LEARNING PHONEME-LEVEL DISCRETE SPEECH REPRESENTATION WITH WORD-LEVEL SUPERVISION

Anonymous authors
Paper under double-blind review

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 long-standing 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. Under mild assumptions, we prove that the phoneme inventory learned by our approach converges to the true one with exponentially low error rate. Moreover, in experiments on TIMIT and Mboshi benchmarks, our approach consistently learns better phoneme-level representation than previous state-of-the-art self-supervised representation learning algorithms and remains effective even in a low-resource scenario.

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

supervision is available. Such supervision can be a small set of word-level category labels that are available from the word's use in naturalistic settings. One such label might be the name of a visual object semantically related to the spoken word, e.g., as used by infants learning the phoneme inventory of their first language (L1). Another such label might be the translation of the spoken word into a written form in another language, e.g., as used by second-language (L2) learners.

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 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 class of neural networks called *information quantizers* 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 & Glass (2005); Lee & 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 & 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 & 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 phoneme types (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, (Żelasko et al. (2020); 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. (Harwath & 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)).

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: Given a spoken word segment $\boldsymbol{x} = [x_1, \cdots, x_T]$ of word type y, where x_t's are phoneme segments within that word, if we switch any segment x_t to x'_t of a different phoneme type, the segment $\boldsymbol{x}' = [x_{1:t-1}, x'_t, x_{t+1:T}]$ will represent a different word type $y' \neq y$. On the other hand, if x_t and x'_t are of the same phoneme type, $\boldsymbol{x}'$ will have the same word type y as $\boldsymbol{x}$. 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 type that X is part of. 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, the conditional distribution of the semantic variable given any instance of the same phoneme is the same. This property will be referred later as the **distributional property of phonemes**.

This definition is the generalization of Definition 1, as shown in the following proposition.

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 segment $\boldsymbol{x} = [x_1, \cdots, x_T]$, and for any phonetic segment pairs $(x_t, x'_t) \in \mathbb{Z}_k^2, k \in \{1, \cdots, K\}$, changing x_t to x'_t does not alter the identity of the word, i.e.,*

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

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.,

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

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

The proof can be found in Appendix A. Definition 2 preserve the semantic property of phonemes characterized in Definition 1, but does not require expensive pairwise labels of semantic change. Further, such a definition is flexible enough to allow further generalization by relaxing Y to be *any* semantic signal related to the spoken utterance, which may include images or translations to a different language.

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 phoneme segment X is randomly chosen from $\mathbb{X}$ with probability distribution P_X and random variable $Z := z(X)$ is its phoneme label, then by the distributional property of phonemes, 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

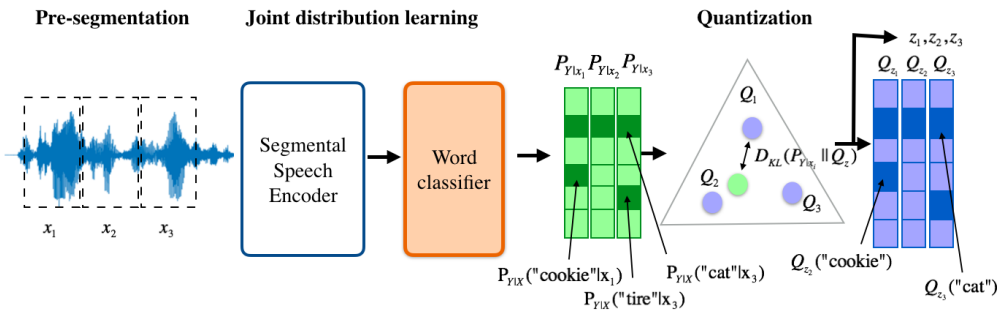


Figure 1: Network architecture of information quantizer

phoneme inventory is thereby completely characterized by the phoneme label function $z(\cdot)$ as well as the set of distributions associated with each class $P_{Y|Z}$.

Further, instead of explicitly minimizing K as required by the definition, we can fix K either based on prior knowledge of the language or by simply gradually decreasing K until the distributional property of phonemes is no longer feasible.

