EXTRACTING BOTTLENECK FEATURES AND WORD-LIKE PAIRS FROM UNTRANSCRIBED SPEECH FOR FEATURE REPRESENTATION

Yougen Yuan[1], *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]ECE Dept, National University of Singapore, Singapore

ABSTRACT

We propose a framework to learn a frame-level speech representation in a scenario where no manual transcription is available. Our framework is based on pairwise learning using bottleneck features (BNFs). Initial frame-level features are extracted from a bottleneck-shaped multilingual deep neural network (DNN) which is trained with unsupervised phoneme-like labels. Word-like pairs are discovered in the untranscribed speech using the initial features, and frame alignment is performed on each word-like speech pair. The matching frame pairs are used as input-output to train another DNN with the mean square error (MSE) loss function. The final frame-level features are extracted from an internal hidden layer of MSE-based DNN. Our pairwise learned feature representation is evaluated on the ZeroSpeech 2017 challenge. The experiments show that pairwise learning improves phoneme discrimination in 10s and 120s test conditions. We find that it is important to use BNFs as initial features when pairwise learning is performed. With more word pairs obtained from the Switchboard corpus and its manual transcription, the phoneme discrimination of three languages in the evaluation data can further be improved despite data mismatch.

Index Terms— pairwise learning, bottleneck features, word-like speech pairs, deep neural network (DNN), feature representation

1. INTRODUCTION

Most applications of using speech recognition technologies require a large amount of transcribed data together with language-specific linguistic knowledge [1]. However, manual annotation is expensive to acquire, and there are many languages in the world which have no written form at all. This prompts us to develop unsupervised learning techniques which usually involve unsupervised discovery of linguistic units in a target language, and these techniques have been shown successful in some downstream applications such as query-by-example spoken term detection (QbE-STD) [2, 3, 4, 5] and topic segmentation of spoken documents [6, 7, 8]. Moreover, the Zero Resource Speech Challenge has recently provided a benchmark to support this research domain [9].

In this paper, we propose to perform pairwise learning based on bottleneck features (BNFs) to obtain an unsupervised frame-level speech representation. We perform unsupervised discovery of phoneme-like units, and derive BNFs from a multilingual deep neural network (DNN) which is trained with the phoneme-like labels. We discover word-like speech pairs using the BNFs, and derive the matching frame pairs using DTW alignment on the word-like speech pairs. We train another DNN with the matching frame pairs as a weak supervision to obtain a new feature representation. Our pairwise learning framework is aimed to enhance the BNF representation by using the word-like speech pairs which are unsupervisedly discovered from clustering of short-term speech frames.

Pairwise learning takes paired examples as input, and maps examples belonging to the same class closer to each other in a fixed dimensional vector space. It has been successfully applied to various tasks [10, 11, 12], whose models are trained to specify whether two input examples are the same in the absence of predefined class labels. The present work is inspired by our previous studies [13, 5] which demonstrate that pairwise learning using cross-lingual or multilingual BNFs can learn more efficient feature representations for acoustic word discrimination and QbE-STD respectively. BNFs instead of spectral features are used in pairwise learning because of the better phoneme discrimination of BNFs. Moreover, this framework is practically useful in the situation where the number of paired examples are limited. In these previous studies, we enhance the BNF representation whose network parameters are trained from transcribed cross-lingual/multi-lingual data, and the paired examples of in-domain speech data are manually verified.

* Corresponding author

This work was supported by the National Natural Science Foundation of China (Grant No. 61571363) and the China Scholarship Council (Grant No. 201706290169).

In our present work, we extend the feasibility of this framework for unsupervised feature learning in the Zero Resource Speech Challenge 2017 (ZeroSpeech 2017[1]) as follows: 1) Since no manual transcription is available for the creation of a BNF extractor, we follow the work in [14] to obtain a BNF representation learned with phoneme-like labels. This type of BNFs have been shown comparable QbE-STD performance by using the cross-lingual BNFs derived from labeled data [4]. 2) As there are no manually identified word pairs in the dataset, we adopt an unsupervised discovery algorithm described in [15] to identify word-like speech pairs.

