Semantic-driven Phoneme Discovery with Information Bottleneck

— ECE 590SIP

Liming Wang

Mar.17th, 2021

Motivation

- **Why discovering phonemes**: A reliable phoneme set can:
  - provide a efficient, low-cost representation to low-resource languages
  - benefit speech processing systems for tasks such as speech recognition, speech synthesis and spoken language understanding
- **Frame**: A sequence of fixed-length consecutive chunk of speech
- **Segment**: A sequence of consecutive speech frames; often variable-length
- **Phoneme**: A set of minimal-length speech segments such that:
  - For any phoneme $\mathcal{X}$ and segments $x_1, x_2 \in \mathcal{X}$, substitute x_1 with x_2 in a sentence has no effect on its meaning
  - For any two distinct phonemes $\mathcal{X}_1, \mathcal{X}_2$ and segments $x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2$, substitute x_1 with x_2 in a sentence alter its meaning
  - Allophone: Any subset of segments $U \in \mathcal{X}$ for a phoneme $\mathcal{X}$

Motivation

Problem Formulation

Method

Experiments

Unsupervised vs weakly supervised phoneme discovery

- Unsupervised phone unit discovery:
  - Input: speech feature frames $\mathbf{x} = x_1, \cdots, x_T$;
  - Output: the phone cluster assignment $\mathbf{z} = z_1, \cdots, z_T$
- Weakly supervised phoneme discovery:
  - Inputs: speech feature frames $\mathbf{x} = (x_1, \cdots, x_T)$ and semantic context $\mathbf{y} = (y_1, \cdots, y_L)$
  - Outputs: same as the unsupervised approach

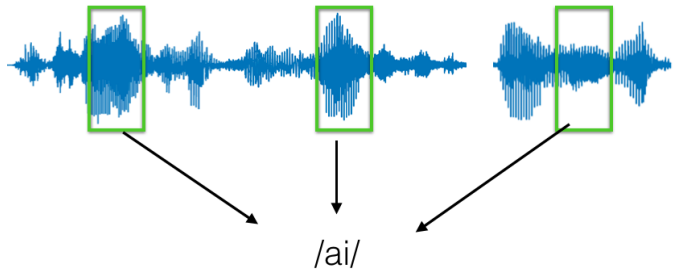


Figure: Unsupervised phone unit discovery

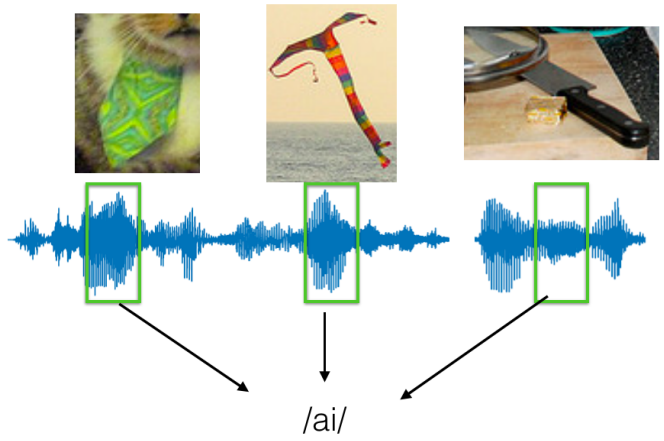


Figure: Weakly supervised phone unit discovery

A distributional view of phonemes

- **Issues with unsupervised models**: Mostly bottom-up, acoustic-driven⇒unable to distinguish between allophones and phonemes⇒large number of redundant clusters
- **Advantage of weakly supervised model**:
  - Will not separate allophones, less redundant clusters
  - Phoneme defined both by its acoustic properties and its semantic context, Closer to the definition of phoneme

