Posterior-Based Features and Distances in Template Matching for Speech Recognition

Guillermo Aradilla [a] Hervé Bourlard [a]

IDIAP–RR 07-41

September 2007

published in
MLMI 2007

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

IDIAP Research Institute www.idiap.ch
Av. des Prés−Beudin 20 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 07-41

Posterior-Based Features and Distances in Template Matching for Speech Recognition

Guillermo Aradilla Hervé Bourlard

September 2007

published in
MLMI 2007

Abstract. The use of large speech corpora in example-based approaches for speech recognition is mainly focused on increasing the number of examples. This strategy presents some difficulties because databases may not provide enough examples for some rare words. In this paper we present a different method to incorporate the information contained in such corpora in these example-based systems. A multilayer perceptron is trained on these databases to estimate speaker and task-independent phoneme posterior probabilities, which are used as speech features. By reducing the variability of features, fewer examples are needed to properly characterize a word. In this way, performance can be highly improved when limited number of examples is available. Moreover, we also study posterior-based local distances, these result more effective than traditional Euclidean distance. Experiments on Phonebook database support the idea that posterior features with a proper local distance can yield competitive results.

1 Introduction

Hidden Markov models (HMMs) constitute the dominant approach for automatic speech recognition (ASR) systems. Their success is mainly based on their efficient algorithms for training and testing. However, these algorithms rely on some assumptions about data that do not hold for speech signals, such as piece-wise stationary or independence of the feature vectors given a state. Template matching (TM) is a different approach for ASR that relies on the fact that a class can be described by a set of examples (templates). Since templates are real utterances, they can better model the dynamics of the trajectories generated by the speech features compared with HMM states in currently used monophone or triphone models. Moreover, TM is preferred in those cases where simplicity and flexibility for training and testing must be considered.

As a non-parametric appraoch, TM requires more training data than parametric models, such as HMM-based systems, to obtain comparable performance. Given the increase of large speech corpora and computational resources, TM has recently drawn new attention. Investigation on this approach has been focused on increasing the number of templates [1, 2, 3] and, hence, improving its generalization capabilities. Since no speech corpora can guarantee to provide many examples for each word, sub-word units are typically used to ensure that a large enough number of templates is available for each possible word. Pronunciation dictionaries are, in this case, needed for concatenating these subword units into words. However, pronunciation of the words is not always easy to obtain, e.g., proper names.

We propose a different method to use the information contained in large speech corpora. Traditional features used in TM are based on short-term spectrum. These features contain linguistic information but also information about the gender[1] and the environment, i.e., they are speaker and task-dependent. In this work, we investigate the use of posterior probabilities of subword units as speech features. These posteriors can be estimated from a multilayer perceptron (MLP) which has been trained on large speech corpora. In this way, the MLP can capture the information contained on large speech corpora to generate speaker and task-independent features. Given the discriminative training procedure of the MLP and the long acoustic context used as input, posterior features are known to be more stable and more robust to noise than spectral-based features [4]. Since these features only contain, in theory, linguistic information, fewer templates are required to represent a word. Hence, in those applications where the number of available templates is few, we can expect to improve the performance. Posteriors estimates from the MLP outputs have already been successfully applied as features for ASR using HMM/GMM as acoustic model, system known as Tandem [4, 5].

TM-based approaches traditionally use Euclidean or Mahalanobis distance as local similarity measure between features. These distances implicitly assume that features follow a Gaussian distribution. This assumption does not hold when using posterior distributions as features. Since posterior features are probability distributions over the space of subword units, more appropriate distances can be considered. In this work, we investigate local distances between frames that take into account the discriminative properties of posterior features.

This work is an extension of a previous experiment where we already applied posterior features to a TM-based ASR system [6]. On that first experiment, posterior features were not task-independent because the data to train the MLP belonged to the same database as the test set. Kullback-Leibler (KL) divergence was applied as local distance for being a natural distance between distributions. In this work, the MLP is trained on a large speech corpus and used for a different recognition task. We also show that other types of local distances can be successfully applied to posterior features which obtain similar performance to KL-divergence but are faster to compute.

