MULTILINGUAL BOTTLE-NECK FEATURE LEARNING FROM UNTRANSCRIBED SPEECH

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

[1] Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]Department of Electrical and Computer Engineering, National University of Singapore, Singapore

ABSTRACT

We propose to learn a low-dimensional feature representation for multiple languages without access to their manual transcription. The multilingual features are extracted from a shared bottleneck layer of a multi-task learning deep neural network which is trained using unsupervised phoneme-like labels. The unsupervised phoneme-like labels are obtained from language-dependent Dirichlet process Gaussian mixture models (DPGMMs). Vocal tract length normalization (VTLN) is applied to mel-frequency cepstral coefficients to reduce talker variation when DPGMMs are trained. The proposed features are evaluated using the ABX phoneme discriminability test in the Zero Resource Speech Challenge 2017. In the experiments, we show that the proposed features perform well across different languages, and they consistently outperform our previously proposed DPGMM posteriorgrams which topped the performance in the same challenge in 2015.

Index Terms— Multi-task learning, multilingual feature, unsupervised feature learning, low/zero-resource

1. INTRODUCTION

Nowadays many state-of-the-art speech applications rely on a huge amount of transcribed speech and linguistic expertise, e.g. pronunciation dictionary. However, manual transcriptions and linguistic resources are expensive to acquire. Even worse, they are absent for some rare languages. Recently there is an increasing research interest in unsupervised speech processing, which usually involves unsupervised discovery of linguistic units [1, 2, 3], and the derived techniques have been used in different applications, such as retrieval of spoken queries in a speech database [4, 5, 6] and spoken document classification or/and clustering [7].

In this paper, we consider the learning of speech features in an unsupervised scenario, where only untranscribed speech is available for target languages and the linguistic knowledge about these languages is not available. The speech features widely studied in this research include posteriorgrams and the features derived from an internal layer of a deep neural network (DNN). Posteriorgrams can be derived from a Gaussian mixture model (GMM) [4, 8], sub-clustered GMM [9], unsupervised hidden Markov models (HMMs) [10, 11], or deep Boltzmann machine (DBM) [12]. When features are derived from an internal layer of a DNN, the DNN can be an autoencoder [13], a correspondence autoencoder [14, 15] or a siamese network [16]. Also the DNN can be trained to predict unsupervised labels [17]. In addition to the above frame-wise features, segment-wise features [18, 19, 20] which are represented by a fixed dimensional vector have been proposed recently, and they are usually associated with the discovery of word-like units.

In this paper, we aim to learn frame-wise speech features that support phoneme discriminability across multiple source languages. To accomplish this, we employ a Dirichlet process Gaussian mixture model (DPGMM) to perform phoneme-like unit modeling on each source language and tokenize the untranscribed speech into sequences of phoneme-like labels. We train a multi-task learning deep neural network (MTL-DNN) in which each task corresponds to a source language and is to predict the unsupervised phoneme-like labels of the source language. Our proposed multilingual features are extracted from a shared bottleneck layer in the MTL-DNN.

The proposed features are inspired by our previous works [8, 17] that support one source language: In [8], we propose to use Dirichlet process Gaussian mixture model (DPGMM) for unsupervised acoustic modeling, in which each Gaussian component aims to model a cluster of sounds from various speakers. DPGMM posteriorgrams are the unsupervised features which perform the best for ABX phoneme discrimination in Zero Resource Speech Challenge 2015, and can perform comparably to the posteriorgrams derived from language-mismatched phoneme recognizers. In [17, 21], we further use DPGMM to derive low-dimensional features by training a bottleneck-shaped DNN to predict the unsupervised DPGMM labels. The DPGMM-derived bottleneck features (BNFs) provide a more compact representation than the corresponding posteriorgrams, and they can perform comparably to cross-lingual BNFs (trained using transcribed data) for retrieval of spoken queries in a speech database. Note that our multilingual bottleneck feature learning is similar to [22] for speech recognition, but our multilingual DNN is trained to predict unsupervised phoneme-like labels instead of supervised labels. To the best of our knowledge, this is the first study to train a multilingual DNN using unsupervised phoneme-like labels, and we demonstrate that it is important to use language-dependent labeling in our feature learning for phoneme discrimination.

