Self-supervised Semantic-driven Phoneme Discovery for Zero-resource Speech Recognition

Liming Wang[1], **Siyuan Feng**[2], **Mark Hasegawa-Johnson**[1], **Chang D. Yoo**[3]
[1]Department of Electrical and Computer Engineering,
University of Illinois at Urbana-Champaign
[2]Multimedia Computing Group, Delft University of Technology
[3]Artificial Intelligence and Machine Learning Lab, KAIST
{lwang114,jhasegaw}@illinois.edu, s.feng@tudelft.nl, cd_yoo@kaist.ac.kr

Abstract

Phonemes are defined by their relationship to words: changing a phoneme changes the word. Learning a phoneme inventory with little supervision has been a longstanding challenge with important applications to under-resourced speech technology. In this paper, we bridge the gap between the linguistic and statistical definition of phonemes and propose a novel neural discrete representation learning model for self-supervised learning of phoneme inventory with raw speech and word labels. Given the availability of phoneme segmentation and some mild conditions, we prove that the phoneme inventory learned by our approach converges to the true one with an exponentially low error rate. Moreover, in experiments on TIMIT and Mboshi benchmarks, our approach consistently learns a better phoneme-level representation and achieves a lower error rate in a zero-resource phoneme recognition task than previous state-of-the-art self-supervised representation learning algorithms.

1 Introduction

Thanks to recent developments in self-supervised speech representation learning (van den Oord et al., 2017, 2019; Chorowski et al., 2019; Baevski et al., 2020), there is new hope for the development of speech processing systems without the need for full textual transcriptions. Supervised speech processing systems for tasks such as automatic speech recognition (ASR) rely on a large amount of textual transcriptions, but self-supervised systems can be applied to under-resourced languages in which such annotation is either scarce or unavailable. A key task of the self-supervised system is to learn a discrete representation. While it is possible to discretize the speech solely on the basis of its acoustic properties, a more desirable discrete representation would serve as a bridge from the continuous acoustic signal toward higher-level linguistic structures such as syntax and semantics. Such a representation would make it possible to repurpose algorithms developed for written languages so that they could be used for unwritten languages in tasks such as speech translation and spoken language understanding. Words are the obvious choice for a discrete, semantic-driven speech representation, but a practical speech understanding system needs at least thousands of words; learning them in an unsupervised manner may be challenging. Phonemes may be a more learnable representation. According to the standard linguistic definition, phonemes are closely linked to words:

Definition 1. *(Linguistic definition of phonemes (Swadesh, 1934)) Phonemes are the smallest units in speech such that given a correct native word, the replacement of one or more phonemes by other phonemes (capable of occurring in the same position) results in a native word other than that intended, or a native-like nonsense word.*

For example, the sentences "he *th*inks" and "he *s*inks" differ by exactly one phoneme but have very different meaning. The optimal compactness of a phoneme inventory as specified in the definition leads to three advantages. First, learning phonemes requires lower sample complexity than learning words since the number of distinct phonemes is much smaller than the number of distinct words in a language. Second, the phonemes are much more abundant and more balanced in classes than words within a speech corpus, which makes sample complexity less of an issue when learning phonemes. Third, phonemes are more generalizable in the sense that knowing the phoneme inventory allows the learner to memorize previously unseen words as sequences of phonemes, and, having memorized them, to begin seeking clues to their meaning. Motivated by the semantic-driven definition of phonemes, we formulate the problem of learning a phoneme inventory as a self-supervised learning problem, where a small amount of semantic su-

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8027 - 8047

pervision is available. The required supervision specifies which acoustic segments are instances of the same word, and which are instances of different words. Such supervision might be acquired in a naturalistic setting by asking native speakers to name objects in a set of standardized images, as is commonly done in primary education classrooms, or by asking for the translations of common words in a second language, a common baseline approach in dialectology and historical linguistics (Swadesh, 1952). Our contributions are threefold: (1) we propose a computationally tractable definition of phoneme that is almost equivalent to the linguistic definition. (2) We propose a finite-sample objective function for learning phoneme-level units and prove that when the phoneme segmentation is available and under mild conditions, the empirical risk minimizer (ERM) of this objective will find the correct phoneme inventory with exponentially low error rate. (3) We propose a novel neural network called *information quantizer* to optimize the proposed objective, which achieve state-of-the-art results in the phoneme inventory discovery task on the TIMIT and low-resourced Mboshi benchmarks with much less training data than previous approaches.

2 Related works

Due to the challenge of learning phonemes, early works on unsupervised speech representation learning (Park and Glass, 2005; Lee and Glass, 2012; Ondel et al., 2016) focus on learning speech segments sharing similar acoustic properties, or *phones*, without taking into account the meaning of the speech they are part of. There are two main approaches in this direction. One approach is to learn discrete phone-like units without *any* textual labels by modeling phone labels of the speech segments as latent variables. In particular, (Park and Glass, 2005; Jansen et al., 2010) first detect segments with recurring patterns in the speech corpus followed by graph clustering using the similarity graph formed by the segments. (Lee and Glass, 2012; Ondel et al., 2016; Kamper et al., 2016) develop probabilistic graphical models to jointly segment and cluster speech into phone-like segments. An extension to the latent variable approach is to introduce additional latent variables such as speaker identity (Ondel et al., 2019) or language identity (Yusuf et al., 2020) and develop mechanisms to disentangle these variables.

With the advance of deep learning, neural network models have also been proposed to learn unsupervised phone-level representation either by first learning a continuous representation (Chung et al., 2019; Feng et al., 2019; Nguyen et al., 2020) followed by off-line clustering, or by learning a discrete representation end-to-end with Gumbel softmax (Eloff et al., 2019b; Baevski et al., 2020) or vector-quantized variational autoencoder (VQ-VAE) (van den Oord et al., 2017; Chorowski et al., 2019; Baevski et al., 2019). However, codebooks learned by the neural approaches tend to be much larger than the number of phonemes (Baevski et al., 2020), leading to low scores in standard phoneme discovery metrics. The second approach utilizes weak supervision such as noisy phone labels predicted by a supervised, multilingual ASR system trained on other languages. Along this direction, early works (Schultz and Waibel, 1998; Lööf et al., 2009; Swietojanski et al., 2012) have showed that phonetic knowledge gained from one language can be leveraged to develop ASR systems for another language using an HMM-based or DNN-HMM hybrid approach. Instead of using phone labels, (Stuker et al., 2003) explores the use of articulatory features as supervision for the multilingual ASR. Recently, (Żelasko et al., 2020a,b; Feng et al., 2021a) systematically study the performance of zero-shot crosslingual ASR on 13 languages trained with international phonetic alphabet (IPA) tokens and found that the system tends to perform poorly on unseen languages. Instead, (Feng et al., 2021b) is able to discover phone-like units by clustering bottleneck features (BNF) from a factorized time-delay neural network (TDNN-f) trained with phone labels predicted by a crosslingual ASR (Feng et al., 2021a).

Several works have since shifted focus toward the more challenging phoneme discovery problem by formulating it as a self-supervised learning problem where the semantics of the speech are known, such as from translation, phoneme-level language models or other sensory modalities such as vision. (Jansen, 2013) has studied the use of pairwise word identity labels for training phoneme discovery models based on Gaussian mixture models (GMM); (Harwath and Glass, 2019) analyzes the hidden layers of a two-branch neural network trained to retrieve spoken captions with semantically related images and finds strong correlation between segment representation and

phoneme boundaries. (Harwath et al., 2020) adds hierarchical vector quantization (VQ) layers in the same retrieval network and is able to find a much smaller codebook than the unsupervised neural approach (Baevski et al., 2020), and achieve high correlation with the phoneme inventory. (Godard et al., 2018; Boito et al., 2019) has studied the possibility of learning semantic units using an attention-based speech-to-text translation system, though the units appear to correlate more with words. Works on unsupervised speech recognition (Chen et al., 2019) attempt to learn to recognize phonemes by leveraging the semantic information from a phoneme language model unpaired with the speech, typically by matching the empirical prior and posterior distributions of phonemes either using cross entropy (Yeh et al., 2019) or adversarial loss (Chen et al., 2019; Baevski et al., 2021). Such models, however, have a slightly different objective as they assume knowledge about the phoneme inventory of the language and instead tries to find the alignment between the speech and phonemes, rather than induce the phoneme inventory from scratch.

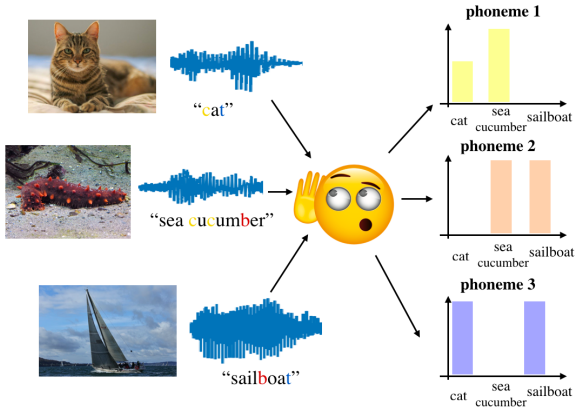


Figure 1: Illustration of semantic-driven phoneme discovery

3 Semantic-driven Phoneme Discovery

3.1 Notation

Throughout the paper, we use $\mathbb{P}\{\cdot\}$ to denote probability. We use capital letters to denote random variables and lower-case letters to represent samples of random variables. We use $P_X := \mathbb{P}\{X = x\}$ to denote both probability mass and density functions of random variable X, depending on whether it is continuous or discrete. Further, denote $P_{Y|X}(y|x) := \mathbb{P}\{Y = y|X = x\}$ as the true conditional probability distribution of random variable $Y = y$ given random variable $X = x$. The probability simplex in $\mathbb{R}^d$ is denoted as Δ^d.

3.2 Statistical Definition of Phonemes

The linguistic definition of phonemes can be rephrased as follows. Define $\mathbb{X}$ to be the set of all physical acoustic segments that can ever be produced as instances of the phonemes of a given language. Definition 1 can be phrased as follows: Two sequences of segments $\boldsymbol{x} = [x_1, \cdots, x_T]$ and $\boldsymbol{x}' = [x_{1:t-1}, x'_t, x_{t+1:T}]$, differing only in that $x'_t \neq x_t$, are instances of different words, $y' \neq y$, if and only if x'_t and x_t are instances of different phonemes. In order to design effective algorithms, we will work with a relaxation of this definition, which we call the statistical definition of phonemes.

Definition 2. *(Statistical definition of phonemes) Let $\mathbb{X}$ be the set of all speech segments in a language, and let X be a random vector taking values in $\mathbb{X}$ and Y be a random variable representing the word of which X is one segment. The phoneme inventory of a language is the minimal partition $\mathbb{Z} = \{\mathbb{Z}_1, \cdots, \mathbb{Z}_K\}$ of $\mathbb{X}$ (i.e., $\mathbb{X} = \cup_{k=1}^K \mathbb{Z}_k, \mathbb{Z}_j \cap \mathbb{Z}_k = \emptyset, \forall 1 \leq j, k \leq K$), such that if a speech segment pair $(x, x') \in \mathbb{X}^2$ satisfies $(x, x') \in \mathbb{Z}_k^2$ for some $k \in \{1, \cdots, K\}$, then their conditional distributions satisfy*

$$P_{Y|X=x} = P_{Y|X=x'}. \tag{1}$$

In other words, given only the knowledge that two acoustic sequences contain instances of the same phoneme, the resulting conditional distributions across possible word labels are the same.

The fundamental intuition of Definition 2 is that different phonemes have different distributions across the words of the language. Two instances of the same phoneme, x and x', might have different likelihoods $P_{X=x|Y}$ and $P_{X=x'|Y}$, e.g., because of allophony; but their posteriors $P_{Y|X=x}$ and $P_{Y|X=x'}$ cannot be different without violating Definition 1. The relationship between Definition 1 and Definition 2 is given by the following proposition, whose proof is in Appendix A.3.

