INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Improving Unsupervised Subword Modeling via Disentangled Speech Representation Learning and Transformation

Siyuan Feng, Tan Lee

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong
siyuanfeng@link.cuhk.edu.hk, tanlee@ee.cuhk.edu.hk

Abstract

This study tackles unsupervised subword modeling in the zero-resource scenario, learning frame-level speech representation that is phonetically discriminative and speaker-invariant, using only untranscribed speech for target languages. Frame label acquisition is an essential step in solving this problem. High quality frame labels should be in good consistency with golden transcriptions and robust to speaker variation. We propose to improve frame label acquisition in our previously adopted deep neural network-bottleneck feature (DNN-BNF) architecture by applying the factorized hierarchical variational autoencoder (FHVAE). FHVAEs learn to disentangle linguistic content and speaker identity information encoded in speech. By discarding or unifying speaker information, speaker-invariant features are learned and fed as inputs to DPGMM frame clustering and DNN-BNF training. Experiments conducted on ZeroSpeech 2017 show that our proposed approaches achieve 2.4% and 0.6% absolute ABX error rate reductions in across- and within-speaker conditions, comparing to the baseline DNN-BNF system without applying FHVAEs. Our proposed approaches significantly outperform vocal tract length normalization in improving frame labeling and subword modeling.
Index Terms: unsupervised subword modeling, disentangled representation, speaker-invariant feature, zero resource

1. Introduction

Recent years have witnessed a huge success in applying deep learning techniques in acoustic and language modeling for automatic speech recognition (ASR). Training deep neural network (DNN) acoustic models requires large amounts of transcribed speech data. For many languages in the world, for which very little or no transcribed speech is available, conventional supervised acoustic modeling techniques cannot be directly applied.

Unsupervised acoustic modeling (UAM) aims at discovering and modeling acoustic units of an unknown language at subword or word level, assuming only untranscribed speech data are available. UAM is a challenging problem with significant practical impact in speech as well as linguistics and cognitive science communities. It has been studied in applications such as ASR for low-resource languages [1], language identification [2] and query-by-example spoken term detection [3]. This problem is also relevant to endangered language protection [4] and understanding infants' language acquisition mechanism [5].

Over the recent past, Zero Resource Speech Challenges (ZeroSpeech) 2015 [6] and 2017 [7] were organized to focus on unsupervised speech modeling. ZeroSpeech 2017 Track one, named unsupervised subword modeling, was formulated as an unsupervised feature representation learning problem, i.e., how to learn frame-level speech features that are discriminative to subword units and robust to linguistically-irrelevant variations such as speaker identity. The present study addresses this problem. It is a fundamental problem in unsupervised speech modeling. Speech simultaneously encodes linguistically-relevant information e.g. subword units and linguistically-irrelevant information e.g. speaker variation that are not easily separable. In supervised acoustic modeling, golden transcription can be relied on to ensure the robustness of the learned subword units towards linguistically-irrelevant information. In the unsupervised scenario, subword units and word patterns can only be inferred from speech features. This makes feature representation learning important in the zero-resource scenario. In the literature, representation learning has been shown beneficial to downstream applications such as spoken query retrieval [8].

In our previous attempt to ZeroSpeech 2017 [9], a DNN was trained with zero-resource speech data to generate bottleneck features (BNFs) as the learned feature representation. Frame labels for supervised DNN training were obtained through Dirichlet process Gaussian mixture model (DPGMM) based frame clustering. This framework is similar to [10]. By employing out-of-domain transcribed speech data for speaker adapted feature learning and DNN frame labeling, the results in [9] significantly outperform [10] in which out-of-domain data were not employed. This improvement is mainly attributed to the advancement of frame label acquisition. Ideally, the learned frame labels should have a full coverage of linguistically-defined phonemes. They should be in good consistency with golden transcription and robust to speaker variation. The quality of frame labels has a significant impact on the performance of subword modeling [11]. Many prior works found out that DPGMM clustering towards speaker adapted features could generate better labels than that towards unadapted features [10–12]. In [10], the authors compared MFCC features with and without vocal tract length normalization (VTLN) for clustering. In [11], MFCCs were first clustered to generate initial tokenization, with which linear transforms such as LDA, MLLT and fMLLR were estimated. The fMLLRs are clustered again to generate the final form of frame labels. This work achieved the best performance in ZeroSpeech 2017. It is worth noting that DPGMM clustering requires high computational costs. Typically, clustering towards 40-hour speech data for 100 iterations using 32 CPU cores takes up to 25 hours. This makes the system in [11] much heavier than [9, 10].

