DEEP LEARNING METHODS FOR UNSUPERVISED ACOUSTIC MODELING - LEAP SUBMISSION TO ZEROSPEECH CHALLENGE 2017

Ansari T K, Rajath Kumar, Sonali Singh and Sriram Ganapathy

Learning and Extraction of Acoustic Patterns (LEAP) Lab, Dept. of Electrical Engg.,
Indian Institute of Science, Bengaluru-560012, India.

ABSTRACT

In this paper, we present our system submission to the ZeroSpeech 2017 Challenge. The track1 of this challenge is intended to develop language independent speech representations that provide the least pairwise ABX distance computed for within speaker and across speaker pairs of spoken words. We investigate two approaches based on deep learning methods for unsupervised modeling. In the first approach, a deep neural network (DNN) is trained on the posteriors of mixture component indices obtained from training a Gaussian mixture model (GMM)-UBM . In the second approach, we develop a similar hidden Markov model (HMM) based DNN model to learn the unsupervised acoustic units provided by HMM state alignments. In addition, we also develop a deep autoencoder which learns language independent embeddings of speech to train the HMM-DNN model. Both the approaches do not use any labeled training data or require any supervision. We perform several experiments using the ZeroSpeech 2017 corpus with the minimal pair ABX error measure. In these experiments, we find that the two proposed approaches significantly improve over the baseline system using MFCC features (average relative improvements of 30-40%). Furthermore, the system combination of the two proposed approaches improves the performance over the best individual system.

Index Terms— Autoencoders, Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), Deep Neural Networks (DNNs), Unsupervised Learning.

1. INTRODUCTION

In the recent years, there has been a growing interest in the task of unsupervised representation learning from raw speech data without any supervision or labels [1, 2, 3]. In particular, zero resource speech technologies operate without the expert labels provided by linguistic knowledge that standard automatic speech recognition (ASR) systems use such as transcribed speech, language models, pronunciation dictionaries etc. A robust zero resource system must instead discover this linguistic knowledge from the speech signal automatically and in a language independent manner. While a zero resource speech challenge was earlier conducted [4] using primarily data from English and Xitsonga, the ZeroSpeech 2017 challenge expands the scope of deriving representations in a language independent fashion [5].

In a zero resource setting, there are no labeled audio resources and the task is to develop speech representations which allow the discovery of word units [4]. The main objective of the feature learning sub-task is to construct a representation of speech sounds which can support word identification robustly for both within and across talkers. The similarity measure used in the evaluation of the challenge is the ABX discriminability between phonemic minimal pairs (e.g. "beg" and "bag"). Some of the earliest approaches explored the HMM framework [1] and the use of posteriograms from the GMM framework [6]. Recently, several approaches have been proposed for this task using various feature representations [7] and neural network models [8, 9]. A hybrid dynamic time warping (DTW) approach has also been previously attempted for the zero resource challenge [10].

In this paper, we address the problem of extracting unsupervised speech representations in a zero resource scenario where the representations need to be learned in a language independent setting. We present our system submission to the zero speech 2017 challenge (Track 1) [5]. The proposed system consists of a combination of two subsystems which are based on a universal background model (UBM) approach. The first system uses a Gaussian mixture model (GMM)-UBM followed by a DNN model which generates posteriors of mixture component indices. The second system uses a hidden Markov model (HMM) in conjunction with a DNN that predicts the posterior probabilities of HMM states. While the GMM/HMM models are learned with mel frequency cepstral coefficients (MFCCs), the HMM-DNN models are trained with either MFCC features or the autoencoder embeddings. A novel method of system combination is also proposed for unsupervised feature learning.

We perform several experiments using the ZeroSpeech 2017 challenge corpus [5] where the speech data from English, French and Mandarin are used for training and devel-

This work was supported by Defense Research and Development Organization (DRDO), Government of India under the grant DRDO0689.

