Non-Parametric Bayesian Subspace Models for Acoustic Unit Discovery

Lucas Ondel, Bolaji Yusuf, Lukáš Burget, Murat Saraçlar

Non-Parametric Bayesian Subspace Models for Acoustic Unit Discovery

Lucas Ondel, *Member, IEEE*, Bolaji Yusuf, *Student Member, IEEE*, Lukáš Burget,*Member, IEEE* and Murat Saraçlar, *Member, IEEE*

***Abstract*—This work investigates subspace non-parametric models for the task of learning a set of acoustic units from unlabeled speech recordings. We constrain the base-measure of a Dirichlet-Process mixture with a phonetic subspace—estimated from other source languages—to build an *educated prior*, thereby forcing the learned acoustic units to resemble phones of known source languages. Two types of models are proposed: (i) the Subspace HMM (SHMM) which assumes that the phonetic subspace is the same for every language, (ii) the Hierarchical-Subspace HMM (H-SHMM) which relaxes this assumption and allows to have a language-specific subspace estimated on the unlabeled target data. These models are applied on 3 languages: English, Yoruba and Mboshi and they are compared with various competitive acoustic units discovery baselines. Experimental results show that both subspace models outperform other systems in terms of clustering quality and segmentation accuracy. Moreover, we observe that the H-SHMM provides results superior to the SHMM supporting the idea that language-specific priors are preferable to language-agnostic priors for acoustic unit discovery.**

I. Introduction

Building a speech recognition system requires a large collection of transcribed data. For instance, recent publications [1], [2], [3] report using tens of thousands hours of recordings paired with their corresponding textual transcription. Such amounts of transcribed data are available for only a handful of languages and stunt the development of speech technologies for many languages. While collecting audio data is relatively easy in our digital word, human-based transcriptions are expensive and too slow to keep pace with the daily production of multimedia content. A tremendous step would be made if one could automatically transcribe this data as it would drastically increase the amount of available resources to build speech technologies in many languages.

In parallel, there is a keen interest to understand how infants learn to recognize speech. Indeed, whereas speech recognition systems are built upon human-transcribed data, toddlers learn seamlessly to structure speech with very distant and noisy supervision. As a remarkable example of this learning capability, children born blind are perfectly able to learn to recognize speech even deprived of any supervision coming from the visual signal. In an attempt to explain this capability using the "reverse-engineering approach" [4], many models for automatic labeling of the data have been proposed by the machine learning community [5], [6], [7].

These two research ideas, while having very distinct objectives, share a common interest: to build a machine learning algorithm that automatically learns a discrete representation of the speech signal in an unsupervised fashion. For the former, this would allow automatic labeling of large collection of data, for the latter, it would serve as a simulation of how infants learn to process speech.

Current acoustic unit discovery (AUD) studies follow two major approaches:

- non-parametric Bayesian models [8], [9], [10], [11] which are usually infinite mixture of time series models such as Hidden Markov Model (HMM)
- neural-network-based models [12], [13], [14] using quantization layers with template vectors that represent the acoustic units' vocabulary.

Note that these approaches are not mutually exclusive and can be combined as was shown in [15], [16].

This work focuses on subspace model techniques applied to non-parametric Bayesian models on the task of discovering a set of pseudo-phones (called acoustic units) from unlabeled audio recordings[1].

Preliminary results on subspace models for AUD have been published in [17], [18] giving rise to two models: the Subspace HMM (SHMM) and Hierarchical-Subspace HMM (H-SHMM). In this paper, we provide a more comprehensive theoretical coverage of these models, their relationship with the Dirichlet process and a complete inference scheme. In addition, we conduct an in-depth performance analysis of the subspace models as well as a comparison with state-of-the-art baselines. Note that we assume readers' familiarity with the Dirichlet process and variational inference [19].

The rest of the paper is organized as follows: in Section II, we introduce a formal Bayesian formulation of the AUD problem, as well as the Dirichlet process HMM model upon which the subspace models are built; in Section III, we introduce the subspace models as HMM-based AUD models with specific prior forcing the model's parameters to dwell in a "phonetic" subspace; in Section IV, we detail the inference and experimental results are presented in Section V.

Lucas Ondel and Bolaji Yusuf contributed equally to this work.

Lucas Ondel is with LISN, CNRS, Universite Paris-Saclay, Orsay, France (e-mail: lucas.ondel@lisn.upsaclay.fr)

Bolaji Yusuf and Lukáš Burget are with the Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia (e-mail: bolaji.yusuf@boun.edu.tr)

Bolaji Yusuf and Murat Saraçlar are with Boğaziçi University, Istanbul, Turkey

[1]All the models and algorithms presented in this work are implemented at: https://github.com/beer-asr/beer

II. Bayesian AUD

A. Probabilistic interpretation of AUD

We first introduce a probabilistic formulation of the AUD task which will motivate the Bayesian approach of this work. Given a speech utterance of N observations $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$, the AUD task amounts to find:

- A collection of U vectors $\mathbf{H} = \{\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_U\}$ best describing the observations, where each $\boldsymbol{\eta}$ represents the parameters of a distribution of observations for a specific sound. These sounds are called *acoustic units* as they represent the basic elements of speech.
- The sequence of indices $\mathbf{u} = u_1, \dots, u_L,\ L < N$ where $u_i \in \{1, \dots, U\}$ is the index of an acoustic unit.

Rather than the *maximum a posteriori* estimate $\mathbf{H}^*, \mathbf{u}^* = \arg\max_{\mathbf{H},\mathbf{u}} p(\mathbf{H}, \mathbf{u}|\mathbf{X})$, we seek to obtain the posterior *distribution* over the embeddings $\mathbf{H}$ and label sequence $\mathbf{u}$:

$$p(\mathbf{H}, \mathbf{u}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{H}, \mathbf{u})p(\mathbf{u}|\mathbf{H})p(\mathbf{H})}{\int_{\mathbf{H}} \sum_{\mathbf{u}} p(\mathbf{X}|\mathbf{H}, \mathbf{u})p(\mathbf{u}|\mathbf{H})p(\mathbf{H}) \mathrm{d}\mathbf{H}}. \quad (1)$$

This allows us to have an estimate of uncertainty. The Bayesian statement of AUD in (1) is analogous to the statistical formulation of ASR [20] with, however, one major difference: in the case of AUD, the inventory of units is unknown and needs to be inferred from the data along with the acoustic unit embeddings $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots$.

It is important to understand the different roles played by the three factors in the numerator of (1):

- $p(\mathbf{X}|\mathbf{H}, \mathbf{u})$: the likelihood of the observations given the collection of acoustic unit parameters and the sequence of labels. This term, referred to as the acoustic model, tells how plausible the sequence of observations is given the sequence of acoustic units.
- $p(\mathbf{u}|\mathbf{H})$: the prior over the label sequence is the language model over the acoustic units' labels. It models the phonotactic constraints of the discovered units.
- $p(\mathbf{H})$: the prior over the collection of embeddings, this term is essential as it defines, before observing data, what are the potential acoustic unit candidates. This term will be the focus of this work.

Under the Bayesian AUD interpretation, the collection of vectors $\mathbf{H}$ bears a particular meaning: they are the parameters of the acoustic model.

B. Non-parametric AUD

In practice, the size of the acoustic units inventory U is not known and we have to pick an appropriate value. This is not an easy task since every language has a unique set of phones and we would like to infer the value of U in light of the data. We achieve this behavior by letting $U \to \infty$ and adding a distribution $\mathcal{P}$ over the parameters of $p(\mathbf{u}, \mathbf{H})$. This approach, referred to as *non-parametric Bayesian* [21], lets the model learn its own complexity from the data. Following, [8], [11] we set $\mathcal{P}$ to be a Dirichlet Process $\mathcal{DP}(\gamma, G_0)$ [22] with *concentration* γ and *base measure* $G_0(\boldsymbol{\eta})$ over the acoustic unit embeddings.

To allow efficient variational inference with the Dirichlet Process [19], we use the stick-breaking process view of the Dirichlet process, expressed as a generative process:

$$\gamma \sim \mathcal{G}(a_0, b_0) \quad (2)$$

$$v_i \sim \mathcal{B}(1, \gamma), \quad i = \{1, 2, \dots\} \quad (3)$$

$$\boldsymbol{\eta}_i \sim G_0(\boldsymbol{\eta}), \quad i = \{1, 2, \dots\} \quad (4)$$

$$\psi_i = v_i \prod_{j=1}^{i-1} (1 - v_j) \quad (5)$$

$$G(\boldsymbol{\eta}) = \sum_{i=1}^{\infty} \psi_i \delta_{\boldsymbol{\eta}_i}(\boldsymbol{\eta}), \quad (6)$$

where $\mathcal{B}$ is a Beta distribution, $\mathcal{G}$ is a Gamma distribution and $\delta_{\boldsymbol{\eta}_i}$ is the Dirac delta function centered at $\boldsymbol{\eta}_i$. Note that we have added a Gamma prior over the concentration of the Dirichlet Process so that we learn the value of the concentration parameter directly from the data.

Finally, we use the base measure $G_0(\boldsymbol{\eta})$ and the constructed distribution $G(\boldsymbol{\eta})$ to build the prior $p(\mathbf{u}|\mathbf{H})p(\mathbf{H})$ in the following way:

$$p(\mathbf{u}|\mathbf{H})p(\mathbf{H}) = \underbrace{\left[\prod_{n=1}^{L} \underbrace{G(\boldsymbol{\eta}_{u_n})}_{p(u_n|\mathbf{H})}\right]}_{p(\mathbf{u}|\mathbf{H})} \underbrace{\left[\prod_{k=1}^{\infty} G_0(\boldsymbol{\eta}_k)\right]}_{p(\mathbf{H})}. \quad (7)$$

One more time, we highlight the different roles played by the two terms in (7). $G_0(\boldsymbol{\eta})$ is a continuous density over the embedding space: it defines which embeddings are likely to be selected as acoustic units. $G(\boldsymbol{\eta}_{u_n})$, on the other hand, is a discrete distribution over an infinite set of atoms (i.e. the samples from the base measure) and it defines how frequently a unit occurs in speech. In other words, $G(\boldsymbol{\eta}_{u_n})$ is a (unigram) language model of the units.

