TUKE at MediaEval 2015 QUESST

Jozef Vavrek, Peter Viszlay, Martin Lojka, Matúš Pleva, Jozef Juhár and Milan Rusko*
Laboratory of Speech Technologies in Telecommunications @ Technical University of Košice
Park Komenského 13, 041 20 Košice, Slovakia
*Institute of Informatics, Slovak Academy of Sciences, Dúbravská cesta 9, 845 07 Bratislava, Slovakia
{Jozef.Vavrek, Peter.Viszlay, Martin.Lojka, Matus.Pleva, Jozef.Juhar}@tuke.sk
*Milan.Rusko@savba.sk

ABSTRACT

In this paper, we present our retrieving system for QUery by Example Search on Speech Task (QUESST), comprising the posteriorgram-based modeling approach along with the weighted fast sequential dynamic time warping algorithm (WFS-DTW). For this year, our main effort was directed toward developing language-dependent keyword matching system, utilizing all available information about spoken languages, considering all queries and utterance files. Despite the fact that the retrieving algorithm is the same as we used in previous year, a big novelty resides in the way of utilizing the information about all languages spoken in the retrieving database. Two low-resource systems using language-dependent acoustic unit modeling (AUM) approaches have been submitted. The first one, called supervised, employs four well-trained phonetic decoders using acoustic models trained on time-aligned and annotated speech. The second one, defined as unsupervised, uses blind phonetic segmentation for the specific language where the information about spoken language is extracted from Mediaeval 2013 and Mediaeval 2014 databases. Considering the influence on the overall retrieving performance, the acoustic model adaptation to the specific language through retraining procedure was investigated for both approaches as well.

1. MOTIVATION

Challenging acoustic conditions and different types of queries led us to explore the area of language adaptation in query-by-example (QbE) retrieving. Therefore, our intention was to built a QbE retrieving system using all the available acoustic models trained solely on languages presented in the provided database.

2. SUPERVISED AUM APPROACH

The low-resource approach allowed us to use external resources (not related to QUESST task) for AUM and building acoustic models (AM) for the target languages in the provided database. We developed four language-dependent (LD) speech recognition systems, each represented by specific LD phonetic decoder and by an external well-trained LD phoneme-based GMM (Gaussian mixture model), trained with the corresponding phone-level transcription.

Four monolingual annotated datasets were used for acoustic model training: Slovak Speechdat (66 hours of read speech, 54 phonemes) [5], Czech Speechdat (89 hours of read speech, 42 phonemes) [5], Romanian anonymous speech corpus[1] (4.6 hours of read speech, 28 phonemes) and Portuguese (3 hours of BN recordings from COST278 DB [6], 1 hour of Laps Benchmark corpus from Fala Brasil project[2], 34 phonemes).

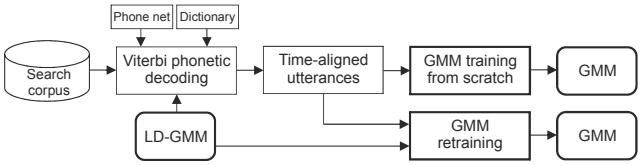

Figure 1: Supervised acoustic unit modeling framework.

The phonetic decoders and the LD AMs were intended to perform phonetic transcription and time alignment of search data utilizing the Viterbi algorithm. Each decoder employed a phone-level vocabulary and a phone network. The time-aligned utterances were used in the supervised training of the final multilingual GMM. In the presented work, we exploited two different ways of building multilingual GMMs.

The first one is oriented to training of a new GMM from scratch using the utterances and the time alignment, needed to initialize the GMM. The initialized GMMs were then simultaneously updated and expanded to higher mixtures (up to 1024 mixtures) using the Baum-Welch estimation procedure [9]. In this case, the external AM operates as an initial AM needed to bootstrap the recognition system, which is further supposed to proceed without an external input.

