BUT QUESST 2015 System Description

Miroslav Skácel, Igor Szöke
BUT Speech@FIT, Brno University of Technology, Czech Republic
{iskacel, szoke}@fit.vutbr.cz

ABSTRACT

All our systems are based on Dynamic Time Warping (DTW). These systems use bottle-neck features (BN) as input. The bottle-neck feature extractors were trained on GlobalPhone Czech, Portuguese, Russian and Spanish languages, so our approach is in low-resource category. We also aimed on T1/T2/T3 types of query search for late submission systems. System calibration and fusion were based on binary logistic regression.

1. MOTIVATION

We developed one (single) system for on-time submission and two more systems for late submission. The system schema is in Figure 1. Similarly to last year, we used feature extractors already available at BUT (so-called Atomic Systems). We aimed only at bottle-neck features and DTW search approach this year. Our goal was to build a simple system and aim on word reordering in queries (T2/T3) (we addressed this problem in late submission). On the other hand, we have not addressed noise and reverberation in the data (see [1] for details on the task).

2. ATOMIC SYSTEMS

All our subsystems use Artificial Neural Networks (ANN) to estimate per-frame phone-state probabilities (so-called posterior-grams) and bottle-neck features. Subsystems are based on DTW using BNs to calculate distances between query and test segment frames. We re-use ANNs, which were trained for different projects as acoustic models for phone or LVCSR recognizers: 1× **SpeechDat** (Hungarian; monolingual LCRC systems [2]) for phone posterior-grams and 4× **GlobalPhone** (Czech, Portuguese, Russian, Spanish; monolingual stacked-bottleneck systems [3]) for BN features. We didn't exploit phone-state posterior-grams for DTW as in the last year's evaluations due to significant loss of accuracy for noisy data sets. The Hungarian phone recognizer was used only for Speech Activity Detection (SAD). Also, we didn't use Acoustic Keyword Spotting (AKWS) subsystems or ANNs adaptation on target language as we did in previous years.

We ended up with 4 atomic systems and 3 subsystems based on DTW using GlobalPhone features.

MediaEval 2015 Workshop, September 14-15, 2015, Wurzen, Germany

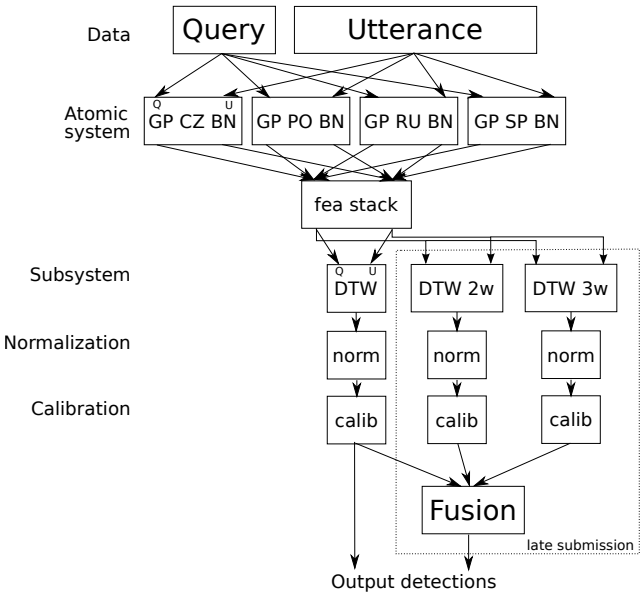

Figure 1: ***BUT QbE 2015 system. Q means queries as an input, U stands for utterances as an input, GP stands for GlobalPhone atomic systems where the output is bottleneck features.***

2.1 Fusion of features

We use concatenation of feature vectors for DTW proposed by GTTS [4]. The feature vectors are simply stacked on each other to create larger feature vector. We tried several combinations of 7 languages and ended up with a concatenation of the Czech, Portuguese, Russian, and Spanish GlobalPhone BNs (denoted as fea_stack).

3. DYNAMIC TIME WARPING

In our implementation, we follow the standard query-by-example recipe – sub-sequence DTW [5]. 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 online normalization. The *online* normalization performs the division by current path length on-the-fly for every step

System	sideinfo	eval		dev	
		C_{nxe}^{act}	C_{nxe}^{min}	C_{nxe}^{act}	C_{nxe}^{min}
p-fea_stack_DTW	QU	0.8452	0.8263 (0.7571/0.8647/0.8337)	0.8580	0.8426 (0.7588/0.8595/0.8706)
l-fea_stack_DTW+slope	-	0.8490	0.8184 (0.7408/0.8524/0.8306)	0.8772	0.8389 (0.7503/0.8470/0.8800)
l-fea_stack_DTW_2w+slope	-	-	-	0.8884	0.8569 (0.8253/0.8519/0.8744)
l-fea_stack_DTW_3w+slope	-	-	-	0.9188	0.8801 (0.8593/0.8716/0.8949)
l-fea_stack_DTW+slope+2w3w_fusion	-	0.8447	0.8124 (0.7423/0.8453/0.8212)	0.8731	0.8321 (0.7526/0.8381/0.8698)

Table 1: *Results for the systems in actual C_{nxe} for ALL types and minimum C_{nxe} with per query type ALL (T1/T2/T3).*

calculation to decide which step (vertical, horizontal or diagonal) is the best to choose. The division is not saved during the calculation, it is performed only to decide the next step. The length normalization is done afterwards as in standard approach. This leads to prefer longer paths over shorter ones.

