# UNSUPERVISED HMM POSTERIOGRAMS FOR LANGUAGE INDEPENDENT ACOUSTIC MODELING IN ZERO RESOURCE CONDITIONS

*Ansari T K[1], Rajath Kumar[1], Sonali Singh[1], Sriram Ganapathy[1] and Susheela Devi[2].*

[1] Learning and Extraction of Acoustic Patterns (LEAP) Lab, Department of Electrical Engineering,
[2] Department of Computer Science and Automation,
Indian Institute of Science, Bengaluru, 560012, India.

## ABSTRACT

The task of language independent acoustic unit modeling in unlabeled raw speech (zero-resource setting) has gained significant interest over the recent years. The main challenge here is the extraction of acoustic representations that elicit good similarity between the same words or linguistic tokens spoken by different speakers and to derive these representations in a language independent manner. In this paper, we explore the use of Hidden Markov Model (HMM) based posteriograms for unsupervised acoustic unit modeling. The states of the HMM (which represent the language independent acoustic units) are initialized using a Gaussian mixture model (GMM) - Universal Background Model (UBM). The trained HMM is subsequently used to generate a temporally contiguous state alignment which are then modeled in a hybrid deep neural network (DNN) model. For the purpose of testing, we use the frame level HMM state posteriors obtained from the DNN as features for the ZeroSpeech challenge task. The minimal pair ABX error rate is measured for both the within and across speaker pairs. With several experiments on multiple languages in the ZeroSpeech corpus, we show that the proposed HMM based posterior features provides significant improvements over the baseline system using MFCC features (average relative improvements of $25\%$ for within speaker pairs and $40\%$ for across speaker pairs). Furthermore, the experiments where the target language is not seen training illustrate the proposed modeling approach is capable of learning global language independent representations.

***Index Terms—*** Unsupervised learning, Hidden Markov Model (HMM) posteriograms, Multilingual Modeling, Zero resource speech.

## 1. INTRODUCTION

Language acquisition and learning is one of the most complex tasks perfected by humans from a very young age. While a major portion of this learning is supervised and top-down, several studies suggest that this learning is also facilitated by unsupervised learning from adult directed speech (for example [1, 2]). On the other hand, the automatic speech recognition technology primarily implements learning mechanisms that rely on large, lexically transcribed corpora and pronunciation dictionaries that relate the underlying signal to a phoneme inventory for subword modeling [3]. In the recent past, there has been renewed interest in the development of low resource multi-lingual speech recognition system [4, 5] where supervised audio data from several languages are used to derive shared representations.

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

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 [6]. The main challenge is to construct a representation of speech sounds which can support word identification robustly for both within and across talkers. The measure used is the ABX discriminability between phonemic minimal pairs (e.g. “beg” and “bag”). Recently, several approaches have been proposed for this task using various feature representations [7] and neural network models [8, 9, 10]. In addition, some approaches were evaluated on data distributed under the Babel program to address the problem of automatic speech recognition (ASR) and keyword spotting (KWS) under a zero acoustic resource scenario [11]. The difference mainly would be - Most of the Babel approaches would use supervised data in some language to generate a bottleneck representation which is used in low resource KWS or ASR. However in our case, we do not assume the presence of any supervised data. The zero resource scenario can further be complicated when the representations need to be learned in an language independent fashion. The ZeroSpeech 2017 corpus provides a corpus for the development of language independent unsupervised sub-word units [12]. While the corpus contains training data for all languages, this paper is also focused on the challenging scenario where there is no training data available from the target language of interest (a true zero resource setting).

The use of Hidden Markov Modeling (HMM) for unsupervised learning has been explored in [13] where the problem is formulated as an optimization over both parameter and transcription space. A similar approach has also been previously attempted for speaker dependent speech clustering [14]. The application of Gaussian posteriograms for keyword spotting using a segmental dynamic time warping (DTW) distance metric has been investigated in [15]. With several examples of a given keyword, the authors use DTW distances to compare the Gaussian posteriorgrams between keyword samples and test utterances. In a recent work, the HMMs have been used along with binarized autoencoder features for zero resource keyword spotting task [16]. However, many of these methods do not address the problem of modeling language independent sub-word units.

