IDIAP RESEARCH REPORT

Using Pitch as Prior Knowledge in Template-Based Speech Recognition

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

IDIAP–RR 05-65

October 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 05-65

Using Pitch as Prior Knowledge in Template-Based Speech Recognition

Guillermo Aradilla Jithendra Vepa Hervé Bourlard

October 2005

Abstract. In a previous paper on speech recognition, we showed that templates can better capture the dynamics of speech signal compared to parametric models such as hidden Markov models. The key point in template matching approaches is finding the most similar templates to the test utterance. Traditionally, this selection is given by a distortion measure on the acoustic features. In this work, we propose to improve this template selection with the use of meta-linguistic information as prior knowledge. In this way, similarity is not only based on acoustic features but also on other sources of information that are present in the speech signal. Results on a continuous digit recognition task confirm the statement that similarity between words does not only depend on acoustic features since we obtained 24% relative improvement over the baseline. Interestingly, results are better even when compared to a system with no prior information but a larger number of templates.

1 Introduction

Hidden Markov models (HMMs) [1] are the most successful approach in automatic speech recognition (ASR). They project the sequence of feature vectors generated from speech utterances onto a state sequence which follows a first order Markov property. Their success is due to their generalisation capability and scalability. However, as any parametric model, HMMs make some assumptions about the data. In particular, HMMs assume that speech signal is a piecewise stationary stochastic process. This property holds true within vowel sounds but it may not within other sounds, such as plosives. Moreover, speech signal is not stationary at transitions from one sound to another since feature vectors describe a continuous trajectory.

Attempts have been investigated to deal with this weakness in the HMM paradigm. One approach has been segmental HMMs [2], where each state generates jointly a set of frames, unlike conventional HMMs, which look at the signal at the frame level. In another approach, experiments have been carried out to exploit the continuity property of the trajectories described by the sequence vectors where dynamic features were used for smoothing the trajectory given by the mean values of the state sequence [3]. This approach has offered good results in speech synthesis and significant improvements in speech recognition.

Templates offer another way to describe trajectories. A template consists of a sequence of feature vectors representing an utterance of a word. No explicit assumption about the data must be made for the template-based approach since there is no model, only utterance representations. In our previous work [4], we used template-based approach for re-scoring N-best hypotheses generated by a state-of-the-art HMM-based system. We showed that templates have some convenient properties, such as a proper description of the continuity of the trajectory, that help HMMs to increase their performance.

The main idea behind template matching is finding the most similar templates for each test word. In our previous work, this process was based only on a distortion measure between the acoustic features of the templates and the test sequences. In this work, we study the use of other types of information in the speech signal for selecting the most appropriate templates. In particular, we use pitch information to decide which templates must be used.

This paper is organized as follows: in Section 2, we explain the template based approach. Section 3 describes the convenience of using extra information and its application to pitch frequency, then Section 4 gives the details about the experiments and reports the results and finally, in Section 5, we present our conclusions and outline some of our plans for future research.

2 Template-based Approach

A template is a sequence of feature vectors which represents the way a certain word is uttered. Templates offer a different way to represent the trajectories that, unlike HMMs, does not make any explicit assumption about the data. As HMMs, they follow the pattern matching approach for recognition, therefore, they need a measure for comparing patterns. This measure is based on the distortion between the trajectories described by the test and the reference sequences and it is efficiently computed using the dynamic time warping (DTW) technique [1].

Indeed, templates can be seen as a kind of HMM where each state is a frame and emission probability consists of a single Gaussian with mean equal to the value of the reference frame and variance equal to one. As templates are longer than the underlying Markov chain in HMMs, the piecewise stationary assumption is weakened since each frame (state) matches with a lower number of frames of the test sequence. Variability inherent to speech which is modeled by a mixture of Gaussians in HMMs is now handled by the variability of templates, since each template represents a unique way of pronunciation of a word. Evidently, the more templates available, the better the accuracy of the system since more variability will be represented. Furthermore, the use of a relative large amount of data does not represent a serious problem given the powerful computer resources available nowadays.

