Exploiting Speaker and Phonetic Diversity of Mismatched Language Resources for Unsupervised Subword Modeling

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 addresses the problem of learning robust frame-level feature representation for unsupervised subword modeling in the zero-resource scenario. Robustness of the learned features is achieved through effective speaker adaptation and exploiting cross-lingual phonetic knowledge. For speaker adaptation, an out-of-domain automatic speech recognition (ASR) system is used to estimate fMLLR features for untranscribed speech of target zero-resource languages. The fMLLR features are applied in multi-task learning of a deep neural network (DNN) to further obtain phonetically discriminative and speaker-invariant bottleneck features (BNFs). Frame-level labels for DNN training can be acquired based on two approaches: Dirichlet process Gaussian mixture model (DPGMM) clustering, and out-of-domain ASR decoding. Moreover, system fusion is performed by concatenating BNFs extracted by different DNNs. Our methods are evaluated by ZeroSpeech 2017 Track one, where the performance is evaluated by ABX minimal pair discriminability. Experimental results demonstrate that: (1) Using an out-of-domain ASR system to perform speaker adaptation of zero-resource speech is effective and efficient; (2) Our system achieves highly competitive performance to state of the art; (3) System fusion could improve feature representation capability.
Index Terms: zero resource, unsupervised learning, robust features, speaker adaptation, multi-task learning

1. Introduction

With the advances of deep neural network (DNN) based acoustic models (AMs) [1] and language models (LMs) [2], state-of-the-art automatic speech recognition (ASR) systems have demonstrated fairly impressive performance in terms of word accuracy [3, 4]. Typically the training of AMs requires hundreds to thousands of hours of transcribed speech. This leads to the fact that high-performance ASR systems are available only for major languages [5]. Even for resource-rich languages like English and Mandarin, preparing transcriptions for the available training speech requires a time-consuming task requiring considerable human effort. For many languages in the world, for which very little or no transcribed speech is available [6], conventional methods of AM training can not be directly applied.

Unsupervised acoustic modeling aims at modeling speech at subword or word level, assuming that only untranscribed raw speech are available [7–10]. This is often referred to as the zero-resource problem. The Zero Resource Speech Challenges 2015 (ZeroSpeech 2015) [11] and 2017 (ZeroSpeech 2017) [6] precisely focused on unsupervised speech modeling without transcription. ZeroSpeech 2017 was organized to tackle two sub-problems, namely *unsupervised subword modeling* (Track 1) and *spoken term discovery (STD)* (Track 2). Track 1 posed a research question of how to learn a frame-level feature representation that is discriminative for subword-level units and robust to linguistically irrelevant variations, e.g., speaker change, emotion, channel, etc. To this end, researchers proposed various feature types for comparison, such as posteriors [5, 12, 13] and BNFs [5,14,15]. Track 2 was focused on developing algorithms for discovering repeated speech patterns in audio streams. The problems concerned in the two tracks are closely related and essential in unsupervised speech modeling. Robust feature representation is found to be preferable as compared with conventional spectral features (e.g. MFCCs) for downstream applications like STD [16]. Whilst accurate STD results could be beneficial to DNN-based supervised learning of feature representation [9, 15, 17]. The present study addresses the problem of Track 1, learning of feature representation for unsupervised subword modeling.

Speaker adaption is critical to robust acoustic modeling. It is widely acknowledged that speaker adaptive training (SAT) is effective in improving ASR performance, especially for large vocabulary tasks [18–20]. Approaches to SAT are divided into two categories: model-based approaches, e.g., maximum likelihood linear regression (MLLR) [21], and feature-based approaches, e.g., feature-space MLLR (fMLLR) [22], i-vectors [23], speaker codes [24], and other appending features [25]. All of these methods are based on the assumption that speech transcription and/or speaker identity information are available. In the zero-resource case, Heck et al. proposed to estimate fMLLRs based on frame-level cluster labels that were obtained by a Dirichlet process Gaussian mixture model (DPGMM) algorithm [12]. The cluster labels are regarded as pseudo transcriptions for the target speech, which facilitate context-dependent GMM-HMM (CD-GMM-HMM) acoustic modeling with fMLLR features. This system demonstrated the best performance among all submitted systems in the ZeroSpeech 2017. Zeghidour et al. regards speaker adaptation as a problem of disentangling speaker information from phoneme information in speech [26]. They proposed to train subword and speaker same-different tasks within a triamese network, and demonstrated the effectiveness of disentanglement between these two types of information. This approach assumes availability of subword same-different supervision, which could be derived from STD results.