As the distance metric, we used the Pearson product moment correlation distance [6]. We applied SAD to drop out non-speech frames in queries (see our previous work [7]). Queries having less than 10 frames after SAD application were discarded. The primary submitted system (denoted as p-fea_stack_DTW) using described algorithm was the winning one in last year's evaluations. We have made a few changes to the primary system for the late submission.

We used different step size conditions during the calculation of accumulated distances to control the slope of paths (systems denoted as *+slope*). Each path has a local slope within the bounds $\frac{1}{2}$ and 2. This limitation allows us to eliminate errors where one query frame maps to the whole utterance perfectly or vice versa. We also experimented with different local weights for vertical, horizontal and diagonal direction but we got no improvement out of it.

3.1 Dealing with T2/T3

We built additional subsystems to deal with T2/T3 type of queries. A query is split into equal parts and for each part, DTW is performed separately (denoted as *bands*). The smallest accumulated distance is chosen from each band of the given query and results are averaged together as a matching score. This approach allows us to search for multiple word queries. For single word queries, the results remain the same. Note that it is not mandatory that two words in T2 query are separated exactly in the middle of the query. We experimented with 1 (baseline system), 2 (denoted as 2w) and 3 (denoted as 3w) bands. These subsystems were used only for fusion and, therefore, were not submitted as separated late systems.

4. SCORE POST-PROCESSING

The global minimum of frame-by-frame detection scores is selected as candidate detection. There might be significant differences between the score distributions corresponding to the different queries and it is important to normalize the scores for each query.

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

As the task expects only one score per query–utterance pair without timing, we find and return the best particular score from a set of detections of a query in an utterance.

5. CALIBRATION

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

We attached a sideinfo to each score (query–utterance pair). The sideinfo consists of: number of phonemes, log of number of phonemes, number of speech frames, log of number of speech frames and average log-posterior of speech frames taken from SAD. The sideinfo was generated for queries and utterances so the final "feature vector" for calibration consists of: 1 detection score (query–utterance pair), 5 query sideinfo, 5 utterance sideinfo. Parameters (11 linear weights and 1 additive constant) were trained on development set. We denoted this 10 sideinfo parameters as *QU*. However, we found sideinfo harms the performance for late submission so we omit it.

6. FUSION

We applied fusion on the level of calibrated systems using the binary logistic regression again.

For improved late system, we fused the primary system with slope limitation (l-fea_stack_DTW+slope) and the 2w and 3w systems. The fused system is denoted as "l-fea_stack_DTW+slope+2w3w_fusion" and was submitted as second late system.

7. CONCLUSION

First, we processed data by the primary system without respect to T2 and T3 type of query. The first general late system using slope constraint improved output score by 0.79% in C_{nxe}^{min} for eval data. The conclusion is that for such noisy data, there could be one or few single query frames fitting perfectly to single or few frames from utterance. This generates path too steep or too shallow, obviously not a matching hit.

To improve accuracy of T2 queries, we used algorithm where queries are split into parts and search is performed "per-partes". The output scores of these subsystems were not significantly better, however, helped in fusion with the best single system. We got improvement of 0.6% in C_{nxe}^{min}. More detailed, the deterioration of T1 query is 0.15% but there is slight improvement for T2 query (0.71%) and T3 query (0.94%).

The real-time factor for the primary system is 0.009, for late system without fusion is 0.009 and for late system with fusion reaches 0.023. The highest memory consumption (high level water mark) is 450MB. The experiments were run on a hybrid cluster with average CPU Intel(R) Xeon(R) CPU X5670 @ 3GHz.

8. REFERENCES

[1] Igor Szöke, Luis J. Rodriguez-Fuentes, Andi Buzo, Xavier Anguera, Florian Metze, Jorge Proença, Martin Lojka, and Xiao Xiong. Query by Example Search on Speech at Mediaeval 2015. In *Working Notes Proceedings of the Mediaeval 2015 Workshop*, Wurzen, Germany, September 14-15 2015.

[2] Petr Schwarz, Pavel Matějka, and Jan Černocký. Towards Lower Error Rates in Phoneme Recognition. In *Proceedings of 7th International Conference Text,Speech and Dialoque 2004*, page 8. Springer Verlag, 2004.

[3] František Grézl and Martin Karafiát. 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] Luis J. Rodriguez-Fuentes, Amparo Varona, Mikel Penagarikano, Germán Bordel, and Mireia Diez. 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] Meinard Müller. *Information Retrieval for Music and Motion*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[6] Igor Szöke, Miroslav Skácel, Lukáš Burget, and Jan Černocký. Coping with Channel Mismatch in Query-by-Example - BUT QUESST 2014. In *Proceedings of ICASSP 2015*, pages –. IEEE Signal Processing Society, 2015.

[7] Igor Szöke, Lukáš Burget, František Grézl, and Lucas Ondel. Calibration and Fusion of Query-by-Example Systems - BUT SWS 2013. In *Proceedings of ICASSP 2014*, pages 7899–7903. IEEE Signal Processing Society, 2014.