Another weakness about HMM paradigm is related with the way emission probabilities are estimated: in state-of-the-art HMMs with a mixture of Gaussians as emission probability distributions, there is no mean to impose continuity constraints on the trajectory. Thus, an observation sequence can be recognized using a sequence of mixtures which has never been observed in the training set [5]. This phenomenon can never happen within template matching framework as templates can be considered as HMMs with unimodal distributions.

In a previous experiment [4], we applied template matching to continuous speech recognition task. The main difficulty for applying templates to continuous speech is that all possible word sequences must be studied. This process is performed with HMMs in an efficient way by using Viterbi algorithm [6] but it becomes practically infeasible when dealing with a large number of templates for each word since all the possible combinations grow exponentially with the size of vocabulary and number of templates. Thus, we reduced the search space by means of a state-of-the-art HMM-based system which generated a list of N best hypotheses [1] $\{H_n\}_{n=1}^{N}$ with their corresponding word boundaries for any given test utterance. Then, template-matching based technique is used. For the sequence of vectors of the ith word in the nth hypothesis X_i^n, a measure $D_{w_i^n}(X_i^n)$ is computed using DTW distortion with all the templates corresponding to that word w_i^n. This measure is an averaged sum of the K lowest DTW distortions:

$$D_c(X) = \min_{\{Y_k^c\}_{k=1}^{K}} \frac{1}{K} \sum_{k=1}^{K} D_{DTW}(X, Y_k^c) \tag{1}$$

where c is the word given by the hypothesis and $\{Y_k^c\}_{k=1}^{K}$ is the set of K templates corresponding to the word c with the lowest DTW distortion. The value of K is greater than one and lower than the total number of templates to reduce the effect of outliers produced by templates with incorrect word boundaries.

The constraint used for the DTW distortion is similar to HMMs, but it also allows to skip one frame. We can, therefore, deal with reference patterns that are longer than the test sequence. Figure 1 shows this constraint graphically. For computing the local distance, we first normalize the data with mean and variance and then, we use Euclidean distance.

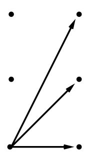

Figure 1: DTW constraint. Horizontal and vertical axis correspond to the test and reference sequences respectively.

A score S_n is assigned to each hypothesis H_n according to the measures given to each of its words $\{w_i^n\}_{i=1}^{M_n}$ where M_n is the number of words in hypothesis H_n.

$$S_n = \sum_{i=1}^{M_n} D_{w_i^n}(X_i^n) \tag{2}$$

The main limitation of this approach is that it is practically infeasible to have all representations of each word. In order to cope with speech variability, the selection of the most similar templates is extended with the incorporation of additional information which is different from the acoustic features. This is the main focus of this work and it will be explained in detail in the next section.

[1] The N hypotheses with the highest likelihoods

3 Template Selection

Similarity between templates and a test word is a crucial point in the template matching approach. The more similar the templates are to the test word, the more likely is that templates belong to the same class as the test word. In our previous work presented in [4], this similarity was based only on the minimum distortion between the trajectories described by the vector sequences of the test and the templates. Hence, we were assuming that similarity between patterns was based on only acoustic information carried by the feature vectors.

We can consider to extend the idea of similarity to other kinds of information present in the speech signal, such as gender, duration or position in the sentence. This meta-linguistic information could be used as a prior knowledge for speech recognition process. In this direction, templates would be clustered according to a criterion given by the extra information. For the decoding phase, test utterances would be compared to those templates which share the same category of the meta-linguistic information.

With the use of meta-linguistic information as a prior knowledge, the concept of similarity is not restricted to acoustic features but also to other features which are also present in the speech signal and cannot be captured by the acoustic feature vector. Indeed, the use of meta-linguistic information is supported by perceptual experiments conducted by Goldinger [7]. These experiments show that utterances spoken by familiar speakers are better recognized by human candidates than those spoken by unfamiliar speakers. This study indicates that prior knowledge about the speaker can influence the speech recognition performance and justifies the inclusion of information in addition to the acoustic features for computing similarity between vector sequences.

The idea of selecting a more specific model given the test conditions has also been used with HMMs with successful results [8]. For instance in [9], they used time information for building long and short duration models which were selected according to the duration conditions of the test utterance. The application of this prior knowledge in this case yielded significant improvement in the recognition performance.

