A HIERARCHICAL SUBSPACE MODEL FOR LANGUAGE-ATTUNED ACOUSTIC UNIT DISCOVERY

Bolaji Yusuf† Lucas Ondel† Lukáš Burget† Jan Černocky† Murat Saraçlar*

* Boğaziçi University, Department of Electrical and Electronics Engineering, Istanbul, Turkey
† Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia

ABSTRACT

In this work, we propose a hierarchical subspace model for acoustic unit discovery. In this approach, we frame the task as one of learning embeddings on a low-dimensional phonetic subspace, and simultaneously specify the subspace itself as an embedding on a hyper-subspace. We train the hyper-subspace on a set of transcribed languages and transfer it to the target language. In the target language, we infer both the language and unit embeddings in an unsupervised manner, and in so doing, we simultaneously learn a subspace of units specific to that language and the units that dwell on it. We conduct our experiments on TIMIT and two low-resource languages: Mboshi and Yoruba. Results show that our model outperforms major acoustic unit discovery techniques, both in terms of clustering quality and segmentation accuracy.

Index Terms— acoustic unit discovery, hierarchical subspace model, unsupervised learning

1. INTRODUCTION

Current machine learning approaches for speech processing rely on large collections of annotated audio recordings. In contrast, infants learn to speak long before they are able to read and write. Augmenting machines with similar capability would have a great impact. First, it would drastically reduce the cost of data annotation, therefore allowing speech technologies to be extended to low-resource languages. Second, by the proposed “reverse-engineering” approach to cognitive science [1], it would pave the way to a better understanding of human learning.

In this paper, we focus on the task of Acoustic Unit Discovery (AUD). This task consists of discovering an inventory of phonetic units—denoted “acoustic units”—from a sequence of untranscribed recordings. This is a simplified model of a language acquisition where we consider learning phonetics rather than the complete structure of speech (phones, syllables, words, ...).

The work was supported by Czech National Science Foundation (GAČR) project “NEUREM3” No. 19-26934X, Czech Ministry of Education, Youth and Sports project No. LTAIN190087 “Multi-linguality in speech technologies”, and by European Union’s Horizon 2020 project No. 870930 - WELCOME.

As in [8, 9], \( p(X|z, H) \) is given by an HMM with parameters \( \eta^v \), and we further factorize the prior:

\[
p(z, H) = p(z|H) \prod_{u=1}^{U} p(\eta^u).
\]

Note that the number of units \( U \) is unknown and also needs to be learned for an unknown language. Prior work [9] addresses this issue by constructing \( p(z|H) \) with a sample from a Dirichlet process [13] with base measure \( p(\eta) \). This leads to an infinite “phone-loop” model where each acoustic unit component is a 3-state left-to-right HMM. The exact relation between the Dirichlet Process and the phone-loop structure of the model is discussed at length in [14]. In this work we focus on the construction of the base measure and we leave the rest of the model unaltered.

The base measure \( p(\eta) \) defines a prior probability that a sound—represented by an HMM with parameters \( \eta \)—is an acoustic unit. Earlier works on Bayesian AUD [8, 9, 15, 16] use exponential family distributions as the base measure. These distributions, while mathematically convenient since they form conjugate priors, do not incorporate any knowledge about phones. For instance, the models \textit{a priori} may consider the sound of a car engine to be as likely an acoustic unit as the “ah” sound. Perhaps more detrimentally, the models are also likely to model other sources of variability such as speaker, emotional state, channel etc.