To evaluate our proposed multilingual BNFs, we conduct the ABX phoneme discriminability test [23, 24] in Zero Resource Speech Challenge 2017 (ZeroSpeech2017)[1], where the corpus consists of five languages. The ABX phoneme discriminability test only requires the generated features and a proper distance metric for the features, providing a straightforward way to measure the discriminability between two sound categories. There is no assumption

This work was supported by the National Natural Science Foundation of China (Grant No. 61571363) and the China Scholarship Council (Grant No. 201606291069). † Corresponding author

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

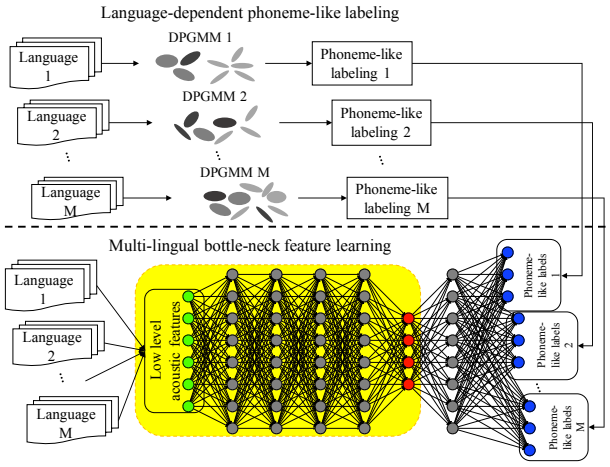


Fig. 1. Multilingual bottle-neck feature learning with language-dependent phoneme-like labeling for M languages. The low-dimensional shared features are extracted from the deep neural network in the dashed box.

on language-specific knowledge (e.g. number of phoneme units) in the generated features and the evaluation metric. Moreover, the test is not specific to a downstream application and can directly reflect the phoneme discriminability of the features, since the defects in the features are not mitigated by some application-specific postprocessing techniques. In our experiments, we report the ABX phoneme discriminability test of the proposed multilingual BNFs. In addition, we compare the multilingual BNFs with the cross-lingual BNFs derived from manual transcription. Meanwhile, we compare the multilingual BNFs with the monolingual BNFs learned using language-dependent phoneme-like labels. Furthermore, we investigate the effect of using language-dependent labeling. We also study the benefit of using bottleneck features over posteriograms, and the effect of vocal tract length normalization (VTLN) in unsupervised acoustic modeling.

2. FEATURE LEARNING FROM UNTRANSCRIBED SPEECH

The proposed feature learning technique is depicted in Fig. 1, which consists of two modules, 1) language-dependent phoneme-like labeling and 2) multilingual bottle-neck feature (BNF) learning. Specifically, language-dependent phoneme-like labeling is to transcribe the speech of each source language with phoneme-like unit labels using Dirichlet process Gaussian mixture model (DPGMM). Multilingual BNF learning is to learn a feature extractor via a multi-task learning deep neural network (MTL-DNN) where each task is to predict the language-dependent phoneme-like unit labels.

2.1. Language-dependent phoneme-like labeling

Many previous studies have shown that Dirichlet process Gaussian mixture model (DPGMM) is practically suitable for scenarios where no knowledge about the model complexity is available. For example, Kamper *et al.* [25] firstly have demonstrated that DPGMM is viable for unsupervised speech word clustering. With a parallel sampler [26], Chen *et al.* [8] have shown the feasibility to apply DPGMM in frame-wise speech clustering and obtained effective posteriorgrams (PGs) for the phoneme discriminability test in ZeroSpeech2015. When using DPGMM to cluster speech frames, one can regard DPGMM as a phoneme-like unit model where each Gaussian component models a phoneme-like unit. Therefore, we can employ DPGMM to transcribe the target untranscribed speech into sequences of phoneme-like labels.

Briefly, DPGMM is a Gaussian mixture model (GMM) extended in a non-parametric Bayesian way in which a Dirichlet process prior is placed over the vanilla GMM with a set of hyperparameters. We adopt a Metropolis-Hastings based split/merge sampler[2] to infer DPGMM parameters. For more in-depth model explanation, please refer to [27]. For detailed information about the sampler used in this work, please refer to [26].