This paper is summarized as follows: Section 2 introduces the TM approach for speech recognition, Section 3 presents the posteriors features, Section 4 describes the local distances investigated in this work, Section 5 presents the experiments and results and finally, Section 6 draws some conclusions.

[1] For instance, speaker recognition systems use spectral-based features as inputs.

2 Template Matching

TM is a non-parametric classifier that relies on the idea that a class w can be identified by a set of N_w examples (templates) $\{\mathbf{Y}_n^w\}_{n=1}^{N_w}$ belonging to that class. Unlike parametric models, TM directly uses all training data at the decoding time and no explicit assumption is made about the data distribution. A test element $\mathbf{X}$ is associated to the same class as the closest sample based on a similarity function φ between samples defined as:

$$\text{class}(\mathbf{X}) = \arg\min_{\{w'\}} \min_{\mathbf{Y}' \in \{\mathbf{Y}_n^{w'}\}} \varphi(\mathbf{X}, \mathbf{Y}') \tag{1}$$

where $\{w'\}$ denotes the set of all possible classes. However, as any non-parametric technique, a large amount of training data is required to obtain a good classification performance. TM has recently received new attention in the ASR field because current computational resources and speech corpora allow to deal with large amount of training data in a practical computational time.

In the case of speech, templates are sequences of feature vectors that correspond to particular pronunciations of a word. When comparing with HMMs, templates can describe in more detail the dynamics of the trajectories defined by speech features because they represent real utterances, whereas HMMs are parametric representations that summarize the information contained on the speech trajectories. Furthermore, the explicit use of non-linguistic information such as gender or speech rate can be easily applied when using templates but this type of long-span information is more difficult to incorporate into a parametric model.

The similarity measure φ between sequences must deal with the fact that utterances usually have different lengths. This measure is based on dynamic time warping (DTW) [7] and it minimizes the global distortion between two temporal sequences. This global distortion is computed as the sum of local distances $d(\mathbf{x}, \mathbf{y})$ between the matched frames. This matching is performed by warping one of the two sequences. In speech, the template sequence is typically warped so that every template frame $\mathbf{y}_m$ matches a frame of the test sequence $\mathbf{x}_n$. Given a template sequence $\{\mathbf{y}_m\}_{m=1}^{M}$ and a test sequence $\{\mathbf{x}_n\}_{n=1}^{N}$, DTW-based distance can be expressed as

$$\varphi(\mathbf{X}, \mathbf{Y}) = \min_{\{\phi\}} \sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{y}_{\phi(i)}) \tag{2}$$

where $\{\phi\}$ denotes the set of all possible warping functions for the template sequence. The warping function must hold some constraints of continuity and boundaries to ensure that the resampled template sequence is realistic. Typical constrains in the ASR field are:

$$\begin{gathered} 0 \leq \phi(i) - \phi(i-1) \leq 2 \\ \phi(1) = 1 \\ \phi(M) = N \end{gathered} \tag{3}$$

These conditions guarantee that no more than one frame from the template sequence will be skipped for each test frame and also, that every test frame will be related to only one template frame.

Although the computation of (2) implies searching among a large set of warping functions, it can be efficiently computed by dynamic programming.

The local distance $d(\mathbf{x}, \mathbf{y})$ is typically chosen as Euclidean or Mahalanobis distance since spectral-based features are normally used for representing the speech signal. However, other types of similarity measures between frames can also be applied depending on the properties of the features. In Section 4, a description of the local distances investigated in this work will be given.