Therefore, we utilize the generalized subspace model (GSM) [12] which provides a solution to the problem of specifying an educated

The problem of acoustic unit discovery can be formulated as that of learning a set of \( U \) discrete units with parameters \( H = \{\eta^1, \ldots, \eta^U\} \) from a sequence of untranscribed acoustic features \( X = (x_1, \ldots, x_N) \), as well as the assignment of frame to unit \( z = (z_1, \ldots, z_N) \). Formally, we seek to maximize:

\[
p(z, H|X) \propto p(X|z, H)p(z, H).
\]
base measure by defining the parameters of each unit \(u\) as:
\[
\eta^u = f(W^u \cdot e^u + b),
\]
where \(e^u\) is a low-dimensional unit embedding, \(W\) and \(b\) are the subspace parameters and \(f(\cdot)\) is a deterministic and differentiable function that ensures that the resulting vector \(\eta^u\) dwells in the HMM parameter space. \(W, e^u,\) and \(b\) are assumed to have Gaussian distributions with diagonal covariance matrices. The posteriors of \(W\) and \(b\) are estimated from other, transcribed, languages and fixed for the target language while the posteriors of \(e^u\) are learned in the target language. Thus, the parameters \(H = \{\eta^1, \ldots, \eta^L\}\) are constrained to a low-dimensional manifold of the parameter space.

### 3. Hierarchical Subspace HMM

The GSM of [12] enforces an educated prior by transferring the subspace parameters \((W, b)\) to a target language. This makes the implicit assumption that the subspace is universal i.e. that the units of all languages lie on the same manifold. We hypothesize that this is too strong an assumption, and that having a language-dependent subspace allows us to better model the units of a specific target language. However, naively training, or even fine-tuning, the subspace on the target language counter its purpose by removing the constraint on the space of units, thereby losing transferred phonetic information.

To deal with the dilemma, we propose a hierarchical subspace model (HSM). The crux of this model is to allow language-dependent subspaces, but only as long as they lie on a manifold of the hyper-space of subspaces, as depicted Fig. 1. Formally:

\[
W^\lambda = M_0 + \sum_{k=1}^K \alpha_k^\lambda M_k
\]

\[
b^\lambda = m_0 + \sum_{k=1}^K \alpha_k^\lambda m_k
\]

\[
\eta^\lambda, u = f(W^\lambda \cdot e^{\lambda, u} + b^\lambda),
\]

where \(W^\lambda\) and \(b^\lambda\) define the subspace for language \(\lambda\) and \(\eta^{\lambda, u}\) is the vector of parameters for unit \(u\) of language \(\lambda\). The unit parameters, \(\eta^{\lambda, u}\), are constructed from a linear combination of the columns of \(W^\lambda\) weighted by unit-specific embedding vectors \(e^{\lambda, u}\) and a bias vector \(b^\lambda\). Similarly, \(W^\lambda\) is defined by a linear combination of basis matrices \([M_1, \ldots, M_K]\) weighted by language-specific embedding vectors \(\alpha^\lambda = [\alpha_1^\lambda, \alpha_2^\lambda, \ldots, \alpha_K^\lambda]\) plus bias matrix \(M_0\). The bias vector \(b^\lambda\) is similarly obtained by a linear combination of \([m_1, \ldots, m_K]\) and \(\alpha^\lambda\) plus bias term \(m_0\).

We assume Gaussian priors for the random variables:

\[
\alpha_k^\lambda \sim \mathcal{N}(0, \sigma_a)
\]

\[
M_{k,ij} \sim \mathcal{N}(0, \sigma_M)
\]

\[
m_{k,i} \sim \mathcal{N}(0, \sigma_m)
\]

\[
e_i^{\lambda, u} \sim \mathcal{N}(0, \sigma_e),
\]

with variances set to 1. Note that the posterior distribution that we seek is modified from (1) to:

\[
p(z, E^\lambda, \alpha^\lambda, \mathcal{M}(X^\lambda)) \propto p(X^\lambda|z, E^\lambda, \alpha^\lambda, \mathcal{M})p(z|E^\lambda, \alpha^\lambda, \mathcal{M})
\]

\[
\cdot p(E^\lambda)p(\alpha^\lambda)p(\mathcal{M})
\]