In the strict zero-resource scenario, out-of-domain speech and language resources are unavailable. This paper proposes to improve DPGMM frame labeling using only in-domain untranscribed speech data, and refrain from performing multiple-pass clustering processes. Specifically, the factorized hierarchical variational AE (FHVAE) model [13] is used to disentangle linguistic content and speaker information in raw speech features in an unsupervised manner. By either discarding or unifying speaker information, speaker-invariant representation is learned and used as the input to DPGMM clustering and DNN-BNF training. The FHVAE is an unsupervised generative model. It

http://dx.doi.org/10.21437/Interspeech.2019-1338

was originally proposed to deal with domain adaptation problems in noise robust ASR [14], distant conversational ASR [15], and later applied to dialect identification [16]. To the best of our knowledge, the use of FHVAEs in unsupervised subword modeling has never been studied before.

2. Speaker-invariant feature learning by FHVAE

Speaker characteristics tends to have a smaller amount of variation than linguistic content within a speech utterance, while linguistic content tends to have similar amounts of variation within and across utterances. The FHVAE model [13], which learns to factorize sequence-level and segment-level attributes of sequential data into different latent variables, is applied in this work to disentangle linguistic content and speaker characteristics.

2.1. FHVAE model

FHVAEs formulate the generation process of sequential data by imposing sequence-dependent priors and sequence-independent priors to different sets of variables. Following notations and terminologies in [13], let $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ denote latent segment variable and latent sequence variable, respectively. $\boldsymbol{\mu_2}$ is sequence-dependent prior, named as *s-vector*. θ and ϕ denote the parameters of generation and inference models of FHVAEs. Let $\mathcal{D} = \{\boldsymbol{X^i}\}_{i=1}^{M}$ denote a speech dataset with M sequences. Each $\boldsymbol{X^i}$ contains N^i speech segments $\{\boldsymbol{x^{(i,n)}}\}_{n=1}^{N^i}$, where $\boldsymbol{x^{(i,n)}}$ is composed of fixed-length consecutive frames. The FHVAE model generates a sequence $\boldsymbol{X}$ from a random process as follows: (1) $\boldsymbol{\mu_2}$ is drawn from a prior distribution $p_\theta(\boldsymbol{\mu_2}) = \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{\mu_2}}\boldsymbol{I})$; (2) $\boldsymbol{z_1^n}$ and $\boldsymbol{z_2^n}$ are drawn from $p_\theta(\boldsymbol{z_1^n}) = \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{z_1}}\boldsymbol{I})$ and $p_\theta(\boldsymbol{z_2^n}|\boldsymbol{\mu_2}) = \mathcal{N}(\boldsymbol{\mu_2}, \sigma^2_{\boldsymbol{z_2}}\boldsymbol{I})$ respectively; (3) Speech segment $\boldsymbol{x^n}$ is drawn from $p_\theta(\boldsymbol{x^n}|\boldsymbol{z_1^n}, \boldsymbol{z_2^n}) = \mathcal{N}(f_{\mu_{\boldsymbol{x}}}(\boldsymbol{z_1^n}, \boldsymbol{z_2^n}), diag(f_{\boldsymbol{\sigma^2_x}}(\boldsymbol{z_1^n}, \boldsymbol{z_2^n}))$. Here $\mathcal{N}$ denotes standard normal distribution, $f_{\mu_{\boldsymbol{x}}}(\cdot,\cdot)$ and $f_{\boldsymbol{\sigma^2_x}}(\cdot,\cdot)$ are parameterized by DNNs. The joint probability for $\boldsymbol{X}$ is formulated as,

$$p_\theta(\boldsymbol{\mu_2}) \prod_{n=1}^{N} p_\theta(\boldsymbol{z_1^n}) p_\theta(\boldsymbol{z_2^n}|\boldsymbol{\mu_2}) p_\theta(\boldsymbol{x^n}|\boldsymbol{z_1^n}, \boldsymbol{z_2^n}). \quad (1)$$