For this first experiment in the use of meta-linguistic information with templates, we have carried out a classification of templates database depending on their pitch. For the time being, two classes have been considered: low and high pitch frequency which represent roughly male and female. In this way, we can say that templates have become gender-dependent. For each test utterance, pitch information will be computed. Then, only templates matching the same pitch category will be used to compute the score of each hypotheses using Equations 1 and 2.

This work can be considered as a first experiment using templates with prior knowledge and only one type of information with two classes have been carried out. Obviously, this method can be extended to more classes and other kinds of information such as duration, position or speech rate.

4 Experiment Description and Results

We use continuous digit recognition task for this experiment. Data is obtained from Numbers95 v1.3 database [10]. We decided to use this database because (a) there is a reasonable large amount of data available, which is an important requirement when dealing with non-parametric techniques and (b) there is no grammar in the utterance structure, hence the recognition task becomes simpler and we can concentrate on the pattern matching issue.

We use 1000 templates for each class word and we generate N-best list from a state-of-the-art HMM system. In this work, we use 10 hypotheses from each test utterance ($N = 10$). As explained in Section 2, the K lowest DTW distortions are chosen, we choose K equal to 10. For the test set, 2824 utterances are analyzed.

The HMM-based system uses context-dependent models with 39-dimensional feature vectors (MFCC[11], delta and delta-delta). Emission probabilities are modeled with a mixture of 10 Gaussians. The whole system was trained and tested with Torch software package[2].

Pitch information is obtained using SIFT algorithm [12] to extract pitch contour, followed by median smoothing. For the unvoiced regions, the pitch values are zero. We compute the average pitch value over the voiced regions. For the use of pitch information, templates are split into two classes given a pitch value threshold.

Unlike our previous work [4] where we use numbers, we decided to use only digits for this experiments to have an equal number of templates for each word in our vocabulary. Templates consists of sequences of 26-dimensional feature vectors (MFCC and delta) obtained from a forced alignment Viterbi decoding on the training set. Delta-delta features are not used because they use too long a context for dealing with templates.

Score S given by Equation 2 is computed for each hypothesis and the one with the lowest score is chosen as correct. This score can be obtained only from DTW distances or combined with the log-likelihood obtained by the HMM system. We distinguish between two different types of experiments:

- In this case, only DTW distances are used for re-scoring the hypotheses. No prior information about pitch is used. Score S can also be combined in a weighted sum with the log-likelihood obtained from the HMM system. Equation 2 then becomes

$$S_n = \sum_{i=1}^{M_n} (w \cdot D_{w_i^n}(X_i^n) - (1-w) \cdot L_{w_i^n}(X_i^n)) \tag{3}$$

  where $L_w(X)$ denotes the log-likelihood of the HMM corresponding to the word w for the sequence of feature vectors X. Actually, log-likelihood is subtracted because we are interested in the hypothesis with the lowest score. The weight w is tuned manually.

- This experiment is similar to the previous one, but pitch information is also used. Templates are classified in two classes: low and high pitch frequency (150 Hz is chosen as threshold). For each test utterance, pith information is also extracted in addition to the acoustic features. Then, words in the hypotheses are compared to those templates which share the same pitch class. To keep the same number of templates as the previous experiment, 1000 templates are chosen for each pitch class.

Given the nature of the experiment, where the correct sentence must be chosen among a set of hypotheses, we believe that it makes sense to give the sentence error rate (SER). Also, it is important to know what the best possible result is, since the correct sentence is not always included among the first N hypotheses. Table 1 shows the baseline given from the state-of-the-art HMM system and the best possible result with 10 hypotheses. In our case, the correct transcription does not appear among the best 10 hypothesis 5.7% of times.

	Baseline SER (WER)	Best Result SER (WER)
HMM	16.1 (4.1)	5.7 (1.4)

Table 1: SER (and WER between parenthesis) of the state-of-the-art HMM system and the best possible result using 10 hypotheses for each utterance.

Results of the experiments are shown in Table 2. It can be seen in the results of the first column (No Pitch) that the use of templates improves the system accuracy significantly when compared to HMM baseline (12.4% relative improvement). This improvement is even higher when combined with the likelihood obtained by HMM system (19.9% relative improvement). Similar to our previous study [4], we observe that the use of templates improves the system accuracy.