For major languages, high-performance ASR systems can be trained with large-scale speech corpora that cover hundreds of speakers [27, 28]. The richness of speaker diversity and linguistic variation in these out-of-domain corpora could be leveraged for learning robust feature representations in the zero-resource scenario. In [5], Shibata et al. made use of a Japanese ASR system to estimate fMLLR features and demonstrated very good performance in Track 1 of ZeroSpeech 2017. In the present study, we propose a framework of unsupervised learning of multilingual bottleneck features. To facilitate fMLLR-based SAT with untranscribed speech, frame-level labels are generated either by DPGMM clustering or using the state-level align-

ment information from an out-of-domain ASR system. BNFs are obtained from a multi-task learning DNN (MTL-DNN), which is trained with these two types of labels. As an alternative approach, the BNFs can be obtained directly from the same out-of-domain AMs. We also investigate on the efficacy of system fusion by concatenating the BNFs obtained from different DNNs.

2. Speaker adaptation with out-of-domain data

Feature-based speaker adaptation, e.g., fMLLR, is an effective approach to improving robustness of speech features. In the case of zero resource, we propose to leverage out-of-domain transcribed and speaker-annotated speech from a resource-rich language to model the speaker variation in target speech. Given the out-of-domain data, context-dependent GMM-HMM (CD-GMM-HMM) AMs are trained with raw spectral features. The models are used to forced-align the training data to provide supervision for vocal tract length normalization (VTLN) [29], linear discriminant analysis (LDA) [30], maximum likelihood linear transforms (MLLT) [31] and fMLLR estimation [32]. Subsequently CD-GMM-HMM models with SAT (CD-GMM-HMM-SAT) are trained, and used to estimate fMLLR transforms for the target zero-resource speech utterances. It is noted that the estimated fMLLR features of target speech can be used directly for subword modeling. The fMLLR features are expected to serve a better baseline than raw spectral features like MFCCs for subsequent feature representation learning and system building.

3. Frame labeling

Frame labeling is an essential step to prepare the target speech utterances for DNN based subword discriminative modeling. While frame labels are not needed in some of the DNN models like autoencoders (AEs), there were studies suggesting that AEs might not be a good approach to improving acoustic features [33]. In this study, two frame-labeling approaches are investigated, namely, DPGMM clustering and out-of-domain ASR decoding.

DPGMM is a non-parametric Bayesian extension to GMM, in which a Dirichlet process prior replaces the vanilla GMM [34]. One advantage of the DPGMM clustering algorithm is that the cluster number does not need to be pre-defined. This makes the algorithm very suitable for the problem of unsupervised acoustic modeling, as the number of basic speech units is unknown for a zero-resource language. Previous studies showed successful application of DPGMM to unsupervised word clustering [35], frame-level feature clustering for subword discriminative modeling [36] and unsupervised fMLLR-based speaker adaptive training [12, 37].

Let us consider M zero-resource languages. For the i-th language, frame-level fMLLR features, estimated as described in Section 2, are denoted as $\{\boldsymbol{x_1^i}, \boldsymbol{x_2^i}, \ldots, \boldsymbol{x_L^i}\}$, where L is the number of frames in the utterance. By applying DPGMM clustering, K Gaussian components are obtained to represent K clusters of frame-level features. The frame-level labels $\{l_1^i, l_2^i, \ldots, l_L^i\}$, are obtained by

$$l_t^i = \underset{1 \le k \le K}{\arg\max}\, p_{i,k}, \tag{1}$$

where $p_{i,k} = P(k|\boldsymbol{x_t^i})$ denotes the posterior probability of $\boldsymbol{x_t^i}$ with respect to the k-th Gaussian component.