Similar to VAE models, FHVAEs introduce an inference model q_ϕ to approximate the intractable true posterior as,

$$q_\phi(\boldsymbol{\mu_2}) \prod_{n=1}^{N} q_\phi(\boldsymbol{z_2^n}|\boldsymbol{x^n}) q_\phi(\boldsymbol{z_1^n}|\boldsymbol{x^n}, \boldsymbol{z_2^n}). \quad (2)$$

Here $q_\phi(\boldsymbol{\mu_2}), q_\phi(\boldsymbol{z_2^n}|\boldsymbol{x^n})$ and $q_\phi(\boldsymbol{z_1^n}|\boldsymbol{x^n}, \boldsymbol{z_2^n})$ are all diagonal Gaussian distributions. The mean and variance values of $q_\phi(\boldsymbol{z_2^n}|\boldsymbol{x^n})$ and $q_\phi(\boldsymbol{z_1^n}|\boldsymbol{x^n}, \boldsymbol{z_2^n})$ are parameterized by two DNNs. For $q_\phi(\boldsymbol{\mu_2})$, during FHVAE training, a trainable lookup table containing posterior mean of $\boldsymbol{\mu_2}$ for each sequence is updated. During testing, maximum a posteriori (MAP) estimation is used to infer $\boldsymbol{\mu_2}$ for unseen test sequences. Details of $\boldsymbol{\mu_2}$ estimation for test sequences are described in [13].

FHVAEs optimize the discriminative segmental variational lower bound $\mathcal{L}(\theta, \phi; \boldsymbol{x^{(i,n)}})$ defined as,

$$\begin{aligned}
&\mathbb{E}_{q_\phi(\boldsymbol{z_1^{(i,n)}}, \boldsymbol{z_2^{(i,n)}}|\boldsymbol{x^{(i,n)}})}[\log p_\theta(\boldsymbol{x^{(i,n)}}|\boldsymbol{z_1^{(i,n)}}, \boldsymbol{z_2^{(i,n)}})] - \\
&\mathbb{E}_{q_\phi(\boldsymbol{z_2^{(i,n)}}|\boldsymbol{x^{(i,n)}})}[\mathrm{KL}(q_\phi(\boldsymbol{z_1^{(i,n)}}|\boldsymbol{x^{(i,n)}}, \boldsymbol{z_2^{(i,n)}})||p_\theta(\boldsymbol{z_1^{(i,n)}}))] \\
&- \mathrm{KL}(q_\phi(\boldsymbol{z_2^{(i,n)}}|\boldsymbol{x^{(i,n)}})||p_\theta(\boldsymbol{z_2^{(i,n)}}|\tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}^{\boldsymbol{i}})) \\
&+ \frac{1}{N^i}\log p_\theta(\tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}^{\boldsymbol{i}}) + \alpha \log p(i|\boldsymbol{z_2^{(i,n)}}),
\end{aligned}$$

where i is sequence index, $\tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}^{\boldsymbol{i}}$ denotes posterior mean of $\boldsymbol{\mu_2}$ for the i-th sequence, α denotes the discriminative weight. The discriminative objective $\log p(i|\boldsymbol{z_2^{(i,n)}})$ is defined as $\log p_\theta(\boldsymbol{z_2^{(i,n)}}|\tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}^{\boldsymbol{i}}) - \log \sum_{j=1}^{M} p_\theta(\boldsymbol{z_2^{(j,n)}}|\tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}^{\boldsymbol{j}})$.

After FHVAE training, $\boldsymbol{z_2}$ encodes factors that are relatively consistent within a sequence. The discriminative objective ensures that $\boldsymbol{z_2}$ captures sequence-dependent information. $\boldsymbol{z_1}$ encodes residual factors that are sequence-independent.

2.2. Extracting speaker-invariant features by FHVAE

In order to apply the FHVAE model to speaker-invariant feature learning, training utterances of the same speaker are concatenated into a single sequence. By this means, $\boldsymbol{z_2}$ is expected to encode speaker identity information and carry little phonetic information. $\boldsymbol{z_1}$ is expected to encode residual information, i.e. linguistic content, and carry little speaker information. This work considers obtaining speaker-invariant feature representations based on a trained FHVAE by two methods. The first method is straightforward to treat latent segment variables $\{\boldsymbol{z_1^{(i,n)}}\}$ as the desired feature representation.