Proposition 1. *Let $\mathbb{Z} = \cup_{k=1}^K \mathbb{Z}_k$ be a partition of $\mathbb{X}$. If, for all possible $\{P_{Y|X=x_s}\}_{s \neq t}$, for any spoken word $\boldsymbol{x} = [x_1, \cdots, x_T]$, and for any segment pairs $(x_t, x'_t) \in \mathbb{Z}_k^2, k \in \{1, \cdots, K\}$, changing x_t*

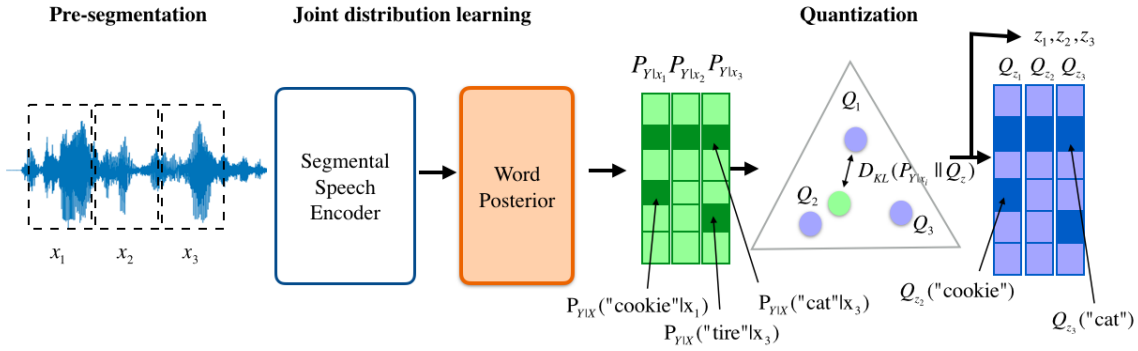


Figure 2: Network architecture of information quantizer

to x_t' does not alter the identity of the word, i.e.,

$$\arg\max_y P_{Y|X_{1:T}}(y|x_{1:t-1}, x_t', x_{t+1:T}) = \arg\max_y P_{Y|X_{1:T}}(y|\boldsymbol{x}), \quad (2)$$

but for any segment pairs $x_t \in \mathbb{Z}_k, x_t'' \in \mathbb{Z}_l$ for $k \neq l$, changing x_t to x_t' alters the identity of the word, i.e.,

$$\arg\max_y P_{Y|X_{1:T}}(y|x_{1:t-1}, x_t'', x_{t+1:T}) \neq \arg\max_y P_{Y|X_{1:T}}(y|\boldsymbol{x}), \quad (3)$$

then $\mathbb{Z}$ is a phoneme inventory from Definition 2.

Define the **phoneme assignment function** $z : \mathbb{X} \to \{1, \cdots, K\}$ such that $z(x) = k$ if $x \in \mathbb{Z}_k$. Suppose a segment X is randomly chosen from $\mathbb{X}$ with probability distribution P_X and its phoneme label is another random variable $Z := z(X)$, then by Definition 2, for any pair $x, x' \in \mathbb{X}$ such that $z(x) = z(x')$, we have $P_{Y|X=x} = P_{Y|X=x'} = P_{Y|Z=z(x)}$. The phoneme inventory is thereby completely characterized by the phoneme label function $z(\cdot)$ as well as the set of distributions associated with each class $P_{Y|Z}$.

3.3 Problem Formulation

Let $z(\cdot)$ be the phoneme assignment function from Definition 2 and assume the size of the phoneme inventory is known to be K.

Given a training set $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$, where each $x^{(i)}$ is an acoustic segment extracted from a spoken word, and each $y^{(i)} \in \mathbb{Y}$ is the corresponding word label, a semantic-driven phoneme discovery (SPD) system tries to find an assignment function that minimizes the **token error rate (TER)**:

$$P_{\text{TER}}(\hat{z}) := \min_{\pi \in \Pi} \mathbb{P}\{z(X) \neq \pi(\hat{z}(X))\}, \quad (4)$$

where Π is the set of all *permutations* of length K, which is used because the problem is unsupervised and $z(\cdot)$ is not available during training. An assignment function $\hat{z}$ is said to achieve **exact discovery** if $P_{\text{TER}}(\hat{z}) = 0$. It can be easily shown that TER is equivalent to standard evaluation metrics for phoneme discovery such as normalized mutual information (NMI) (Yusuf et al., 2020; Harwath et al., 2020; Feng et al., 2021b) and token F1 (Dunbar et al., 2017), as presented in Appendix A.2. Thus, to provide guarantees for NMI and token F1, it suffices to provide a guarantee for TER.

4 Information Quantizer

We solve the SPD problem using a novel type of neural network called an information quantizer (IQ), depicted in Figure 2. An IQ $(\theta, q) \in \Theta \times \mathbb{Q}_K$ consists of four main components: A pre-segmentation network, a speech encoder $e_{\theta_1}(\cdot)$, a word posterior $c_{\theta_2}(\cdot)$ and a quantizer $q : \Delta^{|\mathbb{Y}|} \to \mathbb{C} = \{Q_1, \cdots, Q_K\}$, where $[\theta_1, \theta_2] = \theta$ and $\mathbb{C}$ is the **distribution codebook** and Q_k's are called the **code distributions** of q.

4.1 Phoneme inventory discovery with IQ

IQ performs phoneme discovery in three stages. The pre-segmentation stage takes a raw speech waveform as input and extracts phoneme-level segments $\boldsymbol{x} = [x_1, \cdots, x_T]$ in a self-supervised fashion (Kreuk et al., 2020). Afterwards, in the joint distribution learning stage, the speech encoder extracts phoneme-level representations $e_{\theta_1}(\boldsymbol{x}) = [e_{\theta_1}(x_1), \cdots, e_{\theta_1}(x_T)]$ before passing them into the word posterior network to estimate the distribution of word labels, Y, given the presence in the word of acoustic phonetic segment $X = x$:

$$P^{\theta}_{Y|X=x_t} = c_{\theta_2}(e_{\theta_1}(x_t)), 1 \leq t \leq T. \quad (5)$$

Note that it is crucial that no recurrent connection exists between segments since our goal is to learn the probability of a word label given the presence of one phoneme segment. Finally, in the quantization stage, the quantizer creates the phoneme inventory by assigning each segment x_t an integer index via **codeword assignment function** $\hat{z}(x_t)$ such that $\hat{z}(x_t) = k$ if $q(P^{\theta}_{Y|X=x_t}) = Q_k$.

4.2 Training

The loss function that IQ minimizes has two goals: learn a good estimator for the conditional distribution $P_{Y|X}$ and learn a good quantization function $q(\cdot)$. The first goal is achieved by minimizing the cross entropy loss:

$$\mathcal{L}_{\text{CE}}(P_n, \theta) := -\frac{1}{n}\sum_{i=1}^{n} \log P^{\theta}_{Y|X}(y^{(i)}|x^{(i)}), \quad (6)$$

where P_n is the empirical joint distribution. The second goal is achieved by minimizing the KL-divergence between the estimated conditional distribution before and after quantization:

$$\mathcal{L}_{\text{Q}}(\tilde{P}_n, \theta, q) := \frac{1}{n}\sum_{i=1}^{n} D_{\text{KL}}(P^{\theta}_{Y|X=x^{(i)}} || q(P^{\theta}_{Y|X=x^{(i)}})), \quad (7)$$

where

$$\tilde{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x^{(i)}} P^{\theta}_{Y|X=x^{(i)}}$$

is the *smoothed* version of the empirical distribution. The final loss function of IQ for SPD is then:

$$\mathcal{L}_{\text{IQ}}(P_n, \theta, q) := \mathcal{L}_{\text{CE}}(P_n, \theta) + \lambda \mathcal{L}_{\text{Q}}\left(\tilde{P}_n, \theta, q\right), \quad (P_1)$$

where $\lambda > 0$ is some hyperparameter set to approximately 1 for most experiments. Further, we restrict q to be **nearest-neighbor** so that:

$$q(P) = \underset{Q_k : 1 \le k \le K}{\arg\min} D_{\text{KL}}(P||Q_k). \quad (8)$$

This restriction does not increase the loss (P_1) and serves as a regularization during phoneme discovery, as shown in Appendix A.3.

4.3 Theoretical Guarantee

We show that when the phoneme segmentation is available and under mild assumption, IQ is able to achieve exact discovery of phoneme inventory. First, let us state the main assumptions of the paper.

Assumption 1. *(boundedness of the density ratio) There exist universal constants $C_l < C_u$ such that $\forall\theta \in \Theta, \forall q \in \mathbb{Q}_K, \forall (x, y) \in \mathbb{X} \times \mathbb{Y}, \log \frac{P_{Y|X}(y|x)}{P^{\theta}_{Y|X}(y|x)} \in [C_l, C_u], \log \frac{P_{Y|X}(y|x)}{q(P^{\theta}_{Y|X}(y|x))} \in [C_l, C_u]$.*

Assumption 2. *(log-smoothness of the density ratio) There exists $\rho > 0$ such that $\forall\theta_1, \theta_2 \in \Theta, x, y \in \mathbb{X} \times \mathbb{Y}, \left|\log \frac{P^{\theta_1}_{Y|X}(y|x)}{P^{\theta_2}_{Y|X}(y|x)}\right| \le \rho\|\theta_1 - \theta_2\|$.*

Assumption 3. *(realizability) There exists a nonempty subset $\Theta^* \subset \Theta$ such that $P^{\theta}_{Y|X} = P_{Y|X}, \forall\theta \in \Theta^*$.*

Assumption 4. *The true prior of the phoneme inventory is known to be $P_Z(z) = \frac{1}{K}, 1 \le z \le K$.*

The first two assumptions are similar to the ones in (Tsai et al., 2020). Assumption 3 assumes that the true probability measure is within the function class, which combined with Assumption 1 requires the true distribution to share the same support as the estimated one. However, such assumption can be relaxed so that $D_{\text{KL}}(P^{\theta^*}_{Y|X}||P_{Y|X}) \le \nu, \forall\theta^* \in \Theta^*$ for some small enough $\nu > 0$, which does not affect the essential idea behind our analysis and can be achieved by some rich class of universal approximators such as neural networks (Hornik et al., 1989). The last assumption ensures the inventory to be identifiable by assuming knowledge of the prior of the phoneme inventory.

Next, we will state the theoretical guarantee before giving some intuitive explanation.

Theorem 1. *Given Assumption 1-4, let the information quantizer $(\hat{\theta}, \hat{q})$ with assignment function $\hat{z}$ be an empirical risk minimizer (ERM) of (P_1):*

$$\mathcal{L}_{IQ}(P_n, \hat{\theta}, \hat{q}) = \min_{\theta\in\Theta, q\in\mathbb{Q}_K} \mathcal{L}_{IQ}(P_n, \theta, q). \quad (9)$$

For any $\delta \in (0, 1]$, with probability at least $1 - \delta$, the cluster assignment function $\hat{z}$ of the ERM information quantizer $\hat{q}$ achieves $P_{TER}(\hat{z}) = 0$ if the sample size n satisfies:

$$n \ge O\left(\frac{\log\frac{1}{\delta}}{\min\{\epsilon^{*2}, \log\frac{K}{K-1}\}}\right), \quad (10)$$

where

$$\epsilon^* = \min_{z_1, z_2 : z_1 \ne z_2} c(z_1, z_2) D_{\text{JS}}(P_{Y|Z=z_1}||P_{Y|Z=z_2})^2$$

for some constants $c(z_1, z_2) > 0, 1 \le z_1, z_2 \le K$ independent of n, δ, $O(x)$ is such that $O(x) \le$

αx *for some* $\alpha > 0$ *and* $D_{\mathrm{JS}}(P||Q) := \frac{1}{2}D_{\mathrm{KL}}\left(P||\frac{P+Q}{2}\right) + \frac{1}{2}D_{\mathrm{KL}}\left(Q||\frac{P+Q}{2}\right)$ *is the Jensen-Shannon divergence.*

The bound in Theorem 1 captures two main factors determining the sample complexity of exact phoneme discovery: the first factor is how close the word distributions of phonemes are from each other as measured by their Jensen-Shannon (JS) divergence, and the second factor is how hard it is for the training data to cover all the phonemes. The theorem works essentially because (P_1) can be viewed as an approximation of the mutual information between the codeword $\hat{z}(X)$ and word type Y, $I(\hat{z}(X);Y)$. Suppose $P^{\hat{\theta}}_{Y|X} \approx P_{Y|X}$ and let $H(\cdot|\cdot)$ denotes conditional entropy, we have:

$$\begin{aligned}
&\mathcal{L}_{IQ}(P_n, \hat{\theta}, \hat{q}) \\
&\approx H(Y|X) + D_{\mathrm{KL}}(P_{Y|X}||\hat{q}(P_{Y|X})) \\
&\propto -I(X;Y) + D_{\mathrm{KL}}(P_{Y|X}||\hat{q}(P_{Y|X})) \\
&= -I(\hat{z}(X);Y),
\end{aligned}$$

which is minimized if $\hat{q}(P_{Y|X}) = P_{Y|z(X)}$. In fact, we prove that $\hat{z}$ for such $\hat{q}$ is equivalent to $z(\cdot)$ up to a permutation in Appendix A.3.

