Proceedings of SIGUL2022 @LREC2022, pages 1–9
Marseille, 24-25 June 2022

Unsupervised Word Segmentation from Discrete Speech Units in Low-Resource Settings

Marcely Zanon Boito[1,*], Bolaji Yusuf[2,3], Lucas Ondel[4], Aline Villavicencio[5], Laurent Besacier[6]
[1]Avignon University, FR, [2]Bogazici University, TR
[3]Brno University of Technology, CZ
[4]LISN CNRS, FR [5]University of Sheffield, UK
[6]Naver Labs Europe, FR and University Grenoble Alpes, FR
* Research done while at University Grenoble Alpes.
contact: marcely.zanon-boito at univ-avignon dot fr

Abstract

Documenting languages helps to prevent the extinction of endangered dialects – many of which are otherwise expected to disappear by the end of the century. When documenting oral languages, unsupervised word segmentation (UWS) from speech is a useful, yet challenging, task. It consists in producing time-stamps for slicing utterances into smaller segments corresponding to words, being performed from phonetic transcriptions, or in the absence of these, from the output of unsupervised speech discretization models. These discretization models are trained using raw speech only, producing discrete speech units that can be applied for downstream (text-based) tasks. In this paper we compare five of these models: three Bayesian and two neural approaches, with regards to the exploitability of the produced units for UWS. For the UWS task, we experiment with two models, using as our target language the Mboshi (Bantu C25), an unwritten language from Congo-Brazzaville. Additionally, we report results for Finnish, Hungarian, Romanian and Russian in equally low-resource settings, using only 4 hours of speech. Our results suggest that neural models for speech discretization are difficult to exploit in our setting, and that it might be necessary to adapt them to limit sequence length. We obtain our best UWS results by using Bayesian models that produce high quality, yet compressed, discrete representations of the input speech signal.

Keywords: unsupervised word segmentation, speech discretization, acoustic unit discovery, low-resource settings

1. Introduction

Popular models for speech processing still rely on the availability of considerable amounts of speech data and their transcriptions, which reduces model applicability to a limited subset of languages considered *high-resource*. This excludes a considerable number of *low-resource* languages, including many from oral tradition. Besides, learning supervised representations from speech differs from the unsupervised way infants learn language, hinting that it should be possible to develop more data-efficient speech processing models.

Recent efforts for *zero-resource* processing (Glass, 2012; Jansen et al., 2013; Versteegh et al., 2016; Dunbar et al., 2017; Dunbar et al., 2019; Dunbar et al., 2020) focus on building speech systems using limited amounts of data (hence *zero resource*), and without textual or linguistic resources, for increasingly challenging tasks such as acoustic or lexical unit discovery. Such zero resource approaches also stimulated interest for computational language documentation (Besacier et al., 2006; Duong et al., 2016; Godard et al., 2018; Bird, 2021) and computational language acquisition (Dupoux, 2018).

In this paper we address the challenging task of unsupervised word segmentation (UWS) from speech. This task consists of outputting time-stamps delimiting stretches of speech, associated with class labels corresponding to word hypotheses, without access to any supervision. We build on the work presented in Godard et al. (2018): they proposed a cascaded model for UWS that first generates a discrete sequence from the speech signal using the model from Ondel et al. (2016), and then segments the discrete sequence into words using a Bayesian (Goldwater, 2007) or a neural (Boito et al., 2017) approach. Since then, much progress has been made in automatic speech discretization: efficient Bayesian models for acoustic unit discovery (AUD) emerged (Ondel et al., 2019; Yusuf et al., 2021), and self-supervised models based on neural networks – typically made of an auto-encoder structure with a discretization layer – were also introduced (van den Oord et al., 2017; Baevski et al., 2020a; Chorowski et al., 2019).

Therefore, in this work we revise and extend Godard et al. (2018) by empirically investigating the *exploitability* of five recent approaches for speech discretization for the UWS task in a rather low-resource scenario, using approximately 4 hours of speech (roughly 5k sentences). More precisely, we train three Bayesian speech discretization models (*HMM* (Ondel et al., 2016), *SHMM* (Ondel et al., 2019) and *H-SHMM* (Yusuf et al., 2021)), and two neural models (*VQ-VAE* (van den Oord et al., 2017) and *vq-wav2vec* (Baevski et al., 2020a)). We extract discrete speech units from them using only 4 hours of speech, and we perform UWS from the sequences produced. Our pipeline targets the Mboshi language (Bantu C25), an unwritten language from Congo-Brazzaville. Additionally, we perform experiments in equal data settings for Finnish, Hungarian, Romanian and Russian. This allows us to assess the language-related impact in our UWS pipeline.

Our experiments show that neural models for speech discretization are difficult to exploit for UWS, as they output very long sequences. In contrast to that, the Bayesian speech discretization approaches from Ondel et al. (2019) and Yusuf et al. (2021) are robust and generalizable, producing high quality, yet compressed, discrete speech sequences from the input utterances in all languages. We obtain our best results by using these sequences for training the neural UWS model from Boito et al. (2017).

This paper is organized as follows. Section 2 presents related work, and Section 3 details the speech discretization models we experiment with. Section 4 presents our experimental setup, and Section 5 our experiments. Section 6 concludes our work.