In the second method, the FHVAE model reconstructs speech features of all utterances based on a unified s-vector. The reconstructed features are the desired representation. Specifically, a representative speaker with his/her s-vector $\boldsymbol{\mu_2^*}$ is chosen from the dataset. Next, for each speech segment $\boldsymbol{x^{(i,n)}}$ of an arbitrary speaker i, its corresponding latent sequence variable $\boldsymbol{z_2^{(i,n)}}$ is transformed to $\hat{\boldsymbol{z}}_{\boldsymbol{2}}^{(\boldsymbol{i,n})} = \boldsymbol{z_2^{(i,n)}} - \boldsymbol{\mu_2^i} + \boldsymbol{\mu_2^*}$, where $\boldsymbol{\mu_2^i}$ denotes the s-vector of speaker i. Finally the FHVAE decoder reconstructs speech segment $\hat{\boldsymbol{x}}^{(\boldsymbol{i,n})}$ conditioned on $\boldsymbol{z_1^{(i,n)}}$ and $\hat{\boldsymbol{z}}_{\boldsymbol{2}}^{(\boldsymbol{i,n})}$ using $p_\theta(\hat{\boldsymbol{x}}^{(\boldsymbol{i,n})}|\boldsymbol{z_1^{(i,n)}}, \hat{\boldsymbol{z}}_{\boldsymbol{2}}^{(\boldsymbol{i,n})})$. This method is named as *s-vector unification* in this work. Compared to original features, reconstructed features are expected to keep the linguistic content unchanged and capture speaker characteristics corresponding to the representative speaker. In other words, speech synthesized from $\{\hat{\boldsymbol{x}}^{(\boldsymbol{i,n})}\}$ would tend to sound as if they were all spoken by the representative speaker.

3. Unsupervised subword modeling with speaker-invariant features

3.1. DNN-BNF architecture

A DNN-BNF architecture [9,10] is adopted to perform phonetic discriminative training of untranscribed speech data and generate BNFs for subword modeling. In this architecture, given untranscribed speech data, Dirichlet process Gaussian mixture model (DPGMM) [17] algorithm is applied to cluster frame-level MFCC features for each target language individually. After clustering, each frame is assigned with a cluster label. These frame labels are regarded as pseudo phoneme alignments to support supervised DNN training. A multilingual DNN with a linear bottleneck layer is trained with frame alignments and MFCC features for all the target languages simultaneously, using multi-task learning [18]. After training, multilingual BNFs are extracted as the subword discriminative representation.

3.2. DNN-BNF training with speaker-invariant features

Speaker-invariant features learned by FHVAEs are applied to the DNN-BNF architecture in two aspects. As can be seen in Figure 1, during DPGMM-based frame clustering, input features to DPGMM are reconstructed MFCCs $\{\hat{\boldsymbol{x}}\}$ generated by the FHVAE decoder network using the s-vector unification

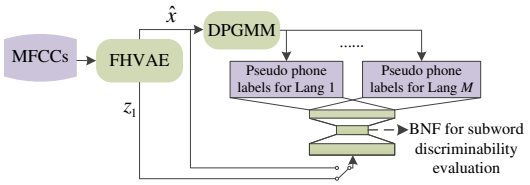

Figure 1: *DNN-BNF architecture with FHVAE-based speaker-invariant features for unsupervised subword modeling*

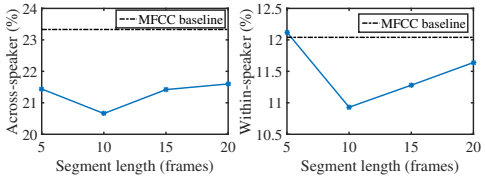

Figure 2: *ABX error rates (%) on $\boldsymbol{z_1}$ with different segment lengths and official MFCC baseline [7] (Avg. over languages)*

method described in Section 2.2, instead of original MFCCs. Compared to original MFCCs, FHVAE reconstructed MFCCs carry speaker information that is more consistent across utterances spoken by different speakers. With the reconstructed features as inputs, DPGMM clustering is expected to generate better phoneme-like labels and less affected by speaker variation.

