

INCREMENTAL ON-LINE SPEAKER ADAPTATION IN ADVERSE CONDITIONS

Olli Viikki, Kari Laurila

Nokia Research Center, Speech and Audio Systems Laboratory, Tampere, Finland
Email: {olli.viikki, kari.laurila}@research.nokia.fi

ABSTRACT

In this paper, we examine the use of speaker adaptation in adverse noise conditions. In particular, we focus on incremental on-line speaker adaptation since it, in addition to its other advantages, enables joint speaker and environment adaptation. First, we show that on-line adaptation is superior to off-line adaptation when realistic changing noise conditions are considered. Next, we show that a conventional left-to-right HMM structure is not well suited for on-line adaptation in variable noise conditions due to unreliable state-frame alignments of noisy utterances. To overcome this problem, we suggest the use of state duration constrained HMMs. Our experimental results indicate that the performance gain due to adaptation is much greater with duration constrained HMMs than obtained with conventional left-to-right HMMs. In addition to the appropriate model structure, we point out that in long-term adaptation, such as incremental on-line adaptation, the supervised approach is a necessity.

1. INTRODUCTION

Recently, the performance of automatic speech recognition systems has improved drastically so that the implementation of practical speech recognition applications has become possible. Despite the progress made in the speech recognition field, however, fully speaker-independent recognition applications are still inferior to corresponding speaker-dependent systems. Due to this fact, it is desired that current speech recognition systems are either completely speaker-trained, or an original speaker-independent system is transformed to be speaker-dependent using speaker adaptation techniques.

Speaker adaptation can be performed either in an off-line or on-line mode. In off-line adaptation the user is aware of adaptation, typically by performing a special adaptation session, while in on-line adaptation the user may not even know that adaptation is carried out. On-line adaptation is usually embedded to the normal use of a speech recognition system. The adaptation data can be aligned with HMMs in two different ways. In supervised adaptation, the identity of the adaptation data is always known, whereas in the unsupervised case, the identity is not verified at all, and hence adaptation utterances are not necessarily correctly aligned. Speaker adaptation can also be classified to be either incremental (continuous), or to have a finite duration after which no adaptation is done.

During the last years, various speaker adaptation techniques have been proposed in the literature. Two the most popular tech-

niques are currently Bayesian adaptation [1] and Maximum Likelihood Linear Regression (MLLR) [4]. Both techniques attempt to adjust the parameters of Continuous Density HMMs (CDHMM) in such a way that the likelihood scores produced by the models are maximized with the given adaptation data. In [3], we showed that Bayesian adaptation is superior to MLLR if there is an environmental mismatch between the adaptation and testing conditions. Therefore, the Bayesian approach has been adopted for the experiments conducted in this study.

In this paper, we focus on the problems associated with incremental on-line speaker adaptation under realistic, variable noise conditions in a car environment. Up to now, little work has been done to apply speaker adaptation methods in practical noise conditions. Adaptation techniques have usually been tested in noise-free, stationary environments. In practice, however, speech recognition applications are often used under changing noise conditions. Our aim is to find a set of methods that enables long-term continuous speaker adaptation to be safe (i.e. a high recognition accuracy can be maintained over the entire life-span of a recognition system) and transparent to the user.

The remainder of this paper is organized as follows. In Section 2, we describe the Bayesian speaker adaptation approach. A state duration constrained HMM topology is presented in Section 3. These sections provide a sufficient background information for Sections 4 and 5 in which we describe the proposed incremental on-line speaker adaptation technique with experimental results that prove the effectiveness of the selected approach.

2. BAYESIAN MEAN ADAPTATION

In this paper, speaker adaptation is carried out according to the Bayesian mean adaptation approach as presented in Formulas (1) and (2). Diagonal covariance matrices of Gaussian mixtures are here left unchanged since speaker differences are mainly characterized by the Gaussian mean vectors, and thus, additional performance improvements due to variance compensation are small. Let us denote a mean vector of the k th mixture component in state j as $\mathbf{m}_{jk}$. The new estimate for the mean vector can be given in the form

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

where τ denotes the learning rate during adaptation, $\mathbf{o}_t$ is a feature vector at time t , and d_{jkt} is the mixture occupation probability for the k th Gaussian mixture in state j at time t , respectively. A small value of τ corresponds to fast adaptation. The mixture occupation probability can be computed as

$$d_{jkt} = \frac{c_k N(\mathbf{o}_t; \mathbf{m}_k, \Sigma_k)}{\sum_{i=1}^K c_i N(\mathbf{o}_t; \mathbf{m}_i, \Sigma_i)}, \quad (2)$$

where K corresponds to the total number of Gaussian mixtures in state j and $N(\cdot)$ characterizes a Gaussian distribution. Here, we have used the Viterbi algorithm to find an appropriate state-frame alignment. Parameter statistics accumulation is thus done according to the observed Viterbi state sequence.