3.3 Problem Formulation

Let $z(\cdot)$ be the phoneme assignment function from Definition 2 and assuming 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 a spoken phonetic segment extracted from a spoken word segment, and each $y^{(i)} \in \mathbb{Y}$ is the corresponding word type label, a semantic-driven phoneme discovery (SPD) system tries to find an assignment function that minimizes the **token error rate (TER)**:

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

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)). Please refer to Appendix A.2 for details. 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 information quantizer (IQ), as depicted in Figure 1. 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 classifier $c_{\theta_2}(\cdot)$ 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 an 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 classifier network to estimate the phoneme-level conditional word distribution as:

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

Note that it is crucial that no recurrent connection exists between segments since our goal is to learn the probability of words given each individual phoneme segments. 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:

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

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:

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

where $\tilde{P}_n := \frac{1}{n}\sum_{i=1}^{n} P^{\theta}_{Y|X}$ is the *smoothed* version of the empirical distribution.

The final loss function of IQ for SPD is then:

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

where $\lambda > 0$ is some hyperparameter set to approximately 1 for most experiments.

Further, we restrict q to be **nearest-neighbor** so that:

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

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 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. The uniform prior is chosen as it approximates phoneme prior sufficiently well, though other prior can also be used.

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):*

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

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:

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

where $\epsilon^* = \min_{z_1, z_2 : z_1 \neq z_2} c(z_1, z_2) D_{JS}(P_{Y|Z=z_1} || P_{Y|Z=z_2})^2$ for some constants $c(z_1, z_2) > 0, 1 \leq z_1, z_2 \leq K$ independent of n, δ, $O(x)$ is such that $\lim_{x \to \infty} \frac{O(x)}{x} = 0$ and $D_{JS}(P||Q) := \frac{1}{2} D_{\mathrm{KL}}\left(P||\frac{P+Q}{2}\right) + \frac{1}{2} D_{\mathrm{KL}}\left(Q||\frac{P+Q}{2}\right)$ 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 phoneme types. 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:

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

which is minimized if $\hat{q}(P_{Y|X}) = P_{Y|z(X)}$. In fact, we can show that $\hat{z}$ for such $\hat{q}$ is equivalent to $z(\cdot)$ up to a permutation. We will defer the proofs to Appendix A.3.

5 Experiments

5.1 Implementation Details

For the pre-segmentation stage in Figure 1, 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 baseline (Feng et al. (2021a)) for experiments on Mboshi language. The segmental speech encoder $e_{\theta_1}(\cdot)$ is a CPC model pretrained on Librispeech (Nguyen et al. (2020)) with 256-dimensional representation for each 10ms frame followed by averaging across each segments. The word classifier $c_{\theta_2}(\cdot)$ 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 & 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:

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

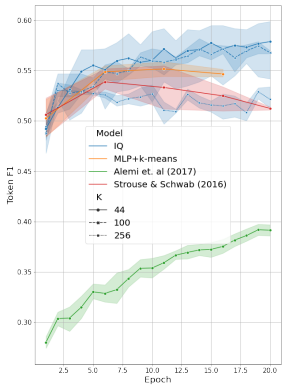
where $\text{sg}[\cdot]$ denotes the stop-gradient operation and $\lambda = 0.5$ for all experiments. Exponential moving average (EMA) codebook update is used with a decay rate of 0.999 to optimize the first KL term. Each code distribution is initialized using a symmetric Dirichlet distribution with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

5.2 Word-level Phoneme Inventory Discovery

Datasets We construct four training datasets consisting of spoken words only. The details can be found in Appendix B. The visual words are extracted by selecting the 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, the spoken word data are extracted from Flickr audio dataset (Harwath & Glass (2015)). The Librispeech and TIMIT word dataset with $|\mathbb{Y}| = 224$ uses the same set of

	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 & Schwab (2016)), K=44	49.4±1.0	52.2±0.2	68.3±1.3	72.8±1.0
(Strouse & Schwab (2016)), K=100	41.7±0.7	52.8±0.1	60.3±0.0	71.0±0.5
(Strouse & 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

(2a)



(2b)

Figure 2: (a) In-domain phoneme discovery results on Flickr audio and Librispeech. (b) Convergence plot of various models on Flickr audio.