where \(\mathcal{M} = \{M_0, \ldots, M_K, m_0, \ldots, m_K\}\), \(E^\lambda = \{e^{\lambda, 1}, \ldots, e^{\lambda, U^\lambda}\}\) and \(X^\lambda\) is language-specific data. To complete the definition of our generative process, we model the likelihood of a speech segment given an acoustic unit \(p(X_k^s|z_s = u, \Theta_H)\) as a 3-state left-to-right HMM with parameter vector:

\[
\eta^{\lambda, u} = [\eta_1^{\lambda, u\top}, \eta_2^{\lambda, u\top}, \eta_3^{\lambda, u\top}]^\top,
\]

and each state has emission probabilities modeled as a GMM with \(K = 4\) Gaussian components:

\[
\pi_i^{\lambda, u} = [\mu_{i,1}^{\lambda, u\top}, \ldots, \mu_{i,K}^{\lambda, u\top}, \text{vec}(\Sigma_{i,1}^{\lambda, u\top}), \ldots, \text{vec}(\Sigma_{i,K}^{\lambda, u\top})]^\top,
\]

where \([\cdot]^\top\) is the transpose operator, vec is the vectorize operator, \(\pi_{i,j}^{\lambda, u}\), \(\mu_{i,j}^{\lambda, u}\) and \(\Sigma_{i,j}^{\lambda, u}\) are the weight, mean and covariance matrix of the \(i\)th HMM state and the \(j\)th Gaussian component of the acoustic unit \(u\) in language \(\lambda\). The function \(f(\cdot)\) in (6) is defined as:

\[
\lambda, u_{i,j} = \frac{\exp\{W_{i,j} \cdot e^{\lambda, u} + b_{i,j}\}_k}{1 + \sum_{k=1}^{K-1} \exp\{W_{k,j} \cdot e^{\lambda, u} + b_{k,j}\}_k}
\]

\[
\Sigma_{i,j}^{\lambda, u} = \text{diag}(\exp\{W_{i,j} \cdot e^{\lambda, u} + b_{i,j}\})
\]

\[
\mu_{i,j}^{\lambda, u} = \Sigma_{i,j}^{\lambda, u} \cdot (W_{i,j} \cdot e^{\lambda, u} + b_{i,j}),
\]

where \(\exp\) is the element-wise exponential function and \(\text{exp}\{\cdot\}\) is the \(j\)th element of the resulting vector. \(W_{i,j}\) is the subset of rows of matrix \(W^\lambda\) assigned to the mixing weights \(\pi_{i,j}^{\lambda, u}\) of the \(i\)th HMM state. Matrices \(W_{i,j}^{\lambda, u}\) and \(W_{i,j}^{\lambda, u}\) are similarly defined for the mean and covariance matrix of the \(j\)th Gaussian component of the \(i\)th HMM state.

Thus we have a Hierarchical Subspace Hidden Markov Model (H-SHMM). Note that the choice of HMM as the likelihood model follows previous work [8,12,12] and is not integral to the proposed hierarchical subspace model.

### 3.1. Inference in the Hierarchical Subspace HMM

The H-SHMM training procedure follows the SHMM training [12] modified to accommodate the alterations made to the model. Given a set of \(L\) languages, our goal is to compute the parameters’ posterior:

\[
p(\{z^\lambda\}, \{E^\lambda\}, \{\alpha^\lambda\}, \mathcal{M}(X^\lambda)), \quad \lambda \in \{1, \ldots, L\},
\]

where \(z^\lambda, E^\lambda, \alpha^\lambda\) and \(X^\lambda\) are language-specific variables. For conciseness, we define \(\theta^\lambda = (\alpha^\lambda, \{E^\lambda\})\). Since (17) is intractable,
We seek an approximate posterior \( q \) by maximizing the empirical approximation of (20):

\[
q(\{z^\lambda\}, \{\theta^\lambda\}, M) = \prod_{\lambda=1}^{L} q(z^\lambda)q(\theta^\lambda)q(M)
\]

\[= q(z^\lambda) q(\theta^\lambda) q(M). \tag{18}\]