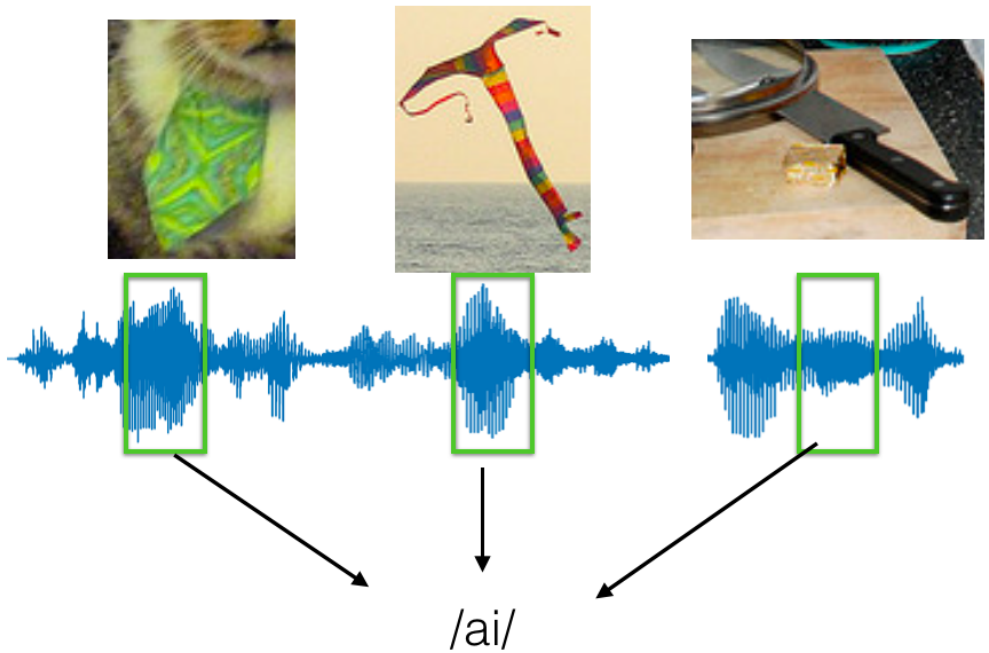


Figure: Distributional representation of the phoneme /ai/ using images

Information bottleneck (IB)

- **Fundamental tradeoff in learning**:
  - Performance: e.g., accuracy, distortion/fidelity metrics, reward
  - Resources: e.g., memory, space, time, energy, material, money
- **Mutual information**:
  $I(X;Y) := \int_{\mathcal{X},\mathcal{Y}} p_{X,Y}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} dxdy$, a nonnegative quantity to measure the amount of information shared by two random variables X and Y with joint distribution $p_{X,Y}$ and marginals p_X, p_Y
- **Information bottleneck principle** (Tishby et al. (1999)): Achieve the highest prediction performance with the lowest amount of information

$$\min \quad I(Z;X) \qquad \text{s.t.} \quad I(Z;Y) \geq R$$

  - X: Source variable containing all the information resources we have;
  - Y: Target variable to predict for a given task;
  - Z: Bottleneck variable containing information from X relevant to predict Y, i.e., form Markov chain $Z - X - Y$;
  - $I(X;Z) \approx$ the amount of information taken by Z from X;
  - $I(Z;Y) \approx$ how accurate can Z predict Y
  - R: Lower bound on the performance, $\leq I(X;Y)$

Phonemes as information bottleneck

- Information bottleneck for sequential data:
  - $X_{1:T} \in \mathcal{X}^T = (X_1, \cdots, X_T)$: Speech feature frames;
  - $Z_{1:T} \in \mathcal{Z}^T = (Z_1, \cdots, Z_T)$: Framewise (latent) phoneme labels;
  - $Y_{1:L} \in \mathcal{Y}^L = (Y_1, \cdots, Y_L)$: Semantic contexts such as nearby frames, allophone segments and images
- Phonemes $Z_{1:T}$ contains all semantic information from X, $\Rightarrow I(Z_{1:T}; Y_{1:L}) = I(X_{1:T}; Y_{1:L})$;
- Any representation $Z'_{1:T}$ with less information will inevitably confuse at least one pair of distinct semantic contexts $(Y, Y') \Rightarrow I(Z_{1:T}; X_{1:T})$ is minimal
- General IB objective for phoneme discovery:

$$\min_{Z_{1:T} - X_{1:T} - Y_{1:L}} \quad I(Z_{1:T}; X_{1:T})$$
$$\text{s.t.} \quad I(Z_{1:T}; Y_{1:L}) = I(X_{1:T}; Y_{1:L})$$

Learning algorithm for IB-based phoneme discovery

Challenges

- $|\mathcal{Z}|^T$ possible combinations of phoneme sequence in a sentence $\Rightarrow I(Z_{1:T}; Y_{1:L})$ intractable in general
- $O(T^2)$ possible phoneme boundaries $\Rightarrow I(X_{1:T}; Z_{1:T})$ hard to compute

Naive IB-based phoneme model

- Decoupling of Z_t's: Z_t's are independent given $X_{1:T}$;
- Decoupling of X_t's: Z_t is independent of $\{X_s\}_{s \neq t}$ given X_t;
- Decoupling of Y_i's: Y_i's are independent given $Z_{1:T}$;
- Replica assumption of Y_i's: One replica $Y_{1:L}^{(t)}$ of $Y_{1:L}$ for each t, with $Y_{1:L}^{(t)}$ independent of $\{(Z_s, Y_{1:L}^{(s)})\}_{s \neq t}$ given Z_t

Naive IB model

Variational IB (VIB) training objective

Let $\beta(R)$ be the *Lagrangian multiplier* of the VIB at rate R:

$$\min_{I(Z_{1:T};Y_{1:L})\geq R} I(Z_{1:T};X_{1:T}) =: \min_{\sum_i I(Z_t;Y_i)\geq R} \sum_t H(Z_t) - H(Z_t|X_t)$$

$$=: \min \mathbb{E}_{x,y}[\sum_t \sum_{z_t} q_t(z_t|x_{1:T}) \log \frac{q_t(z_t|x_{1:T})}{r(z_t)} -$$

$$\beta(R) \sum_{t,i} \sum_{z_t} q_t(z_t|x_{1:T}) \log \frac{q_{ti}(y_i|z_t)}{p(y_i)})] + C =: L_{IB},$$