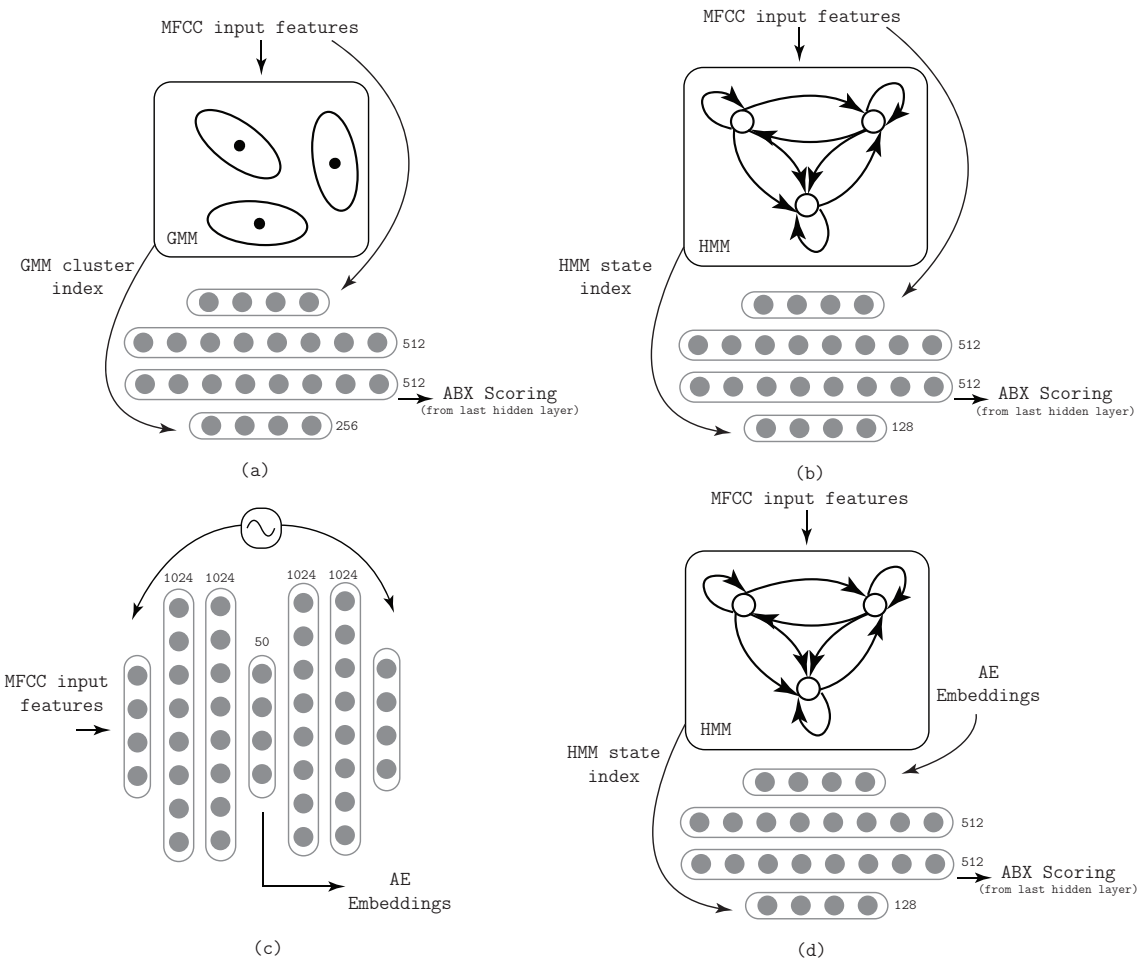

Fig. 1. Schematic illustrating the various components of the proposed system submission. (a) DNN trained using alignments obtained from GMM and input MFCC features (b) DNN trained using alignments obtained from HMM and input MFCC features (c) Extracting autoencoder embeddings from input features (d) DNN trained using alignments obtained from HMM with autoencoder embeddings as features.

opment. During the evaluation, two other surprise languages (denoted as L1 and L2) are also provided which are evaluated using proposed methods. In these experiments, the proposed systems based on autoencoder (AE) embeddings and the HMM-DNN hybrid modeling provide significant improvements over the baseline MFCC based feature representations. The experiments also illustrate that the across speaker conditions provide more relative improvements compared to the within speaker conditions. In addition to the individual system results, the system combination of the two sub-systems further improves the performance and approaches the topline results based on supervised representations from a phone recognition system. These experiments show that the models which are learned without any supervision using multilingual data can generate representations that provide high degree of linguistic similarity in many different languages.

