Zero-resource audio-only spoken term detection based on a combination of template matching techniques

Armando Muscariello, Guillaume Gravier, Frédéric Bimbot

Zero-resource audio-only spoken term detection based on a combination of template matching techniques

Armando Muscariello[*]*, Guillaume Gravier, Frédéric Bimbot*

IRISA CNRS UMR 6074 & INRIA Rennes Bretagne Atlantique

{amuscari, ggravier, bimbot}@irisa.fr

Abstract

Spoken term detection is a well-known information retrieval task that seeks to extract contentful information from audio by locating occurrences of known query words of interest. This paper describes a zero-resource approach to such task based on pattern matching of spoken term queries at the acoustic level. The template matching module comprises the cascade of a segmental variant of dynamic time warping and a self-similarity matrix comparison to further improve robustness to speech variability. This solution notably differs from more traditional *train and test* methods that, while shown to be very accurate, rely upon the availability of large amounts of linguistic resources. We evaluate our framework on different parameterizations of the speech templates: raw MFCC features and Gaussian posteriorgrams, French and English phonetic posteriorgrams output by two different state of the art phoneme recognizers.

Index Terms: spoken term detection, template matching, unsupervised learning, posterior features

1. Introduction

A popular strategy for mining useful information from speech data sets consists in localizing occurrences of known words of interest within the data set, a task known as spoken term detection or keyword spotting. In an era dominated by growing rates of digital media creation and diffusion, keyword spotting plays a prominent role as an indexing tool operating on the audio part of a multimedia database. State of the art approaches rely on automatic speech recognition (ASR) paradigms, whose founding principles are deeply rooted into statistical methods for training complex acoustic and language models. These techniques often require large amounts of annotated speech data, even language and topic specific, and long off-line procedures for performing reliable training estimates. On the one hand, the increase in computational power and the availability of quality training data have largely justified these supervised approaches, that have indeed proven to be very effective and accurate [7]. On the other hand, the drawbacks limiting their attractiveness are also notable. First, performance of supervised systems are all tightly related to the quality and quantity of training data. And the annotation of such data often implies the direct intervention of human experts, a process that is expensive, time consuming and error prone. If the ultimate goal is to automatize the archiving and management of large data sets, conceiving solutions demanding in turn high levels of supervision might seem counterintuitive. Moreover, it is well known that for many languages, indeed for most of the about 7,000 languages and dialects spoken in the world, rich and valuable linguistic resources are scarce or absent.

In the recent years, a few solutions have been proposed that limit as much as possible resorting to prior knowledge in the form of acoustic and linguistic resources. In [2], a keyword spotting framework is proposed, that relies on dynamic time warping (DTW) distances among speech templates obtained by modelling speech through a Gaussian mixture model (GMM), trained without supervision. The same task is carried out in a very similar way in [3], although speech is modeled by posterior features produced by an independently trained phoneme recognizer.

In this paper, we describe one such *zero resource* approach, where a query sample of a given keyword is directly searched on the acoustic data set by template matching techniques, without any prior knowledge, or preliminary, off-line training of model parameters. We claim two main contributions of the present work: first, the computational architecture, that combines the scores of a segmental variant of DTW and self-similarity matrix (SSM) comparison between speech sequences, previously applied in word discovery tasks [1]. Second, a thorough evaluation of the performance on a 4h subset of the French ESTER corpus with respect to different types of speech features: mel-frequency cepstral coefficients (MFCCs), Gaussian posterior features trained without any supervision [2], English HMM-based phone posterior features, and French phonetic posteriorgrams. We show how the additional use of the SSM technique benefits performance for all the feature types with respect to the DTW-based system alone, except for the French posteriorgram case. For this last case, the high level of phonetic knowledge incorporated in the training procedure, performed on the same ESTER corpus, successfully captures the speech variability of the database, making SSM comparison superfluous, if not detrimental to performance. While phonetic posteriorgrams require prior linguistic knowledge to train acoustic models, we justify their use to evaluate the system against the effectively unsupervised (zero-resource) approach implied by the use of MFCCs or Gaussian posteriorgrams.

The remainder of this paper is structured as follows: in Section 2 the architecture of the system is described, by presenting the pattern matching techniques employed and the speech features examined, while in Section 3 the experimental evaluation is thoroughly detailed.

2. The keyword spotting architecture

Template-based spoken term detection relies on two crucial components: the feature extraction and the pattern matching step used to compare speech segments. We tackle these two complementary aspects in this section, by first describing the features and the metric used to measure the pairwise distance of

[*]The present work was partially funded by the Quaero project: http://www.quaero.org/

feature vectors.