	Flickr Audio		Librispeech	
	↑Token F1	↑NMI	↑Token F1	↑NMI
Continuous Representation				
(Nguyen et al., 2020)	35.7±0.6	40.9±0.4	48.6±1.1	60.0±0.4
CPC+MLP+k-means, K=44	49.4±0.8	52.2±0.7	67.5±0.9	71.8±1.1
CPC+MLP+k-means, K=100	40.6±0.5	51.7±0.7	61.3±0.5	71.8±0.6
CPC+MLP+k-means, K=256	28.5±0.4	51.0±0.4	48.4±1.7	68.8±0.7
Discrete Representation				
(Alemi et al., 2017)	43.6±0.7	36.1±1.9	51.0±2.1	56.2±0.9
(Strouse and Schwab, 2016), K=44	49.4±1.0	52.2±0.2	68.3±1.3	72.8±1.0
(Strouse and Schwab, 2016), K=100	41.7±0.7	52.8±0.1	60.3±0.0	71.0±0.5
(Strouse and Schwab, 2016), K=256	31.6±0.1	51.8±0.2	49.1±0.7	68.8±0.2
IQ (Ours), K=44	**53.2**±1.3	55.4±1.1	65.9±2.0	73.0±1.2
IQ (Ours), K=100	51.3±0.4	**56.5**±0.5	68.4±1.5	75.0±1.0
IQ (Ours), K=256	48.2±0.7	53.0±1.9	**69.7**±2.0	**75.8**±1.0

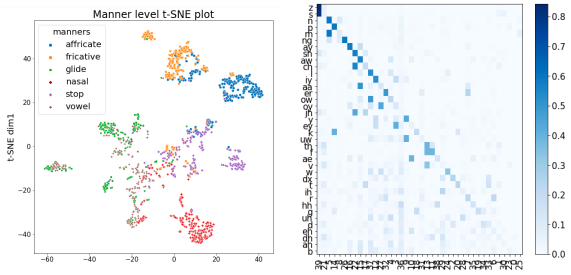
Table 1: Phoneme discovery results using segmented words extracted from Flickr audio and Librispeech.

5 Experimental Setup

Datasets We construct four training datasets consisting of spoken words only. The vocabulary set with $|\mathbb{Y}| = 224$ is selected from head words of noun phrases from the Flickr30kEntities dataset (Hodosh et al., 2010) that appear at least 500 times. For the Flickr audio word dataset, spoken words in the vocabulary are extracted from Flickr audio dataset (Harwath and Glass, 2015). For the Librispeech and TIMIT word dataset with $|\mathbb{Y}| = 224$, spoken words are extracted from Librispeech (Vassil et al., 2015) 460-hour train-clean

TIMIT	↑Token F1	↑NMI	↑Boundary F1
(Yusuf et al., 2020)	-	40.1±0.1	76.6 ±0.5
(Harwath et al., 2020)	-	35.9	54.2
(Feng et al., 2021b)	-	36.8	70.5
+ gold segmentation	-	51.2	97.8
(Ours) IQ, $\|\mathbb{Y}\|$=224, K=39	37.9±1.2	38.6±0.7	77.1±0.1
+ training on TIMIT	50.9±0.8	43.4±0.9	78.6±0.4
+ gold segmentation	62.8±0.8	59.4±0.8	96.9±0.3
(Ours) IQ, $\|\mathbb{Y}\|$=524, K=39	42.4±0.1	43.0±0.5	79.4±0.1
+ training on TIMIT	53.9±0.3	46.7±0.2	80.4±0.2
+ gold segmentation	64.3±0.4	63.4±0.4	98.3±0.3
(Ours) IQ, $\|\mathbb{Y}\|$=824, K=39	43.9±0.1	44.3±0.2	79.2±0.0
+ training on TIMIT	**54.4**±0.4	**47.5**±0.2	**80.5**±0.1
+ gold segmentation	**65.7**±0.7	**65.2**±0.6	**98.6**±0.3

Table 2: The overall phoneme discovery results of all models on TIMIT.



(3a) (3b)

Figure 3: Left: Manner-level t-SNE plot by IQ with $|\mathbb{Y}| = 824$ and gold segmentation on TIMIT. Right: Distribution of codeword assignment for each phoneme by IQ with $|\mathbb{Y}| = 824$ and predicted segmentation on TIMIT. Each row of the plot is the empirical distribution for $P_{\hat{Z}|Z}(\cdot|z), 1 \leq z \leq K$, where the phonemes are sorted top-to-bottom with decreasing $\max_{z'} P_{\hat{Z}|Z}(z'|z)$.

subset, resulting in a dataset of about 6 hours and 0.1 hours; for Librispeech and TIMIT word dataset with $|\mathbb{Y}| = 524$ and $|\mathbb{Y}| = 824$, we supplement the dataset with the speech for the top 300 frequent words and top 600 frequent words respectively (excluding the visual words) in Librispeech, resulting in datasets of about 15 and 21 hours. For Mboshi dataset, we found only about 20 actual words occur more than 100 times, so instead we use n-grams with either $n \geq 3$ (all except uni- and bi-grams) or $n \geq 2$ (all except unigrams) that occur more than 100 times as "words", resulting in a vocabulary size of 161 and 377 respectively. Note that the amount of labeled data we need is much lower than previous works (Yusuf et al., 2020): around 30 hours, (Feng et al., 2021b): around 600 hours) and the vocabulary size used is much smaller than the total vocabulary size in the language. More details of the sets can be found in Appendix B. We also test our

models on two standard phoneme discovery benchmarks, which contain whole-sentence utterances with many words unseen during training. The first dataset is TIMIT (Garofolo et al., 1993), an English corpus consisting of about 5 hours speech and Mboshi (Godard et al., 2017), which contains about 2.4 hours speech from a low-resource language. For both datasets, we follow the split in (Yusuf et al., 2020), (Feng et al., 2021b)

Baselines For phoneme discovery from segmented words, we compare our model (IQ) to four baselines. The first two baselines use continuous representation: the CPC+k-means model performs k-means clustering on the segment-level CPC features, and the k-means model performs k-means clustering after the model is trained on the word recognition task. The last two baselines use discrete representations: the Gumbel variational information bottleneck (Alemi et al., 2017) (Gumbel VIB) is a neural model with a Gumbel softmax (Jang et al., 2016) layer to approximate the codebook assignment function $z(\cdot)$, and we set $\beta = 0.001$ and decay the temperature of the Gumbel softmax from 1 to 0.1 linearly for the first 300000 steps, keeping it at 0.1 afterwards, which works best in our experiments; the deterministic information bottleneck (DIB), a generalization of (Strouse and Schwab, 2016) for continuous feature variable X, which assumes the same deterministic relation between speech X and codebook unit Z as ours, but optimizes the models in a pipeline fashion (first the speech encoder and then the quantizer) by performing clustering on the learned conditional distributions. The CPC features used are trained in a self-supervised fashion on the 960-hour LibriSpeech dataset and released by (Nguyen et al., 2020). All models share the same speech encoder as IQ. For the whole-sentence datasets, we compare our models to three phoneme discovery systems, namely, the unsupervised H-SHMM trained with multilingual speech (Yusuf et al., 2020), the ResDAVEnet-VQ (Harwath et al., 2020) with visual supervision and the TDNN-f system by (Feng et al., 2021b) trained with multilingual speech. To study how well our model performs in extreme low-resource speech recognition compared to other neural speech representation learning models, we compare our models to wav2vec (Schneider et al., 2019), wav2vec 2.0 (Baevski et al., 2020) (small, trained on the 960-hour LibriSpeech), vq-wav2vec with Gumbel softmax and k-means as discretization strategies (Baevski et al., 2019), CPC (van den Oord et al., 2019) and VQ-CPC (van Niekerk et al., 2020), using the pretrained models released by the authors. Implementation details of the baselines and our models are in Appendix C.

Evaluation metrics Standard metrics are used such as NMI and boundary F1 for the quality of codebook and segmentation respectively with the same implementation as in prior works (Yusuf et al., 2020; Feng et al., 2021b). In addition, token F1 (Dunbar et al., 2017) is also reported. To examine the benefit of using our discovered phoneme inventory for low-resource speech recognition, we also evaluate using *equivalent phone error rate* (equiv. PER: Ondel et al. 2019). This metric can be viewed as a proxy for phone error rate (PER) applicable beyond supervised speech recognizers.

	↑Token F1	↑NMI	↑Boundary F1
(Ondel et al., 2019)	-	38.4±1.0	59.5±0.8
(Yusuf et al., 2020)	-	41.1±1.1	59.2±1.5
(Feng et al., 2021b), 5 langs	-	43.5±0.3	62.8±0.0
+ Gold segmentation	-	60.6±0.1	100±0.0
(Feng et al., 2021b), 13 langs	36.4±0.6	44.7±0.6	64.1±0.1
+ Gold segmentation	50.8±0.6	64.6±0.3	100±0.0
(Ours) IQ, $\lvert\mathbb{Y}\rvert = 161$, K=31	46.5±0.4	40.2±0.1	65.5±0.1
+ Multilingual BNF	54.2±1.0	45.1±0.4	**67.5**±0.1
+ Gold segmentation	66.4±0.8	69.7±0.4	100±0.0
+ Multilingual BNF	74.3±0.8	76.9±0.6	100±0.0
(Ours) IQ, $\lvert\mathbb{Y}\rvert = 377$, K=31	50.4±0.5	45.2±0.8	66.8±0.0
+ Multilingual BNF	57.1±1.0	49.3±0.3	67.3±0.1
+ Gold segmentation	69.3±1.0	73.0±0.6	100±0.0
+ Multilingual BNF	**81.7**±0.8	**82.6**±0.3	100±0.0

(3a)

	↓ Equiv. PER		↑ Boundary F1
	Predicted Segments	Gold Segments	
wav2vec+k-means	66.6	64.8	52.4
wav2vec 2.0+k-means	64.5	60.0	55.3
vq-wav2vec (k-means)	77.3	-	31.1
vq-wav2vec (Gumbel)	77.0	-	30.3
CPC+k-means	63.1	57.4	54.7
VQ-CPC	80.3	-	23.0
IQ + Multilingual BNF (Ours)	**44.3**	**25.8**	**67.3**

(3b)

Table 3: (a) Phoneme discovery results of all models on Mboshi dataset. (b) Comparison of IQ with other self-supervised models in zero-resource speech recognition.

6 Results

6.1 Word-level Phoneme Discovery

The results on visual word-only test sets of Flickr audio and Librispeech are shown in Table 1. On both datasets, IQ outperforms both Gumbel VIB and DIB in terms of all metrics, especially on Flickr

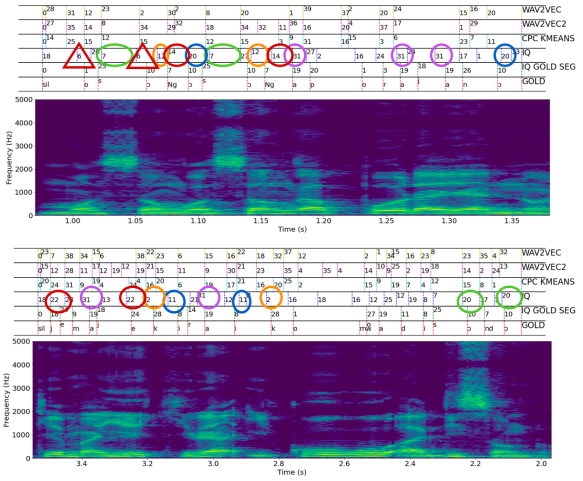


Figure 4: The spectrograms annotated with the gold transcripts and the zero-resource transcriptions by various models for two Mboshi utterances. The spoken segments are in circles of the same colors are identified as the same phoneme by our IQ model and in triangles of the same color if they are but are acoustically similar.

audio, which has more phonemes than Librispeech and a larger test set. Moreover, the performance of IQ is very robust to the codebook size, achieving good results even when the codebook size is very different from the size of the true phoneme inventory, suggesting our theory may be able to work with a relaxed Assumption 4.

6.2 Sentence-level Phoneme Discovery

The results on TIMIT and Mboshi are shown in Table 2 and Table 3a respectively. On TIMIT, our model is able to outperform the visually grounded baseline (Harwath et al., 2020) for all training vocabulary, and all three baselines for $|\mathbb{Y}| = 524$ and $|\mathbb{Y}| = 824$ with and without gold segmentation in terms of all three metrics. Further, we also empirically verify the sample complexity bound in Theorem 1 as IQ performs better in Token F1 and NMI as the training vocabulary size get larger, which generally increases the JS divergence. On Mboshi, IQ with CPC feature consistently outpeforms (Feng et al., 2021b) in token F1 and boundary F1, and IQ with CPC+BNF features consistently outperform (Feng et al., 2021b) in all three metrics under various level of word supervision. The performance of our model on Mboshi compared with other neural self-supervised models are shown in Table 3b. We found that IQ outperforms the best self-supervised model, CPC+k-means in equiv. PER by 34% and 20% absolute with and without gold segmentation respectively and 12% absolute in terms of boundary F1, suggesting that IQ is able to learn consistent phoneme-like sequence useful for zero-resource or extremely low-resource speech recognition.

Effect of segmentation and codebook size The use of unsupervised phoneme segmentation deteriorates the NMI by about 18% and 28% absolute on TIMIT and Mboshi respectively for our models since the distributional property of phonemes does not apply exactly to non-phoneme segments. On the other hand, in Appendix F we show that the quality of codeword assignments by IQs is very robust against varying codebook size, after experimenting with codebook size from 30 to 70 on TIMIT and Mboshi.

