The L2F Spoken Web Search system for Mediaeval 2013

Alberto Abad
INESC-ID Lisboa / Instituto Superior Técnico
alberto@l2f.inesc-id.pt

Ramón F. Astudillo
INESC-ID Lisboa
ramon@l2f.inesc-id.pt

Isabel Trancoso
INESC-ID Lisboa/ Instituto Superior Técnico
imt@l2f.inesc-id.pt

ABSTRACT

The INESC-ID's Spoken Language Systems Laboratory (L^{2}F) primary system developed for the Spoken Web Search task of the Mediaeval 2013 evaluation campaign consists of the fusion of six individual sub-systems exploiting 3 different language-dependent phonetic classifiers. For each phonetic classifier, an acoustic keyword spotting (AKWS) sub-system based on connectionist speech recognition and a dynamic time warping (DTW) based sub-system have been developed. The diversity in terms of phonetic classifiers and methods, together with the efficient fusion and calibration approach applied for heterogeneous sub-systems, are the key elements of the L^{2}F submission. Besides the primary submission, two additional systems based on the fusion of only the AKWS and the DTW sub-systems have been developed for comparison purposes. A final multi-site system formed by the fusion of the L2F and the GTTS primary submissions has been also submitted to explore the potential of the fusion approach for very heterogeneous systems.

1. INTRODUCTION

This document introduces the Spoken Web Search systems developed by the INESC-ID's Spoken Language Systems Laboratory (L^{2}F) for the Mediaeval 2013 campaign. The targeted task in this challenge is query-by-example spoken term detection. Detailed information about the task and the data used can be found in the evaluation plan [5]. One primary and three contrastive systems (one of them in collaboration with another participating group) have been submitted. The primary system consists of the fusion of six individual sub-systems. The proposed systems present three main novelties with respect to the systems developed for the previous year evaluation campaign [1]: 1) the number of language-dependent phonetic networks has been limited to three; 2) DTW-based sub-systems exploiting log-posterior features have been incorporated; and 3) a recently proposed method for discriminative calibration and fusion of heterogeneous spoken term detection systems [4] has been applied.

2. THE L^{2}F SWS SYSTEM DESCRIPTION

Six sub-systems form the core of the L^{2}F SWS system exploiting three different language-dependent phonetic networks trained for European Portuguese (*pt*), Brazilian Portuguese (*br*) and European Spanish (*es*). The phonetic networks are used either as acoustic models in acoustic KWS based on hybrid connectionist methods or as a feature extraction component for DTW based term detection.

2.1 Phonetic network classifiers

L^{2}F systems exploit multi-layer perceptron (MLP) networks that are part of our in-house hybrid connectionist ASR system. The phonetic class posterior probabilities are in fact the result of the combination of four MLP outputs trained with Perceptual Linear Prediction features (PLP, 13 static + first derivative), PLP with log-RelAtive SpecTrAl speech processing features (PLP-RASTA, 13 static + first derivative), Modulation SpectroGram features (MSG, 28 static) and Advanced Font-End from ETSI features (ETSI, 13 static + first and second derivatives). The language-dependent MLP networks were trained using different amounts of annotated data [2]. Each MLP network is characterized by the size of its input layer that depends on the particular parametrization and the frame context size (13 for PLP, PLP-RASTA and ETSI; 15 for MSG), the number of units of the two hidden layers (500), and the size of the output layer. In this case, only monophone units are modelled, resulting in MLP networks of 39 (38 phonemes + 1 silence) soft-max outputs in the case of *pt*, 40 for *br* (39 phonemes + 1 silence) and 30 for *es* (29 phonemes + 1 silence).

2.2 Acoustic KWS systems

AKWS sub-systems exploit the phonetic networks as acoustic models for both phonetic tokenization and query search based on hybrid ANN/HMM approaches for ASR [6]. The decoder used is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition. First, the phonetic transcription of each spoken query is obtained for every sub-system using a phone-loop grammar. Simple *1-best* phoneme chain output has been used. Then, search is carried out with a sliding window of 5 seconds (2.5 seconds time shift) using an equally-likely 1-gram language model formed by the target query and a competing speech background model. On the one hand, keyword/query models are described by the sequence of phonetic units obtained in the tokenization. On the other hand, the likelihood of a background speech unit representing "general speech" is estimated based on the other phonetic classes [3]. The output score for each candidate detection is computed as the average of the phonetic log-likelihood ratios that form the detected query term. More details can be found in [1].

2.3 Dynamic Time Warping systems

DTW sub-systems use the language-dependent phonetic networks to extract log-posterior features. The silence class

Mediaeval 2013 Workshop, October 18-19 2013, Barcelona, Spain

of the phonetic network is also used for voice activity detection. To this end, the segments identified as silence at the beginning and end of each query and document are removed. For each query-document pair, N euclidean distance based DTWs are run on N starting candidate positions of the document. To select the candidate positions, the query-document euclidean distance matrix of the DTW is used. The minimum of each column of the matrix represents the minimum distance among all query feature vectors to a given document feature vector. The average of these minima on a sliding window of query size is used as an approximation of DTW without the warping constraints, from which the best N candidates are selected. The number of candidates N was made equal to the length of the document in feature vectors divided by 100 with a minimum of 100 candidates. In a second stage, DTWs of the size of the query are evaluated at each one of the N candidate positions, and the three candidates with lower normalized cumulative distance, and separated by at least 0.5 seconds, are kept. The reduction of the search space to N candidates as explained above provided a reduction of the search time by a factor of around 5, while having a minimal impact on the performance. It should be noted that the DTW, including the distance matrix, was computed using the R programming language, while the candidate selection and remaining tasks were implemented in Python[1]. This framework benefited particularly from the candidate selection scheme proposed.