Our proposed feature representation is evaluated on the ABX phoneme discrimination task in ZeroSpeech 2017. To the best of our knowledge, the present work is the first attempt to use the BNFs learned with phoneme-like labels for pairwise learning. More importantly, this work demonstrates that the BNFs based on phoneme-like labels provide better phoneme discrimination than the pairwise learned features based on spectral features. It is important to use the BNFs as initial features when pairwise learning is performed. The experimental results show that our pairwise learned features obviously outperform the baseline features. Pairwise learning improves phoneme discrimination in 10s and 120s test conditions. With more word pairs obtained from the Switchboard corpus and its manual transcription, the phoneme discrimination of three languages in the evaluation data can further be improved despite data mismatch. We also investigate how the number of word-like pairs in pairwise learning affects the features' phoneme discrimination and the importance of using BNFs for pairwise learning.

2. PAIRWISE LEARNING FRAMEWORK

We extract BNFs and word-like pairs from untranscribed speech to perform pairwise learning for learning a better feature representation. The diagram of our pairwise learning framework is shown in Fig. 1. Each component is described in the following sections.

2.1. BNFs extraction

Since the Dirichlet process Gaussian mixture models (DPGMMs) outperform pairwise learning based on spectral features in the Zero Resource Speech Challenge 2015 [16, 17], We use language-dependent DPGMMs to obtain unsupervised phoneme-like labels on untranscribed speech, and train a bottleneck-shaped multilingual DNN with the unsupervised phoneme-like labels. The BNFs are extracted from a linear bottleneck layer in the multilingual DNN. We follow the steps provided in [14], and use BNFs as our initial features for pairwise learning.

[1]http://sapience.dec.ens.fr/bootphon/2017/index.html

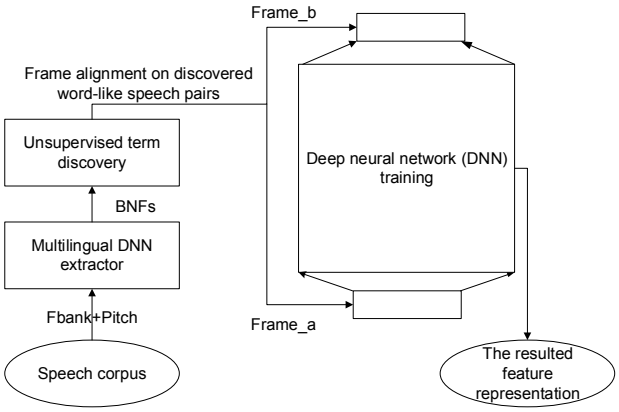

Fig. 1. The diagram of our pairwise learning framework.

2.2. Unsupervised term discovery

We adopt an unsupervised term discovery (UTD) algorithm described in [15] to find out word-like speech pairs from the untranscribed speech. However, we use BNFs instead of spectral features to parameterize the speech. Firstly, voice activity detection (VAD) is applied to discard non-speech frames. Then, locality sensitive hashing (LSH) is used to hash the BNFs of speech frames into bit signatures. Repeated acoustic patterns are searched on the distance matrix of the bit signatures using dynamic time warping (DTW). Finally, a post-processing step is used to select more precise word-like speech pairs according to two parameters: accumulated DTW distance and speech segments duration. In this paper, the unsupervised discovery process is implemented using the code provided for the track 2 of ZeroSpeech 2017[2]. The threshold values of the accumulated DTW distance and speech segment duration in the post-processing step are set as 0.90 and 50 frames to achieve the optimal phoneme discrimination of the pairwise learned features, which are tuned on the ZeroSpeech 2017 development datasets using mel-frequency cepstral coefficients (MFCCs), perceptual linear predictions (PLPs) and BNFs.