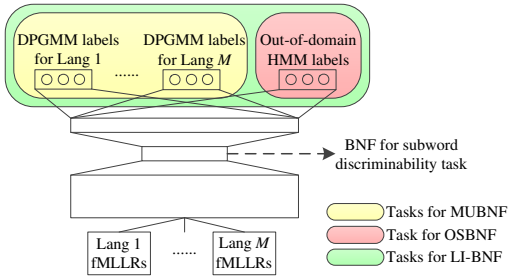

Figure 1: *DNN for extracting LI-BNF, MUBNF and OSBNF*

A Metropolis-Hastings based split/merge sampler is adopted for the inference of DPGMM parameters [34], following other studies on ZeroSpeech Challenges [12, 36].

Frame labeling can also be done with an out-of-domain ASR system, which is typically trained with a large amount of transcribed speech in a resource-rich language. The AM in such ASR system provides fine-grained speech representation of the language. Given an input speech utterance of the target zero-resource language, the out-of-domain ASR system can be applied to assign a language-mismatched state label to each frame of the utterance. It must be noted that the result of ASR decoding depends on the relative weighting of AM and LM. In our application, the LM carries a a very small weight, such that the acquired frame labels mainly reflect the acoustic properties of target speech.

4. Multi-task learning for BNFs

After obtaining frame-level labels for the training utterances of target speech, a DNN is trained with the fMLLR features, with the goal of extracting BNFs for subword modeling. BNFs have been shown to provide a compact and phonetically-discriminative representation, and suppress linguistically-irrelevant variation, e.g., speaker identity, of the input speech [38]. In this study, a multi-task learning (MTL) DNN is adopted in order to leverage the phonetic diversity in different speech tasks and different languages [39]. The proposed DNN architecture is shown as in Figure 1. The DNN supports a total of $M + 1$ learning tasks, which correspond to the M target zero-resource languages and the out-of-domain ASR. For the zero-resource language tasks, the frame labels are obtained by applying DPGMM clustering to training speech of the M target languages, while the out-of-domain ASR system generates an additional frame label. The hidden layers of the DNN, including a low-dimensional linear bottleneck layer, are shared across all tasks, while the soft-max output layers are task-specific. After multi-task training, the DNN is used to generate language-independent BNFs (LI-BNFs) for subword discriminability task.

It must be noted that one could also choose either the M language-dependent DPGMM label prediction tasks or the additional out-of-domain label prediction task for DNN training. As illustrated in Figure 1, the BNFs generated by these sub-tasks are denoted as multilingual unsupervised BNFs (MUBNFs) and the out-of-domain supervised BNFs (OSBNFs).

There are two main reasons why the MTL approach is adopted. First, there are two types of frame labels being investigated in this work, namely the DPGMM cluster labels and ASR decoding labels. The two tasks of label prediction are believed to be positively correlated and therefore are expected to benefit from MTL [39]. Second, one of the requirements of Ze-

roSpeech 2017 is that the learned feature representations for all target languages be generated by exactly the same system input and output. The idea of training a separate DNN for each target language would not satisfy the requirement[1].

5. System fusion

If there are multiple systems developed for feature representation learning which provide complementary information, fusion of these systems is expected to further improve the feature representation capability. System fusion can be done at either model level or output level. MTL is considered a kind of model-level fusion. LI-BNFs, as described in Section 4, can be considered as model-level fusion of MUBNFs and OSBNFs. Output-level system fusion can be realized by concatenating multiple feature representations. In this study, the effectiveness of output-level fusion is validated by:

1. Concatenating language-mismatched BNFs (LM-BNFs), which are obtained from the out-of-domain DNN-HMM AM, and LI-BNFs (LM-BNF + LI-BNF);
2. Concatenating LM-BNFs, MUBNFs and OSBNFs (LM-BNF + MUBNF + OSBNF).

