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 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 un-transcribed 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 DNNs trained with both in-domain and out-of-domain data. 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 in speaker adaptation towards zero-resource speech is effective and efficient; (2) Our system achieves highly competitive performance to state of the art; (3) System fusion brings further performance gain.

Index Terms: zero resource, unsupervised learning, robust features, speaker adaptation, multi-task learning

1. Introduction

With successful application of deep neural network (DNN) in both acoutic models (AMs) [1] and language models (LMs) [2], state-of-the-art automatic speech recognition (ASR) systems have achieved high recognition accuracy [3, 4]. A typical AM is trained by hundreds or even thousands of hours of transcribed speech data. This leads to the fact that high performance ASR systems are available only for some major languages [5]. Even for resource-rich languages such as English and Mandarin, preparing transcription for speech data is highly time-consuming and needs enormous human effort. For the majority of the world's languages, for which few or no transcribed speech is available [6], conventional acoustic modeling techniques cannot be directly applied.

Recently, there has been a growing research interest in unsupervised acoustic modeling in zero-resource scenario [7–9], which aims at modeling speech at phoneme or word level, with the assumption that only raw speech data is available for system development. The Zero Resource Speech Challenge 2015 (ZeroSpeech 2015) [10] and 2017 (ZeroSpeech 2017) [6] precisely focus on unsupervised modeling of zero-resource speech data. The challenge 2017 includes two sub-problems, namely *unsupervised subword modeling* (track one) and *spoken term discovery (STD)* (track two). Track one poses a question as how to learn a frame-level feature representation which is discriminative towards basic speech subword units and robust to linguistically irrelevant variations, such as speaker identity, emotion, channel etc. Track two aims at discovering repeated speech fragments in the audio. These two tracks represent two important research aspects in unsupervised acoustic modeling, and are mutually closely correlated. Well learned feature representation for zero-resource speech is preferable as compared with raw spectral features like MFCCs in downstream applications such as STD [11]; Accurate STD results could serve either as weak same-different supervision [9, 12] or exact acoustic token labels [13] for neural network (NN) based feature representation learning. A joint optimization framework of both problems is proposed in [13]. The challenge 2017 has attracted researchers around the world [5, 12, 14–16]. In this paper, we focus on track one, unsupervised subword modeling.

A critical point in robust acoustic modeling is speaker adaptation. It has been widely acknowledged that speaker adaptive training (SAT) is one of the key issues in improving ASR performance, especially for large vocabulary continuous speech recognition (LVCSR) [17–19]. SAT approaches can be divided into two categories, i.e., model based approaches such as maximum likelihood linear regression (MLLR) [20], and feature based approaches such as feature-space MLLR (fMLLR) [21], i-vectors [22], speaker codes [23] and other appending features [24, 25]. These methods are all based on the assumption that transcribed or at least speaker annotated speech are available. In recent years, there has been a research interest in unsupervised SAT for zero-resource speech data. Heck et al. proposed a clustering based method to perform fMLLR estimation [14]. They first cluster raw spectral features by a Dirichlet process Gaussian mixture model (DPGMM) algorithm [26, 27] to obtain frame-level cluster labels. These labels are regarded as pseudo transcription for target speech and used for supervised context-dependent GMM-HMM (CD-GMM-HMM) acoustic modeling with fMLLR-based SAT. Heck et al. [14] achieved the first place in the challenge 2017 [6]. Shibata et al. [5] performed speaker adaptation by firstly developing a standard CD-GMM-HMM AM with fMLLR-based SAT using out-of-domain transcribed data, followed by estimating fMLLRs for target zero-resource speech.

For some major languages, there are large quantities of transcribed speech corpora covering tens or hundreds of speakers available [28, 29]. The richness of both speaker and phonetic variations in out-of-domain corpora could be leveraged to assist unsupervised feature representation learning and speaker adaptation for zero-resource speech. Motivated by this, our proposed methods can be summarized as in two parts. In the first part, we incorporate fMLLR-based speaker adaptation in unsu-

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

pervised multilingual bottleneck feature (BNF) learning framework [27], where fMLLR features for in-domain speech are estimated by AM of an out-of-domain ASR system. Apart from the DPGMM frame labeling method proposed in [27], we propose to use an out-of-domain ASR system to decode in-domain speech and generate HMM state alignment as the second type of frame labels. A multi-task learning DNN (MTL-DNN) is then trained with two types of labels and used to extract BNFs as the learned representation for evaluation. In the second part, an out-of-domain DNN-HMM AM directly feeds forward in-domain speech and generate BNFs for evaluation. Furthermore, we also investigate on the efficacy of system fusion by concatenating BNFs extracted by multiple systems.