Given M source languages, the speech feature vectors of the m-th source language are $\mathcal{X}^{(m)} = \{\mathbf{x}_i^{(m)}\}_{i=1}^{N}$. After training the DPGMM, we obtain $K^{(m)}$ Gaussian components together with their mixture weights, $\boldsymbol{\pi}^{(m)}$, mean vectors, $\boldsymbol{\mu}^{(m)} = \{\boldsymbol{\mu}_k^{(m)}\}_{k=1}^{K^{(m)}}$, and covariance matrices, $\boldsymbol{\Sigma}^{(m)} = \{\boldsymbol{\Sigma}_k^{(m)}\}_{k=1}^{K^{(m)}}$. We can transcribe the i-th speech frame $\mathbf{x}_i^{(m)}$ with label $l_i^{(m)}$ as follows:

$$l_i^{(m)}(\mathbf{x}_i^{(m)}) = \arg \max_{1 \le k \le K} p_{i,k}, \tag{1}$$

where $p_{i,k} = p(k|\mathbf{x}_i^{(m)})$ is the posterior of k-th Gaussian component given $\mathbf{x}_i^{(m)}$, which can be computed as in [4] using the $\boldsymbol{\pi}^{(m)}$, $\boldsymbol{\mu}^{(m)}$ and $\boldsymbol{\Sigma}^{(m)}$. Since each DPGMM is language-dependent, we refer to this phoneme-like labeling procedure as language-dependent phoneme-like labeling.

2.2. Multilingual bottle-neck feature learning

Similar to the approach proposed by Veselý *et al.* [22], our multilingual bottle-neck features (BNFs) are learned via multi-task learning (MTL) [28] deep neural network (MTL-DNN). The low-dimensional speech representation shared across multiple languages is extracted from a linear bottle-neck layer in an MTL-DNN as shown in Fig. 1. Veselý *et al.* [22] have demonstrated that such a representation captures common acoustic properties across the source languages. However, different from [22], our BNFs are learned without any manual transcription. Our MTL-DNN is to predict unsupervised phoneme-like labels instead of supervised labels [22].

Given M source languages, we have an MTL-DNN with M tasks to learn. The loss function of such an MTL-DNN can be written as:

$$\mathcal{L}_{ce} = \sum_{m=1}^{M} w_m \mathcal{L}_{ce}^{(m)}, \tag{2}$$

where w_m is the weight of the m-th task which is set to $\frac{1}{M}$ in this study, and $\mathcal{L}_{ce}^{(m)}$ is the loss function for the m-th single-task learning (STL) DNN defined as cross-entropy between the predictions and the true labels:

$$\mathcal{L}_{ce}^{(m)} = -\sum_{i}^{N^{(m)}} \sum_{k}^{K^{(m)}} t_{i,k}^{(m)} \log s_{i,k}^{(m)}. \tag{3}$$

Here, $t_{i,k}^{(m)}$ is 1 if $\mathbf{x}_i^{(m)}$ is labeled with phoneme-like unit k from the m-th language-dependent DPGMM according to $l_i^{(m)}$ as defined in

[2] http://people.csail.mit.edu/jchang7/code.php

Eq. (1), otherwise 0. $K^{(m)}$ is the number of output dimensions. The element of the softmax output $s_{i,k}^{(m)}$ is:

$$s_{i,k}^{(m)} = p(k|\mathbf{x}_i^{(m)}) = \frac{e^{z_k^{(m)}}}{\sum_{k'=1}^{K^{(m)}} e^{z_{k'}^{(m)}}} \quad (4)$$

where $z_k^{(m)}$ is the k-th input of the softmax layer, $\mathbf{z}^{(m)}$. The MTL-DNN can be trained using the back-propagation method by using the weighted sum of the back-propagated error in each task.

After training, we obtain BNFs by forward-passing input spectral features through the feature extractor (dashed box in Fig. 1.) which is a DNN without the last two layers of the original MTL-DNN.

3. EXPERIMENTS

3.1. Corpus and ABX phoneme discriminability test

We evaluated multilingual bottle-neck features (BNFs) in the track 1 of ZeroSpeech2017. The goal of this track is to construct a frame-wise representation of speech sounds which supports word/sub-word discrimination. ZeroSpeech2017 corpus consists of five languages as listed in Table 1. Three of them are used for hyperparameter development. The rest two are 'surprise' languages, denoted as LANG1 and LANG2, for performance evaluation only.

Table 1. Datasets in ZeroSpeech2017 corpus.

