IDIAP RESEARCH REPORT

IMPROVING SPEECH RECOGNITION USING EPISODIC MODELS

Jithendra Vepa [a] Guillermo Aradilla [a.b]
Hervé Bourlard [a,b]

IDIAP–RR 04-47

September 2004

submitted for publication

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP Research Institute, P. O. Box 592, Rue du Simplon 4, 1920 Martigny, Switzerland.
[b] Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.

IDIAP Research Report 04-47

IMPROVING SPEECH RECOGNITION USING EPISODIC MODELS

Jithendra Vepa Guillermo Aradilla Hervé Bourlard

September 2004

submitted for publication

Abstract. In this paper, we propose a new technique using episodic models to improve the performance of the current state-of-the-art speech recognition systems. These models allow us to make use of meta data and environmental information such as speaker, gender, accent, noise conditions. We recognise that we can not entirely abandon HMMs which are very powerful and highly scalable models. Hence, we propose one way to combine both models to gain from the advantages of them. In this work, we first obtained N-best hypotheses from HMM models, and then re-scored these hypotheses using episodic models and HMM likelihoods. We carried out recognition experiments on Numbers95 database, and observed that our approach yields more than 1% absolute improvement (22% relative improvement in the word error rate) over the baseline performance. We also applied the K-means clustering technique to the acoustic vectors to speed up the decoding, while still yielding a significant improvement in the recognition accuracy.

1 Introduction

Current state-of-the-art speech recognition systems using hidden Markov models (HMM) yield high recognition accuracies due to their high degree of scalability and great efficiency. However, there are a few weaknesses in this approach due to the assumptions made for optimization, such as, assuming state sequences as the first order Markov chains and frame-based stationarity. Many improvements have been proposed to deal with these assumptions. Recently, an interesting approach was proposed in [1], the motto was: *no models, use all the training data for recognition*. In this approach, first the template database is created using all the training data. These templates can also contain some meta data and environmental information such as speaker, gender, dialect, speech rate, signal-to-noise ratio (SNR). For the recognition, the input speech is compared with the templates in the database and select the closest template sequence using the dynamic time warping (DTW) algorithm. To speed up the search, many pruning or selection techniques could be employed.

The main advantages of the above approach are: no loss of information in the training data and making use of speaker specific and environmental information. But, we do not want to loose the advantages offered by stochastic models, HMMs. Hence, combining both the models may improve the speech recognition accuracy. Very recently, one such approach was presented in [2], where both of these models were combined in an exponential modelling frame work. This combination resulted in better recognition accuracies.

In this paper we present another approach to combine HMMs and episodic models , which can easily fit into large vocabulary speech recognition systems. Our approach is based on rescoring N-best lists using episodic models and HMM likelihoods. In the next section, we describe episodic models, which include DTW algorithms and also briefly explain the clustering technique we used in this study. Then, we discuss our recognition database and experimental setup in Section 3. In Section 4, we present and discuss the results of our recognition experiments. Finally, we conclude the paper with a few directions to the future work.

2 Episodic models

The main motivation of using the data-driven or template-based approach is to use all the available training data instead of training models (e.g. HMMs) [1]. This approach attributes one or more templates to each vocabulary word and these templates are optimally aligned with the incoming speech using a DTW technique for recognition. This system can handle the transient nature of the speech as templates have smooth boundaries and within templates by DTW alignment.

We use the word “episode” for template as we can store many other details along with the acoustic features [3]. This is also in lines with how humans store the past events in episodic memory! We call each episode with the DTW algorithm an *episodic model*.

We used DTW techniques to compute the distance between speech input and reference episodes of different length. Euclidean distance metric was used to compute the local distance between two frames. The acoustic features were 12 MFCCs and the energy extracted from the speech signal using an analysis window of 25ms and a window shift of 10ms.