The feature type of unsupervised subword modeling is not constrained by challenge organizers. Researchers proposed various feature types for comparison such as posteriors [5, 14, 30] and BNFs [5, 12, 16]. This paper focuses on BNFs.

2. Speaker adaptation using resource-rich language

We propose to leverage out-of-domain transcribed and speaker annotated speech data for a resource-rich language to model speaker variation in target zero-resource speech, in order to learn feature representation of target speech that de-emphasizes speaker identity.

Given out-of-domain data for a resource-rich language, the procedure of developing a GMM-HMM AM would normally consist of following steps. At the first step, a CD-GMM-HMM is trained with raw spectral features appended with their derivatives. The CD-GMM-HMM forces align training data to provide supervision for vocal tract length normalization (VTLN) [31], linear discriminant analysis (LDA) [32], maximum likelihood linear transforms (MLLT) [33] and fMLLR [34] estimation. A CD-GMM-HMM model with SAT (CD-GMM-HMM-SAT) is then trained, and used to estimate fMLLR transforms for target zero-resource speech utterances. Note that fMLLR features for target speech can be directly used for subword discriminability task evaluation. Moreover, fMLLR features are expected to serve a better baseline than raw spectral features like MFCCs in subsequent feature representation learning and system building.

3. Frame labeling

Frame labeling aims at finding supervision for target speech frames that can be used for downstream NN based subword discriminative modeling. Although some NN based modeling architectures such as autoencoders (AEs) avoid the need of frame label acquisition, past work has proved that AEs do not show promising results on acoustic feature improvement [35]. This paper adopts two frame labeling approaches, namely, DPGMM clustering and out-of-domain ASR decoding.

3.1. DPGMM clustering based frame labeling

DPGMM is an extension to GMM in a non-parametric Bayesian way in which a Dirichlet process prior replaces the vanilla GMM [26]. One advantage of DPGMM clustering algorithm is that the cluster number does not need to be pre-defined, which is intrinsically suitable for unsupervised acoustic modeling, as prior knowledge on the actual number of basic speech units of a zero-resource language is usually unknown. Several past works have shown successful application of DPGMM to unsupervised speech word clustering [36], frame-level feature clustering for subword discriminative training [27] and unsupervised fMLLR-based SAT to improve feature representation robustness towards speaker variation [14, 37].

Let us consider M target zero-resource languages. For the i-th language, frame-level fMLLR features, estimated as described in Section 2, are denoted as $\{\boldsymbol{x_1^i}, \boldsymbol{x_2^i}, \ldots, \boldsymbol{x_L^i}\}$. After DPGMM clustering, K Gaussian components are obtained. Frame-level fMLLR features are then transcribed to frame labels $\{l_1^i, l_2^i, \ldots, l_L^i\}$, where the t-th frame label l_t^i is generated as

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

here $p_{i,k} = P(k|\boldsymbol{x_t^i})$ is the posterior probability of $\boldsymbol{x_t^i}$ belonging to the k-th Gaussian component.

For the inference of DPGMM parameters, a Metropolis-Hastings based split/merge sampler is adopted [26], as referred to past related works on ZeroSpeech Challenges [14, 27].

3.2. Out-of-domain ASR based frame labeling

A well-developed ASR system trained by transcribed speech of a certain language provides fine-grained representation of speech in that language. For some major language, large quantities of transcribed speech corpora are available, with which a high accuracy ASR system for that language can be developed. In this work, an out-of-domain ASR decodes target speech and assign each frame with a language-mismatched HMM state label which is defined in AM of the out-of-domain ASR. It is worth noting that decoding results of HMM state sequences depend on LM to AM weight ratio. In this paper, the weight ratio is set to a very small value as we expect the acquired frame labels could mainly reflect acoustic properties of target speech represented by HMM states of an out-of-domain AM.

