BUT SWS 2013 - Massive Parallel Approach

Igor Szöke*, Lukáš Burget, František Grézl, and Lucas Ondel
Speech@FIT, Brno University of Technology, Czech Republic
szoke@fit.vutbr.cz

ABSTRACT

We submitted a system composed of 26 subsystems as the required run. 13 subsystems are based on Acoustic Keyword Spotting and 13 on DTW. All of them were using three state phoneme posteriors as input. The underlaying phoneme posterior estimators were both in-language (Czech, English) and out-of-language (other 12 languages). We also performed unsupervised adaptation of the artificial neural network (ANN) on the target data and fusion based on binary logistic regression.

1. MOTIVATION

Our motivation was to use many (mostly) out-of-target-languages systems which can be combined by fusion at the detection level. The goal was to re-use as many trained systems available at BUT as possible. Please bear in mind that reusing all these systems (so-called Atomic Systems) lead to several inconsistencies among them — for example feature extraction and sizes of the ANN.

We performed unsupervised adaptation of ANN on the target data (utterances). Our goal was also to test combination two approaches in the query-by-example task – Acoustic Keyword Spotting (AKWS) and Dynamic Time Warping (DTW). We also explored system calibration with respect to the TWV metric.

2. ATOMIC SYSTEMS

All our subsystems use ANN to estimate per-frame phone-state probabilities (so-called posteriorgrams). The subsystems based on DTW use the posteriorgrams as features for calculating distances between query and test segment frames. The subsystems based on AKWS uses the phone-state posteriors as HMM output probabilities. We re-use ANNs, which were trained for different projects as acoustic models for phone or LVCSR recognizers. One DTW and one AKWS system were built for each of the 13 ANNs trained on the following datasets: 3× **Speechdat** (Czech, Hungarian and Russian; monolingual LCRC systems [5]), 1× **BABEL** (Cantonese, Pashto, Tagalog, Turkish; multilingual stacked-bottleneck system [4]), 1× **SWS2012** (MediaEval SWS2012 development data; multilingual stacked-bottleneck system [7]), 8× **GlobalPhone** (Czech, English, German, Portuguese, Russian, Spanish, Turkish, Vietnamese; monolingual stacked-bottleneck systems [8, 3].

*Igor Szöke was supported by Grant Agency of Czech Republic post-doctoral project No. GP202/12/P567.

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

3. ACOUSTIC KEYWORD SPOTTING

The Acoustic Keyword Spotting (AKWS) systems follow our paper [6]. We build an HMM for each query. For each frame, the detection score is calculated as the log-likelihood ratio between 1) staying in a background HMM (free phoneme loop) and 2) escaping from it through the query HMM.

For standard keyword spotting tasks (in-language task and textual input), the query model is built using a pronunciation dictionary. In SWS task, however, we need to generate the phoneme sequence for each of the query acoustic examples – **query-to-text step**. This is achieved by decoding each example using free phoneme loop. We cut-off initial and final silence labels (if present) and omit queries having less than three non-silence phones, as these short queries could generate huge amounts of false alarms. We experimented with phoneme insertion penalty in the query-to-text step with the conclusion that it has no significant impact and set it to -1 consistently.

4. DYNAMIC TIME WARPING

In our implementation, we follow the standard query-by-example recipe – subsequence DTW. Single DTW is run for each combination of query and test segment, where the query is allowed to start at any frame of the test segment. When selecting the locally optimal path in the standard DTW algorithm, transition from the smallest accumulated distance is chosen. In our implementation, we compare the accumulated distances (including the current local distance) normalized by the corresponding path lengths on-the-fly. This is to avoid the preference for shorter paths. As the distance metric, we used the usual negative logarithm of the dot product of phone-state posterior vectors.

In our late submission, we further improved the DTW systems by applying VAD to cut off the initial and the final silence from the query examples. As can be seen in Table 1, it improved the overall system by 10% relative.

5. DETECTION SCORE POST-PROCESSING

For both DTW and AKWS systems, the local maxima of frame-by-frame detection scores are selected as candidate detections. For overlapping detections, only the best scoring ones are preserved. There might be significant differences between the score distributions corresponding to the different queries and it is important to normalize (calibrate) the scores for each query to allow for a single common threshold maximizing the TWV metric. We adopted a new normalization approach, *m-norm*, which is motivated by the observation that score distributions have very long tails towards the

small scores, which significantly differ in shape from query to query. In *m-norm*, for each query, score corresponding to the mod (maximum) of the score histogram (denoted SM) is found for each query and subtracted from the original scores – all mods are thus aligned to 0. The scores are further divided by standard deviation estimated only on scores larger than SM, to unify the terms' variances.

6. FUSION

Normalized scores from the individual subsystems were fused similarly to [1]. The scores from different subsystems are first aligned in time and then linearly combined. The fusion weights (and the default score for a subsystem with no detection at the given time) are trained to minimize cross-entropy (or binary logistic regression) objective.

