BUT QUESST 2014 System Description

Igor Szöke*, Miroslav Skácel, and Lukáš Burget
BUT Speech@FIT, Brno University of Technology, Czech Republic
szoke@fit.vutbr.cz

ABSTRACT

The primary system we submitted was composed of 11 subsystems as the required run. 3 subsystems are based on Acoustic Keyword Spotting (AKWS) and 8 on Dynamic Time Warping (DTW). The AKWS systems were based only on phoneme posteriors while the DTW subsystems were based on both phoneme posteriors and Bottle-Neck features (BN) as input. The underlying phoneme posterior estimators / bottle-neck feature extractors were both in-language (Czech) and out-of-language (other 4 languages). We also performed experiments on T1/T2/T3 types of query, system calibration and fusion based on binary logistic regression.

1. MOTIVATION

In comparison to last year [7], we decided to use lower number of systems in parallel. Our goal was to further investigate the sensitivity of particular approaches to the language / channel mismatch in the query and utterance data. Also, coping with different types of queries was challenging this year [1]. Similarly to last year, we used systems already available at BUT (so-called Atomic Systems). This led to several inconsistencies — for example, feature extraction and sizes of the Artificial Neural Networks (ANN).

2. ATOMIC SYSTEMS

All our subsystems use ANN to estimate 1) per-frame phone-state probabilities (so-called posterior-grams) 2) bottleneck (BN) features. The subsystems based on DTW use the BN features for calculating distances between query and test segment frames. The subsystems based on AKWS use the phone-state posteriors as HMM output probabilities. We reuse ANNs, which were trained for different projects as acoustic models for phone or LVCSR recognizers: 3× **SpeechDat** (Czech, Hungarian and Russian; monolingual LCRC systems [5]) for phone posterior-grams and 4× **GlobalPhone** (Czech, Portuguese, Russian, Spanish; monolingual stacked-bottleneck systems [3]) for BN features.

We prefer the SpeechDat posterior-grams to GlobalPhone posterior-grams in AKWS due to significantly lower accuracy of "GlobalPhone posterior-grams". For DTW approach, we prefer the GlobalPhone bottle-necks to GlobalPhone posterior-grams also due to significant accuracy deterioration. We have observed even larger deterioration when the GlobalPhone ANNs were adapted on the SWS2013 data in unsupervised manner (as we performed last year with positive impact on accuracy). This holds both for posteriors and BNs.

We ended with 3 atomic AKWS systems based on SpeechDat posteriors and 7 atomic DTW systems based on GlobalPhone bottle-necks.

3. ACOUSTIC KEYWORD SPOTTING

The AKWS systems follow [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 removed all silence labels (if present).

4. DYNAMIC TIME WARPING

In our implementation, we follow the standard query-by-example recipe – sub-sequence 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 Pearson product moment correlation distance.

We applied Speech Activity Detection (SAD) to drop out the silence frames in queries (see our last year's work [7]). We also tried to apply the SAD on utterances, but obtained only tiny improvement therefore SAD was not used. The Hungarian SpeechDat phoneme recognizer was used as the SAD.

4.1 Fusion

We were inspired by GTTS [4] (concatenation of feature vectors going into DTW) and CUHK [8] (averaging of distance matrices). Finally, we found both of these methods comparable, so we followed the feature vector concatenation approach. We concatenate the Czech, Portuguese, Russian, and Spanish GlobalPhone BN features and made a new Atomic DTW system – the eight system.

5. SCORE POST-PROCESSING

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

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

Approach	sideinfo	eval $minC_{nxe}$	eval RT	dev $minC_{nxe}$	dev RT
p-bigfusion	QU	**0.465**(0.310/0.461/0.673)	0.086	0.461(0.309/0.513/0.624)	0.082
g-bigfusionnoside		0.464(0.323/0.470/0.660)	0.086	0.486(0.333/0.554/0.624)	0.082
g-best_single	QU LID	**0.528**(0.374/0.546/0.714)	0.010	0.533(0.376/0.600/0.675)	0.010
g-LID		0.926(0.897/0.946/0.920)	$1e^{-6}$	0.929(0.896/0.961/0.901)	$1e^{-6}$
AKWS-cz	QU LID	**0.648**(0.519/0.645/**0.848**)	-	**0.641**(0.500/0.680/**0.824**)	-
AKWS-T3-cz	QU LID	**0.674**(0.597/0.694/**0.756**)	-	**0.673**(0.581/0.742/**0.718**)	-

Table 1: *Results for the approaches in minimum C_{nxe} with per query type (T1/T2/T3). RT - real-time factor for search step (per second of query). The indexing step RT is* 1.03 *for both bigfusion systems,* 0.18 *for g-best_single, and* 0.04 *for g-LID. The highest memory consumption (high level water mark) is* 450*MB with DTW systems. The experiments were run on a hybrid cluster with average CPU Intel(R) Xeon(R) CPU X5670 @ 3GHz.*

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. We applied m-norm (developed in SWS2013 [7]) to normalize (calibrate) the scores for each query to allow for a single common threshold maximizing the C_{nxe} metric.

