IDIAP RESEARCH REPORT

Improving Speech Recognition Using a Data-Driven Approach

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

IDIAP–RR 05-66

April 2005

[a] IDIAP Research Institute
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL)

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-66

Improving Speech Recognition Using a Data-Driven Approach

Guillermo Aradilla Jithendra Vepa Hervé Bourlard

April 2005

Abstract. In this paper, we investigate the possibility of enhancing state-of-the-art HMM-based speech recognition systems using data-driven techniques, where whole set of training utterances is used as reference models and recognition is then performed through the well-known template matching technique, DTW. This approach allows us to better capture the temporal dynamics of the speech signal while avoiding some of the HMM assumptions such as the piecewise stationarity. Potentially, such data-driven techniques also allow us to better exploit meta-data and environmental information, such as speaker, gender, accent and noise conditions. However, we cannot entirely abandon HMMs, which are very powerful and scalable models. Thus, we investigate one way to combine and take advantage of both the approaches, combining scores of HMMs and reference templates. Experiments on the Numbers95 database showed that this combination yields 22% relative improvement in word error rate over the baseline HMM performance. Applying K-means clustering to the acoustic vectors speeds up the decoding, while still retaining a significant improvement in the recognition accuracy.

1 Introduction

Current state-of-the-art speech recognition systems using Hidden Markov Models (HMMs) yield good recognition accuracies due to their high degree of scalability and great efficiency, and their ability to extract and model statistical properties at the spectral and temporal variability inherent to speech [1]. However, there are a few weaknesses in this approach due to the assumptions made for optimization, such as (1) piecewise stationarity and (2) state conditional independence. Many improvements have been proposed to deal with these assumptions. Recently, an interesting approach was proposed in [2] where all the training data is directly used for recognition. The idea is that since all models make some assumptions about the data which are not always correct, as is the case with HMMs, all the training data is directly used for recognition instead. This approach deals with templates representing linguistic units (words in our case). The template database is created using all the training data and it can also contain some meta data and environmental information such as speaker, gender, dialect, speech rate and signal-to-noise ratio (SNR). For the recognition, the input speech is compared with the templates in the database and the closest template sequence using a distance measure is selected. To speed up the search, many pruning or selection techniques can be employed.

Using template-based approaches result in the following potential advantages: (1) no loss of information in the training data since all the training data is used for decoding, (2) no strong assumptions are required about the data since there is no model, and (3) possible use of speaker specific and environmental information. Although already used as the main recognition approach in the 70's, data-driven approaches were recently revisited given today's availability of very large amounts of training data, as well as increased computational (memory and CPU) resources. While doing this though, we also want to preserve the advantages offered by HMMs, and it thus seems that a principled approach towards exploiting the advantages of both approaches consists in combining both models. Very recently, one such approach was presented in [3], where both of these models were combined in an exponential modelling framework for isolated word digit recognition task. This combination resulted in improved recognition performance.

In this paper, we investigate another approach to combine HMMs and templates, which could easily accommodate large vocabulary speech recognition tasks. The main idea is to re-score N-best lists generated by HMMs, based on a weighted sum of Dynamic Time Warping (DTW) distortion measures and HMM negative log likelihoods.

In the next section, we describe the data-driven approach, involving DTW template matching, and also briefly explain the clustering technique we used in this study to reduce CPU time. Then, we discuss our recognition database and experimental setup in Section 3.1. In Section 3.2, we present and discuss the results of our recognition experiments. Finally, we conclude the paper with a few directions to the future work.

2 Data-Driven Approach

2.1 Template Matching

The main motivation of using a data-driven or template-based approach is to avoid assumptions by using all the available training data instead of training models (e.g. HMMs) [2]. Statistical modeling aims at extracting models by summarizing the training data while also exhibiting good generalization properties. This summarization usually involves making some assumptions about the data, which are not always correct. This is the case with HMMs, which incorrectly assume that a speech utterance can be (1) modeled as a piecewise stationarity process and that (2) acoustic vectors are independent given the output state. By using all the training data for decoding, no specific models are trained and, hence, no explicit assumptions need to be taken. This approach follows the same idea as non-parametric classifiers such as by K-Nearest Neighbor algorithm, which performs classification without making any prior assumption about the data [4]. Indeed, when facing large amounts of training data, template matching can be considered as a particular case of HMMs, allowing us to better capture the temporal dynamics of the speech signal while avoiding some of the HMMs assumptions.

