Irisa MediaEval 2011 Spoken Web Search System

Armando Muscariello, Guillaume Gravier

Irisa MediaEval 2011 Spoken Web Search System

Armando Muscariello
Irisa/Inria
Rennes, France
amuscari@irisa.fr

Guillaume Gravier
Irisa/CNRS
Rennes, France
ggravier@irisa.fr

ABSTRACT

These working notes describe the main aspects of IRISA submission for the Spoken Web Search at the MediaEval 2011 campaign. We test a language-independent audio-only system based on a combination of template matching techniques. A brief overview of the main components of the architecture is followed by reporting on the evaluation on the development and test data provided by the organizers.

Categories and Subject Descriptors

H.3.3 [**Information Systems Applications**]: Spoken Term Detection—*zero-resource speech processing, template matching, posteriorgrams*

1. MOTIVATION

In [1] we have recently proposed a zero-resource audio-only system for spoken term detection (STD), *i.e* a system for performing keyword spotting at the acoustic level, in the absence of any language or domain-specific knowledge, training speech data and models. Main motivation behind our partecipation at the campaign is the opportunity of benchmarking the system on a different, more challenging data set [2], and learn of alternative solutions and respective performance.

2. ARCHITECTURE OF THE SYSTEM

The STD computational system relies on two main components: the acoustic features that represent queries and utterances, and the pattern matching techniques that identify occurrences of the queries within the utterances and provide the respective measure of (dis)-similarity.

2.1 Acoustic features

We have experimented different type of speech parametrizations, namely MFCC features and several type of posteriorgrams, that is 1) posteriors estimated from a Gaussian mixture model (GMM) trained in an unsupervised fashion on the same development data provided [2], and posteriors output by a language-specific (Czech, Hungary, Russian) BUT phoneme recognizer [3], independently trained on (Czech, Hungary, Russian) 8 KHz telephonic data.

MediaEval 2010 Workshop, September 1-2, 2011, Pisa, Italy

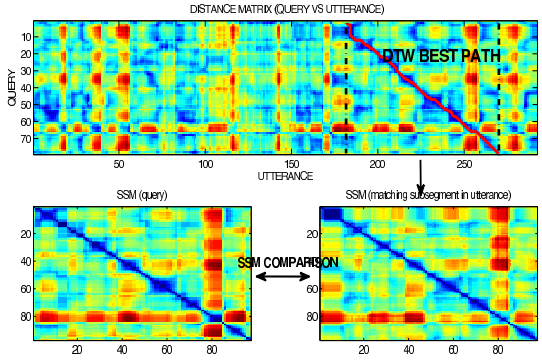

Figure 1: Example of combined use of DTW and SSM-based comparisons for similarity scoring of templates.

We have used the Euclidean distance to computed the pairwise distance between feature frames, and $-log(p \cdot q)$ as a distance-like measure of closeness between two posterior vectors p and q.

2.2 Pattern matching combination

The search for an occurrence of the query within the utterance is performed directly on the feature sequences by a cascade of two different pattern matching techniques. A segmental variant of DTW, named segmental locally-normalized dynamic time warping (SLNDTW) is responsible of selecting the subsegment of the utterance most similar to the searched query, according to a DTW score D_{DTW}. This score can directly be used to decide upon the similarity of the two segments, or refined by the use of additional scores. In our system, the two candidate keyword occurences are further subjected to the comparison of the respective self-similarity matrices (SSMs), and the two SSM scores, D'_{SSM} and D''_{SSM}, resulting from such comparison are then combined with D_{DTW} to obtain a unique dissimilarity score S (see figure 1).

The global score S is computed as:

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

Table 1: Development queries on development utterances: results

DTW+SSM	MFCC	GMM	HU	CZ	RU
P(FA)	0	0.0003	0.02	0.02	0.016
P(Mis)	0.82	0.77	0.66	0.66	0.70
AWTV	0.18	-0.10	-19.9	-20.7	-15.9
MAP (%)	0.26	6.61	1.10	0.82	0.72

so that $S < 1$ implies the detection of a match.