The rest of the paper is organized as follows. Sec. 2 describes the proposed modeling approaches for unsupervised language independent sub-word learning. In Sec. 3, we provide the details of the experimental setup used for the evaluation. The results obtained using the various individual subsystems as well as the system combination are presented in Sec. 4 along with an analysis of the results. This is followed by a summary of the paper in Sec. 5.

2. SYSTEM DESCRIPTION

The individual components of the proposed system submission are given in Figure. 1. The system uses multiple unsupervised models like autoencoders, GMM/HMM based clustering followed by a DNN based backend. The details of these components are given below.

Table 1. Details of the ZeroSpeech 2017 corpus for training and testing [5].

Lang.	Training						Test	
	Relatives		Outsiders		Total		Total	
	#speakers	dur./speaker	#speakers	dur./speaker	dur.	#words	#files	dur.(min)
English	9	165-220min	60	10min	45h	370k	30658	1634
French	10	110-195min	18	10min	24h	220k	23765	1061
Mandarin	4	20-25min	8	10min	2h30min	20k	25383	1522
L1	10	85-150min	20	10min	25h	213k	15243	687
L2	4	37-42min	10	20min	4h	31k	7201	354

2.1. Autoencoder

A deep autoencoder is an artificial neural network which generates non-linear embeddings of the input data while also learning a reconstruction from the embeddings [11]. Autoencoders have been used for unsupervised representation learning [12, 13]. In our work, we use the deep feed-forward autoencoder neural network with rectified linear unit (ReLU) non linearity. The input features are fed frame wise without any context and we use mel frequency cepstral coefficients which are utterance level mean and variance normalized. The non-linear embeddings in our case are set to be higher in dimension compared the input features. However, the identity mapping is avoided with the use of the ReLU non-linearity with real valued input features (mean normalized cepstral features). The relatively high dimensional embeddings from autoencoders were also found to improve the minimal ABX pair classification task. The autoencoder was trained with all the languages and a mean square error (MSE) criterion was used. We use the Theano package [14] for the autoencoder implementation. The autoencoder features are either directly used for minimal pair ABX task using the cosine distance metric or used as features for a second DNN model which learns a mapping to the HMM state alignments (Sec. 2.3).

2.2. GMM-UBM

A Gaussian mixture model (GMM) is trained on the input features with $128/256$ mixture components using the maximum likelihood criterion. The GMM is trained using MFCC features (mean/variance normalized) with the training data from all the five languages. Thus, the GMM functions as a universal background model (UBM) (similar to GMM-UBM used in speaker verification [15]). The GMM posteriograms (consisting of posterior probability vectors containing the probability of cluster indices given the input speech frame) can be directly used for minimal pair ABX classification (similar to [6]).

While the GMM mixture components allow a clustering of the data into a set of language independent units, the independence assumption between the frames makes the clustering process noisy. In HMM-UBM modeling, we use the GMM based cluster alignments to initialize the HMM model.

2.3. HMM-UBM

This section briefly describes the process of growing the GMM-UBM into a HMM. The HMM is also trained using MFCC features. More details about how the GMM is grown into a full-fledged HMM are provided in [16]. The HMM consisting of 128 states is initialized using the 128 mixture component GMM. The HMM model used in this paper has an ergodic architecture (unlike the left-to-right architecture used in supervised HMM training) and the Baum-Welch algorithm is used to learn the parameters of the model with a maximum likelihood criterion [17]. The transition probability matrix to train the HMM is initialized such that the self transition probability for all the states is given a high value (0.7) and the other transitions are distributed equally. This initialization imposes contextual constraints by enabling the same HMM state to generate successive frames of MFCC vectors. The observation density of the states in the model are GMMs with 8 mixture components. The HMM posteriograms can also be used directly for minimal pair ABX classification.

The trained HMM model allows the generation of state alignments from the training data (Viterbi algorithm) which are used in hybrid DNN framework (similar to the hybrid HMM-DNN modeling in ASR [18]). The GMM/HMM model training as well as the generation of cluster/state alignments are done using the HTK toolkit [19].