2.1. Feature extraction

Two main type of features are evaluated in our experiments, namely a) posteriorgrams [6] and b) MFCCs.

Posteriorgrams are a time-vs.-*class* matrix representing the posterior probability of each class for each feature frame (for example, an MFCC frame). The type of class defines the type of posteriorgram. Here we evaluate the following posteriorgrams:

1. Gaussian posteriorgram: each class represents a component of a Gaussian mixture model (GMM) trained in an unsupervised way from a training data set, or, possibly, from the test data itself.
2. Hidden Markov model (HMM) state posteriorgram: each class represents the state of an HMM modelling a language specific phone.

Each posterior vector is a probability vector, *i.e.*, its entries sum up to 1. The probability of two vectors p and q drawing from the same distribution is defined by their dot product $p \cdot q$, that can be assumed as a measure of closeness. To map these scores into the distance-like range $[0, \infty[$, the distance between two posterior vectors is computed as $-log(p \cdot q)$. To avoid zeros while computing the log-probability, each vector is smoothed by zeroing all entries p_i below a threshold $P_{\min}$, then distributing a small portion of the mass probability from the non-zero entries to the zeroed ones (see [2]).

To score the dissimilarity of MFCC vectors, instead, we make use of the Euclidean distance.

Given account of the parameterization of the signal, we are ready to describe the pattern matching techniques employed.

2.2. Template matching techniques

The search for a match of the given query within the test utterance, raises two issues: the primary issue consists in locating the subsegment of the utterance most likely to represent a repetition of the query. The second problem amounts to providing a quantitative measure (*i.e.*, a score) of the (dis)similarity between the query and the extracted segment.

In the system proposed, the first duty is assigned to a segmental variant of DTW, called segmental locally normalized DTW (SLNDTW) [4]. This technique also provides a score that can be used to directly decide on the similarity of the templates, or that can be further combined with additional measures that might better model speech variability and improve scoring.

2.2.1. DTW-based matching

Given the query $u_{i=1}^{M}$ and the test utterance $v_{j=1}^{N}$, the goal is to detect the segment $m = v_{j_s}^{j_e}$ most similar to u. SLNDTW solves the task by a three-stage procedure operating on the distance matrix between u and v, which is a structure gathering the dissimilarity score d between any pair of feature vectors from u and v. The three stages are respectively:

1. a starting point selection heuristic, responsible of selecting the starting points of likely matching subsegments along the first row of the distance matrix.
2. The computation of the alignment paths from the selected starting points, according to recursive dynamic programming relations.
3. The selection of the best path $P^* = \{(1, j_s), \cdots, (M, j_e)\}$ as the one minimizing $W(M, j_e), 1 \leq j_e \leq N$, with W representing the average distance over the $L(P^*)$ entries absorbed in the path P^* (see [4] for more details).

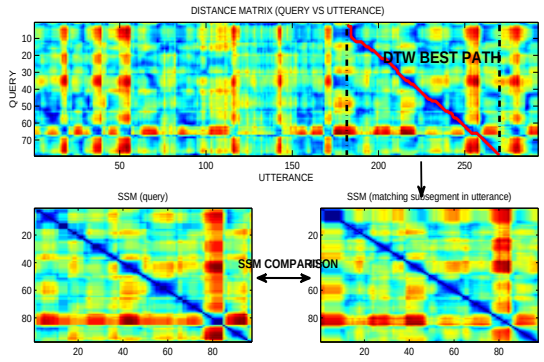

Figure 1: Example of combined use of DTW and SSM-based comparisons for similarity scoring of templates.

The DTW dissimilarity score D_{DTW} between u and v is defined as $D_{\text{DTW}} = W(M, j_e)$. This same score can either be assumed as the final score $S(u, v)$ to deem the similarity of u and v, or be further refined by the SSM-based scores, to better account for speech variability.

2.2.2. SSM-based matching

The SSM of a sequence u is the square symmetric matrix $\Phi(u)$ defined as $\Phi\,(i, j) = d\,(u(i), u(j))$.

The structure of these matrices is strictly dependent on the acoustic-phonetic content of the underlying sequences, and unlike DTW, captures the interaction between all parts of the compared patterns, encoding a richer information on their (dis)similarity, hence accounting for more variability. Indeed, many empirical observations have consistently confirmed the visual resemblance of these matrices across different conditions, that is when instances of a same word are uttered by different speakers, or undergo different channels or are imposed on a noisy background. The main argument advising against the use of SSM comparison is that words that are different at the lexical level might indeed exhibit similar SSMs. As a countermeasure, in our experiments the SSM score is combined with the DTW one, to balance the effect of the SSM score with the more reliable DTW one. The cascade of these two pattern matching methods has been proven to better deal with speech variability in word discovery tasks [1]. We describe in the following how to effectively compare the SSMs.