In addition, we impose the following parametric form on \( q(\{\theta^\lambda\}) \) and \( q(M) \):

\[
q(\{\theta^\lambda\})q(M) = \mathcal{N}(\omega, \text{diag}(\exp\{\psi\})). \tag{19}\]

With the factorization in (18), the variational lower-bound becomes:

\[
\mathcal{L}[q] = \sum_{\lambda=1}^{L} \left[ \left( \ln p(X|z^\lambda, \theta^\lambda, M) \right)_q - \text{D}_{KL} \left( q(z^\lambda)||p(z^\lambda) \right) \right] - \text{D}_{KL} \left( q(\theta^\lambda)||p(\theta^\lambda) \right) - \text{D}_{KL} \left( q(M)||p(M) \right). \tag{20}\]

We optimize (20) through an expectation-maximization procedure where we iteratively re-estimate each of the variational posteriors \( q(\{z^\lambda\}) \) and \( q(\{\theta^\lambda\})q(M) \) given the current estimate of the other.

In the expectation step, we compute \( q(z^\lambda) \) which maximizes (20) using a modified forward-backward algorithm. Instead of the log-likelihood, we use the expectation of the log-likelihood with respect to the posterior of the the HMM parameters. More details on this can be found in [14].

In the maximization step, we compute \( q(\{\theta^\lambda\})q(M) \) which maximizes (20). Since this has no closed-form solution, we instead optimize an empirical approximation of (20):

\[
\mathcal{L}(\omega, \psi) = \frac{1}{S} \sum_{s=1}^{S} \left\{ \sum_{\lambda=1}^{L} \left[ \langle \ln p(X|z^\lambda, \theta^\lambda, M_s) \rangle_q - \text{D}_{KL} \left( q(\theta^\lambda)||p(\theta^\lambda) \right) \right] - \text{D}_{KL} \left( q(M_s)||p(M) \right) \right\}. \tag{21}\]

\[
(\{\theta^\lambda\}, M_s) = \omega + \exp\left( \frac{\psi}{2} \right) \odot \epsilon_s, \hspace{1em} \epsilon_s \sim \mathcal{N}(0, I), \tag{22}\]

where \( \odot \) is the element-wise multiplication operation. Equations (21)-(22) are a special case of the so-called re-parameterization trick [17]. We use stochastic gradient ascent to maximize (21) with respect to \( \omega \) and \( \psi \).

The first term in (21) is a sum over terms computed separately for each Gaussian component \( c = (\lambda, u, i, j) \)–th Gaussian component of the \( u \)th state of unit \( u \) in language \( \lambda \):

\[
\langle \ln p(X|\theta^\lambda, z^\lambda, c) \rangle_q = \frac{N_c}{2} \left( 2\pi \sigma^2 + \ln |\Lambda^\lambda_c| - \mu^\lambda_c^T \Lambda^\lambda_c^{-1} \mu^\lambda_c \right) + \phi^\lambda_c^T \Lambda^\lambda_c^{-1} \mu^\lambda_c - \text{tr} \left( \Phi^\lambda_c \Lambda^\lambda_c^{-1} \right) + \text{const} . \tag{23}\]

For each component, the mean \( \mu^\lambda_c \), precision matrix \( \Lambda^\lambda_c = \left( \Sigma^\lambda_c \right)^{-1} \) and mixing weight \( \pi^\lambda \) are obtained from a sample \( \langle M_s, \theta^\lambda \rangle \) using equations (14)–(16) \( \gamma_{nc} \) are the Gaussian responsibilities for each time frame \( n \) computed in the expectation step. \( N_s = \sum_n \gamma_{nc} \ {}; \ 
\phi^\lambda_c = \sum_n \gamma_{nc} x_n \; \Phi^\lambda_c = \sum_n \gamma_{nc} x_n x_n^T \) are the zeroth, first and second order sufficient statistics respectively for Gaussian component \( c \). Note that (23) is a differentiable function of \( \omega \) and \( \psi \).