Model parameters

- $q_t(z|x)$: Probability of assigning t-th frame to phoneme z;
- $q_{ti}(y|z)$: Probability of predicting the i-th semantic context to be y given the t-th phoneme is z;
- $r(z)$: Prior probability of phoneme z

Naive IB model

Interpretation of β

- Higher performance requirement $R \Rightarrow$ higher $\beta(R) \Rightarrow$ more information taken from speech $X_{1:T}$
- Optimal β for phoneme discovery when $R = I(X_{1:T}; Y_{1:L})$

Decoding

- Maximum a posteriori (MAP) decoding:

$$\begin{aligned} z_t^* &= \max_z p(z|x_t, y_{1:L}) \\ &= \max_z q_t(z|x_t), t = 1, \cdots, T, \end{aligned}$$

by conditional independence between Z_t, $Y_{1:L}$ given X_t.

Naive IB model: Architecture for discrete target variable

- Bottleneck layer: Samples (one-hot representation of z_t's) from $q_t(z_t|x_{1:T})$ with Gumbel softmax trick: Let $b_{tz} := \log q_t(z_t|\mathrm{x})$,

$$z_t = \arg\max_z (b_{tz} + g_{tz})$$

$$\Rightarrow \mathrm{onehot}_z(z_t) \approx \frac{\exp(\frac{b_{tz}+g_{tz}}{\tau})}{\sum_{z'} \exp(\frac{b_{tz'}+g_{tz'}}{\tau})},$$

where $g_{tz} = -\log(-\log u_{tz})$'s are *Gumbel random variables* for standard uniform random variable u_{tz}'s

- Target predictor: A lookup table E with weight e_{zy} to be the logit of $q(y_i = y|z_t = z)$

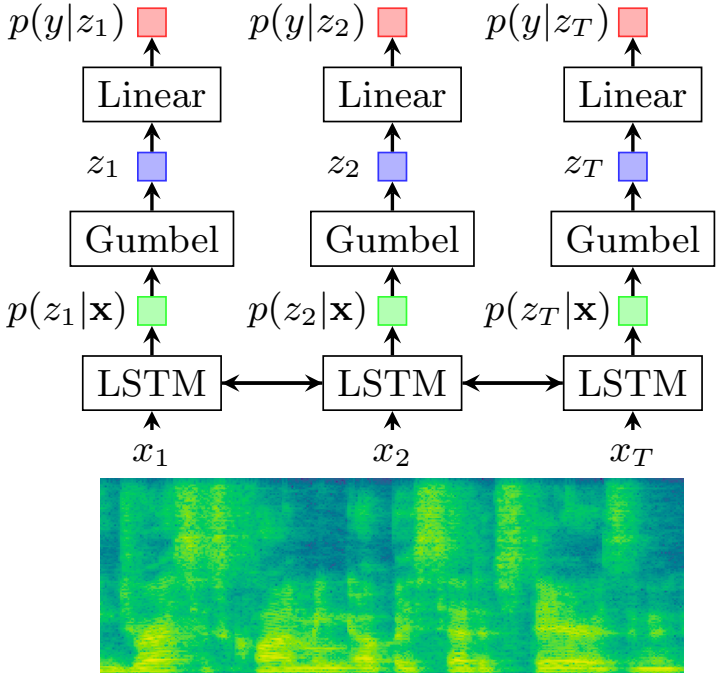


Figure: Naive deep variational IB Model

Naive IB model: Architecture for continuous target variable

- Predictor: Replace logits with a continuous codebook E
- $I(Z;Y) \approx \mathbb{E}_{x,y} \sum_{t=1}^{T} \sum_{z_t \sim q(z_t|x_{1:T})} \log \frac{\exp(s(e_{z_t}, y))}{\sum_{y'} \exp(s(e_{z_t}, y'))} + C =$

  $\mathbb{E}_{x,y} \log \frac{\exp(\sum_{t=1}^{T} \sum_{z_t \sim q(z_t|x_{1:T})} s(e_{z_t}, y))}{\sum_{y'} \exp(\sum_{t=1}^{T} \sum_{z_t \sim q(z_t|x_{1:T})} s(e_{z_t}, y'))}$, for negative samples y''s

