A COMBINATION OF DISCRIMINATIVE AND MAXIMUM LIKELIHOOD TECHNIQUES FOR NOISE ROBUST SPEECH RECOGNITION

Kari Laurila, Marcel Vasilache, Olli Viikki
Nokia Research Center, Speech and Audio Systems Laboratory, Tampere, Finland
Email: {kari.laurila, marcel.vasilache, olli.viikki}@research.nokia.fi

ABSTRACT

In this paper, we study how discriminative and Maximum Likelihood (ML) techniques should be combined in order to maximize the recognition accuracy of a speaker-independent Automatic Speech Recognition (ASR) system that includes speaker adaptation. We compare two training approaches for speaker-independent case and examine how well they perform together with four different speaker adaptation schemes. In a noise robust connected digit recognition task we show that the Minimum Classification Error (MCE) training approach for speaker-independent modelling together with the Bayesian speaker adaptation scheme provide the highest classification accuracy over the whole lifespan of an ASR system. With the MCE training we are capable of reducing the recognition errors by 30% over the ML approach in the speaker-independent case. With the Bayesian speaker adaptation scheme we can further reduce the error rates by 62% using only as few as five adaptation utterances.

1. INTRODUCTION

ASR has not been widely used until quite recently. During the four decades of research many important milestones have been reached. One of the milestones, the change from speaker-dependent speech recognition technology into speaker-independent technology was essential from the general acceptability of ASR point of view. Speaker-independent technology enables the direct use of ASR systems without users having to train the systems to recognize their voice.

A high recognition accuracy is also an essential requirement for an ASR system. Before a particular ASR service or product can be released to the market, the developers must go through a careful self-criticism process. The recognition accuracy must be at a certain level in order to gain users' acceptance. Speaker-independence brings up one problem regarding the recognition accuracy. This is due to the fact that the speaker-independence is not achieved by utilizing truly speaker-independent features in recognition. The speaker-independence is achieved by collecting speech samples from large amounts of people representing well the whole target population. Speaker-independent speech models are then created by effectively averaging the collected speech samples. As a result, the averaged model space becomes more confusable and it is well known that there is about an order of magnitude difference between the speaker-dependent and speaker-independent recognition accuracies.

The recognition accuracy problem of the speaker-independent case becomes even more severe when one practical limitation is still considered. Namely, the amount of collected speech samples is always finite, and all speaker types cannot be well represented in the training material. This means that there will be speakers for whom the recognition accuracy is much smaller than for the others due to language, dialect, pronunciation etc. variations. Nevertheless, a high speaker-independent recognition performance remains a fundamental objective of practical speech recognition systems. There are currently two ways to achieve this objective. The first alternative approach is to collect a huge amount of training data and create very complex speech models that can describe all speakers well enough. The second approach is to have less training data and to rely more on speaker adaptation.

In this paper, we have selected the latter approach. Our target is to cope with simple HMM structures and to find a combination of discriminative and Maximum Likelihood training schemes that maximizes the recognition accuracy during the whole lifespan of an ASR system with a relatively small amount of training and speaker-adaptation data. Noise robust and hands-free voice dialling being an attractive target application, this paper focuses particularly on noise robust connected digit recognition in a car environment.

2. SPEAKER-INDEPENDENT TRAINING

2.1 Maximum Likelihood Training Approach

The speaker-independent models are conventionally trained according to the Maximum Likelihood (ML) estimation principle, given below:

$$\max P(X|\lambda), \tag{1}$$

where λ is the model for the utterances *X*. The widely used ML estimation tries to maximize the likelihood of utterances in the training data independently model by model, and thus, the recognition accuracy is maximized only indirectly.

2.2 Discriminative Training Approach

In the literature, many discriminative training approaches have been suggested that try to maximize the recognition accuracy explicitly for a given vocabulary. In this paper, the Minimum Classification Error (MCE) and the Maximum Mutual Information (MMI) techniques were selected from the class of discriminative methods. Both of them can be regarded as a constrained optimization problem and they can be formulated similarly [1,3,7,9].