6. Experiments

6.1. Dataset and evaluation metric

Experiments are carried out with development data of ZeroSpeech 2017 Track one [6]. The development data 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 made publicly known for train sets while unknown for test sets. Test sets are organized into subsets of differing utterance length (1s, 10s and 120s). Detailed information of the dataset is listed in Table 1.

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

	Training		Test
	Duration	#speakers	Duration
English	45 hrs	60	27 hrs
French	24 hrs	18	18 hrs
Mandarin	2.5 hrs	8	25 hrs

The evaluation metric of ZeroSpeech 2017 is ABX subword discriminability. Briefly speaking, 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 (DTW) and cosine distance are used to measure segment-level and frame-level dissimilarity, respectively.

6.2. Out-of-domain ASR system

A Cantonese ASR is selected as the out-of-domain ASR system. The ASR is trained with CUSENT, a read speech corpus developed by The Chinese University of Hong Kong [28]. There are $20,378$ training utterances from 34 male and 34 female speakers, with a total of 19.3 hours speech. Kaldi [40] is used to train two AMs, one is CD-GMM-HMM-SAT, the other is DNN-HMM. Target labels for DNN-HMM training are state alignments of CUSENT training data generated by CD-GMM-HMM-SAT model. Input features are 40-dimensional fMLLRs for CD-GMM-HMM-SAT, or fMLLRs by splicing with context size ± 5 for DNN-HMM. The fMLLR features are generated by performing VTLN towards 39-dimensional MFCCs+Δ+$\Delta\Delta$, and processed by splicing with context size ± 3 to estimate 40-dimensional LDA and MLLT, followed by fMLLR estimation. The total number of CD-HMM states are 2462. DNN-HMM has 7 hidden layers, with dimensions 440-1024×5-40-1024-2462, and sigmoid activation function except for the 40-dimensional linear bottleneck layer. A syllable trigram language model trained with transcriptions of CUSENT training data is used during decoding. The language model is trained with SRILM toolkit [41].

6.3. Speaker adaptation of target speech

The Cantonese ASR is used to perform fMLLR-based speaker adaptation of target zero-resource speech in a two-pass procedure. In the first-pass, target speech utterances are decoded by the ASR in a speaker-independent manner using unadapted features, from which initial fMLLR transforms are estimated. In the second-pass, target speech features transformed by the initial fMLLRs are decoded by the ASR in a speaker-adaptive manner. Subsequently, the final fMLLR transforms for the target speech are estimated.

6.4. Frame labeling and MTL-DNN training

Two frame labeling approaches are implemented. DPGMM clustering based frame labeling for target zero-resource speech is implemented with tools developed by Chang et al. [34]. Frame-level features for clustering are 40-dimensional fMLLRs for ZeroSpeech 2017 training sets. Frames of each language are clustered individually. The numbers of clustering iterations for English, French and Mandarin corpora are 120, 200 and 3000. After clustering, the numbers of obtained DPGMM clusters are 1118, 1345 and 596, respectively. Each frame is assigned with a DPGMM label.

The out-of-domain ASR based frame labeling is implemented by decoding target zero-resource speech by the Cantonese DNN-HMM ASR. After decoding, lattices are converted to one best path for each utterance, with LM to AM weight ratio set to 0.001. Each best path comprises a sequence of CD-HMM states of the Cantonese AM. These CD-HMM states are regarded as out-of-domain ASR based frame labels.

