ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414899

A HIERARCHICAL SUBSPACE MODEL FOR LANGUAGE-ATTUNED ACOUSTIC UNIT DISCOVERY

Bolaji Yusuf[⋆†] *Lucas Ondel*[†] *Lukáš Burget*[†] *Jan Černocký*[†] *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 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 phone-like units—denoted "acoustic units"—from a set 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 AUD task has been the subject of numerous publications [2, 3, 4]. Nowadays, two major approaches are widely used: (i) neural-network-based models which typically use auto-encoder structure with a discretization layer [5, 6, 7] (ii) non-parametric Bayesian generative-based models which can be seen as infinite mixtures time series models [8, 9, 10], or hybrids of both as in [11].

This work follows the Bayesian paradigm and is a direct extension of [12], where the target language's acoustic units parameters are forced to lie on a language-independent phonetic subspace that is estimated from several transcribed languages.

We propose a subspace that is adapted to the target-language in an unsupervised fashion. We achieve this by learning a language-independent *hyper*-subspace from transcribed data in other languages, and a low-dimensional embedding vector for the target (low-resource) language. The hyper-subspace is a set of matrices which can be thought of as subspace "templates" and the embedding determines how these templates are combined for the target language. Thus we have *hierarchical* structure in which the lower level constrains units and the higher level constrains subspaces.

2. PROBLEM DEFINITION

The problem of acoustic unit discovery can be formulated as that of learning a set of U discrete units with parameters $\mathbf{H} = \{\boldsymbol{\eta}^1, \dots, \boldsymbol{\eta}^U\}$ from a sequence of untranscribed acoustic features $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$, as well as the assignment of frame to unit $\mathbf{z} = (z_1, \dots, z_N)$. Formally, we seek the posterior:

$$p(\mathbf{z}, \mathbf{H}|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{z}, \mathbf{H}) p(\mathbf{z}, \mathbf{H}). \tag{1}$$

As in [8, 9], $p(\mathbf{X}|z_n, \mathbf{H})$ is given by an HMM with parameters $\boldsymbol{\eta}^{z_n}$, and we further factorize the prior:

$$p(\mathbf{z}, \mathbf{H}) = p(\mathbf{z}|\mathbf{H}) \prod_{u=1}^{U} p(\boldsymbol{\eta}^u). \tag{2}$$

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 $p(\mathbf{z}|\mathbf{H})$ with a sample from a Dirichlet process [13] with base measure $p(\boldsymbol{\eta})$. 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 $p(\boldsymbol{\eta})$ defines a prior probability that a sound—represented by an HMM with parameters $\boldsymbol{\eta}$—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 *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 work was supported by Czech National Science Foundation (GACR) project "NEUREM3" No. 19-26934X, Czech Ministry of Education, Youth and Sports project No. LTAIN19087 "Multi-linguality in speech technologies", and by European Union's Horizon 2020 project No. 870930 - WELCOME.

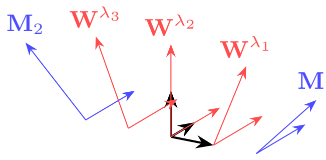

Fig. 1. Illustration of a hierarchical subspace model. For each language λ, acoustic unit embeddings (encoding the parameters of a probabilistic model) are assumed to live in a language-specific subspace $\mathbf{W}^\lambda$ of the total parameter space. This subspace is given by a weighted sum of matrix bases $\mathbf{M}_1, \mathbf{M}_2, \ldots$ (shared across languages) and language-specific weights $\boldsymbol{\alpha}^\lambda$: $\mathbf{W}^\lambda = \alpha_1^\lambda \mathbf{M}_1 + \alpha_2^\lambda \mathbf{M}_2 + \ldots$.

base measure by defining the parameters of each unit u as:

$$\boldsymbol{\eta}^u = f(\mathbf{W} \cdot \mathbf{e}^u + \mathbf{b}), \tag{3}$$

where $\mathbf{e}^u$ is a low-dimensional unit embedding, $\mathbf{W}$ and $\mathbf{b}$ are the subspace parameters and $f(\cdot)$ is a deterministic and differentiable function that ensures that the resulting vector $\boldsymbol{\eta}^u$ dwells in the HMM parameter space. $\mathbf{W}$, $\mathbf{e}^u$, and $\mathbf{b}$ are assumed to have Gaussian distributions with diagonal covariance matrices. The posteriors of $\mathbf{W}$ and $\mathbf{b}$ are estimated from other, transcribed, languages and fixed for the target language while the posteriors of $\mathbf{e}^u$ are learned in the target language. Thus, the parameters $\mathbf{H} = \{\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^U\}$ are constrained to a low-dimensional subspace of the parameter space.