2.4. Deep Neural Network

Using the UBM model either from GMM or HMM, the mixture component/state alignments of the acoustic features are generated at the frame level. We use these alignments with the corresponding MFCC input frames to train a feed forward neural network using the stochastic gradient descent algorithm. We also experiment with training the neural network using autoencoder embeddings along with the HMM-UBM state alignments (the autoencoder embeddings are generated at the same sampling as the input features). The DNN model is learned with a cross entropy cost function and a softmax target layer over 128 classes. After the model training, the hidden layer features before the softmax layer are used in DTW based minimal pair ABX classification with the cosine distance scoring. The DNN training requires careful frame

Table 2. Baseline and topline results for the corpus measured using minimal pair ABX error rate (%).

		English			French			Mandarin			Avg.	L1			L2			Avg.
		1s	10s	120s	1s	10s	120s	1s	10s	120s		1s	10s	120s	1s	10s	120s	
Baseline	within	12.0	12.1	12.1	12.5	12.6	12.6	11.5	11.5	11.5	12.0	10.3	9.3	9.4	14.1	14.3	14.1	11.9
	across	23.4	23.4	23.4	25.2	25.5	25.2	21.3	21.3	21.3	23.3	23.6	23.2	23.0	30.0	29.5	29.5	26.5
Topline	within	6.5	5.3	5.1	8.0	6.8	6.8	9.5	4.2	4.0	6.2	8.7	7.1	7.0	6.6	4.6	3.4	6.2
	across	8.6	6.9	6.7	10.6	9.1	8.9	12.0	5.7	5.1	8.2	12.8	10.5	10.4	7.1	3.6	4.3	8.1

Table 3. Results for various system components on the development languages measured using the minimal pair ABX error rate (%). All the systems are trained with MFCC features.

System		English			French			Mandarin			Avg.
		1s	10s	120s	1s	10s	120s	1s	10s	120s	
GMM-128	within	9.0	8.2	8.1	12.7	11.6	11.6	13.7	12.4	12.2	11.1
	across	13.3	12.4	12.4	17.9	16.7	16.6	14.9	13.9	13.9	14.7
GMM-128-DNN	within	8.3	7.3	7.1	11.2	9.9	9.8	11.7	10.3	10.2	9.5
	across	13.8	12.4	12.2	17.8	16.0	15.8	14.3	13.3	13.3	14.3
GMM-256	within	8.2	7.3	7.4	11.9	10.9	10.8	12.4	11.5	11.4	10.2
	across	12.8	11.8	11.7	17.2	15.7	15.6	14.0	13.1	13.1	13.9
GMM-256-DNN	within	8.2	7.2	7.0	10.9	9.7	9.6	11.3	10.2	10.1	9.4
	across	13.8	12.4	12.2	17.5	16.0	15.6	14.0	13.2	13.2	14.2
HMM-128	within	8.6	7.2	7.8	11.7	10.9	10.9	13.4	12.7	12.7	10.7
	across	13.8	12.4	11.7	16.3	14.9	14.9	15.0	14.1	14.0	13.9
HMM-128-DNN	within	7.9	7.0	7.0	10.6	9.2	9.3	10.5	9.4	9.3	8.9
	across	13.3	12.0	11.9	17.3	15.9	15.6	13.0	12.1	12.2	**13.7**
AE	within	8.3	7.5	7.6	10.4	9.4	9.3	9.1	8.2	8.2	**8.7**
	across	19.2	18.0	17.9	22.0	20.0	19.9	14.8	14.3	14.4	17.8

selection (exclude the speech frames that do not align well with any of the unsupervised HMM states) to avoid any convergence issues [16]. The DNN was implemented using the Theano package [14].

2.5. System Combination

In our experiments, we find that the autoencoder features provide good representations for intra speaker minimal pair ABX task while the HMM-DNN models provide robust representations for inter speaker ABX tasks (Sec. 4). Thus, we perform a system combination of the two systems. One approach to system combination uses the AE embeddings for the final DNN training (using the HMM state alignments). Another approach for system combination is obtained by modifying the cosine distance metric in the following fashion,

$$C'(A,X) = \alpha\, C_{AE}(A,X) + (1-\alpha)\, C_{HMM}(A,X)$$

where A, X represent a pair of words and C denotes the DTW distance using cosine similarity measure with either the autoencoder embeddings (AE) or the $HMM\text{-}DNN$ based representation. The combined distance C' is used in the final minimal pair ABX scoring. A similar combination is also used to compute $C'(B,X)$. The parameter α is obtained on the development set using the languages provided in training.