C. Acoustic Model

We now turn to the definition of the acoustic model $p(\mathbf{X}|\mathbf{H}, \mathbf{u})$. We denote by $\mathbf{X}_{u_l}$ the sub-sequence of the observed data that belongs to the acoustic unit with index u_l such that $\mathbf{X} = (\mathbf{X}_{u_1}, \dots, \mathbf{X}_{u_L})$. We assume the following factorization of the likelihood:

$$p(\mathbf{X}|\mathbf{H}, \mathbf{u}) = \prod_{l=1}^{L} p(\mathbf{X}_{u_l}|\mathbf{H}, u_l) = \prod_{l=1}^{L} p(\mathbf{X}_{u_l}|\boldsymbol{\eta}_{u_l}). \quad (8)$$

Following [8], we model the likelihood $p(\mathbf{X}_{u_l}|\boldsymbol{\eta}_{u_l})$ by a left-to-right HMM with S hidden states where each state has a GMM emission density with C components:

$$\begin{aligned} p(\mathbf{X}_{u_l}|\boldsymbol{\eta}_{u_l}) &= \sum_{\mathbf{s}_{u_l}} \sum_{\mathbf{c}_{u_l}} p(\mathbf{X}_{u_l}, \mathbf{c}_{u_l}, \mathbf{s}_{u_l}|\boldsymbol{\eta}_{u_l}) \quad (9) \\ &= \sum_{\mathbf{s}_{u_l}} \sum_{\mathbf{c}_{u_l}} \prod_{n_l=1}^{N_l} p(\mathbf{x}_{n_l}^{u_l}, c_{n_l}^{u_l}|s_{n_l}^{u_l}, \boldsymbol{\eta}_{u_l}) p(s_{n_l}^{u_l}|s_{n_l-1}^{u_l}), \end{aligned}$$

where:

- $\mathbf{s}_{u_l} = s_1^{u_l}, \dots, s_{N_l}^{u_l}$ is the sequence of indices of the HMM states for acoustic unit u_l,

- $\mathbf{c}_{u_l} = c_1^{u_l}, \dots, c_{N_l}^{u_l}$ is the sequence of indices of the mixture components for the acoustic unit u_l,
- $p(s_1^{u_l}|s_0^{u_l}) = p(s_1^{u_l})$ is the probability of the initial state,
- N_l is the length of the sequence of observations $\mathbf{X}_{u_l}$.

Finally, the acoustic unit embedding $\boldsymbol{\eta}_{u_l}$ encodes the parameters of the HMM model $\{\boldsymbol{\pi}_{u_l}^s\}, \{\boldsymbol{\mu}_{u_l}^{s,c}\}, \{\boldsymbol{\Sigma}_{u_l}^{s,c}\}, \forall\ s \in \{1, \dots, S\},\ c \in \{1, \dots, C\}$, where:

- $\boldsymbol{\pi}_{u_l}^s$ are the mixing weights of the GMM associated with the sth state of the HMM of the acoustic unit u_l,
- $\boldsymbol{\mu}_{u_l}^{s,c}$, $\boldsymbol{\Sigma}_{u_l}^{s,c}$ are the mean and the covariance matrix of the cth Normal component of the GMM associated with the sth state of the HMM of the acoustic unit u_l.

We have not included any parameters for the within-unit transition probabilities $p(s_{n_l}^{u_l}|s_{n_l-1}^{u_l})$ as it has been empirically observed that they play no significant role when modeling speech [23]. Therefore, we assume the transition probabilities are fixed so that there is a 0.75 probability of remaining in the same state and a 0.25 probability of exiting to the next state of the HMM. Note that this only concerns the transition probabilities within the unit; transition probabilities between units are governed by the distribution sampled from the Dirichlet process.

D. Acoustic Unit embeddings

We detail now the relation between the embedding $\boldsymbol{\eta}_{u_l}$ and the HMM parameters. To keep the notation uncluttered, we drop the subscripts and superscripts u_l and n_l, therefore, we write $\mathbf{x}, \boldsymbol{\eta}, \dots$ instead of $\mathbf{x}_{n_l}^{u_l}, \boldsymbol{\eta}_{u_l}, \dots$. Observe that the distribution of $p(\mathbf{x}, c|s, \boldsymbol{\eta})$ is a product of a Normal and a Categorical distribution and, moreover, each of them is a member of the exponential family of distributions [24]. Consequently, we have:

$$p(\mathbf{x}, c|s, \boldsymbol{\eta}) = p(\mathbf{x}|\boldsymbol{\mu}^{s,c}, \boldsymbol{\Sigma}^{s,c}) p(c|\boldsymbol{\pi}^s) \tag{10}$$

$$\begin{aligned} p(c|\boldsymbol{\pi}^s) &= p(c|\boldsymbol{\omega}^s) \\ &= \exp\{\boldsymbol{\omega}^{s\top} T(c) - A(\boldsymbol{\omega}^s)\} \end{aligned} \tag{11}$$

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\mu}^{s,c}, \boldsymbol{\Sigma}^{s,c}) &= p(\mathbf{x}|\boldsymbol{\theta}^{s,c}) \\ &= \exp\{\boldsymbol{\theta}^{s,c\top} T(\mathbf{x}) - A(\boldsymbol{\theta}^{s,c})\} \end{aligned} \tag{12}$$

where $\boldsymbol{\omega}^s$, $T(c)$ and $A(\boldsymbol{\omega}^s)$ are the natural parameters, the sufficient statistics and the log-normalizer of the Categorical distribution over the GMM components of the sth HMM state. Similarly, $\boldsymbol{\theta}^{s,c}$, $T(\mathbf{x})$ and $A(\boldsymbol{\theta}^{s,c})$ are the natural parameters, the sufficient statistics and the log-normalizer of the Normal distribution associated with state s and the mixture component c. For both distributions, the natural parameters and the sufficient statistics can be derived from their respective definitions [24]:

$$\boldsymbol{\omega}^s = \begin{bmatrix} \ln\left(\frac{\pi_1^s}{\pi_C^s}\right) \\ \vdots \\ \ln\left(\frac{\pi_{C-1}^s}{\pi_C^s}\right) \end{bmatrix} \qquad T(c) = \begin{bmatrix} \mathbb{1}[c=1] \\ \dots \\ \mathbb{1}[c=C-1] \end{bmatrix} \tag{13}$$

$$\boldsymbol{\theta}^{s,c} = \begin{bmatrix} (\boldsymbol{\Sigma}^{s,c})^{-1}\boldsymbol{\mu}^{s,c} \\ -\frac{1}{2}\operatorname{vec}((\boldsymbol{\Sigma}^{s,c})^{-1}) \end{bmatrix} \quad T(\mathbf{x}) = \begin{bmatrix} \mathbf{x} \\ \operatorname{vec}(\mathbf{x}\mathbf{x}^\top) \end{bmatrix}, \tag{14}$$

where "vec" is the vectorization operation and $\mathbb{1}$ is the indicator function. Note that because $\boldsymbol{\pi}$ is constrained such that $\sum_{c=1}^{C} \pi_c = 1$, the natural parameter $\boldsymbol{\omega}$ is a $(C-1)$-dimensional vector whereas $\boldsymbol{\pi}$ is a C-dimensional vector. Finally, the log-normalizers $A(\boldsymbol{\omega}^s)$ and $A(\boldsymbol{\theta}^{s,c})$ are given by:

$$A(\boldsymbol{\omega}^s) = \ln\left(1 + \sum_{k=1}^{K-1} \exp\{\omega_k^s\}\right) \tag{15}$$

$$A(\boldsymbol{\theta}^{s,c}) = -\frac{1}{4}\boldsymbol{\theta}_1^{s,c\top} \operatorname{mat}\left(\boldsymbol{\theta}_2^{s,c}\right)^{-1}\boldsymbol{\theta}_1^{s,c} - \frac{1}{2}\ln|-2\operatorname{mat}\left(\boldsymbol{\theta}_2^{s,c}\right)|, \tag{16}$$

where ω_k^s is the kth element of $\boldsymbol{\omega}_k$, mat is the inverse of the vec operation, $\boldsymbol{\theta}_1^{s,c} = (\boldsymbol{\Sigma}^{s,c})^{-1}\boldsymbol{\mu}^{s,c}$, and $\boldsymbol{\theta}_2^{s,c} = -\frac{1}{2}\operatorname{vec}((\boldsymbol{\Sigma}^{s,c})^{-1})$.

Finally, we define the embedding $\boldsymbol{\eta}$ to be a "super-vector" obtained by concatenating the natural parameters of the Normal and Categorical distributions of all S states of the HMM modeling of an acoustic unit. It has the following layout:

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}^1 \\ \vdots \\ \boldsymbol{\eta}^s \\ \vdots \\ \boldsymbol{\eta}^S \end{bmatrix} = \begin{bmatrix} \boldsymbol{\omega}^s \\ \boldsymbol{\theta}^{s,1} \\ \vdots \\ \boldsymbol{\theta}^{s,C} \end{bmatrix}, \tag{17}$$

where $\boldsymbol{\eta}^s$ is the concatenation of the natural parameters of the Normal and Categorical distributions for the sth state of the HMM of an acoustic unit.

E. Joint Distribution

To conclude the description of the model, we present the complete joint distribution. For simplicity, we introduce the variable $z_n = (u_l, s_{n_l})$ which encodes an acoustic unit index u_l and a particular HMM state s_{n_l}. Notice that the time index n in z_n—which combines both the relative time indices l and n_l—is absolute with respect to the sequence of observations, i.e. $n \in \{1, \dots, N\}$. Similarly, c_n represents the index of a GMM component at time n. We write $\boldsymbol{\eta}_{z_n} = \boldsymbol{\eta}_{u_l}^{s_{n_l}}$ which corresponds to the natural parameters of the s_{n_l}th HMM state of the acoustic unit with index u_l. With this notation, the joint distribution is given by:

$$p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma) = p(\mathbf{H})p(\gamma)p(\mathbf{v}|\gamma)p(\mathbf{X}, \mathbf{c}, \mathbf{z}|\mathbf{H}, \mathbf{v}) \tag{18}$$

$$p(\mathbf{X}, \mathbf{c}, \mathbf{z}|\mathbf{H}, \mathbf{v}) = p(\mathbf{z}|\mathbf{v})p(\mathbf{X}, \mathbf{c}|\mathbf{z}, \mathbf{H}) \tag{19}$$

$$= \prod_{n=1}^{N} p(z_n|z_{n-1}, \mathbf{v})p(\mathbf{x}_n, c_n|\boldsymbol{\eta}_{z_n}), \tag{20}$$

where $\mathbf{v} = \{v_1, v_2, \dots\}$ is the set of stick breaking weights. This is the likelihood of a typical "phone-loop" HMM [25] where $\mathbf{z}$ is the sequence of hidden states. As explained in [26], this interpretation of the model allows an efficient dynamic programming algorithm to evaluate all possible sequences of

