Using Posterior-Based Features in Template Matching for Speech Recognition

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

IDIAP–RR 06-23

June 2006

published in
ICSLP 2006

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

IDIAP Research Report 06-23

Using Posterior-Based Features in Template Matching for Speech Recognition

Guillermo Aradilla Jithendra Vepa Hervé Bourlard

June 2006

published in
ICSLP 2006

Abstract. Given the availability of large speech corpora, as well as the increasing of memory and computational resources, the use of template matching approaches for automatic speech recognition (ASR) have recently attracted new attention. In such template-based approaches, speech is typically represented in terms of acoustic vector sequences, using spectral-based features such as MFCC of PLP, and local distances are usually based on Euclidean or Mahalanobis distances. In the present paper, we further investigate template-based ASR and show (on a continuous digit recognition task) that the use of posterior-based features significantly improves the standard template-based approaches, yielding to systems that are very competitive to state-of-the-art HMMs, even when using a very limited number (e.g., 10) of reference templates. Since those posteriors-based features can also be interpreted as a probability distribution, we also show that using Kullback-Leibler (KL) divergence as a local distance further improves the performance of the template-based approach, now beating state-of-the-art of more complex posterior-based HMMs systems (usually referred to as "Tandem").

1 Introduction

Stochastic modeling and template matching are the two most successful approaches applied to ASR. In particular, the most commonly used method is based on hidden Markov models (HMMs) [1], a parametric stochastic model. HMMs benefit from efficient algorithms for training and decoding. However, they rely on some assumptions about the data distribution which are not always correct in the case of the speech signal.

Template matching offers a different approach. All the training data is used at the decoding time instead of trained models. In this case, no explicit assumption is made about the data distribution. This technique obviously requires many operations at the decoding time but this issue can be alleviated given the powerful computational resources available nowadays. For this reason, template matching has received more attention in the ASR field recently. For instance, DeWachter et al. [2] have investigated a bottom-up strategy for selecting the best templates, Axelrod et al. [3] have studied the combination of HMMs and template matching in an isolated word recognition task and we have carried out experiments for re-scoring the N-best hypotheses given the template matching-based distances [4].

Typical ASR systems use features obtained from short-term spectrum, like MFCC or PLP. Phone posterior probabilities can also be used as features as it has been demonstrated in Tandem system [5]. Studies have been carried out for studying the properties of posterior features [6]. In particular, they benefit from being more stable and robust. Hence, they are very suitable for a pattern recognition task.

To our knowledge, posteriors have never been used in the template matching context. Motivated by their good behavior as features, we study here the use of phone posteriors as features applied to template matching.

Euclidean or Mahalanobis distances have been typically used as local distance between vectors. In this work, we also investigate the use of KL-divergence as a measure of local similarity between two vectors since the posterior vector can be seen as a distribution over the phone space.

The paper is organized as follows: Section 2 describes the template matching technique and its application to ASR, Section 3 explains the posterior features and the proposed KL-divergence measure, Section 4 presents the experiments and their results and finally Section 5 gives conclusions and some ideas for future work.

2 Template Matching

Unlike parametric approaches, where information about the data is summarized into models, template-based approaches use all the information contained in the training data in a direct way. Since there is no modeling, no explicit assumption is made about the data distribution. Training data is formed by a set of templates where a template can be defined as a sequence of feature vectors that represents a particular pronunciation of a word[1]. Recognition is, then, based on finding the template most similar to the sequence of test vectors.

The similarity measure between two sequences has to deal with time warping since they usually have different lengths. The template sequence is, then, resampled to have the same length as the test sequence. The resampling function ϕ must hold some conditions on slope and boundaries, i.e., let $X = \{x_i\}_{i=1}^{N}$ be a test sequence of N frames and let $Y = \{y_j\}_{j=1}^{M}$ be a template sequence of length M, then

$$\begin{aligned} 0 \leq \phi(i) - \phi(i-1) \leq 2 \\ \phi(1) = 1 \\ \phi(M) = N \end{aligned} \tag{1}$$

[1] In this work, we consider words, but other types of linguistic units can also be represented by templates.

These conditions ensure that no more than one vector from the template can be skipped at each time. They are typical in the ASR field and they are also used in this work.

The similarity measure D between a test sequence X and a template Y can, then, be computed as

$$D(X, Y) = \min_{\{\phi\}} \sum_{i=1}^{N} d(x_i, y_{\phi(i)}) \tag{2}$$

where $\{\phi\}$ denotes the set of all possible resampling functions given by the conditions expressed in (1). The term of the sum $d(x_i, y_{\phi(i)})$ defines the local distance between the two acoustic vectors x_i and $y_{\phi(i)}$. The choice of this local distance depends on the properties of the feature space. Traditional features have typically used Euclidean or Mahalanobis distances for computing this similarity between vectors but other types of measures can be used depending on the features; this issue will be further discussed in the next section.