2. Related Work

The work presented here revises the UWS model from speech in low-resource settings presented in Godard et al. (2018). Boito et al. (2019) complemented that work by tackling different neural models for bilingual UWS, but they did not address the discretization portion of the pipeline, working directly from manual phonetic transcriptions. In Kamper and van Niekerk (2021), the authors propose constraining the VQ-VAE model in order to generate a more exploitable output representation for direct application to the UWS task in English. Different from that, in this work we focus on providing an empirical comparison of recent discretization approaches, extending Godard et al. (2018) and providing results in low-resource settings, and in five different languages.

This work falls into the category of computational language documentation approaches. Recent works in this field include the use of aligned translation for improving transcription quality (Anastasopoulos and Chiang, 2018), and for obtaining bilingually grounded UWS (Duong et al., 2016; Boito et al., 2017). We find pipelines for obtaining manual (Foley et al., 2018) and automatic (Michaud et al., 2018) transcriptions, and for aligning transcription and audio (Strunk et al., 2014). Other examples are methods for low-resource segmentation (Lignos and Yang, 2010; Goldwater et al., 2009), and for lexical unit discovery without textual resources (Bartels et al., 2016). Finally, direct speech-to-speech (Tjandra et al., 2019) and speech-to-text (Besacier et al., 2006; Bérard et al., 2016) architectures could be an option for the lack of transcription, but it remains to be seen how exploitable these architectures can be in low-resource settings.

Lastly, we highlight that recent models based on self-supervised learning (Schneider et al., 2019; Baevski et al., 2019; Wang et al., 2020; Liu et al., 2020; Baevski et al., 2020b; Hsu et al., 2021) provide an interesting novel option for reducing the amount of labeled data needed in downstream tasks such as automatic speech recognition and speech translation. In this work we experiment with the vq-wav2vec model, a predecessor of the popular wav2vec 2.0 (Baevski et al., 2020b). We however, do not extend our investigation to the latter, or to models such as HuBERT (Hsu et al., 2021). This is because, while these models do produce a certain discretization of the speech (for wav2vec 2.0 via quantization module, for HuBERT via clustering of MFCC features), we judge this discretization to be insufficiently exploitable for downstream text-based approaches due to their excessive length.[1] We do, however, find promising the integration of self-supervised speech features into Bayesian AUD models as in Ondel et al. (2022).

3. Unsupervised Speech Discretization Models

Speech discretization consists in labeling the speech signal into discrete speech units, which can correspond or not to the language phonetic inventory. This problem can be formulated as the learning of a set of U discrete units with embeddings $\mathbf{H} = \{\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^U\}$ from a sequence of untranscribed acoustic features $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, as well as the assignment of frame to unit $\mathbf{z} = [z_1, \ldots, z_N]$. Depending on the approach, neural (Section 3.1) or Bayesian (Section 3.2), the assumptions and the inference regarding these three quantities will differ.

3.1. Neural (VQ-based) models

VQ-VAE. It comprises an encoder, a decoder, and a set of unit-specific embeddings $\mathbf{H}$. The encoder is a neural network that transforms the data into a continuous latent representation $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_N)$. Each frame is then assigned to the closest embedding in the Euclidean sense (Equation 1). The decoder transforms the sequence of quantized vectors into parameters of the conditional log-likelihood of the data $p(\mathbf{x}_n|\mathbf{z})$, and the network is trained to maximize this likelihood. Since the quantization step is not differentiable, the encoder is trained with a straight through estimator (Bengio et al., 2013). In addition, a pair of ℓ_2 losses are used to minimize the quantization error, and the overall objective function that is maximized is presented in Equation 2, where $\text{sg}[\cdot]$ is the stop-gradient operator. We define the likelihood $p(\mathbf{x}_n|z_n) = \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}(\boldsymbol{\eta}^{z_n}), \mathbf{I})$. Under this assumption, the log-likelihood reduces to the mean-squared error $||\mathbf{x}_n - \boldsymbol{\mu}(\boldsymbol{\eta}^{z_n})||_2^2$.

$$z_n = \arg\min_u ||\mathbf{v}_n - \boldsymbol{\eta}^u||_2. \qquad (1)$$

[1] For instance, wav2vec 2.0 trains on a joint *diversity* loss for inciting the use of its discrete units. Their large codebook of $G = 8; V = 8$ results in an upper-bound of 8^8 units.

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\Big(\ln p(\mathbf{x}_n|z_n) - k_1||\,\mathrm{sg}[\boldsymbol{\eta}^{z_n}] - \mathbf{v}_n||_2^2 - k_2||\boldsymbol{\eta}^{z_n} - \mathrm{sg}[\mathbf{v}_n]||_2^2\Big), \qquad (2)$$