3. EXPERIMENTAL SETUP

3.1. Data

The data used in experiments is the ZeroSpeech 2017 corpus [5]. The details of the database are provided in Table 1 for training languages (English, French and Mandarin) as well as for the two surprise languages (L1 and L2). The training data for Mandarin and L2 are relatively small compared to other languages as well as the number of speakers. The training data is also setup in such a way as to simulate infant language learning data where large amount of speech from a few speakers (Relatives) with higher per speaker data is present compared to large number of other speakers with small amount of per speaker data (Outsiders).

The results for the ABX tasks are measured as an error rate which is the percentage of ABX pairs incorrectly classified (a ABX pair with "bag","beg" and "beg" denoting A, B and X respectively is said to be incorrectly classified if distance $C(A,X)$ is less than $C(B,X)$). Here C denotes the

Table 4. Definition of various system combinations used in ZeroSpeech evaluation.

Name	Definition
S1	Cos. Comb. of AE and GMM-DNN (MFCC)
S2	Cos. Comb. of AE and HMM-DNN (MFCC)
S3	AE feats. for HMM alignments in DNN

DTW based distance computed using the cosine similarity measure (for real valued features), the modified cosine similarity measure (Sec.2.5) or the Kullbeck-Leibler (KL) distance for posterior features. For each language, the error rate is measured under two conditions - within speaker and across speaker pairs (A,B are always from the same speaker while X can be from the same speaker or a different speaker). The words A,B and X come from test utterances that are of varying duration ($1s$, $10s$ and $120s$) and the results are reported separately for different lengths of the test recordings.

3.2. Setup

The features used in all the models are the mel-frequency cepstral coefficients (MFCCs) extracted using 25 ms windows which are shifted every 10 ms. The 13 dimensional coefficients are appended with deltas and acceleration coefficients to provide 39 dimensional features. We perform a speech activity detection using an adaptive energy based thresholding [20]. The speech regions are then normalized at the utterance level using cepstral mean and variance normalization (CMVN). A global CMVN is also applied across the recordings in the training data before the model learning step. All the models in our setup use the same features and the models are trained using the training portions of the five languages in the ZeroSpeech corpus (Table 1).

The autoencoder (AE) model is a 5 layer feed forward DNN with a MSE cost having a configuration of $39 \times 1024 \times 1024 \times 50 \times 1024 \times 1024 \times 39$. The GMM model consist of 128/256 mixture components and the HMM model contains 128 states with 8 mixture component GMMs in each state. The DNN model is trained with MFCC/AE embeddings and it has a configuration of 39/$50 \times 512 \times 512 \times 128$. The 128 targets for DNN come from GMM/HMM mixture-component/state alignments. The hidden activations from the last hidden layer of 512 units is used for the minimal pair ABX task.

4. RESULTS AND DISCUSSION

The baseline and topline results for the ZeroSpeech corpus using the minimal pair ABX error rate measure is shown in Table 2. The first row provides the result for the baseline system which uses MFCC features with cosine distance metric. Among the various languages, the L2 data has the highest error rate while the Mandarin data provides the lowest error. Also, the across speaker error pairs result in error rates that are about twice the error rates for the within speaker pairs. The second row provides the topline results which are obtained using a supervised phone recognition engine based on Kaldi [21] with labeled data from each language. The topline performance for the surprise languages (L1 and L2) are similar on the average to the performance on three known languages.

During the training and development phase, only the three known languages were provided. Thus, most of the system development experiments were performed using these three languages. The results for the system development with various individual system components are reported in Table 3. The system combination results are reported in Table 5.

