The NNI Query-by-Example System for MediaEval 2015*

Jingyong Hou[1], Van Tung Pham[2], Cheung-Chi Leung[3], Lei Wang[3], Haihua Xu[2], Hang Lv[1], Lei Xie[1], Zhonghua Fu[1], Chongjia Ni[3], Xiong Xiao[2], Hongjie Chen[1], Shaofei Zhang[1], Sining Sun[1], Yougen Yuan[1], Pengcheng Li[1], Tin Lay Nwe[3], Sunil Sivadas[3], Bin Ma[3], Eng Siong Chng[2], Haizhou Li[2,3]

[1]School of Computer Science, Northwestern Polytechnical University (NWPU), Xi'an, China
[2]Nanyang Technological University (NTU), Singapore
[3]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
jyhou@nwpu-aslp.org, VANTUNG001@e.ntu.edu.sg, ccleung@i2r.a-star.edu.sg, lxie@nwpu.edu.cn

ABSTRACT

This paper describes the system developed by the NNI team for the Query-by-Example Search on Speech Task (QUESST) in the MediaEval 2015 evaluation. Our submitted system mainly used bottleneck features/stacked bottleneck features (BNF/SBNF) trained from various resources. We investigated noise robustness techniques to deal with the noisy data of this year. The submitted system obtained the actual normalized cross entropy (actCnxe) of 0.761 and the actual Term Weighted Value (actTWV) of 0.270 on all types of queries of the evaluation data.

1. INTRODUCTION

This year's data is more challenging in terms of acoustic and noise conditions [1]. Noise robustness techniques, including adding noise to the training data of tokenizers and a speech enhancement method, were investigated to deal with the noisy data. Our submitted system involves dynamic time warping (DTW) and symbolic search (SS) based approaches as last year. This year, the final submitted system was obtained by fusing 66 systems from our 3 groups, including 15 DTW systems (selected from 26 original systems using FoCal toolkit [2]) from NWPU, 39 DTW systems from I^{2}R, and 8 DTW and 4 SS systems from NTU. Moreover, various voice activity detection (VAD) methods were used in the DTW systems.

2. ADDING NOISE TO TRAINING DATA

To reduce the mismatch problem between the training data of tokenizers and this year's development and test data, noise was added to the training data. We used two methods to obtain two sets of noise from the development data. The method used to obtain the first set of noise (noise1) is summarized as follows [3, 4, 5]:

- Perform voice/unvoice detection on the development data and obtain segments of noise from the utterance.
- Estimate the noise power spectrum of each utterance and generate minimum phase signal according to the power spectrum of each sentence and design the minimum phase filter.
- Use EM algorithm to estimate the parameters of the noise amplitude distribution (empirically select Gaussian distribution and set the number of Gaussian mixtures to 2).
- Generate a random white noise with the target noise amplitude distribution.
- Filter the random white noise using the minimum phase filter.

The second set of noise (noise2) was also estimated from the development data by using a method in [6]. The time domain noise was reconstructed by inverse short-time Fourier transform of the estimated instantaneous noise spectrum. Please refer to [7, 8] for details.

When noise was added, we had to ensure that the signal-to-noise ratio (SNR) distribution of the resultant training data was similar to that of this year's development data. Moreover, since not all the utterances in this year were highly noisy or reverberated, we only added noise to randomly selected 50 percent of training data.

3. SPEECH ENHANCEMENT

A Wiener filter [9] was used to reduce the noise in the data. The noise was reduced in the time domain and the enhanced data was used for VAD and feature extraction. Initial results (detailed in section 8) showed that the enhanced data led to better DTW performance for some tokenizers.

4. VOICE ACTIVITY DETECTION

For exact matching DTW systems, we used two voice activity detectors (VADs), including a frequency band energy based VAD [10] (VAD1) and a statistical model based VAD [11] (VAD2), because we found that they performed the best in different types of queries. For phoneme-sequence based approximate matching DTW systems (detailed in section 5) with phoneme posterior features, we used their single-best decoding hypotheses to perform VAD and obtain phoneme boundary information. For a phoneme-sequence approximate matching DTW systems with SBNF, we simply borrowed the single-best decoding hypothesis of a phoneme recognizer to perform VAD and obtained the phoneme boundary information.

5. DTW SEARCH

*This work was partially supported by the National Natural Science Foundation of China (61175018 and 61571363).

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

Exact matching and approximate matching DTW systems were developed to deal with different types of queries. An exact matching system matched each query with a subsequence of a test utterance using DTW [12, 13]. It found a path on the cosine distance matrix of the speech feature of the query and the test utterance. The system output the similarity score between the query and the matched subsequence of the test utterance.

