PAIRWISE LEARNING USING MULTI-LINGUAL BOTTLENECK FEATURES FOR LOW-RESOURCE QUERY-BY-EXAMPLE SPOKEN TERM DETECTION

Yougen Yuan[1,2], *Cheung-Chi Leung*[2], *Lei Xie*[1*], *Hongjie Chen*[1], *Bin Ma*[2], *Haizhou Li*[2,3]

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]Department of ECE, National University of Singapore, Singapore

ABSTRACT

We propose to use a feature representation obtained by pairwise learning in a low-resource language for query-by-example spoken term detection (QbE-STD). We assume that word pairs identified by humans are available in the low-resource target language. The word pairs are parameterized by a multi-lingual bottleneck feature (BNF) extractor that is trained using transcribed data in high-resource languages. The multi-lingual BNFs of the word pairs are used as an initial feature representation to train an autoencoder (AE). We extract features from an internal hidden layer of the pairwise trained AE to perform acoustic pattern matching for QbE-STD. Our experiments on the TIMIT and Switchboard corpora show that the pairwise learning brings 7.61% and 8.75% relative improvements in mean average precision (MAP) respectively over the initial feature representation.

Index Terms— pairwise learning, bottleneck features, autoencoder, spoken term detection, low-resource speech processing

1. INTRODUCTION

Query-by-example spoken term detection (QbE-STD) is to search for the occurrence of a spoken query in audio archives [1, 2, 3]. Recently, many previous works have been investigated to extract unsupervised acoustic features directly in the target languages [4, 5, 6, 7], or extract posterior or bottleneck features (BNFs) from neural networks (NNs) trained using high-resource non-target languages [8, 9, 10, 11, 12].

In this paper, we propose to perform pairwise learning of NNs based on multi-lingual BNFs in a low-resource target language, and use the pairwise learned feature representation for QbE-STD. Training NNs with paired examples has been proposed for various tasks [13, 14, 15, 16, 17]. For speech in a language without any prior linguistic knowledge, it is difficult to give utterances with appropriate labels. However, it is easy for a native speaker to annotate whether the spoken words in two audio segments are the same. In our previous study [18], pairwise learning based on BNFs has been shown successful in a word discrimination task.

Multi-lingual BNFs are a kind of compact (low-dimensional) representations which can capture rich information to distinguish phonetic classes in multiple languages. They are more language-independent and have been commonly used in low-resource languages [19, 20, 21]. This kind of feature representation is derived from a multi-lingual bottleneck-type NN, in which the internal hidden layers are shared across multiple languages while the softmax layer is language-dependent. Experiments on automatic speech recognition (ASR) show that multi-lingual BNFs are more flexible for rapid language adaptation especially in low-resource languages. Moreover, multi-lingual BNFs have been used for QbE-STD [11]. To our best knowledge, this study is the first attempt to use pairwise learning based on multi-lingual BNFs, and use pairwise learning for QbE-STD.

Our proposed method was evaluated on the TIMIT and Switchboard corpora. To verify the effect of pairwise learning for QbE-STD, we tested our proposed pairwise learning by training an autoencoder (AE) with multi-lingual BNFs. The experiments showed that the multi-lingual BNFs far exceeded mel-frequency cepstral coefficients (MFCCs), and also outperformed cross-lingual BNFs. With pairwise learning, the resulted feature representations were much better than their original feature representations. Moreover, we investigated the effect of the amount of word pairs as supervision on our proposed feature representation, and investigated the effect of different features in AE training and frame alignment for QbE-STD.

2. METHODS

We assume that word pairs identified by humans are available in the low-resource target language. First, a multi-lingual bottleneck-type NN is trained[1]. Then, a deep AE is trained by using the extracted multi-lingual BNFs of frame aligned

This work was supported by the National Natural Science Foundation of China (Grant No. 61571363). * Corresponding author

[1]If only one language resource is used, it is a cross-lingual bottleneck-type NN.

pairs. Finally, this pairwise learned AE feature representation is used for QbE-STD.

2.1. Multi-lingual BNFs

In this paper, a multi-lingual BNF extractor is trained using transcribed data from other languages, and the extracted BNFs are used as an initial feature representation for pairwise learning. We summarize the training process of multi-lingual NN in Fig.1, and more details can be seen in [20, 22, 23]. The multi-lingual NN takes filter-bank with pitch features as input, and outputs phone posteriors for their corresponding language. The characteristic of this NN structure is that the internal hidden layers of the multi-lingual NN are language-independent, while its softmax layer is language-dependent. So all the information of language-dependent is concentrated in the softmax layer, and the rest of NN produces more language-independent feature representation. Moreover, there is one bottleneck layer in the language-independent hidden layers, and its outputs are exactly what we want to extract for pairwise learning in a new target language.