vq-wav2vec. This model is composed of an encoder ($f : \mathbf{X} \to \mathbf{Z}$), a quantizer ($q : \mathbf{Z} \to \hat{\mathbf{Z}}$) and an aggregator ($g : \hat{\mathbf{Z}} \to \mathbf{C}$). The encoder is a CNN which maps the raw speech input $\mathbf{X}$ into the dense feature representation $\mathbf{Z}$. From this representation, the quantizer produces discrete labels $\hat{\mathbf{Z}}$ from a fixed-size codebook $\mathbf{e} \in \mathbb{R}^{V\times d}$ with V representations of size d. Since replacing an encoder feature vector $\mathbf{z}_i$ by a single entry in the codebook makes the method prone to model collapse, the authors independently quantize partitions of each feature vector by creating multiple *groups* G, arranging the feature vector into a matrix $\mathbf{z'} \in \mathbb{R}^{G\times(d/G)}$. Considering each row as an integer index, the full feature vector is represented by the indices $\mathbf{i} \in [V]^G$, with V being the possible number of *variables* for a given group, and each element $\mathbf{i}_j$ corresponding to a fixed codebook vector ($j \in |G|$). For each of the G groups, the quantization is performed by using Gumbel-Softmax (Jang et al., 2017) or online k-means clustering. Finally, the aggregator combines multiple quantized feature vector time-steps into a new representation $\mathbf{c}_i$ for each time step i. The model is trained to distinguish a sample k steps in the future $\hat{\mathbf{z}}_{i+k}$ from *distractor* samples $\tilde{\mathbf{z}}$ drawn from a distribution p_n. This is done by minimizing the contrastive loss for steps $k = \{1, \ldots, K\}$ as in Equation 3, where T is the sequence length, $\sigma(x) = 1/(1 + exp(-x))$, $\sigma(\hat{\mathbf{z}}_{i+k}^{\mathsf{T}} h_k(\mathbf{c}_\mathbf{i}))$ is the probability of $\hat{\mathbf{z}}_{i+k}$ being the true sample, and $h_k(\mathbf{c}_i)$ is the step-specific affine transformation $h_k(\mathbf{c}_i) = W_k\mathbf{c}_i + b_k$. Finally, this loss is accumulated over all k steps $\mathcal{L} = \sum_{k=1}^{K}\mathcal{L}_k$.

$$\mathcal{L}_k = \sum_{i=1}^{T-k}\Big(\log\sigma(\hat{\mathbf{z}}_{i+k}^{\mathsf{T}} h_k(\mathbf{c}_\mathbf{i})) + \lambda\mathbb{E}_{\tilde{\mathbf{z}}\sim p_n}[\log\sigma(-\tilde{\mathbf{z}}^{\mathsf{T}} h_k(\mathbf{c}_i))]\Big) \qquad (3)$$

Training. For **VQ-VAE**, the encoder has 4 Bi-LSTM layers each with output dimension 128 followed by a 16-dimensional feed-forward decoder with one hidden layer. The number of discovered units (quantization centroids) is set to 50. This setting is unusually low but it helps to reduce the length of the output sequence. We set $k_1 = 2$ and $k_2 = 4$ (Equation 2), and train[2] with Adam (Kingma and Ba, 2015) with an initial learning rate of 2×10^{-3} which is halved whenever the loss stagnates for two training epochs.

For **vq-wav2vec**, we use the small model from (Baevski et al., 2020a),[3] but with only 64 channels, residual scale of 0.2, and warm-up of 10k. For vocabulary we set $G = 2$ and experimented with having both $V = 4$, resulting in 16 units (*VQ-W2V-V16*), and $V = 6$, resulting in 36 units (*VQ-W2V-V36*). Larger vocabularies resulted in excessively long sequences which could not be used for UWS.[4] We also experimented reducing the representation by using byte pair encoding (BPE) (Sennrich et al., 2016), hypothesizing that phones were being modeled by a combination of different units. In this setting, BPE serves as a method for identifying and clustering these patterns. Surprisingly, we found that using BPE resulted in a decrease in UWS performance. This hints that this model might not be very consistent during its labeling process.

3.2. Bayesian Generative Models

For generative models, each acoustic unit embedding $\boldsymbol{\eta}_i$ represents the parameters of a probability distribution $p(\mathbf{x}_n|\boldsymbol{\eta}_{z_n}, z_n)$ with latent variables $\mathbf{z}$. Discovering the units amounts to estimating the posterior distribution over the embeddings $\mathbf{H}$ and the assignment variables $\mathbf{z}$ given by:

$$p(\mathbf{z}, \mathbf{H}|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{z}, \mathbf{H})p(\mathbf{z}|\mathbf{H})\prod_{u=1}^{U} p(\boldsymbol{\eta}^u). \qquad (4)$$

From this, we describe three different approaches.

HMM. In this model each unit is a 3-state left-to-right HMM with parameters $\boldsymbol{\eta}^i$. Altogether, the set of units forms a large HMM analog to a "phone-loop" recognition model. This model, described in Ondel et al. (2016), serves as the backbone for the two subsequent models.

SHMM. The prior $p(\boldsymbol{\eta})$ in Equation 4 is the probability that a sound, represented by an HMM with parameters $\boldsymbol{\eta}$, is an acoustic unit. For the former model, it is defined as a combination of exponential family distributions forming a prior conjugate to the likelihood. While mathematically convenient, this prior does not incorporate any knowledge about phones, i.e. it considers all possible sounds as potential acoustic units. In Ondel et al. (2019), they propose to remedy this shortcoming by defining the parameters of each unit u as in Equation 5, where $\mathbf{e}^u$ is a low-dimensional unit embedding, $\mathbf{W}$ and $\mathbf{b}$ are the parameters of the *phonetic subspace*, and the function $f(\cdot)$ ensures that the resulting vector $\boldsymbol{\eta}^u$ dwells in the HMM parameter space. The subspace, defined by $\mathbf{W}$ and $\mathbf{b}$, is estimated from several labeled source languages. The prior $p(\boldsymbol{\eta})$ is defined over the low-dimensional embeddings $p(\mathbf{e})$ rather than $\boldsymbol{\eta}$ directly, therefore constraining the search of units in the relevant region of the parameter space. This model is denoted as the Subspace HMM (SHMM).

$$\boldsymbol{\eta}^u = f(\mathbf{W}\cdot\mathbf{e}^u + \mathbf{b}) \qquad (5)$$

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

[3] Implementation available at: https://github.com/pytorch/fairseq/tree/master/examples/wav2vec

[4] For instance, the dpseg original implementation only processes sequences shorter than 350 tokens.

