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

Speaker Adversarial Training of DPGMM-based Feature Extractor for Zero-Resource Languages

Yosuke Higuchi, Naohiro Tawara, Tetsunori Kobayashi, Tetsuji Ogawa

Department of Communication and Computer Engineering, Waseda University, Tokyo, Japan

Abstract

We propose a novel framework for extracting speaker-invariant features for zero-resource languages. A deep neural network (DNN)-based acoustic model is normalized against speakers via adversarial training: a multi-task learning process trains a shared bottleneck feature to be discriminative to phonemes and independent of speakers. However, owing to the absence of phoneme labels, zero-resource languages cannot employ adversarial multi-task (AMT) learning for speaker normalization. In this work, we obtain a posteriorgram from a Dirichlet process Gaussian mixture model (DPGMM) and utilize the posterior vector for supervision of the phoneme estimation in the AMT training. The AMT network is designed so that the DPGMM posteriorgram itself is embedded in a speaker-invariant feature space. The proposed network is expected to resolve the potential problem that the posteriorgram may lack reliability as a phoneme representation if the DPGMM components are intermingled with phoneme and speaker information. Based on the Zero Resource Speech Challenges, we conduct phoneme discriminant experiments on the extracted features. The results of the experiments show that the proposed framework extracts discriminative features, suppressing the variety in speakers.

Index Terms: speech recognition, zero-resource language, embeddings, Dirichlet process Gaussian mixture model, adversarial multi-task learning

1. Introduction

A number of speech processing technologies rely on a large amount of artificially labeled data. In contrast, having no access to precise transcriptions or parallel data, speech processing on zero-resource languages has attracted increasing attention in recent years. Different types of topics have been studied in such unsupervised setting: spoken query detection [1, 2], discovery of subword units [3, 4], topic segmentation [5], and document classification [6]. In this paper, we focus on speech recognition on the zero-resource language, and to this end, subword units (e.g., phonemes) must be acquired without any specific knowledge of the target language.

Several attempts have been made to identify subword units from unlabeled speech data. The Dirichlet process Gaussian mixture model (DPGMM) has frequently been applied to generate subword units on the target language in an unsupervised clustering manner [7]. Given the speech features, the model automatically estimates the parameters of the clusters and the optimized components can be treated as clusters of subwords in the target language. Deep neural networks (DNNs) have also been adopted to obtain the fine representation of subword units, based on manifold learning [8], auto-encoder [9, 10], and a multilingual bottleneck feature [11, 12]. A posteriorgram from the DPGMM can be used as a feature that is discriminative to phonemes as well.

In general, acoustic signals contain several sources of variation such as phonemes, noises, channels, and speakers. To build a robust speech recognition system, it is important not only to keep the information that contributes to distinguishing the phonemes but also to normalize the factors that are irrelevant or a nuisance for the classification. Vocal tract length normalization (VTLN) [13] and feature-space maximum likelihood linear regression (fMLLR) [14] have been commonly applied to reduce the variety in acoustic features from different speakers. These methods learn transformation matrices that maximize the likelihood of an observed sequence of features given a pre-trained acoustic model. Speaker-normalized features are obtained by converting source features with the learned matrices. An adversarial training technique in DNNs [15, 16] has also been applied to learn robust bottleneck features in a multi-tasking manner. For example, in [17], a DNN-HMM acoustic model is designed to be robust to noise. By installing a noise classification network in an adversarial manner into the bottleneck layer, the bottleneck features are trained to degrade the noise classification accuracy. Adversarial multi-task (AMT) training is also adopted to build a speaker-robust acoustic model [18, 19]. These types of approaches, however, cannot be applied to a zero-resource language because the trained decoder of the target language is generally inaccessible. To tackle this problem, [20, 21] use a DPGMM-based acoustic model as a decoder to estimate the transformation parameters of fMLLR. In this paper, we discuss the use of DPGMM in AMT training for speaker normalization in the zero-resource language.