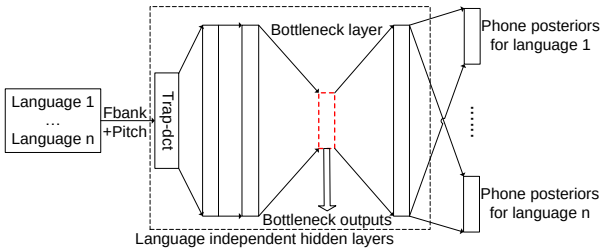

Fig. 1. Multi-lingual BNF extraction.

2.2. Pairwise learning

Pairwise learning is a good idea to capture information of same word pairs for learning an efficient feature representation without linguistic transcriptions. It maps a sequence of input features to a new sequence of feature representations with the same length. Since AE in [14, 18] has been shown successful in pairwise learning to obtain a good NN feature representation, this type of NNs is adopted in our experiments. The procedure of pairwise learning with multi-lingual BNFs is given in Fig.2. First, multi-lingual BNFs are extracted from a multi-lingual NN as an initial feature representation. Then, a stacked AE is trained in an unsupervised way directly on this initial NN feature representation using the mean squared error (MSE) as loss function. Next, frame alignment between a word pair is done by dynamic time warping (DTW). For each matching frame pair (a, a'), a and a' are presented as input and output respectively to fine-tune the previous trained stacked AE. Finally, our pairwise learned final NN feature representation is extracted from an internal hidden layer of the trained AE for QbE-STD.

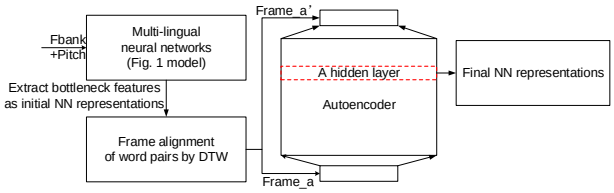

Fig. 2. Pairwise learning with an autoencoder.

2.3. QbE-STD

Fig.3 depicts the NN-based template matching method for QbE-STD, which involves feature extraction and DTW detection. In this paper, we extract features from an internal hidden layer of the pairwise learned AE, and use the subsequence-DTW (SDTW) algorithm described in [24] to search a spoken query in the test utterances. Given the acoustic feature representations $U = (u_1, u_2, ..., u_m)$ from a spoken query and $V = (v_1, v_2, ..., v_n)$ from a test utterance, a cosine distance between two feature vectors u_i and v_j is computed by:

$$D(i,j) = 1 - \frac{u_i^{\mathrm{T}} v_j}{|u_i||v_j|}, \tag{1}$$

where m and n denote the length of each corresponding sequence. In addition, the SDTW uses a dynamic programming (DP) algorithm to find an optimal path with the minimum distance cost, which is based on the matrix of cosine distances between any two feature vectors. Finally, according to these minimum distance costs for each spoken keyword U, all the test utterances are ranked in an ascending order, and three metrics are used to evaluate the detection scores.

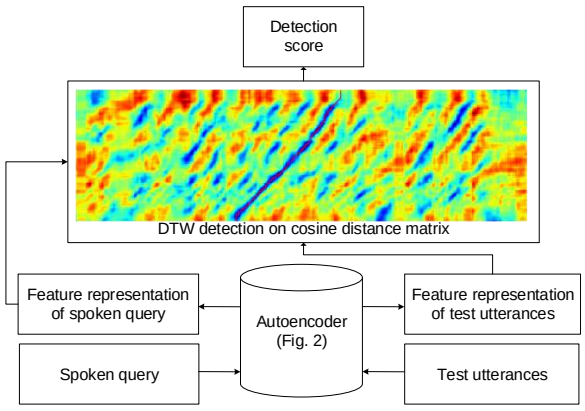

Fig. 3. NN-based template matching method for QbE-STD.

