The L2F Spoken Web Search system for Mediaeval 2012

Alberto Abad
L²F - Spoken Language Systems Lab
INESC-ID Lisboa, Portugal
alberto@l2f.inesc-id.pt

Ramón F. Astudillo
L²F - Spoken Language Systems Lab
INESC-ID Lisboa, Portugal
ramon@l2f.inesc-id.pt

ABSTRACT

This document presents a brief description of INESC-ID's Spoken Language Systems Laboratory (L²F) Spoken Web Search system submitted to the Mediaeval 2012 evaluation campaign. The L²F system consists of the fusion of four individual sub-systems based on hybrid approaches for speech recognition exploiting four different language-dependent phonetic classifiers. The achieved results confirm the proposed system as a simple, efficient and considerably robust approach to the problem of spoken query search.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering, Query formulation, Search process*

General Terms

Algorithms, Experimentation

Keywords

Spoken query search, Acoustic keyword spotting, Connectionist speech recognition

1. INTRODUCTION

The objective of the "Spoken Web Search" (SWS) task at MediaEval 2012 consists of searching for audio content within audio content using an audio content query. This year evaluation introduces a new "African" dataset as the primary evaluation corpus. This data set consists of audio content created over the phone in four South African languages. Detailed information about the task and the data used for this evaluation can be found in the evaluation plan[2].

This document introduces the SWS system developed by the INESC-ID's Spoken Language Systems Laboratory (L²F) for the Mediaeval 2012 campaign. The system is composed by the fusion of four individual phonetic-based SWS subsystems. Each sub-system exploits different language-dependent phonetic networks. The main characteristic of the sub-systems is that they exploit hybrid ANN/HMM connectionist methods for both query tokenization and search. In a first stage, a language-dependent phonetic tokenization of the query is obtained. Then, in a second phase, acoustic keyword spotting (AKWS) of the query in contrast to a filler or background model is performed. Different detection score normalization and fusion strategies were investigated during the development of the systems. The final submitted system applies per-query score normalization (Q-norm) and majority voting (MV) fusion.

Mediaeval 2012 Workshop, October 4-5 2012, Pisa, Italy

2. THE L²F SWS SYSTEM DESCRIPTION

Four sub-systems form the core of the L²F SWS system. Each sub-system is based on our built-in automatic speech recognition (ASR) system named AUDIMUS [3]. The different sub-systems exploit four different language-dependent acoustic models trained for European Portuguese (*pt*), Brazilian Portuguese (*br*), European Spanish (*es*) and American English (*en*).

2.1 Speech search based on ANN/HMM

For each sub-system, a phonetic tokenization of the queries is obtained. Then, AKWS is performed to search for each query in the data. Both the phonetic tokenization and the query search are based on hybrid ANN/HMM approaches for ASR.

2.1.1 The baseline hybrid ASR system

AUDIMUS is a hybrid recognizer that follows the connectionist approach [4]. A Markov process is used to model the basic temporal nature of the speech signal, while an artificial neural network is used to estimate posterior phone probabilities given the acoustic data at each frame.

Feature extraction The speech recognizers used for this task combine 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).

MLP phonetic models The language-dependent MLP networks used are part of our existing ASR systems and were trained using different amounts of annotated data [1]. 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 41 (39 phonemes

+1 silence + 1 respiration) soft-max outputs in the case of *en*, 39 for *pt* (38 phonemes + 1 silence), 40 for *br* (39 phonemes + 1 silence) and 30 for *es* (29 phonemes + 1 silence).

WFST decoder The AUDIMUS decoder is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition.

2.1.2 Spoken Query tokenization

A phone-loop grammar with phoneme minimum duration of three frames is used to obtain a phonetic transcription or tokenization for each query. In our development experiments we could not observe significant benefits using alternative n-best hypothesis for charactering each query. Hence, simple 1-best phoneme chain output has been used. Moreover, it was observed a large influence of the word insertion penalty (*wip*) parameter on the tokenization result. In general, it was found more convenient to set the *wip* term to 0.