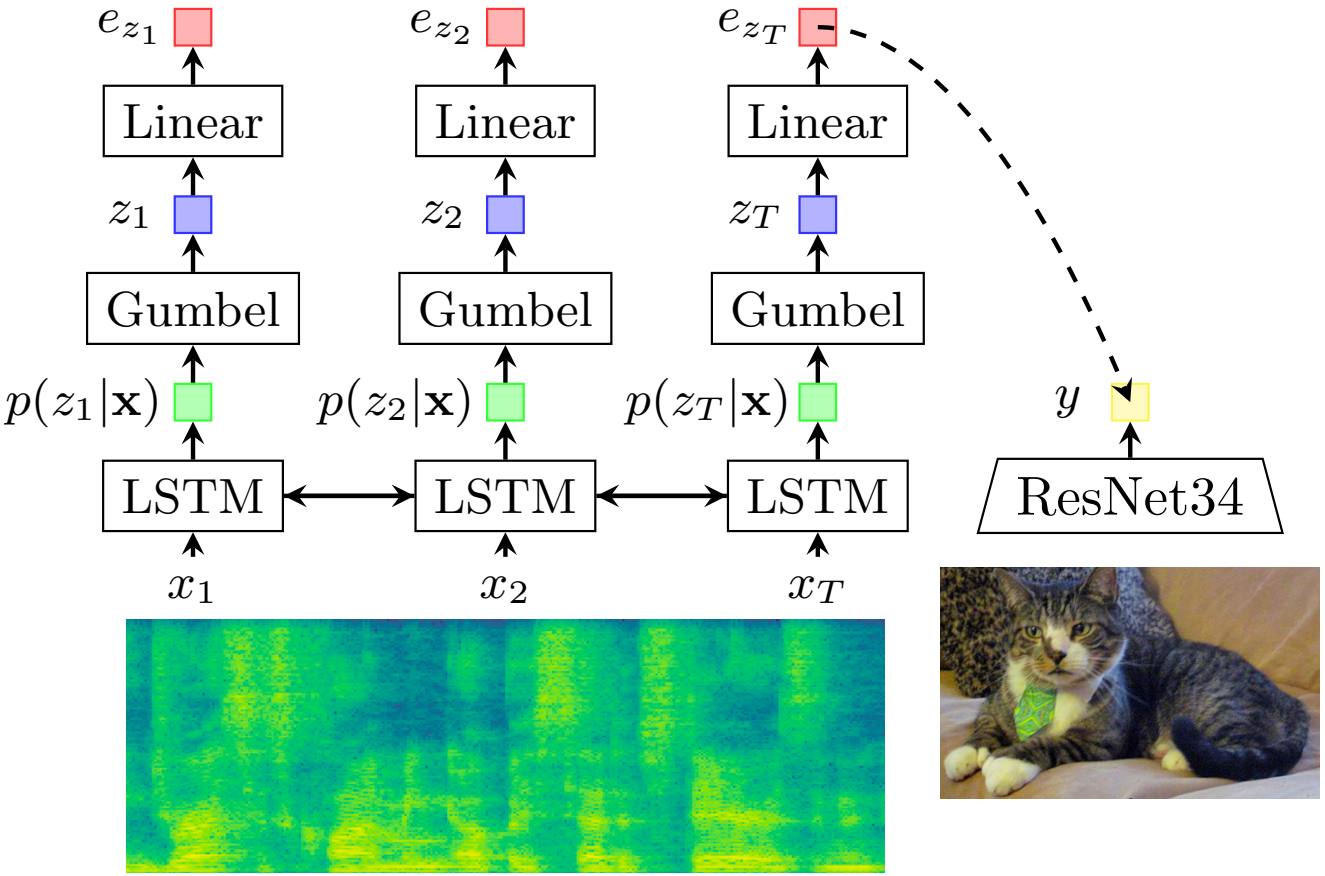


Figure: Naive IB Model for continuous Y

Naive IB model: Alternative model with vector quantization (van den Oord et al. (2017))

- Replace the Gumbel softmax layer with:
  1. An encoder $e : \mathcal{X} \mapsto \mathbb{R}^D$
  2. A codebook: $E \in \mathbb{R}^{D \times K} = [e_1, \cdots, e_K]$
  3. A distance metric: $d(\cdot, \cdot)$, typically L2 squared distance
  4. A quantizer: $z : \mathbb{R}^D \mapsto \{1, \cdots, K\}$, a *deterministic* mapping from x to a phoneme unit index whose corresponding vector is closest to x in the embedding space, i.e., $z(\mathrm{x}) = \arg\min_z d(e(\mathrm{x}), e_z)$
- Phoneme posterior:

$$q_t(z|\mathrm{x}) := \frac{\exp(-d(e_t(\mathrm{x}), e_z))}{\sum_{z'} \exp(-d(e_t(\mathrm{x}), e_z))}$$

- Vector quantization loss:

$$L_{VQ}(\theta) = \mathbb{E}_{\mathrm{x}}[\sum_{t=1}^{T} d(e_t(\mathrm{x}), e_{z_t(\mathrm{x})})]$$

$$L = L_{IB} + L_{VQ}(\theta)$$

Beyond Naive IB model

Limitations of Naive IB model

1. Oversimplified relations between phonemes and its semantic context
2. Not modeling inter-dependencies between phonemes

Incoporating phonotactics: Bag-of-phone IB model

- Modified IB objective: Prediction for future speech features within a window of k_{max}