In this paper, we propose a novel method for unsupervised subword unit modeling with a HMM paradigm using unlabeled multilingual data. The speech recordings from multiple languages are initially used to learn a Gaussian mixture model (GMM) - universal background model (UBM) (similar to speaker/language recognition framework [17]). The GMM provides an initial state clustering for the HMM with the states grown out of the GMM mixture components. The HMM is then trained to learn the distribution of unsupervised language independent units. The HMM-UBM model generates frame level alignments which can further be used in a deep neural network (DNN) framework (similar to hybrid speech recognition

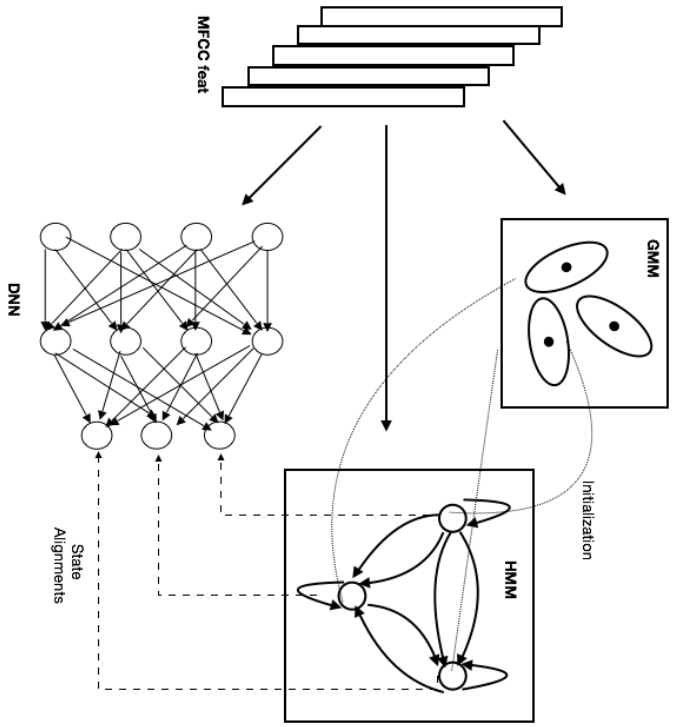


**Fig. 1**. Block Schematic of the proposed unsupervised HMM-DNN modeling where the GMM clustering is used to initialize the HMM.

approach [18] except that the proposed approach uses unsupervised language independent subword units instead of phonetic units). The DNN posterior features are then used for a spoken term detection using the minimal pair ABX classification score for within speaker and across speaker pairs.

We perform several experiments using the ZeroSpeech 2017 corpus where the speech data from English, French and Mandarin are used for training and development while two surprise languages (L1,L2) are provided in evaluation. We experiment with a full training (FT) setting where all the languages are used in training and a leave-one-out-training (LOOT) where the target language is not used in training the background models. In these experiments, the proposed HMM-UBM based posterior features provide significant improvements over the baseline MFCC based feature representations. We also illustrate that the proposed model performs equally well in the FT and LOOT conditions indicating that the model is capable of learning global representations that are language independent.

The rest of the paper is organized as follows. Sec. 2 describes the proposed model for unsupervised language independent sub-word learning. In Sec. 3, we provide the experimental setup as well as the results obtained using the baseline approach. A detailed analysis of the proposed approach is given in Sec. 4 which is followed by a summary of the paper in Sec. 5.

## 2. UNSUPERVISED HMM-DNN HYBRID MODEL LEARNING

The proposed approach of unsupervised HMM-DNN modeling is illustrated by the block schematic shown in Fig. 1. The model learning consists of three parts, GMM-UBM, HMM-UBM and HMM-DNN based hybrid modeling.

### 2.1. GMM-UBM

The input speech features are mel frequency cepstral coefficients (MFCCs) and are extracted from the training data consisting of multiple languages and speakers. These features are used to train a Gaussian mixture model (GMM). The GMM density function is given by,

$$p(\mathbf{x}) = \sum_{n=1}^{N} \alpha_n \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$$

where $\mathbf{x}$ corresponds to the MFCC features of dimension $D$, $N$ is the number of mixture components and $\mathcal{N}$ represents the Gaussian density function. The variables $\{\alpha_n, \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\}$ represent the model parameters of the GMM. We use the standard expectation maximization (EM) algorithm to train the GMM with a maximum likelihood (ML) criterion.

One of the main drawbacks of the GMM is the independence assumption of the speech frames. Due to this assumption, we find that the cluster assignments for speech frames are not consistent in time (the assignments are not smooth over successive speech frames). In the subsequent HMM learning, the GMM mixture components are used as the states and temporal consistency is introduced using the HMM transition matrix that is biased to encourage the within state transitions.