		#Types	#Tokens	Avg Token Length	Avg #Tokens per Sentence
MB-FR	**MB***	6,633	30,556	4.2	6.0
	FR	5,162	42,715	4.4	8.3
MaSS	**FI***	12,088	70,226	6.0	13.2
	HU*	12,993	69,755	5.9	13.1
	RO*	6,795	84,613	4.5	15.9
	RU*	10,624	67,176	6.2	12.6
	FR	7,226	94,527	4.1	17.8

Table 1: Statistics for the datasets, computed over the text (FR), or over the phonetic representation (*).

		HMM	SHMM	H-SHMM
RAW	**# Units**	77 (+9)	76 (+8)	49 (-19)
	Avg #Units per sequence	27.5 (+8.7)	24.0 (+5.2)	21.7 (+2.9)
	Max Length	68 (+17)	69 (+18)	63 (+12)
+SIL	**# Units**	75 (+7)	75 (+7)	47 (-21)
	Avg #units per sequence	20.9 (+2.1)	19.9 (+1.1)	19.4 (+0.6)
	Max Length	69 (+18)	62 (+11)	60 (+9)
		VQ-VAE	**VQ-W2V-16**	**VQ-W2V-36**
RAW	**# Units**	50 (-18)	16 (-52)	36 (-32)
	Avg #units per sequence	65.2 (+46.4)	81.7 (+62.9)	111.0 (+92.2)
	Max Length	217 (+166)	289 (+238)	361 (+310)
+SIL	**# Units**	50 (-18)	16 (-52)	36 (-32)
	Avg #units per sequence	43.4 (+24.6)	52.6 (+33.8)	76.2 (+57.4)
	Max Length	143 (+92)	229 (+178)	271 (+220)

Table 2: Statistics for the discrete speech units produced for the Mboshi, with the difference between the produced and reference representation between parentheses. RAW is the original output from speech discretization models, +SIL is the result after silence post-processing. Other languages follow the same trend.

H-SHMM. While the SHMM significantly improves results over the HMM, it also suffers from an unrealistic assumption: it assumes that the phonetic subspace is the same for all languages. Yusuf et al. (2021) relax this assumption by proposing to adapt the subspace for each target language while learning the acoustic units. Formally, for a given language λ, the subspace and the acoustic units' parameters are constructed as in Equation 6-8, where the matrices $\mathbf{M}_0, \dots, \mathbf{M}_K$ and vectors $\mathbf{m}_0, \dots, \mathbf{m}_K$ represent some "template" phonetic subspace linearly combined by a language embedding $\boldsymbol{\alpha}^\lambda = [\alpha_1^\lambda, \alpha_2^\lambda, \dots, \alpha_K^\lambda]^\top$. The matrices $\mathbf{M}_i$ and the vectors $\mathbf{m}_i$ are estimated from labeled languages – from multilingual transcribed speech dataset for instance. The acoustic units' low-dimensional embeddings $\{\mathbf{e}_i\}$ and the language embedding $\boldsymbol{\alpha}$ are learned on the target (unlabeled) speech data. We refer to this model as the Hierarchical SHMM (H-SHMM).

$$\mathbf{W}^\lambda = \mathbf{M}_0 + \sum_{k=1}^{K} \alpha_k^\lambda \mathbf{M}_\mathbf{k} \tag{6}$$

$$\mathbf{b}^\lambda = \mathbf{m}_0 + \sum_{k=1}^{K} \alpha_k^\lambda \mathbf{m}_\mathbf{k} \tag{7}$$

$$\boldsymbol{\eta}^{\lambda,u} = f(\mathbf{W}^\lambda \cdot \mathbf{e}^{\lambda,u} + \mathbf{b}^\lambda) \tag{8}$$

Inference. For the three generative models, the posterior distribution is intractable and cannot be estimated. Instead, one seeks an approximate posterior $q(\{\boldsymbol{\eta}_i\}, \mathbf{z}) = q(\{\boldsymbol{\eta}_i\})q(\mathbf{z})$ that maximizes the variational lower-bound $\mathcal{L}[q]$. Concerning the estimation of $q(\mathbf{z})$, the *expectation* step is identical for all models and is achieved with a modified *forward-backward* algorithm described in Ondel et al. (2016). Estimation of $q(\boldsymbol{\eta})$, the *maximization* step, is model-specific and is described in Ondel et al. (2016) for the HMM, in Ondel et al. (2019) for SHMM models, and in Yusuf et al. (2021) for the H-SHMM model. Finally, the output of each system is obtained from a modified Viterbi algorithm that uses the expectation of the log-likelihoods with respect to $q(\{\boldsymbol{\eta}_i\})$, instead of point estimates.

Training. The models are trained with 4 Gaussians per HMM state and using 100 for the Dirichlet process' truncation parameter. SHMM and H-SHMM use an embedding size of 100, and H-SHMM models have a 6-dimensional language embedding. For the methods that use subspaces estimation (SHMM and H-SHMM), this estimation uses the following languages: French, German, Spanish, Polish from the Globalphone corpus (Schultz et al., 2013), as well as Amharic (Abate et al., 2005), Swahili (Gelas et al., 2012) and Wolof (Gauthier et al., 2016) from the ALFFA project (Besacier et al., 2015). We use 2-3 hours subsets of each, for a total of roughly 19 hours.