3. STATE DURATION CONSTRAINED HMM TOPOLOGY

Duration constrained HMMs have been found useful in speech recognition, particularly in speaker-dependent isolated word recognition, e.g., name dialling applications [2]. Strict time domain constraints of duration HMMs substantially reduce the number of possible state sequences provided by a model. Because of these constraints, Viterbi decoded state sequences are less prone to environmental changes.

The topology of our state duration constrained HMMs is illustrated in Fig. 1. Each HMM state has two new parameters, namely the state minimum and maximum duration limits. Duration modelling is implemented with the help of sub-states. The sub-states share the Gaussian parameter values of their "mother" state. The maximum state duration corresponds to the overall number of states within a certain state (e.g. the state 1 in Fig. 1 has the maximum duration of 5 frames). Both minimum and maximum state duration bounds can be estimated according to the Maximum Likelihood (ML) principle as shown in [2].

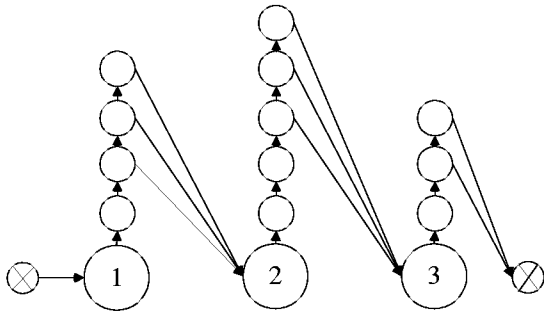


Figure 1: State duration constrained HMM structure.

4. TEST DATABASES AND SETTINGS

A Finnish language connected digit database was used for training the initial whole-word speaker-independent HMMs. Training utterances were spoken in a car environment under various noise conditions. For multi-environment type of

speaker-independent training, all utterances were pooled together and a set of CDHMMs (2 Gaussian mixtures in each state) were estimated according to the ML criterion. The number of male and female speakers, the number of digits, and the transitions between the digits were balanced in the training database.

For speaker adaptation, we had a separate connected digit database which was also recorded in a car environment. This database consisted of 10 speakers (5 males and 5 females) and each speaker spoke at least 1,000 utterances. Recordings were carried out during four recording days over a time-span of one month. Each recording session took approximately one hour during which a test speaker spoke connected digit triples in continuously changing noise conditions depending on the speed of a car, road, and traffic conditions. In the recognition experiments, the order of the test utterances was further randomized so that consecutive utterances were *not* necessarily spoken in similar noise conditions. This arrangement enabled us to simulate a practical usage pattern. In addition to these test utterances, each speaker also uttered 30 connected digit triples in a noise-free car environment. These utterances were used in the off-line adaptation experiments.

Throughout the experiments in this paper, feature vectors consisted of 12 FFT based MFCCs, log-energy, and their first- and second-order time derivatives. Feature vectors were further normalized to have similar parameter statistics in all noise conditions as described in [5].

5. ADAPTATION FRAMEWORK EVALUATION IN PRACTICAL CONDITIONS

In practice, speaker adaptation can be implemented in two different ways:

- By embedding an incremental on-line adaptation step to the normal recognition mode of a system
- By performing a separate off-line adaptation session

From the usability point of view, incremental on-line adaptation provides several advantages over the off-line approach making it very attractive for practical applications. First, by means of on-line adaptation, one can hide the adaptation process from the user. Secondly, the use of on-line adaptation allows us to perform adaptation simultaneously both to a speaker and environment. In addition to follow the short- and long-term variation in the user's voice, this approach is also effective to improve robustness against changing noise conditions, channels, and microphones. Off-line type of adaptation is usually done as an additional training session in a single environment, and thus, it is not possible to incorporate any information on the environment to HMMs.

All the tests in this paper were conducted in unknown length connected digit recognition in a car environment which is one of the key speech recognition tasks in voice dialling applications.