3. EXPERIMENTS

3.1. Data and Experimental Setup

To evaluate our proposed feature representation on speech recordings in different conditions, our QbE-STD experiments were conducted on the TIMIT and Switchboard corpora. On TIMIT, we followed the data setup as in [6, 7]. The training set consists of 10k word pairs, and each word has at least 5 English letters and 0.5 seconds. The keyword set consists of 346 spoken term examples, and each of them has at least 6 English letters and 0.35 seconds. The test set of 944 utterances is used as a speech archive. On Switchboard, we followed the data setup as in [14, 16, 18, 25], The training set consists of 100k word pairs, and each example has at least 5 English letters and 0.5 seconds. The keyword set also consists of 346 spoken term examples with the same requirement. The test set of 100 utterances is used as a speech archive.

We considered Mandarin Chinese and Spanish as high-resource non-target languages. A multi-lingual BNF extractor was trained by using about 170 hours of data from the HKUST Mandarin Chinese telephone speech corpus (LDC2005S15) and 152 hours of data from the Fisher Spanish telephone speech corpus (LDC2010S01), and each corpus was used to train a cross-lingual BNF extractor. We considered English as a low-resource target language in the TIMIT and Switchboard corpora. For multi-lingual or cross-lingual BNF extraction, the input features are 39-dimensional filterbank with pitch features. The model of multi-lingual NN used the configuration of 1500-1500-40-1500-[L_1...L_n], where n denotes the number of languages involved, and L_n is the number of tied triphone states of the corresponding language[2]. We followed the configuration of AE in [14, 18]. The AE consists of 13 hidden layers with 100 units in each layer. We extracted features from an internal hidden layer of the pairwise learned AE[3] as our proposed feature representation.

Three different evaluation metrics are used for QbE-STD evaluation as in [2, 5, 7]: (1) mean average precision (MAP), which is the mean of the average precision for each query in the test set; (2) the average precision of the top N utterances in the test set (P@N), where N is the number of target utterances containing the query term in the test set; (3) the average precision of the top five or ten utterances in the test set (P@5/P@10);

3.2. Comparison of feature representations

Table 1 and Table 2 show the evaluation results of different feature representations for QbE-STD on the TIMIT and Switchboard corpora. As illustrated in both tables, the multi-lingual BNFs are much better than MFCCs in terms of the three aforementioned evaluation metrics. This indicates that the information captured in the multiple languages for phone classification helps to learn an efficient acoustic feature representation for QbE-STD. Moreover, the multi-lingual BNFs usually outperform the cross-lingual BNFs. We believe that the multi-lingual BNFs capture more language-independent information, and they can provide a better feature representation than cross-lingual BNFs in the limited target language resources. In addition, the cross-lingual BNFs trained on Spanish outperform those trained on Mandarin, which is in line with our earlier results in a word discrimination task [18]. This demonstrates again that the cross-lingual BNFs are influenced by the selected language.

Table 1. *Comparison of different feature representations for QbE-STD on TIMIT. 10k word pairs are used for pairwise training.*

Representations	No pairwise training (MAP/P@N/P@10)	Pairwise training (MAP/P@N/P@10)
MFCCs	0.285/0.289/0.247	0.297/0.293/0.257
BNFs (Mandarin)	0.494/0.459/0.413	0.571/0.538/0.467
BNFs (Spanish)	0.540/0.512/0.446	**0.594/0.561**/0.484
BNFs (Multi-lingual)	**0.552/0.524/0.461**	**0.594/0.561/0.490**

Table 2. *Comparison of different feature representations for QbE-STD on Switchboard. 100k word pairs are used for pairwise training.*

Representations	No pairwise training (MAP/P@N/P@5)	Pairwise training (MAP/P@N/P@5)
MFCCs	0.232/0.200/0.232	0.258/0.236/0.260
BNFs (Mandarin)	0.370/0.338/0.446	0.417/0.382/0.451
BNFs (Spanish)	0.388/0.358/0.475	0.430/0.398/**0.484**
BNFs (Multi-lingual)	**0.400/0.365/0.485**	**0.435/0.404**/0.473

When pairwise supervision on the target language data is performed, the resulted feature representations have a significant improvement. This result indicates that pairwise learning provides a more efficient feature representation for QbE-STD. From both tables we can see that the pairwise learned feature representation based on the multi-lingual BNFs usually hold the best performance in the QbE-STD tasks.