2.3. Pairwise learning with a DNN

The phoneme discriminability of our BNFs is possibly limited by the clustering of short-term spectral features in DPGMM. When word-like speech pairs are available, there is higher chance to discover repeated speech patterns from different speakers and multiple recordings of the same speakers [18]. We would use pairwise learning to further map the speech frames belonging to the sound class (derived from the discovered word-like speech pairs) close to each other in the final feature representation. We train another DNN with mean square error (MSE) criterion for pairwise learning, which

[2]https://github.com/bootphon/zerospeech2017/tree/master/track2

was first proposed by [19, 20]. In our previous studies [13, 5] which assume the availability of manual annotation of paired examples, we find that the pairwise learned features can be better if BNFs instead of the MFCCs are used as initial features. In the present work, a deep stacked autoencoder network [21] is layer-wise trained to initialize the parameters using BNFs. Given the discovered word-like speech pairs, frame alignment based on DTW is used to obtain the matching frame pairs. We use each aligned frame pair ($(Frame\_a, Frame\_b)$ in Fig. 1) as input-output to fine tune the initialized DNN with the MSE loss function. After finishing the training, the pairwise learned feature representation is obtained from an internal hidden layer of the MSE-trained DNN.

3. EXPERIMENTS

3.1. Data and setup

Our experiments were conducted on the development datasets provided by ZeroSpeech 2017. The development datasets consist of three languages including English, French and Mandarin. All the datasets contain read speech in 16 kHz, 16 bits and single channel. Table 1 lists the duration of each development dataset.

The multi-lingual DNN took filterbank and fundamental frequency (F0) features as the input, and it had the network topology of 1024-1024-1024-1024-40-1024-$[L_1, L_2, L_3]$, where $[L_1, L_2, L_3] = [1148, 1070, 451]$ and it represents the number of unsupervised phoneme-like labels discovered from the English, French and Mandarin training datasets respectively. Each language-dependent DPGMM was trained using mel-frequency cepstral coefficients (MFCC) with Δ + $\Delta\Delta$ (39-dimensional) post-processed by cepstral mean and variance normalization (CMVN).

The word-like speech pairs were discovered on the development (training) datasets (3,324 in English; 1,167 in French; 14 in Mandarin). Since the English training dataset has the largest amount of word-like speech pairs, we trained a DNN on this dataset. The DNN took 40-dimensional BNFs as input, and it was trained by minibatch stochastic gradient descent (SGD) [22]. We followed the configuration in [19, 13, 5], and implemented on Theano [23]. The DNN consisted of 13 hidden layers with 100 units in each layer. We initialized the DNN over 30 epochs per layer with the minibatch size of 256 and the learning rate of 0.00025. All the training frames were shuffled prior to each epoch of training. We fine-tuned the DNN over 120 epochs with the minibatch size of 256 and the learning rate of 0.002. The matching frame pairs were used twice as in [20] when fine-tuning the DNN. The features from the third layer of the pairwise trained DNN formed our proposed feature representation.

To evaluate the effectiveness of our proposed speech representation, we performed the ABX phoneme discrimination test using the track 1 evaluation toolkit provided in ZeroSpeech 2017. This test is to decide whether an unknown speech segment X is closer to the category A or the category B in terms of their DTW distance. The within- and across-talker error rates are reported for each representation. A lower error rate indicates a better performance. (see [9] for details).

Table 1. Development datasets in ZeroSpeech 2017.

Language	Duration (hrs) of Training set	Duration (hrs) of Test set
English	45	27
French	24	18
Mandarin	3	25

3.2. Comparison of different feature representations

Table 2 summarizes the within- and across-talker error rates of our proposed features, BNFs, and the baseline features (MFCCs) and the topline features (GMM-HMM posteriorgram; the language-specific model trained using speech data with manual transcription). The BNFs derived from untranscribed data, as the initial features in pairwise learning, show consistently lower error rates than MFCCs on the three languages. This observation is consistent with that in [4] for QbE-STD.