Development Dataset			Surprise Dataset		
	Training	Test		Training	Test
English	45 hrs	27 hrs	LANG1	25 hrs	11 hrs
French	24 hrs	18 hrs	LANG2	11 hrs	6 hrs
Mandarin	3 hrs	25 hrs			

We performed a set of ABX tests to evaluate the phoneme discrimination ability of our proposed feature representation. For further detailed definition, please refer to [23, 24] or the website of ZeroSpeech2017. We used the official ABX phoneme discriminability test toolkit of the track 1 in ZeroSpeech2017. The test instances were extracted from utterances with the average size of 1s, 10s, and 120s. The distance of instance pairs is their DTW distance normalized by the length of the aligned path. In DTW, we used cosine distance for BNFs and J-divergence for posteriorgrams for frame-by-frame comparison. The correct rates were averaged over all found contexts for a given central phoneme and then over all central phonemes.The error rates are reported in Section 4.

3.2. Training of DPGMMs

For each language in ZeroSpeech2017, we trained a DPGMM on the training set using mel-frequency cepstral coefficients (MFCC) with $\Delta + \Delta\Delta$ (39-dimensional) post-processed by cepstral mean and variance normalization (CMVN). The parallel split/merge sampler [26] is employed to infer these DPGMMs. All the hyperparameters for the inference of all DPGMMs are set following our previous work [8] and are listed in Table 2, where $\mathrm{mean}(\mathcal{X})$ and $\mathrm{cov}(\mathcal{X})$ is the mean vector and covariance matrix of the input feature vectors, $\mathcal{X}$. Codes and parameter configuration are available at [29]. The numbers of Gaussian components of the resultant language-dependent DPGMMs are shown in Table 2. As shown in Table 2, the number of Gaussian components varies according to the number of speech frames presented in each language.

Table 2. The hyperparameters in the training of the DPGMMs and the number of Gaussian components ($K^{(m)}$) in each DPGMM after 3000 iterations of the sampler.

α_0	κ_0	ν_0	μ_0	Σ_0	K_0	#iteration
1	1	41	$\mathrm{mean}(\mathcal{X})$	$\mathrm{cov}(\mathcal{X})$	800	3000

m	English	French	Mandarin	LANG1	LANG2
$K^{(m)}$	1148	1070	451	1108	578

3.3. Training of DNNs

Two MTL-DNNs were trained on the training sets with language-dependent unsupervised phoneme-like labels in ZeroSpeech2017. One was trained using all languages to extract the proposed multilingual BNFs for the challenge evaluation and the other was trained using the development datasets for experimental analysis due to the time limitation in the evaluation. Note that DPGMM posteriorgrams (PGs) [8] performed the best in the track 1 of ZeroSpeech2015 while the monolingual BNFs proposed by Chen *et al.* [17] performed better than DPGMM PGs in the low-resource query-by-example spoken term detection. We believed monolingual BNFs would also outperform PGs in ZeroSpeech2017. To verify this guess and whether multilingual BNFs outperform monolingual BNFs in terms of generalization performance across multiple source languages when using unsupervised phoneme-like labels, we trained a set of STL-DNNs with BN layers on each language in ZeroSpeech2017 to extract monolingual BNFs. Moreover, since one may transfer cross-lingual knowledge when faced with a language without manual transcription, we extract cross-lingual BNFs from a DNN trained on the Fisher Spanish Speech Corpus[3] (LDC2010S01, 163 hrs) to evaluate the proposed multilingual BNFs.

The data configuration for the DNNs are shown in Table 3. ①-⑤ share the same hidden topology, four 1024-unit hidden layers, one 40-unit linear bottle-neck layer and one 1024-unit hidden layer. The dimensions of their softmax layers depend on the numbers of the Gaussian components as shown in Table 2. The hidden topology of ⑥ consists of two 1500-unit hidden layers and one 40-unit linear bottle-neck layer. All the DNNs were trained using Kaldi [30] and took filter banks plus pitch (FBank+F0) as the input feature. 90% of training set was used as training subset and the rest 10% was used as cross-validation subset when training the DNNs.