Multilingual and word supervision are complimentary In all vocabulary sizes, concatenating the multilingual BNF from (Feng et al., 2021b) to the CPC output representation from the segmental speech encoder in Figure 2 significantly improves token F1 and NMI to allow our best models to outperform baselines in all three metrics.

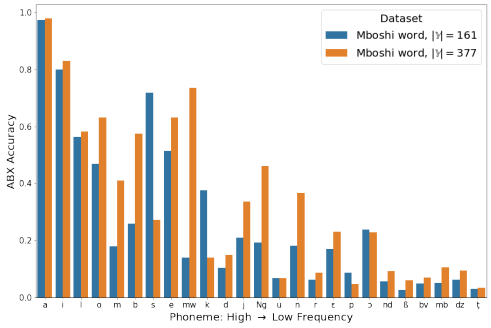


Figure 5: ABX phoneme identification accuracy vs phoneme frequency on the Mboshi dataset for IQ trained with vocabulary size 161 and 377.

6.3 Analysis

IQ codebook resembles true phonemes From Figure 3b, we observe that the codeword assignments by IQ correlates well with the actual phonemes, but tends to confuse the most between phonemes within the same manner class, such as nasals /n/ and /m/. This is also confirmed by the t-SNE plot in Figure 3a, where the embeddings of most manner classes are well-clustered, except for related manner classes such as affricate and fricative, or glide and vowel. Further, from the examples shown in Figure 4, we can see that IQ is not only better at grouping segments of the same

phonemes but also at detecting segment boundaries than the baselines. Also, across different examples, IQ assign the same codes to phonemes such as /a/ (31) and /s/ (7) more consistently than other models do. Please check Appendix G for more speech examples.

Limitation While our theory predicts that with gold segmentation, the TER of IQ is asymptotically zero, in practice TER is nonzero due to the violation of Assumption 4, i.e., the phonemes are not uniformly distributed for languages such as Mboshi. As a result, the model often discards information of the rare phonemes by merging them into a more frequent phoneme cluster. Evidently, from Figure 5, where we use ABX accuracy (Munson and Gardner, 1950) to score how reliable the IQ codebook can identify segments of the same phoneme, we observe a strong correlation is observed between ABX accuracy and the frequency of the phonemes.

7 Conclusion

Motivated by the linguistic definition of phonemes, we propose information quantizer (IQ), a new neural network model for self-supervised phoneme discovery that can take advantage of word-level supervision. We demonstrate in two ways that word-level supervision is beneficial for phoneme inventory discovery: theoretically, we prove that IQ can achieve zero token error rate asymptotically with the help of word labels; empirically, we show that IQ out-performs various speech-only algorithms in phoneme discovery tasks under both simulated (English) and realistic (Mboshi) low-resource settings. In the future, we would like to apply the discovered phoneme inventory to develop better low-resource speech technologies such speech translation and speech synthesis systems.

References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*.

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021. Unsupervised speech recognition. In *ArKiv*.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Neural Information Processing System*.

Marcely Zanon Boito, Aline Villavicencio, and Laurent Besacier. 2019. Empirical evaluation of sequence-to-sequence models for word discovery in low-resource settings. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*.

K.-Y. Chen, C.-P. Tsai, D.-R. Liu, H.-Y. Lee, and L. shan Lee. 2019. Completely unsupervised speech recognition by a generative adversarial network harmonized with iteratively refined hidden markov models. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*.

Jan Chorowski, Ron J. Weiss, Samy Bengio, and Aäron van den Oord. 2019. Unsupervised speech representation learning using wavenet autoencoders. *IEEE Transactions on Audio, Speech and Language Processing*.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James R. Glass. 2019. An unsupervised autoregressive model for speech representation learning. In *In Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA.

Ewan Dunbar, Xuan-Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017. *CoRR*, abs/1712.04313.

Ryan Eloff, André Nortje, Benjamin van Niekerk, Avashna Govender, Leanne Nortje, Arnu Pretorius, Elan van Biljon, Ewald van der Westhuizen, Lisa van Staden, and Herman Kamper. 2019b. Unsupervised acoustic unit discovery for speech synthesis using discrete latent-variable neural networks. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*.

Siyuan Feng, Tan Lee, and Zhiyuan Peng. 2019. Combining adversarial training and disentangled speech representation for robust zero-resource subword modeling. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*.

Siyuan Feng, Piotr Żelasko, Laureano Moro-Velázquez, Ali Abavisani, Mark Hasegawa-Johnson, Odette Scharenborg, and Najim Dehak. 2021a. How phonotactic affect multilingual and zero-shot asr performance. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Siyuan Feng, Piotr Żelasko, Laureano Moro-Velázquez, and Odette Scharenborg. 2021b. Unsupervised acoustic unit discovery by leveraging a language-independent subword discriminative feature representation. In *INTERSPEECH*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform. In *arXiv*.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathon G. Fiscus, David S. Pallett, and Nancy L. Dahlgren. 1993. *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM*. Linguistic Data Consortium.

Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-No"el Kouarata, Lori Lamel, H'el'ene Maynard, Markus M"uller, Annie Rialland, Sebastian St"uker, François Yvon, and Marcely Zanon Boito. 2017. A very low resource language speech corpus for computational language documentation experiments. *CoRR*, abs/1710.03501.

Pierre Godard, Marcely Zanon Boito, Lucas Ondel, Alexandre Berard, Aline Villavicencio, and Laurent Besacier. 2018. Unsupervised word segmentation from speech with attention. In *Interspeech*.

David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. *Automatic Speech Recognition and Understanding*.

David Harwath and James Glass. 2019. Towards visually grounded subword speech unit discovery. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

David Harwath, Wei-Ning Hsu, and James Glass. 2020. Learning hierarchical discrete linguistic units from visually-grounded speech. In *International Conference on Learning Representation*.

M. Hodosh, P. Young, and J. Hockenmaier. 2010. Framing image description as a ranking task: data, models and evaluation metrics. In *Journal of Artificial Intelligence Research*.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Aren Jansen. 2013. Weak top-down constraints for unsupervised acoustic model training. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Aren Jansen, Kenneth Church, and Hynek Hermansky. 2010. Toward spoken term discovery at scale with zero resources. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*.

Herman Kamper, Aren Jansen, and Sharon Goldwater. 2016. Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE Transaction on Audio, Speech and Language Processing*, 24:669–679.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*.

Felix Kreuk, Joseph Keshet, and Yossi Adi. 2020. Self-supervised contrastive learning for unsupervised phoneme segmentation. In *INTERSPEECH*.

Chiaying Lee and James Glass. 2012. 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*, pages 40–49, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Lööf, C. Gollan, and H. Ney. 2009. Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a polish speech recognition system. In *10th Annual Conference of the International Speech Communication Association*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

W. A. Munson and Mark B. Gardner. 1950. Standardizing auditory tests. *The Journal of the Acoustical Society of America*, 22.

Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. 2020. The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. In *Self-Supervised Learning for Speech and Audio Processing Workshop @ NeurIPS*.

Benjamin van Niekerk, Leanne Nortje, and Herman Kamper. 2020. Vector-quantized neural networks for acoustic unit discovery in the zerospeech 2020 challenge. In *Interspeech*.

L. Ondel, L. Burget, and J. Cernocký. 2016. Variational inference for acoustic unit discovery. In *Spoken Language Technology for Underresourced Languages*.

L. Ondel, H. K. Vydana, L. Burget, and J. Cernocký. 2019. Bayesian subspace hidden Markov model for acoustic unit discovery. In *INTERSPEECH*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. In *ArXiv*.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Park and James Glass. 2005. Towards unsupervised pattern discovery in speech. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

David Qiu, Anuran Makur, and Lizhong Zheng. 2019. Probabilistic clustering using maximal matrix norm couplings. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv*.

T. Schultz and A. Waibel. 1998. Multilingual and crosslingual speech recognition. In *Proc. DARPA Workshop on Broadcast News Tran- scription and Understanding*, page 259–262.

Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning*. Cambridge University Press.

DJ Strouse and David Schwab. 2016. The deterministic information bottleneck. In *Association for Uncertainty in Artificial Intelligence*.

S. Stuker, T. Schultz, F. Metze, and A. Waibel. 2003. Multilingual articulatory features. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*.

Morris Swadesh. 1934. The phonemic principle. *Language*, 10(2):117–129.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proc. American Philosophical Society*, 96(4):452–463.

P. Swietojanski, A. Ghoshal, and S. Renals. 2012. Unsupervised crosslingual knowledge transfer in dnn-based lvcsr. In *2012 IEEE Spoken Language Technology Workshop (SLT)*.

Yao-Hung Hubert Tsai, Han Zhao, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2020. Neural methods for point-wise dependency estimation. In *Neural Information Processing System*.

Panayotov Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *ICASSP*, page pp. 5206–5210.

Roman Vershynin. 2018. *High-Dimensional Probability–An Introduction with Applications in Data Science*. Cambridge University Press.

Chih-Kuan Yeh, Jianshu Chen, Chengzhu Yu, and Dong Yu. 2019. Unsupervised speech recognition via segmental empirical output distribution matching. In *International Conference on Learning Representations*.

B. Yusuf, L. Ondel, L. Burget, J. Cernocký, and M. Saraclar. 2020. A hierarchical subspace model for language-attuned acoustic unit discovery. In *CoRR*.

P. Żelasko, L. Moro-Velázquez, M. Hasegawa-Johnson, O. Scharenborg, and N. Dehak. 2020a. That sounds familiar: an anal- ysis of phonetic representations transfer across languages. In *Proc. INTERSPEECH*.

Piotr Żelasko, Laureano Moro-Velázquez, Mark Hasegawa-Johnson, Odette Scharenborg, and Najim Dehak. 2020b. That sounds familiar: an analysis of phonetic representations transfer across languages. In *Interspeech*.

A Proofs of Theoretical Results

A.1 Statistical definition of phonemes

Proof of Proposition 1. Without loss of generality, suppose $(x_1, x_1') \in \mathbb{X}^2$, suppose there exists y_1 such that

$$P_{Y|X}(y_1|x_t) > P_{Y|X}(y_1|x_t'),$$

then there exists y_2 such that

$$P_{Y|X}(y_2|x_t) < P_{Y|X}(y_2|x_t'),$$

which means there exists $0 \leq \alpha_1, \alpha_2 \leq 1, \alpha_1 + \alpha_2 \leq 1$, such that

$$\frac{P_{Y|X}(y_1|x_t')}{P_{Y|X}(y_2|x_t')} \leq \frac{\alpha_2}{\alpha_1} < \frac{P_{Y|X}(y_1|x_t)}{P_{Y|X}(y_2|x_t)}.$$

Now, since Equation 2 holds for arbitrary $P_{Y|X=x_s} \in \Delta^{|\mathbb{Y}|}, s \neq t$, we can set

$$P_{Y|X}(y_1|x_2) = \alpha_1, P_{Y|X}(y_2|x_2) = \alpha_2,$$
$$P_{Y|X}(y_1|x_t) = P_{Y|X}(y_2|x_t) = \frac{1}{2}, \forall t > 2,$$

in which case Equation 2 boils down to

$$\arg\max_{i \in \{1,2\}} \alpha_i P_{Y|X}(y_i|x_1) = \arg\max_{i \in \{1,2\}} \alpha_i P_{Y|X}(y_i|x_1').$$

However, by the choice of α_i's, the left-hand side is y_1 since $\alpha_1 P_{Y|X}(y_1|x_1) > \alpha_2 P_{Y|X}(y_2|x_1)$ and the right-hand side is y_2 since $\alpha_2 P_{Y|X}(y_2|x_1) > \alpha_1 P_{Y|X}(y_1|x_1')$, and therefore Equation 2 cannot hold. Therefore, Equation 2 is true only if $P_{Y|X}(y|x_1) = P_{Y|X}(y|x_1'), \forall (x_1, x_1') \in \mathbb{X}^2, y \in \mathbb{Y}$. □

A.2 Equivalence of TER and standard phoneme discovery metrics

Consider the groundtruth assignment $z(\cdot)$ and a codebook assignment $\hat{z}(\cdot)$ with $\hat{K}$ code words, the NMI of $\hat{z}$ is defined as:

$$\text{NMI}(\hat{z}) = \frac{2I(z(X); \hat{z}(X))}{H(z(X)) + H(\hat{z}(X))}, \tag{11}$$

where $H(\cdot)$ denotes the entropy and $I(\cdot;\cdot)$ denotes the mutual information.

which is also related to the **token F1** used for acoustic unit discovery (Dunbar et al., 2017). Since SPD is an unsupervised learning problem and ground truth phoneme labels are not available, matching between codebook indices and phoneme units is needed. When computing token F1, we consider two different many-to-one mappings $\pi_{\text{rec}} : \{1, \cdots, K\} \to \{1, \cdots, \hat{K}\}$ and $\pi_{\text{prec}} : \{1, \cdots, \hat{K}\} \to \{1, \cdots, K\}$ to compute the token recall and precision respectively as:

$$\text{Rec}(\hat{z}) := \max_{\pi_{\text{rec}}} \mathbb{P}\{\hat{z}(X) = \pi_{\text{rec}}(z(X))\} \tag{12}$$

$$\text{Prec}(\hat{z}) := \max_{\pi_{\text{prec}}} \mathbb{P}\{z(X) = \pi_{\text{prec}}(\hat{z}(X))\}, \tag{13}$$

before computing the harmonic mean between the two to obtain token F1: $\text{F1}(\hat{z}) := \frac{2\text{Prec}(\hat{z})\text{Rec}(\hat{z})}{\text{Prec}(\hat{z})+\text{Rec}(\hat{z})}$. The following proposition relates TER with token F1 and NMI.