In DTW techniques, usually a set of local continuity constraints are imposed on the wrapping function in order to ensure proper time alignment while keeping any potential loss of information to a minimum [4]. There are many local constraints proposed in the literature, which are mainly based on heuristics. In this study, we considered three DTW local continuity constraints: symmetric, type-2 [4, 5] and Itakura constraint [6]. These are shown in Figure 1, the numbers on top of each path are corresponding weighting coefficients, for the symmetric and the Itakura cases these are equal for all segments except for type-2. The symmetric[1] constraint can be expressed as:

$$D_{i,j} = d_{i,j} + min(D_{i-1,j}, D_{i-1,j-1}, D_{i,j-1}) \tag{1}$$

where, $D_{i,j}$ is the partial path DTW distance at test frame i and episode frame j and $d_{i,j}$ is the local distance between test frame i and episode frame j.

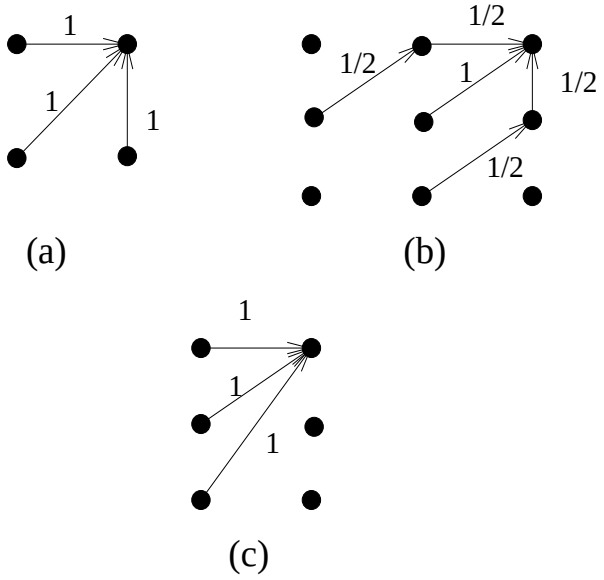

Figure 1: DTW local continuity constraints with weighting coefficients: (a) Symmetric (b) Type-2 (c) Itakura

In the type-2 constraint, we have two paths with double segments and the diagonal path. To avoid strong bias towards diagonal path movements, we set the weights on the paths with double segments. These weights are obtained by using the (smoothed) weighting function suggested in [7]. For the type-2 constraint, we use:

$$D_{i,j} = min \begin{cases} D_{i-2,j-1} + \frac{1}{2} * (d_{i-1,j} + d_{i,j}) \\ D_{i-1,j-1} + d_{i,j} \\ D_{i-1,j-2} + \frac{1}{2} * (d_{i,j-1} + d_{i,j}) \end{cases} \tag{2}$$

Finally, the Itakura constraint, which is an asymmetric[2] constraint rule, can be shown as below:

$$D_{i,j} = d_{i,j} + min(D_{i-1,j}, D_{i-1,j-1}, D_{i-1,j-2}) \tag{3}$$

[1]Though type-2 is also symmetric, we refer "symmetric" to case (a) in Figure 1
[2]Can not interchange the speech input and reference episodes

2.1 Clustering Episodes

To reduce the computational complexity, we used the K-Means algorithm to cluster the acoustic vectors of all the training data [8, 9]. This clustering algorithm partitions N acoustic vectors into K disjoint subsets S_j by minimising the sum-of-squares criterion:

$$\sum_{j=1}^{K} \sum_{n \in S_j} \mid x_n - \mu_j \mid^2 \tag{4}$$

where x_n is n^{th} acoustic vector and μ_j is the geometric centroid of the data points in S_j. We used K values of 50, 100 and 200 to build the clustered episode databases, where we only store the cluster indexes for the acoustic vectors of episodes. So, to compute the DTW path distances, we only need to compute the distances between input vectors and the centroids, hence saving lot of CPU time [9].