4. RESULTS AND ANALYSIS

4.1. Comparison with the baseline and the topline

Table 4 summarizes the performance of the proposed multilingual BNFs and the features including the baseline feature (MFCC) and the topline feature (PGs) provided by ZeroSpeech2017 challenge organization in the ABX phoneme discriminability test. As shown in this table, multilingual BNFs outperformed the baseline with absolute 0.8%-4.8% and 7.6%-13.2% in the within-talker and across-talker test respectively. Our proposed multilingual BNFs showed comparable performance across different source languages in within-talker phoneme discrimination. Note that the topline PGs of each language are obtained from a language-specific model trained using speech data with manual transcription. In across-talker phoneme discrimination, however, there are still obvious gaps between our proposed features and the topline features. This may be caused by the lack of consideration to mitigate across-talker varia-

[3] https://catalog.ldc.upenn.edu/LDC2010S01

Table 3. Data configurations for DNNs. ①-② denote the MTL-DNNs for the proposed multilingual BNFs; ③-⑤ denote the DNNs trained with unsupervised phoneme-like labels for monolingual BNFs proposed by Chen *et al.* [17]; ⑥ denotes the DNN trained on cross-lingual corpus with manual transcription. Here, *Sup* (*Unsup*) means that the corresponding DNN is trained with (without) manual transcription.

①	Multilingual (Unsup, 5)
②	Multilingual (Unsup, 3)
③	Monolingual (Unsup, English)
④	Monolingual (Unsup, French)
⑤	Monolingual (Unsup, Mandarin)
⑥	Monolingual (Sup, Spanish)

	ZeroSpeech2017 training set					Fisher
	English	French	Mandarin	LANG1	LANG2	Spanish
①	√	√	√	√	√	
②	√	√	√			
③	√					
④		√				
⑤			√			
⑥						√

tion in our phoneme-like unit modeling. Thus we investigate the use of vocal tract length normalization in DPGMM in Section 4.5.

Fig. 2 illustrates the comparison between the proposed multilingual BNFs and the cross-lingual BNFs learned on the Fisher corpus. Here the cross-lingual BNFs are not tuned for the target zero-resource languages. We can see that for within-talker phoneme discrimination, multilingual BNFs approached the cross-lingual counterparts. However, similar to the comparison between multilingual BNFs and the topline PGs in Table 4, multilingual BNFs performed worse than cross-lingual BNFs and the increase in across-talker error rates is larger than that in the within-talker error rates. Again, we think this is not surprising, since DPGMMs only cluster speech frames without explicit consideration of speaker information, while the cross-lingual DNN is steered by manual transcriptions to ignore the talker variation.

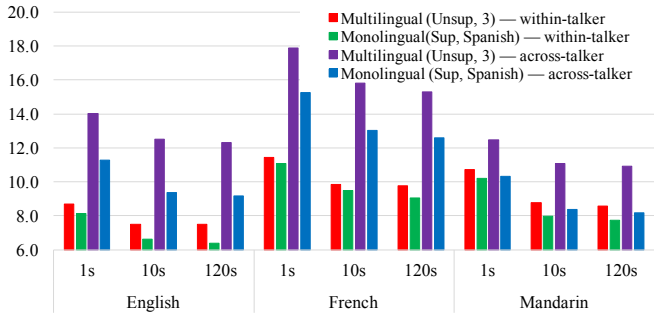
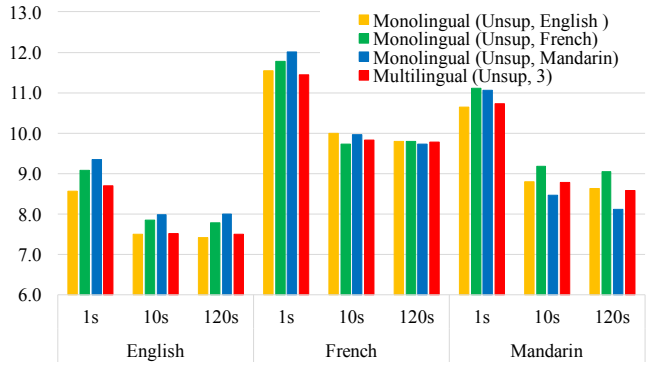


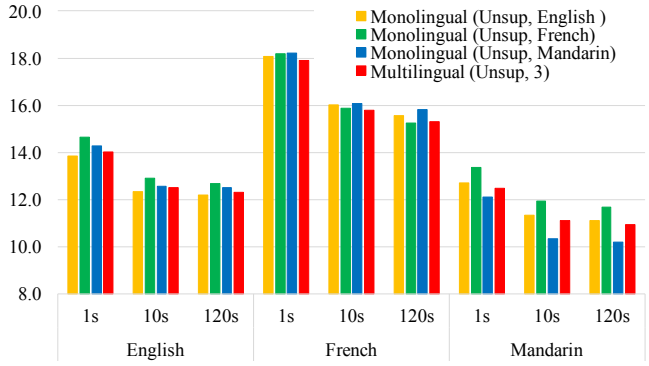
Fig. 2. Comparison between the proposed BNFs and the cross-lingual BNFs learned in a supervised manner.