4. Experimental Setup

From the discrete speech units produced by the presented speech discretization models, we produce segmentation in the symbolic domain by using two UWS models. A final speech segmentation is then inferred using the units' time-stamps and evaluated by using the *Zero-Resource Challenge* 2017 evaluation suite, track 2 (Dunbar et al., 2017)[5]. We now detail the UWS models used in this work, which are trained with the same parameters from Godard et al. (2018). We also detail the datasets and the post-processing for the discrete speech discrete units.

Bayesian UWS approach (monolingual). Nonparametric Bayesian models (Goldwater, 2007; Johnson and Goldwater, 2009) are statistical approaches for UWS and morphological analysis, known to be robust in low-resource settings (Godard et al., 2016). In these models, words are generated by a unigram or bigram model over an infinite inventory, through the use of a Dirichlet process. In this work, we use the unigram model from *dpseg* (Goldwater et al., 2009)[6], which was shown to be superior to the bigram model in low-resource settings (Godard, 2019).

Neural UWS approach (bilingual). We follow the bilingual pipeline from Godard et al. (2018). The discrete speech units and their sentence-level translations are fed to an attention-based neural machine transla-

[5] Resources are available at http://zerospeech.com/2017

[6] Implementation available at http://homepages.inf.ed.ac.uk/sgwater/resources.html

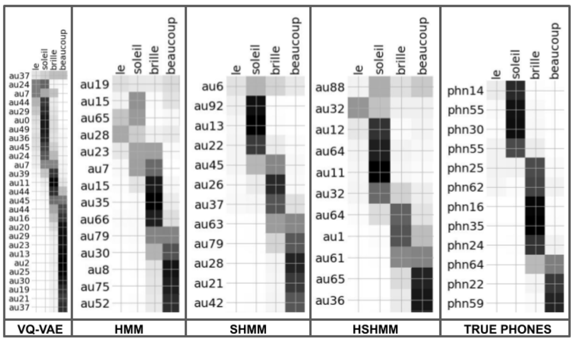


Figure 1: Heatmaps for the soft-alignment probability matrices generated by the neural UWS models (bilingual) trained on different discrete speech units, for the same French-Mboshi sentence. The darker the square, the higher the pair probability. The rows present the automatically generated units from the different discretization models, informed in the bottom.

tion system that produces soft-alignment probability matrices between source and target sequences. For each sentence pair, its matrix is used for clustering together (segmenting) neighboring phones whose alignment distribution peaks at the same source word. Examples of these matrices are provided in Figure 1. We refer to this model as *neural*.

Datasets. We use the Mboshi-French parallel corpus (MB-FR) (Godard et al., 2018), which is a 5,130 sentence corpus from the language documentation process of Mboshi (Bantu C25), an oral language spoken in Congo-Brazzaville. We also report results using an extract from the MaSS corpus (Boito et al., 2020), a multilingual speech-to-speech and speech-to-text dataset. We use the down-sampling from Boito et al. (2020), which results in 5,324 aligned sentences. We exclude French and Spanish, as these languages are present in the subspace prior from SHMM and H-SHMM models, and we exclude English as it was used as to tune the hyperparameters of the subspace models and the VQ-VAE. We also exclude Basque, as the sequences produced were too long for UWS training. The final set of languages is: Finnish (FI), Hungarian (HU), Romanian (RO) and Russian (RU). In all cases, the French (FR) translations are used as supervision for the neural UWS approach. Statistics are presented in Table 1.

Discrete Speech Units Post-processing. We experiment with reducing the representation by removing units predicted in silence windows. For this, we use the gold references' silence annotations. Removing these allow us to focus the investigation on the quality of the units generated in *relevant* portions of the speech. We see in Table 2 that removing windows that we *know* correspond to silence considerably reduces the number of units generated by all models. Before UWS evaluation, the silence windows are reintroduced to ensure that their segmentation boundaries are taken into account. This approach is justified because a silence detector is an inexpensive resource to obtain. For instance, popular software such as Praat (Boersma, 2006) are able to handle this task in any language. Figure 2 exemplifies the discrete speech units discovered by the models before applying this post-processing.

		dpseg		*neural*	
		RAW	+SIL	RAW	+SIL
1	**HMM**	32.4	59.9	35.1	61.2
2	**SHMM**	43.7	**61.4**	41.4	**64.7**
3	**H-SHMM**	**45.3**	**61.4**	**44.8**	63.9
4	**VQ-VAE**	39.0	52.7	32.1	60.1
5	**VQ-W2V-V16**	37.4	52.2	32.0	50.6
6	**VQ-W2V-V36**	-	48.0	-	49.8
7	**True Phones**	-	**77.1**	-	74.5

Table 3: UWS Boundary F-scores for the MB-FR dataset.