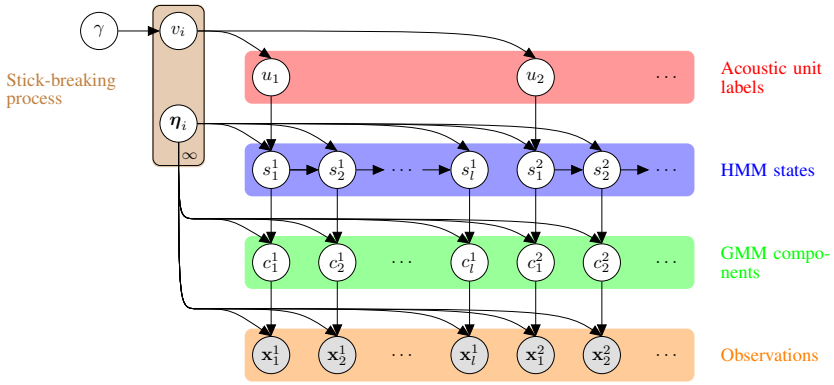


Fig. 1: Bayesian network of the non-parametric acoustic unit clustering model. The model is composed of 3 layers of hidden variables: (i) the GMM components which acts as a quantization layer, (ii) the HMM states layer which captures the temporal dynamic of the data, and (iii) the acoustic units' layer which encodes the phonetic information.

units (see Appendix A-A). The per-state emission likelihood in (20) is given by:

$$p(\mathbf{x}_n, c_n|\boldsymbol{\eta}_{z_n}) = p(\mathbf{x}_n|\boldsymbol{\theta}_{z_n}^{c_n})p(c_n|\boldsymbol{\omega}_{z_n}), \tag{21}$$

where $\boldsymbol{\theta}_{z_n}^{c_n}$ is the vector of natural parameters of the c_nth mixture components of the s_n state, and the two factors on the right hand side are defined in (11) and (12). The transition probability combines the within and across units' transition in the following way:

$$\prod_{n=1}^{N} p(z_n|z_{n-1}, \mathbf{v}) = \prod_{l=1}^{L} p(u_l|\mathbf{v}) \prod_{n=1}^{N_l} p(s_n^{u_l}|s_{n-1}^{u_l}), \tag{22}$$

where the transition probability within a unit's HMM is fixed: $p(s_n|s_{n-1}) = \text{const}$ and the probability of the unit index $p(u_l|\mathbf{v})$ is given by the stick-breaking process as defined in Section II-B:

$$p(u_l|\mathbf{v}) = v_{u_l} \prod_{i=1}^{u_l-1} (1 - v_i). \tag{23}$$

The priors over the stick-breaking process parameters $\mathbf{v}$ and the prior over the concentration parameter γ are given by:

$$p(\mathbf{v}|\gamma) = \prod_{i=1}^{\infty} p(v_i|\gamma) \tag{24}$$

$$p(v_i|\gamma) = \mathcal{B}(1, \gamma) \tag{25}$$

$$p(\gamma) = \mathcal{G}(a_0, b_0), \tag{26}$$

the prior over embeddings $\mathbf{H}$ is defined from the base measure:

$$p(\mathbf{H}) = \prod_{u=1}^{\infty} G_0(\boldsymbol{\eta}_u), \tag{27}$$

and its exact construction will be addressed in the next section.

Finally, Figure 1 gives a graphical perspective of the complete model. It is composed of 3 layers of hidden variables: (i) the GMM components layer which acts as a quantization layer, i.e. it transduces a sequence of continuous features into a sequence of discrete symbols, (ii) the HMM states layer which captures the temporal dynamic of the data, and (iii) the acoustic units' layer which encodes phonetic information.

III. Prior over the embeddings

We have formulated a probabilistic interpretation of the AUD problem. From this, we have seen that 3 terms emerge: (i) the likelihood defining the acoustic model, (ii) the language model and (iii) the prior over the embeddings. We have detailed the two first terms in the previous section and, now, we draw our attention to the last term: the prior over the embeddings $G_0(\boldsymbol{\eta})$.

A. Conjugate prior

Early Bayesian AUD models [8], [11], [16] set $G_0(\boldsymbol{\eta})$ to be the conjugate prior of the conditional HMM likelihood:

$$G_0(\boldsymbol{\eta}) = \prod_{s=1}^{S} p(\boldsymbol{\omega}^s) \prod_{c=1}^{C} p(\boldsymbol{\theta}^{s,c}) \tag{28}$$

$$p(\boldsymbol{\omega}^s) = \exp\left\{\boldsymbol{\xi}_0^\top \begin{bmatrix} \boldsymbol{\omega}^s \\ -A(\boldsymbol{\omega}^s) \end{bmatrix} - A(\boldsymbol{\xi}_0)\right\} \tag{29}$$

$$p(\boldsymbol{\theta}^{s,c}) = \exp\left\{\boldsymbol{\vartheta}_0^\top \begin{bmatrix} \boldsymbol{\theta}^{s,c} \\ -A(\boldsymbol{\theta}^{s,c}) \end{bmatrix} - A(\boldsymbol{\vartheta}_0)\right\}, \tag{30}$$

where $\boldsymbol{\xi}_0$ and $\boldsymbol{\vartheta}_0$ are the natural parameters of the conjugate prior of the states' emission density. The conjugacy implies that the prior $p(\boldsymbol{\omega}^i)$ over the natural form of the mixing weights is a Dirichlet distribution and the prior $p(\boldsymbol{\theta}^{i,j})$ over the natural form of the mean and the precision matrix (inverse of the covariance matrix) is a Normal-Wishart distribution. This choice is convenient since it greatly simplifies the inference; it is, however, difficult to control precisely which type of sounds the prior will emphasize. In previous works, $p(\boldsymbol{\omega}^i)$ and $p(\boldsymbol{\theta}^{i,j})$ were set to be *vague* priors (i.e. priors that play a mininal role in the estimation of the posterior distribution) leading the AUD model to consider, say, the phone /aw/ and the sound of a slamming door as equally good candidate acoustic units.

B. Phonetic Subspace

Vague priors are easy to define but they fail to provide a reasonable selection of "good" candidates. Recent works [17], [18] have proposed to remedy this shortcoming by introducing

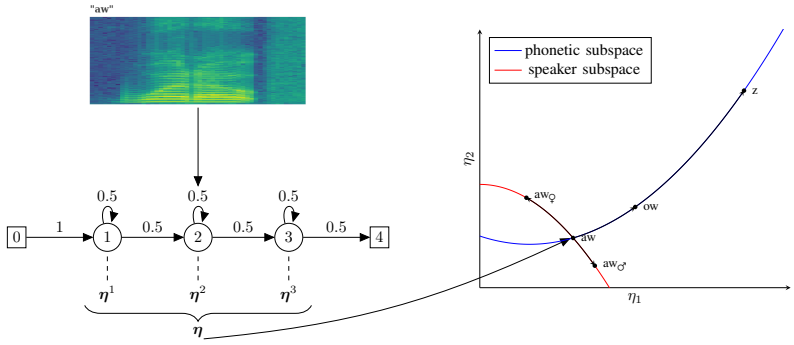


Fig. 2: Illustration of the concept of *phonetic subspace*: each phone is represented as a vector $\boldsymbol{\eta}$ encoding the parameters of a probabilistic model (an HMM in this example). Ideally, moving away from the subspace only changes the characteristics of the phone (speaker, channel, loudness, ...) but not the phone itself. For illustration purposes, the red line represents one of such factor of variability: the speaker subspace. Finally, not that in this example, the parameter space has only 2 dimensions (η_1 and η_2) but in practice it will have several thousands of dimensions.

subspace-based priors which act as *informative* (or *educated*) priors over the space of acoustic unit embeddings. These works rely upon the concept of phonetic subspace which we'll illustrate with an example.

Let's consider that we fit an HMM to a set of recordings of the English phone /aw/ which gives the embedding vector $\boldsymbol{\eta}_{\text{aw}}$. Moving $\boldsymbol{\eta}_{\text{aw}}$ in any direction in the embedding space will affect the parameters of the HMM and, consequently, the phone it represents. As an example, a particular displacement may lead to changing the phone /aw/ to /ow/. Moving the $\boldsymbol{\eta}_{\text{aw}}$ even further will change the original /aw/ phone more profoundly and yield, say, the consonant /z/. In this thought experiment, we have assumed that there is a continuum between all phones or, expressed in another way, that we can smoothly transition from one phone to another. Generally, we can envision all the phones of a language as points on a low-dimensional manifold which represents this continuum. This manifold is depicted by the blue line in Figure 2 and it is what we call the *phonetic subspace*. Importantly, this concept of phonetic subspace is independent of the choice of the phone model: GMM, HMM, Linear Dynamical Model, etc. However, the type of model used will influence how well the continuity between phones is represented.

C. SHMM

The Subspace HMM (SHMM) [17] defines the base measure $G_0(\boldsymbol{\eta})$ as the probability distribution induced by the following sampling process:

$$\mathbf{W} \sim p(\mathbf{W}) = \prod_{r,c} p(W_{r,c}) \tag{31}$$

$$p(W_{r,c}) = \mathcal{N}(0,1) \tag{32}$$

$$\mathbf{b} \sim p(\mathbf{b}) = \mathcal{N}(\mathbf{0},\mathbf{I}) \tag{33}$$

$$\mathbf{e}_u \sim p(\mathbf{e}_u) = \mathcal{N}(\mathbf{0},\mathbf{I}) \tag{34}$$

$$\boldsymbol{\eta}_u = f(\mathbf{W}\mathbf{e}_u + \mathbf{b}), \tag{35}$$

where $\mathbf{e}_u$ is a Q-dimensional embedding of the acoustic unit u on the subspace, the weights matrix $\mathbf{W}$ and the bias vector $\mathbf{b}$ are the parameters of the phonetic subspace, and $f(\cdot)$ is a function that takes a real vector and projects it to the HMM parameter space. In this work, it is set such that:

$$\pi^s_{u,j} = \frac{\exp\{\mathbf{W}^s_\pi \cdot \mathbf{e}_u + \mathbf{b}^s_\pi\}_j}{1 + \sum_{k=1}^{K-1} \exp\{\mathbf{W}^s_\pi \cdot \mathbf{e}_u + \mathbf{b}^s_\pi\}_k} \tag{36}$$

$$\boldsymbol{\Sigma}^{s,c}_u = \text{diag}(\exp\{\mathbf{W}^{s,c}_\Sigma \cdot \mathbf{e}_u + \mathbf{b}^{s,c}_\Sigma\}) \tag{37}$$

$$\boldsymbol{\mu}^{s,c}_u = \boldsymbol{\Sigma}^u_{s,c} \cdot \left(\mathbf{W}^{s,c}_\mu \cdot \mathbf{e}_u + \mathbf{b}^{s,c}_\mu\right), \tag{38}$$

where $\text{diag}(\cdot)$ returns a diagonal matrix from an input vector, $\exp$ is the element-wise exponential function and $\exp\{...\}_j$ is the jth element of the resulting vector. $\mathbf{W}^s_\pi$ is the subset of rows of matrix $\mathbf{W}$ assigned to the mixing weights $\boldsymbol{\pi}^s$ of the sth HMM state. Matrices $\mathbf{W}^{s,c}_\mu$ and $\mathbf{W}^{s,c}_\Sigma$ are similarly defined for the mean and covariance matrix of the cth Gaussian component and sth HMM state. Note that (36) holds for $j \in \{1, ..., K-1\}$ and $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$.