3. SYSTEM TUNING

The data set described in [2] is particulary challenging for such a system, because it is 8 KHz telephonic quality, presents portion of silences in the queries and a large pronunciation variability due to non native English speakers. We have preliminarily removed silences from the queries thanks to a speech detector, both for the development end evaluation queries. The thresholds $th_{\text{DTW}}, th'_{\text{SSM}}, th''_{\text{SSM}}$ have been tuned on word samples from a different data set (see [1]) and the pattern matching weights have been set to $\alpha_{\text{DTW}} = 0.50, \alpha'_{\text{SSM}} = 0.20, \alpha''_{\text{SSM}} = 0.30$ following [1]. Despite the availability of the ground truth for the development data set, reliable tuning of the thresholds on this data has not been successful, as many true hits exhibit a dissimilarity score higher than false alarms. This highlights the poor discriminative properties of the employed features in this task. The results for the different features are shown on table 1, for the system jointly employing the DTW and SSM-based comparisons, and the metrics: P(FA), that is the average false alarm rate, P(Mis), the average false rejection rate, the average weighted term value AWTV (the primary performance indicator), and the mean average precision MAP. The posterior features estimated by the BUT recognizer are the least performing according to the AWTV, as their P(FA), weighted by a factor $\beta = 1000$, is greater by order of magnitudes than the P(FA) for the MFCC and GMM features. Gaussian posteriorgrams yield the highest MAP value among the features tested, although very disappointing if compared to the values reported by this same system and features in the evaluation conducted in [1]. While yielding the highest miss detection rate P(Mis), the raw MFCC features report the best AWTV, as no false alarm has been collected. According to this metric, the MFCC-based system has been selected as the primary one.

4. RESULTS ON EVALUATION DATA

The results of the evaluation of the system on the test data are summarized by table 2, as for the primary runs and table 3, as for the secondary runs, where Gaussian posteriorgrams have been used. Not suprisingly, the figures reflect substantially the poor results of the experiments on the development data set. The system operates in a completely unsupervised fashion and the knowledge of the performance on the development data are not exploited in any way, and therefore do not bear any impact on the result. Indeed, the only parameteres needed to be tuned were estimated on a different data set.

Table 2: Evaluation runs: primary system

DTW+SSM	DEV-EVAL	EVAL-EVAL	EVAL-DEV
P(FA)	0.0003	0.00007	0.00006
P(Mis)	0.999	0.831	0.962
AWTV	-0.29	0.10	-0.022

Table 3: Evaluation runs: secondary system

DTW+SSM	DEV-EVAL	EVAL-EVAL	EVAL-DEV
P(FA)	0.00019	0.00013	0.00017
P(Mis)	0.97	0.788	0.97
AWTV	-0.17	-0.10	-0.14

It is worth noting that searching for the evaluation queries on the evaluation utterances perform better than conducting a cross-dataset spoken term detection, which is likely due to the limited variability among patterns extracted from the same set.

5. CONCLUSION

The IRISA architecture for spoken term detection, presented in [1], was evaluated on the data set provided by the MediaEval 2011 Spoken Web Search. This dataset has proven extremely challenging for the system in its current form, yielding poor results for all type of acoustic features employed. For this particular data set, given the presence of many English keywords, training a phone recognizer based on English phone models would have likely improved performance, although our team did not dispose of such training data (indeed one of the reasons why pursuing research on zero-resource systems would benefit the community). One possible idea is to combine posteriors from different recognizers to increase robustness to multiple languages, although in this specific case the results for Hungarian, Czech and Russian-based posteriorgrams were bad enough to prevent any satisfying application of this solution. Also, the Gaussian posteriors were only estimated from models trained on the development utterances; performance could have been, at least slightly, improved by training the GMM on the combined development-evaluation data set, in particular for the cross-data detection that yielded the poorest results.

6. REFERENCES

[1] A. Muscariello, G. Gravier, and F. Bimbot. Zero-resource audio-only spoken term detection based on a combination of template matching techniques. In *Interspeech*, 2011.

[2] N. Rajput and F. Metze. Spoken web search. In *MediaEval Workshop*, 2011.

[3] P. Schwartz, P. M. P., and J. Černocký. Towards lower error rates in phoneme recognition. In *International Conference on Text, Speech and Dialogue, 2004.*, 2004.