The following experiments were carried out in order to justify our conclusions:

- Incremental (continuous) on-line vs. off-line adaptation session
- Conventional vs. duration constrained HMMs
- Supervised vs. unsupervised adaptation

5.1. On-line vs. Off-line Adaptation

First, the performance of incremental on-line adaptation was compared to that of off-line adaptation. The goal of this experiment was to demonstrate the effect of environment adaptation. In both experiments, adaptation started from the same state duration constrained speaker-independent HMMs. In the case of supervised off-line adaptation, the parameter update was based on 30 adaptation utterances spoken in a noise-free environment. After all adaptation utterances were processed, an updated set of HMMs was computed, and the new model set was subsequently used to recognize the test data. In the case of incremental on-line adaptation, only correctly recognized test utterances were utilized in adaptation (supervised approach). A new set of digit HMMs was always computed after 10 correctly recognized utterances had been processed. The recognition results obtained are given in Fig. 2. Recognition rates have been averaged over all test speakers.

Fig. 2 clearly shows the superiority of incremental on-line adaptation over the off-line adaptation approach. Even though one can improve the performance by means of off-line adaptation, the obtained performance improvements are still marginal compared to those of incremental on-line adaptation. The use of off-line speaker adaptation provides 32.1% error rate reduction over the initial speaker-independent HMMs, whereas in the case joint speaker and environment adaptation, the error rate reduction is as high as 80.2%.

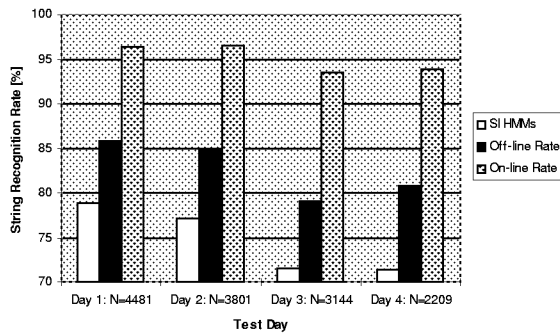


Figure 2: Digit string recognition rates with speaker-independent HMMs, using supervised off-line adaptation, and supervised incremental on-line adaptation (N = total number of utterances spoken during each recording day).

5.2. Effect of HMM Topology on Adaptation Performance

Realistic, continuously varying noise conditions are a challenge for speaker adaptation algorithms. Obtaining appropriate state-frame alignments needed in adaptation is feasible in noise-free environments, but the task gets more and more difficult as the noise level in the speech signal increases. In the case of conventional left-to-right HMMs, loose time domain constraints at the state level are one major reason producing occasionally inappropriate Viterbi state sequences, particularly in adverse noise conditions. Since consecutive utterances in practical speech recognition applications are usually spoken under a wide range of noise conditions, poor state-frame alignments are obtained every now and then. Eventually, these inconsistent alignments result in a non-optimal set of adapted HMMs, and make the incremental adaptation process ineffective.

An efficient method to restrict the number of possible Viterbi state sequences so that the state-frame alignments of the same utterance are more similar in different noise conditions is to use duration constrained HMMs. Our duration constrained HMM topology was briefly outlined in Section 3. The objective of the following experiment was to study the effect of duration constrained HMMs on the adaptation performance. Fig. 3 illustrates the adaptation performance with conventional left-to-right and duration constrained HMMs in incremental supervised on-line adaptation. A cumulative recognition accuracy of all test utterances is averaged over all speakers.

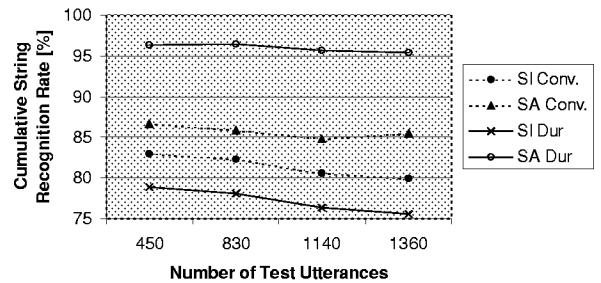


Figure 3: Duration constrained (solid line) vs. conventional (dotted line) HMMs using original speaker-independent and incrementally adapted HMMs.

With both type of HMMs, the test settings were the same as in Section 5.1. and the initial speaker-independent HMMs were estimated as described in Section 4. Fig. 3 shows that conventional HMM structure performs better when using the original speaker-independent HMMs. However, by means of incremental on-line adaptation, one can achieve over 95.8% average digit string recognition accuracy with duration constrained HMMs. For conventional left-to-right HMMs, the corresponding recognition rate is only 85.7%.