Importantly, the construction of an acoustic unit embedding $\boldsymbol{\eta}_u$ relies upon a phonetic subspace parameterized by $\mathbf{W}$ and $\mathbf{b}$. Since these parameters are unknown in practice, they need to be inferred prior to utilizing them for AUD. This issue will be addressed in Section IV-B.

D. H-SHMM

The SHMM introduced in the above section has made the implicit assumption that the phonetic subspace is universal, i.e. it is the same for all the languages. [18] argues that this assumption is unrealistic and proposes to have a language-specific base measure $G^\lambda_0(\boldsymbol{\eta})$ defined by the following generative process:

$$\mathcal{M} = \{\mathbf{M}_0, \ldots, \mathbf{M}_K, \mathbf{m}_0, \ldots, \mathbf{m}_K\} \tag{39}$$

$$\boldsymbol{\alpha}^\lambda \sim \mathcal{N}(\mathbf{0},\mathbf{I}) \tag{40}$$

$$\mathbf{M}_i \sim p(\mathbf{M}_i) = \prod_{r,c} p(M_{i,r,c}), \quad p(M_{i,r,c}) = \mathcal{N}(0,1) \tag{41}$$

$$\mathbf{m}_i \sim p(\mathbf{m}_i) = \mathcal{N}(\mathbf{0},\mathbf{I}) \tag{42}$$

$$\mathbf{W}^\lambda = \mathbf{M}_0 + \sum_{k=1}^{K} \alpha^\lambda_k \mathbf{M}_k, \quad \mathbf{b}^\lambda = \mathbf{m}_0 + \sum_{k=1}^{K} \alpha^\lambda_k \mathbf{m}_k \tag{43}$$

$$\mathbf{e}_u \sim \mathcal{N}(\mathbf{0},\mathbf{I}), \quad \boldsymbol{\eta}_u = f(\mathbf{W}^\lambda \cdot \mathbf{e}_u + \mathbf{b}^\lambda). \tag{44}$$

This generative process incorporates a G-dimensional language embedding $\boldsymbol{\alpha}^\lambda$ which is used to built a language specific phonetic subspace by a linear combination of bases $\{\mathbf{M}_i\}$ and $\{\mathbf{m}_i\}$. These bases define a hyper-subspace of languages, as depicted in Figure 3. Because of the hierarchical nature of the generative process, the resulting model is termed the Hierarchical Subspace HMM (H-SHMM).

The bases $\{\mathbf{M}_i\}$ and $\{\mathbf{m}_i\}$ are shared across languages and act as "template" phonetic subspaces. In this view, each language specific phonetic subspace is a weighted combination of these generic subspaces. Similar to the SHMM subspace parameters, these parameters are unknown in practice and need to be estimated prior the AUD task.

IV. Inference

We now turn to the problem of inference for the SHMM and the H-SHMM. Since these models include many parameters, the derivation of the update equations is long and tedious. Therefore, we have opted to only give a general overview of the training in the main text with more technical details left to Appendix A.

A. Variational Bayes Inference

As discussed in Section II-A, from a Bayesian perspective, the AUD task amounts to finding the *a posteriori* distribution:

$$p(\mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma | \mathbf{X}) = \frac{p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma)}{p(\mathbf{X})}. \tag{45}$$

Note that $\mathbf{H}$ refers to the acoustic unit parameters, incorporating the subspace parameters as well as the low-dimensional embeddings. Since the denominator $p(\mathbf{X}) = \int_\cdot p(\mathbf{X}|\cdot)p(\cdot)\mathrm{d}\cdot$ is intractable, we resort to the Variational Bayes framework [27] to find an approximate posterior $q(\mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma)$. This entails maximizing the following lower-bound:

$$\ln p(\mathbf{X}) \geq \left\langle \ln \frac{p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma)}{q(\mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma)} \right\rangle_{q(\mathbf{c},\mathbf{z},\mathbf{H},\mathbf{v},\gamma)} \triangleq \mathcal{L}, \tag{46}$$

where we write: $\langle f(x) \rangle_{q(x)} = \int_x f(x) q(x) \mathrm{d}x$.

To be able to maximize (46), we use the following *structured mean-field* factorization:

$$q(\cdot) = q(\mathbf{c}|\mathbf{z})q(\mathbf{z}) \left[\prod_{i=1}^{\infty} q(\boldsymbol{\eta}_i) \right] \left[\prod_{i=1}^{\infty} q(v_i) \right] q(\gamma) \tag{47}$$

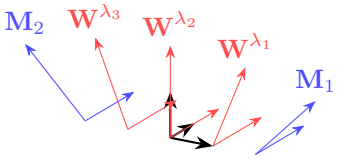


Fig. 3: Illustration of a hierarchical subspace model. For each language λ, acoustic unit embeddings (encoding the parameters of a probabilistic model) are assumed to live in a language-specific subspace of the total parameter space spanned by $\mathbf{W}^\lambda$. This subspace is given by a weighted sum of matrix bases $\mathbf{M}_1, \mathbf{M}_2, \ldots$ (shared across languages) and language-specific weights $\boldsymbol{\alpha}^\lambda$: $\mathbf{W}^\lambda = \alpha_1^\lambda \mathbf{M}_1 + \alpha_2^\lambda \mathbf{M}_2 + \ldots$.

Algorithm 1 Training of phone-loop model for acoustic unit discovery. Detailed coverage of the update equations can be found in Appendix A

function VB_MSTEP($\mathbf{X}, q^*(\mathbf{c}|\mathbf{z}), q^*(\mathbf{z}), q^*(\gamma)$)
  ▷ No closed-form solution, stochastic optimization described in Appendix A-B
  $q^*(\mathbf{H}) \leftarrow \arg\max_{q(\mathbf{H})} \mathcal{L}$
  ▷ Update defined in (68)
  $q^*(\mathbf{v}) \leftarrow \arg\max_{q(\mathbf{v})} \mathcal{L}$
  ▷ Update defined in (71)
  $q^*(\gamma) \leftarrow \arg\max_{q(\gamma)} \mathcal{L}$
  return $q^*(\mathbf{H}), q^*(\mathbf{v}), q^*(\gamma)$

function VB_ESTEP($\mathbf{X}, q(\mathbf{H}), q(\mathbf{v})$)
  ▷ Update defined in (51)
  $q^*(\mathbf{c}|\mathbf{z}) \leftarrow \arg\max_{q(\mathbf{c}|\mathbf{z})} \mathcal{L}$
  ▷ Update defined in (56)
  $q^*(\mathbf{z}) \leftarrow \arg\max_{q(\mathbf{z})} \mathcal{L}$
  return $q^*(\mathbf{c}|\mathbf{z}), q^*(\mathbf{z})$

procedure TRAIN($\mathbf{X}, E$)
  ▷ E: number of epochs (i.e. 1 epoch = E-step + M-step)
  ▷ initialize the variational posteriors (Appendix A-E)
  $q^*(\mathbf{v}), q^*(\gamma), q^*(\mathbf{H}) \leftarrow \ldots$
  for $e \leftarrow 1$ to E **do**
    $q^*(\mathbf{c}|\mathbf{z}), q^*(\mathbf{z}) \leftarrow$ VB_ESTEP($\mathbf{X}, q^*(\mathbf{H}), q^*(\mathbf{v})$)
    $q^*(\mathbf{H}), q^*(\mathbf{v}), q^*(\gamma) \leftarrow$
      VB_MSTEP($\mathbf{X}, q^*(\mathbf{c}|\mathbf{z}), q^*(\mathbf{z}), q^*(\gamma)$)

leading to an optimization algorithm analogous to the Expectation-Maximization (EM) algorithm [28] where we alternately estimate $q(\mathbf{c}|\mathbf{z})$ and $q(\mathbf{z})$ (E-step) and $q(\mathbf{H})$, $q(\mathbf{v})$ and $q(\gamma)$ (M-step). The complete training procedure is summarized in Algorithm 1.

B. Learning from other languages

The idea behind the SHMM is to supply prior information via the subspace parameters $\{\mathbf{W}, \mathbf{b}\}$ to the AUD system before observing the data. Hence, training the subspace parameters on the target language defeats the purpose of the model. In practice, we infer a set of variational posteriors $q_0(\{\mathbf{z}\})$, $q_0(\{\mathbf{e}_u\})$ and $q_0(\mathbf{W}, \mathbf{b})$ on phonetically transcribed source languages. This is the *supervised* phase of the training, where the system learns the notion of phone from transcribed data. At this stage, the phone-loop of the AUD model is replaced with a *forced alignment* graph since the variable $\mathbf{u}$ is observed in this case. Then, on the target language, we infer new variational posteriors $q_1(\{\mathbf{z}\})$, $q_1(\{\mathbf{e}_u\})$ using $q_0(\mathbf{W}, \mathbf{b})$. Note that $q_0(\mathbf{W}, \mathbf{b})$ is not updated during this stage, but transferred from the source languages as is. The H-SHMM is trained with a similar procedure: first we estimate $q_0(\mathcal{M})$ on several languages and then we use this posterior to learn $q_1(\{\mathbf{e_u}\}, \boldsymbol{\alpha}^\lambda)$ (and the other variational posteriors) on the target language λ while keeping $q_0(\mathcal{M})$ fixed.