Proposition 2. *For any assignment function $\hat{z} : \{1, \cdots, K\} \to \{1, \cdots, K\}$, $P_{TER}(\hat{z}) = 0$ if and only if $F1(\hat{z}) = NMI(\hat{z}) = 1$.*

Proof. First of all, for such $\hat{z}$, we have

$$\begin{aligned} 1 \geq \text{F1}(\hat{z}) &\geq \min\{\text{Prec}(\hat{z}), \text{Rec}(\hat{z})\} \\ &\geq 1 - P_{\text{e, TER}}(\hat{z}) = 1, \end{aligned}$$

where the third inequality comes from the fact that the set of permutations is a smaller set than the set of all many-to-one mappings $\pi : \{1, \cdots, K\} \to \{1, \cdots, K\}$. Further, using the fact that z and $\hat{z}$ are functions of each other when $P_{TER}(\hat{z}) = 0$, it can be shown that $NMI(\hat{z}) = \frac{2I(z(X),\hat{z}(X))}{H(z(X))+H(\hat{z}(X))} = 2H(z(X))/2H(z(X)) = 1$. □

A.3 Exact Discovery Guarantee

First, we prove the claim made in Section 4.2 about nearest neighbor information quantizers. Recall the definition of general and nearest-neighbor information quantizers as follows.

Definition 3. *(Information quantizer) A K-point information quantizer is a function $q : \Delta^{|\mathbb{Y}|} \to \mathbb{C} = \{Q_1, \cdots, Q_K\} \subset \Delta^{|\mathbb{Y}|}$, where $\mathbb{C}$ is called the codebook and Q_k's are called the code distributions. Further, define $\mathbb{Q}_K$ to be the class of such functions.*

Definition 4. *(Nearest-neighbor Information quantizer) A K-point information quantizer is called nearest-neighbor if, $\forall P \in \Delta^{|\mathbb{Y}|}$, $D_{\text{KL}}(P||q(P)) = \min_{1\leq k\leq K} D_{\text{KL}}(P||Q_k)$. Further, define $\mathbb{Q}_K^{NN} \subset \mathbb{Q}_K$ to be the class of such functions.*

Then we have the following lemma.

Lemma 1. *There exists an information quantizer $\hat{\theta}_n \in \Theta, \hat{q}_n \in \mathbb{Q}_K^{NN}$ such that*

$$\mathcal{L}_{IQ}(P_n, \hat{\theta}_n, \hat{q}_n) = \min_{\theta\in\Theta, q\in\mathbb{Q}_K} \mathcal{L}_{IQ}(P_n, \theta, q). \tag{14}$$

Therefore, $(\hat{\theta}_n, \hat{q}_n)$ is an ERM of (P_1).

Proof of Lemma 1. Notice that only the $\mathcal{L}_Q$ term of Equation P_1 depends on q, so it suffices to show that $\min_{q\in\mathbb{Q}_K^{\text{NN}}} \mathcal{L}_Q(\tilde{P}_n, q) \leq \min_{q\in\mathbb{Q}_K} \mathcal{L}_Q(\tilde{P}_n, q)$. This is true since

$$\begin{aligned} &\min_{q\in\mathbb{Q}_K} \mathcal{L}_Q(\tilde{P}_n, q) \\ =& \min_{q\in\mathbb{Q}_K} \mathbb{E}_{\tilde{P}_n}[D_{\text{KL}}(P^\theta_{Y|X}||q(P^\theta_{Y|X}))] \\ \geq& \mathbb{E}_{\tilde{P}_n}[\min_{1\leq k\leq K} D_{\text{KL}}(P^\theta_{Y|X}||Q_k)] \\ =& \min_{q\in\mathbb{Q}_K^{\text{NN}}} \mathbb{E}_{\tilde{P}_n}[D_{\text{KL}}(P^\theta_{Y|X}||q(P^\theta_{Y|X}))] \\ =& \min_{q\in\mathbb{Q}_K^{\text{NN}}} \mathcal{L}_Q(\tilde{P}_n, q), \end{aligned}$$

where the inequality holds since

$$D_{\text{KL}}(P^\theta_{Y|X}||q(P^\theta_{Y|X})) \geq \min_{1\leq k\leq K} D_{\text{KL}}(P^\theta_{Y|X}||Q_k)$$

for any $q \in \mathbb{Q}_K$. □

Next, we show under the condition $P^\theta_{Y|X} = P_{Y|X}$ and $n \to \infty$, (P_1) recovers $z(\cdot)$ up to a permutation.

Proposition 3. *The pair $(z^*, P^*_{Y|Z})$ is a minimizer to the following optimization problem:*

$$\max_{\hat{z}:\mathbb{X}\to\{1,\cdots,K\}, P_{Y|Z}\in\Delta^{|\mathbb{Y}|}} I(\hat{z}(X); Y), \tag{P_0}$$

if and only if z^* is equal to the true assignment function z up to a permutation.

Proof. $\Rightarrow$: First, $z(\cdot)$ is a feasible solution by definition. By data processing inequality, we have

$$I(z'(X);Y) \le I(X;Y) = I(z(X);Y).$$

Therefore, $z(\cdot)$ is also the optimal solution.

$\Leftarrow$: Suppose there exists some optimal $(\hat{z}, \hat{P}_{Y|Z})$ with $\hat{P}_{Y|\hat{z}(x)} \neq P_{Y|z(x)}$ for at least one $x \in \mathcal{X}$. Since such discrepancies are independent with each other, it suffices to show that each such discrepancy leads to lower $I(Z;Y)$. Indeed, for $(\hat{z}, \hat{P}_{Y|Z})$ with $\hat{P}_{Y|Z=\hat{z}(x)} \neq P_{Y|Z=z(x)}$ only at x,

$$\begin{aligned}
&I(\hat{z}(X);Y) - I(z(X);Y)\\
=&P_X(x)\sum_y P_{Y|X}(y|x)\log\frac{\hat{P}_{Y|Z=\hat{z}(x)}}{P_{Y|Z=z(x)}}\\
=&-P_X(x)D(P_{Y|Z=z(x)}||\hat{P}_{Y|Z=\hat{z}(x)}) < 0,
\end{aligned}$$

which contradicts the optimality of $\hat{z}$. Therefore, $\hat{P}_{Y|\hat{z}(x)} = P_{Y|z(x)}$ for all optimal solution of (P_0). □

To prove Theorem 1, we also need the following lemma.

Lemma 2. *Under Assumption 3, for any bounded parameter set Θ, there exists $\gamma > 0$ and some optimal parameter $\theta^* \in \Theta^*$ such that*

$$D_{\mathrm{KL}}(P^{\theta}_{Y|X}||P^{\theta^*}_{Y|X}) \ge \gamma\|\theta - \theta^*\|,\ \forall\theta\in\Theta.$$

Proof. We prove the lemma by contradiction. First, we assume $\theta \notin \Theta^*$ since the inequality satisfies trivially for any $\theta \in \Theta^*$. By boundedness, there exists some $R > 0$ such that $\|\theta\| \le R$. Suppose for any $\gamma > 0$, there exists some $\theta \in \Theta$ such that $D_{\mathrm{KL}}(P^{\theta}_{Y|X}||P^{\theta^*}_{Y|X}) \le \gamma\|\theta - \theta^*\| \le 2\gamma R$, then we have $D_{\mathrm{KL}}(P^{\theta}_{Y|X}||P^{\theta^*}_{Y|X}) \le \inf_{\gamma>0}\gamma R = 0$. However, since $D_{\mathrm{KL}}(P^{\theta}_{Y|X}||P^{\theta^*}_{Y|X}) \ge 0$, we have $D_{\mathrm{KL}}(P^{\theta}_{Y|X}||P^{\theta^*}_{Y|X}) = 0$, which implies $\theta \in \Theta^*$ and leads to contradiction. □

Note it is crucial that the parameter set is bounded, which is the case for neural nets. Further, Assumption 3 is needed or the inequality can be easily violated when the optimal parameter set Θ^* is empty.

Next, we need the following lemma, which is based on (Tsai et al., 2020):

Lemma 3. *Under Assumptions 1-3, and consider $\hat{\theta}$ to be part of the ERM of (P_1) with conditional distribution $\hat{P}_{Y|X} := P^{\hat{\theta}}_{Y|X}$. Then for any $\epsilon > 0$, the following inequality holds:*

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{x\in\mathbb{X}} D_{\mathrm{KL}}(P_{Y|X=x}||\hat{P}_{Y|X=x}) > \epsilon\right\}\\
\le&2\left|\mathcal{N}(\Theta,\frac{\epsilon}{4\rho})\right|\exp\left(-\frac{\gamma^2 n\epsilon^2}{2\rho^2(C_u-C_l)^2}\right), \quad (15)
\end{aligned}$$

where $\mathcal{N}(A,\epsilon)$ is the ϵ-net of set A.

Proof. For notational ease, we drop the dependence of $\mathcal{L}_{\mathrm{CE}}$ on P if the context is clear. Using Assumption 3, let $P_{Y|X} = P^{\theta^*}_{Y|X}$. Define $D_n(P||Q)$ as the empirical KL divergence. Further, notice that for $P_{Y|X}$, $\mathcal{L}_Q$ can always be made 0 and therefore, the ERM of P_1 needs to satisfy $\mathcal{L}_{\mathrm{CE}}(\hat{\theta}) \le \mathcal{L}_{\mathrm{CE}}(\theta^*)$. As a result,

$$\begin{aligned}
&D_n(P_{Y|X}||\hat{P}_{Y|X})\\
:=&\mathbb{E}_{P_n}\left[\log\frac{P_{Y|X}(Y|X)}{\hat{P}_{Y|X}(Y|X)}\right]\\
=&\mathcal{L}_{\mathrm{CE}}(\hat{\theta}) - \mathcal{L}_{\mathrm{CE}}(\theta^*) \le 0.
\end{aligned}$$

Note that $D_n(P||Q)$ is an unbiased estimator of the conditional KL divergence between distributions P and Q: $\mathbb{E}_{P_{X^n,Y^n}}\mathbb{E}_{P_n}\log\frac{P_{Y|X}(Y|X)}{Q_{Y|X}(Y|X)} = D_{\mathrm{KL}}(P_{Y|X}||Q_{Y|X})$. Therefore, let $\Delta_n(\theta) := D_n(P_{Y|X}||P^{\theta}_{Y|X}) - D_{\mathrm{KL}}(P_{Y|X}||P^{\theta}_{Y|X})$,

$$\begin{aligned}
&\mathbb{P}\left\{D_{\mathrm{KL}}(P_{Y|X}||\hat{P}_{Y|X}) > \epsilon\right\} \le\\
&\mathbb{P}\left\{D_{\mathrm{KL}}(P_{Y|X}||\hat{P}_{Y|X}) - D_n(P_{Y|X}||\hat{P}_{Y|X}) > \epsilon\right\}\\
&\le \mathbb{P}\left\{|\Delta_n(\theta)| > \epsilon\right\} \le \mathbb{P}\left\{\sup_{\theta\in\Theta}|\Delta_n(\theta)| > \epsilon\right\}.
\end{aligned}$$

To bound the last probability, consider an $\frac{\epsilon}{4\rho}$-net in the parameter space $\mathcal{N}(\Theta,\frac{\epsilon}{4\rho})$ and $\Theta = \cup_{k=1}^{|\mathcal{N}(\Theta,\frac{\epsilon}{4\rho})|}\Theta_k$, where Θ_k is the $\frac{\epsilon}{4\rho}$-ball surrounding $\theta_k \in \mathcal{N}(\Theta,\frac{\epsilon}{4\rho})$, we have $\forall\theta\in\Theta_k$,

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\theta\in\Theta}|\Delta_n(\theta)| > \epsilon\right\}\\
\le&\sum_{k=1}^{\left|\mathcal{N}(\Theta,\frac{\epsilon}{4\rho})\right|}\mathbb{P}\left\{\sup_{\theta\in\Theta_k}|\Delta_n(\theta)| > \epsilon\right\} \le\\
&\left|\mathcal{N}(\Theta,\frac{\epsilon}{4\rho})\right|\sup_k\mathbb{P}\left\{\sup_{\theta\in\Theta_k}|\Delta_n(\theta)| > \epsilon\right\}. \quad (16)
\end{aligned}$$