SSM scores. To ease the comparison of the SSMs, the templates u and m are warped into u' and m' according to the path length $L(P^*)$, to obtain SSMs of equal size. Then two different dissimilarity scores are computed:

1. $D'_{\text{SSM}} = \|\Phi(u') - \Phi(m')\|_1 / L(P^*)^2$
2. $D''_{\text{SSM}} = \|V(\Phi(u')) - V(\Phi(m'))\|_1 / L(V)^2$

where V denotes a type of *visual descriptor* of length $L(V)$ extracted from each SSM. The score D'_{SSM} represents the simple entry-by-entry distance between the SSMs, but this measure is dependent on the absolute values of the SSMs and does not encode well their implicit spatial pattern. That is why the visual descriptors V, based on local histograms of oriented gradients [5], are computed, assuming that local objects' appearances and

Table 1: The 4h dataset keyword collection with the occurrence counts.

mobilisation (19)	Jean Marie Le Pen (35)	final (25)
vingt-et-un avril (47)	Lionel Jospin (13)	syndicat (48)
Saddam Hussein (36)	important (14)	Iraq (127)
personne (45)	gauche (48)	general (29)
week end (48)	president (77)	droite (26)
politique (67)	journal (35)	france (284)
responsable (21)	americain (58)	

shapes can be well characterized by the distribution of local intensity gradients and edge orientations. The score D''_{SSM} does not depend on entries (or *pixels*) magnitudes but rather on local gradients' magnitudes, as they measure the strength and directions of local edges. Moreover, they do not provide just a punctual information, *i.e.*, confined to each single pixel, but are computed over dense, overlapping grids of pixels, encoding a more complex information on the self-similarity visual patterns.

Scores fusion. The different scores are finally combined into the global score $S(u, v)$ according to the following expression:

$$S(u,v) = \alpha_{\text{DTW}} \cdot \frac{D_{\text{DTW}}}{th_{\text{DTW}}} + \alpha'_{\text{SSM}} \cdot \frac{D'_{\text{SSM}}}{th'_{\text{SSM}}} + \alpha''_{\text{SSM}} \cdot \frac{D''_{\text{SSM}}}{th''_{\text{SSM}}} \quad (1)$$

That means each score is normalized by a proper threshold value for each pattern matching technique, and weighted by a factor so that $\alpha_{\text{DTW}} + \alpha'_{\text{SSM}} + \alpha''_{\text{SSM}} = 1$. A schematic representation of the combination of the pattern matching techniques is provided by Fig. 1.

3. Experimental results

Experiments were performed on a 4h subset of the ESTER corpus [9], comprising four different French broadcast news shows, sampled at 16 KHz. The shows were recorded on the same days, at different channels, to build a data set including several speakers, likely covering similar topics, thus inducing frequent repetitions. The file has been segmented into 2,915 utterances separated by silences. From the same data set, 20 keywords with at least 10 occurrences, listed in Table 1, were considered and a unique sample for each was randomly selected as the actual query (and obviously disregarded during performance evaluation). The average number of speakers per keyword is 22.55.

3.1. Evaluation criteria

The scores between each acoustic query and each test utterance are computed according to Eq. 1 and utterances are ranked accordingly in ascending order. The various threshold have been tuned, for each type of feature, on a separated set of word occurrences. The weights have been varied during different runs of the system and the respective performance are to be described in subsection 3.3. We opted for three sets of weight distribution: a) $\alpha_{\text{DTW}} = 0.34, \alpha'_{\text{SSM}} = 0.33, \alpha''_{\text{SSM}} = 0.33$, b) $\alpha_{\text{DTW}} = 0.4, \alpha'_{\text{SSM}} = 0.2, \alpha''_{\text{SSM}} = 0.4$ and c) $\alpha_{\text{DTW}} = 0.5, \alpha'_{\text{SSM}} = 0.2, \alpha''_{\text{SSM}} = 0.3$. The logic behind this choice was to check the behaviour a) with all the scores equally weighted, and b) and c) with the DTW score privileged over the others because of the more reliable similarity information provided, and with D''_{SSM} preferred over D'_{SSM}, as showing better results in previous word discovery experiments.