As the task was document retrieval rather than keyword spotting this year, only one score per query–utterance pair without timing was requested. That is why we find and return the best particular score from a set of detections of a query in an utterance.

6. CALIBRATION

The post-processed scores were calibrated to respect the C_{nxe} scoring metric using binary logistic regression.

We attached a side info to each score (query–utterance pair). The side-info consists of: number of phonemes, log of number of phonemes, number of speech frames, log of number of speech frames, average log-posterior of speech frames taken from SAD and optionally the LID i-vector score. The side-info was generated for queries and utterances so the final "feature vector" for calibration consists of: 1 detection score (query–utterance pair), 5 query side-info, 5 utterance side-info. Parameters (11 linear scales and 1 additive constant) were trained on development set. We denoted this 10 side-info parameters as *QU*.

The language identification system is a state-of-the-art system based on i-vectors [2]. As acoustic features, we used Shifted Delta Cepstra. Gaussian mixture model with 2048 Gaussians serves as Universal Background Model for 600 dimensional, gender-independent, i-vector extractor. Our goal here was to calculate distance (Pearson product moment correlation distance) between particular query and utterance i-vectors. This distance should provide us similarity measure on the level of language (as Czech queries do not exist in Basque utterances for example) and was used also as side-info (denoted as *LID*).

7. FUSION

Finally, we applied fusion on the level of calibrated systems using the binary logistic regression again. We took all 11 systems (3 AKWS, 7 DTW, 1 fused DTW) and found the best linear combination of them.

8. CONCLUSION

We tried to approach the T2 and T3 queries to improve accuracy of our system. However, we ended up with conclusion that slight improvement accuracy of T2 / T3 queries largely degrades accuracy of T1 queries. This leads to overall score degradation. Our conclusion here was, that it does not make sense to cover T2 queries by a special approach (search algorithm), as these queries are covered enough by "softness" of standard DTW algorithm. We found tiny improvement of 0.4% on T2 while we got overall 1% C_{nxe} deterioration. This T2 improvement was observed with AKWS approach when we allowed the last phoneme to be any phoneme.

To improve accuracy of T3 queries, we split queries longer than 7 phonemes in the middle. Then, we searched for these two particular sub-queries independently. Finally, we merged the sub-query results by forbidding sub-queries overlap longer than 10 frames. Results of this experiment are in table 1. System AKWS-cz is reference system where we search for T3 in the same way as for T1. We implement the above mentioned split to sub-queries in system AKWS-T3-cz. We got improvement 9% on T3 but overall deterioration is 2.4% of C_{nxe} on eval queries.

We built a QbE system making use of phoneme posteriors and bottlenecks as input features. We found DTW superior to AKWS event in cross channel environment (this year data set). Our conclusion on different types of query is, that it does not make sense to aim at T2 queries due to tiny 0.4% improvement on the T2 but significant 1% deterioration on overall score. The same holds for T3, where the improvement is significant (9%) but overall deterioration is (2.4%). The T3 queries need more investigation to overcome the overall deterioration.

9. REFERENCES

[1] X. Anguera et al. Query by Example search on speech at MediaEval 2014. In *Working Notes Proceedings of the Mediaeval 2014 Workshop*, Barcelona, Spain, October 16-17 2014.

[2] N. Brummer et al. Description and analysis of the brno276 system for LRE2011. In *Proceedings of Odyssey 2012: The Speaker and Language Recognition Workshop*, pages 216–223. International Speech Communication Association, 2012.

[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] L. J. Rodriguez-Fuentes et al. GTTS systems for the SWS task at MediaEval 2013. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, volume 2013, pages 1–2, 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. Calibration and fusion of Query-by-Example systems - BUT SWS 2013. In *Proceedings of ICASSP 2014*, pages 7899–7903. IEEE Signal Processing Society, 2014.

[8] H. Wang and T. Lee. CUHK system for the Spoken Web Search task at MediaEval 2012. In *Proceedings of the MediaEval 2012 Multimedia Benchmark Workshop*, volume 2012, pages 1–2, 2012.