音声信号から文字記号を創り出す —深層ベイズに基づく教師なし表現学習によるアプローチ—

髙橋　舜[1]　金崎朝子[2]　須田仁志[3]　サクリアニ・サクティ[1]
[1] 奈良先端科学技術大学院大学 [2] 東京科学大学 [3] 産業技術総合研究所
takahashi.shun.tq9@naist.ac.jp kanezaki@c.titech.ac.jp
suda.h@aist.go.jp ssakti@is.naist.jp

概要

如何にして対象の音声言語の音声データから，その言語において言語学的に妥当な文字記号の体系を機械に創り出させるか．この問いに答えるべく，本研究では深層ベイズに基づく機械学習手法を提案する．提案手法では，世界の言語の音素数の報告データをもとに，文字記号の種類数の弱情報事前分布を導入する．これにより，従来研究のように，文字記号の種類数に関する事前の仮定に無限や定数を無理に持ち出すことなく，対象言語が持ち得る文字記号の種類数を推定しながら，その言語の文字記号の体系を創出することが可能になる．実験により，提案手法の文字体系と人手で創られた文字体系に，従来手法と比較してより強い対応が示された．

1　はじめに

世界の 7,168 言語の 93 %程度が文字体系を確立しておらず，話しことば（音声）を主とする口頭言語に分類される [1, 2, 3]．このことから，自然言語の総体として最も「自然」な状態は音声の形式といっても過言でない．しかし，自然言語処理の分野はこれまで，専らテキストという人手で整形された二次的なデータ形式で自然言語を扱ってきた．結果として，その殆どの技術がテキスト資源の豊かなごく一部の言語にしか適用できない現状がある．

より包摂的な言語技術の実現を目標に，音声から直接，音素に相当するような一種の文字記号（以下，文字とする）の体系を機械に創出させる「音響単位発見」という基礎タスクがある．それを通じて得られた文字体系をもとに，ブートストラップ方式に自然言語処理を展開することが見込まれる [4]．また，その過程が人間の第一言語獲得と相通ずるため，発達認知科学への貢献も期待されている [5]．音響単位発見でモデル化の対象となる音素体系は，図 1 のように，どの言語でも有限小数個で構成される．しかし，第 2 節で詳述するように，従来手法には，文字の種類数を潜在的に無限，あるいは事前指定の定数とするなど，極端な仮定があった．そこで本研究では，音素の実態により即した文字体系の獲得手法を提案する．より具体的には, 本研究では深層ベイズというベイズ推論と深層学習を融合させたアプローチにより，「音素らしい種類数」の事前知識を考慮しつつ，文字体系の獲得を可能にする手法を提案する．そして従来手法との比較実験により，提案手法が創り出す文字体系が，人手で創られた文字体系とより強く対応することを示す．

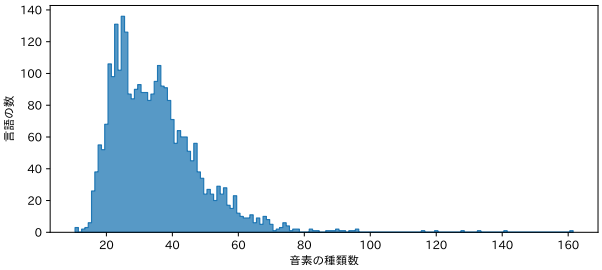

図 1　Phoible [6] 内の 3,020 言語の音素の種類数分布

2　関連研究

ディリクレ過程 (DP)　音響単位発見の先駆的な研究のひとつに Lee らの研究 [7] がある．Lee らはディリクレ過程（DP）という無限次元のカテゴリー分布を生成する確率過程をもとに，混合分布数がデータ依存のガウス混合モデルを提案している．その各混合分布が音響単位（文字）に対応する．後続研究には例えば，変分推論を導入した手法 [8] や，より複雑な階層構造を取り入れた手法 [9] がある．

ニューラル量子化　近年，より主流となった手法がニューラル量子化オートエンコーダー（VQVAE）[10] という，DNN の中間表現を量子化する手法で

ある．Tjandra ら [11] や Niekerk ら [12] が音響単位発見に応用した．同手法は，文字の種類数に相当するコードブックの大きさを事前指定する必要があるが，DP ベースの手法に対して音素情報の特徴抽出性能に関して優位性が報告されている [13]．

無限のカテゴリー数を想定する DP は，例えば任意のテキスト上の単語や漢字の頻度分布のモデルには適切と考えられるが，音素体系のモデルにはそう考え難い．また，各言語・方言で音素の種類数が異なることから，コードブックを事前指定するニューラル量子化も妥当性を欠く．これらに対して，本研究では，言語学者がつくる音素目録のように，有限小数で構成される文字体系の獲得を可能にする．

3 提案手法

本節では深層ベイズに基づく提案手法について説明する．提案手法は，深層ベイズの代表手法 VAE [14] や従来研究 [12] の VQVAE [10] と同様にオートエンコーダーの一種とみなせる．つまり，入力データを所望の形質の表現に変換するエンコーダーと，その表現から元の入力データを再構成するデコーダーで構成される．第 3.1 節の認識モデルが前者に，第 3.2 節の生成モデルが後者に相当する．