4.2. Multilingual BNFs vs. monolingual BNFs

Fig. 3 shows the comparison between multilingual BNFs and monolingual BNFs when they were trained with unsupervised phoneme-like labels. As shown in Fig. 3, the proposed BNFs performed the best or the second best across different languages, i.e. multilingual BNFs has a better generalization performance across multiple source languages than monolingual BNFs. This is similar to the comparison between multilingual BNFs and monolingual BNFs when using manual transcription in [22]. We also observed that our multilingual BNFs brought more obvious errors than the best monolingual (Mandarin) BNFs in Mandarin test data (0.4%-0.8% in Mandarin, and 0.1%-0.2% in English and French). This is possibly because the amount of training data in Mandarin is much less than that of the other two languages. We also obtained the similar result in the Mandarin test data even when we used a larger weight for Mandarin in the weighted cross-entropy loss function, and further investigation will be needed in the future.



(a) Within-talker error rate (%)



(b) Across-talker error rate (%)

Fig. 3. The comparison between multilingual BNFs and monolingual BNFs, which are derived from unsupervised phoneme-like labels.

4.3. Effect of language-dependent phoneme-like labeling

We investigate the importance of language-dependent labeling when learning multilingual BNFs from the untranscribed speech. We trained a DPGMM by pooling all the datasets used in the training of the MTL-DNN and then trained a DNN to extract BNFs, which we refer to as *Pooling* BNFs. As shown in Fig. 4, multilingual BNFs outperformed *Pooling* BNFs significantly. This observation, on one hand, suggests that multiple languages have considerably different acoustic characteristics so that it is not desirable to represent all the languages by a single language-independent DPGMM. On the other hand, there are indeed acoustic properties shared by multiple languages and it is plausible to learn a multilingual representation, e.g. the proposed multilingual BNFs, even though no manual transcription is available.

Table 4. Error rate (%) of ABX phoneme discriminability test on the proposed multilingual BNFs, MFCC and posteriorgrams (PGs). The PGs are provide by ZeroSpeech2017 trained with manual transcription.

	Within-talker														
	Development dataset									Surprise dataset					
	English			French			Mandarin			LANG1			LANG2		
	1s	10s	120s	1s	10s	120s	1s	10s	120s	1s	10s	120s	1s	10s	120s
Multilingual (Unsup, 5)	8.5	7.4	7.3	11.1	9.6	9.4	10.6	8.6	8.5	7.6	6.3	6.3	11.8	10.0	9.8
Baseline (MFCC)	12.0	12.1	12.1	12.5	12.6	12.6	11.5	11.5	11.5	10.3	9.3	9.4	14.1	14.3	14.1
Topline (Sup, PGs)	6.5	5.3	5.1	8.0	6.8	6.8	9.5	4.2	4.0	8.7	7.1	7.0	6.6	4.6	3.4

	Across-talker														
	Development dataset									Surprise dataset					
	English			French			Mandarin			LANG1			LANG2		
	1s	10s	120s	1s	10s	120s	1s	10s	120s	1s	10s	120s	1s	10s	120s
Multilingual (Unsup, 5)	13.8	12.2	12.1	17.6	15.6	14.9	12.4	10.8	10.7	15.5	13.0	12.7	17.7	16.9	16.3
Baseline (MFCC)	23.4	23.4	23.4	25.2	25.5	25.2	21.3	21.3	21.3	23.6	23.2	23.0	30.0	29.5	29.5
Topline (Sup, PGs)	8.6	6.9	6.7	10.6	9.1	8.9	12.0	5.7	5.1	12.8	10.5	10.4	7.1	3.6	4.3

Note that, a concatenation of monolingual BNFs is also a multilingual feature representation where each source of monolingual BNFs is derived from the language-dependent phoneme-like labeling. Thus, we also tested phoneme discriminability of feature concatenation. The feature concatenation outperformed *Pooling* BNFs as well. We note that feature concatenation performed slightly better (absolute error reduction 0.05%-0.4%) than our multilingual BNFs. However, our proposed BNFs can keep the compact representation. The dimension of our proposed multi-lingual BNFs does not depend on the number of source languages while the feature dimension in feature concatenation grows linearly with the number of source languages.