2.2.1 Minimum Classification Error Approach

In the MCE approach, the misclassification measure has the following formulation:

$$d(X) = -P_{\log}\left(X|\lambda_C\right) + \log\left[\frac{1}{N-1}\sum_{\lambda,\lambda\neq\lambda_C}\exp\left(\delta P_{\log}\left(X|\lambda\right)\right)\right]^{\frac{1}{\delta}} \tag{2}$$

The loss function is traditionally selected to be of a sigmoid type in the form:

$$l(X) = \frac{1}{1+\exp\left[-\alpha\left(d(X)+\beta\right)\right]}. \quad (3)$$

In the previous formulas $P_{\log}\left(X|\lambda\right)$ represents the log-likelihood in the Viterbi sense for the model λ on utterance *X*, *N* is the total number of models, $\alpha > 0, \beta, \delta > 0$ are optimization control parameters and λ_c represents the correct model for *X*.

The objective function is defined as

$$Obj = E\left[l(X)\right], \quad (4)$$

and the target is to minimize the expected loss over all the utterances from the adaptation data.

It can be seen from the the limit $\alpha \to \infty, \delta \to \infty$ of the loss function that it approaches a decision step function and the objective will be the minimization of the classification error rate.

2.2.2 Maximum Mutual Information Approach

The objective of the MMI approach is to maximize the following expression for each of the training utterances:

$$mi(X) = \log\left(\frac{P\left(X|\lambda_C\right)}{P(X)}\right). \quad (5)$$

The main difference between the discriminative approaches presented above is that MCE is focused on the classification boundaries while MMI assigns greater weight in training to the most incorrect classifications [9].

3. SPEAKER ADAPTATION

Speaker adaptation can be applied in such cases in which the user is known by the system. A typical example would be the voice control of a mobile phone or a PC that are considered highly personal devices. In the case of multi-user systems, speaker adaptation can be applied if the user can be identified.

The speaker adaptation process can be performed in several different ways depending on the use and identity of the adaptation data. If the identity of the adaptation data is known by the system, i.e., the system knows what words are spoken, the adaptation is called *supervised*. Otherwise the adaptation process is called *unsupervised*. The adaptation data can also be utilized in two different ways. In *static* or batch adaptation, all data is collected before the models are converted to be speaker-dependent. If the models are continuously updated whenever new data becomes available, the adaptation process is called *incremental* or dynamic.

Speaker adaptation can be done either for the front-end or for the back-end of the recognizer (or even for both). Front-end speaker adaptation schemes usually attempt to perform feature space normalization by estimating the vocal tract length and computing the spectral shift [8]. Due to difficulties associated with finding proper mappings, front-end adaptation has been found ineffective. Much better results have been obtained in the back-end domain [2] where the HMM parameters are tuned to better characterize the new speaker. Due to the success of adapting model parameters, we have chosen the back-end adaptation approaches for this paper.

3.1 Maximum Likelihood Adaptation Approach

As in the speaker-independent case, the target of ML based speaker adaptation is to modify the model parameters so that the likelihood of the adaptation utterances is maximized. Two widely known ML based adaptation methods were chosen to be studied in this paper, namely, the Bayesian adaptation approach [2] (Maximum a Posteriori, MAP), and the Maximum Likelihood Linear Regression (MLLR) technique [5].

Since reliable variance estimation from a limited amount of data is difficult, only Gaussian mean vectors are updated in the experiments presented here. Moreover, the Viterbi algorithm was used throughout this paper to provide the frame-state alignments.

3.1.1 Bayesian Mean Adaptation

In Bayesian adaptation, the new estimate for the *k*'th mean vector $\mathbf{m}_{jk}$ in state *j* can be expressed in the form:

$$\hat{\mathbf{m}}_{jk} = \frac{\tau \cdot \mathbf{m}_{jk} + \sum_{t=1}^{T} d_{jkt}\mathbf{o}_t}{\tau + \sum_{t=1}^{T} d_{jkt}} \quad (6)$$

where τ can be regarded as the step-size controlling the learning rate, and d_{jkt} denotes the probability of being in state *j* and observing the *k*'th mixture at time *t*, respectively.