V. Results

In this section, we experimentally validate the benefits of subspace models on the AUD task. In Sections V-A and V-B, we describe the experimental setup and the evaluation metrics respectively. Then, we analyze the improvement brought by the SHMM and the H-SHMM compared to a Bayesian HMM AUD baseline in Sections V-D and V-E. Finally, we compare the SHMM and the H-SHMM models with two alternative approaches: cross-lingual phonetic decoders, i.e. phone recognizers each trained on a different language than the target one, and neural-networks with discretization layers.

To account for model stochasticity, we run our AUD systems 5 times and we report the mean and standard deviation of the results.

A. Data and features

We use the following languages to evaluate the performance of our models:

1) Mboshi [29]: 4.4 hours with 5130 utterances by 3 speakers.
2) Yoruba [30]: 4 hours with 3583 utterances by 36 speakers.

In keeping with the AUD problem definition, we assume that we lack any transcribed data at training time, and our test data constitute our training data. Specifically, we do not assume the existence of a separate, transcribed development set for the target language for hyper-parameter selection. Therefore, we train and test on the entirety of each corpus as we would have to do for a real target language.

In place of a language-specific development set, we use English (from TIMIT [31] excluding the `sa` utterances) as a development *language*. Any hyper-parameter selection is done by picking the model which maximizes the task metrics on this set, and we transfer the model directly to any new target languages. The use of English as a development language has the added advantage that it facilitates comparison with baselines that can only be constructed for English, e.g. because they require training data that is only available for English.

In addition to the target languages, we also need a set of source languages for training the subspace of the SHMM and the hyper-subspace of the H-SHMM. We use seven transcribed source languages: German, Spanish, French and Polish from Globalphone [32]; and Amharic [33], Swahili [34] and Wolof [35] from the ALFFA project [36]. For each of these, we use only a subset of 1500 utterances; the resulting durations are shown in Table I. The difference in durations is due to the varying length of utterances for each corpus.

Finally, each system is trained on 13-dimensional MFCC features (12 coefficients and the per-frame energy) along with their first and second order derivatives.

TABLE I: Amounts of data for each source languages, i.e. languages used to pre-train the (H-)SHMM.

Language	AM	FR	GE	PO	SP	SW	WO
Data size (hours)	2.73	3.84	2.71	3.45	2.72	1.43	1.81

TABLE II: AUD performance of the SHMM.

Language	System	NMI	F-score
English	HMM	35.42 ± 0.18	63.50 ± 0.81
	SHMM	**38.96** ± 0.07	**74.03** ± 0.49
	Topline-U	44.02 ± 0.20	78.67 ± 0.13
	Topline-S	45.24	74.66
Mboshi	HMM	37.14 ± 0.26	48.63 ± 0.91
	SHMM	**38.95** ± 0.60	**60.13** ± 0.43
	Topline-U	52.14 ± 0.37	77.01 ± 0.11
	Topline-S	55.52	78.05
Yoruba	HMM	36.20 ± 0.31	53.97 ± 0.22
	SHMM	**38.98** ± 0.15	**63.77** ± 0.39
	Topline-U	45.25 ± 0.22	74.10 ± 0.21
	Topline-S	48.71	71.73

TABLE III: Effect of the phonetic subspace at initialization.

Language	System	NMI	F-score
English	HMM	1.80 ± 0.03	0.20 ± 0.00
	SHMM	**18.50** ± 1.14	**54.06** ± 2.08
Mboshi	HMM	1.60 ± 0.08	0.05 ± 0.00
	SHMM	**21.04** ± 1.11	**36.19** ± 1.97
Yoruba	HMM	1.43 ± 0.05	0.46 ± 0.03
	SHMM	**22.84** ± 0.81	**44.31** ± 1.70

B. Metrics

We use F-score and normalize mutual information (NMI) as the metrics for evaluating AUD performance.

1) F-score measures phone segmentation performance. We get precision and recall rates by comparing phone boundaries detected by the system of interest to reference phone boundaries with a tolerance ± 20 milliseconds. We report the harmonic mean of precision and recall as the F-score.
2) Normalized mutual information measures phone clustering quality. We compute the NMI from a frame level alignment of discovered units U and actual reference phones P, resulting in a matrix containing the joint probabilities $p(U, P)$. From this, we compute NMI as:

$$\text{NMI}(P, U) = 200 \times \frac{I(P;U)}{H(P) + H(U)}\%, \qquad (48)$$

where $H(\cdot)$ is the Shannon entropy functional [37] and $I(P;U)$ is the mutual information [37]. Since $0 \leq I(P;U) \leq \min\big(H(P), H(U)\big)$, the NMI takes on values between 0 and 100. An NMI of 0 is obtained when $I(P;U) = 0$ and the discovered acoustic units are completely unrelated to the actual phones. An NMI of 100 is obtained when $I(P;U) = H(P) = H(U)$ which only occurs when discovered units have a one-to-one correspondence with the actual phones. Note that the $H(U)$ term in the denominator penalizes representations with too many units. Without it, we could artificially inflate the NMI by increasing the number of units.

C. Hyper-parameters

Unless stated otherwise, the hyper-parameters of the SHMM and H-SHMM are set as follows:

- each acoustic unit HMM has 3 states left-to-right topology with 4 Gaussians per state with diagonal covariance matrix.
- The truncation τ (Appendix A-D)—the upper bound on the number of units discovered—is set to 100
- the parameters of the concentration prior in (2) are set to $a_0 = 1$ and $b_0 = \frac{1}{\tau}$
- the dimension of the phonetic subspace Q is set to 100
- the dimension of the language embedding G is set to 5.
- the Normal-Wishart distributions of the non-informative conjugate prior of HMM system (Section III-A) are set as: $\mathbf{m}_0 = \hat{\boldsymbol{\mu}}$, $\beta_0 = 1$, $\mathbf{W}_0 = \mathrm{diag}(\hat{\boldsymbol{\Sigma}})$ and $\nu_0 = D + 1$, where $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$ are the empirical mean, covariance of the data. Similarly, the Dirichlet distributions are set to have concentration parameters equal to 1.
- for the HMM baseline, the variational posteriors are initialized to have the same parameters as the prior, and we break the symmetry of the mixture components by perturbing their means with Gaussian noise with standard deviation of 0.01.

D. SHMM

Our first experiment compares the effect of using an *educated* prior—as implemented by the SHMM—against the non-informative conjugate prior as described in Section III-A. We refer to the later model as the HMM-based AUD system or simply HMM. Our implementation of the HMM follows [11].

From Table II, we observe that the SHMM outperforms the HMM baseline in terms of clustering quality and segmentation accuracy. We also report the results of two oracle systems:

- Topline-U, the unsupervised topline, is an SHMM AUD system whose phonetic subspace is pre-trained on the target language using the reference transcription. The concomitant phone embeddings are discarded and new embeddings are inferred with the AUD procedure without any transcription. This "cheating" experiment shows us the best performance achievable if we could estimate the perfect phonetic subspace.
- Topline-S, the supervised topline, is an HMM phone-recognizer with a uniform phonotactic language model trained on the target data. This model reveals the best clustering results we could obtain by using an HMM to model each acoustic unit.

In terms of clustering quality (NMI metric), we observe that Topline-U is quite close to Topline-S. This highlights the soundness of using a phonetic subspace. However, it also shows that estimating the phonetic subspace from other languages is not optimal and leaves room for improvement.

The goal of an *educated prior* is to provide the system with some information *before observing the data*. In the context of the SHMM, this prior information is the phonetic subspace which encodes the notion of phone for the AUD system. To verify that the phonetic subspace brings relevant information *a priori*, we report in Table III the performance of the HMM and the SHMM AUD systems at initialization. For the SHMM, this means after the subspace has been pre-trained on the source languages; for the HMM, this means after random initialization of the variational posteriors). As expected, the SHMM has much better performance at initialization compared to the HMM system which has a vague prior.

TABLE IV: Adapting the phonetic subspace to the target language.

Language	System	NMI	F-score
English	SHMM	38.96 ± 0.07	74.03 ± 0.49
	SHMM-finetune	37.66 ± 0.29	72.28 ± 0.56
	H-SHMM	**39.75** ± 1.14	**76.38** ± 0.49
Mboshi	SHMM	38.95 ± 0.60	60.13 ± 0.43
	SHMM-finetune	37.50 ± 0.37	54.30 ± 0.63
	H-SHMM	**42.73** ± 0.97	**64.63** ± 1.74
Yoruba	SHMM	38.98 ± 0.15	63.77 ± 0.39
	SHMM-finetune	36.86 ± 0.24	58.12 ± 0.72
	H-SHMM	**39.52** ± 0.46	**66.27** ± 0.6

TABLE V: Cross-lingual phone-recognizer for AUD.

	NMI			F-Score		
System	English	Mboshi	Yoruba	English	Mboshi	Yoruba
Cross-AM	27.94	25.49	25.69	53.18	43.02	48.39
Cross-FR	35.34	33.58	30.19	70.82	54.16	56.21
Cross-GE	32.69	29.44	25.57	68.43	48.57	53.75
Cross-PO	33.30	31.05	28.15	66.90	56.16	56.78
Cross-SP	33.43	29.83	25.58	67.20	54.10	55.20
Cross-SW	31.93	30.01	24.40	67.10	44.66	49.46
Cross-WO	30.37	35.66	33.03	60.18	59.82	61.12

E. H-SHMM

We have seen that building an AUD system with an educated prior such as the SHMM brings a significant improvement. This performance boost can be explained partly by the added information brought by the phonetic subspace. However, this information may not always be accurate: for instance, the set of languages used for learning the phonetic subspace may not be "relevant" (phonetically speaking) for the target languages. This phonetic mismatch between the source languages and the target language results in the observed performance gap between the SHMM model and Topline-U (see Table II).

As explained in previous sections, the H-SHMM attempts to reduce the mismatch between the source and target language by adapting the phonetic subspace to the target data. In Table IV, we compare the H-SHMM to the SHMM. We observe that if we update the SHMM phonetic subspace posterior $q_0(\mathbf{W}, \mathbf{b})$ on the target language rather than freezing it as learned from the source languages, the clustering and segmentation performance degrades ("SHMM-finetune" in Table IV). The H-SHMM, on the other hand, by constraining the adaptation of the phonetic subspace by its hyper-subspace, successfully adapts the phonetic subspace on the target data. However, despite the improvement brought by the H-SHMM, our best system remains far from Topline-U suggesting that there is still potential for adapting the subspace to the target language.

F. Comparison with other methods

We have shown that subspace models, as implemented by the SHMM and the H-SHMM, offer a significant improvement over the Bayesian HMM baseline. We now broaden the comparison with non-Bayesian approaches.