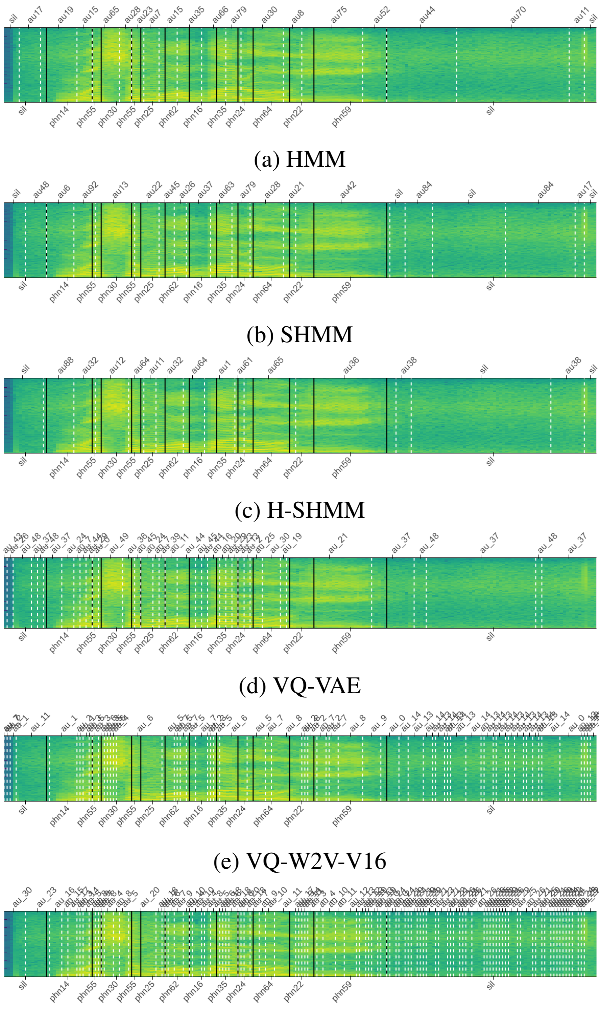
5. Experiments

We first present our results for the MB-FR dataset, the language which corresponds to the true low-resource scenario that we are interested in. Table 3 presents UWS Boundary F-scores for UWS models (dpseg and neural) trained using different discrete speech units for the MB-FR dataset. We include results for both the direct output (RAW) and the post-processed version (+SIL). The RAW VQ-W2V-V36 is not included as its output sequences were excessively large for training our UWS models (Table 2).

We observe that in all cases, post-processing the discrete speech units with the silence information (+SIL) creates *easier* representations for the UWS task. We believe this is due to the considerable reduction in average length of the sequences (Table 2). For Bayesian models, we also observe a reduction in the number of units, meaning that some units were modelling silence windows, even though these models already produce an independent token for silence, which we remove before UWS training.

Looking at the results for UWS models trained using the output of VQ-based models (rows 4-6), we see that the best segmentation result is achieved using the one with the smallest average sequence length (VQ-VAE). In general, we believe that all VQ-based models underperform due to the excessively long sequences produced, which are challenging for UWS. Figure 2 illustrates this difference in representation length, by presenting the discrete speech units produced by Bayesian and neural models for a given utterance: the latter produce considerably more units.

Overall, we find that UWS models trained using the discrete speech units from Bayesian models produce better segmentation, with models trained with SHMM and H-SHMM presenting the best results. In Yusuf et al. (2021) both systems showed competitive results for the AUD task. A noticeable difference between these two models is the compression level: H-SHMM



(a) HMM

(b) SHMM

(c) H-SHMM

(d) VQ-VAE

(e) VQ-W2V-V16

(f) VQ-W2V-V36

Figure 2: Speech discrete units produced by the five models for the same Mboshi sentence. Black lines denote the true boundaries, while dashed white lines denote the discovered units boundaries. For each example, discrete speech units (top) and reference (bottom).

	dpseg				*neural*			
	FI	**HU**	**RO**	**RU**	**FI**	**HU**	**RO**	**RU**
HMM	45.6	49.9	53.5	47.1	53.4	51.2	56.6	54.9
SHMM	49.0	52.3	53.5	50.5	56.0	**53.9**	57.7	**57.7**
H-SHMM	50.5	52.9	58.0	52.9	**56.1**	53.3	**59.6**	56.0
True Phones	<u>87.1</u>	<u>83.3</u>	<u>88.0</u>	<u>85.9</u>	68.4	63.4	75.7	68.4

Table 4: UWS Boundary F-scores for the MaSS dataset using Bayesian models (+SIL only). Best UWS results from speech discrete units (**bold**) and from true phones (<u>underlined</u>) are highlighted.

uses 27 fewer units than SHMM. Regarding type retrieval, the models scored 12.1% (SHMM), 10.7% (H-SHMM), and 31% (topline). We also find that SHMM models produced more types and fewer tokens, reaching a higher Type-Token Ratio (0.63) compared to H-SHMM (0.55).

Focusing on the generalization of the presented speech discretization models, we trained our models using four languages from the MaSS dataset. We observed that due to the considerably larger average length of the sentences (Table 1), the VQ-based models produced sequences which we were unable to directly apply to UWS training. This again highlights that these models need some constraining, or post-processing, in order to be directly exploitable for UWS. Focusing on the Bayesian models, which performed the best for generating exploitable discrete speech units for UWS in low-resource settings, Table 4 present UWS results. We omit results for RAW, as we observe the same trend from Table 3. Looking at the results for the four languages, we again observe competitive results for SHMM and H-SHMM models, illustrating that these approaches generalize well to different languages.