3.1.2 MLLR Mean Adaptation

In MLLR adaptation, an affine transformation is applied to all Gaussian mean vectors as follows:

$$\hat{\mathbf{m}}_{jk} = \mathbf{A} \cdot \mathbf{m}_{jk} + \mathbf{b}. \quad (7)$$

The actual adaptation task in MLLR is to estimate the transformation parameters which maximize the likelihood of the adaptation data. To guarantee robust parameter estimation, a high degree of transformation parameter tying is usually preferred. Due to the limited adaptation data, we use in this paper a global transformation matrix and an offset vector that are shared by all the mixture densities.

3.2 Discriminative Adaptation Approach

Speaker adaptation can also be done so that the target is better linked to the maximization of the recognition accuracy, like in [6]. Again, as in the speaker-independent case, the ML based speaker adaptation approaches maximize the recognition accuracy only indirectly. Thus, the so called discriminative approaches presented earlier for speaker-independent case can also be applied for adaptation. However, there are some significant aspects to be considered. Especially the optimization control parameters must be readjusted for the adaptation purpose. One important change required by the reduced amount of adaptation data is that the correct models should have a higher learning rate than the competing (incorrect) ones.

4. RECOGNITION EXPERIMENTS

In the experiments, our target was to find out the combination of a speaker-independent training scheme and a speaker adaptation scheme that maximizes the overall recognition accuracy. First we compared two speaker-independent HMM sets estimated according to the ML and the MCE criteria. Then we applied four different discriminative and ML based speaker adaptation schemes for both of these HMM sets and studied the achieved recognition accuracies.

4.1 Databases

We used an English language connected digit database for training the initial whole-word speaker-independent HMMs. The training utterances were spoken in a car environment under the following noise conditions: parking place (motor off), city (moving car), and highway (moving car, 120 km/h). The database consisted of 57 male and 57 female speakers, about 45,000 spoken digits altogether.

Another database consisting of 5 speakers was used for the adaptation tests. There were about 1,800 test digits for each speaker distributed in 400 strings of 3 or 6 digits each. This data was recorded in a clean environment. To test the performance in the presence of noise, we added car noise to the original clean waveforms at 0 dB and -10 dB SNRs. For adaptation purposes we had a separate set of clean digit sequences.

In all the experiments, we used feature vectors consisting of 12 FFT-based MFCCs, log-energy and their first and second order time derivatives. The sampling rate was 8 kHz for both databases.

4.2 Speaker-Independent Experiment

State duration constrained [4], speaker-independent, multi-environment HMMs were estimated from the initial training data according to the ML and MCE principles. Two sets of models were estimated, with one and three mixtures per state.

Table 1 shows the error rates for the initial ML and MCE HMMs when using single mixture and three mixture Gaussian densities. The results indicate the importance of having several mixtures characterizing each state. In particular, in the presence of noise, additional mixtures are needed. Moreover, the MCE sets of HMMs were always superior to the corresponding ML HMMs. In the single mixture case an overall 29% error rate reduction (e.r.r.) was achieved and in the three mixture case an overall 30% error rate reduction was achieved due to the MCE approach.

	Clean		**SNR = 0 dB**		**SNR = -10 dB**	
	string	e.r.r.	string	e.r.r.	string	e.r.r.
ML1	11.27	-	11.37	-	16.78	-
MCE1	6.21	44.90	8.42	25.95	14.23	15.20
ML3	4.71	-	6.66	-	11.62	-
MCE3	3.66	22.29	3.81	42.79	8.62	25,82

Table 1: String error rates with initial speaker-independent ML and MCE HMMs using one and three Gaussian mixtures in each HMM state.