Further, by Assumption 2, we have

$$\sup_{\theta\in\Theta_k}|\Delta_n(\theta)-\Delta_n(\theta_k)|\le$$
$$\sup_{\theta\in\Theta_k}\left|D_n(P_{Y|X}||P^{\theta}_{Y|X})-D_n(P_{Y|X}||P^{\theta_k}_{Y|X})\right|$$
$$+\left|D_{\mathrm{KL}}(P_{Y|X}||P^{\theta}_{Y|X})-D_{\mathrm{KL}}(P_{Y|X}||P^{\theta_k}_{Y|X})\right|=$$
$$\mathbb{E}_{P_n}\left|\log\frac{P^{\theta_k}_{Y|X}(Y|X)}{P^{\theta}_{Y|X}(Y|X)}\right|$$
$$+\mathbb{E}_{P_{XY}}\left|\log\frac{P^{\theta_k}_{Y|X}(Y|X)}{P^{\theta}_{Y|X}(Y|X)}\right|\le 2\rho\|\theta_k-\theta\|\le\frac{\epsilon}{2}.$$

As a result,

$$\mathbb{P}\left\{\sup_{\theta\in\Theta_k}|\Delta_n(\theta)|>\epsilon\right\}$$
$$\le\mathbb{P}\left\{|\Delta_n(\theta_k)|+\sup_{\theta\in\Theta_k}|\Delta_n(\theta)-\Delta_n(\theta_k)|>\epsilon\right\}$$
$$\le\mathbb{P}\left\{|\Delta_n(\theta_k)|>\frac{\epsilon}{2}\right\}$$
$$\le 2\exp\left(-\frac{n\epsilon^2}{2(C_u-C_l)^2}\right),$$

by Assumption 1 and Hoeffding's inequality. Plugging this into (16), we arrive at

$$\mathbb{P}\left\{D_{\mathrm{KL}}(P_{Y|X}||\hat{P}_{Y|X})>\epsilon\right\}$$
$$\le 2\left|\mathcal{N}(\Theta,\frac{\epsilon}{4\rho})\right|\exp\left(-\frac{n\epsilon^2}{2(C_u-C_l)^2}\right). \quad (17)$$

To prove uniform convergence, use Assumption 2 to conclude that:

$$D_{\mathrm{KL}}(P_{Y|X=x}||\hat{P}_{Y|X=x})$$
$$=\sum_y P_{Y|X}(y|x)\log\frac{P^{\theta^*}_{Y|X}(y|x)}{P^{\hat{\theta}_n}_{Y|X}(y|x)}$$
$$\le\sup_y\left|\log\frac{P^{\theta^*}_{Y|X}(y|x)}{P^{\hat{\theta}_n}_{Y|X}(y|x)}\right|\le\rho\|\theta^*-\hat{\theta}_n\|,$$

for some $\theta^*\in\Theta^*$. Therefore, using Lemma 2, we arrive at the desired result:

$$\mathbb{P}\left\{\sup_{x\in\mathbb{X}}D_{\mathrm{KL}}(P_{Y|X=x}||\hat{P}_{Y|X=x})\ge\epsilon\right\}$$
$$\le\mathbb{P}\left\{\|\theta^*-\hat{\theta}_n\|\ge\frac{\epsilon}{\rho}\right\}$$
$$\le\mathbb{P}\left\{D_{\mathrm{KL}}(P_{Y|X}||\hat{P}_{Y|X})\ge\frac{\gamma\epsilon}{\rho}\right\}$$
$$\le 2\left|\mathcal{N}(\Theta,\frac{\epsilon}{4\rho})\right|\exp\left(-\frac{\gamma^2 n\epsilon^2}{2\rho^2(C_u-C_l)^2}\right).$$

□

Next, we prove the following lemma by performing a perturbation analysis on (P_1) inspired by (Qiu et al., 2019).

Lemma 4. *Consider some subset of speech segments $\mathcal{D}\subset\mathbb{X}$ such that for any $1\le z\le K$, there exists $x\in\mathbb{X}$ such that $z(x)=z$. Further, suppose there exists $\epsilon>0$ such that $\|\hat{P}_{Y|X=x}-P_{Y|X=x}\|_1\le\epsilon,\forall x\in\mathcal{D}$. Then, $\forall x\in\mathbb{X}$, $\|\hat{q}(\hat{P}_{Y|X=x})-P_{Y|X=x}\|_1\le c_1\epsilon^{1/2}$ for some constant $c_1>0$.*

Proof. We first prove the statement for the segments from the set $\mathcal{D}$. By the definition of ERM,

$$\mathcal{L}_Q(P_n,\hat{q})-\mathcal{L}_Q(P_n,q^*) \quad (18)$$
$$=\mathbb{E}_{\tilde{P}_n}\left[\log\frac{P_{Y|X}(Y|X)}{\hat{q}(\hat{P}_{Y|X}(Y|X))}\right]\le 0. \quad (*)$$

From the condition in the lemma, we have $\hat{P}_{Y|X=x}=P_{Y|X=x}+\epsilon\phi_x$ for some $\epsilon\in[0,1]$ and $\phi_x\in\mathbb{R}^{|\mathbb{Y}|},\phi_x^\top\mathbf{1}=0,\|\phi_x\|_1\le 1,\forall x\in\mathcal{D}$. Further, suppose $q(\hat{P}_{Y|X})=P_{Y|X}+\delta\psi_x$ for some $\delta\in[0,1]$ and $\psi_x\in\mathbb{R}^{|\mathbb{Y}|},\psi_x^\top\mathbf{1}=0,\|\psi_x\|_1\le 1,\forall x\in\mathbb{X}$. Using Assumption 1 and the inequality $\log(1+x)\le x-\frac{x^2}{4},\forall x\in(-1,1]$, we have

$$\sum_y\hat{P}_{Y|X}(y|x)\log\frac{P_{Y|X}(y|x)}{\hat{q}(\hat{P}_{Y|X}(y|x))}$$
$$=-\sum_y P_{Y|X}(y|x)\log\left(1+\delta\frac{\psi_x(y)}{P_{Y|X}(y|x)}\right)$$
$$-\sum_y\epsilon\phi_x(y)\log\frac{P_{Y|X}(y|x)}{\hat{q}(\hat{P}_{Y|X}(y|x))}$$
$$\ge\sum_y\frac{\delta^2\psi_x^2(y)}{4P_{Y|X}(y|x)}-C_u\epsilon\ge\frac{\delta^2\|\psi_x(y)\|^2}{4}$$
$$\ge\frac{\delta^2}{4|\mathbb{Y}|}-C_u\epsilon,$$

for every $x\in\mathcal{D}$. Therefore, to maintain (18), we need $\delta^2\le 4C_u|\mathbb{Y}|\epsilon$ for the training examples X^n and the inequality in the lemma holds for examples from $\mathcal{D}$ with coefficient $c_1':=2\sqrt{C_u|\mathbb{Y}|}$.

To show the same claim holds for any unseen segments $x'\in\mathbb{X}\backslash\mathcal{D}$, we first use Lemma 1 to conclude that there always exists a nearest-neighbor information quantizer $\hat{q}$ that is an ERM. Further, since every phoneme class occurs in $\mathcal{D}$, we can always find $x\in\mathcal{D}$ such that $z(x)=z(x')$. Therefore, using the inequality $\log(1+x)\ge x-\frac{x^2}{1+x},\forall x>-1$,

we have

$$\begin{aligned}
&\frac{1}{2}\|\hat{P}_{Y|X=x'} - \hat{q}(\hat{P}_{Y|X=x'})\|_1^2 \\
&\le D(\hat{P}_{Y|X=x'}||\hat{q}(\hat{P}_{Y|X=x'})) \\
&\le D(\hat{P}_{Y|X=x'}||q(\hat{P}_{Y|X=x})) \\
&\le D(P_{Y|X=x'}||q(\hat{P}_{Y|X=x})) \\
&+ \epsilon|D(P_{Y|X=x'}||q(\hat{P}_{Y|X=x'})) \\
&\quad - D(P_{Y|X=x'}||\hat{P}_{Y|X=x'})| \\
&\le D(P_{Y|X=x'}||q(\hat{P}_{Y|X=x})) + \epsilon(C_u - C_l) \\
&\le \sum_y \frac{\delta^2 \psi_{x'}(y)^2}{\hat{P}_{Y|X}(y|x_j)} + \epsilon(C_u - C_l) \\
&\le \frac{e^{C_u}\delta^2}{\min_{y:P_{Y|X}(y|z(x'))>0} P_{Y|Z}(y|z(x'))} \\
&\quad + \epsilon(C_u - C_l) \le a_1\epsilon,
\end{aligned}$$

where

$$a_1 := \frac{e^{C_u}c_1'^2}{\min_{y:P_{Y|Z}(y|z(x'))>0} P_{Y|Z}(y|z(x'))} + C_u - C_l > c_1'^2.$$

Notice that the minimum is taken over y's with nonzero probabilities due to the boundedness conditions in Assumption 1, which asserts $\phi_x(y) = \psi_x(y) \equiv 0$ for y's with zero probabilities. Finally, using triangular inequality:

$$\begin{aligned}
&\|P_{Y|X=x'} - \hat{q}(\hat{P}_{Y|X=x'})\|_1 \\
&\quad \le \|\hat{P}_{Y|X=x'} - \hat{q}(\hat{P}_{Y|X=x'})\|_1 + \\
&\qquad \|\hat{P}_{Y|X=x'} - P_{Y|X=x'}\|_1 \\
&\qquad\quad \le \sqrt{2a_1\epsilon} + \epsilon \le c_1\sqrt{\epsilon}
\end{aligned}$$

where $c_1 := \sqrt{2a_1} + 1$ is the coefficient in the lemma. □

Now we are ready to prove Theorem 1.

Proof of Theorem 1. Define the event $C_\epsilon := \{\sup_{x\in\mathbb{X}} D(P_{Y|X=x}||\hat{P}_{Y|X=x}) < \epsilon\}$. Further, suppose Θ is within the ball of radius R in $\mathbb{R}^d$. By Lemma 3, we have:

$$P(C_\epsilon) \ge 1 - \exp(-c_2 n\epsilon^2 + c_3(\epsilon)), \tag{19}$$

where $c_2 := \frac{\gamma^2}{2\rho^2(C_u - C_l)^2}$, $c_3(\epsilon) := d\log R(1 + \frac{8\rho}{\epsilon}) + \log 2 \ge \log 2|\mathcal{N}(\Theta, \frac{\epsilon}{4\rho})|$ (see e.g., (Vershynin, 2018), Section 4.2). For the subsequent discussion, suppose C_ϵ occurs. To prove that $\hat{z}$ achieves zero TER, it suffices to prove that $\hat{z}(x) = \hat{z}(x') \Leftrightarrow z(x) = z(x'), \forall x, x' \in \mathbb{X}$. To prove the "⇒" direction, suppose for some segment pairs $(x_1, x_2) \in \mathbb{X}^2$, $\hat{z}(x_1) = \hat{z}(x_2) = z'$ but $z(x_1) = z_1 \ne z(x_2) = z_2$. Invoke Lemma 4 and write $Q_{\hat{z}(x_j)} = P_{Y|X=x_j} + \delta\psi_{x_j}, \delta = c_1\epsilon^{1/4}, \psi_{x_j}^\top \mathbf{1} = 0, \|\psi_{x_j}\|_1 \le 1, j \in \{1,2\}$. Use the inequality $\log(1+x) \ge x - \frac{x^2}{1+x}, \forall x > -1$ we have

$$\begin{aligned}
&D_{\mathrm{KL}}(P_{Y|X=x_j}||Q_{\hat{z}(x_j)}) \\
&= -\sum_y P_{Y|X}(y|x_j)\log\left(1 + \frac{\delta\psi_{x_j}(y)}{P_{Y|X}(y|x_j)}\right) \\
&\le \sum_y \frac{e^{C_u}\delta^2\psi_{x_j}(y)^2}{P_{Y|X}(y|x_j)} \le a_2(z_1, z_2)\delta^2,
\end{aligned}$$

where $a_2(z_1, z_2) = \max_{j\in\{1,2\}} e^{C_u} / \min_{y:P_{Y|Z}(y|z_j)>0} P_{Y|Z}(y|z_j)$. As a result,

$$\begin{aligned}
&2a_2(z_1, z_2)\delta^2 \ge \\
&D_{\mathrm{KL}}(P_{Y|X=x_1}||Q_{z'}) + D_{\mathrm{KL}}(P_{Y|X=x_2}||Q_{z'}) \ge \\
&\qquad 2D_{\mathrm{JS}}(P_{Y|X=x_1}||P_{Y|X=x_2}),
\end{aligned} \tag{20}$$

which cannot be true if $\delta^2 \le \frac{D_{\mathrm{JS}}(P_{Y|Z=z_1}||P_{Y|Z=z_2})}{a_2(z_1,z_2)}$, or $\epsilon \le \frac{D_{\mathrm{JS}}(P_{Y|Z=z_1}||P_{Y|Z=z_2})^2}{c_1(z_1,z_2)^2 a_2(z_1,z_2)^2}$.

