BUT-HCTLab APPROACHES FOR SPOKEN WEB SEARCH - MEDIAEVAL 2011

Igor Szöke
Speech@FIT, Brno University
of Technology, Czech Republic
szoke@fit.vutbr.cz

Javier Tejedor
HCTLab, Universidad
Autónoma de Madrid, Spain
javier.tejedor@uam.es

Michal Fapšo
Speech@FIT, Brno University
of Technology, Czech Republic
ifapso@fit.vutbr.cz

José Colás
HCTLab, Universidad
Autónoma de Madrid, Spain
jose.colas@uam.es

ABSTRACT

We present the three approaches submitted to the Spoken Web Search. Two of them rely on Acoustic Keyword Spotting (AKWS) while the other relies on Dynamic Time Warping. Features are 3-state phone posterior. Results suggest that applying a Karhunen-Loeve transform to the log-phone posteriors representing the query to build a GMM/HMM for each query and a subsequent AKWS system performs the best.

Categories and Subject Descriptors

H.3.3 [**Information systems**]: Information Storage and Retrieval, Information Search and Retrieval, Search process

General Terms

Experimentation

Keywords

query-by-example spoken term detection, acoustic keyword spotting, dynamic time warping, spoken web search

1. MOTIVATION

The Spoken Web Search (SWS) task aims at building a language-independent query-by-example spoken term detection system without any knowledge of the target language and query transcriptions. In so doing, our approaches are based on the combination of as many language-dependent "recognizers" as possible. [1]

2. FEATURE EXTRACTION

Our feature extractor outputs 3-state phone posteriors as features [2]. The phone posterior estimator [5] contains a Neural Network (NN) classifier with a hierarchical structure called *bottle-neck universal context network*. It consists of a context network, trained as a 5-layer NN, and a merger which employs 5 context net outputs. Relevant parameters are in Table 1 and more details are in [5].

Table 1: **Language parameter specifications: Number of phones, $|UC|$ which represents the size of the hidden layer in the universal context NN, $|Mer|$ which represents the size of the hidden layers in the merger with 3-state phone posteriors output, and $|out|$ which represents the size of the 3-state phone posteriors output layer.**

Language	Phones	$\|UC\|$	$\|Mer\|$	$\|Out\|$
Czech	37	1373	495	114
English	44	1298	488	135
Hungarian	64	1128	470	193
Levantine	32	1432	500	99
Polish	33	1408	498	105
Russian	49	1228	481	157
Slovak	41	1305	489	133

3. APPROACHES

3.1 Parallel Acoustic Keyword Spotting (PAKWS)

We combined decisions from 6-language dependent Acoustic Keyword Spotters (AKWS) (from all the languages in Table 1 except the Polish one due to its worse performance on dev data). One AKWS consists of two steps: *Query recognition* done by a phone recognizer and *query detection* done by AKWS. They only differ in the decoder. Features (3-state phone posteriors) extracted from the audio are fed into a phone decoder – unrestricted phone loop without any phone insertion penalty. AKWS filler model-based recognition networks [4] are built according to the detected phone string per each query. The filler/background models are represented by a phone loop. Each phone model is represented by a 3-state HMM tied to 3-state phone posteriors. The output of the AKWS is a set of putative hits. The score is logarithm of likelihood ratio normalized by the length of the detection. These detections are converted into a matrix for each utterance. The size of this matrix is $\#queries \times \#frames$. Next, matrices for all 6 languages are "log added". Finally, the combined matrix is converted back to the list of detections and the detection for which all 6 detectors agree has a higher score.

3.2 GMM/HMM term modeling

Inspired in the previous approach, this relies on a single AKWS as *query detection* with these differences: (1) the background model is a GMM/HMM with 1-state modeled with 10-GMM components, (2) the query model is rep-

[1] Part of this work was done will JT was a visitor research at Speech@FIT, BUT.

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

resented with a GMM/HMM whose number of states is 3 times the number of phones according to the phone recognition with 1 GMM component each and (3) all the languages in Table 1 have been employed to produce the final feature super-vector. Queries represented by a single phone have been modeled with 6 states, as if the query contained 2 phones. We used the number of phones output by the Slovak recognizer due to its best performance in terms of the Upper-bound Term Weighted Value metric (UBTWV) [3]. Features used for background and query modeling were got as follows: (1) the log-phone posteriors got from the feature extractor are applied a Karhunen-Loeve transform (KLT) for each invididual language, (2) we keep the features that explain up to 95% of the variance after KLT for each individual language, (3) we build a 152-dimensional feature super-vector, from them. The KLT statistics have been computed from the dev data and next applied over both the dev/eval queries and dev/eval data.

3.3 Dynamic Time Warping (DTW)

A similarity matrix from phonetic posteriorgrams [5] stores the similarity between each query frame and each utterance frame with the cosine distance as similarity function and a DTW search hypothesises putative hits. DTW is run iteratively starting in every frame in the utterance and ending in a frame on the utterance [5]. The features that represent the phonetic posteriorgrams are the concatenation of the 3-state phone posteriors corresponding to every language in Table 1.