### 2.2. HMM-UBM

The HMM configuration consist of $N$ states which are ergodic in nature (any state can follow any other state which is different from the strict left-to-right architecture used in speech recognition). The HMM state observation densities are GMMs with $C$ mixture components per state. The model parameters of the HMM are denoted as $\lambda = \{\pi, \boldsymbol{A}, \boldsymbol{B}\}$, where $\pi$ are the initial state probabilities, $\boldsymbol{A}$ is

**Table 1**. Details of the ZeroSpeech Training and Testing Setup [12].

| Language | Training | | | | | | Test | |
|---|---|---|---|---|---|---|---|---|
| | Relatives | | Outsiders | | Total | | Total | |
| | #speakers | duration/speaker | #speakers | duration/speaker | duration | #words | #files | duration(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 |

**Table 2**. Baseline results using MFCC features and Topline results using supervised phoneme posteriors for the ZeroSpeech 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 |

the state transition probability matrix of size $N \times N$ and $\boldsymbol{B}$ contains the parameters of the state observation densities denoted by $\{\alpha_n^c, \boldsymbol{\mu}_n^c, \boldsymbol{\Sigma}_n^c\}$ for $n = \{1, ..., N\}$ and $c = \{1, .., C\}$.

In growing the GMM into a full fledged HMM with unsupervised state units, we found that the initialization plays a crucial role. The GMM mean parameters $\boldsymbol{\mu}_n^c$ for state $n$ are initialized with a perturbation of the Gaussian mean parameters representing the GMM-UBM $\boldsymbol{\mu}_n$. The covariance matrix is initialized as $\boldsymbol{\Sigma}_n^c = \boldsymbol{\Sigma}_n$ for $c = 1, .., C$. The weights of the GMM observation density are initially chosen equally $\alpha_n^c = \frac{1}{C}$. The increase in Gaussian mixture components (for example, from 1 to 2, 2 to 4 etc) in each state of the HMM is done sequentially by mixture splitting technique. This involves copying the mixture component, dividing the weights of both copies by 2, followed by perturbation of means by 0.2 standard deviations [19]. The transition probability matrix $\boldsymbol{A}$ is initialized in such a way self transition probability for all the states is given a high value (0.7) and the other transitions are distributed equally. This way of initializing the HMM allows to grow the initial GMM-UBM into a HMM and the assignment of higher transition probability for self transitions encourages temporal smoothness by enabling the same HMM state to generate successive frames of MFCC vectors.

Using the above mentioned initialization, the HMM is retrained using the Baum-Welch algorithm [20] for several iterations to learn all the model parameters $\lambda$. After the HMM training, the best state alignment for the speech frames is estimated and this is used in hybrid modeling (similar to hybrid modeling in standard ASR [18]).

### 2.3. DNN Model Learning

The DNN model is learned using the MFCC features with the state alignments obtained from the HMM. For instance, in the model "HMM-128-DNN 2 mix-GMM" reported in Table 3 , the DNN is trained on state alignments from HMM-128 2 mix-GMM. In all the DNN models, the number of target nodes is 128 corresponding to the number of HMM states. The target units in the DNN model are unsupervised HMM states (unlike the force aligned HMM states obtained using transcripts in an ASR system). The DNN modeling of HMM-states improves the estimation of state posterior probabilities (similar to the supervised Hybrid modeling in ASR) [18]. This also improves the ABX performance as seen in Table 3. Given the unsupervised nature of the targets, we found that the learning of the DNN parameters suffers from convergence issues when all the speech frames are used in the training. In order to address this issue, we propose a frame selection method for the DNN model learning.

The main objective of the frame selection is to reject the speech frames that do not align well with any of the unsupervised HMM states. Thus, an estimate of the confidence of the frame alignment with the HMM states is needed. For this purpose, we compute the posterior probabilities of HMM states denoted as $\gamma_n(\mathbf{x}) = p(n|\mathbf{x})$ using the trained HMM-UBM $\lambda$. This can be efficiently estimated using the forward-backward algorithm [20]. Then, a $N$ dimensional vector of state posterior probabilities is constructed for every frame $[\gamma_1(\mathbf{x}), .., \gamma_N(\mathbf{x})]$ and the entropy of the posterior vector as well as the maximum value of this vector are computed. A low entropy value as well as a high value for the maximum state posterior probability can indicate a high confidence in the HMM state alignment. We use pre-set thresholds on the entropy as well as the maximum posterior value and the speech frames which have a lower entropy and higher maximum value (compared to the respective threshold values) are selected for the DNN training. This way of frame selection allows the DNN model to learn efficiently and we found that the frame selection helps in avoiding convergence issues in the DNN training using the unsupervised HMM state targets. In our setup, about 70 % of the training data frames were selected for the DNN training using the above mentioned criterion.