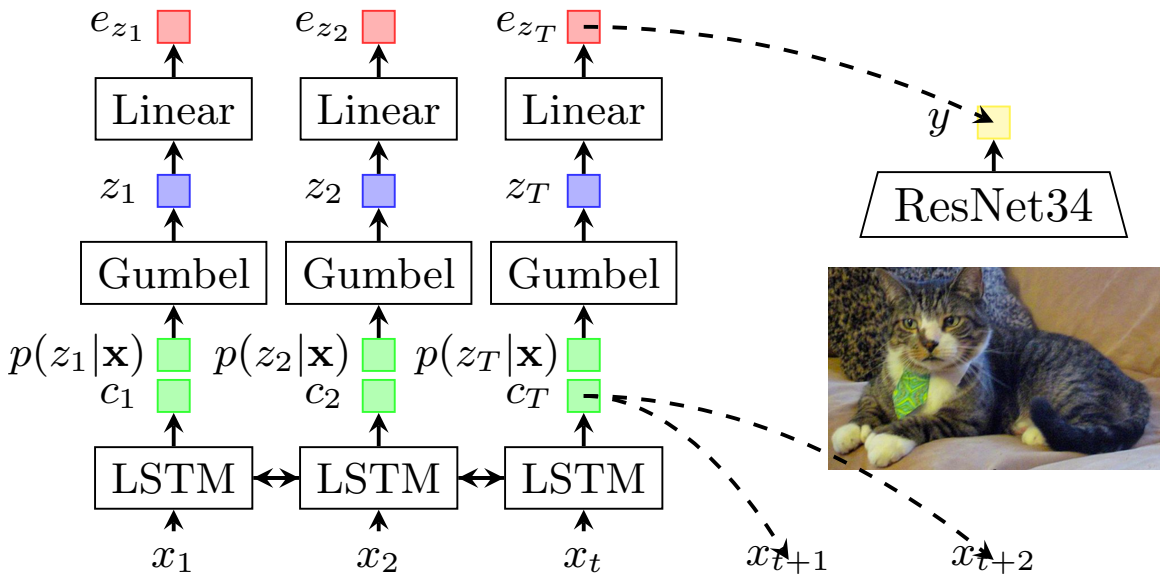
$$\min \quad \sum_{t=1}^{T} I(Z_t; X_t)$$

$$\text{s.t.} \quad \sum_{t=1}^{T} I(Z_t; Y_i^{(t)}) \geq R_Y, \; \sum_{t=1}^{T-k} I(Z_t; X_{t+k}) \geq R_k,$$

$$k = 1, \cdots, k_{\max}$$

Bag-of-phone IB model: Architecture

- Phonotactic predictor: use contrastive predictive coding (CPC) (van den Oord et al.) and its variants
- $I(Z; X_k) \approx \mathbb{E}_{x,y} \sum_{t=1}^{T-k_{\max}} \sum_{k=1}^{k_{\max}} \log \frac{\exp(s(c_t, x_{t+k}))}{\sum_{n \in \mathcal{N}_t} \exp(s(c_t, n))} + C$, with some sets of negative samples $\mathcal{N}_t$'s



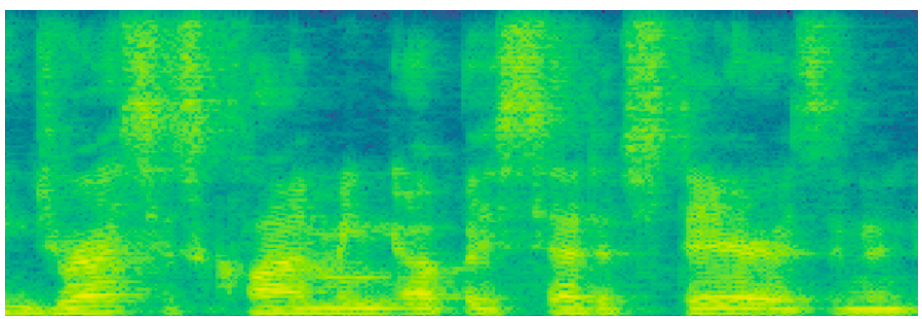


Figure: IB+CPC Model

Experimental setting: Weakly supervised phoneme discovery

Dataset

MSCOCO2k: 12500 isolated words with paired images for 65 visual concepts from validation set of SpeechCOCO dataset, 200 instances per concept, 10000/2500 train test split

Task

Phoneme discovery with:

1. Word-level supervision: provide ground truth word labels, but no character-level info
2. Visual supervision: provide image features from hidden layer of ResNet34 for the paired images

Evaluation metrics

- Word error rate (WER)
- Token F1: $\frac{2TR \times TP}{TR+TP}$, $TR = \sum_k \frac{\text{max \# of allophones in cluster k}}{\text{\# of phones in cluster k}}$, $TP = \sum_k \frac{\text{max \# of phones of type k from the same cluster}}{\text{\# of phones of type k}}$;
- ABX: based on discrimination task between triphones "b**a**g" (A) vs another "b**a**g" (B) and vs "b**e**g" (X); 50% chance