The second, improved training scheme is related to AM retraining[3]. The main idea is to re-estimate the acoustic likelihoods of the well-trained AM iteratively, using the utterances and the time alignments described above. We performed always three re-estimation cycles in the retraining to achieve the convergence of estimation. We found that the retraining brings higher precision over the standard training from scratch. The newly prepared language-dependent GMMs were used to generate posteriorgrams, which were finally fed to DTW-based search. The low-resource acoustic unit modeling is conceptually illustrated in Fig. 1. In the whole experimental setup, we used the standard 39-dim. MFCCs (Mel-Frequency Cepstral Coefficients).

[1] http://rasc.racai.ro
[2] http://www.laps.ufpa.br/falabrasil/
[3] characteristic for late submission

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

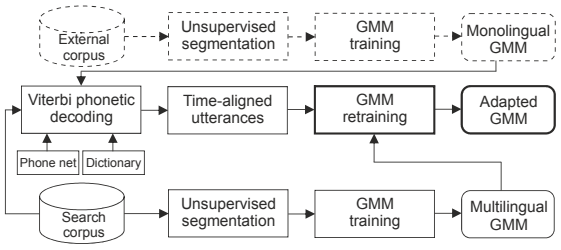

Figure 2: Unsupervised acoustic unit modeling framework.

3. UNSUPERVISED AUM APPROACH

The multilinguality problem and missing knowledge about acoustic units led us to employ two different unsupervised acoustic modeling approaches. For both types, we extracted an additional acoustic information about six spoken languages from Mediaeval 2013 and Mediaeval 2014 databases according to the available language tags.

The first type is focused on unsupervised building of acoustic model from unlabelled speech data. We re-employed our well-established procedures from the previous year and we built an acoustic model with up to 1024 mixtures for each language. This process included PCA-based voice activity detection [7], feature extraction and selection [2], K-means clustering ($K = 50$), Euclidean segmentation and GMM training, respectively [8]. This concept is depicted in Fig. 2 with dashed line. Each LD AM was intended to generate LD posteriorgrams, whereas the score results from all subsystems were finally fused together.

The second, advanced acoustic modeling technique used language adaptation[4] in acoustic (phonetic) sense performed through a retraining procedure similar to that used in low-resource modeling. The main idea is to use the already prepared LD AMs and feed them to phonetic decoding of search data in order to obtain LD time alignments. Inevitably, it was necessary to build an initial multilingual AM intended to adaptation, utilizing the same, already mentioned unsupervised segmentation and GMM training. The multilingual AM is then iteratively retrained on the LD time-aligned utterances using the search data (Fig. 2). Compared to the low-resource retraining, we retrained here the multilingual GMM instead of the language-specific GMM. The resulting six language-adapted GMMs practically match the probability distributions of the acoustic units of the specific language as good as possible.

4. POST-PROCESSING: SCORE NORMALIZATION AND FUSION

The average cumulative distance parameter (ACD), represented by mean value of cumulative distance matrix elements within each warping region and multiplied by $\alpha = 0.1$, was used as score parameter. Scaling the ACD parameter within the values 0-1 helped us to unify score ranges for the first 500 detection candidates per each query. Then the score fusion for different subsystems was carried out, employing a max-score merging strategy and z-normalization, similarly as we did last year [8]. The final set was obtained by keeping all fused detections per each query.

[4]characteristic for late submission

Table 1: Results for the primary supervised (p-S) and general unsupervised (g-U) systems (* late subm.)

system	*eval*		*dev*	
	C_{nxe} *(act/min)*	*TWV (act/max)*	C_{nxe} *(act/min)*	*TWV (act/max)*
p-S	**0.971**/0.953	**0.002**/0.022	**0.970**/0.947	**0.022**/0.036
g-U	**0.973**/0.953	**−0.01**/0.023	**0.974**/0.953	**0.0001**/0.031
p-S*	**0.963**/0.940	**0.046**/0.049	**0.962**/0.940	**0.055**/0.059
g-U*	**0.974**/0.954	**0.028**/0.032	**0.970**/0.951	**0.032**/0.035