Our proposed features consistently bring further error reductions on top of BNFs in 10s and 120s test conditions. No error reduction is found in 1s test condition. This may be because a larger portion of speech frames in this condition (leading or trailing frames in the segments) cannot benefit from the long temporal context for more accurate data projection by the DNN. This may also lead to the fact that BNFs bring more error reductions in 10s and 120s test conditions than in 1s test condition over the baseline MFCCs.

Another interesting point is that, although our proposed features were learned only using the word-like speech pair in the English training data, we found that the knowledge of aligned frame pairs could bring error reductions in the test data of the other two languages. Although English is supposed to be acoustically different from the other two languages, similar error reductions in the test data of French and Mandarin were observed. It is encouraging to observe that the knowledge obtained in pairwise learning can be reused into other languages.

We also performed pairwise learning by using the real word pairs (98,957) provided from the Switchboard English corpus. As shown in the last row of Table 2, the within- and across-talk error rates in the test data of all the three languages are further reduced. Notice that there is data mismatch between the Switchboard English corpus and the datasets in ZeroSpeech 2017. We believe that our proposed features could provide more obvious error reductions on top of BNFs when

Table 2. Error rates (%) of ABX phoneme discrimination test in ZeroSpeech 2017. MFCCs are the Baseline and supervised GMM-HMM posteriorgrams are the Topline. BNFs are the initial features for pairwise learning. System 1 is a pairwise learning system with the word-like pairs discovered from English training dataset in ZeroSpeech 2017. System 2 is a pairwise learning system with the word pairs from Switchboard.

	Use manual labels		English			French			Mandarin		
			1s	10s	120s	1s	10s	120s	1s	10s	120s
Baseline (MFCC)	No	within talkers	12.0	12.1	12.1	12.5	12.6	12.6	11.5	11.5	11.5
		across talkers	23.4	23.4	23.4	25.2	25.5	25.2	21.3	21.3	21.3
Topline (GMM-HMM)	Yes	within talkers	6.5	5.3	5.1	8.0	6.8	6.8	9.5	4.2	4.0
		across talkers	8.6	6.9	6.7	10.6	9.1	8.9	12.0	5.7	5.1
BNFs	No	within talkers	8.6	7.4	7.3	11.5	9.7	9.6	10.6	8.6	8.4
		across talkers	14.0	12.4	12.3	18.1	15.7	15.2	12.4	11.0	10.9
System 1	No	within talkers	9.0	**7.1**	**7.0**	11.9	**9.5**	**9.5**	11.1	**8.5**	**8.2**
		across talkers	**14.0**	**11.9**	**11.7**	18.6	**15.5**	**14.9**	12.7	**10.8**	**10.7**
System 2	Yes	within talkers	8.9	7.1	6.9	11.9	9.3	9.1	11.2	8.5	8.1
		across talkers	13.6	11.5	11.2	17.7	14.8	14.4	12.8	10.6	10.4

more in-domain speech data were available to discover more word-like pairs.