Implementation details

- Speech feature: Mel filterbank with 80 mels, 25ms window size and 15ms overlap
- Source encoder: single-layer BiLSTM with hidden size 256
- Gumbel softmax: 49 categories, temperature varies from 1 to 0.1 with an anneal rate of 3×10^{-6} every 100 steps
- VQ: codebook with 65 512-dim embeddings uniformly initialized between $[-1/512, 1/512]$; exponential mean average with a decay rate of 0.995 for codebook update,
- CPC: positive step$=6$, number of negative samples per step$=17$; dot product predictor score s
- Adam optimizer, starting learning rate 10^{-3}, batch size of 32, 150 epochs

Overall results

	WER	Token F1	ABX
Word-level supervision			
CPC	-	-	24
CPC+VQ (van Niekerk et al. (2020))	-	-	20
IB	1.5	42	**5.2**
IB+CPC	1.5	**48**	12.5
Image-level supervision			
BLSTM	24	-	21
IB+VQ	31	-	10.5
IB+CPC+VQ	**29**	-	**10**

Table: Weakly supervised phoneme discovery performance on MSCOCO2k

Phoneme discovery performance vs level of compression for models with word supervision

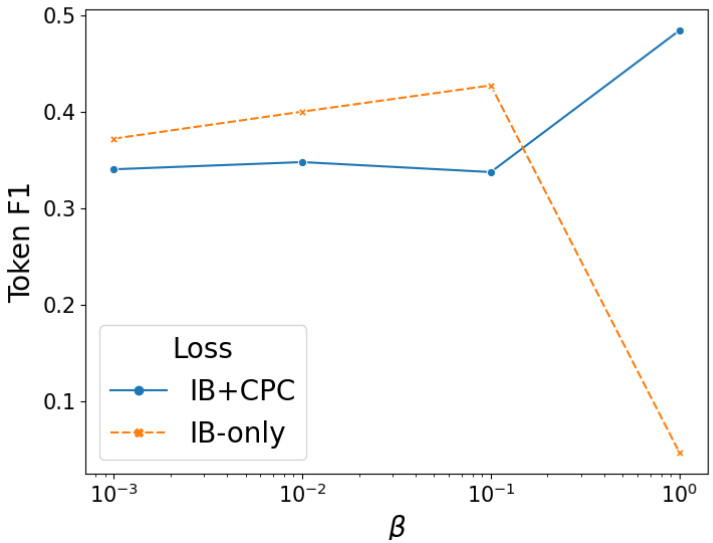


Figure: Token F1 vs β

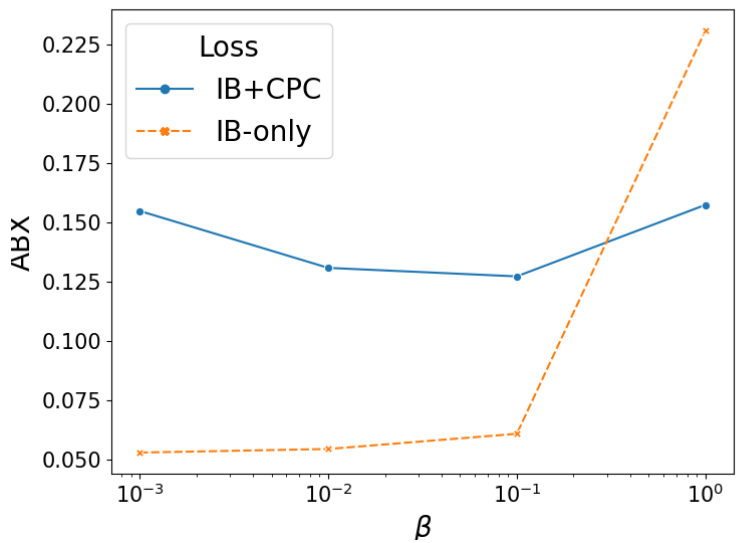


Figure: ABX vs β

Phoneme discovery performance vs level of compression for models with visual supervision

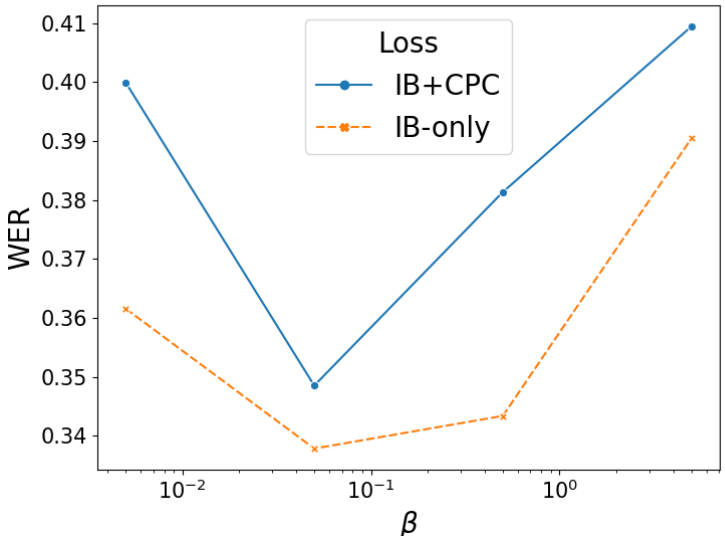


Figure: WER vs β (Trained only 50 epochs)