2.1.3 Spoken Query search

Spoken query search is based on AKWS with our hybrid speech recognizer. A search sliding window of 5 seconds with 2.5 seconds of time shift is used to process each file. An equally-likely 1-gram language model formed by the target query and a competing speech background model is used. The minimum duration for the background speech word is set to 250 msec. On the one hand, keyword/query models are described by the sequence of phonetic units obtained in the previous tokenization stage. On the other hand, rather than re-training the whole phonetic network, it is possible to estimate the likelihood of a background speech unit representing "general speech" based on the other phonetic classes. In practice, we compute the likelihood of the background class as the mean likelihood of the top-6 most likely classes at each time frame. In order to control the weight of the background speech competing model in the search process, we introduce a background scale term β in the computation of the acoustic score of the background phonetic class. This β scale together with the *wip* term strongly affects searching results. These two parameters were adjusted during the development of the systems following a non-exhaustive greedy search. The detection score for each candidate detection is computed as the average of the phonetic log-likelihood ratios that form the detected query term.

2.2 Score normalization

Two different mean and variance score normalization strategies named Q-norm and F-norm have been investigated. Q-norm assumes that the scores may be dependent of the queries and applies a by-query normalization to compensate for this effect. Alternatively, F-norm was explored under the assumption that different data utterances may introduce different biases to the detection scores, consequently, per-utterance normalization is applied. In addition, QF-norm and FQ-norm combinations were investigated. According to our development experiments, consistent improvements are obtained using Q-norm with respect to the other strategies. Thus, Q-norm was applied to each sub-system.

2.3 System combination and score fusion

We have explored three simple schemes for sub-system combination: AND, OR and majority voting (MV). In the AND fusion only candidate detections detected by the four sub-systems are kept, in the OR fusion any candidate detection is kept, while the MV fusion keeps only the candidates given by at least two sub-systems. In the three cases, the fusion score is computed as the mean of the sub-system scores that detected that term. In practice, MV showed the best results in all our development experiments and it has been applied to our proposed system.

Table 1: L^{2}F SWS2012 performance scores

Task	dev-dev	dev-eval	eval-dev	eval-eval
ATWV	0.5313	0.4846	0.6185	0.5195
maxATWV	0.5313	0.4861	0.6327	0.5229

3. RESULTS

Table 1 shows the actual and maximum ATWV official scores obtained by the L^{2}F SWS system in the four conditions *query-collection*: *dev-dev*, *dev-eval*, *eval-dev* and *eval-eval*. Notice that the threshold associated to the maximum ATWV point for the *dev-dev* condition ($th = -0.117932$) was used as the decision threshold for the other conditions. Regarding the evaluation results, the L^{2}F proposed system is able to perform quite robustly in the four conditions. Surprisingly, the best performance is obtained in the *eval-dev* condition, a mismatched condition. In any case, it seems that the system is quite independent on the combination of query terms and utterances and the performance differences may be due to some particular characteristics of the data sets. Moreover, the difference between actual and maximum performance scores shows a well-calibrated system.

4. CONCLUSIONS

The L^{2}F Spoken Web Search system fully exploits hybrid ANN/HMM speech recognition for both query tokenization and query search. The resulting system formed by the fusion of four language-dependent sub-systems achieves an actual ATWV score of 0.5195 in the *eval-eval* condition. We consider this result very promising given the simplicity of the proposed system. Moreover, the L^{2}F system is well-calibrated and achieves also satisfactory results in the other evaluation conditions.

5. 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 project PEst-OE/EEI/LA0021/2011 and the grant number SFRH/BPD/68428/2010.

6. REFERENCES

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

[2] F. Metze et al. The Spoken Web Search Task Overview, 2012. MediaEval 2012 Workshop.

[3] H. Meinedo, A. Abad, T. Pellegrini, I. Trancoso, and J. Neto. The L2F Broadcast News Speech Recognition System. In *Fala 2010*, November 2010.

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