3. HIERARCHICAL SUBSPACE HMM

The GSM of [12] enforces an educated prior by transferring the subspace parameters $(\mathbf{W}, \mathbf{b})$ 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 subspace. We hypothesize that this is too strong an assumption, and that having a language-dependent subspace will allow us to better model the units of a specific target language. However, naively training, or even fine-tuning, the subspace on the target language counters 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 subspace of the *hyper-space* of subspaces, as depicted Fig. 1. Formally:

$$\mathbf{W}^\lambda = \mathbf{M}_0 + \sum_{k=1}^{K} \alpha_k^\lambda \mathbf{M_k} \tag{4}$$

$$\mathbf{b}^\lambda = \mathbf{m}_0 + \sum_{k=1}^{K} \alpha_k^\lambda \mathbf{m_k} \tag{5}$$

$$\boldsymbol{\eta}^{\lambda,u} = f(\mathbf{W}^\lambda \cdot \mathbf{e}^{\lambda,u} + \mathbf{b}^\lambda), \tag{6}$$

where $\mathbf{W}^\lambda$ and $\mathbf{b}^\lambda$ define the subspace for language λ and $\boldsymbol{\eta}^{\lambda,u}$ is the vector of parameters for unit u of language λ. The unit parameters, $\boldsymbol{\eta}^{\lambda,u}$ are constructed from a linear combination of the columns of $\mathbf{W}^\lambda$ weighted by unit-specific embedding vectors $\mathbf{e}^{\lambda,u}$ and a bias vector $\mathbf{b}^\lambda$. Similarly, $\mathbf{W}^\lambda$ is defined by a linear combination of basis matrices $[\mathbf{M}_1, \ldots, \mathbf{M}_K]$ weighted by language-specific embedding vectors $\boldsymbol{\alpha}^\lambda = [\alpha_1^\lambda, \alpha_2^\lambda, \ldots, \alpha_K^\lambda]^\top$ plus bias matrix $\mathbf{M}_0$. The bias vector $\mathbf{b}^\lambda$ is similarly obtained by a linear combination of $[\mathbf{m}_1, \ldots, \mathbf{m}_K]$ and $\boldsymbol{\alpha}^\lambda$ plus bias term $\mathbf{m}_0$.

We impose Gaussian priors on the random variables:

$$\alpha_k^\lambda \sim \mathcal{N}(0, \sigma_\alpha) \tag{7}$$

$$M_{k,ij} \sim \mathcal{N}(0, \sigma_M) \tag{8}$$

$$m_{k,i} \sim \mathcal{N}(0, \sigma_m) \tag{9}$$

$$e_i^{\lambda,u} \sim \mathcal{N}(0, \sigma_e), \tag{10}$$

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

$$p(\mathbf{z}, \mathbf{E}^\lambda, \boldsymbol{\alpha}^\lambda, \mathcal{M} | \mathbf{X}^\lambda) \propto p(\mathbf{X}^\lambda | \mathbf{z}, \mathbf{E}^\lambda, \boldsymbol{\alpha}^\lambda, \mathcal{M}) p(\mathbf{z} | \mathbf{E}^\lambda, \boldsymbol{\alpha}^\lambda, \mathcal{M}) \cdot p(\mathbf{E}^\lambda) p(\boldsymbol{\alpha}^\lambda) p(\mathcal{M}) \tag{11}$$