MTL-DNN is trained with 40-dimensional fMLLRs with context size ± 5 for training sets of three target zero-resource languages. There are 4 equally weighted tasks in MTL, 3 language-dependent DPGMM label prediction tasks and an out-of-domain Cantonese CD-HMM state prediction task. The neural network structure is 440-1024×5-40-1024-"*Block output layer*", where block softmax output layer dimensions for 4 tasks are $1118, 1345, 596$ and 2462, respectively. After MTL-DNN training, 40-dimensional language-independent BNFs (LI-BNFs) for test sets of target languages are extracted and used for ABX task. Similarly, multilingual unsupervised BNFs (MUBNFs), extracted by MTL-DNN with only the first 3 DPGMM label prediction tasks, and out-of-domain supervised BNFs (OSBNFs), extracted by STL-DNN with only the Cantonese CD-HMM state prediction task, are also used for ABX task. The dimensions of both MUBNFs and OSBNFs are 40.

[1] Although better performance was found by language-specific BNFs during our experiments, we do not report it in this paper.

Table 2: *ABX error rate (%) on the proposed methods, MFCC and state of the art of ZeroSpeech 2017*

	Within-speaker										Across-speaker										Avg.
	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 (MFCC) [6]	12.0	12.1	12.1	12.5	12.6	12.6	11.5	11.5	11.5	12.0	23.4	23.4	23.4	25.2	25.5	25.2	21.3	21.3	21.3	23.3	17.7
fMLLR	8.0	8.2	7.3	10.3	10.3	9.1	9.3	9.3	8.4	8.9	13.4	12.0	11.3	17.2	15.8	14.8	10.7	10.2	9.4	12.8	10.8
MUBNF	7.4	6.9	6.3	9.6	9.0	8.1	9.8	8.8	8.1	8.2	10.9	9.5	8.9	15.2	13.0	12.0	10.5	8.9	8.2	10.8	9.5
OSBNF	7.2	7.1	6.3	10.2	9.7	8.7	9.1	8.6	7.6	8.3	10.0	9.7	8.6	13.9	13.4	11.6	9.0	8.4	7.5	10.2	9.3
LI-BNF	6.9	6.6	6.1	9.5	9.2	8.4	9.2	8.5	7.9	8.0	10.0	**8.9**	**8.2**	14.3	12.9	11.5	9.5	8.5	7.7	10.2	9.1
LM-BNF	7.2	6.8	6.1	9.6	9.0	8.0	8.7	7.6	6.8	7.8	10.6	9.6	8.7	14.2	13.2	11.5	8.5	**7.6**	**6.7**	10.1	8.9
LM-BNF + LI-BNF	7.0	6.6	6.0	9.3	**8.8**	7.9	8.6	**7.5**	**6.7**	7.6	10.3	9.3	8.4	13.9	12.9	11.4	8.5	**7.6**	**6.7**	9.9	8.7
LM-BNF + MUBNF + OSBNF	**6.8**	**6.4**	**5.8**	**9.0**	**8.8**	**7.8**	**8.5**	7.7	6.8	**7.5**	**9.9**	9.0	**8.2**	**13.6**	**12.6**	**11.1**	**8.4**	7.7	**6.7**	**9.7**	**8.6**
Heck et al. [12]	6.9	6.2	6.0	9.7	8.7	8.4	8.8	7.9	7.8	7.8	10.1	8.7	8.5	13.6	11.7	11.3	8.8	7.4	7.3	9.7	8.8
System 1, Shibata et al. [5]	6.7	6.5	5.7	9.7	9.2	7.9	9.8	9.2	8.2	8.1	10.1	9.2	8.2	13.7	12.4	10.8	10.4	9.5	8.0	10.3	9.2

6.5. System fusion

For model-level system fusion approach, LI-BNFs can be considered as fusion of MUBNFs and OSBNFs. For output-level system fusion approach, two types of feature concatenation are implemented, i.e., concatenating LM-BNFs and LI-BNFs, resulting in 80-dimensional features, and concatenating LM-BNFs, MUBNFs and OSBNFs, resulting in 120-dimensional features. LM-BNFs are generated by feeding forward fMLLRs for target zero-resource languages to bottleneck layer of the Cantonese DNN-HMM AM. Attributes of the concerned BNFs are listed in Table 3. In this Table, unsupervised DPGMM labels are denoted as "Unsup.", while Cantonese CD-HMM state labels are denoted as "Sup.".