During DNN-BNF model training, FHVAE-based speaker-invariant features are fed as inputs to the DNN. As seen in Figure 1, in this study we consider two feature types, i.e. reconstructed MFCCs with s-vector unification $\{\hat{\boldsymbol{x}}\}$ and latent segment variables $\{\boldsymbol{z_1}\}$, as DNN inputs. The effectiveness of these two types of features is compared in this study.

4. Experimental setup

4.1. Dataset and evaluation metric

Experiments are carried out with ZeroSpeech 2017 Track one [7]. Speaker identity information is released only for train sets. Detailed information is listed in Table 1.

Table 1: *Development data in ZeroSpeech 2017 Track one*

	Training			Test
	Duration	#speakers-R[1]	#speakers-L[1]	Duration
English	45 hrs	9	60	27 hrs
French	24 hrs	10	18	18 hrs
Mandarin	2.5 hrs	4	8	25 hrs

The evaluation metric is ABX subword discriminability. The ABX task is to decide whether X belongs to x or y if A belongs to x and B belongs to y, where A, B and X are three speech segments, x and y are two phonemes that differ in the central sound (e.g., "beg"-"bag"). Each pair of A and B are generated by the same speaker. ABX error rates for *within-speaker* and *across-speaker* are evaluated separately, depending on whether X and $A(B)$ belong to the same speaker.

4.2. FHVAE setup and parameter tuning

FHVAE model parameters are determined by reference to [14]. The encoder and decoder networks of FHVAE are both 2-layer LSTMs with 256 neurons per layer. The dimensions of $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ are 32. Training data for the three target languages are merged to train the FHVAE. Input features are fixed-length speech segments randomly chosen from utterances. The determination of segment length l is discussed in the next paragraph. Each frame is represented by a 13-dimensional MFCC with cepstral mean normalization at speaker level. During the inference of reconstructed feature representation, input segments are shifted by 1 frame. To match the length of extracted features with original MFCCs, the first and last frame are padded. Adam [19] with $\beta_1 = 0.95$ and $\beta_2 = 0.999$ is used to train the FHVAE. A 10% subset of training data is randomly selected for cross-validation. The training process is terminated if the lower bound on the cross-validation set does not improve for 20 epochs. Open-source tools [13] are used to train FHVAEs.

[1] "speakers-R/-L" denotes speakers with rich/limited speech data.

In our preliminary experiments, the ABX performance of $\boldsymbol{z_1}$ was found to be sensitive to the input segment length l. This could be explained as: a too large l would reduce the capability of $\boldsymbol{z_1}$ in modeling linguistic content at subword level; a too small l would restrict the FHVAE from capturing sufficient temporal dependencies which are essential in modeling speech. ABX error rates on $\boldsymbol{z_1}$ with different values of l are shown in Figure 2. The optimal value of l is 10. For the remaining experiments in this work, l is fixed to 10.

4.3. Selecting representative speaker for reconstructed feature extraction

The extraction of reconstructed MFCCs $\{\hat{\boldsymbol{x}}\}$ using s-vector unification assumes a pre-defined representative speaker. In order to validate the generalization ability of our proposed s-vector unification method and evaluate its sensitivity to the gender of the representative speaker, 6 English speakers {s0107, s3020, s4018, s0019, s1724, s2544}, 4 French speakers {M02R, M03R, F01R, F02R} and 2 Mandarin speakers {A08, C04} are randomly chosen from 'speaker-R' sets of ZeroSpeech 2017 training data. The first half speakers inside each language set are male and the second half are female. During the extraction of $\{\hat{\boldsymbol{x}}\}$, s-vectors $\{\boldsymbol{\mu_2^i}\}$ of all three target languages' utterances are modified to the same $\boldsymbol{\mu_2^*}$ corresponding to one of the 12 speakers mentioned above. The performance of the 12 groups of $\{\hat{\boldsymbol{x}}\}$ is evaluated by the ABX discriminability task.