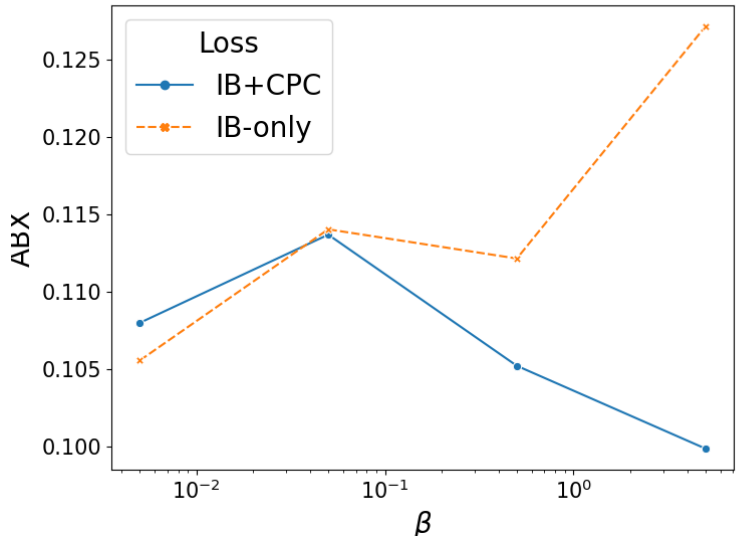


Figure: ABX vs β (Trained only 50 epochs)