Our work aims to obtain frame-wise speech features that are discriminative to phonemes and independent of speakers. The proposed feature extraction process mainly consists of two parts, the acquisition of phoneme supervision by DPGMM clustering and the speaker normalization by AMT training. First, a DPGMM is trained on frame-wise speech features and a posteriorgram of the DPGMM given the observed feature is calculated for each frame. Then, using posteriorgrams as the supervision of phonemes for each speech frame, a speaker adversarial network is trained so that its bottleneck feature becomes speaker-invariant. [22] labels speech frames using the index of the DPGMM components with the highest likelihood. We use the posteriorgram itself for the supervision of phonemes, considering the loss of information that describes phonemes. Due to the variability of factors in speech features, the components estimated by DPGMM clustering may be mixed up with phonemes and speaker clusters, making the posteriorgrams unreliable in terms of phoneme information. Based on this assumption, it may not be appropriate to use posteriorgrams for the supervision of phonemes in multi-task learning. With this consideration in mind, we redesigned the conventional AMT network by directly applying speaker-adversarial training to the posteriorgrams obtained through the DPGMM. In this network, the posteriorgrams are expected to be embedded into a feature-

http://dx.doi.org/10.21437/Interspeech.2019-2052

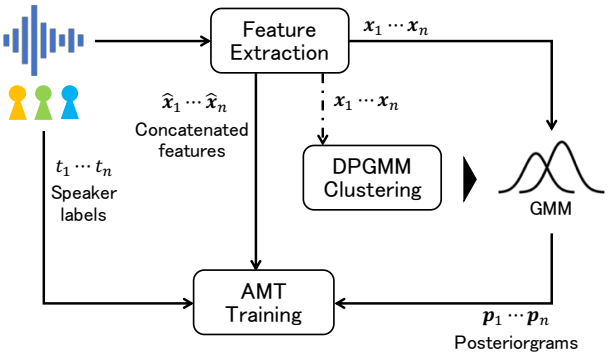

Figure 1: *Feature extraction framework*

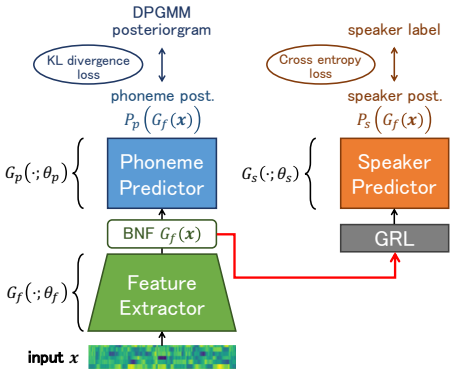

Figure 2: *Conventional AMT network*

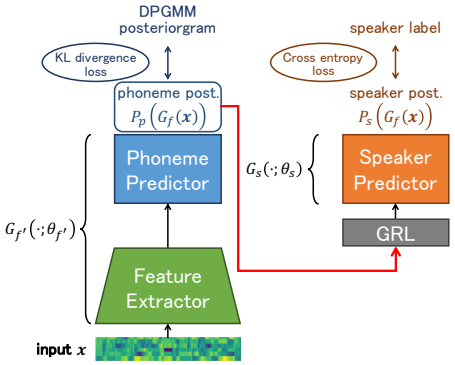

Figure 3: *Proposed AMT network*

space that is independent of speakers.

To evaluate the features extracted by our framework, we conducted the ABX phone discriminability test of the Zero Resource Challenge 2015, and 2017[1] - Track 1. These experiments demonstrate that our approach is applicable to speech data, whether it is zero-resource or not.

2. Speaker-invariant feature extraction

Figure 1 represents an overview of the proposed feature extraction framework for zero-resource languages. The model is constructed using the following processes:

1. Extract features from speech data in a target language
2. Construct a DPGMM with the speech features
3. Based on calculations for the posterior probabilities of the DPGMM components given the speech features, obtain frame-wise DPGMM posteriorgrams.
4. Given speaker labels, train an AMT network with the DPGMM posteriorgrams.