1) Cross-lingual decoders: We compare our subspace AUD models against cross-lingual decoders. To make the comparison fair, the cross-lingual decoders are structurally equivalent to the AUD models: each of them is an HMM phone-loop (with the same number of Gaussians per state) trained on phonetically transcribed data. We use the same languages and data (Table I) as for estimating the phonetic subspace for the SHMM and H-SHMM models. Results are shown in Table V: we observe that these cross-lingual decoders are much less accurate both in terms of clustering and segmentation. Note that the SHMM and H-SHMM use the data of all source languages whereas the cross-lingual decoders are trained on a single language. To assess that the benefits of the subspace methods are not due to having more data, we report in Table VI the performance of the SHMM[2] using only one language to estimate the phonetic subspace, we observed that for any given language, the SHMM AUD system outperforms the cross-lingual decoder trained on the same source languages.

Additionally, an interesting insight is highlighted by the results in Table VI: for Mboshi, we observe that training the phonetic subspace of the SHMM on a single source language is better than training on all source languages. This indicates that some combination of source languages can be detrimental for the AUD SHMM. However, the H-SHMM, which adapts the phonetic subspace (in an unsupervised fashion) to the target language achieves better results than any SHMM.

2) Neural-network based AUD systems: In recent years, several neural-network-based systems have been proposed for discovering acoustic units from speech. While architecture and objective function differ across models, all of them follow the same principle: an encoding-decoding architecture with one or more discretization hidden layers. We compare our subspace-based models against the following neural-network baselines:

- VQ-VAE [38]: a variational auto-encoder with a quantization layer; variations of this model were successfully used for AUD by several teams in recent iterations of the Zero Resource Challenge [12], [7], [39], [40]. Keeping with our theme of using English as a development language, we tuned the VQ-VAE hyper-parameters to maximize the NMI on English and transferred them to the other languages
- constrained VQ-VAE [41]: a recently proposed post-processing method for VQ-VAE which encourages temporally consecutive frames to be quantized to the same class; this was shown to provide a significant improvement over the the plain VQ-VAE [41]
- ResDAVEnet-VQ [14]: neural network with quantization layers trained to correlate images with their associated audio captions; we choose this baseline to compare our method against an AUD system with a weak supervision signal
- VQ-WAV2VEC [13]: a convolutional neural network with a quantization layer trained with a contrastive prediction objective on the 960 hour Librispeech corpus [42].

The VQ-VAE[3] and the constrained VQ-VAE are trained on the same target data as the SHMM and the H-SHMM. For ResDAVENet and VQ-WAV2VEC, we used the pre-trained model directly and use their quantization output for evaluation purposes. Note that ResDAVENet and VQ-WAV2VEC were trained only on English data which explains some of the degradation in performance when they are used for AUD for other languages.

The results are presented in Table VII. We observe that the best performing neural network baseline is the constrained VQ-VAE, showing that a temporal constraint is an important feature in any AUD models. Nevertheless, the Bayesian subspace models perform significantly better.

The Bayesian AUD models may seem simple compared to large convolutional neural networks. However, they benefit from a well-structured prior which guides them during the clustering. Conversely, the neural network-based AUD models are very potent but lack structured priors and are easily trapped in sub-optimal solutions, limiting how well they can utilize their potential.

G. Effect of the subspaces dimensionality

In this last part, we provide an analysis of the effect of phonetic and language subspace dimensionality.

1) SHMM: In Figure 4a, we illustrate the effect of the subspace dimensionality for the SHMM. We observe that, in terms of clustering, the behavior varies by target language. For English and Yoruba, the optimal subspace dimension is around 250 while for Mboshi, it is between 50 and 100.

This performance variance highlights the major drawback of the SHMM: the assumption of a universal phonetic subspace. Indeed, we see that there is no unique setting that fits well for all possible languages.

Segmentation wise, a low (50-100) dimensional subspace leads to more accurate segmentation. This suggests that a coarser phonetic representation is preferable when the segmentation accuracy is concerned.

2) H-SHMM: The H-SHMM has two subspaces: the language subspace and the phonetic subspace. Figure 4b shows the performance of the H-SHMM as the language subspace dimension is varied (the phonetic subspace dimension is fixed to 100). We observe that having larger dimension is globally preferable though the effect is only significant for the Mboshi data. Notice that the curves in Figure 4b are somewhat noisy, indicating that the H-SHMM is affected by random initialization.

The performance of the H-SHMM as the phonetic subspace dimension is varied is shown in Figure 4c. In this experiment, the language dimension is fixed to 5. We observe that the behavior is now homogeneous across languages: both for the

[2]We make this comparison only with the SHMM, as it is not sensible to estimate the "hyper-subspace" of the H-SHMM with only one lanugage.

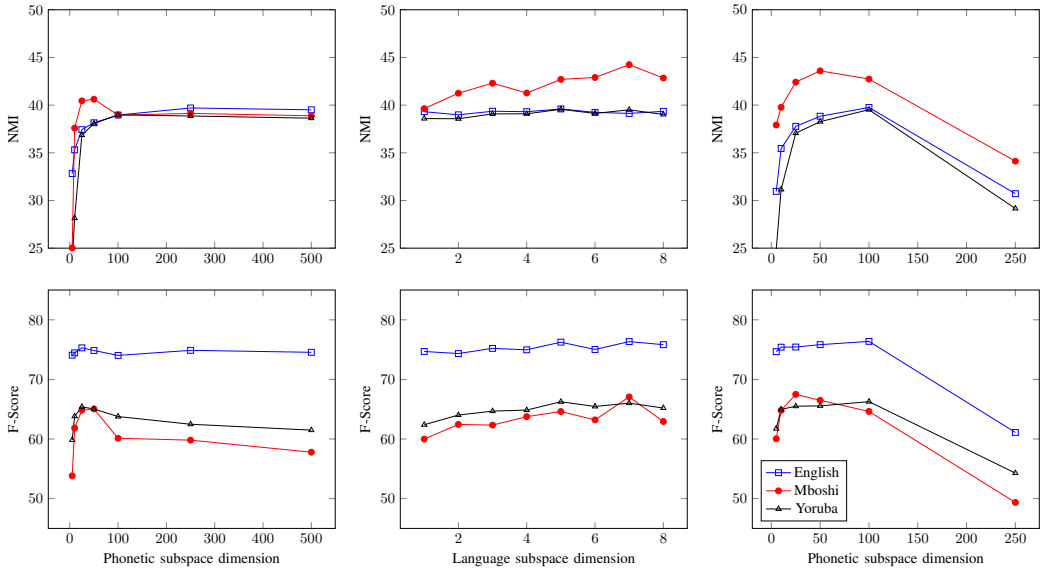
[3]Implementation and training details for the VQ-VAE can be found at https://github.com/BUTSpeechFIT/vq-aud

TABLE VI: SHMM performance pre-trained with only one source language.

System	NMI English	NMI Mboshi	NMI Yoruba	F-Score English	F-Score Mboshi	F-Score Yoruba
SHMM-AM	36.60 ± 0.39	41.17 ± 0.36	38.22 ± 0.15	68.11 ± 0.33	57.90 ± 1.02	62.12 ± 0.29
SHMM-FR	37.13 ± 0.08	41.36 ± 0.40	36.70 ± 0.42	76.27 ± 0.13	63.66 ± 1.03	65.73 ± 0.70
SHMM-GE	35.79 ± 0.08	41.91 ± 0.56	34.64 ± 0.44	76.97 ± 0.52	65.27 ± 1.07	63.62 ± 0.59
SHMM-PO	38.21 ± 0.10	42.04 ± 0.14	37.34 ± 0.10	76.93 ± 0.23	64.34 ± 0.78	64.94 ± 0.28
SHMM-SP	38.01 ± 0.14	41.27 ± 0.41	35.54 ± 0.31	75.00 ± 0.15	64.82 ± 0.44	64.95 ± 0.20
SHMM-SW	34.57 ± 0.19	40.63 ± 0.16	33.91 ± 3.16	75.83 ± 0.13	62.16 ± 0.42	62.30 ± 1.17
SHMM-WO	32.19 ± 0.74	42.70 ± 0.17	34.02 ± 0.08	70.78 ± 1.18	67.62 ± 0.43	66.77 ± 0.16

TABLE VII: Comparison with neural-network-based AUD.

System	NMI English	NMI Mboshi	NMI Yoruba	F-Score English	F-Score Mboshi	F-Score Yoruba
VQ-VAE	35.30 ± 0.50	35.88 ± 0.69	31.74 ± 0.65	52.50 ± 3.17	34.74 ± 2.76	35.26 ± 0.94
constrained VQ-VAE	36.01 ± 0.59	36.49 ± 0.79	32.30 ± 0.62	71.33 ± 2.98	50.47 ± 1.39	49.04 ± 1.44
ResDAVEnet-VQ	34.39	33.67	34.07	64.36	52.85	50.90
VQ-WAV2VEC	35.20	28.79	30.66	26.84	14.94	15.84
SHMM	38.96 ± 0.07	38.95 ± 0.60	38.98 ± 0.15	74.03 ± 0.49	60.13 ± 0.43	63.77 ± 0.39
H-SHMM	**39.75**± 0.58	**42.73** ± 0.97	**39.52** ± 0.46	**76.38** ± 0.49	**64.63** ± 1.74	**66.27** ± 0.60



(a) (b) (c)

Fig. 4: NMI and F-Score metrics when varying the (a) dimension the phonetic subspace of SHMM, (b) dimension of the language subspace of the H-SHMM, and (c) dimension of the phonetic subspace of the H-SHMM.

clustering and segmentation the optimal phonetic subspace dimension is between 50 and 100 dimensions. This shows the benefit of adapting the phonetic subspace on the target language: we can have better clustering and segmentation while using lower dimension to represent each language-specific phonetic subspace.

VI. Conclusion

This work provides a theoretical treatment of subspace models for the task of Acoustic Units Discovery (AUD). It shows how the paradigm of subspace models naturally fits within the non-parametric Bayesian framework: an *educated prior* is formed by constraining the base measure to a subspace that is estimated on phonetically transcribed data from a set of source languages. Thus, the acoustic unit parameters are constrained to live in a phonetic subspace forcing the model to learn units that resemble the phones of the source languages.

This work focuses on two specific models: the Subspace HMM (S-HMM) and the Hierarchical Subspace HMM (H-SHMM). The SHMM assumes that the phonetic subspace is language agnostic: it is the same for every language whereas the H-SHMM assumes that the phonetic subspace is language dependent and has to be adapted on the target language.