t-SNE visualization for phoneme discovery with word supervision

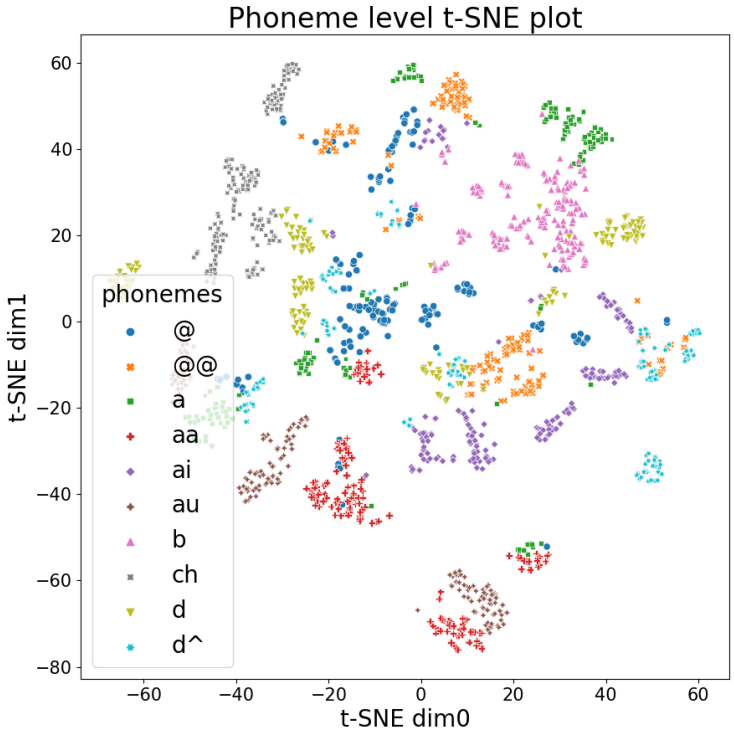


Figure: IB only

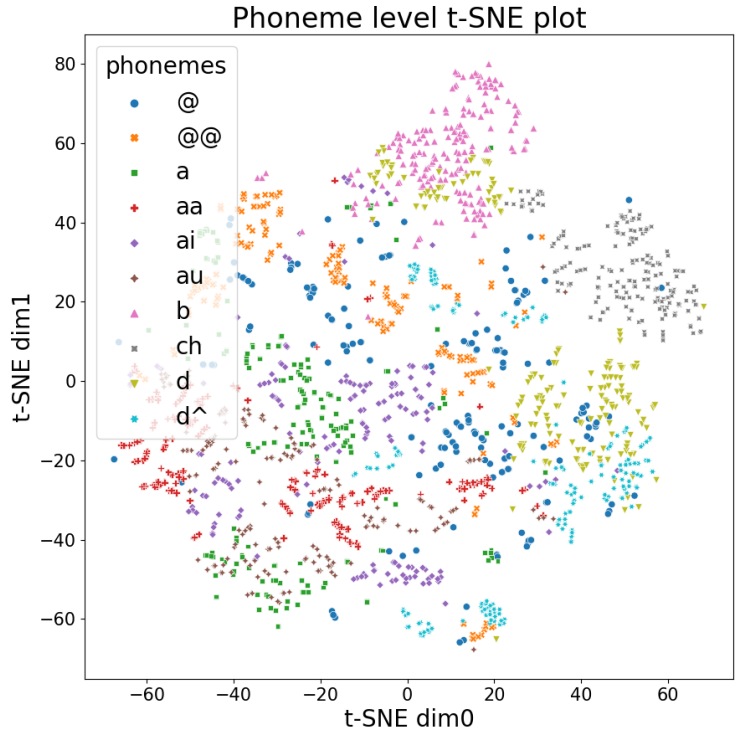


Figure: IB+CPC

t-SNE visualization for phoneme discovery with visual supervision

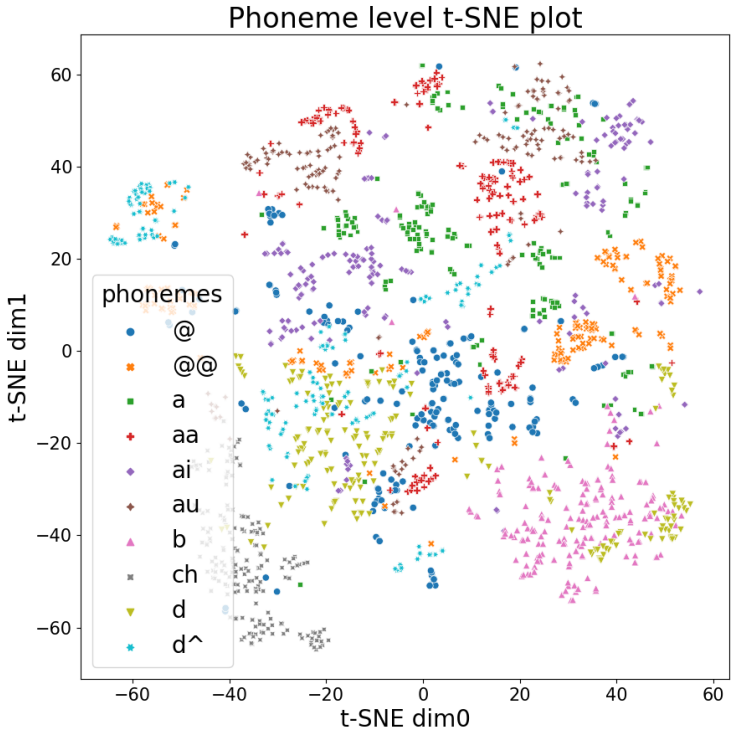


Figure: IB+VQ

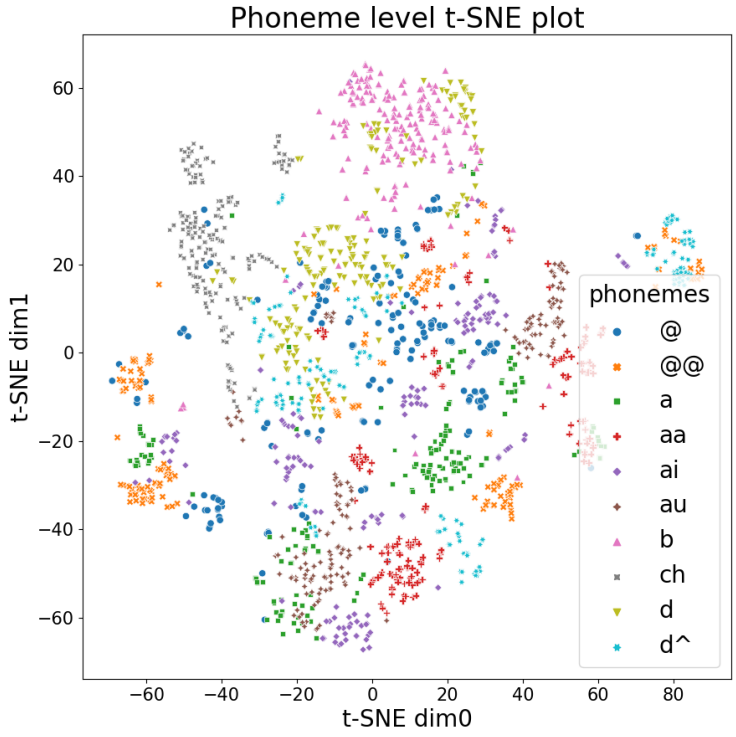


Figure: IB+CPC+VQ

Conclusion and future work

- Information bottleneck as a general framework for weakly supervised phoneme discovery
- Need models beyond naive IB to further remove intra-phone variabilities
- Need better methods to train continuous codebook
- Need to test on larger dataset with richer vocabularies

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

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

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

N. Tishby, F.C. Pereira, and W. Bialek. 1999. The information bottleneck method. In *The 37th annual Allerton Conf. on Communication, Control, and Computing*, page pp. 368–377.