As described before, recent investigation to improve the performance of TM-based ASR systems is to take advantage of the current large speech corpora and computational resources by increasing the number of templates. TM becomes then a search problem among all possible templates [1]. In order to increase the speed and efficiency of the search, non-linguistic information can be used at the decoding

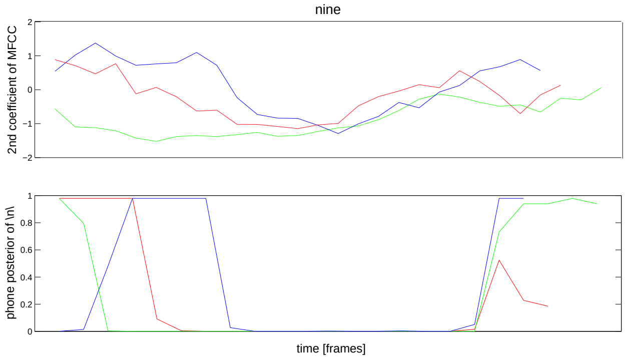

Figure 1: *The value of the second component of the feature vector in the case of MFCC features and phone posterior corresponding to the phoneme /n/ are plotted for three different templates of the word "nine". It can be seen that values from spectral-based feature vectors are more variable within a phone than posterior features, which follow a more stationary trajectory.*

time [8]. As templates and HMMs convey different types of information since they are different types of models, investigation has also been carried out for combining both approaches [2, 3] with successful results. However, this technique requires a large amount of samples per word (or linguistic unit). In this work, we will focus on the situation where a few samples are given for every word. In this case, the goal is to reduce as much as possible the variability within a class so that a few samples will be enough to represent a class word. This variability reduction will be performed at the feature level and will be explained in detail in the next section.

3 Posterior Features

The posterior probability $p(q_k|z_t)$ of a phoneme q_k given a spectral-based acoustic feature z_t at time t can be estimated from a MLP. A posterior vector x_t can then be obtained where each dimension corresponds to a phoneme posterior $x_t = \{p(q_k|z_t)\}_{k=1}^{K}$. K corresponds to the total number of phonemes and is also the number of MLP outputs[2].

If posterior estimates were correct, these features could be considered as optimal speech features by assuming that words are formed by phonemes since, in theory, they only carry linguistic information and also, they can be seen as optimal phone detectors as it is demonstrated in [9]. This reduction of the undesirable information makes posterior features more stable as it is illustrated in Figure 1.

Traditional features, like MFCC [10] or PLP [11], contain information about the spectrum and hence, about the speaker and its environment. However, posterior features can be considered speaker and task-independent since they only contain information about the phoneme that has been pronounced. Rigorously speaking, posterior features are not task-independent since the MLP is implicitly learning the prior probability of each phoneme, which will be dependent of the database. However, when using large vocabulary corpora, these probabilities converge to phoneme priors of the language of the database. In this way, posterior features are language-dependent.

[2] We are using this notation for the sake of simplicity, but in fact an acoustic context (typically 4 frames) is used as input of the MLP, hence, rigorous notation should be $p(q_k|z_{t-\Delta}^{t+\Delta})$

4 Local Distance

From (2), it can be observed that DTW-based distance requires a distance $d(\mathbf{x}, \mathbf{y})$ between reference and test samples of the observation space. Since any local distance assumes a particular geometry of the observation space, the choice of the local distance plays a crucial role on the performance of the system. Traditionally, these distances are based on Euclidean and Mahalanobis distances. In the TM-based approach, investigation has been recently carried out to estimate the parameters of the weighting matrix of the Mahalanobis distance to improve the performance. A maximum-likelihood estimation was described in [12] and a discriminative procedure was presented in [13]. However, these methods require a large amount of data to properly estimate the weights.

Since posterior vectors can be seen as distributions over the space of subword units (e.g., phonemes), measures from the information theory field can be applied. These measures can capture higher order statistics from the data than Euclidean-based distances. Furthermore, they can explicitly consider the particular properties of posterior vectors (i.e., values must be non-negative and sum must be equal to one).

In the following, we will consider that y represents a frame from the template and x denotes a frame from the test sequence. As explained before, x and y can be considered discrete distribution on the $\mathbb{R}^K$ space (i.e. there are K different phonemes).

