used in testing. The error rate did not improve significantly, perhaps because the trained speaker means were over-fitted to the data. To avoid that a smoother speaker means could be obtained with *MAP* techniques, and will be explored in future work.

6. CONCLUSION

We presented a speaker-cluster normalization algorithm that we applied to both gender-normalization and speaker-normalization. The rational for speaker-cluster normalization is to be able to model basic speaker differences, and yet to make use of all the available training data. The acoustic space is partitioned into several classes, where we assume that the class means are mostly speaker dependent and the difference between the state means and the class means are mostly speaker-dependent. Using a *Maximum Likelihood* estimation technique has shown a decrease in error rate for both gender and speaker normalization.

REFERENCES

[1] Cox, S. "Predictive Speaker Adaptation in Speech Recognition". *Computer, Speech and Language*. Jan 1995.

[2] Gauvain J.L., and Lee C.H. "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains". *IEEE Transactions on Speech and Audio Processing*. Apr 1994.

[3] Huang X., Lee K.F., Hon H.W. and Hwang M.Y. "Improved Acoustic Modeling with the SPHINX Speech Recognition System". *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Toronto, May 1991.

[4] Huang X., Acero A., Alleva A., Hwang M.Y., Jiang L. and Mahajan M. "Microsoft Windows Highly Intelligent Speech Recognizer: Whisper". *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Detroit, May 1995

[5] Huang X., Hwang M.Y, Jiang L. and Mahajan M. "Deleted Interpolation and Density Sharing for Continuous Density Hidden Markov Models". *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Atlanta, May 1996.

[6] Leggetter, C.J. and Woodland P.C. "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models". *Computer, Speech and Language*. Apr 1995.

[7] Liu F., Stern R., Huang X. and Acero A. "Efficient Cepstral Normalization for Robust Speech Recognition". *Proceedings of ARPA Human Language Technology Workshop*, March 1993.

[8] Sankar A. and Lee C.H. "Robust Speech Recognition Based on Stochastic Matching". *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Detroit, May 1995.

[9] Zavaliagkos G. "Batch, Incremental and Instantaneous Adaptation Techniques for Speech Recognition". *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Detroit, May 1995.