4.4. DNN-BNF setup

For the baseline system without using FHVAE-based speaker-invariant features, input features to DPGMM are 39-dimensional MFCCs+Δ+$\Delta\Delta$. The numbers of clustering iterations for English, French and Mandarin sets are 120, 200 and 3000. After clustering, each frame is assigned with a label. A DNN-BNF is trained with all three languages' cepstral mean normalized MFCCs+Δ+$\Delta\Delta$ and frame labels using multi-task learning with equal task weights. The dimensions of hidden layers are $\{1024 \times 5, 40, 1024\}$. After training, 40-dimensional BNFs for test sets are extracted and evaluated by the ABX task. DPGMM is implemented using tools developed by [17]. DNN-BNF training is implemented using Kaldi `nnet1` recipe [20].

For the systems employing FHVAE-based speaker-invariant features, input features to DPGMM are reconstructed MFCCs $\{\hat{\boldsymbol{x}}\}$ with s-vector unification and further appended by Δ+$\Delta\Delta$. The representative speaker is selected from the 12 speakers mentioned in Section 4.3. The numbers of clustering iterations for the three languages are 80, 80 and 1400. DNN-BNFs are trained with either reconstructed MFCCs $\{\tilde{\boldsymbol{x}}\}$ or latent segment variables $\{\boldsymbol{z_1}\}$. The extraction of $\{\tilde{\boldsymbol{x}}\}$ is slightly different from $\{\hat{\boldsymbol{x}}\}$. During the inference of $\{\tilde{\boldsymbol{x}}\}$ for training sets, s-vector unification is not applied; during the inference for test sets, s-vector unification is applied within every test subset with a subset-specific $\boldsymbol{\mu_2^*}$. The reason is that DNN-BNFs trained with $\{\tilde{\boldsymbol{x}}\}$ were found to outperform those trained with

Table 2: *ABX error rates (%) on DNN-BNFs trained with/without FHVAE-based speaker-invariant features*

ID		Across-speaker										Within-speaker									
		English			French			Mandarin			Avg.	English			French			Mandarin			Avg.
		1s	10s	120s	1s	10s	120s	1s	10s	120s		1s	10s	120s	1s	10s	120s	1s	10s	120s	
	Baseline	13.5	12.4	12.4	17.8	16.4	16.1	12.6	11.9	12.0	13.90	8.0	7.3	7.3	10.3	9.4	9.3	10.1	8.8	8.9	8.82
	CA-Sup [9]	10.9	9.5	8.9	15.2	13.0	12.0	10.5	8.9	8.2	10.79	7.4	6.9	6.3	9.6	9.0	8.1	9.8	8.8	8.1	8.22
	MFCC [10]	13.7	12.1	12.0	17.6	15.6	14.8	12.3	10.8	10.7	13.29	8.5	7.3	7.2	11.1	9.5	9.4	10.5	8.5	8.4	8.93
	MFCC+VTLN [10]	12.7	11.0	10.8	17.0	14.5	14.1	11.9	10.3	10.1	12.49	8.5	7.3	7.2	11.2	9.4	9.4	10.5	8.7	8.5	8.97
①	$\boldsymbol{z_1}$ Orig.	12.9	11.7	11.7	17.2	15.5	15.2	12.5	11.4	11.5	13.29	8.2	7.0	7.0	10.7	9.2	9.1	10.4	8.8	8.7	8.79
②	$\tilde{\boldsymbol{x}}$ Orig.	12.8	11.7	11.5	17.8	15.5	15.1	12.3	10.9	10.7	13.14	8.2	7.3	7.0	10.6	9.3	8.9	10.5	8.8	8.7	8.81
③	$\boldsymbol{z_1}$ $\hat{\boldsymbol{x}}$-s0107	11.2	10.1	10.1	15.5	13.8	13.7	11.5	10.2	10.0	11.79	7.3	6.4	6.6	10.1	8.9	8.8	10.4	8.5	8.4	8.38
④	$\tilde{\boldsymbol{x}}$ $\hat{\boldsymbol{x}}$-s0107	11.6	10.4	10.1	16.1	13.9	13.7	11.9	10.2	10.4	12.03	7.8	6.7	6.5	10.5	9.6	9.3	10.8	8.6	8.7	8.72
⑤	$\boldsymbol{z_1}$ $\hat{\boldsymbol{x}}$-s4018	11.0	9.8	9.8	14.9	13.4	13.0	11.4	10.1	10.0	**11.49**	7.3	6.3	6.3	9.7	8.6	8.4	10.1	8.5	8.4	**8.18**
⑥	$\tilde{\boldsymbol{x}}$ $\hat{\boldsymbol{x}}$-s4018	11.3	10.0	9.8	15.7	13.6	13.3	11.8	10.0	10.4	11.77	7.8	6.5	6.5	10.1	9.1	8.8	10.6	8.7	8.7	8.53