In addition, local distance directly affects the decoding time since computing the local distance is the most frequent operation on the DTW algorithm. Hence, the choice of the local distance should also take into account its computational time.

4.1 Squared Euclidean Distance

This is the traditional distance used as local distance between frames. However, it is related with the Gaussian distribution. Indeed, when taking the logarithm of a Gaussian distribution with unity covariance matrix, it becomes the squared Euclidean distance plus a constant factor.

$$D_{Eucl}(x, y) = \sum_{k=1}^{K} (x(k) - y(k))^2 \tag{4}$$

However, when measuring the similarity between posterior features, Euclidean distance is not very appropriate since posterior space holds some special properties which are not taken into account by this distance.

4.2 Kullback-Leibler Divergence

KL divergence (or relative entropy) comes from the information theory field 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 [14].

$$D_{KL}(x \,||\, y) = \sum_{k=1}^{K} y(k) \log \frac{y(k)}{x(k)} \tag{5}$$

KL-divergence is a natural measure between distributions. The fact that it is not symmetric must not affect its application to DTW algorithm. In this case, the reference distribution y is considered to be the template frame whereas x corresponds to the test frame.

4.3 Bhattacharyya Distance

This distance was initially motivated by geometrical considerations since it computes the cosine between two distributions [15]. It is also a particular case of the Chernoff bound (an upper bound for the Bayes error) [16].

$$D_{Bhatt}(x,y) = -\log \sum_{k=1}^{K} \sqrt{x(k)y(k)} \tag{6}$$

Bhattacharyya distance is symmetric and also it is faster to compute than KL divergence because less logarithms must be computed. This distance has been used already in speech processing for phone clustering [17].

4.4 Distance Based on Bayes Risk

Bhattacharyya distance is originated from an upper bound of the Bayes risk. However, the exact probability of error can be easily computed for discrete distributions [18]:

$$\text{Bayes Error} = \sum_{k=1}^{K} \min\{x(k), y(k)\} \tag{7}$$

A distance can be derived similar to Bhattacharyya distance by taking the negative logarithm:

$$D_{Bayes}(x,y) = -\log \sum_{k=1}^{K} \min\{x(k), y(k)\} \tag{8}$$

This distance is even simpler to compute than (6) because it avoids the square root function.

5 Experiments

5.1 Description

In this work, Phonebook database has been used to carry out word recognition experiments using the TM-based approach. This database consists of 47455 utterances of isolated words. There are 3992 different words pronounced by around 12 different speaker in average. Experiments with different lexicon sizes have been carried out: 5, 10, 20, 50 and 100 different words were selected randomly from the global lexicon. For each experiment and each word, one or two utterances have been selected as templates and the rest of utterances containing the selected words have been used for test. Since lexicon has been selected at random, experiments have been repeated ten times using a different lexicon at each time. Results have been consistent, i.e., similar results have been obtained at each time and average results are shown.

Two types of features have been considered: PLP and phoneme posterior probabilities. PLP features also contain delta features. Posterior features have been obtained from a MLP trained on 30 hours of the CTS database following the MRASTA procedure [19]. The MLP contains 2000 hidden units and 46 phonemes (including silence) have been considered.

Constraints for DTW are the same as described in Formula 3. Euclidean, KL-divergence, Bhattacharyya and Bayes-based distance are considered as local distances. PLP features only use Euclidean distance (the rest of local distance can only be applied to discrete posterior vectors).

Experiments on decoding time have been carried out on a workstation with a Athlon64 4000+ processor.