We train a 3 layer feed forward DNN with a rectified linear unit (ReLU) activation. A cross entropy cost function is used and softmax output layer non-linearity is applied. The model is learned using stochastic gradient descent. We use the Theano package [21] for the DNN model training. For testing, the trained DNN model is used to generate posteriors of HMM states given the input MFCC feature vector. The posterior features are input to the DTW based minimal pair ABX scoring using a Kullbeck-Leibler (KL) divergence distance metric [12].

**Table 3**. Performance in terms of minimal pair ABX error rate (%) for different model configurations in the FT setup.

| 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-256 | within | 8.2 | 7.3 | 7.1 | 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 |
| HMM-128 | within | 9.1 | 8.0 | 8.0 | 12.5 | 11.4 | 11.4 | 14.2 | 12.8 | 12.7 | 11.1 |
| 1 mix-GMM | across | 13.2 | 12.4 | 12.3 | 17.1 | 15.8 | 15.7 | 15.6 | 14.3 | 14.3 | 14.5 |
| HMM-128 | within | 8.9 | 8.1 | 8.0 | 12.4 | 11.4 | 11.2 | 14.0 | 12.6 | 12.4 | 11.0 |
| 2 mix-GMM | across | 12.9 | 12.1 | 12.0 | 16.7 | 15.6 | 15.5 | 15.3 | 14.3 | 14.2 | 14.3 |
| HMM-128 | within | 8.7 | 8.0 | 7.9 | 11.7 | 10.9 | 11.1 | 13.9 | 12.6 | 12.5 | 10.8 |
| 4 mix-GMM | across | 12.7 | 11.8 | 11.8 | 16.5 | 15.3 | 15.4 | 15.2 | 14.1 | 14.1 | 14.1 |
| HMM-128 | within | 8.6 | 7.8 | 7.8 | 11.7 | 10.9 | 10.9 | 13.4 | 12.7 | 12.7 | 10.7 |
| 8 mix-GMM | across | 12.5 | 11.7 | 11.7 | 16.3 | 14.9 | 14.9 | 15.0 | 14.1 | 14.0 | 13.9 |
| HMM-128-DNN | within | 7.9 | 7.0 | 6.9 | 10.6 | 9.2 | 9.2 | 10.9 | 9.6 | 9.5 | 9.0 |
| 2 mix-GMM | across | 13.2 | 11.9 | 11.8 | 17.3 | 15.7 | 15.4 | 13.1 | 12.3 | 12.4 | 13.7 |
| HMM-128-DNN | within | 7.9 | 7.0 | 7.0 | 10.6 | 9.2 | 9.2 | 10.7 | 9.5 | 9.3 | **8.9** |
| 4 mix-GMM | across | 13.3 | 12.1 | 11.9 | 17.3 | 15.9 | 15.6 | 13.1 | 12.3 | 12.3 | **13.7** |
| HMM-128-DNN | within | 7.9 | 7.0 | 7.0 | 10.6 | 9.2 | 9.3 | 10.5 | 9.4 | 9.3 | **8.9** |
| 8 mix-GMM | across | 13.3 | 12.0 | 11.9 | 17.3 | 15.9 | 15.6 | 13.0 | 12.1 | 12.2 | **13.7** |

## 3. BASELINE EXPERIMENTAL SETUP

### 3.1. Data

The development data of ZeroSpeech corpus [12] comprises of three languages (English, French and Mandarin) as shown in Table 1. In order to simulate natural language learning conditions in infants, the training data is split into Relatives and Outsiders, where Relatives are small number of speakers with more speech data and Outsiders consist of more number of speakers with less duration per speaker. The development test set includes speech recordings of different durations (1s, 10s and 2min) from a large number of speakers that are different from the training speakers. The effect of duration may factor in some adaptation and normalization techniques that use the whole recording duration (for example, the topline results). The evaluation data includes two surprise languages (L1 and L2).

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 [22]. 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 model learning.