$\mathcal{M} = (\mathbf{M}_0, \ldots, \mathbf{M}_K, \mathbf{m}_0, \ldots, \mathbf{m}_K)$, $\mathbf{E}^\lambda = \{\mathbf{e}^{\lambda,1}, \ldots, \mathbf{e}^{\lambda,U^\lambda}\}$ and $\mathbf{X}^\lambda$ is language-specific data. To complete the definition of our generative process, we model the likelihood of a speech segment $\mathbf{X}_s^\lambda$ given an acoustic unit $p(\mathbf{X}_s^\lambda | \mathbf{z}_s = u, \mathbf{H})$ as a 3-state left-to-right HMM with parameter vector:

$$\boldsymbol{\eta}^{\lambda,u} = \left[\boldsymbol{\eta}_1^{\lambda,u\top}, \boldsymbol{\eta}_2^{\lambda,u\top}, \boldsymbol{\eta}_3^{\lambda,u\top}\right]^\top, \tag{12}$$

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

$$\boldsymbol{\eta}_i^{\lambda,u} = \left[\boldsymbol{\mu}_{i,1}^{\lambda,u\top}, \ldots, \boldsymbol{\mu}_{i,K}^{\lambda,u\top}, \mathrm{vec}(\boldsymbol{\Sigma}_{i,1}^{\lambda,u})^\top, \ldots, \mathrm{vec}(\boldsymbol{\Sigma}_{i,K}^{\lambda,u})^\top, \pi_{i,1}^{\lambda,u} \ldots \pi_{i,K}^{\lambda,u}\right]^\top, \tag{13}$$

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

$$\pi_{i,j}^{\lambda,u} = \frac{\exp\{\mathbf{W}_\pi^{\lambda,i} \cdot \mathbf{e}^{\lambda,u} + \mathbf{b}_\pi^{\lambda,i}\}_j}{1 + \sum_{k=1}^{K-1} \exp\{\mathbf{W}_\pi^{\lambda,i} \cdot \mathbf{e}^{\lambda,u} + \mathbf{b}_\pi^{\lambda,i}\}_k} \tag{14}$$

$$\boldsymbol{\Sigma}_{i,j}^{\lambda,u} = \mathrm{diag}(\exp\{\mathbf{W}_\Sigma^{\lambda,i,j} \cdot \mathbf{e}^{\lambda,u} + \mathbf{b}_\Sigma^{\lambda,i}\}) \tag{15}$$

$$\boldsymbol{\mu}_{i,j}^{\lambda,u} = \boldsymbol{\Sigma}_{i,j}^{\lambda,u} \cdot \left(\mathbf{W}_\mu^{\lambda,i,j} \cdot \mathbf{e}^{\lambda,u} + \mathbf{b}_\mu^{\lambda,i}\right), \tag{16}$$

where exp is the element-wise exponential function and $\exp\{\ldots\}_j$ is the jth element of the resulting vector. $\mathbf{W}_\pi^{\lambda,i}$ is the subset of rows of matrix $\mathbf{W}^\lambda$ assigned to the mixing weights $\boldsymbol{\pi}_i^{\lambda,\cdot}$ of the ith HMM state. Matrices $\mathbf{W}_\mu^{\lambda,i,j}$ and $\mathbf{W}_\Sigma^{\lambda,i,j}$ are similarly defined for the mean and covariance matrix of the jth Gaussian component of ith HMM state.

Thus we have a Hierachical Subspace Hidden Markov Model (H-SHMM). Note that the choice of HMM as the likelihood model follows previous work [8, 9, 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:

$$p(\{\mathbf{z}^\lambda\}, \{\mathbf{E}^\lambda\}, \{\boldsymbol{\alpha}^\lambda\}, \mathcal{M} | \{\mathbf{X}^\lambda\}), \quad \lambda \in \{1, \ldots, L\}, \tag{17}$$

where $\mathbf{z}^\lambda$, $\mathbf{E}^\lambda$, $\boldsymbol{\alpha}^\lambda$ and $\mathbf{X}^\lambda$ are language-specific variables. For conciseness, we define $\boldsymbol{\theta}^\lambda = (\{\boldsymbol{\alpha}^\lambda\}, \{\mathbf{E}^\lambda\})$. Since (17) is intractable, we seek an approximate posterior q by maximizing the variational lower-bound $\mathcal{L}[q]$ subject to the following mean-field factorization:

$$
\begin{aligned}
q(\{\mathbf{z}^\lambda\}, \{\boldsymbol{\theta}^\lambda\}, \mathcal{M}) &= \Big[\prod_{\lambda=1}^{L} q(\mathbf{z}^\lambda) q(\boldsymbol{\theta}^\lambda)\Big] q(\mathcal{M}) \\
&= q(\{\mathbf{z}^\lambda\}) q(\{\boldsymbol{\theta}^\lambda\}) q(\mathcal{M}). \qquad (18)
\end{aligned}
$$

