ゼロ資源状況におけるサブワード単位の獲得にむけて グラフニューラルネットワークを用いた手法

Towards Sub-word Unit Discovery in Zero Resource Scenario: An Approach Based on Graph Neural Networks

高橋舜 [1]　サクリアニサクティ [1,2]　中村哲 [1,2]

Shun Takahashi[1], Sakti Sakriani[1,2], and Satoshi Nakamura[1,2]

[1] 奈良先端科学技術大学院大学
[1]Nara Institute of Science and Technology
[2] 理化学研究所革新知能統合研究センター
[2]RIKEN AIP Center

Abstract: Zero resource speech technology aims for discovering discrete units in the limited amount of unannotated, raw speech data. The previous studies have mainly focused on learning the discrete units from acoustic features, segmented by fixed small time-frame. While achieving high unit quality, they suffer from high bitrate due to the time-frame encoding. In this work, in order to lower the bitrate, we propose a novel approach based on discrete autoencoder and graph convolutional networks. We exploit the speech features discretized by vector-quantization encoding. Since the maximum number of the discretized features is predetermined, we consider a directed graph where each node represents a discretized acoustic feature, and each edge transition from one feature to another. Using graph convolution, we extract and encode the topological feature of the graph into each node, and then we symmetrize the graph to apply spectral clustering on the node features. In terms of ABX error rate and bit rate estimation, we demonstrate that our model successfully decreases the bitrate, while retaining the unit quality.

1. はじめに

ゼロ資源音声技術は限られた量の音声データから教師ラベルを利用せずに離散的な音響単位や言語学的記号を獲得することを目的とする．これまで Zero Resource Speech Challenge[1], [2]を通じて音響単位の獲得に関して共通のデータセットと評価手法によって統一的な評価がなされてきた．特に著しいパフォーマンスを記録しているのはベクトル量子化[3]を利用した離散オートエンコーダーによる手法[4], [5]である．

一方でこれらの手法は固定長の時間単位(タイムフレーム)における音響特徴量の離散的表現の学習に重きが置かれているため，獲得される表現はビットレートが非常に高いという問題がある．そこで本研究ではより粗く，タイムフレームから独立した表現を獲得することを目的として，離散オートエンコーダーをベースとしたグラフニューラルネットワーク(GNNs)による新たな手法を提案する．

本研究で利用する Vector-Quantized Contrastive Predictive Coding (VQCPC)は音声データをコードブックと呼ばれる予め決められた数の離散単位からなる系列データに変換する．われわれは VQ-CPC のエンコーダーによって離散化された各音響単位をグラフにおけるノードとし，それらの遷移を辺として考える．そして GCN を利用して各ノードについて近傍の特徴量を畳み込み，ノードの特徴量に基づくスペクトラル・クラスタリングを行う．

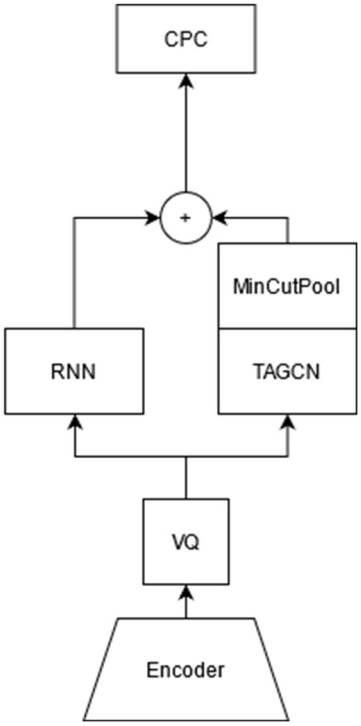
本稿ではこの手法によって獲得される表現を ABX 誤り率及びビットレートの観点から評価し，その結果 ABX 誤り率を抑えながら，ビットレートを半減させることに成功したことを報告する．

連絡先:
高橋舜: takahashi.shun.tp0@is.naist.jp
サクティサクリアニ: ssakti@is.naist.jp,
中村哲: s-nakamura@is.naist.jp

2. 提案手法

本研究では VQVAE のエンコーダーの出力であるコード系列からグラフを構築し，スペクトラル・クラスタリングを行う離散オートエンコーダーのモデルを提案する．以下の図 1 は本研究で提案するモデルの概略である．