### 3.2. Evaluation

The performance of the proposed features are evaluated using minimal pair ABX discriminability scoring [6]. Here, A, B and X represent similar sounding phone triplets where linguistically either A=X or B=X and A! =B. The triplet B is same as A except for the center phoneme in the triplet (for example, A="b-a-g", B="b-e-g", X="b-a-g"). The pairwise distance between (A,X) and (B,X) are measured and any pair for which the distance between linguistically same triplet is more than the distance between linguistically different triplet is counted as error. The error rate (measured as percentage of erroneous pairs in the test data) is used as an evaluation metric. The distance measure is based on the DTW method with frame-wise distance computation using cosine distance (for real valued features like MFCC vectors) or Kullbeck-Leibler (KL) distance (for posterior features). The evaluation is done on 2 conditions, when all of ABX belongs to the same speaker, which is termed as within speaker and when X is from a different speaker which is termed as across speaker condition. The triplets AB always come from the same speaker.

The baseline results makes use of 13 dimensional MFCC features along with delta and acceleration and the topline system is a supervised phone recognition engine implemented in Kaldi toolkit [23]. These results are tabulated in Table 2.

## 4. PROPOSED SYSTEM RESULTS AND ANALYSIS

The performance of the proposed models are measured in two conditions, full-training (where all the five languages consisting of three development languages and two surprise languages are used in training) denoted as FT and leave-one-out-training where the target language is not used in model training denoted as LOOT. For LOOT conditions, only two development languages are used in the training of the models and the target language is the unseen development language.

### 4.1. Full Training (FT)

The results for the different modeling methods described in Sec. 2 is given in Table 3. The GMM-128 system consists of $N = 128$ component GMM and the posteriograms from the GMM are used in the ABX scoring using the KL distance measure. As seen here, the GMM model provides significant improvements over the baseline features particularly for the across speaker conditions (average relative improvements of 37 % in the across speaker conditions).

The HMM-UBM which is initialized using the GMM also provides state level posterior features which can be used in the ABX scoring. This system provides moderate gains over the GMM system. We experiment with different number of mixture components

**Table 4**. Performance in terms of minimal pair ABX error rate (%) in LOOT condition for different development languages.

| LOOT Language | | English | | | French | | | Mandarin | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1s | 10s | 120s | 1s | 10s | 120s | 1s | 10s | 120s |
| English | within | 7.8 | 6.9 | 6.9 | 10.2 | 8.6 | 8.6 | 10.1 | 9.0 | 9.0 |
| | across | 13.1 | 11.8 | 11.7 | 16.6 | 15.2 | 14.9 | 12.4 | 11.7 | 11.7 |
| French | within | 7.6 | 6.8 | 6.8 | 10.4 | 9.0 | 9.1 | 10.4 | 9.4 | 9.3 |
| | across | 13.2 | 12.0 | 11.8 | 17.1 | 15.7 | 15.4 | 12.8 | 12.2 | 12.1 |
| Mandarin | within | 7.8 | 6.9 | 6.9 | 10.5 | 9.1 | 9.0 | 10.4 | 9.4 | 9.2 |
| | across | 13.3 | 12.0 | 11.8 | 17.3 | 15.7 | 15.4 | 13.0 | 12.3 | 12.3 |