We used two different kinds of approximate matching DTW systems in total, including fixed-window [12, 14] and phoneme-sequence [15] approximate matching systems, to deal with type 2 and type 3 queries. In fixed-window approximate matching systems, when the window was shifted, the corresponding segment of the query was matched with a test utterance. The highest similarity score which corresponds to a query segment and the test utterance was used as the score of the query-utterance pair of the system. The window sizes were set between 70 and 90 frames and the window shifts were set between 5 and 10 frames. In phoneme-sequence approximate matching systems, the size of the window was determined by the phoneme boundary information derived from phoneme recognizers. The window size was set to 8 phonemes, as it provided best results on the development data.

6. SYMBOLIC SEARCH

Weighted finite state transducer (WFST) based symbolic search systems were used as last year [12]. Phoneme-sequence approximate matching [14] was used to faciliate type 2 and type 3 queries, and to reduce the miss rate. A sequence length of 6 phonemes was chosen, as it provided best matching results on the development data.

7. TOKENIZERS AND SYSTEMS

Spectral features, phoneme-state posterior features and BNF/SBNF were used in our DTW systems.

NWPU extracted truncated PLP [16] (a1), posterior features from 3 BUT phoneme recognizers [17] (Czech, Hungarian and Russian; a2-a4), 3 sets of SBNF (1 being monophone state using original training data and 2 being triphone state with noise1 and noise2 added in training data respectively; a5-a7) trained from the English Switchboard corpus (SWBD), and 1 set of triphone state SBNF (a8) trained from the SEAME corpus [18].

I^{2}R extracted 4 sets of BNF (b1-b4) and 4 sets of SBNF (b5-b8) trained from four LDC corpora (SWBD, Fisher Spanish, HKUST Mandarin and CallHome Egyptian), and 5 sets of BNF (b9-b13) (4 language-dependent and one language-independent [19]) trained from 4 development languages in the OpenKWS evaluation [20].

NTU extracted 3 sets of BNF (c1-c3) trained (1 being triphone state with original training data and 2 being triphone state with Noisex92 [21] added in training data once and twice respectively) from SWBD, and 1 set of BNF (c4) trained from the 6 development languages in the OpenKWS evaluation.

NWPU's 26 DTW systems consisted of 9 exact matching systems (using a1-a8, c4) and 4 phoneme-sequence approximate matching systems (using a2-a4, a6). The rest 13 systems were exactly the same as the previous 13 systems except the enhanced data was used in VAD and feature extraction.

Table 1: Performance gain of an exact matching DTW system on the development set when different data (s1: original SWBD data; s2: noise1 is added; s3: noise2 is added) is used to train a tokenizer. The tokenizer is used to extract triphone state SBNF. Result Form: minCnxe, maxTWV

	All	T1	T2	T3
s1	0.891,0.111	0.762,0.227	0.934,0.024	0.918,0.093
s2	**0.875,0.133**	**0.733**,0.258	0.925,**0.041**	**0.901,0.101**
s3	0.877,0.132	0.735,**0.270**	**0.923**,0.038	0.907,0.095

Table 2: Performance on different types of queries in development and evaluation datasets.

	dev	eval
	All(T1, T2, T3)	All(T1, T2, T3)
actCnxe	0.773(0.629,0.813,0.829)	0.761(0.609,0.854,0.783)
minCnxe	0.757(0.601,0.793,0.810)	0.747(0.577,0.831,0.769)
actTWV	0.286(0.439,0.203,0.200)	0.270(0.436,0.189,0.203)
maxTWV	0.286(0.447,0.208,0.205)	0.274(0.444,0.194,0.215)

I^{2}R's 39 DTW systems consisted of 13 exact matching systems (using b1-b13) and 13 fixed-window approximate matching systems (using b1-b13) with VAD1, and 13 exact matching systems (using b1-b13) with VAD2.

NTU's 12 systems consisted of 4 exact matching (using c1-c4) and 4 fixed-window approximate matching (using c1-c4) DTW systems with VAD1, and 4 phoneme-sequence approximate matching SS systems with 4 acoustic models trained from SWBD and a Malay speech corpus [22].

The scores of all systems in each group were fused to a single system internally and the 3 resultant systems were further fused to obtain the final submitted system. In each fusion step, scores were first normalized to zero mean and unit variance, and then fused with the FoCal toolkit [2].

8. RESULTS AND CONCLUSION

Table 1 shows the performance gain of an exact matching DTW system on the development set when noise1 and noise2 were added to the SWBD data for training triphone SBNF. The results show that adding the noise to the training data gives 1.8% relative improvement on all query types and 3.8% relative improvement on type 1 queries in minCnxe.

When the enhanced data was used to extract SWBD monophone SBNF, BUT Czech and Hungarian phoneme-state posterior features for our DTW systems, we observed relative improvements of 1.9-3.1% on all query types and relative improvements of 2.7-6.3% on type 1 queries in minCnxe.

Table 2 shows the performance of our final submitted system on this year's data. In the intra-group fusion, each group experienced performance gains by fusing exact matching and approximate matching systems, and fusing sytems using different speech preprocessing techniques and different tokenizers. Compared with our single best exact matching DTW system (s2 in table 1), system fusion brings around 13.5% relative improvement in minCnxe on the development data (all query types).