4.3 Speaker Adaptation Experiments

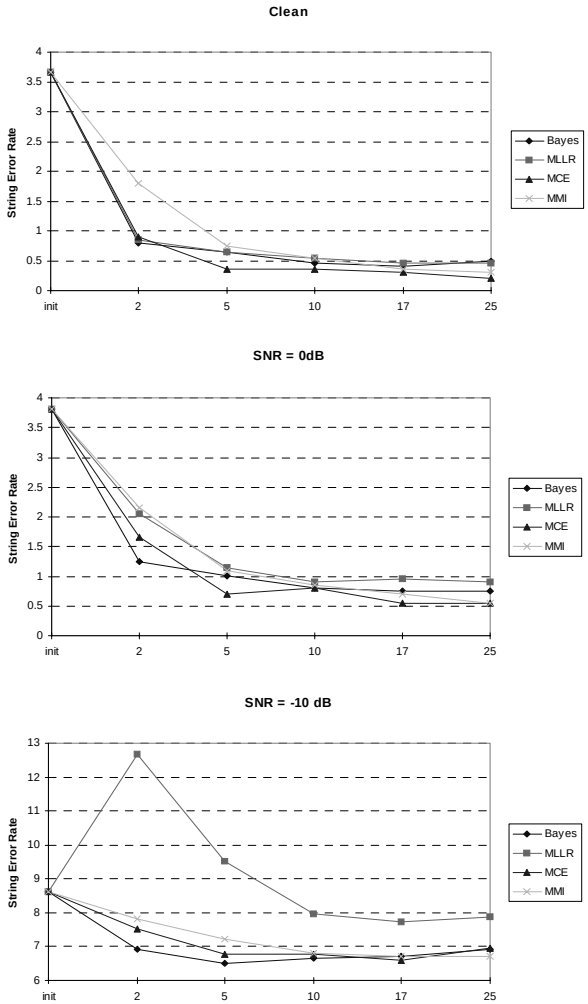
For speaker adaptation experiments, we selected the supervised static approach. All adaptation utterances were from a clean environment. Each adaptation utterance consisted of six digits spoken in a connected manner. The number of adaptation utterances varied from 2 to 25.

4.3.1 Adaptation Schemes Comparison

Figs. 1-3 illustrate the recognition rates obtained with various adaptation schemes in different noise conditions. As an initial speaker-independent model set we selected the three mixture MCE HMM set due to its best speaker-independent performance.

All the adaptation approaches (Bayes, MLLR, MCE, and MMI) were capable of decreasing the error rates. The Bayesian and MCE methods seemed to work best, but no dramatic performance differences between these methods could be observed. However, regarding the computational requirements of MCE, we could conclude that the Bayesian method was the best choice for speaker adaptation. With only 5 adaptation utterances Bayesian adaptation decreased the error rates by 62%. In clean environment the error rate reduction was as high as 82%.

Although having a good performance in clean, MLLR was not a good choice for noisy environments. Among the discriminative methods, MCE was superior to MMI.



Figs. 1-3: Recognition accuracy as a function of adaptation utterances for discriminative and ML adaptation approaches in different noise conditions.

4.3.2 Simple vs. Accurate HMMs

The objective of this experiment was to find out whether comparable recognition rates can be obtained with single mixture HMMs and with three mixture HMMs after speaker adaptation. Single mixture HMMs are particularly preferred in practical ASR systems where the memory consumption is a critical aspect. Based on the results in 4.2 and 4.3.1, the MCE speaker-independent initial model set and the Bayesian adaptation scheme were selected to be used in this experiment.

Fig. 4 shows that multi mixture HMMs clearly outperform single mixture HMMs in all noise conditions. With the initial speaker-independent models the usage of three mixtures gave 45% error rate reduction over the single mixture case. After 5-utterance Bayesian adaptation the advantage reduced, but the usage of three mixtures still gave over 35% error rate reduction over the single mixture case. The results indicate the importance of having more accurate models even in the case of speaker adaptation, though the performance penalty because of single mixture HMMs is not very severe in the absolute scale.

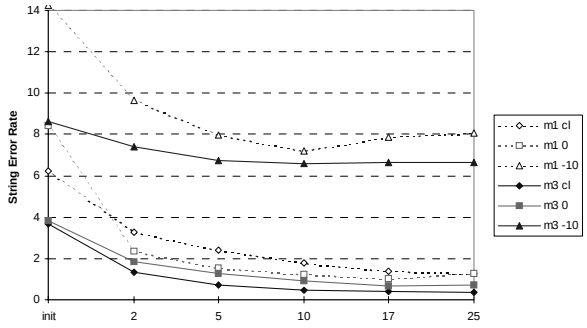


Fig. 4: Recognition performance comparison between single- and multi mixture HMMs.