図 1 本研究で提案するモデル



2.1. 音声データの離散化

ベクトル量子化 (VQ) を通じて音声データの離散化を行う．VQ はコードブックと呼ばれる埋め込みベクトル$e_i \in R^D, i \in \{1,2,\dots,N\}$に最近傍法に基づいて入力の音声データ$z_i \in R^{T\times D}$をマッピングする．

$$N = \arg\min_j \|z_i - e_j\|^2$$

逆伝搬の際にエンコーダーの勾配はストレートスルー推定 (STE) によって近似される[3]．この層から埋め込みベクトルの系列データ及びコード ID 系列情報$c \in \{1,2,\dots,N\}^T$を GCN クラスタリングへ，また埋め込みベクトルの系列データを RNN へ伝搬させる．

2.2. コード系列からのグラフの構築

VQVAE の予め決められたN個の埋め込みベクトル（コードブック）に着目して，本研究ではまず各頂点がN個の要素に対応する有向グラフGを考える．VQVAE の出力の各コード系列をこのグラフにおける経路としてみなすと，このグラフにおける辺Eはコード間の遷移として考えることができる．また本研究では以下のように各辺の重みとしてミニバッチtごとに遷移頻度の指数移動平均(EMA)を逐次算出して，各辺に付与する．

$$EMA_t(w^{i,j}) = EMA_{t-1}(w^{i,j}) + \alpha\{w_t^{i,j} - EMA_{t-1}(w^{i,j})\}$$

この際，定数αはハイパーパラメーターであり，$\alpha \in [0,1]$とする．さらにスペクトラル・クラスタリングを適用するためにこうして得られた重み付き隣接行列$W_{dir} \in R^{V\times V}$に対して$W_{dir} + W_{dir}^T$として対称化する．

2.3. グラフ畳み込み

各コードの時間的隣接性をもとに構築したグラフからその特徴を抽出する．グラフ畳み込みには Topology-Adaptive Graph Convolutional Networks (TAGCN) [6]を用いる．TAGCN は，頂点領域においてグラフ信号処理に基づく畳み込みを行うことで，各頂点のグラフのトポロジーに基づく特徴を抽出する．

TAGCN のl層目の隠れ層における頂点iのグラフ畳み込みは以下の式によって表される．

$$\boldsymbol{y}_f^{(l)}(i) = \sum_{k=1}^{K_l}\sum_{c=1}^{C_l}\sum_{j\in\{j|j \text{ is } k \text{ paths to } i\}} g_{c,f,k}^{(l)}\omega(p_{j,i}^k)\boldsymbol{x}_c^{(l)}(j) + b_f\mathbf{1}_{N_l}$$

ここでKはフィルターサイズ，Cは各頂点の入力特徴マップの数，$g_{c,f,g}^{(l)}$はグラフフィルターの多項式係数，$\omega(p_{j,i}^k)$は頂点jから頂点iまで経路をその各辺の重みの積とし，そのk通りの和，$\boldsymbol{x} \in R^{N_l}$は頂点の特徴量，$b$は学習バイアスとする．

また CNN と同様に各グラフ畳み込みのあとに非線形変換がなされる．

$$\boldsymbol{x}_f^{(l+1)} = \sigma(\boldsymbol{y}_f^{(l)})$$

2.4. グラフ・クラスタリング

時間的隣接性が高いコードをクラスタリングする．手法としては GNN ベースのスペクトラル・クラスタリングである MinCutPool[7] を利用する．MinCutPool は GNN の出力をもとに各頂点を MLP および Softmax を通じてK個のクラスターへ振り分けるように以下の損失関数を最適化する．

$$\mathcal{L}_{MC} = \mathcal{L}_C + \mathcal{L}_o = -\frac{Tr(\boldsymbol{S}^T\tilde{\boldsymbol{A}}\boldsymbol{S})}{Tr(\boldsymbol{S}^T\tilde{\boldsymbol{D}}\boldsymbol{S})} + \left\| \frac{\boldsymbol{S}^T\boldsymbol{S}}{\|\boldsymbol{S}^T\boldsymbol{S}\|_F} - \frac{I_K}{\sqrt{K}} \right\|_F$$