Comparing the UWS results present in Table 3 (Mboshi) and Table 4 (languages from MaSS), we notice overall lower results for the languages from the MaSS dataset (best result: 59.6) compared to Mboshi (best result: 64.7). We believe this is due to the MaSS data coming from read text, in which the utterances correspond to verses that are consistently longer than sentences (Table 1). This results in a more challenging setting for UWS and explains the lower results. Lastly, our results over five languages show that the neural UWS model produces better segmentation results from discrete speech units than dpseg, which in turn performs the best with the true phones (topline). This confirms the trend observed by (Godard et al., 2018). The neural UWS models have the advantage of their word-level aligned translations for grounding the segmentation process, which might be attenuating the difficulty of the task in this noisier scenario, with longer sequences and more units. Moreover, a benefit of these models is the potentially exploitable bilingual alignment discovered during training. Boito et al. (2019) used these alignments for filtering the generated vocabulary, increasing type retrieval.

6. Conclusion

In this paper we compared five methods for speech discretization, two neural models (VQ-VAE, VQ-WAV2VEC), and three Bayesian approaches (HMM, SHMM, H-SHMM), with respect to their performance serving as direct input to the task of unsupervised word segmentation (UWS) in low-resource settings. Our motivation for such a study lies in the need of processing oral and low-resource languages, for which obtaining transcriptions is a known bottleneck (Brinckmann, 2009).

In our UWS setting, and using five different languages (Finnish, Hungarian, Mboshi, Romanian and Russian), we find that VQ-based methods are not a good fit for our pipeline, as they output very long and inconsistent sequences, which are difficult to treat. This was also recently observed in Kamper and van Niekerk (2021).

In contrast to that, the Bayesian SHMM and H-SHMM models perform the best, as they produced concise yet highly exploitable representations from just few hours of speech. We believe this difference in performance is due to HMM-based models explicitly performing acoustic unit discovery. This means the discretization produced by them aims not only to summarize the speech signal, but to closely match the language's phonology. Moreover, the subspace estimation performed by both SHMM and H-SHMM, might also play a significant role. This is because these models are able to learn from an additional 19 hours of data in different languages. The other models (HMM and VQ-based models) do not have access to any form of pretraining or prior.

Finally, comparing the neural and Bayesian UWS approaches, we notice that the neural model is competitive in the *noisier* setting, reaching better UWS boundary scores working with the output of speech discretization models. The Bayesian model is however better at segmenting true phones (topline scenario). Concluding, this work updates Godard et al. (2018) by using more recent speech discretization models, and presenting better UWS results for Mboshi.

7. Bibliographical References

Anastasopoulos, A. and Chiang, D. (2018). Leveraging translations for speech transcription in low-resource settings. In *Proc. Interspeech 2018*, pages 1279–1283.

Baevski, A., Auli, M., and Mohamed, A. (2019). Effectiveness of self-supervised pre-training for speech recognition. *arXiv preprint arXiv:1911.03912*.

Baevski, A., Schneider, S., and Auli, M. (2020a). vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations (ICLR)*.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020b). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Bartels, C., Wang, W., Mitra, V., Richey, C., Kathol, A., Vergyri, D., Bratt, H., and Hung, C. (2016). Toward human-assisted lexical unit discovery without text resources. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*, pages 64–70. IEEE.

Bengio, Y., Léonard, N., and Courville, A. (2013). Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv:1308.3432*.

Bérard, A., Pietquin, O., Servan, C., and Besacier, L. (2016). Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on End-to-End Learning for Speech and Audio Processing*.

Besacier, L., Zhou, B., and Gao, Y. (2006). Towards speech translation of non written languages. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 222–225. IEEE.

Besacier, L., Gauthier, E., Mangeot, M., Bretier, P., Bagshaw, P., Rosec, O., Moudenc, T., Pellegrino, F., Voisin, S., Marsico, E., et al. (2015). Speech technologies for african languages: example of a multilingual calculator for education. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Bird, S. (2021). Sparse transcription. *Computational Linguistics*.

Boersma, P. (2006). Praat: doing phonetics by computer. *http://www. praat. org/*.

Boito, M. Z., Bérard, A., Villavicencio, A., and Besacier, L. (2017). Unwritten languages demand attention too! word discovery with encoder-decoder models. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 458–465. IEEE.

Boito, M. Z., Villavicencio, A., and Besacier, L. (2019). Empirical evaluation of sequence-to-sequence models for word discovery in low-resource settings. In *Proc. Interspeech 2019*, pages 2688–2692.

Boito, M. Z., Villavicencio, A., and Besacier, L. (2020). Investigating language impact in bilingual approaches for computational language documentation. In *Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020)*.

Brinckmann, C. (2009). Transcription bottleneck of speech corpus exploitation. In *Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics (LULCL II). Combining efforts to foster computational support of minority languages*.

Chorowski, J., Weiss, R. J., Bengio, S., and van den Oord, A. (2019). Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM transactions on audio, speech, and language processing*, 27(12):2041–2053.

Dunbar, E., Cao, X. N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X., and Dupoux, E. (2017). The zero resource speech challenge 2017. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 323–330. IEEE.

Dunbar, E., Algayres, R., Karadayi, J., Bernard, M., Benjumea, J., Cao, X.-N., Miskic, L., Dugrain, C., Ondel, L., Black, A. W., Besacier, L., Sakti, S., and Dupoux, E. (2019). The Zero Resource Speech Challenge 2019: TTS Without T. In *Proc. Interspeech 2019*, pages 1088–1092.

Dunbar, E., Karadayi, J., Bernard, M., Cao, X.-N., Algayres, R., Ondel, L., Besacier, L., Sakti, S., and Dupoux, E. (2020). The Zero Resource Speech Challenge 2020: Discovering Discrete Subword and Word Units. In *Proc. Interspeech 2020*, pages 4831–4835.

