# Improving Purely 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

For your paper to be published in the conference proceedings, you must use this document as both an instruction set and as a template into which you can type your own text. If your paper does not conform to the required format, you will be asked to fix it. Please do not reuse your past papers as a template. To prepare your paper for submission, please always download a fresh copy of this template from the conference website and please read the format instructions in this template before you use it for your paper. Conversion to PDF may cause problems in the resulting PDF or expose problems in your source document. Before submitting your final paper in PDF, check that the format in your paper PDF conforms to this template. Specifically, check the appearance of the title and author block, the appearance of section headings, document margins, column width, column spacing, and other features such as figure numbers, table numbers and equation number. In summary, you must proofread your final paper in PDF before submission.

**Index Terms**: speech recognition, human-computer interaction, computational paralinguistics

## 1. Introduction

Recent years have witnessed a huge success in applying deep learning models in acoustic and language modeling for automatic speech recognition (ASR). Training deep neural network (DNN) acoustic models (AMs) requires large amounts of transcribed speech data. This leads to the fact that high-performance ASR systems are only available for major languages. 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 automatically discovering and modeling acoustic units of a target 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. Researchers proposed various approaches for comparison [8–13]. In between, most of the works follow the *purely unsupervised* condition, i.e., only in-domain untranscribed speech data are available for building their systems. Cluster posteriorgram [8, 14, 15], DNN bottleneck features (BNFs) [9, 10], autoencoders (AEs) and their variants [11, 16] are among the widely adopted approaches. There are also works relaxing the purely unsupervised condition by assuming transcribed speech data for out-of-domain languages (different from in-domain ones) available, and building their systems in a transfer learning manner [12, 13]. The present study addresses the problem of ZeroSpeech 2017 Track one, unsupervised feature representation learning, and follows the purely unsupervised condition.

In our previous attempt to the task of ZeroSpeech 2017 [13], 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 frame clustering. This framework is similar to [9]. By employing out-of-domain transcribed speech data for speaker adapted feature learning and DNN frame labeling, the results in [13] significantly outperform [9] in which out-of-domain data were not employed. This naturally poses a question: Without exploiting any out-of-domain speech resources, how could we achieve the same performance in [13] or at least approach it to a great extent?

Speaker variation is a major difficulty in unsupervised subword modeling. The huge phoneme discriminability performance degradation caused by speaker variation in ZeroSpeech 2017 results [7] implies that speaker-invariant feature learning is both crucial and unsolved in the concerned task. Many works precisely focused on investigating speaker adaptation and speaker-invariant training methods in this task. Heck et al. [8] proposed to estimate fMLLR features. The transcription needed for fMLLR estimation was obtained by a first-pass frame clustering with MFCCs. Works in [12, 13] estimated fMLLRs by an out-of-domain ASR system. Chen et al. [9] applied vocal tract length normalization (VLTN) [17] on top of raw MFCCs. Tsuchiya et al. [18] applied speaker adversarial multi-task learning. Zeghidour et al. [19] proposed to train subword and speaker same-different tasks within a triamese network, and untangle linguistic and speaker representations. This model requires subword same-different information.

In this paper, we propose to improve our DNN-BNF based unsupervised subword modeling framework [13] by performing speaker-invariant feature learning using the factorized hierarchical variational AE (FHVAE) model [20]. Specifically, the FHVAE model is used to disentangle linguistic content and speaker information in speech in an unsupervised manner. By either discarding or unifying speaker information, speaker-invariant representation is learned and used as the input to DNN-BNF frame labeling and model training. The FHVAE is an unsupervised generative model, making it suitable for the zero-resource scenario. It was originally proposed for