3 Recognition experiments

We used OGI Numbers95 connected digits telephone speech database [10] for our recognition experiments. This database is described by a lexicon of 30 words and 27 phonemes. Our speech recognition system is based on HTK [11]. We trained triphone HMM models on MFCC features of 39 dimension, including 13 static coefficients, 13 delta coefficients and 13 delta-delta coefficients. Our training set consists of 3233 utterances and test set consists of 1206 utterances.

First, we obtained forced-alignments of training data from our HMM models to construct episode database. Currently, we use whole-word episodes which contain only acoustic features (12 MFCCs and the energy). We, then generated N-best ($N = 20$) hypotheses list with word segmentations and likelihoods. We computed DTW distances using one of the three local continuity constraints. Then we re-scored our N-best list using the HMM likelihoods and DTW distances. We set the weights for likelihoods and distances empirically.

4 Results & discussion

Table 1 presents word recognition rates of our baseline system and combined HMM & episodic models. We have achieved 1.3% absolute improvement over baseline, i.e. 22% relative improvement in the word error rate, which is quite high for this task. We performed a statistical significance test presented in [12], which is a non-parametric test based on a bootstrap method. This test indicates that the difference between the combined HMM & episodic system and the baseline HMM system is highly significant.

system	*Word recognition rate in %*
Baseline HMM system	94.1
Combined HMM & Episodic	95.4

Table 1: Recognition accuracies of baseline system and combined HMM & episodic system

We experimented with different weights on HMM likelihoods and DTW distances. We also used three different DTW local path constraints described in Section 2, while computing DTW distance between the test frames and episodes. The word recognition rates are shown in Table 2. These results indicate that we can achieve a significant improvement over baseline by using any of our three local constraints. We performed a one-way analysis of variance (ANOVA) on word recognition rates with three levels: symmetric, type-2 and Itakura. This test indicated that there is no significant difference among means of the three DTW constraints.

weight on DTW dist.	*weight on likelihood*	*Word recognition rate in %*		
		symmetric	*type-2*	*Itakura*
0.25	0.75	95.2	**95.4**	94.8
0.20	0.80	95.2	95.2	95.1
0.15	0.85	**95.4**	95.3	95.2
0.10	0.90	95.1	95.1	95.1

Table 2: Recognition accuracies obtained by rescoring the 20-best list using different weights on DTW distances and HMM likelihoods

As mentioned in Section 2.1, we built the clustered episode databases and carried out recognition experiments, where the DTW distances were computed using the clustered databases. This reduced the computation expenses for decoding quite significantly, for example by around 20 times using 100-clustered episode database.

Table 3 presents the recognition accuracies obtained by rescoring the 20-best list using both HMM likelihoods and DTW distances, which were computed using 100-clustered database. Though the recognition accuracies are slightly lower compared to those obtained using unclustered episode database, these are still significantly better than the baseline. Here also, we compared the three DTW local continuity constraints and ANOVA test confirmed that there is no significant difference among them.

weight on DTW dist.	*weight on likelihood*	*Word recognition rate in %*		
		symmetric	*type-2*	*Itakura*
0.25	0.75	94.8	95.0	94.6
0.20	0.80	95.0	95.1	94.8
0.15	0.85	95.2	95.1	94.9
0.10	0.90	95.2	95.0	94.9

Table 3: Recognition accuracies obtained by rescoring 20-best list using different weights on DTW distances, computed using the 100-clustered episode database, and HMM likelihoods

We also carried out recognition experiments using 50-clustered episode database and 200-clustered episode database. In Figure 2, we show the word recognition rates for the three DTW local continuity constraints for weights of 0.15 on DTW distance and 0.85 on HMM likelihood. We plot the four episode databases with different number of clusters: 50, 100, 200 and using all the acoustic vectors (i.e. no clustering). As expected, the recognition rates are increasing with respect to the number of clusters. But clustering is really helpful in the case of large databases, so from Figure 2 we can say that 100-clustered database is a good speed-accuracy trade-off.