3.3. Dependence on the amount of word-pair supervision

From the above results, we can find that pairwise learning always provides gain in QbE-STD tasks. To investigate the dependence on the amount of word-pair supervision, we varied the number of word pairs N=0.1k,1k,10k,100k (k=1000) by taking random subsets of the full training set on Switchboard as in [14, 18]. Fig.4 shows the QbE-STD results of different feature representations on Switchboard with different number of word pairs. We can find that the multi-lingual BNFs always outperform MFCCs with the same amount of word-pair supervision. This verifies that our proposed pairwise learning based on multi-lingual BNFs can consistently provide a better feature representation for QbE-STD. When more word

[2]In this paper, n equals 2, L1 equals 412 and L_2 equals 420.

[3]In this paper, the 9th layer of AE was used on TIMIT and Switchboard corpora.

Table 3. *Effect of different features in AE training and frame alignment. QbE-STD is performed on TIMIT.*

Input features of AE training \ Features for alignment	MFCC (MAP/P@N/P@10)	BNFs (Multi-lingual) (MAP/P@N/P@10)
MFCCs	0.285/0.289/0.247	0.320/0.314/0.274
BNFs (Multi-lingual)	0.587/0.556/0.486	**0.594/0.561/0.490**

Table 4. *Effect of different features in AE training and frame alignment. QbE-STD is performed on Switchboard.*

Input features of AE training \ Features for alignment	MFCC (MAP/P@N/P@5)	BNFs (Multi-lingual) (MAP/P@N/P@5)
MFCCs	0.258/0.236/0.260	0.273/0.248/0.286
BNFs (Multi-lingual)	0.432/0.395/**0.483**	**0.435/0.404**/0.473

pairs are given, the pairwise learned NN feature representation can get a better performance. In addition, with 10k word pairs (1/10 of the whole word pairs), the pairwise learned representations give comparable performance to those using all the word pairs. This indicates that we can learn an efficient feature representations with 10k word pairs, and it would be practical for the scenario when limited word pairs are available.

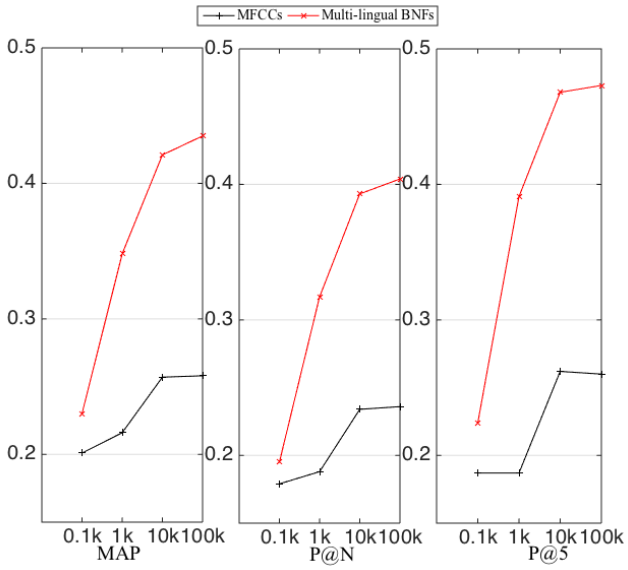

Fig. 4. Comparison of different feature representations with different number of word pairs in pairwise learning. QbE-STD is performed on Switchboard.

3.4. Effect of different features in AE training and frame alignment

We observed that using multi-lingual and cross-lingual BNFs for parameterization of word pairs brought more obvious improvement in pairwise learning than using MFCCs. This led us to investigate whether the more obvious improvement is because of using more efficient BNFs in the training of AE, or because of more accurate frame-level DTW alignment provided by BNFs. To investigate this issue, we performed pairwise learning with different combination of these features for frame-level DTW alignment and the training of AE. The QbE-STD results on the two corpora are shown in Table 3 and Table 4. Regardless of either MFCCs or multi-lingual BNFs are used for frame-level DTW alignment, multi-lingual BNFs consistently provide much better QbE-STD results than MFCCs when presented to the AE. This indicates the importance of using multi-lingual BNFs in the training of AE.

4. CONCLUSION

We have proposed to perform pairwise learning using multi-lingual BNFs of word pairs for QbE-STD. Pairwise learning facilitates supervision with data in a target language, even though no linguistic knowledge is available in the target language. Multi-lingual BNFs, which capture rich information of phonetic discrimination from other language resources, have been shown successful for QbE-STD. Pairwise learning makes the resulted feature representation more capable in phonetic discrimination for a new target language, which brings further performance improvement on low-resource QbE-STD tasks. In future work, we will investigate methods of word-level pairwise learning for this task, which avoids frame-level alignment of word pairs.