domain adaptation problems with applications to noise robust ASR [21], distant conversational ASR [22], and later applied to dialect identification [23]. To the best of our knowledge, the use of FHVAEs in unsuperivsed 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 [20], 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 [20], let $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ denote latent segment variable and latent sequential variable, respectively. $\boldsymbol{\mu_2}$ is sequence-dependent prior, named as *s-vector*. $\theta$ and $\phi$ 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 $p_\theta(\boldsymbol{\mu_2}) = \mathcal{N}(\boldsymbol{0}, \sigma_{\mu_2}^2\boldsymbol{I})$; (2) $\boldsymbol{z_1^n}$ and $\boldsymbol{z_2^n}$ are drawn from $p_\theta(\boldsymbol{z_1^n}) = \mathcal{N}(\boldsymbol{0}, \sigma_{z_1}^2\boldsymbol{I})$ and $p_\theta(\boldsymbol{z_2^n}|\boldsymbol{\mu_2}) = \mathcal{N}(\boldsymbol{\mu_2}, \sigma_{z_2}^2\boldsymbol{I})$ respectively; (3) Speech segment $\boldsymbol{x^n}$ is drawn from $p_\theta(\boldsymbol{x^n}|\boldsymbol{z_1^n}, \boldsymbol{z_2^n}) = \mathcal{N}(f_{\mu_x}(\boldsymbol{z_1^n}, \boldsymbol{z_2^n}), diag(f_{\sigma_x^2}(\boldsymbol{z_1^n}, \boldsymbol{z_2^n}))$. Here $\mathcal{N}$ denotes standard normal distribution, $f_{\mu_x}(\cdot,\cdot)$ and $f_{\sigma_x^2}(\cdot,\cdot)$ 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 $q_\phi$ 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 $\boldsymbol{\mu_2}$ of unseen test sequences with details described in [20].

FHVAEs optimize 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, $\alpha$ denotes 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

The FHVAE model is applied to disentangle linguistic content and speaker characteristics encoded in speech. To this end, speech utterances belonging to the same speaker are concatenated into a single sequence before FHVAE training. 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 invariant to speaker identity.

This work considers two methods to obtain speaker-invariant feature representation based on a trained FHVAE. The first method is straightforward to treat latent segment variables $\{\boldsymbol{z_1^{(i,n)}}\}$ as the desired feature representation.

In the second method, s-vectors $\{\boldsymbol{\mu_2^i}\}$ of all the speech sequences are modified to the same value. Specifically, a representative speaker with his/her s-vector (denoted as $\boldsymbol{\mu_2^*}$) is chosen from the dataset. Next, for each speech segment $\boldsymbol{x^{(i,n)}}$ of an arbitrary speaker $i$, its corresponding latent sequence variable $\boldsymbol{z_2^{(i,n)}}$ inferred from $\boldsymbol{x^{(i,n)}}$ is transformed to $\hat{\boldsymbol{z}}_{\boldsymbol{2}}^{\boldsymbol{(i,n)}} = \boldsymbol{z_2^{(i,n)}} - \boldsymbol{\mu_2^i} + \boldsymbol{\mu_2^*}$. Finally the decoder network of the FHVAE model reconstructs segment $\hat{\boldsymbol{x}}^{\boldsymbol{(i,n)}}$ based on $\boldsymbol{z_1^{(i,n)}}$ and $\hat{\boldsymbol{z}}_{\boldsymbol{2}}^{\boldsymbol{(i,n)}}$. This method is named as *s-vector unification* in the present work. The reconstructed features $\{\hat{\boldsymbol{x}}^{\boldsymbol{(i,n)}}\}$ are our desired speaker-invariant representation. They are expected to retain the linguistic content in $\{\boldsymbol{x^{(i,n)}}\}$ while capturing speaker characteristics corresponding to the representative speaker. In other words, the speech synthesized by $\{\hat{\boldsymbol{x}}^{\boldsymbol{(i,n)}}\}$ would ideally sound as if they were all spoken by the representative speaker.

# 3. Unsupervised subword modeling with speaker-invariant features

A DNN-BNF architecture [9,13] is adopted to perform phonetic discriminative training of untranscribed speech data and generate BNFs for subword modeling. In the literature, input features for DNN-BNF were MFCCs+VTLN [9] or fMLLRs estimated by a pre-trained out-of-domain ASR [13]. In this study, speaker-invariant features obtained by an FHVAE model are used as inputs to the DNN-BNF architecture.