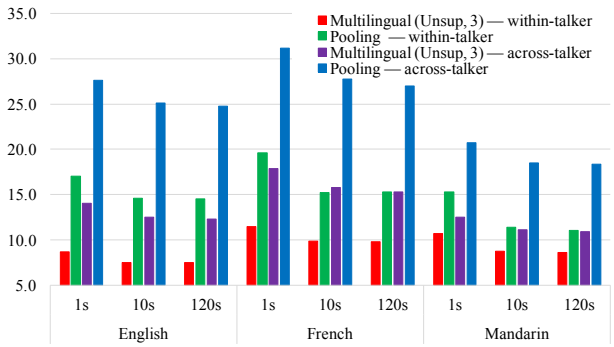


Fig. 4. The performance of multilingual BNFs and the BNFs learned from a DNN using language-independent DPGMM (*Pooling*). The training data for these features are the training sets of English, French and Mandarin from ZeroSpeech2017 altogether.

4.4. BNFs vs. PGs

Fig. 5 illustrates the comparison between the BNFs and PGs which are both derived from the language-dependent DPGMM trained on Mandarin in ZeroSpeech2017. It is similar to the observation in [17] that the BNFs consistently outperformed DPGMM PGs in both within-talker and across-talker tests. Note that the BNFs are 40-dimension while the PGs can be several hundreds of dimensions, e.g. 451-1148, as shown in Table 2. This suggests that the proposed multilingual BNFs are a more efficient representation than PGs.

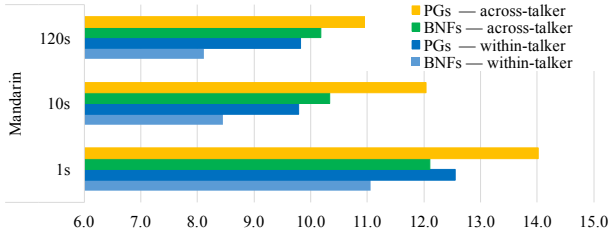


Fig. 5. The comparison between the DPGMM posteriorgrams (PGs) [8] and the BNFs learned using unsupervised phoneme-like labels on Mandarin datasets.

4.5. Effect of VTLN

Note that the across-talker error rates (10.9%-17.9%) were much higher than the within-talker ones (8.6%-11.5%) and our phoneme-like model considers no talker information. Within the DPGMM phoneme-like modeling, a direct way to improve the across-talker performance is to improve the input feature of the DPGMMs so that the unsupervised labels are more phoneme-like. Thus we further applied linear vocal tract length normalization (VTLN) similar to [31] implemented in Kaldi on the input features of DPGMMs. As shown in Fig. 6, after VTLN, the across-talker error rates were reduced 4%-12% relatively. This further reduced the gap between the proposed BNFs and the cross-lingual BNFs derived from manual transcription.

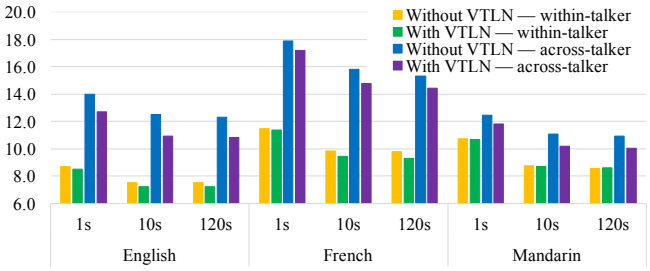


Fig. 6. The performance of the proposed BNFs using DPGMMs learned with/without vocal tract length normalization (VTLN).

5. CONCLUSION AND FUTURE WORK

We have shown that it is viable to learn multilingual BNFs from the untranscribed speech by training an MTL-DNN using unsupervised phoneme-like labels. The unsupervised phoneme-like labels are obtained from language-dependent DPGMMs. Our experimental results showed that the proposed BNFs support phoneme discriminability across multiple source languages, and their within-talker phoneme discriminability shows competitive to the official topline features of each source language. In the future, we would investigate the ways to improve the across-talker phoneme discriminability of our proposed features. For example, talker normalization in phoneme-like unit modeling will further be studied. We would also improve our proposed BNFs by considering the techniques adopted by Heck *et al.* [32, 33] to improve DPGMM posteriorgrams.

References

[1] A. Park and J. R. Glass, "Towards unsupervised pattern discovery in speech," in *Proc. ASRU*, 2005, pp. 53–58.