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

$$
q(\{\boldsymbol{\theta}^\lambda\}) q(\mathcal{M}) = \mathcal{N}\Big(\boldsymbol{\omega}, \mathrm{diag}(\exp\{\boldsymbol{\psi}\})\Big). \qquad (19)
$$

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

$$
\begin{aligned}
\mathcal{L}[q] = \sum_{\lambda=1}^{L} \Bigg[&\Big\langle \ln p(\mathbf{X}^\lambda | \mathbf{z}^\lambda, \boldsymbol{\theta}^\lambda, \mathcal{M}) \Big\rangle_q - \mathrm{D}_{\mathrm{KL}} \Big(q(\mathbf{z}^\lambda) || p(\mathbf{z}^\lambda) \Big) \\
&- \mathrm{D}_{\mathrm{KL}} \Big(q(\boldsymbol{\theta}^\lambda) || p(\boldsymbol{\theta}^\lambda) \Big) \Bigg] - \mathrm{D}_{\mathrm{KL}} \Big(q(\mathcal{M}) || p(\mathcal{M}) \Big). \qquad (20)
\end{aligned}
$$

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

In the *expectation* step, we compute $q(\{\mathbf{z}^\lambda\})$ which maximizes (20) using the forward-backward (FB) algorithm. Where conventional FB uses log-likelihoods to compute this posterior, we use the expectation of the log-likelihood wrt the posterior of the HMM parameters. More details on this can be found in [14].

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

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\omega}, \boldsymbol{\psi}) = &\frac{1}{S} \sum_{s=1}^{S} \Bigg\{ \sum_{\lambda=1}^{L} \Bigg[\Big\langle \ln p(\mathbf{X}^\lambda | \mathbf{z}^\lambda, \boldsymbol{\theta}_s^\lambda, \mathcal{M}_s) \Big\rangle_{q(\mathbf{z}^\lambda)} \\
&- \mathrm{D}_{\mathrm{KL}} \Big(q(\boldsymbol{\theta}_s^\lambda) || p(\boldsymbol{\theta}_s^\lambda) \Big) \Bigg] - \mathrm{D}_{\mathrm{KL}} \left(q(\mathcal{M}_s) || p(\mathcal{M}_s) \right) \Bigg\}, \qquad (21)
\end{aligned}
$$

$$
(\{\boldsymbol{\theta}_s^\lambda\}, \mathcal{M}_s) = \boldsymbol{\omega} + \exp\{\frac{\boldsymbol{\psi}}{2}\} \odot \boldsymbol{\epsilon}_s, \quad \boldsymbol{\epsilon}_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad (22)
$$

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 $\boldsymbol{\omega}$ and $\boldsymbol{\psi}$.

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

$$
\begin{aligned}
\Big\langle \ln p(\mathbf{X}^\lambda | \boldsymbol{\theta}_s^\lambda, \mathbf{z}^\lambda, c) \Big\rangle_{q(\mathbf{z}^\lambda)} = &\frac{N_c}{2} \left(2\pi_s^c + \ln |\boldsymbol{\Lambda}_s^c| - \boldsymbol{\mu}_s^{c\top} \boldsymbol{\Lambda}_s^c \boldsymbol{\mu}_s^c \right) \\
&+ \boldsymbol{\phi}_c^\top \boldsymbol{\Lambda}_s^c \boldsymbol{\mu}_s^c - \frac{\mathrm{tr}\{\boldsymbol{\Phi}_c \boldsymbol{\Lambda}_s^c\}}{2} + \mathrm{const}\,. \qquad (23)
\end{aligned}
$$