4.1. Sub-system Results

The first and third row of Table 3 report the results for the GMM system using the mixture component posterior features with 128 and 256 mixture components respectively. The minimal pair ABX score in this case is computed using the DTW distance with a KL distance measure. The GMM-128 and GMM-256 provides significant improvements over the baseline system with average relative improvements of 8% and 16% for within speaker and 37% and 40% for across speaker conditions respectively. The second and fourth rows consist of the results for the DNN model trained with GMM cluster alignments. The GMM alignments are obtained by finding the mixture component index which has the maximum posterior value and these are used as labels for the DNN. The activations from the final hidden layer of the DNN model are used in the ABX scoring along with the cosine distance metric (Fig. 1(b)). The DNN model using GMM based alignments improves the performance over the GMM posterior features for both within and across speaker conditions (average relative improvements of 14% are achieved for within and 3% for across speaker conditions). Using the DNN backend, we also find that both 128 and 256 mixtures gives similar performance. Thus, we use 128 mixture component GMM for initializing the HMM.

The fifth row of Table 3 shows the results for the HMM with 128 states using the state posterior features along with the KL based minimal pair ABX scoring. This improves over GMM-128 in both within and across conditions by 4% and 5% respectively. Further, the hybrid DNN model (Fig. $1(c)$) improves the HMM-128 system and provides the best across speaker performance among all the individual system components (average relative improvements of 41% over the baseline system).

The results for the autoencoders (AE) embeddings using a cosine distance based minimal pair ABX scoring are shown in the last row of Table 3. The AE features provide the best within speaker performance among all the individual systems considered here (average relative improvements of 28% over the baseline system). However, AE embeddings do not gen-

Table 5. Results for various system combinations submitted to the ZeroSpeech 2017 challenge on evaluation set measured using minimal pair ABX error rate (%).

		English			French			Mandarin			Avg.	L1			L2			Avg.
		1s	10s	120s	1s	10s	120s	1s	10s	120s		1s	10s	120s	1s	10s	120s	
S1	within	7.4	6.6	6.6	9.8	8.7	8.5	9.3	8.5	8.3	8.2	6.9	6.1	6	9.9	9.2	9.1	**7.9**
	across	14.5	13.3	13.2	17.8	16.4	16.2	13.2	12.8	12.7	14.5	16.9	14.7	14.7	18.8	17.7	17.7	16.8
S2	within	7.4	6.6	6.6	9.8	8.5	8.4	9.2	8.3	8.2	**8.1**	6.8	6	6	10.1	9.6	9.6	8.0
	across	13.7	12.5	12.4	17.2	15.8	15.6	12.6	12.0	12.0	13.8	16	14	13.9	17.9	16.9	16.6	15.9
S3	within	7.7	6.8	6.7	10.4	8.9	8.8	10.4	9.3	9.1	8.7	7.3	6.2	6.1	11.1	10.3	10.2	8.5
	across	13.2	12.0	11.9	17.2	15.6	15.4	13.0	12.2	12.3	**13.6**	15.5	13.5	13.4	17.6	16	16	**15.3**

eralize well for the across speaker conditions compared to the HMM-DNN representations.

4.2. System Combination Results

The definition of various systems used in the system combination experiments is given in Table 4. For S1, we combine the representations of individual sub-systems AE with $GMM\text{-}DNN$ using the combination method mentioned in Sec. 2.5. A similar cosine distance combination of AE embeddings with $HMM\text{-}DNN$ is denoted as S2. The system S3 refers to the use of AE embeddings in training the hybrid DNN using the HMM state alignments (Fig. 1 (d)). The best value of the parameter α (obtained on the development set) is 0.8 for the modified cosine distance metric of S1 and S2 .

Table 5 provides the summary of the evaluation results for all the conditions in the five languages. As seen here, the cosine distance based combination (S1 and S2) improves the within speaker conditions of ABX pair error rate compared to the best individual system result in Table 3 for the known languages. Among the three system combination methods for the surprise languages (L1, L2), the best within speaker condition result is obtained for S1. However, for the across speaker conditions, the S3 system which is the feature level combination (using AE features for DNN training with HMM state targets), provides the best results for both known languages and the surprise languages. In terms of average relative improvements over the baseline features for the surprise languages, the proposed system S3 improves 29% for within speaker pairs and 42% for the across speaker pairs. It is also interesting to note that, for language L1 in all the durations (1s, 10s and 120s) for within speaker pairs, the results for all the proposed systems (S1,S2,S3) are better than the topline results (supervised phone recognition system). These results highlight that the proposed approaches for unsupervised representation learning in language independent settings can approach the linguistic similarity results achieved by the supervised phoneme posterior features for the matched speaker pairs.