Duong, L., Anastasopoulos, A., Chiang, D., Bird, S., and Cohn, T. (2016). An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 949–959.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*.

Foley, B., Arnold, J., Coto-Solano, R., Durantin, G., Mark, E., van Esch, D., Heath, S., Kratochvil, F., Maxwell-Smith, Z., Nash, D., et al. (2018). Building speech recognition systems for language documentation: the coedl endangered language pipeline and inference system (elpis).

Glass, J. (2012). Towards unsupervised speech processing. In *Information Science, Signal Processing and their Applications (ISSPA)*. IEEE.

Godard, P., Adda, G., Adda-Decker, M., Allauzen, A., Besacier, L., Bonneau-Maynard, H., Kouarata, G.-N., Löser, K., Rialland, A., and Yvon, F. (2016). Preliminary experiments on unsupervised word discovery in mboshi. In *Proc. Interspeech*.

Godard, P., Boito, M. Z., Ondel, L., Bérard, A., Yvon, F., Villavicencio, A., and Besacier, L. (2018). Unsupervised word segmentation from speech with attention. In *Proc. Interspeech 2018*, pages 2678–2682.

Godard, P. (2019). *Unsupervised word discovery for computational language documentation*. Ph.D. thesis, Paris Saclay.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Goldwater, S. J. (2007). *Nonparametric Bayesian models of lexical acquisition*. Ph.D. thesis, Citeseer.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Jang, E., Gu, S., and Poole, B. (2017). Categorical reparameterization with gumbel-softmax. In *ICLR*.

Jansen, A., Dupoux, E., Goldwater, S., Johnson, M., Khudanpur, S., Church, K., Feldman, N., Hermansky, H., Metze, F., Rose, R., et al. (2013). A summary of the 2012 jhu clsp workshop on zero resource speech technologies and models of early language acquisition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8111–8115. IEEE.

Johnson, M. and Goldwater, S. (2009). Improving nonparameteric bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proc. NAACL-HLT*, pages 317–325. Association for Computational Linguistics.

Kamper, H. and van Niekerk, B. (2021). Towards Unsupervised Phone and Word Segmentation Using Self-Supervised Vector-Quantized Neural Networks. In *Proc. Interspeech 2021*, pages 1539–1543.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Yoshua Bengio et al., editors, *ICLR 2015, Conference Track Proceedings*.

Lignos, C. and Yang, C. (2010). Recession segmentation: simpler online word segmentation using limited resources. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 88–97.

Liu, A. T., Yang, S.-w., Chi, P.-H., Hsu, P.-c., and Lee, H.-y. (2020). Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423. IEEE.

Michaud, A., Adams, O., Cohn, T. A., Neubig, G., and Guillaume, S. (2018). Integrating automatic transcription into the language documentation workflow: Experiments with na data and the persephone toolkit.

Ondel, L., Burget, L., and Černockỳ, J. (2016). Variational inference for acoustic unit discovery. *Procedia Computer Science*, 81:80–86.

Ondel, L., Vydana, H. K., Burget, L., and Černocký, J. (2019). Bayesian Subspace Hidden Markov Model for Acoustic Unit Discovery. In *Interspeech*, pages 261–265.

Ondel, L., Yusuf, B., Burget, L., and Saraclar, M. (2022). Non-parametric bayesian subspace models for acoustic unit discovery. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, pages 3465–3469.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Strunk, J., Schiel, F., Seifart, F., et al. (2014). Untrained forced alignment of transcriptions and audio for language documentation corpora using webmaus. In *LREC*, pages 3940–3947.

Tjandra, A., Sakti, S., and Nakamura, S. (2019). Speech-to-speech translation between untranscribed unknown languages. *arXiv:1910.00795*.

van den Oord, A., Vinyals, O., and kavukcuoglu, k. (2017). Neural discrete representation learning. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6306–6315. Curran Associates, Inc.

Versteegh, M., Anguera, X., Jansen, A., and Dupoux, E. (2016). The zero resource speech challenge 2015: Proposed approaches and results. *Procedia Computer Science*, 81:67–72.

Wang, W., Tang, Q., and Livescu, K. (2020). Unsupervised pre-training of bidirectional speech encoders via masked reconstruction. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6889–6893. IEEE.

Yusuf, B., Ondel, L., Burget, L., Černockỳ, J., and Saraclar, M. (2021). A hierarchical subspace model for language-attuned acoustic unit discovery. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3710–3714. IEEE.

8. Language Resource References

Abate, S. T., Menzel, W., and Tafila, B. (2005). An amharic speech corpus for large vocabulary continuous speech recognition. In *Ninth European Conference on Speech Communication and Technology*.

Boito, M. Z., Havard, W. N., Garnerin, M., Ferrand, É. L., and Besacier, L. (2020). Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible. *Language Resources and Evaluation Conference (LREC)*.

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

Gelas, H., Besacier, L., and Pellegrino, F. (2012). Developments of Swahili resources for an automatic speech recognition system. In *SLTU - Workshop on Spoken Language Technologies for Under-Resourced Languages*, Afrique Du Sud.

Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G.-N., Lamel, L., Maynard, H., Mueller, M., Rialland, A., Stueker, S., Yvon, F., and Boito, M. Z. (2018). A very low resource language speech corpus for computational language documentation experiments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Schultz, T., Vu, N. T., and Schlippe, T. (2013). Globalphone: A multilingual text & speech database in 20 languages. In *International Conference on Acoustics, Speech and Signal Processing*. IEEE.