7. RESULTS

Approach	eval MTWV	eval RT	dev MTWV	dev RT
AKWSDTW-vad (late)	0.3776(0.4835)	0.177	0.4373(0.5310)	0.181
DTW-vad (notsub)	0.3557(0.4585)	0.166	0.4199(0.5153)	0.170
AKWS	0.3041(0.4165)	0.011	0.3644(0.4713)	0.011
AKWSDTW	0.2969(0.4081)	0.276	0.3710(0.4719)	0.281
AKWSDTW-treefus	0.2787(0.3934)	0.276	0.3560(0.4622)	0.281
AKWSDTW-notarlang	0.2562(0.3685)	0.213	0.3237(0.4264)	0.216
AKWS-notarlang	0.2778(0.3840)	0.009	0.3285(0.4351)	0.009

Table 1: *Results for the approaches in Maximum TWV, with Upper Bound (UB) TWV in parenthesis. RT - real-time factor for search step (per second of query). The indexing step RT is* 1.996 *for all systems except *-notarlang where the RT factor is* 1.545*. (notsub) means not submitted, (late) means late submission. The highest memory consumption (high level water mark) is* 210MB *with DTW systems. The experiments were run on a hybrid cluster with average CPU Intel(R) Xeon(R) CPU X5670 @ 3GHz.*

8. LESSONS LEARNED

8.1 NN adaptation

We have experimented with three types of NN adaptation using BABEL system NN. This network was initially trained with 4 independent softmax non-linearities in the output layer – one softmax per language (Cantonese, Pashto, Tagalog, Turkish). The original network had 1065 phoneme-state outputs (355 phonemes for all the 4 languages). We decoded SWS-dev data [2] using free phoneme loop phone recognizer based on this network and we found, that 37 out of 355 were never activated. We also filtered out 95 phonemes having less occurrences than 10 seconds in total. We ended up with 220 "active" phonemes – denoted as *orig*. Next, we used this *orig* network to label the SWS-dev data again. Using this labeling, we 1) adapted (re-trained) the original NN on SWS-dev data (denoted as *adapt*) and 2) we completely re-trained the NN using the SWS-dev data (denoted as *rtfs*). In the stacked bottleneck NN hierarchy, only the merger was adapted in *adapt* case.

In terms of MTWV (UBTWV), our results on SWS dev with BABEL AKWS subsystem are as follows : *orig* 0.0443 (0.1154), *adapt* 0.0569 (0.1355), and *trfs* 0.0769 (0.1630).

8.2 Calibration

As the TWV metric was set to drastically penalize false alarms [2], the proper calibration and good choice of global threshold was very important this year. We experimented with two approaches for score normalization. First, we have experimented with *z-norm* that worked well for the last year SWS evaluations [9].

Next, we tried to calibrate the scores using binary logistic regression, where the input to the logistic regression was a vector of z-normed scores augmented with different per-term side-information scores [1] – denoted as *z-norm-sideinfo*. The best tested side information, which significantly improved MTWV, was *the logarithm of the number of detections* of a particular term. This indicates that z-norm is not sufficient to properly normalize score distributions over different queries.

Finally, we tested *m-norm* (see section 5), which we found to be superior to *z-norm-sideinfo*. Furthermore, the additional side information based calibration resulted in no further MTWV improvements.

In terms of MTWV (UBTWV), our results are: *raw* 0.000 (0.1012), *z-norm* 0.0330 (0.1434), *z-norm-sideinfo* 0.0603 (0.1436), *m-norm* 0.0769 (0.1611).

9. CONCLUSION

We successfully built a QbE system making use of a high number of already trained phoneme posterior estimators and applied unsupervised adaptation of ANNs. DTW with application of VAD (VAD is really important) is still superior to AKWS approach. On the other hand, AKWS is level-of-magnitude faster compared to DTW. Adaptation of ANN is also important, so it makes sense to take as much as possible "black-boxed" phoneme posterior estimators, label the target data and train a new ANN. Finally, we found *m-norm* calibration is promising in the area of high FA rates and non-posterior scores.

10. REFERENCES

[1] M. Akbacak et al. Rich system combination for keyword spotting in noisy and acoustically heterogenous audio streams. In *Proceedings of ICASSP 2013*, pages 8267–8271. IEEE Signal Processing Society, 2013.

[2] X. Anguera et al. The spoken web search task. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[3] F. Grézl et al. Hierarchical neural net architectures for feature extraction in asr. In *Proceedings of INTERSPEECH 2010*, volume 2010, pages 1201–1204. International Speech Communication Association, 2010.

[4] M. Karafiát et al. BUT babel system for spontaneous cantonese. In *Proceedings of Interspeech 2013*, number 8, pages 2589–2593. International Speech Communication Association, 2013.

[5] P. Schwarz et al. Towards lower error rates in phoneme recognition. In *Proceedings of 7th International Conference Text,Speech and Dialoque 2004*, page 8. Springer Verlag, 2004.

[6] I. Szöke et al. Phoneme based acoustics keyword spotting in informal continuous speech. *Lecture Notes in Computer Science*, 2005(3658):8, 2005.

[7] I. Szöke et al. BUT2012 approaches for spoken web search - mediaeval 2012. In *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012. CEUR Workshop Proceedings, Vol. 2012, No. 927, DE.

[8] K. Veselý et al. The language-independent bottleneck features. In *Proceedings of IEEE 2012 Workshop on Spoken Language Technology*, pages 336–341. IEEE Signal Processing Society, 2012.

[9] H. Wang et al. CUHK system for the spoken web search task at mediaeval 2012. In *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012. CEUR Workshop Proceedings, Vol. 2012, No. 927, DE.