Table 3: *Attributes of LM-BNF, MUBNF, OSBNF and LM-BNF*

	LI-BNF	MUBNF	OSBNF	LM-BNF
Training method	MTL	MTL	STL	STL
Training data	ZeroSpeech 2017 fMLLR			CUSENT fMLLR
Label type	Sup. & Unsup.	Unsup.	Sup.	Sup.
Dimension	40			

7. Results and analyses

Experimental results of our proposed methods and state of the art of ZeroSpeech 2017 are summarized in Table 2. Baseline (MFCC) is released by challenge organizers. The sign "+" in Table 2 denotes output-level system fusion, i.e., feature concatenation. From Table 2, several observations are made.

(1) The fMLLR features consistently outperform MFCCs on all target zero-resource languages, with relative ABX error rate reduction 25.8% in within-speaker and 45.1% in across-speaker conditions. Note that in this system, training sets of ZeroSpeech 2017 data are not required. The results demonstrate that speaker adaptation based on an out-of-domain ASR system is effective and efficient for unsupervised subword modeling. The learned fMLLRs achieve larger ABX error rate reductions on long test utterances than on short ones. This is probably because fMLLR-based speaker adaptation does not work well on very short speech.

(2) MTL-DNN training with fMLLR features followed by system fusion brings the best performance. LI-BNFs, trained with both DPGMM labels and out-of-domain HMM state labels, outperform fMLLRs with relative ABX error rate reduction 10.1% in within-speaker and 20.3% in across-speaker conditions. Our best system concatenates LM-BNFs, MUBNFs and OSBNFs and achieves 7.5%/9.7% average ABX error rates in within/across-speaker conditions. This performance is highly competitive with the best submitted system for the challenge by Heck et al. [12] (7.8%/9.7%). It must be noted that system development in [12] does not rely on any out-of-domain resources, while our system uses a 19.3-hour Cantonese transcribed speech corpus. Our best system outperforms System 1 of Shibata et al. [5] (8.1%/10.3%) in both conditions. Note that a 240-hour Japanese transcribed speech corpus is used to develop System 1 of [5].

(3) Improved feature representation capability could be achieved by combining in-domain and out-of-domain resources with system fusion methods. Compared with MUBNFs, the advance of LI-BNFs is probably because the additional task of predicting out-of-domain CD-HMM state labels serves as a supplement to in-domain DPGMM label prediction tasks. DPGMM labels are generated in an unsupervised, purely data-driven manner, whilst out-of-domain CD-HMM state labels regularize in-domain data in a phonetically-aware form. On the other hand, the system of concatenating LM-BNFs, MUBNFs and OSBNFs achieves better ABX task performance than each of these single systems. The LM-BNFs, extracted by an out-of-domain DNN-HMM AM, provide language-mismatched phonetically-discriminative representation. By concatenating LM-BNFs, MUBNFs and OSBNFs, phonetic information in both domains is combined. The advance of feature concatenation method demonstrates the complementarity among BNFs extracted by in-domain and out-of-domain DNNs.

8. Conclusions

This paper presents a study on exploiting speaker and phonetic diversity of mismatched language resources for unsupervised subword modeling of zero-resource speech. Out-of-domain transcribed and speaker-annotated speech resources are employed to perform speaker adaptation of zero-resource speech. Frame labeling methods including DPGMM clustering and out-of-domain ASR decoding are adopted to provide frame-level labels for multi-task learning DNN (MTL-DNN) training. Bottleneck features (BNFs) extracted by MTL-DNN are used for ABX subword discriminability task. Moreover, system fusion is performed by concatenating BNFs extracted by different DNNs. Experiments are carried out with Zero Resource Speech Challenge 2017 Track one. Experimental results show that: (1) Speaker adaptation based on out-of-domain ASR system is effective and efficient; (2) Our best system achieves highly competitive performance to state of the art; (3) Model and output-level system fusion methods could improve feature representation capability.