To prove the other direction, we use "⇒" to conclude that every phoneme occurs in at least one distinct cluster from other classes, since every cluster in $\hat{C}$ contains only a unique phoneme class. Further, define $E = \left\{\frac{1}{n}\min_z \sum_{i=1}^n \mathbf{1}_{Z_i=z} = 0\right\}$. Using Sanov's theorem (see e.g., (Cover and Thomas, 2006)), we have:

$$P(E) \le (n+1)^K \exp\left(-n \min_{P\in\mathbb{P}_E} D_{\mathrm{KL}}(P||P_Z)\right),$$

where $\mathbb{P}_E := \{P \in \Delta^K : \min_z P(z) = 0\}$. Use Assumption 4 and optimize the bound, we obtain

$$\begin{aligned}
&\min_{P\in\mathbb{P}_E} D_{\mathrm{KL}}(P||P_Z) \\
&\quad = \min_{P\in\mathbb{P}_E} D_{\mathrm{KL}}\left(P||\frac{1}{K}\mathbf{1}\right) \\
&\quad = \log K - \max_{P\in\mathbb{P}_E} H(P) = \log\frac{K}{K-1}
\end{aligned}$$

and

$$P(E) \le \exp\left(-n\log\frac{K}{K-1} + K\log(n+1)\right).$$

As a result, phonemes of each class occur at least once in the training set with high probability. If this is the case and if there exists some $x, x' \in \mathbb{X}$ such that $z(x) = z(x')$ but $\hat{z}(x) \neq \hat{z}(x')$, $\hat{C}$ contains at least $K+1$ clusters, which contradicts Assumption 4. Therefore, define the event $R := \{\hat{z}(X) = \hat{z}(X') \Leftrightarrow z(X) = z(X')\}$, the token error rate can be upper bounded as

$$\begin{aligned}
&P_{\text{TER}}(\hat{z}) \\
\leq &P(C_\epsilon \cap E^c)\mathbb{P}\left\{R|C_\epsilon \cap E^c\right\} + P(C_\epsilon^c \cup E) \\
= &\exp\left(-n \min\{e_1(n, \epsilon^*), e_2(n, K)\}\right),
\end{aligned}$$

where

$$\begin{aligned}
\epsilon^* &:= \min_{z_1 \neq z_2} \frac{D_{\text{JS}}(P_{Y|Z=z_1}||P_{Y|Z=z_2})^2}{c_1(z_1, z_2)^2 a_2(z_1, z_2)^2} \\
&=: \min_{z_1 \neq z_2} c(z_1, z_2) D_{\text{JS}}(P_{Y|Z=z_1}||P_{Y|Z=z_2})^2 \\
e_1(\epsilon^*) &:= c_2 \epsilon^{*2} - \frac{c_3(\epsilon^*)}{n} \\
e_2 &:= \log \frac{K}{K-1} - \frac{K \log(n+1)}{n}.
\end{aligned}$$

Therefore, $P_{\text{TER}}(\hat{z}) \leq \delta$ amounts to

$$\begin{aligned}
c_2 n \epsilon^{*2} - c_3(\epsilon^*) &\geq \log \frac{1}{\delta} \\
n \log \frac{K}{K-1} - K \log(n+1) &\geq \log \frac{1}{\delta}.
\end{aligned}$$

The first inequality implies

$$n \geq \frac{\log c_3(\epsilon^*) + (1/\delta)}{c_2 \epsilon^{*2}} = O\left(\frac{\log(\frac{1}{\delta})}{\epsilon^{*2}}\right).$$

For the second inequality, rearranging the terms we obtain:

$$n \geq \frac{K}{\log \frac{K}{K-1}} \log n + \frac{\log \frac{1}{\delta}}{\log \frac{K}{K-1}}, \tag{21}$$

which by Lemma A.2 from (Shalev-Shwartz and Ben-David, 2014) holds if

$$\begin{aligned}
n \geq &\frac{4K \log \frac{2K}{\log \frac{K}{K-1}} + 2 \log \frac{1}{\delta}}{\log \frac{K}{K-1}} \\
&= O\left(\frac{\log \frac{1}{\delta}}{\log \frac{K}{K-1}}\right).
\end{aligned} \tag{22}$$

Combining Equation 21 and Equation 22 proves the theorem. □

B Collection Process and Statistics of the Spoken Word Datasets

The dataset statistics of all the datasets used for our experiments are shown in Table 4. We collect all the spoken word datasets from existing datasets in the following steps:

1. *Decide the train-test split*: For Flickr audio, we use the original training and validation set to extract spoken words for the training set and the test set to extract words for test set; for LibriSpeech, we use train-clean-100 and train-clean-360 for training set and dev-clean for test set; for TIMIT and Mboshi, we use the whole dataset without `SA` utterances to extract spoken words, to be consistent with prior works. For the latter, it will not lead to overfitting since our setting is unsupervised in a sense that the target label, phoneme, is not available during training.

2. *Decide the phoneme inventory*: The phoneme inventory of the English corpora such as Flickr audio, LibriSpeech and TIMIT are the standard 61 phonemes from TIMIT merged into 39 classes for LibriSpeech and 44 classes for Flickr Audio, due to slightly different phoneme set required for the forced alignment systems used to extract phoneme and word boundaries. The phoneme inventory of Mboshi is provided in (Godard et al., 2017).

3. *Decide the vocabulary*: For English corpora, we use a neural dependency parser (Gardner et al., 2017) to extract head words of noun phrases from the Flickr30kEntities and choose those with frequency more than 500 times in the entire Flickr30k corpus. For Mboshi, we use the bigrams and trigrams as proxy for words.

4. *Word and phoneme boundary detection*: For evaluation purposes, we need to extract word and phoneme boundaries for the utterances. While TIMIT and Mboshi has provided frame-level phoneme transcriptions, such labels are not available for Flickr Audio and LibriSpeech. Therefore, we use the Montreal forced aligner to extract word and phoneme boundaries for LibriSpeech and another HMM-DNN hybrid ASR system to extract segment boundaries for Flickr audio

	Flickr Audio	LibriSpeech			TIMIT			Mboshi	
$\lvert\mathbb{Y}\rvert$	224	224	524	824	224	524	824	161	377
K	44	39	39	39	39	39	39	31	31
#train words	46569	50073	143512	188863	1289	1678	2348	30290	82606
#test words	6557	595	595	595	1289	1678	2348	30290	82606
#phonemes	318756	223821	590647	816754	5501	7692	11874	93236	165212
#hours	6.1	6.3	15.4	21.2	0.1	0.1	0.2	2.2	4.1

Table 4: Statistics of four spoken word datasets used for experiments. Mboshi has the same number of training and test words since the whole datasets are used for both training and evaluation, consistent with prior works (Yusuf et al., 2020; Feng et al., 2021b).

5. *Extract spoken word utterances*: To keep the dataset as balanced as possible, we set a cutoff on the maximal number of word utterances per class, which is set to be 200 for Flickr Audio and 1000 for LibriSpeech, TIMIT and Mboshi.

C Model Implementation

For the pre-segmentation stage in Figure 2 of IQ, we use the self-supervised model proposed in (Kreuk et al., 2020) to predict the phoneme-level segmentation for English datasets, and the segmentation generated by one of our baselines (Feng et al., 2021a) for experiments on Mboshi language. The segmental speech encoder $e_{\theta_1}(\cdot)$ is a CPC model pretrained on the whole 960h LibriSpeech (Nguyen et al., 2020) with 256-dimensional representation for each 10ms frame followed by averaging across each segments. The word posterior $c_{\theta_2}(\cdot)$ for the joint distribution learning stage consists of four hidden layers and 512 ReLU units per layer with layer normalization and one softmax output layer. All our models are trained for 20 epochs using Adam optimizer (Kingma and Ba, 2014) with learning rate of 0.001 decayed by 0.97 every 2 epochs and a batch size of 8. We slightly modify (P_1) analogous to the VQ-VAE (van den Oord et al., 2017) to make it more suitable for gradient-based optimization:

$$\mathcal{L}_{\text{IQ-VAE}}(P_n, \theta, q) := \mathcal{L}_{\text{CE}}(P_n, \theta) + \lambda \mathbb{E}_{P_n}[D_{\text{KL}}(\text{sg}[P^{\theta}_{Y|X}] || q(P^{\theta}_{Y|X})) + D_{\text{KL}}(P^{\theta}_{Y|X} || \text{sg}[q(P^{\theta}_{Y|X})])]$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operation and $\lambda = 0.5$ for all experiments. Exponential moving average (EMA) codebook update is used with a decay rate of 0.999 to optimize the first KL term. Each code distribution is initialized using a symmetric Dirichlet distribution with a concentration parameter of 100.

For CPC, wav2vec and wav2vec 2.0, we extract discrete units using the same predicted and gold segmentations as our IQ model using k-means clustering with the same number of clusters ($K = 31$).

D Convergence Plot for Word-Level Phoneme Discovery

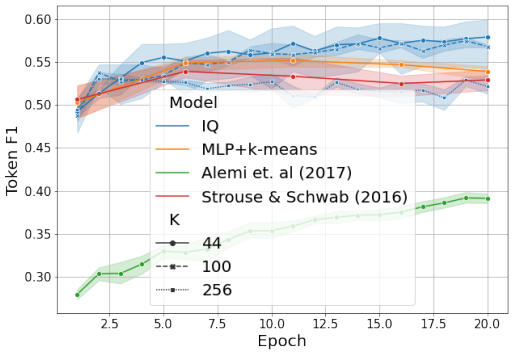


Figure 6: Token F1 convergence plot of various models on Flickr audio.

The convergence plot of Token F1 during training of IQ on Flickr Audio compared to the baselines is shown in Figure 6.

E Further Analysis of Representations Learned by IQ

The visualizations of the estimated distributions $P^{\theta}_{Y|X}$ using t-SNE (van der Maaten and Hinton, 2008) on Mboshi are shown in Figure 7. We again observe that IQ is capable of clustering phonemes from the same manner class as shown in the t-SNE plots for TIMIT in the main text. We also show the most confusing phoneme pairs for both datasets in Table 8a and Table 8b respectively, where the *error*

probability for a phoneme pair is defined as the probability that segments of different phonemes in the pair are assigned to the same cluster. While we can see that most phoneme pairs confused by the model are acoustically very similar such as (/ae/, /aa/), (/z/, /s/) in TIMIT and (/e/, /a/), (/bv/, /b/) in Mboshi, we also observe some non-obvious pairs such as the pair (/ch/, /ah/) in TIMIT. From the confusion matrix shown in Fig 33b, we can see that this is due to the high variability and potentially lack of samples for the vowel /ah/, which makes its cluster more likely to be merged by other bigger clusters. A more general reason for the model to confuse between such non-obvious pairs may be that distinguishing such phonemes is not very useful in discriminating those words used during the IQ training, which is possible since the vocabulary size during training is relatively small (<1000).