Overall, the training of our AUD system comprises two stages. First, we infer a set of variational posteriors \( q_0(\{z^\lambda\}), q_0(\{\theta^\lambda\}) \) and \( q_0(M) \) on transcribed source languages where \( \lambda \in \{1, \ldots, L\} \). At this stage, the phone-loop of the AUD model is replaced with a forced alignment graph. Then, on the target language \( t \), we infer new variational posteriors \( q_1(\{z^\lambda\}), q_1(\{\theta^\lambda\}) \) using \( q_0(M) \) to compute the expectation and maximization steps. Note that \( q_0(M) \) is not updated during this stage, but transferred as is from the source languages.

Finally, the output of our AUD system is obtained from a modified Viterbi algorithm [14] which uses the expectation of the log-likelihoods with respect to \( q_1(\{\theta^\lambda\})q_0(M) \) instead of point estimates.

## 4. Related Work

Our proposed model builds heavily on the generalized subspace model of [14]. Our novelty is that we introduce a hyper-subspace to allow unsupervised adaptation of the subspace itself. In particular, if the language embedding \( \alpha \) is set to 0, then the H-SHMM becomes identical to the SHMM. Another related model is the Subspace Gaussian Mixture Model (SGMM) used for ASR acoustic modeling [18]. While the SGMM incorporates a subspace of means of individual Gaussian components and mixture weights, following the SHMM, our subspace models entire 3-state HMMs including covariance matrices. Moreover, where the SGMM uses maximum likelihood training, we use a variational Bayes framework to infer the posterior distributions of the parameters.

## 5. Experiments

### 5.1. Data and Features

We test the performance of our models on the following languages:

1. Mbooshi [19]: 4.4 hours with 5130 utterances by 3 speakers.
2. Yoruba [20]: 4 hours with 3583 utterances by 36 speakers.
3. English: from TIMIT [21] excluding the sas utterances. 3.6 hours with 4288 utterances by 536 speakers.

Note that we train and test on the entirety of each corpus since we are doing unsupervised learning. We include English in order to have a control language to facilitate comparison with other baselines.

We use seven transcribed languages for training the hypersubspace: German, Spanish, French and Polish from Globalphone [24]; and Amharic [23]. Swahili [24] and Wolof [25] from the ALPFA project [26]. For each of these, we use only a subset of 1500 utterances which corresponds to 3-5 hours per language.

### 5.2. Metrics

We evaluate the performance of our models with two metrics for phone segmentation and phone clustering. For segmentation, we report the F-score on phone boundary detection with a tolerance of ±20 milliseconds. For clustering, we report the normalized mutual information (NMI) which is computed from a frame-level confusion matrix of discovered units (U) and actual phones (P) as:

\[
\text{NMI}(P, U) = 200 \times \frac{I(P; U)}{H(P) + H(U)} \%	ag{24}\]

where \( H(\cdot) \) is the Shannon entropy and \( I(P; U) \) is the (un)normalized mutual information. An NMI of 0 means that the discovered acoustic units are completely unrelated to the actual phones, while an NMI of 100 means that the units have a one-to-one correspondence with the actual phones. Note that the \( H(U) \) term in the denominator rewards more compact representations.
5.3. Experiment setup

Although the Dirichlet process prior allows us to model an arbitrary number of units, in practice, we set the truncation parameter $T$ to 100, and we find that number of discovered units tends to be fewer. We set the dimension of the H-SHMM unit embeddings $\alpha^{\lambda,u} = 100$. We set the dimension of the language embeddings $\alpha^{\lambda} = 6$. We use 5 samples for the re-parametrization trick and train with Adam optimizer [28] with a learning rate of $5 \times 10^{-3}$.