5. REFERENCES

[1] Alex S Park and James R Glass, "Unsupervised pattern discovery in speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[2] Timothy J Hazen, Wade Shen, and Christopher White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. ASRU*, 2009, pp. 421–426.

[3] Gautam Mantena, Sivanand Achanta, and Kishore Prahallad, "Query-by-example spoken term detection using frequency domain linear prediction and non-segmental Dynamic Time Warping," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 5, pp. 946–955, 2014.

[4] Yaodong Zhang and James R Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *Proc. ASRU*, 2009, pp. 398–403.

[5] Haipeng Wang, Cheung-Chi Leung, Tan Lee, Bin Ma, and Haizhou Li, "An acoustic segment modeling approach to query-by-example spoken term detection," in *Proc. ICASSP*, 2012, pp. 5157–5160.

[6] Peng Yang, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li, "Intrinsic spectral analysis based on temporal context features for query-by-example spoken term detection," in *Proc. INTERSPEECH*, 2014, pp. 1722–1726.

[7] Hongjie Chen, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li, "Unsupervised bottleneck features for low-resource query-by-example spoken term detection," in *Proc. INTERSPEECH*, 2016, pp. 923–937.

[8] Javier Tejedor et al., "Comparison of methods for language-dependent and language-independent query-by-example spoken term detection," *ACM Transactions on Information Systems*, vol. 30, no. 3, pp. 18, 2012.

[9] Luis J Rodriguez-Fuentes, Amparo Varona, Mikel Penagarikano, Germán Bordel, and Mireia Diez, "High-performance query-by-example spoken term detection on the SWS 2013 evaluation," in *Proc. ICASSP*, 2014, pp. 7819–7823.

[10] Yang Peng et al., "The NNI query-by-example system for mediaeval 2014," in *Proc. MediaEval Workshop*, 2014.

[11] Hou Jingyong et al., "The NNI query-by-example system for mediaeval 2015," in *Proc. MediaEval Workshop*, 2015.

[12] Cheung-Chi Leung et al., "Toward high-performance language-independent query-by-example spoken term detection for mediaeval 2015: post-evaluation analysis," in *Proc. INTERSPEECH*, 2016, pp. 3703–3707.

[13] Sumit Chopra, Raia Hadsell, and Yann LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. CVPR*, 2005, pp. 539–546.

[14] Herman Kamper, Micha Elsner, Aren Jansen, and Sharon Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *Proc. ICASSP*, 2015, pp. 5818–5822.

[15] Daniel Renshaw, Herman Kamper, Aren Jansen, and Sharon Goldwater, "A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge," in *Proc. INTERSPEECH*, 2015, pp. 3199–3203.

[16] Herman Kamper, Weiran Wang, and Karen Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *Proc. ICASSP*, 2016, pp. 4950–4954.

[17] Jonas Mueller and Aditya Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *Proc. AAAI*, 2016, pp. 2786–2792.

[18] Yougen Yuan, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li, "Learning neural network representation using cross-lingual bottleneck features with word-pair information," in *Proc. INTERSPEECH*, 2016, pp. 788–792.

[19] Ngoc Thang Vu, Florian Metze, and Tanja Schultz, "Multilingual bottle-neck features and its application for under-resourced languages," in *Proc. SLTU*, 2012.

[20] Karel Veselỳ, Martin Karafiát, František Grézl, Miloš Janda, and Ekaterina Egorova, "The language-independent bottleneck features," in *Proc. SLT*, 2012, pp. 336–341.

[21] Frantisek Grézl, Martin Karafiát, and Karel Vesely, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *Proc. ICASSP*, 2014, pp. 7654–7658.

[22] František Grézl, Martin Karafiat, and Miloš Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proc. ASRU*, 2011, pp. 359–364.

[23] Frantisek Grezl and Petr Fousek, "Optimizing bottle-neck features for LVCSR," in *Proc. ICASSP*, 2008, pp. 4729–4732.

[24] Armando Muscariello, Guillaume Gravier, and Frédéric Bimbot, "Audio keyword extraction by unsupervised word discovery," in *Proc. INTERSPEECH*, 2009, pp. 2843–2846.

[25] Aren Jansen, Samuel Thomas, and Hynek Hermansky, "Weak top-down constraints for unsupervised acoustic model training," in *Proc. ICASSP*, 2013, pp. 8091–8095.