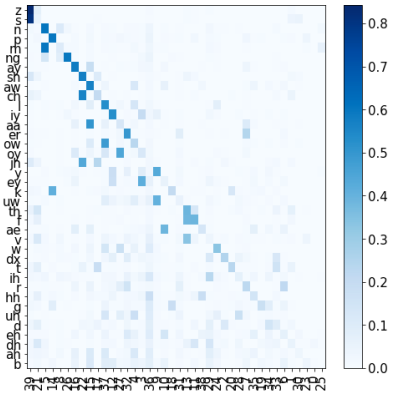
word types from Flickr30k and spoken word tokens from Librispeech (Vassil et al. (2015)) 460-hour train-clean subset, resulting in a dataset of about 6 hours and 0.1 hours; for Librispeech and TIMIT word dataset with $|\mathbb{Y}| = 524$ and $|\mathbb{Y}| = 824$, we supplement the dataset with the speech 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 all n-grams except uni- and bigrams or bigrams+trigrams that occur more than 100 times as "words", resulting in a vocabulary size of 161 and 377 respectively. Note that the amount of 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. For each training dataset, we test our models on spoken words whose types have appeared during training. The best results during testing are reported for all models.

Evaluation metrics Standard metrics such as NMI for the quality of the codebook and boundary F1 for the quality of segmentation respectively are used to evaluate the models. We used the same evaluation code used in (Yusuf et al. (2020), Feng et al. (2021b)) with a tolerance of 20ms for boundary F1. In addition, we also report token F1 (Dunbar et al. (2017)), which is more sensitive to the purity of the clusters.

Baselines We compare our model (IQ) with 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 (Gumbel VIB) (Alemi et al. (2017) 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 & 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. All models share the same speech encoder as IQ.

Results The results on visual word-only test sets of Flickr audio and Librispeech are shown in Table 2a. On both datasets, IQ outperforms both Gumbel VIB and DIB in terms of all metrics, especially on Flickr audio, which has more phoneme types than Librispeech and a larger test set. Further, models assuming a deterministic relation between phonemes and speech tend to perform better than models without such an inductive bias such as Gumbel VIB, suggesting the relation

between phonemes and speech is approximately deterministic. 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.

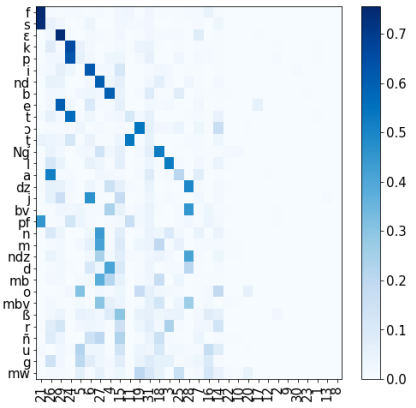


(3a)

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

(3b)

Figure 3: (a) 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 $\max_{z'} P_{\hat{Z}|Z}(z'|z)$. (b) The overall phoneme discovery results of all models on TIMIT.



(4a)

	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, $\|\mathbb{Y}\| = 161$, 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, $\|\mathbb{Y}\| = 377$, 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

(4b)

Figure 4: (a) The distribution of codeword assignment for each phoneme by IQ with CPC+BNF features, vocab size $|\mathbb{Y}| = 377$ and predicted segmentation on Mboshi. 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 $\max_{z'} P_{\hat{Z}|Z}(z'|z)$. (b) Phoneme discovery results of all models on Mboshi dataset.

5.3 Whole-sentence Phoneme Inventory Discovery

Datasets We also test our models on two standard 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 we follow the split in (Yusuf et al. (2020)) to use the entire corpus excluding `SA` utterance for both training and testing, which is acceptable since our setting is unsupervised. On this dataset, we experiment with models pretrained on spoken word subsets of Librispeech as well as the combined subsets of Librispeech and TIMIT. Another benchmark dataset, Mboshi (Godard et al. (2017)), contains about 2.4 hours speech from

a low-resource language and we again follow (Yusuf et al. (2020), Feng et al. (2021b)) to use the entire corpus for training and testing. The best results during testing are reported for all models.

Baselines For the whole-sentence datasets, we compare our models with the state-of-the-art unsupervised and weakly supervised phoneme discovery systems, namely, the unsupervised H-SHMM trained with untranscribed 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 transcribed multilingual speech.