4.3.3 MCE vs. ML Speaker-Independent HMMs

In this experiment we wanted to find out what kind of effect the selection of initial speaker-independent models have on the results after speaker adaptation. We selected Bayesian adaptation and applied that to the MCE and the ML initial model sets. Because of different objectives there was a mismatch when applying Bayesian adaptation for the MCE initial HMMs. Thus, it was not guaranteed that this combination of different techniques would provide the best results after the adaptation despite of the superiority of the initial MCE model set.

Figure 5 depicts the recognition performance for the ML and MCE trained speaker-independent three mixture HMMs when using the Bayesian adaptation approach. It can be noted that the MCE model set always provided higher recognition rates than the corresponding ML models.

In the case of the initial speaker-independent HMMs, the MCE model set provided 30% error rate reduction over the ML model set. After 5-utterance Bayesian speaker adaptation the advantage reduced, but the MCE model set still gave over 16% error rate reduction over the ML model set. The results indicate that it is possible to combine discriminative and ML techniques in noise robust speech recognition in an advantageous way.

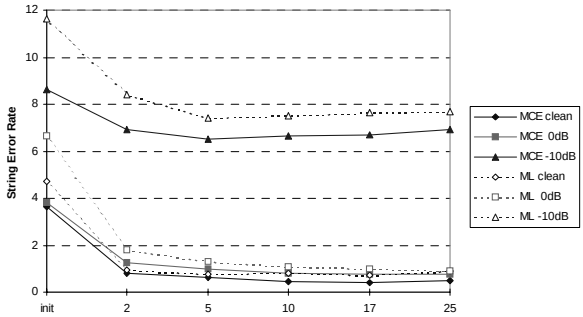


Fig. 5: Effect of the initial speaker-independent model set selection on the speaker adaptation.

5. CONCLUSIONS

In this paper, we showed that discriminative and Maximum Likelihood (ML) techniques can be successfully combined in noise robust speech recognition. In the speaker-independent case, the discriminative training approach performed significantly better than the ML approach, resulting in 30% error rate reduction. We also showed that a rapid and effective speaker adaptation is achievable, resulting in a further 62% error rate reduction. Moreover, we showed that significant error rate reductions in noisy conditions were achieved by performing adaptation only with clean data. For speaker adaptation, the performance of discriminative and ML approaches were shown to be comparable. However, the Bayesian approach was considered as the most suitable from the implementation point of view.

REFERENCES

[1] W. Chou, B.-H. Juang, C.-H. Lee, "Segmental GPD Training of HMM Based Speech Recognizer", *Proc. ICASSP*, pp. 471-476, San Francisco, USA, 1992.

[2] J. L. Gauvain, C.-H. Lee, "Maximum a Posteriori Estimation of Multivariate Gaussian Mixture Observations of Markov Chains", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 2, pp. 291-298, April 1994.

[3] B.-H. Juang, S. Katagiri, "Discriminative Learning for Minimum Error Classification", *IEEE Trans. on Signal Processing*, Vol. 40, No. 12, pp.3043-3054, 1992

[4] K. Laurila, "Noise Robust Speech Recognition with State Durations Constraints", *Proc. ICASSP*, pp. 871-874, Munich, Germany, 1997.

[5] C. J. Leggetter, P. C. Woodland, "Speaker Adaptation of Continuous Density HMMs Using Linear Regression", *Proc. ICSLP*, pp. 451-454, Yokohama, Japan, 1994.

[6] T. Matsui, S. Furui, "A Study of Speaker Adaptation Based on Minimum Classification Error Training", *Proc. EUROSPEECH*, Madrid, Spain, pp.81-84, 1995.

[7] Y. Normandin, R. Cardin, R. De Mori "High-Performance Connected Digit Recognition Using Maximum Mutual Information Estimation", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 2, pp. 299-311, April 1994.

[8] Y. Ono, H. Wakita, Y. Zhao, "Speaker Normalization Using Constrained Spectra Shifts in Auditory Filter Domain", *Proc. EUROSPEECH*, Berlin, Germany, pp. 355-358, 1993.

[9] W. Reichl, G. Ruske, "Discriminative Training for Continuous Speech Recognition", *Proc. EUROSPEECH*, Madrid, Spain, pp. 537-540, 1995