4. FILTERING AND CALIBRATION

To deal with the score calibration and some problematic query length under certain approaches issues, detections were post-processed in the following steps: 1]"Filtering" detections according to length difference from "average length". Average length of a query is calculated as the average length of speech (phones) across the 6 phone recognizers used in the PAKWS approach. It was applied on all the approaches except the DTW as follows:

$$ScF(det) = \begin{cases} Sc(det) - \frac{L^Q_{min} - L(det)}{L^Q_{min}}, & L^Q_{min} > L(det) \\ Sc(det) - \frac{L(det) - L^Q_{max}}{L^Q_{max}}, & L^Q_{max} > L(det) \\ Sc(det), & otherwise \end{cases} \quad (1)$$

where Q identifies the query to which the detection belongs, $Sc(det)$ is the original score, $ScF(det)$ is the "filtered" score, $L(det)$ is the length of the detection in frames, $L^Q_{min} = 0.8L^Q_{aver}$ is 80% of the average query length and $L^Q_{max} = 1.4L^Q_{aver}$ is 140% of the average query length. The detection score remains the same if the detection length is longer than 80% and shorter than 140% of the average query length. Otherwise the score is lowered the shorter/longer the detection is according to the original query.

2] Calibration, applied only in our PAKWS approach, produces the final score of each detection as follows:

$$ScC(det) = ScF(det) + A_1 + A_2 * Occ(Q), \quad (2)$$

where $Occ(Q)$ is number of query detection occurrences in the data, and $A_1 = -1.0807$ and $A_2 = -0.0001$ are calibration parameters. These were estimated from best thresholds (UBTWV) on dev data using linear regression.

5. RESULTS AND DISCUSSION

Results for the required runs [1] are given in Table 2. The PAKWS approach has two versions (with and without score calibration). We clearly see that the GMM/HMM term modeling approach outperforms the two other in a great extent for unseen queries/data even with score calibration applied on the PAKWS approach. We consider this is due to: (1) The KLT statistics, computed from dev data and applied on queries and data, plays the role of "adaptation" towards the target domain, which differs from that used to train the phone estimators from which the 3-state phone posteriors are computed, (2) the use of a single example to train the query model is more robust against uncertainties than the set of features itself (used in the DTW approach) and the phone transcription got from phone decoding (used in the PAKWS approach), (3) the *prior combination* of the most relevant features after KLT given to the GMM/HMM approach, opposite to the PAKWS approach, based on a *posterior combination* from the detections got from each individual AKWS system and (4) by comparing the MTWV (pooled) and UBTWV (non pooled) for PAKWS Qdev-Ddev 0.133 and 0.253, Qeval-Ddev 0.002 and 0.056, Qdev-Deval 0.030 and 0.157 and Qeval-Deval 0.033 and 0.223 respectively, suggests that the PAKWS system is the most sensitive to data mismatch. The GMM/HMM-based term modeling approach is less sensitive to data mismatch with the following values: Qdev-Ddev 0.103 and 0.238, Qeval-Ddev 0.019 and 0.035, Qdev-Deval 0.010 and 0.179 and Qeval-Deval 0.131 and 0.267 respectively. For the DTW similar pattern as that of GMM/HMM is observed: Qdev-Ddev 0.020 and 0.106, Qeval-Ddev 0 and 0.011, Qdev-Deval 0 and 0.099 and Qeval-Deval 0.014 and 0.055.

Table 2: *ATWV results for the approaches. "PAKWS-cal" denotes the PAKWS approach with score calibration and "PAKWS-nocal" denotes the PAKWS approach without score calibration. "Qx-Dy" denotes the set of "x" queries searched on the set of "y" data.*

Approach	Qdev-Ddev	Qeval-Ddev	Qdev-Deval	Qeval-Deval
PAKWS-cal	0.133	−0.221	−1.141	−0.307
PAKWS-nocal	0.093	−0.359	0.009	−0.110
GMM/HMM	0.103	0.008	0.024	0.101
DTW	0.020	−0.005	−0.032	−0.115

6. CONCLUSIONS

Our GMM/HMM-based term modeling approach achieves the best performance, whereas the two other, PAKWS and DTW, fail due to the unreliable phone transcription derived in the former and the "meaningless" phone posteriors by themselves used in the latter when facing to the language-independency issue. Future work will investigate new features to enhance the performance of the best approach.

7. REFERENCES

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

[2] P. Schwarz et al. Towards lower error rates in phoneme recognition. In *Proc. of TSD*, pages 465–472, 2004.

[3] I. Szöke. *Hybrid word-subword spoken term detection.* PhD thesis, 2010.

[4] I. Szöke et al. Phoneme based acoustics keyword spotting in informal continuous speech. *LNAI*, 3658(2005):302–309, 2005.

[5] J. Tejedor et al. Novel methods for query selection and query combination in query-by-example spoken term detection. In *Proc. of SSCS*, pages 15–20, Florence, Italy, 2010.