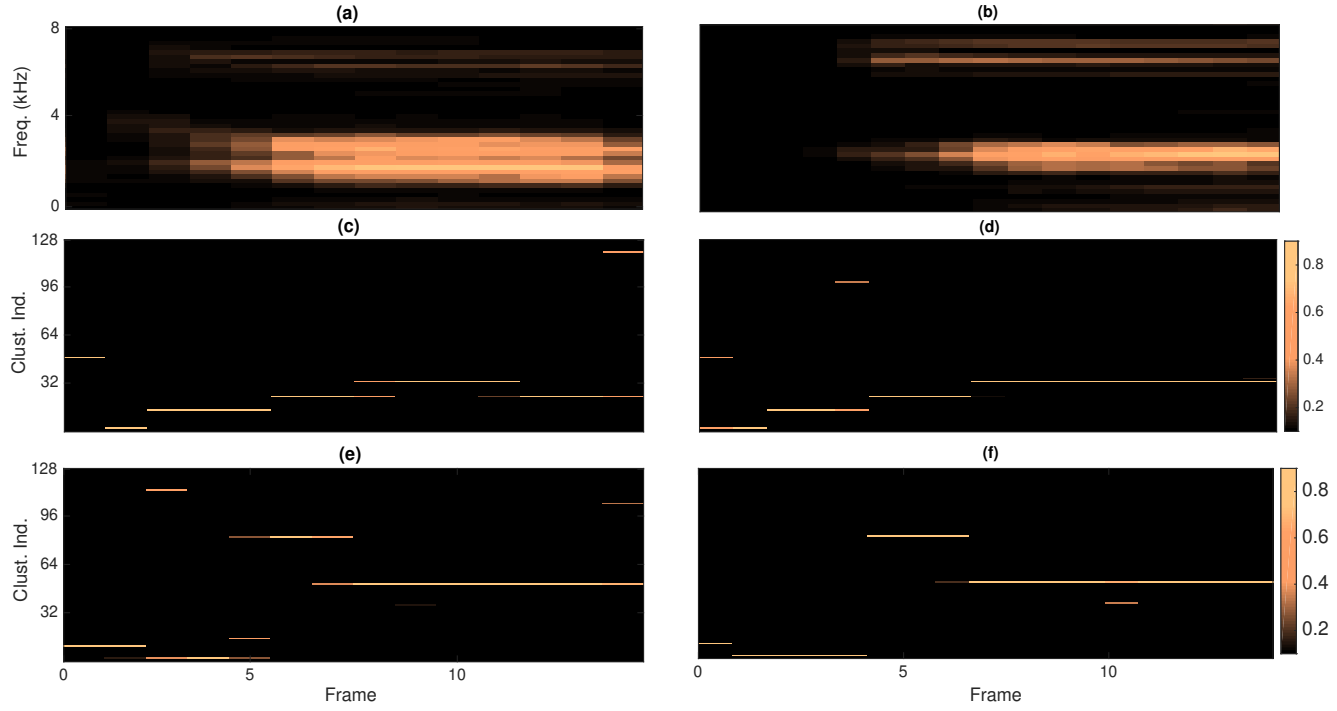


**Fig. 2**. "Plots illustrating the representations for word "*b-iy-ax*" from two speakers (S1,S2) in the development data. (a) Mel spectrogram for S1, (b) Mel spectogram for S2, (c) HMM-DNN posteriogram using FT for S1, (d) HMM-DNN posteriogram using FT for S2, (e) HMM-DNN posteriogram using LOOT for S1, (f) HMM-DNN posteriogram using LOOT for S2.

**Table 5**. Result of the HMM-DNN system in FT mode submitted to the ZeroSpeech 2017 challenge on the surprise languages.

| | L1 | | | L2 | | | Avg. |
|---|---|---|---|---|---|---|---|
| | 1s | 10s | 120s | 1s | 10s | 120s | |
| within | 7.6 | 6.4 | 6.2 | 11.6 | 10.9 | 10.7 | 8.9 |
| across | 15.7 | 13.7 | 13.5 | 17.5 | 16.1 | 16.1 | 15.4 |

($C = 2, 4, 8$) for the HMM state observation density. The addition of DNN based hybrid modeling further improves the performance. In particular, significant improvements are observed in the within speaker condition using the HMM-DNN framework. The best overall result is obtained for the HMM-DNN configuration with $N = 128$ and $C = 4$ where we observe average relative improvements (over the baseline) of about 25% for within speaker conditions and 41% for the across speaker conditions.

### 4.2. Leave One Out Training (LOOT)

In this scenario, the models are trained using two out of the three development languages (English, French and Mandarin) while the third language is used as the target language for testing. This represents the most challenging "zero" resource setting when there is no training data available in the language of interest. This framework also allows us to investigate the language independent nature of the speech representations generated by the proposed modeling framework.

The results for the LOOT condition are reported in Table 4. Since the LOOT setup requires the training of a new system for each testing language, we only experiment with the HMM-DNN system. The shaded results along the diagonal in Table 4 report the results for the language not seen in training. For example, the first row of the table reports the results when the HMM-DNN model is trained using French and Mandarin data. In this case, English represents the language not seen in training.

As seen in Table 4, the results along the diagonal are quite comparable to the FT results (using 5 training languages) reported in the last row of Table 3. The other results presented in the non-diagonal entries of Table 4 are indicative of the change in performance (if any) due to a particular choice of the languages used in training. This setup of LOOT shows that the proposed approach modeling using HMM-DNN posteriograms provides robust representations of language independent units for the underlying speech signal. While the baseline MFCC features are also language independent, the proposed approach yields significant improvements over the MFCC features in the spoken term discovery of an unseen language (measured using the minimal pair ABX error rate). It is also important to note the proposed representations approach the supervised language representations used in the topline results.