4. Multi-task learning for BNFs

After frame labeling, a DNN is trained with fMLLRs of target speech and frame-level labels, in order to extract BNFs for subword discriminability task. BNFs are shown by previous works to carry abundant and compact phonetically-discriminative information and suppress linguistically irrelevant variations such as speaker identity [38]. This paper investigates the use of multi-task learning (MTL) [39] in DNN training. Our proposed DNN structure is shown in Figure 1. There are in

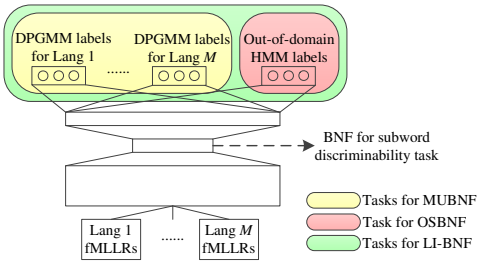

Figure 1: *DNN for extracting LI-BNF, MUBNF and OSBNF*

total $M + 1$ tasks, where M denotes the number of target zero-resource languages. The first M tasks predict language-dependent DPGMM labels for M languages, while the additional language-independent task predicts HMM state labels generated by decoding results of an out-of-domain ASR system. Hidden layers, including a low-dimensional linear bottleneck layer, are shared between all target languages, while

soft-max output layers are task-specific. After training, the language-independent BNFs (LI-BNFs) are used for subword discriminability task. Note that one could also choose either M language-dependent tasks or the language-independent task for DNN training, and extract BNFs for evaluation. To differentiate from LI-BNFs, as illustrated in Figure 1, in this paper BNFs extracted by a DNN trained in the former case is denoted as multilingual unsupervised BNFs (MUBNFs), while in the latter case is denoted as out-of-domain supervised BNFs (OSBNFs).

There are two main reasons for us to perform MTL instead of single-task learning (STL). First, there are two frame labeling approaches investigated in this work, one based on DPGMM, the other based on an out-of-domain ASR. The tasks of predicting two types of labels are believed to be positively correlated, which satisfies the requirements of benefiting from MTL [39]. Second, one requirement of the challenge 2017 is that the learned feature representation for all target languages be generated by the exact same system input-output. The idea of training a separate DNN for each target language is on the contrary with this[1].

5. System fusion

If multiple feature representation learning systems provide complementary information for target speech subword discriminability, fusion of these systems is believed to improve feature representation capability. System fusion can be made in different levels, e.g., model-level and output-level. MTL is one kind of model-level system fusion, as it composes several STL or MTL sub-systems. LI-BNFs, as described in Section 4, can be considered as model-level fusion of MUBNFs and OSBNFs. Output-level system fusion is realized by directly concatenating features for the same speech frame extracted by multiple original systems. In this work, the effectiveness of output-level system fusion is validated by concatenating language-mismatched BNFs (LM-BNFs), which are extracted by an out-of-domain DNN-HMM AM, and LI-BNFs, or concatenating LM-BNFs, MUBNFs and OSBNFs together.

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 un-transcribed speech. Speaker identities are made publicly known for train sets while unknown for test sets. Test utterances lie in one of three lengths, 1s, 10s and 120s. Detailed information of data 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	6 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", etc). Each pair of segments A and B belong to the same speaker. ABX error rates for *within-talker* and *across-talker* are evaluated separately, depending on whether X belongs to the same speaker as $A(B)$. Dynamic time warping (DTW) and cosine distance are used to measure segment-level and frame-level dissimilarity, respectively.

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

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 [29]. There are $20,378$ training utterances from 34 male and 34 female speakers, with total 19.3 hours data size. Kaldi [40] is used to train two versions of acoustic models, one is CD-GMM-HMM-SAT, the other is DNN-HMM. DNN-HMM is trained with alignment generated by CD-GMM-HMM-SAT model. Input features are 40-dimensional fMLLRs for CD-GMM-HMM-SAT model, or fMLLRs by splicing with context size ± 5 for DNN-HMM. The fMLLR features are estimated by firstly performing VTLN towards 39-dimensional MFCCs+Δ+$\Delta\Delta$, followed 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. Sigmoid is chosen as nonlinear activation function. A syllable trigram LM is trained with CUSENT training data transcription by SRILM [41].