5 Conclusions & future work

We have proposed a new approach to enhance the performance of HMM-based speech recognition system by combining with episodic models. In this approach, we re-scored the N-best lists generated from HMM-based system using the HMM likelihoods and DTW distances obtained from episodic models. The results from recognition experiments carried out on Numbers95 are very promising. We achieved more than 1% absoulte improvement (22% relative improvement) over baseline performance. We compared three different DTW local continuity constraints: symmetric, type-2 and Itakura and observed that symmetric is performing better than other two, although the difference is not statistically significant. We also carried out experiments using the clustered databases, which significantly speed up our decoding while keeping the improvements of around 1% over the baseline.

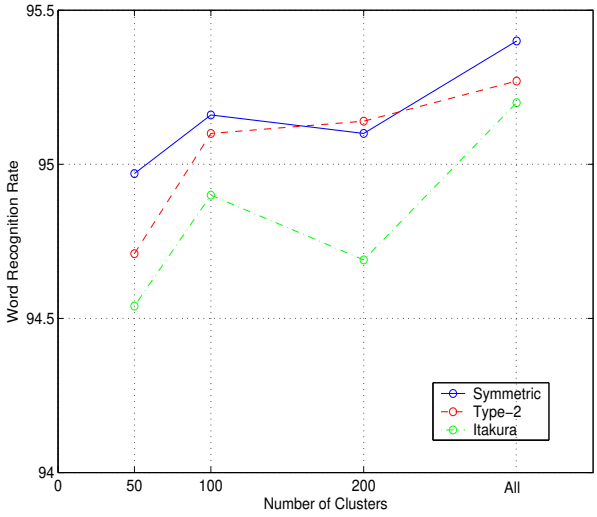

Figure 2: Effect of clustering on word recognition rates, showed for three DTW local continuity constraints using weights of 0.15 and 0.85 for DTW distance & HMM likelihood

Currently we are using intuitive weights on HMM likelihoods and DTW distances but in future we would like to use an entropy motivated combination scheme. Also, we want to implement this idea in large vocabulary recognition system built in DARPA EARS framework with the addition of some meta-data information to our episodic models.

6 Acknowledgements

This work was supported by the EU 6th FWP IST integrated project AMI (FP6-506811). The authors thank DARPA for supporting through the EARS (Effective, Affordable, Reusable Speech-to-Text). The authors also would like to thank Hynek Hermansky and John Dines for useful discussions during this work.

References

[1] Mathias De Watcher, Kris Demuynck, Dirk Van Compernolle, and Patrick Wambacq, "Data driven example based continuous speech recognition," in *Proc. Eurospeech*, Geneva, Switzerland, 2003.

[2] Scott Axelrod and Benoit Maison, "Combination of hidden markov models with dynamic time warping for speech recognition," in *Proc. ICASSP*, Montreal, Canada, 2004.

[3] H. Strik, "Speech is like a box of chocolates...," in *Proc. ICPhS*, Barcelona, Spain, 2003.

[4] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey, USA, 1993.

[5] C. Myers, L.R. Rabiner, and A.E. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 6, pp. 623–635, 1980.

[6] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 67–72, 1975.

[7] H. Sakoe and S. Chiba, "Dynamic programming optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 43–49, 1978.

[8] H. Bourlard, H. Ney, and C.J. Wellekens, "Connected digit recognition using vector quantization," in *Proc. ICASSP*, 1984, pp. 26.10.1–4.

[9] H. Bourlard, Y. Kamp, H. Ney, and C.J. Wellekens, "Speaker-dependent connected speech recognition via dynamic programming and statistical methods," in *Speech and Speaker Recognition*, M.R. Schroeder, Ed., pp. 115–148. Karger (Basel), 1985.

[10] R. Cole, M. Noel, T. Lander, and T. Durham, "New telephone speech corpora at CSLU," in *Proc. of European Conference on Speech Communication Technology*, 1995.

[11] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Version 2.1. Cambridge University, Entropic Cambridge Research Laboratory, UK, 1997.

[12] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. ICASSP*, Montreal, Canada, 2004.