After model construction, the information gathered using the feature extractor of the adversarial network is expected to be speaker-invariant. The following subsections explain DPGMM clustering for phonetic information extraction and AMT training for speaker normalization.

2.1. Acquisition of phonetic information through DPGMM clustering

Regarded as an infinite Gaussian mixture model, DPGMM is adopted for the unsupervised acoustic modeling of zero-resource languages [7]. The model is constructed using a non-parametric fully-Bayesian approach, in which the complexity and the number of Gaussian components are optimized automatically. Based on the assumption that each component can be considered as a phoneme-like unit, the posterior probabilities of the components for a given speech frame can be used as features containing phonetic information.

The posterior probability of the k-th cluster, given a speech feature at the i-th frame is computed as follows:

$$p_{i,k} = p(c_k|\boldsymbol{x}_i) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}, \qquad (1)$$

[1] https://zerospeech.com/index.html

where $X = \{\boldsymbol{x}_i\}_{i=1}^{N}$ denotes feature vectors; K, the number of components; $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^{K}$, mixture weights; $\boldsymbol{\mu} = \{\boldsymbol{\mu}_k\}_{k=1}^{K}$, mean vectors; $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_k\}_{k=1}^{K}$, covariance matrices. From Eq. (1), the DPGMM posteriorgram of $\boldsymbol{x}_i$ at the i-th frame is represented as follows:

$$P(\boldsymbol{x}_i) = \{p_{i,1}, ..., p_{i,K}\} \quad (i = 1, ..., N). \qquad (2)$$

2.2. Conventional AMT learning for speaker invariant bottleneck feature

Through speaker-adversarial training for the DNN classifying the phonemes from the feature vectors, the variability in the speakers is suppressed from the bottleneck feature, the output of a hidden layer [18, 19]. The model consists of three networks: feature extractor, phoneme predictor, and speaker predictor. The feature extractor converts an input vector into a bottleneck feature. Given a bottleneck feature, the phoneme and the speaker predictors calculate posterior probabilities for each of their target classes. The predictors are optimized simultaneously to minimize classification losses in a multi-task manner. Given the classification losses, the feature extractor is optimized to maximize the accuracy of the phoneme classification as well as to minimize the accuracy of the speaker classification. This adversarial training against speakers leads to a bottleneck feature that is easy to classify for phonemes and difficult for speakers, yielding a speaker-invariant feature.

To leverage the conventional AMT training for speaker normalization, input features must be labeled with both phonemes and speakers. However, in the absence of acoustic models, zero-resource languages have no phoneme labels while speak-

ers can easily be distinguished. To overcome this problem, we use DPGMM posteriorgrams as the target posterior distribution of the phoneme predictor. Figure 2 shows the AMT training for a speaker invariant bottleneck feature using DPGMM posteriorgrams. The loss function for the phoneme predictor is designed using KL divergence between the DPGMM posteriorgram and the phoneme posterior. Let $\boldsymbol{\theta}_f$, $\boldsymbol{\theta}_p$, and $\boldsymbol{\theta}_s$ be the parameters of the feature extractor G_f, the phoneme predictor G_p, and the speaker predictor G_s, respectively. The phoneme prediction loss $\mathcal{L}_p$ is calculated as follows:

$$\mathcal{L}_p(\boldsymbol{\theta}_f, \boldsymbol{\theta}_p) = D_{KL}\big(P(\boldsymbol{x})||P_p(G_f(\boldsymbol{x}; \boldsymbol{\theta}_f); \boldsymbol{\theta}_p)\big), \quad (3)$$

where $P(\cdot)$ denotes the DPGMM posteriorgram computed from Eq. (2) and $P_p(\cdot)$ the phoneme posterior. The cross-entropy loss $\mathcal{L}_s$ for the speaker predictor is defined as follows:

$$\mathcal{L}_s(\boldsymbol{\theta}_f, \boldsymbol{\theta}_s) = -\mathbb{E}_{t\sim p(t|\boldsymbol{x})}\big[\log P_s(t|G_f(\boldsymbol{x}; \boldsymbol{\theta}_f); \boldsymbol{\theta}_s)\big], \quad (4)$$

where $P_s(\cdot)$ denotes the speaker posterior and $t \sim p(t|\boldsymbol{x})$ denotes the speaker label t of the input feature $\boldsymbol{x}$.

