TUKE System for MediaEval 2014 QUESST

Jozef Vavrek, Peter Viszlay, Martin Lojka, Matúš Pleva, and Jozef Juhár
Laboratory of Speech Technologies in Telecommunications @ Technical University of Košice
Park Komenského 13, 041 20 Košice, Slovakia
{Jozef.Vavrek, Peter.Viszlay, Martin.Lojka, Matus.Pleva, Jozef.Juhar}@tuke.sk

ABSTRACT

Two approaches to QbE (Query-by-Example) retrieving system, proposed by the Technical University of Košice (TUKE) for the query by example search on speech task (QUESST), are presented in this paper. Our main interest was focused on building such QbE system, which is able to retrieve all given queries with and without using any external speech resources. Therefore we developed posteriorgram-based keyword matching system, which utilizes a novel weighted fast sequential variant of DTW (WFS-DTW) algorithm in order to detect occurrences of each query within the particular utterance file, using two GMM-based acoustic units modeling approaches. The first one, referred as low-resource approach, employs language-dependent phonetic decoders to convert queries and utterances into posteriorgrams. The second one, defined as zero-resource approach, implements combination of unsupervised segmentation and clustering techniques by using only provided utterance files.

1. MOTIVATION

The motivation for developing our system was to assess the ability of proposed WFS-DTW algorithm to detect various spoken query terms by implementing low and zero-resource posteriorgram-based matching approach.

2. WFS-DTW SEARCHING ALGORITHM

Searching algorithm for QUESST task follows the one used in our paper [8]. Proposed solution is a modification of segmental DTW algorithm we applied in spoken web search task last year [7]. There are three main contributions to this algorithm: 1) one step forward moving strategy, when each DTW search is carried out sequentially, block by block, with size equal to the length of query; 2) linear time-aligned accumulated distance for speeding up sequential DTW without considerable loss in retrieving performance; 3) optimization of global minimum for set of alignment paths by implementing weighted cumulative distance (WCD) parameter.

3. LOW-RESOURCE APPROACH

The low-resource approach includes 4 language-dependent subsystems, each represented by GMM-based acoustic model. The acoustic models were trained previously using four databases: 2× Speechdat (Slovak, 66h and Czech, 89h) [6], Slovak ParDat1 (40h) [3] and English TIMIT (10h) [4].

The well-trained models were intended to generate time-aligned and labelled segments for each utterance through Viterbi decoding. The phonetic decoder employed a phone-level vocabulary and a phone network. We found that the phoneme insertion log probability p in Viterbi segmentation has significant impact to time-alignment. Since the best results were obtained with $p = 0$, we used this value in the whole setup. The time-alignments were used to train a new GMM-based acoustic model using the development data. It means that each language-dependent model was replaced by its refined version, which was finally used to generate the posteriorgrams for utterances and queries.

Note that we used 39-dimensional MFCC (Mel-Frequency Cepstral Coefficients) features for Viterbi segmentation and GMM training. In low-resource approach we did not need any voice activity detector (VAD) because the silent parts of the audio stream were identified in the Viterbi segmentation.

4. ZERO-RESOURCE APPROACH

In keeping with the zero-resource approach, we did not assume any prior knowledge of the acoustic units or pronunciation lexicon. In order to train the acoustic models, it was firstly necessary to identify the acoustic speech units in the audio data automatically. In this work, we utilized four different zero-resource approaches to address this problem.

Type 1: This one uses a PCA-based VAD to discriminate the voice active segments from the silent ones [8]. The initial feature selection, based on simple PCA (principal component analysis) [5], is carried out after extracting first 13 MFCCs. Only those speech active feature vectors are selected, whose variance achieves values greater than 90% at the first principal component. Then, K-means clustering with $K = 75$ clusters and correlation distance metric is computed on the reduced data. The clustering starts by selecting K points uniformly. Finally, speech segmentation is performed by computing the squared Euclidean distance between feature vectors and K mean vectors, where the label of the mean vector with minimum distance is assigned in collaboration with VAD.

Type 2: Type 2 approach comes directly out from the Type 1 and is further extended by Viterbi segmentation and new GMM training. These two steps are identical to those already described in Section 3. The main difference is that the acoustic model from the Type 1 is used to generate the time-alignments through Viterbi segmentation.