ここで$\boldsymbol{S} \in [0,1]^{N\times K}$はクラスター振り分け行列，$\tilde{\boldsymbol{A}}=\boldsymbol{D}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\boldsymbol{D}^{-\frac{1}{2}} \in R^{N\times N}$は正規化した隣接行列，$\tilde{\boldsymbol{D}} \in R^{N\times N}$はその次数行列，$\|\cdot\|_F$をフロベニウスノルムとする．

クラスタリングされた頂点(コードブック)の特徴量は次のようにして得られる．

$$\bar{\boldsymbol{X}} = \boldsymbol{S}^T\boldsymbol{X}\boldsymbol{S} \in R^{N\times F}$$

$\bar{\boldsymbol{X}}$を入力のコードID系列$\in \{1,2,\dots,N\}^T$をもとに並べ替えて時間的隣接性に基づいて再解釈されたコード系列として次の層へ出力する．

2.5. 対照的予測符号化による学習

本研究では VQCPC[8]と同様に VQVAE のデコーダーによる再構築誤差の代わりに対照的予測符号化(CPC) によりエンコーダー及び GCN クラスタリングの学習を行う．

$$\mathcal{L}_{CPC} = -\frac{1}{M}\sum_{m=1}^{M}\left[\frac{\exp\left(\hat{\boldsymbol{z}}_{\boldsymbol{t+m}}^{\boldsymbol{T}}\boldsymbol{W}_{\boldsymbol{m}}\boldsymbol{c}_{\boldsymbol{t}}\right)}{\sum_{\bar{z}\in\mathcal{N}_{t,m}}\exp\left(\bar{\boldsymbol{z}}^{\boldsymbol{T}}\boldsymbol{W}_{\boldsymbol{m}}\boldsymbol{c}_{\boldsymbol{t}}\right)}\right]$$

ここでMは予測タイムステップ，$\hat{z}$は予測したいコード (埋め込みベクトル) ，$\bar{z} \in \mathcal{N}_{t,m}$は区別すべき負例のコード，$\boldsymbol{c}_{\boldsymbol{t}} \in R^{(D_{RNN}+D_{GCN})\times T}$は RNN の出力と GCN クラスタリングの出力を連結させた文脈系列，$\boldsymbol{W}_{\boldsymbol{m}}$はその文脈系列をもとに先のコードを予測する学習行列とする．

3. 実験評価

Zero Resource Speech Challenge 2019 において提供された英語のデータセットと評価手法を用いて評価実験を行った．

3.1. データセットと評価手法

訓練データには 100 名の発話者による 15 時間 40 分の発話データを，テストデータには 24 名の発話者による 28 分の発話データをそれぞれ含まれている．

ABX 音素弁別テスト[9]はモデルに依存しない音韻単位の評価手法で，X が A と B のどちらに近いか，を獲得された音韻単位をもとに動的時間縮約法をもとに編集距離やコサイン類似度などで測る．本稿ではコサイン類似度と編集距離をもとに評価を行った．

ビットレートは以下の式に基づいて推定される．

$$B(U) = \frac{P}{D}\sum_{i=1}^{P} p(\boldsymbol{s}_i)log_2 p(\boldsymbol{s}_i)$$

ここで P は系列長，D は時間長，s は音響単位を表すシンボルベクトルとする．

3.2. 実験設定

バッチサイズを 64 とし，8 人の話者から 8 つの発話をそれぞれ 1.28 秒サンプリングしたものである．学習率は $1\cdot10^{-5}$ からはじめの 150 エポックを通じて $2\cdot10^{-4}$ に直線的に引き上げるウォームアップをおこなった．最適化には Adam[10]を利用した．

VQCPC は著者らが公開している実装を利用した[i]．TAGCN および MinCutPool は Pytorch Geometric[11]の実装を参考にした．TAGCN は 2 層重ね，フィルターサイズをそれぞれ 2 と 3，入力データの次元は 64 次元，隠れ層では 32 次元とした．MinCutPool のクラスター数は 64 とした．

3.3. 実験結果と議論

以下の表 1 は Zero Resource Speech Challenge における Topline である正解データで学習した ASR, Zero Resource Speech Challenge 2019 でトップの VQVAE 及び 2020 でトップの VQCPC，そして本研究で提案する手法の比較である．