The parameters are updated using Stochastic Gradient Decent (SGD) using the minibatch algorithm as follows:

$$\boldsymbol{\theta}_p \leftarrow \boldsymbol{\theta}_p - \mu\frac{\partial \mathcal{L}_p}{\partial \boldsymbol{\theta}_p}, \quad (5)$$

$$\boldsymbol{\theta}_s \leftarrow \boldsymbol{\theta}_s - \mu\frac{\partial \mathcal{L}_s}{\partial \boldsymbol{\theta}_s}, \quad (6)$$

$$\boldsymbol{\theta}_f \leftarrow \boldsymbol{\theta}_f - \mu\left(\frac{\partial \mathcal{L}_p}{\partial \boldsymbol{\theta}_f} - \lambda\frac{\partial \mathcal{L}_s}{\partial \boldsymbol{\theta}_f}\right), \quad (7)$$

where μ denotes the learning rate, and λ the parameter of the gradient reversal layer (GRL) [15, 16], which causes the training to be adversarial to speaker classification. GRL acts as an identity function in forward propagation and inverses the sign of losses in back propagation.

2.3. Proposed AMT learning for speaker invariant phoneme posterior

Due to speaker variability, the DPGMM clusters generated observing speech data are not necessarily subword units. Based on this assumption, using DPGMM posteriograms for the supervision of the phonemes is inconsistent in conventional multi-task learning. To analyze the problem, we redesign the adversarial network architecture to directly suppress speaker information in DPGMM posteriorgrams.

Figure 3 shows the concept of the proposed multi-task network. The feature extractor is combined with the phoneme predictor to convert an input vector directly into a phoneme posterior, with the aim of approximating the DPGMM posteriorgram. Taking this approximated posteriorgram as an input, the speaker predictor calculates posterior probabilities for speaker classes. Using training that is adversarial to the speaker predictor, the feature extractor can eliminate feature factors that contribute to speaker classification, yielding a speaker-invariant DPGMM posteriorgram.

Let $\boldsymbol{\theta}_{f'}$, and $\boldsymbol{\theta}_s$ be the parameters of the feature extractor $G_{f'}$, and the speaker predictor G_s, respectively. Using KL divergence, the loss of the feature extractor's approximation of the DPGMM posteriogram is defined as follows:

$$\mathcal{L}_{f'}(\boldsymbol{\theta}_{f'}) = D_{KL}\big(P(\boldsymbol{x})||P_{f'}(\boldsymbol{x}; \boldsymbol{\theta}_{f'})\big), \quad (8)$$

where $P_{f'}(\cdot)$ denotes the posteriorgram of the modified feature extractor. The cross-entropy loss for the speaker predictor is defined as follows:

$$\mathcal{L}_s(\boldsymbol{\theta}_{f'}, \boldsymbol{\theta}_s) = -\mathbb{E}_{t\sim p(t|\boldsymbol{x})}\big[\log P_s(t|P_{f'}(\boldsymbol{x}; \boldsymbol{\theta}_{f'}); \boldsymbol{\theta}_s)\big]. \quad (9)$$

With the same conditions as in Eq. (5) $\sim$ (7), the parameters are updated as follows:

$$\boldsymbol{\theta}_s \leftarrow \boldsymbol{\theta}_s - \mu\frac{\partial \mathcal{L}_s}{\partial \boldsymbol{\theta}_s}, \quad (10)$$

$$\boldsymbol{\theta}_{f'} \leftarrow \boldsymbol{\theta}_{f'} - \mu\left(\frac{\partial \mathcal{L}_{f'}}{\partial \boldsymbol{\theta}_{f'}} - \lambda\frac{\partial \mathcal{L}_s}{\partial \boldsymbol{\theta}_{f'}}\right). \quad (11)$$