[2]More information about this package is available at `www.torch.ch`

On the second column (Pitch), we can see the results with templates classified by pitch. Improvement over the non-classified templates can be noted in both situations: using only DTW distortion (6.5% relative improvement) and combining with likelihood (5.5% relative improvement). These results confirm the idea that similarity between templates and test utterances can be better achieved by using information which is not present in acoustic features.

	No Pitch SER (WER)	Pitch SER (WER)
DTW	14.1 (4.5)	13.2 (3.3)
DTW & Likelihood	12.9 (3.3)	12.2 (3.1)

Table 2: SER (and WER between parenthesis) of the experiments

Moreover, the use of prior information improves system accuracy when compared to an experiment with no prior information but a larger number of templates. In particular we have tried 2000 and 2500 templates. As results in Table 3 show, error rate is still higher than the experiment with the use of pitch information but using only 1000 templates. Even if the difference in relative errors is not significant in this case, the computational time is. The system with 2000 templates takes twice the time for recognition.

	2000 templates SER (WER)	2500 templates SER (WER)
DTW	13.9 (3.5)	13.7 (3.4)
DTW & Likelihood	12.5 (3.2)	12.5 (3.2)

Table 3: SER (and WER between parenthesis) of the experiments without pitch information but using more templates

We have also investigated the case where pitch information is inferred automatically by the system. The test utterance is compared to both types of templates independently. Two possible transcriptions are, then, obtained. Each of them corresponds to a pitch category. The one with the lowest score S is chosen as the correct transcription to the test utterance. In this experiment, pitch class is not extracted previously to the decoding phase but it is estimated. Similar results to those obtained in Table 2 were obtained. This experiment can be useful in the case of noisy conditions, where pitch information cannot be estimated properly.

5 Conclusion and Future Work

In this work, we have investigated the use of meta-linguistic information as a prior knowledge in a template-based approach for speech recognition. Experiments with pitch frequency have shown that extra information which is not included in the acoustic features can help in the selection of the most appropriate templates. Pitch frequency has been used to cluster templates database and test utterances are compared only with those templates that share the same pitch category. This experiment results in a 24% relative improvement over the HMM baseline.

Interestingly, the use of prior knowledge gives better results than the blind method even when the latter experiment is carried out with more template. In particular 2500 templates per class cannot yield better results than 1000 templates with previous pitch classification. The error difference of both the systems is not significant but there is large saving in computational time.

This work is a first attempt towards the use of prior knowledge with templates. Future work should be oriented to other types of meta-linguistic information such as duration, speech rate or position in the sentence.

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)". The authors also would like to thank Mathew Magimai Doss for useful discussions during this work.

References

[1] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, vol. 247. Prentice Hall, 1993.

[2] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, "From HMM's to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 360–378, 1996.

[3] K. Tokuda, H. Zen, and T. Kitamura, "Trajectory Modeling based on HMMs with the Explicit Relationship between Static and Dynamic Features," *Proceedings of Eurospeech*, pp. 865–868, 2003.

[4] G. Aradilla, J. Vepa, and H. Bourlard, "Improving Speech Recognition Using a Data-Driven Approach," *Proceedings of Interspeech*, 2005.

[5] I. Illina and Y. Gong, "Elimination of Trajectory Folding Phenomenon: HMM, Trajectory Mixture HMM and Mixture Stochastic Trajectory Model," *ICASSP*, pp. 1395–1398, 1998.

[6] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Transactions on Information Theory*, vol. 13, pp. 260–269, 1967.

[7] S. D. Goldinger, "Words and Voices: Episodic Traces in Spoken Word Identification and Recognition Memory," *Journal of Experimental Psychology: Learning Memory and Cognition*, pp. 1166–1183, 1996.

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

[9] C. Chesta, P. Laface, and F. Ravera, "Connected Digit Recognition Using Short and Long Duration Models," *Proceedings of ICASSP*, pp. 557–560, 1999.

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

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

[12] J. Markel, "The SIFT algorithm for Fundamental Frequency Estimation," *IEEE Transactions on Audio and Electroacoustics*, pp. 367–377, 1972.