The peak memory usage (PMU) of all DTW systems is 1.45GB when 1 set of 30 dimensional SBNF are loaded, and the searching speed factor (SSF) is around 0.0044 in each DTW system. The PMU of all SS systems is 45GB, and the SSF is around 0.0012 in each SS system.

We adopted noise robustness techniques to deal with the noise condition of data, which led to better search performance. We also experienced performance gains by fusing systems using different tokenizers, different VADs and different search algorithms.

9. REFERENCES

[1] I. Szoke, L. J. Rodriguez-Fuentes, A. Buzo, X. Anguera, F. Metze, J. Proenca, M. Lojka, and X. Xiao, "Query by example search on speech at mediaeval 2015," *Working Notes Proceedings of the MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany*, 2015.

[2] N. Brümmer, "FoCal: Toolkit for Evaluation, Fusion and Calibration of statistical pattern recognizers," https://sites.google.com/site/nikobrummer/focal.

[3] W. Yao and T. Yao, "Analyzing classical spectral estimation by MATLAB," *Journal of Huazhong University of Science and Technology*, vol. 4, p. 021, 2000.

[4] M. H. Hayes, J. S. Lim, and A. V. Oppenheim, "Signal reconstruction from phase or magnitude," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 6, pp. 672–680, 1980.

[5] M. H. Gruber, "Statistical digital signal processing and modeling," *Technometrics*, vol. 39, no. 3, pp. 335–336, 1997.

[6] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1218–1234, 2006.

[7] J. Chen, Y. Huang, and J. Benesty, "Filtering techniques for noise reduction and speech enhancement," in *Adaptive Signal Processing.* Springer, 2003, pp. 129–154.

[8] E. J. Diethorn, "Subband noise reduction methods for speech enhancement," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems.* Springer, 2004, pp. 91–115.

[9] J. Chen, J. Benesty, Y. Huang, and T. Gaensle, "On single-channel noise reduction in the time domain," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.* IEEE, 2011, pp. 277–280.

[10] E. Cornu, H. Sheikhzadeh, R. L. Brennan, H. R. Abutalebi, E. C. Tam, P. Iles, and K. W. Wong, "ETSI AMR-2 VAD: evaluation and ultra low-resource implementation," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 2. IEEE, 2003, pp. II–841.

[11] M. Huijbregts and F. De Jong, "Robust speech/non-speech classification in heterogeneous multimedia content," *Speech Communication*, vol. 53, no. 2, pp. 143–153, 2011.

[12] P. Yang, H. Xu, X. Xiao, L. Xie, C.-C. Leung, H. Chen, J. Yu, H. Lv, L. Wang, S. J. Leow *et al.*, "The NNI query-by-example system for MediaEval 2014," *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, Oct*, pp. 16–17, 2014.

[13] A. Muscariello, G. Gravier, and F. Bimbot, "Audio keyword extraction by unsupervised word discovery," in *INTERSPEECH 2009: 10th Annual Conference of the International Speech Communication Association*, 2009.

[14] H. Xu, P. Yang, X. Xiao, L. Xie, C.-C. Leung, H. Chen, J. Yu, H. Lv, L. Wang, S. J. Leow *et al.*, "Language independent query-by-example spoken term detection using n-best phone sequences and partial matching," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on.* IEEE, 2015, pp. 5191–5195.

[15] J. Hou, L. Xie, P. Yang, X. Xiao, C.-C. Leung, H. Xu, L. Wang, H. Lv, B. Ma, E. S. Chng, and H. Li, "Spoken term detection technology based on DTW(to be published)," *Journal of Tsinghua University (Sci and Tech)*, 2015.

[16] A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, S. Khudanpur, K. Church, N. Feldman, H. Hermansky, F. Metze, and R. Rose, "A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.* IEEE, 2013, pp. 8111–8115.

[17] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2006 IEEE International Conference on.* IEEE, 2006, pp. 325–328.

[18] D. C. Lyu, T. P. Tan, E. Chng, and H. Li, "SEAME: a Mandarin-English code-switching speech corpus in South-East Asia." *INTERSPEECH 2010: 11th Annual Conference of the International Speech Communication Association*, 2010.

[19] K. Vesely, M. Karafiát, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Spoken Language Technology Workshop (SLT), 2012 IEEE.* IEEE, 2012, pp. 336–341.

[20] "Open keyword search 2015 evaluation," http://www.nist.gov/itl/iad/mig/openkws15.cfm.

[21] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[22] T. Tan, X. Xiao, E. K. Tang, E. S. Chng, and H. Li, "MASS: A Malay language LVCSR corpus resource," in *Speech Database and Assessments, 2009 Oriental COCOSDA International Conference on.* IEEE, 2009, pp. 25–30.