The evaluation is conducted on a per-utterance basis, *i.e.*, an utterance is deemed a correct hit if containing the desired search term, regardless of its location. Moreover, a match is considered correct even if the acoustic query appears as the subword of a longer word at the lexical level.

As for the performance indicators, we have resorted to 1) the average P@10, that is the precision over the first 10 ranked utterances, 2) the average P@N, the precision over the best N utterances (N being the occurrence count for each keyword), 3) the average MAP, the mean average precision, the mean of precision scores after each keyword occurrence is retrieved and 4) the average EER, the equal error rate, where the false acceptance and false rejection rates are equivalent.

3.2. Details on the features

All the features are extracted at a 100 Hz frame rate. The MFCCs are 39 dimensional vectors, comprising 12 MFCCs, the log energy coefficient and their first and second order derivatives, refined by mean subtraction and variance normalization. As for the posteriorgrams:

1. the Gaussian posterior features are 50 dimensional vectors obtained by a GMM with 50 components trained on the same MFCCs extracted from the data set. The number of components was decided upon the evaluation conducted in [2].
2. The French state posteriorgram representation comprises 115 dimensional vectors corresponding to 38 phones, each modelled by a three-state HMM with 16 component GMM emission distributions, independently trained on the 150h ESTER corpus. Each posterior frame has been computed from the same MFCCs used for the experiments.
3. The English state posteriorgrams are output by the BUT phoneme recognizer [8], with 120 classes corresponding to the states of 40 three-state tandem HMM English phone models, discriminatively trained on the TIMIT database. In this case, the posteriorgrams are computed over TRAP-based features extracted from the signal.

3.3. Results and discussion

The results of the experiments are concisely summarized in Table 2, respectively accounting for the DTW-based system ($\alpha_{\text{DTW}} = 1$) and the SSM-driven one, for the three different weight distributions examined.

According to the evaluation framework described, the English phonetic posteriorgram is outperformed by all the other ones, in any configuration setting. For this type of feature, P@10, P@N and MAP barely fall above the 50% and 30% respectively in the best cases, figures significantly improved by all the other feature types in any setting scenario. The EER, instead, is comparable with the MFCC case, and, in the DTW-only system, significantly better (18.9% against 21.6%). Such disappointing results might look surprising at first, as this representation of the speech data relies on probability estimates output by a state of the art phoneme recognizer, based on long temporal context features and neural network classifiers, proven to successfully capture the speech variability over a data set by the training procedure. We suspect that the likely reason for this relatively poor performance is the application of English specific phone models and parameters to a French data set. This interpretation seems confirmed by the impressive numbers (about 90% for P@10, about 65-75% for P@N and MAP, 6-7% for the EER) yielded by the posterior features output by the French GMM-HMM recognizer, which not only is based on French

specific phone models, but has also been trained on the same ESTER corpus including our test data[1]. While the English phonetic posteriorgram still produces valuable results, it does not even perform as well as the raw MFCC features (from 55% to 67% depending on score combination for P@10, 30-40% for P@N and MAP), except for slightly better EERs (around 20%), or as well as the Gaussian posteriorgrams (about 70% for P@10, 40-45% for P@N and MAP, 15% for EER).

One very interesting conclusion that can be drawn is the beneficial effect induced by the use of the SSM comparison stage, for all features but for the French phone-state posteriorgram. The aim of using this technique in addition to the DTW-based one is to enhance the robustness of the system to speech variability. Its capability to better account for variability is demonstrated by an improvement of about 12% for P@10, of 6-9% (depending on score weighting) for P@N and MAP in the MFCC case (while failing to substantially impact the EER), and a minor one for the Gaussian and English phonetic posteriorgram case (in these two cases, it is mostly the P@10 that is improved). The addition of the SSM scores appears, instead, slightly detrimental to performance in the French posteriorgram case (even if, for at least one weight distribution, this worsening might be negligible, amounting to about 2% for P@10, P@N and MAP, with the EER unchanged). The reason for this behaviour resides evidently in the modelling capability of the posteriorgram representation; the SSM matching technique does not specifically account for a particular type of speech variability, for example the inter-speaker one, while phonetic posteriorgrams are speaker-independent by construction. While this is also true for the Gaussian and English phone posteriorgrams, the far more precise prior knowledge available for the French recognizer during training clearly improves results. This underlines how SSM might be of significant value in many situations where language-specific phonetic knowledge is absent (indeed the zero-resource scenario depicted while using MFCCs or Gaussian posteriorgrams, or even when language-specific phone recognizers are available, even if not matching the target language).