## 3.1. DNN-BNF architecture

In the DNN-BNF architecture, given untranscribed speech of each target zero-resource language, Dirichlet process Gaussian mixture model (DPGMM) [24] algorithm is applied to cluster frame-level MFCC features. After clustering, cluster indices of speech frames are treated as pseudo phoneme-like transcription. A multilingual DNN with a linear bottleneck layer is trained with pseudo transcriptions and MFCCs of all the target languages using multi-task learning [25]. After training, subword discriminative BNFs are extracted for performance evaluation.

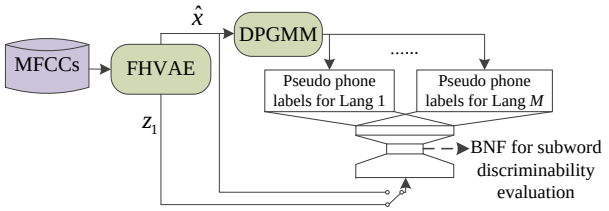


Figure 1: *DNN-BNF architecture with FHVAE-based speaker-invariant features for subword discriminability task*

### 3.2. DNN-BNF training with speaker-invariant features

Speaker-invariant features learned by FHVAEs are applied in 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 method described in Section 2.2. Compared to original MFCCs, FHVAE reconstructed MFCCs carry speaker information that is more consistent across utterances spoken by different speakers. DPGMM clustering is therefore able to generate better phoneme-like labels for target untranscribed speech and less affected by speaker variation.

During DNN-BNF model training, FHVAE-based speaker-invariant features are used as inputs instead of raw MFCCs. As seen in Figure 1, in this study we consider both types of features as DNN-BNF inputs, i.e. reconstructed MFCCs with s-vector unification $\{\hat{\boldsymbol{x}}\}$ and latent segment variables $\{\boldsymbol{z_1}\}$, and experimentally verify and compare their effectiveness.

## 4. Experimental setup

### 4.1. Dataset and evaluation metric

Experiments are carried out with development data of ZeroSpeech 2017 Track one [7]. The dataset consists of three languages, i.e., English, French and Mandarin. Each language contains separate training and test sets of untranscribed speech. Speaker identity information is only released for train sets. Test sets are organized into subsets of differing utterance length (1s, 10s and 120s). Detailed information is listed in Table 1. Note that in this Table, 'speakers-R/-L' denotes speakers with rich/limited speech data.

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

| | Training | | | Test |
|---|---|---|---|---|
| | Duration | #speakers-R | #speakers-L | 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 of ZeroSpeech 2017 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 segments $A$ 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. Dynamic time warping and cosine distance are used to measure segment- and frame-level dissimilarity, respectively.

### 4.2. FHVAE setup and parameter tuning

Most of the FHVAE model parameters are determined as suggested by [21]. 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 of 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. During the extraction of speaker-invariant features, 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 [26] with $\beta_1 = 0.95$ and $\beta_2 = 0.999$ is used to optimize the FHVAE discriminative segment variational lower bound. 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. Experiments are implemented in Tensorflow [27].

In our preliminary experiments, the ABX evaluation results of $\boldsymbol{z_1}$ were found to be sensitive to the input segment length $l$. This could be explained as: a too large $l$ would reduce $\boldsymbol{z_1}$'s capability in modeling linguistic content at subword-level; a too small $l$ would limit FHVAE encoder and decoder in capturing sufficient temporal dependencies which are essential in modeling speech. ABX performance results of $\boldsymbol{z_1}$ with different values of $l$ are shown in Fig. 2. It can be seen that the optimal value of $l$ is 10. For the remaining experiments in this work, $l$ is fixed to 10.

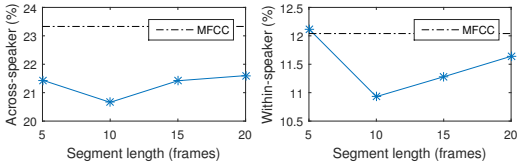


Figure 2: *ABX error rates (%) on $\boldsymbol{z_1}$ with different segment lengths and official MFCC baseline [7]. The results are averaged over three target languages.*

### 4.3. Selecting representative speaker for extracting reconstructed MFCCs

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 measure 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 ZeroSpeech 2017 training sets. All the above speakers belong to 'speakers-R'. The first half speakers inside each group are male and the second half are female[1]. During the extraction of $\{\hat{\boldsymbol{x}}\}$, $\{\boldsymbol{\mu_2^i}\}$ 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 task.