Type 3: The third approach is based on the well-known *flat start* training procedure [9]. It does not need any seg-

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

Table 1: Evaluation of primary low-resource (p-low) and general zero-resource (g-zero) systems (* indicates late submission)

system	*eval*		*dev*	
	C_{nxe} *(act/min)*	*TWV (act/max)*	C_{nxe} *(act/min)*	*TWV (act/max)*
p-low	**0.959**/0.891	**0.154**/0.154	**0.960**/0.892	**0.161**/0.162
g-zero	**0.973**/0.934	**0.075**/0.077	**0.974**/0.934	**0.091**/0.091
p-low*	**0.947**/0.853	**0.168**/0.169	**0.948**/0.854	**0.191**/0.191
g-zero*	**0.970**/0.921	**0.102**/0.103	**0.971**/0.922	**0.106**/0.107

mentation or clustering because the utterances are uniformly segmented using the Baum-Welch embedded re-estimation. Therefore, an alternative GMM initialization strategy is applied, where all phone models are initialized identically with state means and variances equal to the global mean and variance. The phone models are then moved straight to embedded training and simultaneously updated and expanded to the higher GMs (Gaussian Mixtures) [9]. The key element in flat start training is the phone-level transcription, obtained from the phone-based recognition using the acoustic model acquired from the first type zero-resource approach.

Type 4: Type 4 approach implements GMM-based segmentation and ergodic HMM (EHMM) training. Firstly, an unsupervised GMM training is performed on whole database, where each acoustic unit is represented by one GM. Each GM is then associated with one of the 64 states in EHMM and new GMs for each acoustic unit are trained iteratively.

Note that we used conventional 39-dimensional MFCCs for each zero-resource processing (except the Type 1). We did not use any VAD here (except the Type 1) because the `<sil>` labels were available from the Viterbi segmentation.

5. POST-PROCESSING: SCORE NORMALIZATION AND FUSION

Score parameter was represented by WCD, normalized by scaling factor 0/1, similarly as we used in [8]. This step helped us to unified score ranges for the first 500 detection candidates per each query. Then the score fusion for four different subsystems was carried out, employing a simple max-score merging strategy, similarly as Anguera et al. did in [1]. Detection candidates from each individual subsystem were merged together, keeping the one with the highest score in case of overlap. Merged candidates for each query were subsequently normalized by z-normalization and aligned according to the score value. The final set was obtained by keeping first 45-150 candidates, according to the length of query (the shorter query the lower number of candidates).

6. RESULTS AND CONCLUSION

We submitted four runs obtained from low-resource (primary) and zero-resource (general) systems for QUESST 2014 task [2]. The primary systems employ language-dependent acoustic modeling using Viterbi segmentation with 128 GMs (ParDat1, TIMIT) and 256 GMs (Speechdat SK, CZ). The general systems use 32 GMs for Type 1,2,3 and 64 GM for Type 4. The best-one-win strategy was used at first runs (on time). Thus, only the subsystem with best performance was submitted, namely p-low using Speechdat SK and g-zero Type 2 subsystem. Late submissions include max-score merging fusion of four subsystems for both primary and general approaches. Results in Tab. 1 show that there are still big differences in performance between p-low and g-zero approaches, even if the score fusion technique was applied. Even more, there is also considerable gap between *act* and *min* C_{nxe} despite the fact that the *act* and *max* TWV are perfectly calibrated. Therefore, an improved calibration/fusion models based on affine transformation and linear-regression will be investigated in the future.

Table 2: Processing resources measures

system	*ISF*	*SSF*	PMU_I	PMU_S	*PL*
p-low (dev)	0.61	0.0034	0.05	2.46	0.0106
g-zero (dev)	1.5	0.0042	1.4	3.92	0.225

The indexing was done using 2xIBM x3650 (Intel E5530 @ 2.4 GHz, 8 cores), 28 GB RAM, under Debian OS. Searching algorithm was running on 52xIBM dx360 M3 cluster (Intel E5645 @ 2.4GHz, 624 cores), 48 GB RAM per node, running on Scientific Linux 6 and Torque (see Tab. 2).

7. 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%).

8. REFERENCES

[1] X. Anguera et al. The Telefonica Research Spoken Web Search System for MediaEval 2013. In *Working Notes Proc. of the MediaEval 2013*, 2013.

[2] X. Anguera, F. Metze, A. Buzo, I. Szoke, and L. J. Rodriguez-Fuentes. Query by Example Search on Speech at Mediaeval 2014. In *Working Notes Proc. of the MediaEval 2014 Workshop*, Barcelona, Spain, 16-17 October 2014.

[3] S. Darjaa et al. Rule-based Triphone Mapping for Acoustic Modeling in Automatic Speech Recognition. In *Proc. of the 14th Intl. Conf. on Text, Speech and Dialogue, TSD'11*, pages 268–275, 2011.

[4] J. S. Garofolo et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1993. Linguistic Data Consortium, Philadelphia.

[5] 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.

[6] 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.

[7] J. Vavrek et al. TUKE at MediaEval 2013 Spoken Web Search Task. In *Working Notes Proc. of the MediaEval 2013*, 2013.

[8] 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.

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