3.3. Dependence of pairwise learning on the amount of word-pair supervision

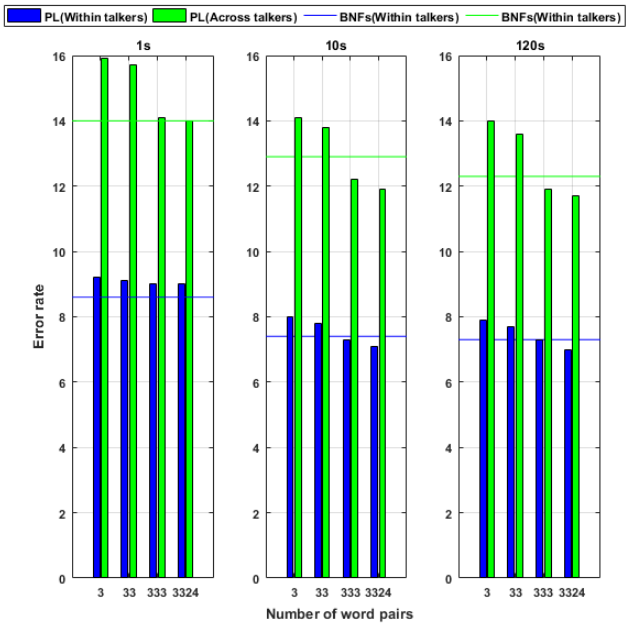

Fig. 2. Dependence of pairwise learning on the amount of word-pair supervision from English training dataset in ZeroSpeech 2017. PL represents pairwise learned features trained on the BNFs.

To investigate how the number of word-like pairs in pairwise learning affects the phoneme discrimination of our proposed features, we randomly selected subsets of 3, 33, 333 and 3,324 word-like pairs discovered from the English training data for pairwise learning. The results are shown in Fig 2. We observed that our proposed features could give better within- and across-talker phoneme discrimination when more word-like speech pairs were available. Unfortunately, our proposed features performed no better than BNFs when there were fewer than three hundred word-like pairs. However, together with the experiment result obtained using the real word pairs from the Switchboard corpus, we believe that the performance gain of our proposed features over BNFs could be increased when more in-domain speech data were available for the unsupervised discovery of word-like speech pairs.

3.4. Different features in pairwise learning

Our previous study [5] shows that pairwise learning using multi-lingual or cross-lingual BNFs derived from transcribed speech gives better features for QbE-STD than that using spectral features. We further compared between BNFs and MFCCs as initial features in pairwise learning where the BNFs in our present work were derived from untranscribed speech. In the comparison, the discovery of word-like pairs in the English training data and the frame-level DTW alignment in pairwise learning were performed using BNFs, the BNFs and MFCCs of the aligned frame pairs were presented to the DNN in pairwise learning. The within- and across-talker errors in the three test conditions are shown in Fig 3. As the observation in [5], we found that better pairwise learned features were obtained when BNFs were presented to the DNN as input features. Moreover, regardless of either BNFs or MFCCs were used for frame-level DTW alignment in pairwise learning, similar performance were obtained if BNFs were presented to DNN as input features. In addition, BNFs alone performed obviously better than the pairwise learned features based on spectral features. This would be attributed to the

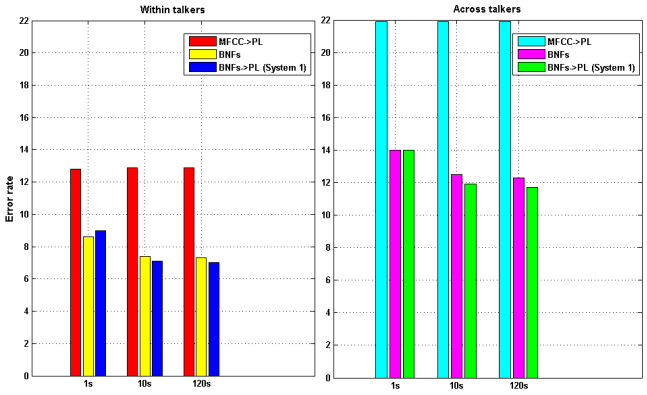

Fig. 3. Different initial features in pairwise learning. BNFs→PL and MFCCs→PL represent pairwise learned features based on the BNFs and MFCCs respectively. English training dataset in ZeroSpeech 2017 is used for word-pair supervision.

fact that not many speech frames (out of 310,510 frames) were involved in the creation of DNN in pairwise learning. On the other hand, it is worth noting that all speech frames in the training data were involved to derive phoneme-like labels and the BNF representation.