The significance of the results in Table 4 is further highlighted by analyzing the posteriogram representations of the same word from two different speakers. This is illustrated in Fig. 2 where we plot the posteriogram representations (using 128 HMM-DNN target classes) for the word "b-iy-ax" spoken by two speakers. Here, two sets of posteriograms are compared - generated using FT setup as well as the LOOT setup. A comparison with the baseline mel spectrogram is also provided for reference. As seen here, the HMM-DNN provides consistent representations for linguistically similar word pairs from two different speakers irrespective of whether the target language is present in the training or not.

In addition, the HMM-DNN representations from the FT mode are also used in the two surprise languages provides in the ZeroSpeech 2017 corpus. These results are shown in Table 5. These results show similar improvements observed in Table 3 for the development languages. The performance of the HMM-DNN system proposed in this paper is advanced using system combinations with other deep learning methods and these results are reported in [24].

## 5. SUMMARY

We have proposed a novel approach to unsupervised language independent acoustic modeling. The proposed model utilizes a GMM based clustering which is followed by a HMM. The mapping of the speech signal to HMM states, which represent unsupervised subword units, can be learned from the input features directly using a DNN model. We show that DNN based modeling of acoustic units from raw speech data is quite capable of discovering linguistically similar words from within and across speaker conditions when the target language is seen in training.

Using a set of leave-one-out-training experiments, we also show that the proposed approach is robust to the lack of any training data from the target language. In particular, the LOOT condition provides comparable results to the FT condition which highlights the usefulness of the HMM posteriograms in language independent acoustic subword modeling.

The performance gap between the within speaker and across speaker conditions indicate that there is further room for improvement using speaker normalization and adaptation methods. In future, we plan to investigate these research directions for improving the robustness of the model to speaker variabilities.

## 6. REFERENCES


[1] Bart De Boer and Patricia K Kuhl, "Investigating the role of infant-directed speech with a computer model," *Acoustics Research Letters Online*, vol. 4, no. 4, pp. 129–134, 2003.

[2] Weiyi Ma, Roberta Michnick Golinkoff, Derek M Houston, and Kathy Hirsh-Pasek, "Word learning in infant-and adult-directed speech," *Language Learning and Development*, vol. 7, no. 3, pp. 185–201, 2011.

[3] Dong Yu and Li Deng, *Automatic speech recognition: A deep learning approach*, Springer, 2014.

[4] Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky, "Multilingual MLP features for low-resource lvcsr systems," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4269–4272.

[5] Karel Veselỳ, Martin Karafiát, František Grézl, Miloš Janda, and Ekaterina Egorova, "The language-independent bottleneck features," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 336–341.

[6] 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.

[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 Communication 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 Communication Association*, 2015.

[9] 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.

[10] Tadahiro Taniguchi, Ryo Nakashima, Hailong Liu, and Shogo Nagasaka, "Double articulation analyzer with deep sparse autoencoder for unsupervised word discovery from speech signals," *Advanced Robotics*, vol. 30, no. 11-12, pp. 770–783, 2016.

[11] Kate M Knill, Mark JF Gales, Anton Ragni, and Shakti P Rath, "Language independent and unsupervised acoustic models for speech recognition and keyword spotting," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[12] 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.

[13] Man-hung Siu, Herbert Gish, Arthur Chan, William Belfield, and Steve Lowe, "Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery," *Computer Speech & Language*, vol. 28, no. 1, pp. 210–223, 2014.

[14] 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.

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

[16] Leonardo Badino, Alessio Mereta, and Lorenzo Rosasco, "Discovering discrete subword units with binarized autoencoders and hidden-markov-model encoders," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

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

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

[19] S Young, G Evermann, M Gales, T Hain, D Kershaw, X Liu, G Moore, J Odell, D Ollason, D Povey, et al., "The htk book (v3. 4)," *Cambridge University*, 2006.

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

[21] 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.

[22] 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.

[23] 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.

[24] Ansari T.K, Rajath Kumar, Sonali Singh, and Sriram Ganapathy, "Deep learning methods for unsupervised acoustic modeling - LEAP submission to ZeroSpeech challenge 2017," in *IEEE 2017 workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2017.