For each component, the mean $\boldsymbol{\mu}_s^c$, precision matrix $\boldsymbol{\Lambda}_s^c = (\boldsymbol{\Sigma}_s^c)^{-1}$ and mixing weight π_s^c are obtained from a sample $(\mathcal{M}_s, \boldsymbol{\theta}_s^\lambda)$ using equations 14-16. γ_{nc} are the Gaussian responsibilities for each time frame n computed in the *expectation* step, $N_c = \sum_n \gamma_{nc}$, $\boldsymbol{\phi}_c = \sum_n \gamma_{nc} \mathbf{x}_n$ and $\boldsymbol{\Phi}_c = \sum_n \gamma_{nc} \mathbf{x}_n \mathbf{x}_n^\top$ are the zeroth, first and second order sufficient statistics respectively for Gaussian component c. Note that (23) is a differentiable function of $\boldsymbol{\omega}$ and $\boldsymbol{\psi}$.

Overall, the training of our AUD system comprises two phases. First, we infer a set of variational posteriors $q_0(\{\mathbf{z}^\lambda\})$, $q_0(\{\boldsymbol{\theta}^\lambda\})$ and $q_0(\mathcal{M})$ on transcribed source languages where $\lambda \in \{1, \ldots, L\}$. At this phase, 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 $q_1(\mathbf{z}^t)$, $q_1(\boldsymbol{\theta}^t)$ using $q_0(\mathcal{M})$ to compute the *expectation* and *maximization* steps. Note that $q_0(\mathcal{M})$ is not updated during this phase, but transferred as is from the source languages.

Finally, the output of our AUD system is obtained from the Viterbi algorithm which is also modified to use the expectation of the log-likelihoods wrt $q_1(\boldsymbol{\theta}^t) q_0(\mathcal{M})$ instead of point estimates [14].

4. RELATED WORK

Our proposed model builds heavily on the generalized subspace model of [12]. Our novelty is that we introduce a hyper-subspace to allow unsupervised adaptation of the subspace itself. In particular, if the language embedding $\boldsymbol{\alpha}$ is set to $\mathbf{0}$, 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. Mboshi [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 `sa` 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 hyper-subspace: German, Spanish, French and Polish from Globalphone [22]; and Amharic [23], Swahili [24] and Wolof [25] from the ALFFA project [26]. For each of these, we use only a subset of 1500 utterances which corresponds to 1.5-3 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:

$$
\mathrm{NMI}(P, U) = 200 \times \frac{I(P; U)}{H(P) + H(U)} \%, \qquad (24)
$$

where $H(\cdot)$ is the Shannon entropy and $I(P; U)$ is the (unnormalized) mutual information. An NMI of 0 means that the 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 [27] to 100, and we find that number of discovered units tends to be fewer. We set the dimension of the H-SHMM unit embeddings $|\mathbf{e}^{\lambda,u}| = 100$. We set the dimension of the language embeddings $|\boldsymbol{\alpha}^{\lambda}| = 6$. We use 5 samples for the re-parametrization trick and train with Adam optimizer [28] with a learning rate of 5×10^{-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 subpace dimension to 100.

We also impemented 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 "$\{3\} \rightarrow \{2,3\}$" model (ResDAVEnet-VQ-I) with units extracted from layer-2 and the "$\{2\} \rightarrow \{2,3\}$" 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.

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	26.84
	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

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 $|\boldsymbol{\alpha}^{\lambda}| = 2$. 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 $\boldsymbol{\alpha}^{\lambda}$ 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 *https://www.fit.vutbr.cz/~iyusuf/hshmm.html*.

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.

[1] https://github.com/beer-asr/beer/tree/master/recipes/hshmm

[2] https://github.com/BUTSpeechFIT/vq-aud

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 Aäron 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] Alexei 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 Černockỳ, "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 bottle-neck feature learning from untranscribed speech," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 727–733.

[11] Thomas Glarner, 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, 2021, Chapter 2.

[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, Lukaš Burget, Jan Černockỳ, 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, Işın Demirşahin, Oddur Kjartansson, Clara Rivera, and Kọ́lá Túbọ̀sún, "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 Schlippe, "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 Lei Ba, "Adam: A method for stochastic gradient descent," in *International Conference on Learning Representations*, 2015.

[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.