While a conventional network aims to learn the speaker-invariant bottleneck representation, the proposed network works to embed the DPGMM posteriorgram into the speaker-invariant feature space. The proposed network can be applied not only to general low-resource problems but also to resource-abundant data to yield a fine representation. In the experiment, we evaluated both the zero-resource-like language and the resource-abundant languages to confirm the versatility of the proposed network.

Table 1: *Training Data*

Language	#speakers	dur. / speaker	Total
Xitsonga	20	3-25 min	314 min
English	9	165-220 min	1695 min
French	10	110-195 min	1334 min
Mandarin	12	10-25 min	156 min

3. Experiments

3.1. Datasets

Comparisons of the experiments were carried out using the corpora provided by the Zero Resource Speech Challenge 2015 and 2017 (ZSC2015, 2017) [23, 24]. We used the Xitsonga dataset from ZSC2015 and the English, French, and Mandarin datasets from ZSC2017. The training sets consist of unsegmented audio files and each one of the files is linked to a unique speaker. Table 1 lists the details of the training data for each language. The total duration of Xitsonga for the test set was 149 min. Segmented into 120 second parts, the test sets for English, French, and Mandarin had a total duration of 1634 min, 1061 min, and 1522 min, respectively. In these experiments, all languages were used as the zero-resource language; training was done without transcriptions.

We used the Kaldi speech recognition toolkit [25] to extract feature vectors from the raw speech data. 13-dimensional mel-frequency cepstral coefficients (MFCCs) and their delta parameters ($\Delta + \Delta\Delta$) were extracted with a 25 ms analysis window and a 10 ms window shift, followed by a cepstral mean and variance normalization (CMVN) for each segment. Then, each frame of the target language was linearly transformed by applying fMLLR transformation and 40-dimensional features were obtained. The transformation matrix was trained using the acoustic model that was built using the Kaldi TIMIT [26] recipe. The extracted features of each language were splitted in a 90-10 ratio to obtain training and development sets in the model training.

3.2. Evaluation metric

To confirm the effectiveness of the proposed framework, extracted features were evaluated based on the ABX phone discriminability between phonemic minimal pairs [27, 28]. The

ABX error rate was calculated using the following equation:

$$\theta(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{m(m-1)n} \sum_{a \in S(\boldsymbol{x})} \sum_{b \in S(\boldsymbol{y})} \sum_{x \in S(\boldsymbol{x}) \setminus \{a\}} \left(\mathbb{1}_{d(a,x) < d(b,x)} + \frac{1}{2} \mathbb{1}_{d(a,x) = d(b,x)} \right), \quad (12)$$

where m and n represent the number of examples that belong to the subword category $S(\boldsymbol{x})$ and $S(\boldsymbol{y})$, and $\mathbb{1}$ is an indicator function. The distance between sounds $\boldsymbol{x}$ and $\boldsymbol{y}$, $d(\boldsymbol{x}, \boldsymbol{y})$, was calculated using dynamic time warping (DTW) on the underlying frame-to-frame distance. DTW was calculated using a cosine distance for fMLLR and bottlneck features and KL-divergence for posterior vectors. In the experiment, to investigate the effect of speaker normalization on phoneme discrimination performance, the score was measured for speech data with a single speaker (within-speaker condition) and with multiple speakers (across-speaker condition).

3.3. Experimental setup

For DPGMM sampling, we used the parallel sampler by Chang et al.[2], as described in [10, 11]. We used default parameters for the priors and set the concatenation parameter α to 1. Inference of DPGMM was repeated until convergence of the number of mixture components and stopped at 1500-th iteration. The number of DPGMM components for Xitsonga, English, French, and Mandarin were 510, 1880, 1623, and 510, respectively.