Table 2: Processing resources measures

system	*ISF*	*SSF*	PMU_I(GB)	PMU_S (GB)	*PL*
p-S (dev)	2.312	0.0061	0.250	1.874	0.068
g-U (dev)	0.383	0.0066	0.515	2.292	0.033

5. RESULTS AND CONCLUSION

We submitted four runs obtained from supervised (primary) and unsupervised (general) systems, including late submissions, for QUESST 2015 task [4]. We did not perform evaluation with each individual type of query T1/T2/T3, but concentrated on the overall detection performance. The maximum number of Gaussian mixtures (GMs), we employed in both primary and general subsystems, was 256 for supervised and 64 for unsupervised AUM. Higher number of GMs did not bring any improvement.

The result obtained from all examined systems are far beyond our expectations (Tab. 1). It can be explained by the quality of audio data that were recorded in degraded acoustic conditions and influenced by background noises. The overall detection accuracy did not increase even though we examined various ways of speech enhancement techniques (DC offset removal, spectral subtraction, minimum mean squared error and Wiener filtering). Even the bottle-neck features developed at Brno University of Technology (BUT) [1] did not work well for our system.

Supervised AUM approach shows slightly better values of C_{nxe} and TWV in comparison with unsupervised AUM. The reason for relatively high performance of general approach is the AM adaptation employed in unsupervised AUM where all spoken languages are covered. Not significant improvement can be observed for late submission systems, that represent retraining procedure employed in AUM. However, the process of retraining did not perform well in case of supervised AUM for eval query set.

The robust statistical model-based speech enhancement methods embedded in the AUM and HMM-based speech segmentation will be investigated in the future. The processing load (PL) [3] for all systems, comprising development query set, is shown in Tab. 2. Considering the same searching set, the processing load is nearly identical for both dev and eval queries. It is obvious that the unsupervised AUM has the advantage of fast processing, mainly due to separate segmentation and acoustic modeling for each language.

6. ACKNOWLEDGMENTS

This publication is the result of the Project implementation: University Science Park TECHNICOM for Innovation Applications Supported by Knowledge Technology, ITMS: 26220220182, supported by the Research & Development Operational Programme funded by the ERDF (100%).

7. REFERENCES

[1] F. Grézl, M. Karafiát, S. Kontár, and J. Černocký. Probabilistic and bottle-neck features for LVCSR of meetings. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, pages 757–760. IEEE Signal Processing Society, 2007.

[2] J. Juhár and P. Viszlay. Linear Feature Transformations in Slovak Phoneme-Based Continuous Speech Recognition. In *Modern Speech Recognition Approaches with Case Studies*, pages 131–154. InTech Open Access, 2012.

[3] L. J. Rodriguez-Fuentes and M. Penagarikano. Mediaeval 2013 spoken web search task:system performance measures. Technical report, Software Technologies Working Group (GTTS, http://gtts.ehu.es), 2013.

[4] I. Szoke, L. J. Rodriguez-Fuentes, A. Buzo, X. Anguera, J. Metze, F. Proenca, M. Lojka, and X. Xiong. Query by Example Search on Speech at Mediaeval 2015. In *Working Notes Proc. of the MediaEval 2015 Workshop*, Germany, Wurzen, 14-15 September 2015.

[5] H. van den Heuvel et al. SpeechDat-E: five eastern european speech databases for voice-operated teleservices completed. In *Proc. of INTERSPEECH*, pages 2059–2062, 2001.

[6] A. Vandecatseye et al. The COST278 pan-european broadcast news database. In *Proc. of the 4th Intl. Conf. on Language Resources And Evaluation, LREC'04*, 2004.

[7] J. Vavrek et al. Query-by-Example Retrieval via Fast Sequential Dynamic Time Warping Algorithm. In *TSP 2014, Berlin, DE*, pages 469–473. IEEE, July 2014.

[8] J. Vavrek et al. TUKE System for MediaEval 2014 QUESST. In *Working Notes Proc. of the MediaEval 2014*, 2014.

[9] S. Young et al. *The HTK Book (for HTK Version 3.4)*. Cambridge University, 2006.