Result on TIMIT The results on TIMIT are shown in Table 3b. First of all, 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. Interestingly, we observe a diminishing return in all metrics as the vocabulary size increases, possibly because words ranked between 524-824th have very low frequency and may not increase the JS divergence between $P_{Y|Z}$'s of phonemes much. As expected the NMI of our model degrades by about 8% on TIMIT compared to that of word-level SPD in Librispeech, due to the challenge of generalization and improves by about 1% improvement as we include words from TIMIT. In addition, the use of unsupervised phoneme segmentation deteriorates the NMI by about 18% absolute for our models since the distributional property of phonemes does not apply exactly to non-phoneme segments. From Figure 3a, we observe that the codeword assignments by IQ correlates well with the actual phonemes. Further, most phonemes confused by the model appears to fall into the same manner class, such as nasals "n" and "m" as well as affricates "ch" and "jh". In addition, vowel appears to be the hardest manner class for the model to learn. From Figure 33a, vowels such as "eh" and "ah" are assigned to many different codewords, exhibiting high within-class variability; vowel pairs such as "aw" and "aa" are often assigned the same codeword, exhibiting high between-class variability for vowels. These observations are confirmed by further analysis using t-SNE (van der Maaten & Hinton (2008)) on the estimated conditional distributions $P^{\theta}_{Y|X}$ and the top-10 most confusing phoneme pairs, both of which can be found in Appendix C.

Result on Mboshi The results on Mboshi are shown in Table 4b. As we can see, when $|\mathbb{Y}| = 377$, IQs with both CPC and CPC+BNF features outperform all baselines in all three metrics with or without gold segmentation; when $|\mathbb{Y}| = 161$ and gold segmentation is available, IQs with CPC and CPC+BNF features outperform (Feng et al. (2021b)) in all three metrics. While being equivalent for perfect assignment, we find token F1 and NMI behave quite differently empirically, as for $|\mathbb{Y}| = 161$, IQ with CPC features and predicted segmentation outperforms the best baseline (Feng et al. (2021b), 13 languages) in terms of token and boundary F1, but worse in NMI. Adding multilingual BNF to the same model significantly improves the result and allows the model to outperform the baselines in terms of NMI as well. In all cases, adding multilingual BNF, that is, concatenating the multilingual BNF from (Feng et al. (2021b)) to the CPC output representation from the segmental speech encoder in Figure 1 further improves the result, suggesting that multilingual information and word-level semantic information are complimentary. By comparing IQ with $|\mathbb{Y}| = 161$ and $|\mathbb{Y}| = 377$, we find that increasing the vocabulary size again improves the performance of our system. From Figure 4a, we observe similar trends as in TIMIT, and the confusion between phonemes within each manner class such as nasals "n", "m" and "mb" is more severe, both because the more diverse set of nasal phonemes and because the relatively small vocabulary size. More visualization and analysis of the phoneme representation can be found in Appendix C.

Effect of codebook size we find that the quality of codeword assignments by IQs is robust against varying codebook size, after experimenting with codebook size from 30 to 70 on TIMIT and Mboshi. The detailed results can be found in Appendix D.

6 Conclusion

Motivated by the linguistic definition of phonemes, we propose information quantizer (IQ), a new neural network model with theoretical guarantee and strong empirical performance for self-supervised learning of phoneme inventory with word-level supervision.

REFERENCES

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

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

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

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Unsupervised speech recognition. In *ArKiv*, 2021. URL https://arxiv.org/pdf/2105.11084.pdf.

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

K.-Y. Chen, C.-P. Tsai, D.-R. Liu, H.-Y. Lee, and L. shan Lee. 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)*, 2019.

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

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

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

Ewan Dunbar, Xuan-Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. The zero resource speech challenge 2017. *CoRR*, abs/1712.04313, 2017. URL http://arxiv.org/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. Unsupervised acoustic unit discovery for speech synthesis using discrete latent-variable neural networks. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019b.

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

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

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

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

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. A very low resource language speech corpus for computational language documentation experiments. *CoRR*, abs/1710.03501, 2017. URL http://arxiv.org/abs/1710.03501.

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

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

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

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

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

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. URL https://www.sciencedirect.com/science/article/pii/0893608089900208.

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

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

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

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

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

Chiaying Lee and James Glass. A nonparametric Bayesian approach to acoustic model discovery. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pp. 40–49, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. 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*, 2020.

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

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

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

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

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

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

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

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

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *ArXiv*, 2019. URL https://arxiv.org/pdf/1807.03748.pdf.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

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

Roman Vershynin. *High-Dimensional Probability–An Introduction with Applications in Data Science*. Cambridge University Press, Sept. 2018. doi: https://doi.org/10.1017/9781108231596.