Experimental results show that, both the SHMM and the H-SHMM outperform state-of-the-art AUD baselines in terms of clustering quality and segmentation accuracy in three different languages: English, Yoruba and Mboshi. Furthermore, the H-SHMM proves to be superior to the SHMM which supports

the idea that each language has a unique phonology that needs to be learn specifically.

Finally, the concept of subspace models for AUD can be expanded in several ways; we list here potential future research on subspace modeling for AUD.

- The quality of the phonetic subspace—how well the subspace models the continuum of phone in a language—is highly dependent on the choice of the acoustic model, an HMM in the present work. Building more refined generative models of phones would allow a qualitatively better acoustic unit embeddings.
- This works uses a single subspace of all the phones assuming implicitly a continuum between any pair of phones. This continuity may not be relevant between phones of distinct category, e.g. vowels and fricatives. To bypass this issue, one can have a specific subspace for different phonetic categories. This would have the added advantage of easing the interpretation of the final acoustic units (for instance an acoustic unit embedding on a vowel-specific subspace is a vowel)
- Adding a speaker subspace to model explicitly the speaker variability would help the AUD model adapt to the speaker and avoid having speaker-specific clusters.

REFERENCES

[1] S. H. K. Parthasarathi and N. Strom, "Lessons from building acoustic models with a million hours of speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6670–6674.

[2] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[3] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: Acoustic-to-word lstm model for large vocabulary speech recognition," in *Proc. Interspeech 2017*, 2017, pp. 3707–3711. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1566

[4] E. 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.

[5] M. Versteegh, R. Thiolliere, T. Schatz, X. N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015," in *Sixteenth annual conference of the international speech communication association*, 2015.

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

[7] ——, "The Zero Resource Speech Challenge 2019: TTS Without T," in *Interspeech*, 2019, pp. 1088–1092. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2904

[8] C.-y. Lee and J. 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] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Parallel inference of dirichlet process gaussian mixture models for unsupervised acoustic modeling: A feasibility study," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[10] M. Heck, S. Sakti, and S. Nakamura, "Iterative training of a dpgmm-hmm acoustic unit recognizer in a zero resource scenario," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 57–63.

[11] L. Ondel, L. Burget, and J. Černockỳ, "Variational inference for acoustic unit discovery," *Procedia Computer Science*, vol. 81, pp. 80–86, 2016.

[12] J. Chorowski, R. J. Weiss, S. Bengio, and A. 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.

[13] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=rylwJxrYDS

[14] D. Harwath, W.-N. Hsu, and J. Glass, "Learning Hierarchical Discrete Linguistic Units from Visually-Grounded Speech," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=B1elCp4KwH

[15] J. Ebbers, J. Heymann, L. Drude, T. Glarner, R. Haeb-Umbach, and B. Raj, "Hidden Markov Model Variational Autoencoder for Acoustic Unit Discovery," in *Interspeech*, 2017, pp. 488–492.

[16] T. Glarner, P. Hanebrink, J. Ebbers, and R. Haeb-Umbach, "Full Bayesian Hidden Markov Model Variational Autoencoder for Acoustic Unit Discovery," in *Interspeech*, 2018, pp. 2688–2692.

[17] L. Ondel, H. K. Vydana, L. Burget, and J. Černocký, "Bayesian Subspace Hidden Markov Model for Acoustic Unit Discovery," in *Interspeech*, 2019, pp. 261–265. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2224

[18] B. Yusuf, L. Ondel, L. Burget, J. Cernocky, and M. Saraclar, "A hierarchical subspace model for language-attuned acoustic unit discovery," *arXiv e-prints*, pp. arXiv–2011, 2020.

[19] D. M. Blei, "Variational methods for the dirichlet process," in *In Proceedings of the Twenty-First International Conference on Machine Learning (ICML) 2004*, 2004.

[20] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532–556, 1976.

[21] P. Orbanz and Y. W. Teh, "Bayesian nonparametric models." *Encyclopedia of machine learning*, no. 1, 2010.

[22] Y. W. Teh, "Dirichlet processes," in *Encyclopedia of Machine Learning*. Springer, 2010.

[23] H. Bourlard, "Reconnaissance automatique de la parole: modélisation ou description," *Journées Etude Parole'96*, pp. 263–272, 1996.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006, ch. 2, pp. 113–120.

[25] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[26] A. F. L. Ondel, "Discovering acoustic units from speech: a bayesian approach," Ph.D. thesis, Brno University of Technology, Faculty of Information Technology, 2021. [Online]. Available: https://www.fit.vut.cz/study/phd-thesis/751/

[27] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[29] P. Godard *et al.*, "A very low resource language speech corpus for computational language documentation experiments," *arXiv preprint arXiv:1710.03501*, 2017.

[30] A. Gutkin, I. Demirşahin, O. Kjartansson, C. Rivera, and K. Túbòsún, "Developing an Open-Source Corpus of Yoruba Speech," in *Interspeech*, Shanghai, China, 2020.

[31] 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.

[32] T. Schultz, N. T. Vu, and T. 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.

[33] S. T. Abate, W. Menzel, and B. Tafila, "An Amharic Speech Corpus for Large Vocabulary Continuous Speech Recognition," in *INTERSPEECH-2005*, 2005.

[34] H. Gelas, L. Besacier, and F. 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. [Online]. Available: http://hal.inria.fr/hal-00954048

[35] E. Gauthier, L. Besacier, S. Voisin, M. Melese, and U. P. Elingui, "Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition: a Case Study of Wolof," *LREC*, 2016.

[36] L. Besacier *et al.*, "Speech technologies for african languages: Example of a multilingual calculator for education," in *Interspeech*, 2015.

[37] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. USA: Wiley-Interscience, 1991.

[38] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6306–6315.
[39] A. Tjandra, B. Sisman, M. Zhang, S. Sakti, H. Li, and S. Nakamura, "VQVAE Unsupervised Unit Discovery and Multi-Scale Code2Spec Inverter for Zerospeech Challenge 2019," in *Interspeech*, 2019, pp. 1118–1122. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3232
[40] E. Dunbar, J. Karadayi, M. Bernard, X.-N. Cao, R. Algayres, L. Ondel, L. Besacier, S. Sakti, and E. Dupoux, "The zero resource speech challenge 2020: Discovering discrete subword and word units," *arXiv preprint arXiv:2010.05967*, 2020.
[41] H. Kamper and B. van Niekerk, "Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks," *arXiv preprint arXiv:2012.07551*, 2020.
[42] V. Panayotov, G. Chen, D. Povey, and S. 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.
[43] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
[44] M. J. Beal, "Variational algorithms for approximate bayesian inference," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, University College London, 2003. [Online]. Available: http://www.cse.buffalo.edu/faculty/mbeal/thesis/index.html
[45] H. Ishwaran and L. F. James, "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 161–173, 2001.

APPENDIX A
INFERENCE

In this appendix, we detailed the optimization of the variational posterior defined in (47). Note that $q(\mathbf{H})$ and $q(\mathbf{v})$ are distributions over infinite set of variables and, therefore, cannot be used directly in any practical implementation. We derive the optimal factors ignoring this issue and we address it specifically in appendix A-D.

A. Latent variables $\mathbf{z}, \mathbf{c}$

We assume $q(\mathbf{H})$, $q(\mathbf{v})$ and $q(\gamma)$ are fixed and we derive the optimal variational posteriors $q^*(\mathbf{c}|\mathbf{z})$ and $q^*(\mathbf{z})$. $q^*(\mathbf{c}|\mathbf{z})$:

$$\ln q^*(\mathbf{c}|\mathbf{z}) \stackrel{+}{=} \sum_{n=1}^{N} \langle \ln p(\mathbf{x}_n, c_n|\boldsymbol{\eta}_{z_n}) \rangle_{q(\boldsymbol{\eta}_{z_n})}, \tag{49}$$

where $\stackrel{+}{=}$ means equality up to a constant. Furthermore, (49) implies that:

$$q^*(\mathbf{c}|\mathbf{z}) = \prod_{n=1}^{N} q^*(c_n|z_n) \tag{50}$$

$$q^*(c_n|z_n) = \frac{\exp\{\langle \ln p(\mathbf{x}_n, c_n|\boldsymbol{\eta}_{z_n}) \rangle_{q(\boldsymbol{\eta}_{z_n})}\}}{\sum_{j=1}^{C} \exp\{\langle \ln p(\mathbf{x}_n, j, \boldsymbol{\eta}_{z_n}) \rangle_{q(\boldsymbol{\eta}_{z_n})}\}}. \tag{51}$$

From (10), the expected likelihood has the following form:

$$\langle \ln p(\mathbf{x}_n, c_n|\boldsymbol{\eta}_{z_n}) \rangle_{q(\boldsymbol{\eta}_{z_n})} = \begin{bmatrix} \langle \boldsymbol{\omega}_{z_n} \rangle_q \\ -\langle A(\boldsymbol{\omega}_{z_n}) \rangle_q \end{bmatrix}^\top \begin{bmatrix} T(c_n) \\ 1 \end{bmatrix} + \begin{bmatrix} \langle \boldsymbol{\theta}_{z_n}^{c_n} \rangle_q \\ -\langle A(\boldsymbol{\theta}_{z_n}^{c_n}) \rangle_q \end{bmatrix}^\top \begin{bmatrix} T(\mathbf{x}_n) \\ 1 \end{bmatrix}. \tag{52}$$

In practice, the expectations are estimated empirically using the variational posterior $q(\boldsymbol{\eta}_{z_n})$ derived in appendix A-B.

Using (51), we derive the optimal posterior of the HMM state sequence:

$$\ln q^*(\mathbf{z}) \stackrel{+}{=} \sum_{n=1}^{N} \langle \ln \frac{p(\mathbf{x}_n, c_n|\boldsymbol{\eta}_{z_n})}{q(c_n|z_n)} \rangle_{q(c_n|z_n)q(\boldsymbol{\eta}_{z_n})} + \langle \ln p(z_n|z_{n-1}\mathbf{v}) \rangle_{q(\mathbf{v})}. \tag{53}$$

For conciseness, we introduce the following placeholders:

$$\phi_n(z_n) = \langle \ln \frac{p(\mathbf{x}_n, c_n|\boldsymbol{\eta}_{z_n})}{q(c_n|z_n)} \rangle_{q(c_n|z_n)q(\boldsymbol{\eta}_{z_n})} \tag{54}$$

$$A_{z_{n-1},z_n} = \langle \ln p(z_n|z_{n-1}, \mathbf{v}) \rangle_{q(\mathbf{v})}. \tag{55}$$

Rewriting (53) with $\phi_n(z_n)$ and A_{z_{n-1},z_n} we get:

$$q^*(\mathbf{z}) = \frac{1}{\zeta} \prod_{n=1}^{N} \exp\{\phi_n(z_n) + A_{z_{n-1},z_n}\} \tag{56}$$

$$\zeta = \sum_{\mathbf{z}} \prod_{n=1}^{N} \exp\{\phi_n(z_n) + A_{z_{n-1},z_n}\}. \tag{57}$$

The normalization constant ζ in (57) requires to sum over all possible state sequences $\mathbf{z}$. Despite the astronomical number of possible sequences, this sum can be computed exactly and efficiently using dynamic programming [43], [44], [26].