We report results on five baselines: HMM [9], SHMM [12], VQ-VAE [7], VQ-wav2vec [6] and ResDAVenet [29]. We use 13-dimensional MFCC features along with their first and second derivatives as features for the H-SHMM, HMM, SHMM and VQ-VAE.

The HMM [9] and SHMM [12] are the most comparable models to the H-SHMM. We set the truncation parameter to 100 for both. Furthermore, we train the SHMM subspace with the same source languages that we use to train the H-SHMM hyper-subspace and we set the subspace dimension to 100.

We also implemented VQ-VAE [7] baselines as they have been shown to learn good discrete representations of speech [5]. The most critical choices we made were: (i) having a big encoder (5 BLSTM layers) but weak decoder (feed-forward with one hidden layer) as stronger decoders resulted in better reconstruction error but worse latent representations, (ii) low-dimensional latent space (16-d) to discard irrelevant information, (iii) relatively few units (50 centroids) and down-sampled encoder (by factor of 2) to prevent over-segmentation, and (iv) concatenating a learned speaker-embedding (32-d) to the decoder input to help the encoder focus more on phonetic information. We tune these parameters to maximize the NMI and F-score on English and kept them for the other two languages.

We report results for two other neural discrete representation learning models: VQ-wav2vec [6] and ResDAVenet-VQ [29]. Note that we cannot replicate these models on our low-resource languages as the former is trained with 960 hours of Librispeech [30] data while the latter requires paired captioned images for training [31]. Therefore, in both cases, we use the authors’ own code and pre-trained models to extract discrete representations for the TIMIT utterances and report those results. We tested the various pre-trained models available for each system and report only the best performing ones.

For VQ-wav2vec, we report results for the Gumbel softmax variant since it gave higher NMI and F-score than the K-means variant.

For ResDAVenet-VQ, we report results on the “{3} → {2,3}” model (ResDAVenet-VQ-I) with units extracted from layer-2 and the “{2} → {2,3}” model (ResDAVenet-VQ-II) with units extracted from layer-3 as we found they had the highest NMI and F-score respectively of the available pre-trained models.

Our H-SHMM, HMM and SHMM code are publicly available [1].

Our implementation of the VQ-VAE is also public [2].

5.4. Experiment results

We train each system with 5 random initializations and report the means and standard deviations of the results in Table 1. The SHMM and H-SHMM subspace and hyper-subspace respectively are trained once, and only the AUD is repeated 5 times. Since we use pre-trained VQ-wav2vec and ResDAVenet-VQ models, we only run them once per language. From the results, we take the SHMM as our main baseline since it outperforms the other baselines on all metrics. Moreover, it provides the best direct comparison as it is the most structurally similar baseline to our proposed model.

We achieve significant NMI improvements over the SHMM with the H-SHMM. We also get similar F-score improvements in English and Yoruba, but get a slightly worse average F-score on Mboshi.

The novelty of the H-SHMM is that we introduce a way of adapting the subspace to the target language. We tested whether simply fine-tuning the SHMM subspace parameters on the target language would achieve the same results. The result of fine-tuning (SHMM+finetune) is not just worse than H-SHMM, it is in fact worse than SHMM. This is intuitive because fine-tuning the subspace parameters relaxes the constraint on the HMM parameters.

Another difference is that the H-SHMM has more transferred parameters than the SHMM. Therefore, we experimented with increasing the number of transferred SHMM parameters by changing the subspace dimension to 300 (SHMM+300d). The results show no significant benefit over the SHMM with dimension 100.

To visualize the learned language embeddings, we train an H-SHMM with $|\alpha^{\lambda}| = 2$. For this experiment, we split each corpus into four subsets, so that the model a priori treats each subset as a different language. After inference, we find that $\alpha^{\lambda}$ of different subsets of the same language converge with small within-language variance for source languages and higher variance for target languages, so the model is able to cluster subsets that come from the same language without being told. The images can be found at [1].