Obviously, the more training data, the more accurate will be the recognition since the more speech variability is represented. Also, unlike parametric methods, e.g. HMMs, data-driven methods will consume a lot of computational resources at the decoding time since a large amount of data to be dealt with. These drawbacks were encountered by the DTW-based approaches which were carried out in the 70's in the Automatic Speech Recognition (ASR) field [5], but nowadays they can be conveniently handled today because of the computational power and the huge amount of data available.

This approach assigns one or more templates to each vocabulary word and these templates are aligned with the incoming speech. The most used alignment is the one which is optimal in the sense of minimum distortion between the trajectories described by the feature sequences, and it can be efficiently implemented by the well-known DTW technique. This measurement does not have the mathematical properties required for being a distance strictly speaking [6], and this is the reason why we prefer to call it as DTW *distortion*.

Templates can also be associated with additional information [7], such as gender speaker, environmental information, etc. It is sometimes argued that this is also in lines with how humans store the past events in episodic memory [8]. This meta-information can be used to classify templates which have the similar characteristics and these class-specific templates can then be used for recognition. This will enhance the performance recognition, since it has been shown that lower word error rates can be achieved for class-specific recognition [9] and furthermore, will reduce the decoding time as the number of templates to compare will be lower.

In DTW, a set of local continuity constraints are usually imposed on the wrapping function to ensure proper time alignment while keeping any potential loss of information to a minimum [1]. Many local constraints, mainly based on heuristics, have been proposed in the literature. In the present work, we considered two DTW local continuity constraints[1]: symmetric and Itakura constraint [10]. These constraints are illustrated in Figure 1. They can be expressed as

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

for the symmetric constraint and

$$D_{i,j} = d_{i,j} + \min\{D_{i-1,j}, D_{i-1,j-1}, D_{i-1,j-2}\}$$

for the Itakura constraint. $D_{i,j}$ is the partial path DTW distortion at test frame i and template frame j and $d_{i,j}$ is the local distance between test frame i and template frame j. The DTW distortion then will be $D_{I,J}$ where I and J are the number of frames of the test and template sequences.

The local distance used in this work was simply the Euclidean distance

$$d_{i,j} = \|x_i - y_j\|^2$$

where x_i and y_j denotes the ith and jth frames of the input and template sequences respectively. Other distances, such as Mahalanobis distance, can be investigated[2]

2.2 Clustering Acoustic Vectors

To reduce the computational complexity, we used the K-Means algorithm to cluster the acoustic vectors of all the training data [11, 12]. This clustering algorithm partitions N acoustic vectors into K disjoint subsets S_j by minimizing the sum-of-squares criterion:

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

[1]Other constrains have also been considered yielding similar results.

[2]In the case of Mahalanobis distance, results shoult not be significantly different since features have been previously liftered in order to approximate equal variance in all dimensions.

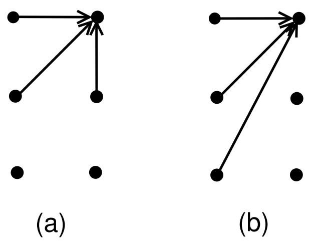

Figure 1: DTW local continuity constraints: (a) Symmetric (b) Itakura

where x_n is n^{th} acoustic vector and μ_j is the geometric centroid of the data points in S_j. Reference templates are then replaced by label sequences, where each acoustic vector has been replaced by its closest centroid.

In this case, computation of the local distance $d(x_i, y_j)$ between an input frame x_i and a reference frame y_j is replaced by $d(x_i, c_\ell)$, where

$$c_\ell = \arg\max_{c_k} \|x_i - c_k\|^2$$

with $c_k \in C = \{c_1, \ldots, c_k, \ldots, c_K\}$, the set of centroid vectors. Consequently, for every input vector x_n, computation of DTW local distances can be replaced by the computation of distances between input vectors and the centroids, hence saving a lot of CPU time since many acoustic vectors from the template database will correspond to the same centroid [12].