[2] C. Lee, T. O'Donnell, and J. Glass, "Unsupervised lexicon discovery from acoustic input," *Trans. ACL*, vol. 3, pp. 389–403, 2015.

[3] H. Kamper, A. Jansen, and S. Goldwater, "Unsupervised word segmentation and lexicon discovery using acoustic word embeddings," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 669–679, 2016.

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

[5] 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*. IEEE, 2012, pp. 5157–5160.

[6] K. Levin, A. Jansen, and B. V. Durme, "Segmental acoustic indexing for zero resource keyword search," in *Proc. ICASSP*, 2015, pp. 5828–5832.

[7] C. Liu, J. Yang, M. Sun, S. Kesiraju, A. Rott, L. Ondel, P. Ghahremani, N. Dehak, L. Burget, and S. Khudanpur, "An empirical evaluation of zero resource acoustic unit discovery," in *Proc. ICASSP*. IEEE, 2017, pp. 5305–5309.

[8] 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*. ISCA, 2015, pp. 3189–3193.

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

[10] C.-Y. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proc. ACL*. ACL, 2012, pp. 40–49.

[11] A. H. H. N. Torbati and J. Picone, "A nonparametric Bayesian approach for spoken term detection by example query," in *Proc. Interspeech*. ISCA, 2016, pp. 928–932.

[12] Y. Zhang, R. Salakhutdinov, H.-A. Chang, and J. Glass, "Resource configurable spoken query detection using deep Boltzmann machines," in *Proc. ICASSP*. IEEE, 2012, pp. 5161–5164.

[13] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "An auto-encoder based approach to unsupervised learning of subword units," in *Proc. ICASSP*, 2014, pp. 7634–7638.

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

[15] Y. Yuan, C.-C. Leung, L. Xie, H. Chen, B. Ma, and H. Li, "Extracting bottleneck features and word-like pairs from untranscribed speech for feature representation," in *Proc. ASRU*. IEEE, 2017.

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

[17] 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*. ISCA, 2016, pp. 923–927.

[18] K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in *Proc. ASRU*, 2013, pp. 410–415.

[19] H. Kamper, W. Wang, and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *Proc. ICASSP*, 2016, pp. 4950–4954.

[20] Y. Chung, C. Wu, C. Shen, H. Lee, and L. Lee, "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," in *Proc. Interspeech*, 2016, pp. 765–769.

[21] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Multi-task feature learning for low-resource query-by-example spoken term detection," *IEEE J. Sel. Topics Signal Process. (submitted)*, 2017.

[22] K. Veselý, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Proc. SLT*. IEEE, 2012, pp. 336–341.

[23] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline," in *Proc. Interspeech*. ISCA, 2013, pp. 1–5.

[24] T. Schatz, V. Peddinti, X.-N. Cao, F. Bach, H. Hermansky, and E. Dupoux, "Evaluating speech features with the Minimal-Pair ABX task (ii): Resistance to noise," in *Proc. Interspeech*. ISCA, 2014, pp. 915–919.

[25] H. Kamper, A. Jansen, S. King, and S. Goldwater, "Unsupervised lexical clustering of speech segments using fixed-dimensional acoustic embeddings," in *Proc. SLT*. IEEE, 2014, pp. 100–105.

[26] J. Chang and J. W. F. III, "Parallel sampling of DP mixture models using sub-cluster splits," in *Proc. NIPS*, 2013, pp. 620–628.

[27] C. E. Rasmussen, "The infinite Gaussian mixture model," in *Proc. NIPS*, 1999, pp. 554–560.

[28] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[29] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, https://doi.org/10.5281/zenodo.808915, 2017.

[30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz et al., "The Kaldi speech recognition toolkit," in *Proc. ASRU*, no. EPFL-CONF-192584. IEEE, 2011.

[31] D. Y. Kim, S. Umesh, M. J. F. Gales, T. Hain, and P. C. Woodland, "Using VTLN for broadcast news transcription," in *Proc. Interspeech*, 2004.

[32] M. Heck, S. Sakti, and S. Nakamura, "Unsupervised linear discriminant analysis for supporting DPGMM clustering in the zero resource scenario," in *Proc. SLTU*, 2016, pp. 73–79.

[33] ——, "Supervised learning of acoustic models in a zero resource setting to improve DPGMM clustering," in *Proc. Interspeech*. ISCA, 2016, pp. 1310–1314.