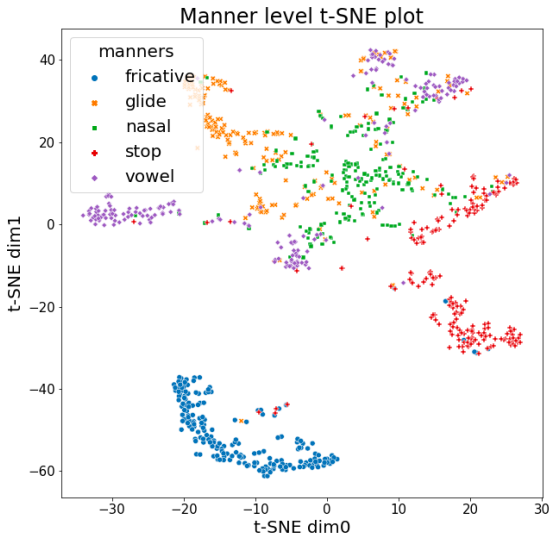


Figure 7: Manner-level t-SNE plots of phoneme clusters discovered by IQ with $|\mathbb{Y}| = 161$ and gold segmentation on Mboshi

Phoneme Pair	Error Prob.
ae, aa	1.00
ch, ah	0.85
sh, s	0.82
ah, aa	0.82
aw, aa	0.77
z, s	0.75
n, m	0.73
p, k	0.70
r, er	0.67
iy, ey	0.60

(8a) Top-10 most confusing phoneme pairs by IQ with $|\mathbb{Y}| = 824$ and predicted segmentation on TIMIT

Phoneme Pair	Error Prob.
a, Ng	1.00
bv, b	0.82
e, a	0.79
t̪, s	0.77
i, e	0.73
b, Ng	0.68
p, k	0.68
f, a	0.59
g, a	0.59
o, mw	0.56

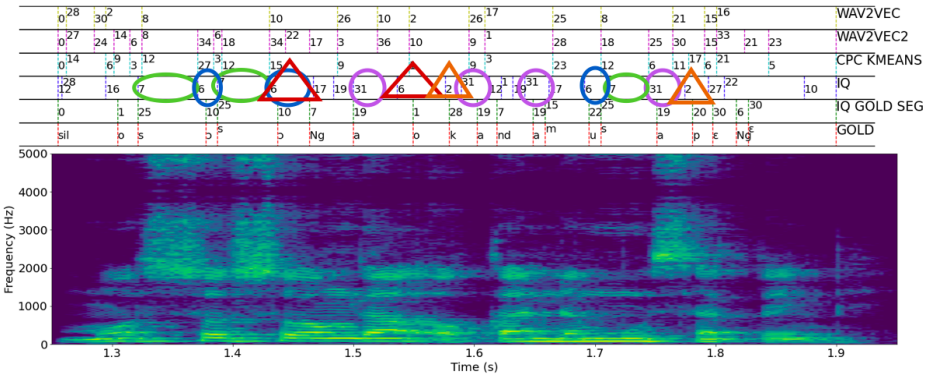
(8b) Top-10 most confusing phoneme pairs by IQ with $|\mathbb{Y}| = 161$ with predicted segmentation on Mboshi

F Effect of Codebook Size for IQ

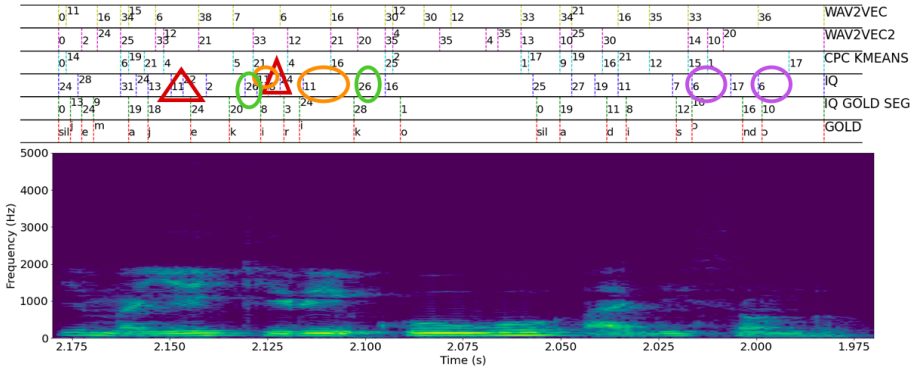
The phoneme discovery results of IQ with different codebook sizes on Mboshi and TIMIT are shown in Table 5 and Table 6 respectively. As discussed in the paper, our IQ model achieving equally good NMI and boundary F1 and is thus robust to the codebook size on both datasets.

G More Speech Examples

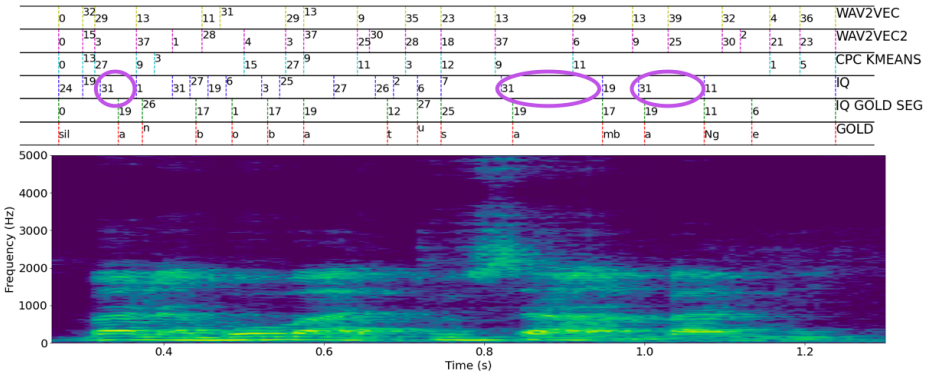
Lastly, we provide eight more spoken utterances annotated with phoneme discovery results.



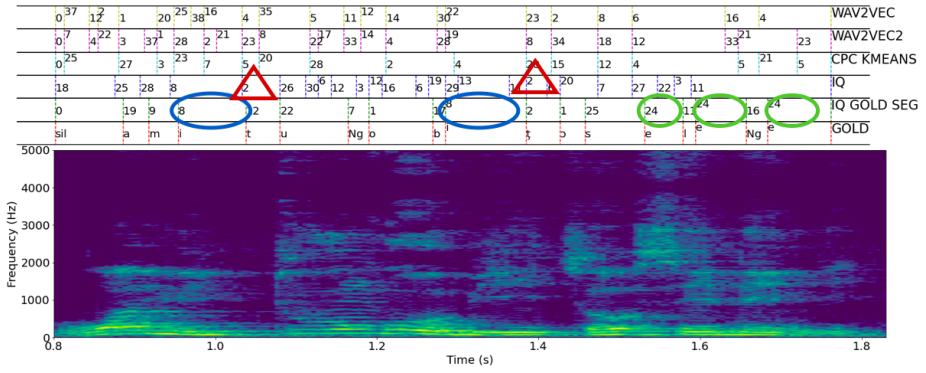
(9a)



(9b)

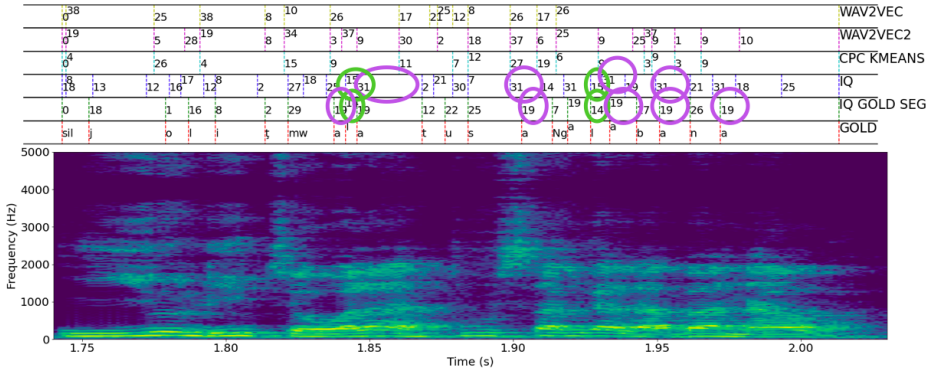


(9c)

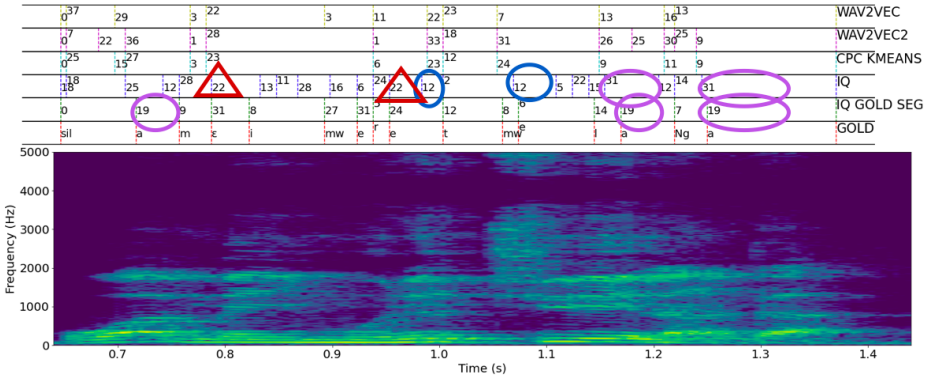


(9d)

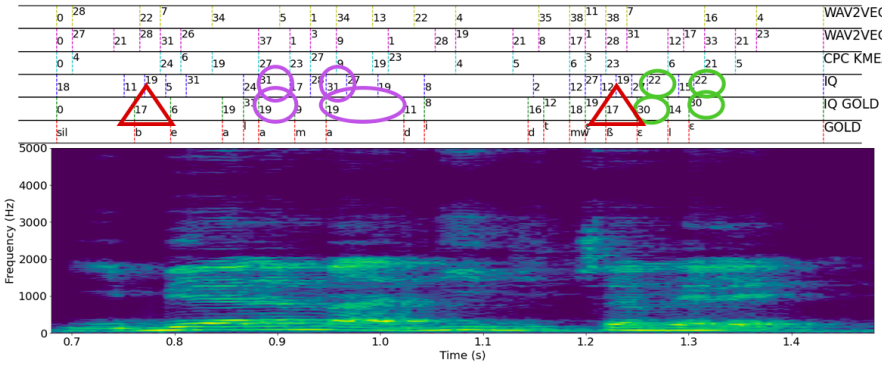
Figure 9: The spectrograms annotated with the gold transcripts and the zero-resource transcriptions by various models for four more utterances from Mboshi. The spoken segments in circles of the same colors are phonemes correctly identified by our IQ model without gold segmentation, and those in triangles of the same color are incorrect pairs that are acoustically similar.



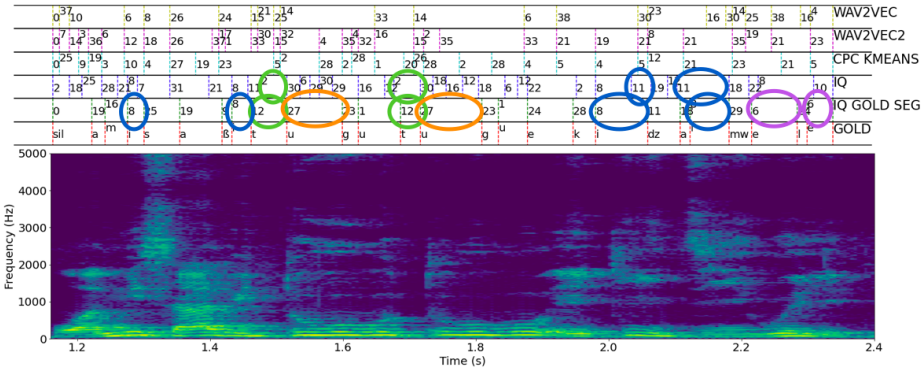
(10a)



(10b)



(10c)



(10d)

Figure 10: (Continued) the spectrograms annotated with the gold transcripts and the zero-resource transcriptions by various models for four more utterances from Mboshi. The spoken segments in circles of the same colors are phonemes correctly identified by our IQ model without gold segmentation, and those in triangles of the same color are incorrect pairs that are acoustically similar.

Codebook size		30	40	50	60	70
$\|\mathbb{Y}\| = 224$	Token F1	**51.2**±1.0	50.9±0.8	50.3±0.6	49.0±1.2	49.0±0.4
	NMI	43.0±0.7	43.4±0.9	**43.6**±0.3	43.1±0.7	43.5±0.5
	Boundary F1	77.7±0.5	**78.6**±0.4	78.2±0.3	78.1±0.6	78.3±0.6
$\|\mathbb{Y}\| = 524$	Token F1	53.5±0.8	**53.9**±0.3	53.0±0.9	52.0±0.9	52.5±0.7
	NMI	46.8±0.6	46.7±0.2	46.7±0.4	46.9±0.3	**47.3**±0.2
	Boundary F1	**80.4**±0.2	**80.4**±0.2	80.3±0.1	80.2±0.1	80.3±0.1
$\|\mathbb{Y}\| = 824$	Token F1	53.7±0.5	**54.4**±0.4	53.3±0.4	52.6±0.8	50.7±0.9
	NMI	47.1±0.4	**47.5**±0.2	47.3±0.2	47.4±0.4	47.1±0.4
	Boundary F1	80.6±0.0	**80.5**±0.1	80.4±0.1	80.3±0.0	80.3±0.0

Table 5: Phoneme discovery performance vs. codebook size on TIMIT. The models used are IQs trained on LibriSpeech+TIMIT.

Codebook size		30	40	50	60	70
$\|\mathbb{Y}\| = 161$	Token F1	**54.2**±1.0	54.2±0.2	51.1±0.9	54.0±0.7	45.9±0.8
	NMI	**45.1**±0.4	44.0±0.4	44.7±0.2	44.3±0.7	44.3±0.5
	Boundary F1	**67.5**±0.0	67.4±0.1	67.3±0.1	67.3±0.1	66.8±0.0
$\|\mathbb{Y}\| = 377$	Token F1	57.1±1.0	**57.2**±1.1	56.7±1.6	56.8±1.1	55.2±0.4
	NMI	49.3±0.3	49.0±0.1	**49.8**±0.2	49.6±0.4	49.5±0.6
	Boundary F1	**67.3**±0.1	**67.3**±0.1	**67.3**±0.1	67.1±0.2	67.0±0.0

Table 6: Phoneme discovery performance vs codebook size on Mboshi. The models used are IQs with CPC+BNF features.