Although the computation of D from (2) implies searching among a large set of resampling functions, it can be efficiently computed by the dynamic time warping (DTW) algorithm [7].

In the case of isolated word recognition, the distance D as defined in (2) is computed between the test sequence and all the possible training templates. The test sequence is, then, classified as the same class as the template with the lowest distance D.

In the case of continuous speech, there is a variant of DTW known as one-pass DTW [8]. This algorithm relies on the same principle of finding the resampling function that yields the lowest total distance. In this case, though, the best resampling function results from a concatenation of templates since the test utterance usually contains more than a word. A word insertion penalty is then used to control the number of words per utterance.

The main weakness of this approach is that, if a large amount of templates is required to represent all the variability of a word, the system can be impractical since the decoding time increases exponentially with the number of templates.

3 Posterior Features

Short-term spectral-based features, such as MFCC or PLP, are traditionally used in ASR. They have been successfully applied because they can be modeled by a mixture of Gaussians, which is the typical function used to estimate the emission distribution of a standard HMM system (HMM/GMM). However, in addition to the lexical information, spectral-based features also contain knowledge about the speaker or environmental noise[2]. This extra information is cause of unnecessary variability in the feature vector, which may decrease the performance of the ASR system.

A transformation of traditional acoustic vectors can also be used as features for ASR. In particular, a multi-layer perceptron (MLP) can be trained to estimate the phone posterior probabilities based on spectral-based features. In this case, the MLP performs a non-linear transformation. Because of this discriminant projection, posteriors are known to be more stable [6] and more robust to noise (chapter 6 of [9]). These characteristics are illustrated in Figure (1). Moreover, the databases for training the MLP and for testing do not have to be the same so it is possible to train the MLP on a general-purpose database and use this posterior estimator to obtain features for more specific tasks; this approach has been studied in [10].

Also, phone posterior probabilities can be seen as phone detectors as it has been demonstrated in [11], this interpretation makes posteriors a very suitable set of features for speech recognition systems since words are formed by phones.

Despite their good properties, posterior features cannot be easily modeled by a mixture of Gaussians. In the Tandem approach [5], posteriors are used as input features for a standard HMM/GMM system. However, a PCA transform on the logarithm of the posteriors has to be done previously to

[2] For instance, there are speaker recognition systems that use MFCC features.

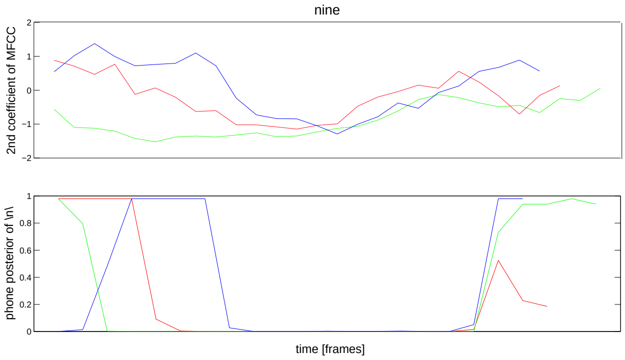

Figure 1: *This figure plots the value of one component of the feature vector in the case of MFCC features and phone posteriors for three different templates of the word 'nine'. Phone posteriors are more stable than MFCC features because of their discriminant nature.*

Gaussianize and decorrelate the feature vector. In the template matching approach, since no distribution has to be modeled, posteriors can be used directly as feature vectors.

A local distance between vectors must be defined for applying posteriors to the template matching framework. Since the vector feature of posteriors is a probability distribution over the phone space, it is appropriate to use KL-divergence when measuring the similarity between vectors. Given two distributions x and y with K classes (i.e. two feature vectors of dimension K, where each component corresponds to a particular phone), KL-divergence is defined as

$$KL(x \,||\, y) = \sum_{k=1}^{K} y(k) \log \frac{y(k)}{x(k)} \tag{3}$$

KL-divergence comes from information theory and can be interpreted as the amount of extra bits that are needed to code a message generated by the a reference distribution y, when the code is optimal for a given test distribution x [12].

KL-divergence can be used in the template matching framework as the local distance appearing in Equation (2). As this local distance is always computed between a vector from the test sequence and a vector from a template, KL-divergence fits naturally in the local distance definition by taking the reference distribution y as the vector from the template and the test distribution x as the vector from the test sequence. In our case, then, we can apply (2) as

$$D(X,Y) = \min_{\{\phi\}} \sum_{i=1}^{N} KL(x_i \,||\, y_{\phi(i)}) \tag{4}$$

4 Experiments and Results

This work must be considered as a first experiment to evaluate the effectiveness of the phone posteriors when applied to template matching. With this purpose, we have chosen a continuous digit recognition task to test our hypothesis that posterior features can outperform traditional features.