B. Acoustic units' embeddings

We focus now on deriving the acoustic units' posterior $q(\mathbf{H})$ using first the SHMM and then H-SHMM. In both cases, we assume the other variational factors $q(\mathbf{c}|\mathbf{z})$, $q(\mathbf{z})$, $q(\mathbf{v})$ and $q(\gamma)$ to be fixed.

1) SHMM: Recall from section III-C that each acoustic unit vector is constructed from a low-dimensional embedding in a subspace. Because the prior and the likelihood are not-conjugate, we cannot obtain closed-form solution and, consequently, we add the following parametric constraints to the variational posterior:

$$q(\mathbf{W}, \mathbf{b}) \prod_{i=1}^{\infty} = \mathcal{N}(\boldsymbol{\nu}, \mathrm{diag}(\exp\{\boldsymbol{\xi}\})), \tag{58}$$

where $\boldsymbol{\nu}$ is a vectorized form of $\mathbf{W}$ and $\mathbf{b}$, and we optimize the empirical expectation of (46) with respect to the parameters $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$:

$$\mathcal{L} \stackrel{+}{=} \frac{1}{J} \sum_{j=1}^{J} \sum_{n=1}^{N} \langle \ln p(\mathbf{x}_n, c_n|\boldsymbol{\eta}_{z_n,j}) \rangle_{q(c_n|z_n)q(z_n)} - \mathrm{D}_{\mathrm{KL}}\left(q(\mathbf{S}_j, \mathbf{E}_j)||p(\mathbf{S}_j, \mathbf{E}_j)\right) \tag{59}$$

$$(\mathcal{S}_j, \mathbf{E}_j) = \boldsymbol{\nu} + \exp\{\boldsymbol{\xi}\} \odot \epsilon_j, \quad \epsilon_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{60}$$

where $\odot$ represents element-wise multiplication. In (59), the expectation of the log-likelihood can be computed exactly by using (51) and $q(z_n)$:

$$\langle \ln p(\cdot|\boldsymbol{\eta}_{z_n,j}) \rangle_q = q(z_n) \begin{bmatrix} \boldsymbol{\omega}_{z_n,j} \\ -A(\boldsymbol{\omega}_{z_n,j}) \end{bmatrix}^\top \begin{bmatrix} q(c_1|z_n) \\ \vdots \\ q(c_{C-1}|z_n) \\ 1 \end{bmatrix} + q(z_n)q(c_n|z_n) \begin{bmatrix} \boldsymbol{\theta}_{z_n,j}^{c_n} \\ -A(\boldsymbol{\theta}_{z_n,j}^{c_n}) \end{bmatrix}^\top \begin{bmatrix} T(\mathbf{x}_n) \\ 1 \end{bmatrix}. \tag{61}$$

In practice, we optimize (59) using stochastic gradient ascent.

2) H-SHMM: Optimization of the acoustic units' posterior in the H-SHMM is very similar to the SHMM case. However, we need to take into account that each subspace is language specific. Let's consider that we have a set of L languages and we would like to learn an inventory of acoustic units $\mathbf{H}^\lambda$ for each language $\lambda \in \{1, \dots, L\}$. From the definition of the H-SHMM ((40)-(44)), we have:

$$q(\{\mathbf{H}^\lambda\}) \triangleq q(\mathcal{M}) \prod_{\lambda=1}^{L} \left[\prod_{i=1}^{\infty} q(\mathbf{e}_i^\lambda) \right] q(\boldsymbol{\alpha}^\lambda). \tag{62}$$

(63)

Similarly as before, we introduce the following parametric constraint:

$$q(\mathcal{M}, \{\mathbf{E}^\lambda\}, \{\boldsymbol{\alpha}^\lambda\}) = \mathcal{N}(\boldsymbol{\nu}, \mathrm{diag}(\exp\{\boldsymbol{\xi}\})), \tag{64}$$

and we optimize the empirical expectation of (46):

$$\begin{aligned} \mathcal{L} \stackrel{+}{=} \frac{1}{J} \sum_{j=1}^{J} \sum_{\lambda=1}^{L} \Bigg[& \sum_{n=1}^{N} \langle \ln p(\mathbf{x}_n^\lambda, c_n^\lambda | \boldsymbol{\eta}_{z_n,j}^\lambda) \rangle_{q(c_n^\lambda | z_n^\lambda) q(z_n^\lambda)} \\ & - \mathrm{D}_{\mathrm{KL}} \left(q(\mathbf{E}_j^\lambda, \boldsymbol{\alpha}_j^\lambda) || p(\mathbf{E}_j^\lambda, \boldsymbol{\alpha}_j^\lambda) \right) \Bigg] \\ & - \mathrm{D}_{\mathrm{KL}}(q(\mathcal{M}_j) || q(\mathcal{M}_j)) \end{aligned} \tag{65}$$

$$(\mathcal{M}_j, \{\mathbf{E}_j^\lambda\}, \{\boldsymbol{\alpha}_j^\lambda\}) = \boldsymbol{\nu} + \exp\{\boldsymbol{\xi}\} \odot \epsilon_j, \quad \epsilon_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{66}$$

using stochastic gradient ascent.

C. Update of the stick-breaking process

We address now the last part of the inference: the update of the variational posteriors of the stick-breaking process. For this stage, we consider that the variational posteriors $q(\mathbf{c}|\mathbf{z})$, $q(\mathbf{z})$, $q(\mathbf{H})$ are fixed. The following updates equations are based on the the variational treatment of the stick-breaking process presented in [19].

1) Stick-breaking parameters: We first start to estimate the optimal $q^*(\mathbf{v})$ assuming $q(\gamma)$ is fixed:

$$\ln q^*(\mathbf{v}) \stackrel{+}{=} \langle \ln p(\mathbf{z}|\mathbf{v}) \rangle_{q(\mathbf{z})} + \ln p(\mathbf{v}). \tag{67}$$

From (22) and (23) we have $p(\mathbf{z}|\mathbf{v}) = p(\mathbf{s}|\mathbf{u})p(\mathbf{u}|\mathbf{v})$ which leads to:

$$q^*(\mathbf{v}) = \prod_{k=1}^{\infty} \mathcal{B}(\alpha_k, \beta_k) \tag{68}$$

$$\alpha_k = 1 + \Big\langle \sum_{u_i \in \mathbf{u}} \mathbb{1}[u_i = k] \Big\rangle_{q(\mathbf{u})} \tag{69}$$

$$\beta_k = \langle \gamma \rangle_{q(\gamma)} + \Big\langle \sum_{u_i \in \mathbf{u}} \mathbb{1}[u_i > k] \Big\rangle_{q(\mathbf{u})}), \tag{70}$$

where $\mathbb{1}[...]$ is the indicator function. The expectations in (69) adn (70) requires summing over all the units of all possible sequences $\mathbf{u}$ from $q(\mathbf{z})$. Once again, this large summation can be calculated exactly using dynamic programming [26].

2) Concentration parameters: Finally, the optimal variational posterior $q^*(\gamma)$ while assuming $q(\mathbf{v})$ is fixed is given by:

$$q^*(\gamma) = \mathcal{G}(a, b) \tag{71}$$

$$a = a_0 + \sum_{k=1}^{\infty} 1, \quad b \quad = b_0 - \sum_{k=1}^{\infty} \langle \ln(1 - v_k) \rangle_{q(v_k)}. \tag{72}$$

D. Truncation

In our derivation of the optimal variational posteriors, we have ignored issues raised by the sum or product of infinitely many terms. Following [19], we address this by introducing a truncation parameter τ such that $q(v_\tau = 1) = 1$. This approximation, motivated by the almost sure truncation of the Dirichlet Process [45], ensures that $q(u_i > \tau) = 0$, $\forall i$ and, therefore, truncates all infinite sum and product to τ terms in the solution of the optimal variational posteriors.

E. Initialization

Because of the constraints imposed on on the variational posterior (47), the optimization is prone to converge to a local optimum. To avoid this, we initialize the model for the *supervised* phase of the training by the following procedure:

1) we train a standard HMM with C-components GMM emissions for each phone using the Baum-Welch training and the provided phonetic transcription.
2) for each state of each phone's HMM
   a) we set the mixing weights $\boldsymbol{\pi}$ such that $\pi_k = \frac{1}{C}$
   b) compute the per-state global mean $\hat{\boldsymbol{\mu}} = \frac{1}{C} \sum_{c=1}^{C} \boldsymbol{\mu}_c$ and global diagonal covariance matrix $\hat{\boldsymbol{\Sigma}} = \frac{1}{C} \sum_{c=1}^{C} \boldsymbol{\Sigma}_c$
   c) we set each Gaussian component to have mean $\hat{\boldsymbol{\mu}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$.
3) using the HMM estimated in step 1, we initialize $q^*(z_n)$ using the Baum-Welch algorithm and we set $q^*(c_n|z_n) = \mathrm{const}$
4) we set $\boldsymbol{\nu} = \mathbf{0}$ and $\exp\{\boldsymbol{\xi}\} = \frac{1}{D}\mathbf{1}$ then, we burn-in the model by optimizing $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$ until convergence while keeping other factors fixed.

The initialization for the *unsupervised* phase—the actual AUD task—is easier:

- we initialize the posterior of the stick-breaking process by setting $q^*(\mathbf{v}) \coloneqq p(\mathbf{v})$ and $q^*(\gamma)$
- we use the variational posteriors estimated during the supervised phase to initialize the new variational posteriors as explained in section IV-B.
- finally, we set $\boldsymbol{\nu}, \boldsymbol{\xi}$ such that $q_1(\mathbf{E}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ (respectively $q_1(\mathbf{E}, \boldsymbol{\alpha}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ for the H-SHMM).