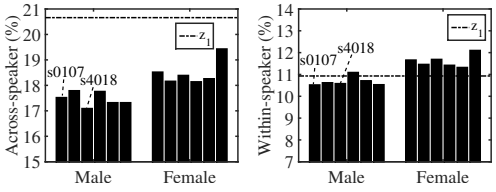

Figure 3: *ABX error rates (%) on $\hat{\boldsymbol{x}}$ using s-vector unification with different representative speakers (Avg. over languages)*

$\{\hat{\boldsymbol{x}}\}$. The DNN-BNF mentioned here has the same structure and loss function as that in the baseline system.

5. Results and analyses

5.1. Effectiveness of reconstructed MFCCs

ABX error rates on the 12 groups of reconstructed MFCCs $\{\hat{\boldsymbol{x}}\}$ using s-vector unification is shown in Figure 3. Each group is presented as a bar inside a bar graph. The reference line denotes ABX error rate on latent segment variables $\{\boldsymbol{z_1}\}$. It can be observed that, $\{\hat{\boldsymbol{x}}\}$ outperform $\{\boldsymbol{z_1}\}$ in across-speaker condition regardless of choosing any of the 12 speakers as the representative. In within-speaker condition, $\{\hat{\boldsymbol{x}}\}$ perform slightly better than $\{\boldsymbol{z_1}\}$ in most of the male cases, and are worse in all female cases. Further studies are needed to explain why male speakers are more suitable than females for s-vector unification.

5.2. DNN-BNFs trained with reconstructed MFCCs

Experimental results of the baseline DNN-BNF system and systems adopting FHVAE-based speaker-invariant features are summarized in Table 2. The second and third columns of IDs ①-⑥ denote inputs to DNN-BNF training and DPGMM clustering, respectively. 'Orig.' denotes original MFCCs without reconstruction. '$\hat{\boldsymbol{x}}$-s0107/-s4018' denotes reconstructed MFCCs with representative speaker s0107 or s4018. Here, $\hat{\boldsymbol{x}}$-s4018 is used to represent the ideal case as s4018 performs the best among the 12 speakers in across-speaker condition (see Figure 3). $\hat{\boldsymbol{x}}$-s0107 represents the general case as s0107 performs moderately among the male speakers. The system exploiting a Cantonese ASR for fMLLR estimation [9] is denoted as 'CA-Sup'. From this Table, several observations can be made:

(1) The comparison between baseline and ① & ② shows that without improving frame labels, the DNN-BNF model trained with $\{\tilde{\boldsymbol{x}}\}$ or $\{\boldsymbol{z_1}\}$ outperforms that trained with raw MFCCs, especially in across-speaker condition.

(2) The reconstructed MFCC features $\{\hat{\boldsymbol{x}}\}$ significantly outperform original MFCCs in DPGMM frame labeling. In the ideal case where the representative speaker 's4018' is selected, by comparing ⑤ and ①, frame labeling based on $\{\hat{\boldsymbol{x}}\}$ contributes to 13.5% and 6.9% relative ABX error rate reductions in across- and within-speaker conditions, compared to that based on original MFCCs. In the general case where 's0107' is selected, by comparing ③ and ①, the relative error rate reductions are 11.3% and 4.7% in across- and within-speaker conditions. The results demonstrate the importance of applying FHVAE-based speaker-invariant features in frame labeling.

(3) Our best system ⑤ achieves 2.4% and 0.6% absolute (17.3% and 7.3% relative) ABX error rate reductions compared to the baseline DNN-BNF system in across- and within-speaker conditions. The error rate reductions are attributed to better frame labeling and more speaker-invariant input features. As can be seen from baseline, ① and ⑤, the improvement in frame labeling is more prominent than that in input features. Compared to system CA-Sup in which out-of-domain transcribed data are exploited, ⑤ is slightly better in within-speaker condition while slightly inferior in across-speaker condition.