[1]Gender information is not provided in ZeroSpeech training data. We confirmed it through listening.

Table 2: *ABX error rates (%) on baseline and systems trained with 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 |
| | Topline [13] | | 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 [9] | | 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 [9] | | 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 |

### 4.4. DNN-BNF setup

For the baseline system without using FHVAE-based speaker-invariant features, input features to DPGMM are 39-dimensional MFCCs+$\Delta$+$\Delta\Delta$. 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 model is trained with all three languages' frame labels and cepstral mean normalized MFCCs+$\Delta$+$\Delta\Delta$ using multi-task learning [28]. The loss function is cross-entropy with equally-weighted language tasks. The dimensions of hidden layers are $\{1024 \times 5, 40, 1024\}$. After training, 40-dimensional BNFs for test sets are extracted and evaluated by ABX sub-word discriminability. DNN-BNF training is implemented in Kaldi. [29]

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 $\Delta$+$\Delta\Delta$. 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 $1400$. The DNN-BNF model is 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 ZeroSpeech 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 $\{\tilde{\boldsymbol{x}}\}$ were found to outperform $\{\hat{\boldsymbol{x}}\}$ in training the DNN-BNF model. The hidden layer dimensions and loss function of DNN-BNF are kept the same as in the baseline system.

## 5. Results and analyses

### 5.1. Effectiveness of reconstructed MFCCs

ABX performance on the 12 groups of reconstructed MFCCs $\{\hat{\boldsymbol{x}}\}$ using s-vector unification is shown in Figure 3. Each group of $\{\hat{\boldsymbol{x}}\}$ is presented as a bar inside each sub-figure. As a reference, the performance on latent segment variables $\{\boldsymbol{z_1}\}$ is marked as a dash-dotted line. 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 representative speaker cases, and are worse in all the female representative speaker cases. The results also indicate that selecting a male speaker as the representative is more effective than a female speaker in extracting $\{\hat{\boldsymbol{x}}\}$.

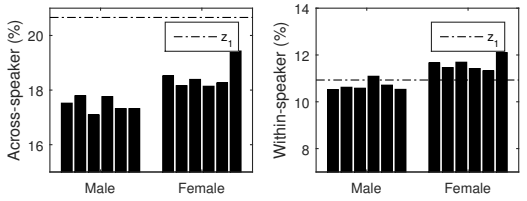


Figure 3: *ABX error rates (%) on $\hat{\boldsymbol{x}}$ using s-vector unification with different representative speakers. The results are averaged over three target languages.*

### 5.2. Baseline and systems trained with FHVAE-based speaker-invariant features

Experimental results of the baseline system and systems trained with FHVAE-based speaker-invariant features are summarized in Table 2. In the DPGMM clustering process, two groups of reconstructed MFCCs $\{\hat{\boldsymbol{x}}\}$ are tested as inputs. One group is corresponding to the representative speaker 's4018', while the other is corresponding to 's0107'. The former group is used to mimic the ideal scenario as it performs the best in across-speaker ABX evaluation ($17.09\%$, 3-rd bar from left in Fig. 3). The latter group mimics an average scenario as it performs moderately ($17.52\%$, 1-st bar from left in Fig. 3) among all the male speakers. In Table 2, the second and third columns of ID ① $\sim$ ⑥ denote input features to DNN-BNF training and DPGMM clustering, respectively. 'Orig.' means original MFCCs without reconstruction. The system exploiting a Cantonese ASR for fMLLR estimation [13] serves as the topline. From this Table, several observations are made:

(1) The DNN-BNF models trained with features $\{\tilde{\boldsymbol{x}}\}$ and $\{\boldsymbol{z_1}\}$ both outperform that trained with MFCCs. The baseline system and systems ① & ② differ only in DNN-BNF inputs. The results demonstrate the effectiveness in performing speaker-invariant feature learning towards DNN-BNF inputs.