4. CONCLUSION

The experiment results suggest that it is effective to use BNFs which provide better phoneme discrimination than the spectral features as initial features for pairwise learning. Pairwise learning improves within- and across-talker phoneme discrimination mainly in 10s and 120s test conditions. When more real word pairs from the Switchboard corpus are used in pairwise learning, the phoneme discrimination can be further improved despite data mismatch. This probably indicates that the performance of our proposed features is limited by the small amount of word-like pairs discovered from the dataset in ZeroSpeech 2017, and we believe that their performance gain over the BNFs can be increased when more in-domain speech data is available for the unsupervised discovery of word-like pairs. Our future work will investigate the use of more sophisticated neural networks [24, 25] in pairwise learning to improve our proposed features.

5. REFERENCES

[1] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *Proc. ASRU*, 2009, pp. 398–403.

[3] H. Wang, C.-C. Leung, T. Lee, B. Ma, and H. Li, "An acoustic segment modeling approach to query-by-example spoken term detection," in *Proc. ICASSP*, 2012, pp. 5157–5160.

[4] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Unsupervised bottleneck features for low-resource query-by-example spoken term detection," in *Proc. INTERSPEECH*, 2016, pp. 923–927.

[5] Y. Yuan, C.-C. Leung, L. Xie, H. Chen, B. Ma, and H. Li, "Pairwise learning using multi-lingual bottleneck features for low-resource query-by-example spoken term detection," in *Proc. ICASSP*, 2017, pp. 5645–5649.

[6] I. Malioutov, A. Park, R. Barzilay, and J. Glass, "Making sense of sound: Unsupervised topic segmentation over acoustic input," in *Proc. ACL*, 2007, p. 504.

[7] L. Zheng, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Acoustic texttiling for story segmentation of spoken documents," in *Proc. ICASSP*, 2012, pp. 5121–5124.

[8] H. Chen, L. Xie, W. Feng, L. Zheng, and Y. Zhang, "Topic segmentation on spoken documents using self-validated acoustic cuts," *Soft Computing*, vol. 19, no. 1, pp. 47–59, 2015.

[9] M. Versteegh *et al.*, "The zero resource speech challenge 2015." in *Proc. INTERSPEECH*, 2015, pp. 3169–3173.

[10] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," in *Proc. NIPS*, 1994, pp. 737–744.

[11] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. CVPR*, 2005, pp. 539–546.

[12] G. Synnaeve, T. Schatz, and E. Dupoux, "Phonetics embedding learning with side information," in *Proc. SLT*, 2014, pp. 106–111.

[13] Y. Yuan, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Learning neural network representation using cross-lingual bottleneck features with word-pair information," in *Proc. INTERSPEECH*, 2016, pp. 788–792.

[14] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Multilingual bottle-neck feature learning from untranscribed speech," in *Proc. ASRU*, 2017.

[15] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *Proc. ASRU*, 2011, pp. 401–406.

[16] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Parallel inference of dirichlet process gaussian mixture models for unsupervised acoustic modeling: A feasibility study," in *Proc. INTERSPEECH*, 2015, pp. 3189–3193.

[17] M. Versteegh, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015: Proposed approaches and results," *Procedia Computer Science*, vol. 81, pp. 67–72, 2016.

[18] A. Jansen, S. Thomas, and H. Hermansky, "Weak top-down constraints for unsupervised acoustic model training," in *Proc. ICASSP*, 2013, pp. 8091–8095.

[19] H. Kamper, M. Elsner, A. Jansen, and S. Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *Proc. ICASSP*, 2015, pp. 5818–5822.

[20] D. Renshaw, H. Kamper, A. Jansen, and S. Goldwater, "A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge," in *Proc. INTERSPEECH*, 2015, pp. 3199–3203.

[21] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[22] L. Bottou, "Stochastic gradient learning in neural networks," *Neuro-Nımes*, vol. 91, no. 8, 1991.

[23] F. Bastien *et al.*, "Theano: new features and speed improvements," in *Proc. a deep learning workshop at NIPS*, 2012.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

[25] W. He, W. Wang, and K. Livescu, "Multi-view recurrent neural acoustic word embeddings," in *Proc. ICLR*, 2016.