Templates with clustered acoustic vectors can be considered as a kind of degenerated HMM where temporal information has not been parameterized. In this sense, this kind of templates does not strictly follow the data-driven philosophy because all the information from the data is intended to be in the centroids, hence some summarization (modeling) has been carried out.

2.3 HMM-DTW Combination

A combination approach has been investigated to exploit the advantages of both HMM modeling and data-driven techniques. More specifically, we first use HMM to generate an N-best hypothesis list. However, instead of using more and more detailed HMM models for re-scoring this N-best list [13], we then perform DTW on the templates associated to the words corresponding to each of the N-best hypothesis. Recognition is then performed on the basis of a weighted sum of the HMM (negative log likelihoods) and DTW distortion. Hence, with each hypothesis is associated a score S_n defined as:

$$S_n = w \cdot D_n + (1 - w) \cdot L_n$$

where w is the combination weight, L_n denotes minus log likelihood of the n-th hypothesis, and D_n corresponds to its total DTW distortion:

$$D_n = \sum_{m=1}^{M(n)} d_m(n)$$

$M(n)$ represents the number of words of the n-th hypothesis and $d_m(n)$ corresponds to the DTW distortion associated to the m-th word of the n-th hypothesis.

In this present work, the weight w for computing the total score was set empirically. The hypothesis associated with the resulting lowest score is then chosen as the correct one.

This combinational approach also offers the advantage that language modeling constrains, if they exist, are preserved since all the possible hypothesis are generated by the HMM system, which can incorporate a grammar.

3 Recognition Experiments

3.1 Database

OGI Numbers95 connected digits telephone speech database [14] was used for our recognition experiments. This database is described by a lexicon of 30 words and 27 phonemes. Our speech recognition system was based on HTK [15]. 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 template database. Currently, we use whole-word templates which contain only acoustic features (12 MFCCs and the energy). We generated N-best ($N = 20$) hypotheses list with word segmentations and likelihoods. We computed DTW distortion using one of the two local continuity constraints described previously. Then we re-scored our N-best list as explained in Section 2.3

3.2 Results & Discussion

Table 1 presents word recognition rates of our baseline system and combined HMM and data-driven approach, showing 1.3% absolute improvement over baseline, i.e. 22% relative improvement in the word error rate. We performed a statistical significance test presented in [16], which is a non-parametric test based on a bootstrap method. This test indicates that the difference between the combined HMM & data-driven approach and the baseline HMM system is highly significant.

System	*Word Recognition Rate*
Baseline HMM System	94.1%
Combined HMM & DTW	95.4%

Table 1: Recognition accuracies of baseline system and combined HMM & data-driven approach

We experimented with different weights on HMM log likelihoods and DTW distortions. We also used two different DTW local path constraints described in Section 2, while computing DTW distance between the test frames and templates. 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 two local constraints. We performed a one-way analysis of variance (ANOVA) on word recognition rates with two levels: symmetric and Itakura. This test indicated that there is no significant difference among the two DTW constraints.

Weight	*Symmetric*	*Itakura*
0.25	95.2%	94.8%
0.20	95.2%	95.1%
0.15	**95.4%**	95.2%
0.10	95.1%	95.1%

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

As mentioned in Section 2.2, we built template database with clustered acoustic vectors and carried out recognition experiments, where the DTW distortion were computed using the clustered databases. This reduced the computation expenses for decoding quite significantly, for example by around 20 times using 100-clustered template database.

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

Weight	*Symmetric*	*Itakura*
0.25	94.8%	94.6%
0.20	95.1%	94.8%
0.15	95.2%	94.9%
0.10	95.2%	94.9%

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

We also carried out recognition experiments using 50-clustered template database and 200-clustered template database. In Figure 2, we show the word recognition rates for the two DTW local continuity constraints for weights of 0.15 on DTW distortion and 0.85 on HMM likelihood. We plot the four template 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.