9. Acknowledgements

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

10. References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] A. Ragni, E. Dakin, X. Chen, M. J. Gales, and K. M. Knill, "Multi-language neural network language models." in *Proc. INTERSPEECH*, 2016, pp. 3042–3046.

[3] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall, "English conversational telephone speech recognition by humans and machines," in *Proc. INTERSPEECH*, 2017, pp. 132–136.

[4] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM," in *Proc. INTERSPEECH*, 2017, pp. 949–953.

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

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

[7] J. Glass, "Towards unsupervised speech processing," in *Proc. ISSPA*, 2012, pp. 1–4.

[8] H. Kamper, A. Jansen, and S. Goldwater, "Fully unsupervised small-vocabulary speech recognition using a segmental bayesian model," in *Proc. INTERSPEECH*, 2015, pp. 678–682.

[9] R. Thiollière, E. Dunbar, G. Synnaeve, M. Versteegh, and E. Dupoux, "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling," in *Proc. INTERSPEECH*, 2015, pp. 3179–3183.

[10] S. Feng, T. Lee, and H. Wang, "Exploiting language-mismatched phoneme recognizers for unsupervised acoustic modeling," in *Proc. ISCSLP*, 2016, pp. 1–5.

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

[12] 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.

[13] 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.

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

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

[16] 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.

[17] C.-T. Chung, C.-Y. Tsai, H.-H. Lu, C.-H. Liu, H.-y. Lee, and L.-s. Lee, "An iterative deep learning framework for unsupervised discovery of speech features and linguistic units with applications on spoken term detection," in *Proc. ASRU*, 2015, pp. 245–251.

[18] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: A maximum likelihood approach to speaker normalization," in *Proc. ICASSP*, vol. 2, 1997, pp. 1043–1046.

[19] Y. Miao, H. Zhang, and F. Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors," *IEEE/ACM Trans. ASLP*, vol. 23, no. 11, pp. 1938–1949, 2015.

[20] X. Cui, V. Goel, and G. Saon, "Embedding-based speaker adaptive training of deep neural networks," in *Proc. INTERSPEECH*, 2017, pp. 122–126.

[21] M. J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[22] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems," in *Proc. INTERSPEECH*, 2010, pp. 526–529.

[23] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. ASRU*, 2013, pp. 55–59.

[24] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," *IEEE/ACM Trans. ASLP*, vol. 22, no. 12, pp. 1713–1725, 2014.

[25] X. Xie, X. Liu, T. Lee, and L. Wang, "RNN-LDA clustering for feature based DNN adaptation," in *Proc. INTERSPEECH*, 2017, pp. 2396–2400.

[26] 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.

[27] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. workshop on Speech and Natural Language*, 1992, pp. 357–362.

[28] T. Lee, W. K. Lo, P. C. Ching, and H. Meng, "Spoken language resources for Cantonese speech processing," *Speech Communication*, vol. 36, no. 3, pp. 327–342, 2002.

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

[30] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. ICASSP*, vol. 1, 1992, pp. 13–16.

[31] M. J. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Trans. SAP*, vol. 7, no. 3, pp. 272–281, 1999.

[32] D. Povey and K. Yao, "A basis representation of constrained MLLR transforms for robust adaptation," *Computer Speech & Language*, vol. 26, no. 1, pp. 35–51, 2012.

[33] 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.

[34] 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.

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

[36] 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.

[37] M. Heck, S. Sakti, and S. Nakamura, "Supervised learning of acoustic models in a zero resource setting to improve DPGMM clustering," in *Proc. INTERSPEECH*, 2016, pp. 1310–1314.

[38] F. Grézl, M. Karafiát, and L. Burget, "Investigation into bottleneck features for meeting speech recognition," in *Proc. INTERSPEECH*, 2009, pp. 2947–2950.

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

[40] 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*, 2011.

[41] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. ICSLP*, 2002, pp. 901–904.