Test utterances and templates have been extracted from the OGI Numbers v1.3 database [13]. This data has been recorded through a telephone channel and a large variety of speakers is represented. For testing, we have chosen 2820 utterances where all the digits appear in a similar number. The number of templates is the same for every word in the lexicon. Templates were obtained by a force alignment process given by a state-of-the-art HMM system. The lexicon has 12 different words (from 'zero' to 'nine' plus 'oh' and 'silence').

MFCC features contain 26 dimensions[3], 13 static features (12 MFCC coefficients and the log energy) plus their delta features. These features are normalized in mean and variance.

Posterior features were obtained using a MLP trained on a smaller version of OGI Numbers, the version 1.0. The MLP has one hidden layer with 1800 units. PLP features jointly with delta and acceleration features are used as inputs. There are 27 output units, each of them corresponding to a different phone. The MLP was trained using the relative entropy criterion.

Since we are working with a continuous speech database, our template matching system is based on one-pass DTW [8]. Constraints for the resampling function are the same as defined in (1) and a word insertion penalty is used to equalize insertion and deletion errors.

A comparison between MFCC features and posteriors was first carried out. Two types of local distances were used: Euclidean and KL-divergence (KL-divergence cannot be applied to MFCC features since they are not distributions). Table 1 presents the results.

Templates per word	MFCC Euclidean	Posteriors Euclidean	Posteriors KL-divergence
10	60.6	93.2	95.6
20	72.4	93.5	95.4
30	73.4	94.0	95.5
40	78.7	93.6	95.6
50	80.0	93.2	95.6

Table 1: *System accuracy using one-pass DTW. The first column shows the number of templates per word available. Three different experiments are presented: MFCC features using Euclidean distance, posteriors using Euclidean distance and posteriors using KL-divergence as local distance.*

We can observe that, when using MFCC features with Euclidean distance, the accuracy increases with the number of templates, but still the performance is far below state-of-the-art for this particular task. The high variability present in MFCC features decreases the performance of the system. However, there is a significant improvement when using posterior features still using Euclidean distance. This supports the evidence that posteriors are more stable and hence, more suitable for being used as features. There is still a very significant improvement when KL-divergence is used as a local measure between vectors, in this case, results can start to be comparable to state-of-the-art systems on this task (in this case, a standard HMM/GMM system achieves 96.4% of accuracy).

From Table 1, we can also observe that the accuracy remains stable when increasing the number of templates. To study the influence of the amount of templates, we carry out a second experiment where we vary the number of templates. Results are shown in Table 2.

Templates per word	Posteriors KL-divergence
1	76.4
2	89.7
4	94.8
6	95.2
8	95.7
10	95.6

Table 2: *System accuracy using one-pass DTW. The first column indicates the number of templates per word used for decoding.*

[3] Feature vectors with 13 and 39 dimensions were also used but the performance was worse. Dynamic features always improve the accuracy but acceleration features use a too wide context in the case of DTW.

In this case, we can see that one template per word is not enough for obtaining the maximum accuracy given by this template matching approach. Results get better when increasing the number of templates until we reach 8 representations per word. Then, system accuracy remains stable. From this experiment we can observe that a few examples are enough to represent properly all the variations of a particular word because of the high stability of posterior features. This issue is very important since the decoding time of DTW increases exponentially with the number of templates. A reduced number of templates makes the system feasible in practice.

We also compare one-pass DTW approach with Tandem system [5] because both systems use posteriors as input features. Tandem system uses post-processed posterior features with a HMM/GMM-based acoustic model. The HMM/GMM part has been trained using 8000 utterances from the OGI Numbers v1.3 database and a HMM has been trained for each word. Table 3 presents the results of this comparison. A HMM/GMM system using MFCC features has also been trained. MFCC acoustic vectors contain delta and acceleration features (39 dimensions).

MFCC	96.4
TANDEM	94.2
DTW	95.6

Table 3: *System accuracy for a standard HMM/GMM system using MFCC features, a Tandem system and one-pass DTW using 10 templates per word.*

One-pass DTW with posteriors and KL-divergence outperforms Tandem system even if both systems use the same input features. This result suggests that one-pass DTW is able to use the information given by the posteriors more efficiently that Tandem system, mainly because it does not assume a distribution of the input vectors. In spite of using only 10 templates per words, one-pass DTW achieves comparable results to the HMM/GMM system using MFCC features.