我々の提案手法は非常に低い ABX 誤り率を記録している VQCPC をベースとしているため，どれほど ABX 誤り率を下げずにビットレートを下げることができるかという点が重要であるが，コサイン類似度における誤り率 (Cos.) を 20 以下に抑えつつ，かつより厳しいメトリックである編集距離において VQCPC より良い結果を出している．また目的通りにビットレートに関して半分ほどではあるが，減少させることに成功した．一方でトップラインと比較するとビットレートの差は依然としてかなり開いている．しかし ABX 誤り率とビットレートのトレードオフにおいて我々の提案手法は現状，最良ともといえる．

表 1 各モデルの性能比較

Model	ABX error rate (Cos.)	ABX error rate (Lev.)	Bit rate
supervised ASR[1]	29.41	29.85	37.73
VQVAE[12]	20.71	37.95	167.02
VQCPC[8]	13.25	27.70	417.89
提案手法	17.93	22.02	271.76

4. まとめ

本稿では離散オートエンコーダーとグラフニューラルネットワークを利用することで異なるアプローチによるクラスタリングを階層的に行うモデルを提案し，より抽象的な表現の獲得を目指した．今後，離散化された音声データをグラフ構造としてより効果的に扱えるモデルへ改良を続けていきたいと考えている．

謝辞

本研究は科研費 [JP17H06101] の 助成を受けております。

参考文献

[1] E. Dunbar *et al.*, “The Zero Resource Speech Challenge 2019: TTS without T,” *ArXiv190411469 Cs Eess*, Jul. 2019, Accessed: Sep. 27, 2020. [Online]. Available: http://arxiv.org/abs/1904.11469.

[[2] E. Dunbar, et al, “The Zero Resource Speech Challenge 2020: Discovering discrete subword and word units,” Proc. Interspeech 2020, pp. 4831–4835, 2020.

[3] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural Discrete Representation Learning,” *ArXiv171100937 Cs*, May 2018, Accessed: May 19, 2020. [Online]. Available: http://arxiv.org/abs/1711.00937.

[4] A. Tjandra, B. Sisman, M. Zhang, S. Sakti, H. Li, and S. Nakamura, “VQVAE Unsupervised Unit Discovery and Multi-Scale Code2Spec Inverter for Zerospeech Challenge 2019,” in *Interspeech 2019*, Sep. 2019, pp. 1118–1122, doi: 10.21437/Interspeech.2019-3232.

[5] B. van Niekerk, L. Nortje, and H. Kamper, “Vector-Quantized Neural Networks for Acoustic Unit Discovery in the ZeroSpeech 2020 Challenge,” in *Interspeech 2020*, Oct. 2020, pp. 4836–4840, doi: 10.21437/Interspeech.2020-1693.

[6] J. Du, S. Zhang, G. Wu, J. M. F. Moura, and S. Kar, “Topology Adaptive Graph Convolutional Networks,” *ArXiv171010370 Cs Stat*, Feb. 2018, Accessed: Dec. 08, 2020. [Online]. Available: http://arxiv.org/abs/1710.10370.

[7] F. M. Bianchi, D. Grattarola, and C. Alippi, “Spectral Clustering with Graph Neural Networks for Graph Pooling,” *ArXiv190700481 Cs Stat*, Jul. 2020, Accessed: Dec. 07, 2020. [Online]. Available: http://arxiv.org/abs/1907.00481.

[8] B. van Niekerk, L. Nortje, and H. Kamper, “Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge,” *ArXiv200509409 Cs Eess*, Aug. 2020, Accessed: Oct. 22, 2020. [Online]. Available: http://arxiv.org/abs/2005.09409.

[9] T. Schatz, et al., “Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline,” In Interspeech, 2013.

[10] D. P. Kingma and J. Lei, “Adam: A Method for Stochastic Optimization,” p. 15, 2015.

[11] M. Fey and J. E. Lenssen, “Fast Graph Representation Learning with Pytorch Geometric,” in ICLR, 2019

[12] A. Tjandra, S. Sakti, and S. Nakamura, “Transformer VQ-VAE for Unsupervised Unit Discovery and Speech Synthesis: ZeroSpeech 2020 Challenge,” in Interspeech, 2020

[i] https://github.com/bshall/VectorQuantizedCPC