5. SUMMARY AND FUTURE WORK

In summary, we have shown that using the GMM and HMM cluster assignments followed by the training of a hybrid DNN model provides considerable benefits for the task of unsupervised subword modeling. Also, the features derived from the autoencoder are beneficial for language generalization in within speaker conditions. Thus the combination of the two approaches further improves the results for within speaker pairs. This may help in reducing the performance gap between within speaker and across speaker pairs. In addition, using the AE embeddings with the HMM state alignments in a DNN framework generates representations that are relatively robust to speaker variability. With experiments on surprise languages, we have also showcased that the proposed systems generalizes across multiple languages which is one of the intended goals for the ZeroSpeech 2017 challenge.

While the experiments reported in this paper are encouraging, we believe that further progress can be achieved by employing Vocal Track Length Normalization (VTLN) and Feature space Maximum Likelihood Linear Regression (fMLLR) techniques using the state alignments. In addition, DNN training methods using soft alignment instead of hard alignments might further prove to be beneficial. In the future, we also plan to investigate other neural network architectures such as convolutional and recurrent neural networks that incorporate more contextual information.

6. REFERENCES

[1] Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux, "Unsupervised learning of acoustic sub-word units," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics, 2008, pp. 165–168.

[2] Aren Jansen, Kenneth Church, and Hynek Hermansky, "Towards spoken term discovery at scale with zero resources," in *Eleventh Annual Conference of the International Speech Communtion Association*, 2010.

[3] David F Harwath, Timothy J Hazen, and James R Glass, "Zero resource spoken audio corpus analysis," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (SSP)*. IEEE, 2013, pp. 8555–8559.

[4] Maarten Versteegh, Roland Thiolliere, Thomas Schatz, Xuan-Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux, "The zero resource speech challenge 2015.," in *INTERSPEECH*, 2015, pp. 3169–3173.

[5] Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadyi, Mathieu Bernard, Laurent Besacier, Xavier Anguerra, and Emmanuel Dupoux, "The zero resource speech challenge 2017," in *IEEE 2017 workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2017.

[6] Yaodong Zhang and James R Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2009, pp. 398–403.

[7] Thomas Schatz, Vijayaditya Peddinti, Xuan-Nga Cao, Francis Bach, Hynek Hermansky, and Emmanuel Dupoux, "Evaluating speech features with the minimal-pair ABX task (ii): Resistance to noise," in *Fifteenth Annual Conference of the International Speech Communtion Association*, 2014.

[8] Daniel Renshaw, Herman Kamper, Aren Jansen, and Sharon Goldwater, "A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge," in *Sixteenth Annual Conference of the International Speech Communtion Association*, 2015.

[9] Hongjie Chen, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li, "Unsupervised bottleneck features for low-resource query-by-example spoken term detection.," in *INTERSPEECH*, 2016, pp. 923–927.

[10] Roland Thiolliere, Ewan Dunbar, Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux, "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling.," in *INTERSPEECH*, 2015, pp. 3179–3183.

[11] Yoshua Bengio et al., "Learning deep architectures for AI," *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[12] Pierre Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 37–49.

[13] Quoc V Le, "Building high-level features using large scale unsupervised learning," in *Acoustics, Speech and Signal Processing (SSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8595–8598.

[14] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio, "Theano: A CPU and GPU math compiler in python," in *Proc. 9th Python in Science Conf*, 2010, pp. 1–7.

[15] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, "Speaker veriftion using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[16] Ansari T.K, Rajath Kumar, Sonali Singh, Sriram Ganapathy, and Susheela Devi, "Unsupervised HMM posteriograms for language independent acoustic unit modeling in zero resource conditions," in *IEEE 2017 workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2017.

[17] Lawrence R Rabiner and Biing-Hwang Juang, "Fundamentals of speech recognition," 1993.

[18] Herve A Bourlard and Nelson Morgan, *Connectionist speech recognition: a hybrid approach*, vol. 247, Springer Science & Business Media, 2012.

[19] Steve J Young and Sj Young, *The HTK hidden Markov model toolkit: Design and philosophy*, University of Cambridge, Department of Engineering, 1993.

[20] Zheng-Hua Tan and Børge Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 798–807, 2010.

[21] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.