6.3. Speaker adaptation of target speech

The Cantonese ASR is used to perform fMLLR-based speaker adaptation of target speech in a two-pass procedure. In the first-pass, target speech is 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. Finally the 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 unsupervised speech is implemented using tools developed by Chang et al. [26]. The features used in clustering are 40-dimensional fMLLRs for ZeroSpeech 2017 training sets. Frames of each language are clustered individually. Clustering iterations for English, French and Mandarin corpora are $120, 200$ and 3000, resulting in cluster numbers $1118, 1345$ and 596, respectively. After training, each frame is assigned with a DPGMM label. On the other hand, the out-of-domain ASR based frame labeling is implemented by decoding target speech of three languages using 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 defined in AM of the Cantonese ASR. 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 languages. There are 4 equally weighted tasks in MTL, 3 language-dependent DPGMM label prediction tasks and a language-mismatched Cantonese CD-HMM state prediction task. The NN structure

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

	Within-talker										Across-talker										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. [14]	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

is 440-1024 $\times$ 5-40-1024-"Block softmax layers", where block softmax layer dimensions for 4 tasks are $1118, 1345, 596$ and 2462, in sequence. After MTL-DNN training, 40-dimensional language-independent BNFs (LI-BNFs) for test sets of three languages are extracted and used for ABX task. Similarly, multilingual unsupervised BNFs (MUBNFs), extracted by MTL-DNN with only 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.

6.5. System fusion

For model-level system fusion approach, LI-BNFs can be regarded as a fusion of MUBNFs and OSBNFs. For output-level system fusion approach, two configurations are made, i.e., appending LM-BNFs to LI-BNFs, resulting in 80-dimensional features, and concatenating LM-BNFs, MUBNFs and OSBNFs together, resulting in 120-dimensional features. LM-BNFs are extracted by feeding forward fMLLRs of ZeroSpeech 2017 test sets to bottleneck layer of Cantonese DNN-HMM AM. Attributes of the concerned BNFs are listed in Table 3. In this

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			

Table, unsupervised DPGMM labels are denoted as "Unsup.", while Cantonese CD-HMM state labels are denoted as "Sup.".

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 by feature concatenation. From this Table, several observations are made.

(1) FMLLR features consistently outperform MFCCs on all target languages, reducing average ABX error rates with relative 33.3% and 45.1% in within-talker and across-talker evaluation conditions, respectively. At this stage training sets of ZeroSpeech 2017 are not used. Speaker adaptation based on an out-of-domain ASR system is effective and efficient for unsupervised subword modeling. The learned fMLLRs achieve more ABX error rate reduction 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 bring the best performance. LI-BNFs, trained with both DPGMM labels and out-of-domain HMM state labels, reduce within and across-talker average ABX error rates by relative 10.1% and 20.3%, as compared with fMLLRs. Our best system concatenates LM-BNFs, MUBNFs and OSBNFs and achieves 7.5%/9.7% average ABX error rates in within/across-talker conditions, which is highly competitive with the best submitted system for the challenge by Heck et al. [14] (7.8%/9.7%). It must be noted that system development in [14] does not rely on any out-of-domain resources, while our system uses 19.3 hours Cantonese transcribed speech. Our best system outperforms System 1 of Shibata et al. [5] (8.1%/10.3%) in both conditions. Note that [5] uses 240 hours Japanese transcribed speech in System 1 development.

(3) Better unsupervised feature representation can be learned when the combination of in-domain and out-of-domain resources is realized by any of model and output-level system fusion. Compared with MUBNF, the advantage of LI-BNF is probably because the additional task of predicting out-of-domain CD-HMM state labels serves as 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, LM-BNF, extracted by an out-of-domain DNN, gives compact language-mismatched phonetically-discriminative representation. By concatenating LM-BNFs, MUBNFs and OSBNFs, phonetically-discriminative information in both domains is combined and lead to better feature representation than any one of LM-BNFs, MUBNFs and OSBNFs.

8. Conclusions

This paper presents a study on exploiting knowledge from resource-rich languages for unsupervised feature representation learning of zero-resource speech. Out-of-domain transcribed and speaker annotated speech is employed to perform speaker adaptation towards zero-resource speech. Frame labeling methods including DPGMM clustering and out-of-domain ASR decoding are adopted to provide supervision 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 in-domain and out-of-domain trained 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 both improve subword discriminability of the learned feature representation.

9. 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] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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.

[23] 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.

[24] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP*, 2013, pp. 7398–7402.

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

[27] 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.

[28] 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*. Association for Computational Linguistics, 1992, pp. 357–362.

[29] 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.

[30] 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.

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

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[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.