2.4 Discriminative calibration and fusion

The combination of systems is based on a recently proposed method for discriminative calibration/fusion of heterogeneous spoken term detection (STD) systems [4][2]. Under this approach, missing scores for systems that do not detect a given candidate are hypothesized based on heuristics. In this way, the original problem of several unaligned detection candidates is converted into a verification task. As for other verification tasks, system weights and offsets are then estimated through linear logistic regression. As a result, the combined scores are well calibrated, and the detection threshold is automatically given by application parameters (priors and costs). The method permits easy integration with majority voting schemes and it is convenient if scores from heterogeneous systems are in the same ranges (we apply a per-query zero-mean and unit-variance normalization *q-norm* [1]). Moreover, the maximum number of detection candidates for a certain query provided by any sub-system was limited to 200 before score normalization and fusion.

3. SUBMITTED SYSTEMS AND RESULTS

One primary and two contrastive "on-time" systems were submitted. The *primary* system consists of the fusion of the six sub-systems previously described, while the *contrastive1* and *contrastive2* submissions correspond to the fusion of only the DTW and only the AKWS sub-systems, respectively. Additionally, a "late" *contrastive3* system based on the fusion of the primary systems of the L^{2}F and GTTS[8] teams was also submitted. All the submitted systems are expected to generate well-calibrated log-likelihood ratios, such that the theoretical minimum expected cost Bayes threshold can be used ($\theta_{\text{Bayes}} = \log \beta$, see [4] for more details).

Table 1: L^{2}F SWS2013 performance scores

System	dev		eval	
	mtwv	atwv	mtwv	atwv
primary	0.3905	0.3883	0.3420	0.3376
contrastive1	0.3205	0.3071	0.2515	0.2364
contrastive2	0.2753	0.2743	0.2463	0.2459
contrastive3	0.4865	0.4850	0.4658	0.4639

Table 1 shows the actual and maximum TWV official scores obtained by the L^{2}F SWS systems for the two query sets: *dev* and *eval*. Notice that the theoretical Bayes threshold has been used in both *dev* and *eval* experiments. It is worth noticing the remarkable performance improvements when very heterogeneous (from different sites) systems are combined, like in the case of the *contrastive3* system. Regarding the amount of processing resources, we have used a cluster of machines with 90 nodes. The estimated cost figures [7] are pessimistic since the cluster was not exclusively used for the challenge. For each AKWS sub-system, the indexing speed factor (ISF), searching speed factor (SSF), maximum memory indexing (MMI) and maximum memory searching (MMS) values are 0.75, 77.33, 0.17 GBytes and 0.073 GBytes, respectively. For the DTW based sub-systems, the ISF, SSF, MMI and MMS are 0.17, 193.34, 0.18 GBytes and 0.43 GBytes, respectively. Considering these values, the total processing load (PL) is 239.76: 3 times the PL of AKWS (5.09) and DTW (74.83) sub-systems.

4. ACKNOWLEDGEMENTS

This work was partially funded by the DIRHA European project (FP7-ICT-2011-7-288121) and the Portuguese Foundation for Science and Technology (FCT), through the projects PEst-OE/EEI/LA0021/2013 and PTDC/EIA-CCO/122542/2010, and the grant number SFRH/BPD/68428/2010.

5. REFERENCES

[1] A. Abad and R. F. Astudillo. The L2F Spoken Web Search system for Mediaeval 2012. In *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012.

[2] A. Abad, J. Luque, and I. Trancoso. Parallel Transformation Network features for Speaker Recognition. In *ICASSP*, May 2011.

[3] A. Abad, A. Pompili, A. Costa, and I. Trancoso. Automatic word naming recognition for treatment and assessment of aphasia. In *Interspeech 2012*, Sep 2012.

[4] A. Abad, L. J. Rodriguez Fuentes, M. Penagarikano, A. Varona, M. Diez, and G. Bordel. On the Calibration and Fusion of Heterogeneous Spoken Term Detection Systems. In *Interspeech 2013*, August 25-29 2013.

[5] X. Anguera, F. Metze, A. Buso, I. Szoke, and L. J. Rodriguez-Fuentes. The Spoken Web Search Task. In *MediaEval 2013 Workshop*, October 18-19 2013.

[6] N. Morgan and H. Bourlad. An introduction to hybrid HMM/connectionist continuous speech recognition. *IEEE Signal Processing Magazine*, 12(3):25–42, 1995.

[7] L. Rodriguez-Fuentes and M. Penagarikano. MediaEval 2013 Spoken Web Search Task: System Performance Measures. Technical report, 2013.

[8] L. J. Rodriguez-Fuentes, A. Varona, M. Penagarikano, G. Bordel, and M. Diez. GTTS Systems for the SWS Task at MediaEval 2013. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[1] https://www.l2f.inesc-id.pt/wiki/index.php/DTW

[2] https://www.l2f.inesc-id.pt/wiki/index.php/STDfusion