The adversarial network models were implemented with Chainer [29]. Regarding the conventional network, the number of units in each layer of the feature extractor, phoneme, and speaker predictors were set to {440 - 1024 - 1024 - 1024 - 1024}, {1024 - K}, and {512 - C}, respectively. K and C denote the number of DPGMM components and the number of speakers, respectively. Regarding the proposed network, the number of units in each layer of the feature extractor and speaker predictor were set to {440 - 1024 - 1024 - 1024 - 1024 - 1024 - K}, {512 - C}, respectively. The minibatch size was set to 1024 and ReLU activation was applied to the output of each hidden layer. The dropout layer was inserted after each hidden layer to prevent over-fitting. The dropout ratios were all set to 0.2. Network models were all trained with SGD with a learning rate of 0.01. The parameter λ for adversarial training was increased as follows:

$$\lambda = \lambda_{\max} \left\{ \frac{2}{1 + \exp(-\gamma p)} - 1 \right\}, \quad (13)$$

where p denotes the training progress, which increases linearly from zero to one. γ is the parameter controlling the convergence speed and is set to 10, as in [16]. $\lambda_{\max}$ represents the maximum value for the GRL parameter λ. Based on preliminary experiments, $\lambda_{\max}$ values of the conventional AMT training for Xitsonga, English, French, and Mandarin were 1.0, 9.0, 9.0, and 1.0, respectively. $\lambda_{\max}$ values of the proposed AMT training for each language were 50.0, 5.0, 7.0, and 9.0, respectively.

3.4. Features evaluated

For the purpose of evaluation, the following types of features were extracted from the test speech data of each target language.

- **fMLLR**: fMLLR features
- **DPGMM-post**: DPGMM posteriorgrams [7]
- **AMT-BNF**: bottleneck features derived from conventional AMT training [18]
- **AMT-post**: posteriorgrams obtained with the proposed AMT training

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

Table 2: *ABX error rate for within speakers*

Feature	Xitsonga	English	French	Mandarin
fMLLR	17.42	6.85	8.96	8.74
DPGMM-post	9.19	6.35	9.12	9.75
AMT-BNF	13.98	5.94	8.16	**8.22**
AMT-post	**8.41**	**5.88**	**8.09**	10.06

Table 3: *ABX error rate for across speakers*

Feature	Xitsonga	English	French	Mandarin
fMLLR	25.70	10.83	14.83	10.35
DPGMM-post	14.00	8.77	12.28	9.46
AMT-BNF	19.68	8.51	12.04	**8.95**
AMT-post	**12.59**	**8.18**	**11.37**	9.55

3.5. Results

Tables 2 and 3 list ABX error rates in the within-speaker and across-speaker conditions, respectively. Here, we can see that the proposed framework effectively suppresses the effect of speakers, as the results of AMTs point to better scores than those of fMLLR and DPGMM-post. The proposed adversarial network AMT-post outperformed the conventional network AMT-BNF for English, French, and especially Xitsonga with a high reduction rate. This indicates that AMT-post works well in terms of eliminating speaker information in DPGMM-post in both low-resource and resource-abundant languages. Regarding Mandarin, however, AMT-BNF generated a better score than AMT-post. There is a unique characteristic for Mandarin, in which some phonemes are distinguished by a change of tone [30]. Since DPGMM is generated with frame-wise features, contextual information is not considered, leading to difficulties in distinguishing Mandarin phoneme units. In fact, the score gets worse when applying DPGMM to fMLLR features.

4. Conclusions

This paper proposed a novel framework for extracting speaker-invariant features from zero-resource languages. DPGMM was chosen to yield phoneme information and AMT learning was applied to extract speaker-independent features. The novel AMT network was proposed as a modification of the conventional network suitable for DPGMM posteriorgrams. The experimental comparisons on ABX phone discriminability demonstrated that the proposed framework successfully suppressed the effects of speaker variability. Moreover, the proposed adversarial network suppressed speaker variability more effectively than an AMT-based conventional network. In the future, we plan to apply more sophisticated DPGMM modeling [21] to obtain fine DPGMM posteriorgrams for our network. Furthermore, we plan to extend our model to utilize a longer context in order to improve the performance of the feature extraction for Mandarin and extract the characteristics of a speaker from an acoustic feature.

5. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 17K12718.

6. References

[1] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on gaussian posteriorgrams," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2009.

[2] G. Mantena and K. Prahallad, "Use of articulatory bottle-neck features for query-by-example spoken term detection in low resource scenarios," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[3] C.-y. Lee and J. Glass, "A nonparametric bayesian approach to acoustic model discovery," in *Proceedings of the Association for Computational Linguistics (ACL)*, 2012.

[4] L. Ondel, P. Godard, L. Besacier, E. Larsen, M. Hasegawa-Johnson, O. Scharenborg, E. Dupoux, L. Burget, F. Yvon, and S. Khudanpur, "Bayesian models for unit discovery on a very low resource language," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

[5] I. Malioutov, A. Park, R. Barzilay, and J. Glass, "Making sense of sound: Unsupervised topic segmentation over acoustic input," in *Proceedings of the Association of Computational Linguistics (ACL)*, 2007.

[6] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, "NLP on spoken documents without ASR," in *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, 2010.

[7] 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 *Proceedings of the INTERSPEECH*, 2015.

[8] R. Thiolliere, E. Dunbar, G. Synnaeve, M. Versteegh, and E. Dupoux, "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling," in *Proceedings of the INTERSPEECH*, 2015.

[9] L. Badino, A. Mereta, and L. Rosasco, "Discovering discrete subword units with binarized autoencoders and hidden-markov-model encoders," in *Proceedings of the INTERSPEECH*, 2015.

[10] H. Kamper, M. Elsner, A. Jansen, and S. Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[11] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Multilingual bottle-neck feature learning from untranscribed speech," in *Proceedings of the IEEE International Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2017.

[12] E. Hermann and S. Goldwater, "Multilingual bottleneck features for subword modeling in zero-resource languages," in *Proceedings of the INTERSPEECH*, 2018.

[13] M. Pitz and H. Ney, "Vocal tract normalization as linear transformation of MFCC," in *Proceedings of the EUROSPEECH*, 2003.

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

[15] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

[16] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[17] Y. Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition." in *Proceedings of the INTERSPEECH*, 2016.

[18] T. Tsuchiya, N. Tawara, T. Ogawa, and T. Kobayashi, "Speaker invariant feature extraction for zero-resource languages with adversarial learning," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

[19] Z. Meng, J. Li, Z. Chen, Y. Zhao, V. Mazalov, Y. Gang, and B.-H. Juang, "Speaker-invariant training via adversarial learning," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

[20] M. Heck, S. Sakti, and S. Nakamura, "Supervised learning of acoustic models in a zero resource setting to improve DPGMM clustering." in *Proceedings of the INTERSPEECH*, 2016.

[21] M. Heck, S. Sakti, and S. Nakamura, "Feature optimized DPGMM clustering for unsupervised subword modeling: A contribution to zerospeech 2017," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2017.

[22] 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 *Proceedings of the INTERSPEECH*, 2016.

[23] M. Versteegh, R. Thiolliere, T. Schatz, X. N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015," in *Proceedings of the INTERSPEECH*, 2015.

[24] 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 *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2017.

[25] 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 *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[26] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.

[27] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline," in *Proceedings of the INTERSPEECH*, 2013.

[28] T. Schatz, V. Peddinti, X.-N. Cao, F. Bach, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair ABX task (ii): Resistance to noise," in *Proceedings of the INTERSPEECH*, 2014.

[29] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: a next-generation open source framework for deep learning," in *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*, 2015.

[30] C. Y. Suen, "Computational analysis of Mandarin sounds with reference to the English language," in *Proceedings of the Computational Linguistics (COLING)*, 1982.