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

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

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

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 $1 \geq \text{F1}(\hat{z}) \geq \min\{\text{Prec}(\hat{z}), \text{Rec}(\hat{z})\} \geq 1 - P_{\text{e, TER}}(\hat{z}) = 1$, 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 $\mathbb{Q}_K^{NN}$ 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*

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

Therefore, $(\hat{\theta}, \hat{q})$ 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 $\min_{q \in Q_K^{\text{NN}}} \mathcal{L}_Q(\tilde{P}_n, q) \leq \min_{q \in Q_K} \mathcal{L}_Q(\tilde{P}_n, q)$. This is true since

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

□

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:*

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

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,

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

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

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

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 $\theta_k \in \mathcal{N}(\Theta, \frac{\epsilon}{4\rho})$, and let $\Delta_n(\theta) := D_n(P_{Y|X}||P^{\theta}_{Y|X}) - D_{\mathrm{KL}}(P_{Y|X}||P^{\theta}_{Y|X})$ we have $\forall\theta\in\Theta_k$,

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

Further, by Assumption 2, we have

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

As a result,

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

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

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

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

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

for some $\theta^* \in \Theta^*$. Therefore, using the local convexity property of the KL divergence around minima in Lemma 2, we arrive at the desired result:

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

□

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 $1 \leq z \leq K$, there exists $x \in \mathbb{X}$ such that $z(x) = z$. Further, suppose there exists $\epsilon > 0$ such that $\|\hat{P}_{Y|X=x} - P_{Y|X=x}\|_1 \leq \epsilon, \forall x \in \mathcal{D}$. Then, $\forall x \in \mathbb{X}$, $\|\hat{q}(\hat{P}_{Y|X=x}) - P_{Y|X=x}\|_1 \leq c_1 \epsilon^{1/2}$ 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,

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

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

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

for every $x \in \mathcal{D}$. Therefore, to maintain (*), we need $\delta^2 \leq 4C_u|\mathbb{Y}|\epsilon$ 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 $\log(1+x) \geq x - \frac{x^2}{1+x}, \forall x > -1$, we have

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

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

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

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:

$$P(C_\epsilon) \geq 1 - \exp(-c_2 n\epsilon^2 + c_3(\epsilon)), \tag{18}$$

where $c_2 := \frac{\gamma^2}{2\rho^2(C_u - C_l)^2}, c_3(\epsilon) := d\log R(1+\frac{8\rho}{\epsilon}) + \log 2 \geq \log 2|\mathcal{N}(\Theta, \frac{\epsilon}{4\rho})|$ (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 "$\Rightarrow$" direction, suppose for some segment pairs $(x_1, x_2) \in \mathbb{X}^2$, $\hat{z}(x_1) = \hat{z}(x_2) = z'$ but $z(x_1) = z_1 \neq z(x_2) = z_2$. Invoke Lemma 4 and write $Q_{\hat{z}(x_j)} = P_{Y|X=x_j} + \delta\psi_{x_j}, \delta = c_1\epsilon^{1/4}, \psi_{x_j}^\top \mathbf{1} = 0, \|\psi_{x_j}\|_1 \leq 1, j \in \{1,2\}$. Use the inequality $\log(1+x) \geq x - \frac{x^2}{1+x}, \forall x > -1$ we have

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

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,

$$2a_2(z_1, z_2)\delta^2 \geq D_{\mathrm{KL}}(P_{Y|X=x_1}||Q_{z'}) + D_{\mathrm{KL}}(P_{Y|X=x_2}||Q_{z'}) \geq 2D_{\mathrm{JS}}(P_{Y|X=x_1}||P_{Y|X=x_2}), \tag{19}$$

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

To prove the other direction, we use "$\Rightarrow$" to conclude that every phoneme type 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 & Thomas (2006)), we have:

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

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

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

and

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

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, the token error rate can be upper bounded as

$$\begin{aligned} P_{\mathrm{TER}}(\hat{z}) &\leq P(C_\epsilon \cap E^c)\mathbb{P}\left\{\hat{z}(X) = \hat{z}(X') \Leftrightarrow z(X) = z(X')|C_\epsilon \cap E^c\right\} + P(C_\epsilon^c \cup E) \\ &= \exp\left(-\min\left\{c_2 n\epsilon^{*2} - c_3(\epsilon^*), n\log\frac{K}{K-1} - K\log(n+1)\right\}\right), \end{aligned}$$

with $\epsilon^* := \min_{z_1\neq z_2} \frac{D_{\mathrm{JS}}(P_{Y|Z=z_1}||P_{Y|Z=z_2})^2}{c_1(z_1,z_2)^2 a_2(z_1,z_2)^2} =: \min_{z_1\neq z_2} c(z_1, z_2) D_{\mathrm{JS}}(P_{Y|Z=z_1}||P_{Y|Z=z_2})^2$. Therefore, $P_{\mathrm{TER}}(\hat{z}) \leq \delta$ 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:

$$n \geq \frac{K}{\log \frac{K}{K-1}} \log n + \frac{\log \frac{1}{\delta}}{\log \frac{K}{K-1}}, \tag{20}$$

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

$$n \geq \frac{4K \log \frac{2K}{\log \frac{K}{K-1}} + 2 \log \frac{1}{\delta}}{\log \frac{K}{K-1}} = O\left(\frac{\log \frac{1}{\delta}}{\log \frac{K}{K-1}}\right). \tag{21}$$

Combining Equation 20 and Equation 21 proves the theorem. □

B Spoken Word Dataset Statistics

The dataset statistics of all the datasets used for our experiments are shown in Table 1.

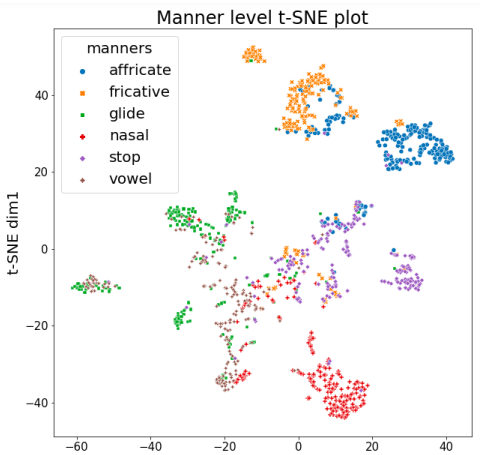
Table 1: Statistics of four spoken word datasets used for experiments. TIMIT and Mboshi have the same number of training and test words since the whole datasets are used for both training and evaluation, consistent with prior works (Yusuf et al. (2020), Feng et al. (2021b)).

	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

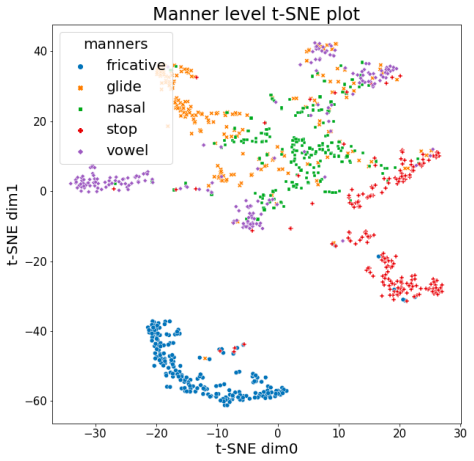
C Further analysis of representations learned by IQ

The visualizations of the estimated distributions $P^{\theta}_{Y|X}$ using t-SNE (van der Maaten & Hinton (2008)) on TIMIT and Mboshi are shown in Figure 5.

D Effect of Codebook Size for IQ



(5a) Manner-level t-SNE plot by IQ with $|\mathbb{Y}| = 824$ and gold segmentation on TIMIT



(5b) Manner-level t-SNE plots of phoneme clusters discovered by IQ with $|\mathbb{Y}| = 161$ and gold segmentation on Mboshi

Figure 5: t-SNE plot by IQ on TIMIT and Mboshi

(6a) Top-10 most confusing phoneme pairs by IQ with $|\mathbb{Y}| = 824$ and predicted segmentation on TIMIT

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

(6b) Top-10 most confusing phoneme pairs by IQ with $|\mathbb{Y}| = 161$ with predicted segmentation on Mboshi

Phoneme Pair	Error Prob.
a, Ng	1.00
bv, b	0.82
e, a	0.79
ţ, 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

Table 2: Phoneme discovery performance vs. codebook size on TIMIT. The models used are IQs trained on Librispeech+TIMIT.

Codebook size		30	40	50	60	70
$\|\mathbb{Y}\| = 224$	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 3: Phoneme discovery performance vs codebook size on Mboshi. The models used are IQs with CPC+BNF features.

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