We also compare the effectiveness of our proposed approaches with [10], in which VTLN was adopted to improve frame labeling. As seen in Table 2, in across-speaker condition, while our baseline system is inferior to their baseline (MFCC), our best system consistently outperforms their system MFCC+VTLN in all test subsets. In within-speaker condition, our proposed approaches also achieve better performance. The comparison shows that FHVAE-based speaker-invariant feature learning is more effective than VTLN in improving the quality of frame labels and the robustness of subword modeling.

6. Conclusions

This paper presents a study on improving the quality of frame labels for unsupervised subword modeling without any out-of-domain resources. Frame labels are generated by clustering towards speaker-invariant features learned from FHVAEs. The speaker-invariant features are further fed as inputs to DNN-BNF training. Experiments conducted on ZeroSpeech 2017 show that our proposed approaches achieve 2.4%/0.6% absolute ABX error rate reductions in across-/within-speaker conditions, compared to the baseline without applying FHVAEs. Compared with a DNN-BNF system in which out-of-domain transcribed data are used for speaker adapted feature learning, our approaches perform slightly better in within-speaker condition while slightly worse in across-speaker condition. Our approaches significantly outperform VTLN in improving the quality of frame labels and the robustness of subword modeling.

7. Acknowledgements

This research is partially supported by the Major Program of National Social Science Fund of China (Ref:13&ZD189), a GRF project grant (Ref: CUHK 14227216) from Hong Kong Research Grants Council and a direct grant from CUHK Research Committee.

8. References

[1] H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, "Acoustic segment modeling with spectral clustering methods," *IEEE/ACM Trans. ASLP*, vol. 23, no. 2, pp. 264–277, 2015.

[2] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 271–284, 2007.

[3] 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 *INTERSPEECH*, 2016, pp. 923–927.

[4] A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, S. Khudanpur, K. Church *et al.*, "A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition," in *Proc. ICASSP*, 2013, pp. 8111–8115.

[5] E. Dupoux, "Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner," *arXiv*, vol. abs/1607.08723, 2016.

[6] M. Versteegh, R. Thiollière, T. Schatz, X.-N. Cao, X. Anguera, A. Jansen *et al.*, "The zero resource speech challenge 2015." in *Proc. INTERSPEECH*, 2015, pp. 3169–3173.

[7] E. Dunbar, X.-N. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier *et al.*, "The zero resource speech challenge 2017," in *Proc. ASRU*, 2017, pp. 323–330.

[8] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Multitask feature learning for low-resource query-by-example spoken term detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1329–1339, 2017.

[9] S. Feng and T. Lee, "Exploiting speaker and phonetic diversity of mismatched language resources for unsupervised subword modeling," in *Proc. INTERSPEECH*, 2018, pp. 2673–2677.

[10] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Multilingual bottle-neck feature learning from untranscribed speech," in *Proc. ASRU*, 2017, pp. 727–733.

[11] M. Heck, S. Sakti, and S. Nakamura, "Feature optimized DPGMM clustering for unsupervised subword modeling: A contribution to zerospeech 2017," in *Proc. ASRU*, 2017, pp. 740–746.

[12] ——, "Unsupervised linear discriminant analysis for supporting DPGMM clustering in the zero resource scenario," in *Proc. SLTU*, 2016, pp. 73–79.

[13] W. Hsu, Y. Zhang, and J. R. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Proc. NIPS*, 2017, pp. 1876–1887.

[14] W. Hsu and J. R. Glass, "Extracting domain invariant features by unsupervised learning for robust automatic speech recognition," in *Proc. ICASSP*, 2018, pp. 5614–5618.

[15] W. Hsu, H. Tang, and J. R. Glass, "Unsupervised adaptation with interpretable disentangled representations for distant conversational speech recognition," in *Proc. INTERSPEECH*, 2018, pp. 1576–1580.

[16] S. Shon, W. Hsu, and J. R. Glass, "Unsupervised representation learning of speech for dialect identification," in *arXiv*, 2018.

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

[18] R. Caruana, "Multitask learning," in *Learning to learn*. Springer, 1998, pp. 95–133.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv*, vol. abs/1412.6980, 2014.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel *et al.*, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.