5.3. Supervised or Unsupervised Adaptation?

Our final remark focuses on the selection of the appropriate state-frame alignment approach for incremental on-line adaptation. Based on our experiments, we can say that if a short enrollment time for a speaker is required, unsupervised adaptation may be preferred. Unsupervised adaptation means that we rely on the HMM recognizer to classify the spoken words correctly. The identity of the recognition result is not verified but the optimal Viterbi decoded state-frame alignment is directly used in adaptation. This property makes unsupervised adaptation a little risky. In the case of recognition errors (misclassifications) wrong models are adapted. If adaptation uses only some incorrectly recognized utterances, no serious damage is done to the models. However, as the number of incorrectly classified utterances increases, HMMs get "corrupted", and eventually the recognition rate collapses. Incorrectly updated models usually start to produce too high log-likelihood values with respect to other HMMs, and they are thus dominating the recognition process. Tolerance to incorrect adaptation utterances depends very much on the used vocabulary and speaker characteristics. Therefore, it is very difficult to predict in advance how many incorrect utterances the system can tolerate.

In on-line adaptation, the supervised adaptation approach is implemented by verifying the correctness of the recognition result before the utterance is used for adaptation. Supervised adaptation can be slower than its unsupervised counterpart, particularly with such speakers whose recognition accuracy is poor with the original speaker-independent models. It may sometimes take a long time before a sufficient number of correctly recognized digit strings are provided to the adaptation process. However, the recognition performance will remain more stable as only correctly classified utterances are utilized in adaptation.

Fig. 4 compares the recognition rates between supervised and unsupervised incremental on-line adaptation. Duration constrained HMMs were used in these experiments. Results are again averaged over all test speakers.

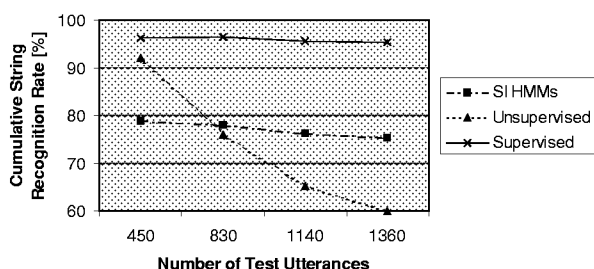


Figure 4: Supervised and unsupervised approaches in incremental on-line speaker adaptation.

Clearly, the use of incorrectly classified utterances for adaptation leads to a performance drop after many enough utterances have been processed. At the beginning, the recognition accuracy

improves, but eventually the recognition performance starts to decrease. Depending on the speaker, this drop of recognition accuracy may occur as late as after processing more than 1,000 utterances. Based on the results in Fig. 4, we can conclude that the supervised adaptation approach is the only viable option for incremental on-line adaptation.

6. CONCLUSIONS

In this paper, we discuss on the use of incremental on-line adaptation in realistic noise conditions. We underline two important remarks which must be considered in order to guarantee a high recognition accuracy over the whole life-span of a recognition system. Our first observation deals with the proper selection of the HMM structure. The experimental results indicate that the conventional left-to-right HMM topology is not very well suited for long-term incremental on-line speaker adaptation in realistic changing noise conditions. Due to the lack of duration constraints, conventional left-to-right HMMs sometimes tend to produce unrealistic state-frame alignments in the presence of noise which further reduce the adaptation performance. This difficulty can be alleviated by using duration constrained HMMs which significantly limit the number of possible alignments, and thus, a high adaptation performance is always obtained. Our experimental results also show that the unsupervised technique cannot be applied to long-term on-line speaker adaptation. After many enough incorrectly classified utterances are used in adaptation, the recognition accuracy of a speech recognition system collapses.

REFERENCES

1. Gauvain, J. L., Lee, C.-H. "Maximum a Posteriori Estimation of Multivariate Gaussian Mixture Observations of Markov Chains", *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, pp. 291-298, April 1994.
2. Laurila, K. "Noise Robust Speech Recognition with State Duration Constraints", *Proc. of ICASSP'97*, pp. 871-874, Munich, 1997.
3. Laurila, K., Vasilache, M., Viikki, O. "A Combination of Discriminative and Maximum Likelihood Techniques for Noise Robust Speech Recognition", *Proc. of ICASSP'98*, pp. 85-88, Seattle, USA.
4. Leggetter, C. J., Woodland, P. C. "Speaker Adaptation of Continuous Density HMMs Using Linear Regression", *Proc. of ICSLP'94*, pp. 451-454, Yokohama, Japan, 1994.
5. Viikki, O., Bye, D., Laurila, K. "A Recursive Feature Vector Normalization Approach for Robust Speech Recognition in Noise", *Proc. of ICASSP'98*, pp. 733-736, Seattle, USA.