In Section 3 we explained that, when computing the KL-divergence, the vectors belonging to the template should play the role of the reference distribution while the test vectors should be considered as the test distribution. We consider to do some small variations in the computation of the KL-divergence to test our natural interpretation. We use the symmetric version of KL-divergence:

$$KL_{sym}(x\,||\,y) = \frac{1}{2}[KL(x\,||\,y) + KL(y\,||\,x)] \tag{5}$$

and we also try the reverse KL, i.e. we consider the test distribution as the templates vectors and the reference as the test vectors. As we can see in Table 4, our assumption is the one which yields the best result.

KL	95.6
Symmetric KL	95.1
Reverse KL	93.2

Table 4: *System accuracy when using 10 templates per word. Symmetric KL uses the symmetric version of this measure. In reverse KL, we switched the test and the reference vectors.*

5 Conclusions and Future Work

In this work, we have carried out some experiments to test the convenience of posterior features in a template matching approach for ASR. The following conclusions can be drawn:

- Posterior features outperform MFCC features in the template matching approach. Their good properties on stability and robustness are supported by the results of our experiments.

- KL-divergence is able to better estimate the similarity between two posterior vectors. Moreover, test and reference distributions play a different and significant role on the computation.
- Given the high stability of posterior features, a reduced number of templates is required to represent all the variability of a word. Hence, the system is practical in terms of decoding time.

Template matching offers a very interesting approach for recognizing speech because no distribution must be modeled and, hence, no explicit assumption has to be made about the data. However, generalization to larger vocabulary recognition tasks has not been investigated yet. This was unfeasible when using traditional features because the huge amount of templates that was required was making the decoding time prohibitive. From the results of this work, only a reduced number of templates per word is necessary to achieve good performance when using posterior features. Therefore, application of template matching approach to large vocabulary systems is now practical. Furthermore, strategies based on pruning or re-scoring can be used to reduce the decoding time.

We ran another experiment where we chose a different set of 10 templates per word. In this case the one-pass DTW system was able to achieve 96.0% of accuracy. This result shows that the choice of templates is important and future work should be focused on investigating criteria for selecting the most representative templates. These criteria could come from the information theory field since, as we have seen with the application of KL-divergence, it fits very well in this approach.

Another possibility offered by posterior features is that it is possible to train a language independent MLP for obtaining the posteriors. Then, we can generate the templates depending on each specific task. In this way, the MLP need not to be trained for each different system. Multi-lingual recognition tasks would fit very well in this framework.

6 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)”.

References

[1] L. R. Rabiner, “A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” *Proceedings of the IEEE*, pp. 257–286, 1989.

[2] M. De Wachter, K. Demuynck, D. Van Compernolle, and P. Wambacq, “Data Driven Example Based Continuous Speech Recognition,” *Proceedings of Eurospeech*, pp. 1133–1136, 2003.

[3] S. Axelrod and B. Maison, “Combination of Hidden Markov Models with Dynamic Time Warping for Speech Recognition,” *Proceedings of ICASSP*, vol. I, pp. 173–176, 2004.

[4] G. Aradilla, J. Vepa, and H. Bourlard, “Improving Speech Recognition Using a Data-Driven Approach,” *Proceedings of Interspeech*, pp. 3333–3336, 2005.

[5] H. Hermansky, D. Ellis, and S. Sharma, “Tandem Connectionist Feature Extraction for Conventional HMM Systems,” *Proceedings of ICASSP*, 2000.

[6] Q. Zhu, “On Using MLP features in LVCSR,” *Proceedings of ICSLP*, 2004.

[7] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[8] H. Ney, “The Use of One-Stage Dynamic Programming Algorithm for Connected Word Recognition,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, pp. 263–271, 1984.

[9] S. Ikbal, *Nonlinear Feature Transformations for Noise Robust Speech Recognition*, Ph.D. thesis, Ecole Polytechnique Fédéral de Lausanne, 2004.

[10] S. Sivadas and H. Hermansky, "On the Use of Task Independent Training Data in Tandem Feature Extraction," *Proceedings of ICASSP*, 2004.

[11] P. Niyogi and M. M. Sondhi, "Detecting Stop Consonants in Continuous Speech," *The Journal of the Acoustic Society of America*, vol. 111, no. 2, pp. 1063–1076, 2002.

[12] T. M. Cover and J. A. Thomas, *Information Theory*, John Wiley, 1991.

[13] R. Cole, M. Fanty, Noel M., and T. Lander, "New Telephone Speech Corpora at CSLU," *Proceedings of Eurospeech*, pp. 821–824, 1995.