5.2 Results

Results on Table 1 show the effectiveness in using posterior features for TM. PLP features contain information about the speaker and since the task is speaker-independent, results when using these spectral-based features are far from being competitive. This explains why TM is mainly focused on speaker-dependent tasks with small vocabulary. On the other hand, posterior features have been

		one template						two templates						
lexicon		PLP	Posteriors					PLP	Posteriors					# test
size		Eucl	Eucl	KL	Bhatt	Bayes		Eucl	Eucl	KL	Bhatt	Bayes		utts
5		79.3	93.2	98.2	98.7	98.0		90.8	96.6	98.9	98.9	98.5		55
10		74.7	91.9	97.8	98.3	97.5		85.4	95.7	98.9	98.9	98.4		104
20		69.8	89.5	95.6	96.5	95.7		81.9	94.2	98.4	97.9	97.5		212
50		59.7	83.1	92.9	94.1	92.9		74.2	90.2	96.6	96.8	96.1		545
100		53.2	78.5	89.7	91.4	89.7		68.0	87.5	94.9	95.1	94.2		1079

Table 1: System accuracy when using one or two templates per word. The size of the lexicon has been varied to observe the effect of increasing the lexicon. The last column shows the average number of test utterances.

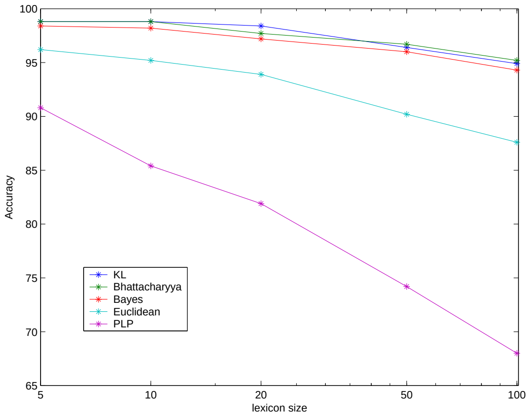

Figure 2: *Accuracy of the system using 2 templates per word.*

estimated by taking into account the information captured by the MLP from the large speech corpus used for training. This, jointly with the discriminative training of the MLP make posterior features robust to speaker and environment conditions.

Moreover, posterior-based distances such as KL divergence, Bhattacharyya and Bayes-based distance yield better results than traditional Euclidean distance since they explicitly deal with the space topology of the posterior features.

Figure 2 plots the system accuracy with two templates per word and also shows the effect of increasing the size of the lexicon. When using 100 different words, the performance of the system is still around 95%, which is reasonable result given the complexity of the task and the limited amount of samples per word[3].

Experiments have been carried out to investigate the effect of the local distance on the decoding time. Results are shown in Figure 3. It can be observed that KL-divergence takes a long time for decoding because of the logarithm function. Bhattacharyya distance replaces the logarithm function by a square root function, which takes less time than the logarithm. Bayes-based distance is faster than the previous since selecting the minimum value is a very simple operation. Finally, Euclidean distance is faster than the rest but its accuracy is significantly worse than the other distances.

[3] Experiments comparing templates and hybrid HMM/MLP [20] have been carried out using the test set described in [21]. There are 8 different test sets consisting each one of 75 different words. In this case, we obtained similar results in both systems, i.e. around 95% accuracy.

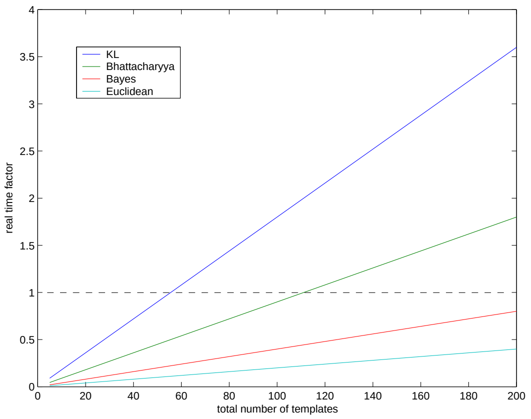

Figure 3: *This figure shows the real time factor depending on the total number of templates. Real time factors is defined as the ratio between the decoding time and the duration of the test sequence. A dashed line indicates when the decoding time is equal to the duration of the sequence.*

6 Conclusion

In this work we have tested the effectiveness of posterior features on a TM-based approach. Since these features have been trained using a large vocabulary database, they can be considered speaker and task-independent. These properties make these features very suitable for those conditions where a word must be represented by a few examples. Moreover, the choice of the local distance has been investigated since it both assumes a topology on the feature space and also directly affects the decoding time. Though KL-divergence is a very appropriate local distance when using posterior features, it takes a long time to be computed because it requires a logarithm function for each dimension of the posterior vector. Other types of distances based on the probability of error have also been investigated which are simpler to compute and yield similar performance.