(2) The reconstructed MFCC features $\{\hat{\boldsymbol{x}}\}$ significantly outperform original MFCCs in DPGMM-based frame labeling. In the ideal scenario where the representative speaker ('s4018') is carefully selected, by comparing ⑤ and ①, frame labeling based on $\{\hat{\boldsymbol{x}}\}$ contributes to $13.5\%/6.9\%$ relative improvements in across-/within-speaker conditions compared to that based on original MFCCs. Even in an average scenario where 's0107' is chosen, by comparing ③ and ①, the relative improvements are $11.3\%/4.7\%$. The results demonstrate the importance of employing speaker-invariant features as the input representation for DPGMM clustering.

(3) Our best system is able to close the topline and baseline across-speaker performance gap by $77\%$. The baseline-topline gap measures the impact of exploiting out-of-domain

transcribed speech resources as additional linguistic knowledge in unsupervised subword modeling. With FHVAE-based speaker-invariant learning, 77% of the performance gap can be compensated without requiring any out-of-domain data. Besides, our best system slightly outperforms topline in within-speaker condition.

We also compare the effectiveness of our proposed methods with VTLN implemented in [9], in which a similar DNN-BNF architecture was applied. As seen in Table 2, in across-speaker test condition, even though our baseline system is inferior to their baseline (MFCC), our best system outperforms their system (MFCC+VTLN) by relative 8%. The comparison shows that the FHVAE-based feature learning and transformation methods are more effective than VTLN in learning speaker-invariant features for unsupervised subword modeling.

## 6. Conclusions

This paper presents a study on improving unsupervised subword modeling by learning speaker-invariant features with FHVAE models. The FHVAEs disentangle linguistic content and speaker information encoded in speech in an unsupervised manner. By discarding or unifying speaker information, speaker-invariant features are learned and used as inputs to DNN-BNF frame labeling and model training. Experiments are conducted on ZeroSpeech 2017. Experimental results demonstrate the effectiveness of our proposed methods, especially in across-speaker evaluation scenario. The proposed methods, without requiring any out-of-domain resources, are able to make up 77% of the across-speaker performance gain benefited from exploiting out-of-domain transcribed speech as additional resources. The proposed methods also outperform VTLN in improving unsuperivised subword modeling.

## 7. Acknowledgements

This research is partially supported by a GRF project grant (Ref: CUHK 14227216) from Hong Kong Research Grants Council.

## 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] 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.

[9] 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.

[10] 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 representations," in *Proc. ASRU*, 2017, pp. 734–739.

[11] T. K. Ansari, R. Kumar, S. Singh, and S. Ganapathy, "Deep learning methods for unsupervised acoustic modeling - LEAP submission to zerospeech challenge 2017," in *Proc. ASRU*, 2017, pp. 754–761.

[12] H. Shibata, T. Kato, T. Shinozaki, and S. Watanabe, "Composite embedding systems for zerospeech2017 track 1," in *Proc. ASRU*, 2017, pp. 747–753.

[13] 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.

[14] 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.

[15] T. K. Ansari, R. Kumar, S. Singh, S. Ganapathy, and S. Devi, "Unsupervised HMM posteriograms for language independent acoustic modeling in zero resource conditions," in *Proc. ASRU*, 2017, pp. 762–768.

[16] 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.

[17] D. Kim, S. Umesh, M. J. Gales, T. Hain, and P. Woodland, "Using VTLN for broadcast news transcription," in *Proc. ICSLP*, 2004.

[18] T. Tsuchiya, N. Tawara, T. Ogawa, and T. Kobayashi, "Speaker invariant feature extraction for zero-resource languages with adversarial learning," in *Proc. ICASSP*, 2018, pp. 2381–2385.

[19] N. Zeghidour, G. Synnaeve, N. Usunier, and E. Dupoux, "Joint learning of speaker and phonetic similarities with siamese networks," in *Proc. INTERSPEECH*, 2016, pp. 1295–1299.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

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

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

[27] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean *et al.*, "Tensorflow: A system for large-scale machine learning," in *in Proc. OSDI*, 2016, pp. 265–283.

[28] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. ICASSP*, 2013, pp. 7304–7308.

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