Finally, the analysis of the results for the three different distributions of weights shows that the uniform weight setting is the least performing. This is an expected outcome since, like previously remarked, the DTW score is indeed the most reliable and should be given more relevance in the global score computation. As far as the alternative weights is concerned, a revealing trend is observed that makes the distribution $\alpha_{\text{DTW}} = 0.50, \alpha'_{\text{SSM}} = 0.20, \alpha''_{\text{SSM}} = 0.30$ the more convenient over the choice $\alpha_{\text{DTW}} = 0.40, \alpha'_{\text{SSM}} = 0.20, \alpha''_{\text{SSM}} = 0.40$ (except for the Gaussian posteriorgram case and the P@10 measure).

4. Conclusion and future work

In this paper a computational framework has been described for audio-only keyword spotting by combining template matching techniques, based on DTW and SSM comparison. The system is easily applicable and well performing in a zero-resource approach, where no prior linguistic or acoustic knowledge is needed to model variability. For evaluation purposes, a comparison of the performance has been carried out against more supervised frameworks based on language-specific features. The evaluation has shown that a) the benefit of SSM comparison in a zero-resource approach and b) the crucial impact of language-specific posteriorgrams on keyword spotting tasks based on template matching.

In the near future we plan to extend this evaluation to word discovery experiments, and investigate the use of the described pattern matching techniques to alternative signal processing tasks.

[1] the system, though, can hardly be considered *biased* towards the 4h file, indeed a very small sample of the 150h ESTER corpus

Table 2: Results for the DTW-only and the SSM-based system, for different weight distributions: a) $\alpha_{\text{DTW}} = 0.34, \alpha'_{\text{SSM}} = 0.33, \alpha''_{\text{SSM}} = 0.33$, b) $\alpha_{\text{DTW}} = 0.40, \alpha'_{\text{SSM}} = 0.20, \alpha''_{\text{SSM}} = 0.40$ and c) $\alpha_{\text{DTW}} = 0.50, \alpha'_{\text{SSM}} = 0.20, \alpha''_{\text{SSM}} = 0.30$.

DTW	MFCC(%)	Gauss(%)	ENG(%)	FRA(%)
P@10	55.5	67.5	48.5	90
P@N	30.8	46.6	30.8	70.5
MAP	29.2	47.9	29.2	74.8
EER	21.6	15.1	18.5	6.4
SSM a)	MFCC(%)	Gauss(%)	ENG(%)	FRA(%)
P@10	64	67.5	50	86
P@N	36.1	46.6	29.7	63.9
MAP	34.6	38.0	27.9	66.9
EER	22.9	16.1	21.8	7.8
SSM b)	MFCC(%)	Gauss(%)	ENG(%)	FRA(%)
P@10	67.5	68.5	53.3	87
P@N	38.5	44.1	32.6	67.3
MAP	37.1	45.6	30.8	71.1
EER	20.9	15.1	21.0	6.5
SSM c)	MFCC(%)	Gauss(%)	ENG(%)	FRA(%)
P@10	67.5	70	51.1	88
P@N	39.8	45.8	33.1	68.5
MAP	38.0	47.0	31.6	72.5
EER	20.8	15.0	19.9	6.5

5. References

[1] Muscariello A., Gravier G. and Bimbot F., "Towards robust word discovery by self-similarity matrix comparison", IEEE ICASSP, 2011.

[2] Zhang Y. and Glass J.R., "Unsupervised spoken keyword spotting via Segmental DTW on Gaussian Posteriorgrams", IEEE ASRU, 2009.

[3] Hazen T.J., Shen W. and White C., "Query-by-example spoken term detection using phonetic posteriorgram templates" IEEE ASRU, 421–426, 2009.

[4] Muscariello A., Gravier G. and Bimbot F., "Audio keyword extraction by unsupervised word discovery", in Interspeech, 2843-2846, 2009.

[5] Dalal N. and Triggs B., "Histograms of oriented gradients for human detection", in IEEE CVPR, 886–893, 2005.

[6] Aradilla G., Bourlard H. and Magimai-Doss M., "Posterior features applied to speech recognition tasks with user-defined vocabulary" IEEE ICASSP, 3809–3812, 2009.

[7] Miller D., *et al*, "Rapid and accurate spoken term detection", in Interspeech, 2007.

[8] Schwarz P., Matějka P. and Černocký J., "Towards lower error rates in phoneme recognition", Proc. Int. Conf. on Text, Speech and Dialogue, 2004.

[9] Galliano S., *et al*, "The ESTER evaluation campaign for the rich transcription of French broadcast news" Interspeech, 2005.