Future work should be focused on investigating other ways to incorporate information of large speech corpora on TM-based approach. A possible way would be to combine the posterior features from different MLPs. Initial experiments have already been carried out with successful results.

Acknowledgments.

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] Wachter, M.D., Demuynck, K., Compernolle, D.V., Wambacq, P.: Data Driven Example Based Continuous Speech Recognition. Proceedings of Eurospeech (2003) 1133–1136

[2] Aradilla, G., Vepa, J., Bourlard, H.: Improving Speech Recognition Using a Data-Driven Approach. Proceedings of Interspeech (2005) 3333–3336

[3] Axelrod, S., Maison, B.: Combination of Hidden Markov Models with Dynamic Time Warping for Speech Recognition. Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP) **I** (2004) 173–176

[4] Zhu, Q., Chen, B., Morgan, N., Stolcke, A.: On Using MLP features in LVCSR. Proceedings of International Conference on Spoken Language Processing (ICSLP) (2004)

[5] Hermansky, H., Ellis, D., Sharma, S.: Tandem Connectionist Feature Extraction for Conventional HMM Systems. Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2000)

[6] Aradilla, G., Vepa, J., Bourlard, H.: Using Posterior-Based Features in Template Matching for Speech Recognition. Proceedings of International Conference on Spoken Language Processing (ICSLP) (2006)

[7] Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice Hall (1993)

[8] Aradilla, G., Vepa, J., Bourlard, H.: Using Pitch as Prior Knowledge in Template-Based Speech Recognition. Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2006)

[9] Niyogi, P., Sondhi, M.M.: Detecting Stop Consonants in Continuous Speech. The Journal of the Acoustic Society of America **111**(2) (2002) 1063–1076

[10] Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Audio, Speech and Signal Processing **28** (1980) 357–366

[11] Hermansky, H.: Perceptual Linear Predictive (PLP) Analysis of Speech. The Journal of the Acoustic Society of America **87** (1990)

[12] Wachter, M.D., Demuynck, K., Wambacq, P., Compernolle, D.V.: A Locally Weighted Distance Measure For Example Based Speech Recognition. Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2004) 181–184

[13] Matton, M., Wachter, M.D., Compernolle, D.V., Cools, R.: A Discriminative Locally Weighted Distance Measure for Speaker Independent Template Based Speech Recognition. Proceedings of International Conference on Spoken Language Processing (ICSLP) (2004)

[14] Cover, T.M., Thomas, J.A.: Information Theory. John Wiley (1991)

[15] Bhattacharyya, A.: On a Measure of Divergence between Two Statistical Populations Defined by their probability distributions. Bull. Calcutta Math. Soc. **35** (1943) 99–109

[16] Fukunaga, K.: Introduction to Statistical Pattern Recogntion. Morgan Kaufmann, Academic Press (1990)

[17] Mak, B., Barnard, E.: Phone Clustering Using the Bhattacharyya Distance. Proceedings of International Conference on Spoken Language Processing (ICSLP) (1996) 2005–2008

[18] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley-Interscience (2001)

[19] Hermansky, H., Fousek, P.: Multi-Resolution RASTA Filtering for TANDEM-based ASR. Proceedings of Interspeech (2005)

[20] Bourlard, H., Morgan, N.: Connectionist Speech Recognition: A Hybrid Approach. Volume 247. Kluwer Academic Publishers, Boston (1993)

[21] Dupont, S., Bourlard, H., Deroo, O., Fontaine, V., Boite, J.M.: Hybrid HMM/ANN Systems for Training Independent Tasks: Experiments on Phonebook and Related Improvements. Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP) (1997)