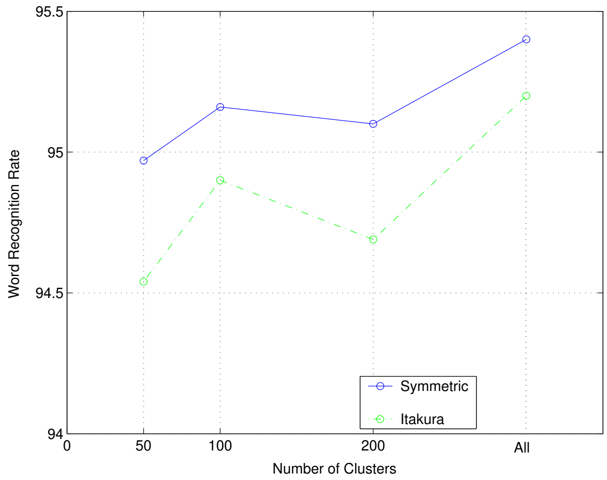

Figure 2: Effect of clustering on word recognition rates, showed for two DTW local continuity constraints using weights of 0.15 and 0.85 for DTW distortion & HMM log likelihood

4 Conclusions & Future Work

In this paper, we have investigated a new approach to enhance the performance of HMM-based speech recognition systems by combining it with a data-driven approach used to re-score HMM generated N-best hypothesis. New matching scores are estimated as a weighted sum recombination of HMM log-likelihoods and DTW scores obtained from templates associated with the words present in the N-best list.

In the present work, we did not make use of meta-data, which is one of the additional advantages offered by data-driven approaches, but results from continuous speech recognition experiments carried out on Numbers95 are already very promising. We achieved more than 1% absolute improvement (22% relative improvement) over baseline performance. We compared two different DTW local continuity constraints: symmetric and Itakura and observed that symmetric is performing better than the other two constraints, 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.

Currently, we are using hand-tuned combination weights, but in the future we intend to use an entropy motivated combination scheme. Also, we want to implement this idea in large vocabulary recognition system with the addition of some meta-data information to our templates in order to prune the search and to save computing time.

Another direction for future work is the type of distance used for measuring similarity between the input sequence and templates. In this work, we have used the well-known DTW algorithm but other more perceptually-based distances should be investigated. Word templates are used in this experiment but other types of linguistic units can be considered for larger vocabulary tasks.

5 Acknowledgements

This work was supported by the EU 6th FWP IST integrated project AMI (FP6-506811). The authors want to thank the Swiss National Science Foundation for supporting this work through the National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)". The authors also would like to thank Hynek Hermansky and John Dines for useful discussions during this work.

References

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

[2] M. D. Watcher, K. Demuynck, D. V. Compernolle, and P. Wambacq, "Data driven example based continuous speech recognition," in *Proc. Eurospeech*, (Geneva, Switzerland), 2003.

[3] S. Axelrod and B. Maison, "Combination of hidden markov models with dynamic time warping for speech recognition," in *Proc. ICASSP*, (Montreal, Canada), 2004.

[4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley-Interscience, 2001.

[5] J. S. Bridle and M. D. Brown, "Connected Word Recognition Using Whole Word Templates," *Proc. Inst. Acoust. U.K.*, pp. 25–29, 1979.

[6] E. Vidal, F. Casacuberta, J. M. Benedi, M. J. Lloret, and H. Rulot, "On the Verification of Triangle Inequality by Dynamic Time Warping Dissimilarity Measures," *Speech Communication*, pp. 67–79, 1988.

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

[8] S. D. Goldinger, "Echoes of Echoes? An Episodic Theory of Lexical Access," *Psychological Review*, vol. 105, pp. 251–279, 1998.

[9] O. E. Scharenborg, A. G. G. Bouwman, and L. Boves, "Connected Digit Recognition with Class Specific Word Models," *COST249 Workshop on Voice Operated Telecom Services*, pp. 71–74, 2000.

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

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

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

[13] M. Padmanabhan, G. Saon, and G. Zweig, “Lattice-Based Unsupervised MLLR for Speaker Adaptation,” *ASR2000 - Automatic Speech Recognition: Challenges for the new Millenium*, 2000.

[14] 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.

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

[16] M. Bisani and H. Ney, “Bootstrap estimates for confidence intervals in ASR performance evaluation,” in *Proc. ICASSP*, (Montreal, Canada), 2004.