![Image](https://www.fit.vutbr.cz/~iyusuf/hshmm.html)

Table 1. Acoustic unit discovery results.

| Corpus     | System          | NMI       | F-score |
|------------|-----------------|-----------|---------|
| English    | ResDAVenet-VQ-I | 35.93     | 54.19   |
|            | ResDAVenet-VQ-II| 34.39     | 64.36   |
| VQ-wav2vec |                 | 35.20     | 68.84   |
| VQ-VAE     |                 | 32.03 ± 0.30 | 59.05 ± 0.34 |
| HMM        |                 | 35.91 ± 0.27 | 63.86 ± 0.95 |
| SHMM       |                 | 39.17 ± 0.16 | 74.65 ± 0.60 |
| SHMM+finetune |               | 37.83 ± 0.25 | 72.20 ± 0.65 |
| SHMM+300d  |                 | 39.62 ± 0.20 | 73.62 ± 0.84 |
| H-SHMM (ours) |               | 40.04 ± 0.51 | 76.60 ± 0.54 |
| Mboshi     | VQ-VAE          | 31.27 ± 0.26 | 39.19 ± 0.71 |
|            | HMM             | 35.85 ± 0.62 | 47.92 ± 1.56 |
|            | SHMM            | 38.38 ± 0.97 | 59.50 ± 0.78 |
|            | SHMM+finetune   | 36.09 ± 0.49 | 53.06 ± 1.06 |
|            | SHMM+300d       | 37.51 ± 0.45 | 53.71 ± 1.41 |
|            | H-SHMM (ours)   | 41.07 ± 1.09 | 59.15 ± 1.51 |
| Yoruba     | VQ-VAE          | 29.90 ± 0.40 | 37.52 ± 0.79 |
|            | HMM             | 36.38 ± 0.22 | 54.47 ± 0.64 |
|            | SHMM            | 38.99 ± 0.08 | 64.46 ± 0.51 |
|            | SHMM+finetune   | 36.97 ± 0.38 | 58.59 ± 0.34 |
|            | SHMM+300d       | 39.08 ± 0.13 | 61.09 ± 1.01 |
|            | H-SHMM (ours)   | 40.06 ± 0.11 | 66.95 ± 0.36 |

6. CONCLUSION

In this paper, we have proposed a hierarchical subspace model for unsupervised phonetic discovery in which a phonetic subspace constrains the parameters and ensures that the learned parameters define a plausible phone. Similarly, a hyper-subspace constrains the parameters of the subspace itself. We have shown that the proposed model outperforms the non-hierarchical baseline as well as other neural network-based AUD models.

Going forward, we hope to explore better generative models than the HMM, as well as different subspace hierarchies, such as having different subspaces for various phone classes or speaker subspaces.
7. REFERENCES

[1] Emmanuel Dupoux, “Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner,” *Cognition*, vol. 173, pp. 43–59, 2018.

[2] Maarten Versteegh et al., “The Zero Resource Speech Challenge 2015,” in *Sixteenth annual conference of the international speech communication association*, 2015.

[3] Ewan Dunbar et al., “The Zero Resource Speech Challenge 2017,” in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 323–330.

[4] Ewan Dunbar et al., “The Zero Resource Speech Challenge 2019: TTS Without T,” in *Interspeech*, 2019, pp. 1088–1092.

[5] Jan Chorowski, Ron J Weiss, Samy Bengio, and Aaron van den Oord, “Unsupervised speech representation learning using wavenet autoencoders,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019.

[6] Alexey Baevski, Steffen Schneider, and Michael Auli, “vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations,” in *International Conference on Learning Representations*, 2020.

[7] Aaron Van Den Oord, Oriol Vinyals, et al., “Neural discrete representation learning,” in *Advances in Neural Information Processing Systems*, 2017, pp. 6306–6315.

[8] Chia-ying Lee and James Glass, “A nonparametric Bayesian approach to acoustic model discovery,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 40–49.

[9] Lucas Ondel, Lukáš Burget, and Jan Černocký, “Variational inference for acoustic unit discovery,” *Procedia Computer Science*, vol. 81, pp. 80–86, 2016.

[10] Hongjie Chen, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li, “Multilingual bottleneck feature learning from untranscribed speech,” in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 727–733.

[11] Thomas Grannes, Patrick Hanebrink, Janek Ebbers, and Reinhold Haeb-Umbach, “Full Bayesian Hidden Markov Model Variational Autoencoder for Acoustic Unit Discovery,” in *Interspeech*, 2018, pp. 2688–2692.

[12] Lucas Ondel, Hari Krishna Vydana, Lukáš Burget, and Jan Černocký, “Bayesian Subspace Hidden Markov Model for Acoustic Unit Discovery,” in *Interspeech*, 2019, pp. 261–265.

[13] Y. W. Teh, “Dirichlet processes,” in *Encyclopedia of Machine Learning*. Springer, 2010.

[14] Lucas Ondel, *Discovering Acoustic Units from Speech: A Bayesian Approach*, Ph.D. thesis, Brno University of Technology, Faculty of Information Technology, to appear.

[15] Lucas Ondel et al., “Bayesian models for unit discovery on a very low resource language,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5939–5943.

[16] Lucas Ondel, Lukáš Burget, Jan Černocký, and Santosh Kesiraju, “Bayesian phonotactic language model for acoustic unit discovery,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5750–5754.

[17] Diederik P Kingma and Max Welling, “Auto-encoding Variational Bayes,” in *ICLR*, 2014.

[18] Daniel Povey et al., “The subspace gaussian mixture model—a structured model for speech recognition,” *Computer Speech & Language*, vol. 25, no. 2, pp. 404 – 439, 2011. Language and speech issues in the engineering of companionable dialogue systems.

[19] Pierre Godard et al., “A very low resource language speech corpus for computational language documentation experiments,” arXiv preprint arXiv:1710.03501, 2017.

[20] Alexander Gutkin, İşın Demişşahin, Oddur Kjartansson, Clara Rivera, and Kólá Túbisín, “Developing an Open-Source Corpus of Yoruba Speech,” in *Interspeech*, Shanghai, China, 2020.

[21] J Garofolo, L Lamel, W Fisher, J Fiscus, D Pallet, and N Dahlgren, “The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM. NTIS order number PB91-505065,” 1990.

[22] Tanja Schultz, Ngoc Thang Vu, and Tim Schlipke, “Globalphone: A multilingual text & speech database in 20 languages,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8126–8130.

[23] Solomon Teferra Abate, Wolfgang Menzel, and Bairu Tafila, “An Amharic Speech Corpus for Large Vocabulary Continuous Speech Recognition,” in *INTERSPEECH-2005*, 2005.

[24] Hadrien Gelas, Laurent Besacier, and Francois Pellegrino, “Developments of Swahili resources for an automatic speech recognition system,” in *SLTU - Workshop on Spoken Language Technologies for Under-Resourced Languages*. Cape-Town, Afrique Du Sud, 2012.

[25] Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui, “Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition: a Case Study of Wolof,” *LREC*, 2016.

[26] Laurent Besacier et al., “Speech technologies for african languages: Example of a multilingual calculator for education,” in *Interspeech*, 2015.

[27] David M Blei, Michael I Jordan, et al., “Variational inference for Dirichlet process mixtures,” *Bayesian analysis*, vol. 1, no. 1, pp. 121–143, 2006.

[28] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[29] David Harwath, Wei-Ning Hsu, and James Glass, “Learning Hierarchical Discrete Linguistic Units from Visually-Grounded Speech,” in *International Conference on Learning Representations*, 2020.

[30] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[31] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, “Learning deep features for scene recognition using places database,” in *Advances in neural information processing systems*, 2014, pp. 487–495.