ここで提案手法の概略を述べる．提案手法の認識モデルは対象言語の音声データ，系列長 $T \in \mathbb{N}$ の対数メルスペクトログラム $\boldsymbol{x}_{1:T} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T]$ を，漸進的に時刻 $t \in \{1, 2, ..., T\}$ ごとに入力として受け取る．そして，認識モデルと生成モデルの逐次的な協働の過程を経て，N 通りの，各入力 $\boldsymbol{x}_t$ に対応し得る文字（の ID），$z_t \in \{0, 1, 2, ..., K\}_{K \in \mathbb{N}}$ が連なった文字列 $\{z_{1:T}^{(n)}\}_{n=1}^{N}$ を，それぞれの尤もらしさを表す確率 $\{w_T^{(n)}\}_{n=1}^{N}$ とともに返す．それらで構成される確率分布こそが提案手法で獲得される表現である．

以下，提案手法を構成する認識モデルと生成モデル，そして提案手法の学習・推論手法を説明する．

3.1 認識モデル

提案手法において，認識モデルは時刻ごとに入力音声データに対応し得る文字の候補を N 個立てる役割を持つ．認識モデルは式 (1) に表したように，時刻に合わせて因数分解された確率分布をベースとする．時刻 t の音声データ $\boldsymbol{x}_t$ に対応する文字の候補は式 (1) の分布から抽出される．

$$q_\phi(z_{1:T} \mid \boldsymbol{x}_{1:T}) = \prod_{t=1}^{T} q_\phi(z_t \mid \boldsymbol{x}_{1:t}, z_{1:t-1}) \tag{1}$$

式 (1) において，各分布のパラメーターは時刻 t までの $\boldsymbol{x}_{1:t}$ と直前までの $z_{1:t-1}$ に依存している．この依存関係は毎時刻，$\boldsymbol{x}_t$ と z_{t-1} の埋め込み $\boldsymbol{e}_{z_{t-1}} \in \mathbb{R}^D$ を入力に更新される IndRNN [15] の内部状態 $\boldsymbol{r}_\phi(t)$ によって捉えられる．

提案手法の認識モデルは文字 z_t の候補を，そもそも対象言語で必要となる文字の種類数が未知の状況下で推定する．本研究ではそれを可能とするために，より高次の潜在変数として，文字の種類数 $K_t \in \{0, 1, 2...\}$ および各文字の存在確率 $\boldsymbol{\pi}_t \in \triangle^{K_t+2}$ を導入する．この階層構造を有した確率分布は時刻 t において式 (2) に表される．

$$\begin{aligned} q_\phi(z_t \mid \boldsymbol{x}_{\leq t}, z_{1:t-1}) &= q_\phi(z_t \mid \boldsymbol{\pi}_t, K_t, \boldsymbol{x}_{1:t}, z_{1:t-1}) \\ &\quad \times q_\phi(\boldsymbol{\pi}_t \mid K_t, \boldsymbol{x}_{1:t}, z_{1:t-1}) \\ &\quad \times q_\phi(K_t \mid \boldsymbol{x}_{1:t}, z_{1:t-1}) \\ &= Categorical(z_t \mid \boldsymbol{\pi}_t, K_t, \boldsymbol{x}_{1:t}, z_{1:t-1}) \\ &\quad \times Dirichlet(\boldsymbol{\pi}_t \mid K_t, \boldsymbol{x}_{1:t}, z_{1:t-1}) \\ &\quad \times Poisson(K_t \mid \boldsymbol{x}_{1:t}, z_{1:t-1}) \end{aligned} \tag{2}$$

式 (2) では K_t を全時刻の共通変数とせず，時刻毎で局所化している．これは全時刻共通の依存関係があると，全体を見る必要性から系列長が長くなりやすい音声データでは計算量が膨大になるためである．

また式 (2) の $Dirichlet$ 分布は，式 (3) の通りにパラメーター化される．K_t の増加に合わせて $\boldsymbol{\pi}_t$ が疎らになり，z_t の種類数の過度な増加が抑制される．

$$\begin{aligned} &Dirichlet(\boldsymbol{\pi}_t \mid K_t, \boldsymbol{x}_{1:t}, z_{1:t-1}) \\ &\quad = Dirichlet\left(\frac{\boldsymbol{\alpha}_{(1)_t}}{(K_t+2)}, ..., \frac{\boldsymbol{\alpha}_{(K_t+2)_t}}{(K_t+2)}\right), \\ &\quad \left[\boldsymbol{\alpha}_{(1)_t}, ..., \boldsymbol{\alpha}_{(K+2)_t}\right]^\intercal = MLP_{\phi,\pi}(\boldsymbol{r}_\phi(t)) \in \mathbb{R}_+^{K+2} \end{aligned} \tag{3}$$

さらに，式 (2) の $Categorical$ 分布のパラメーター計算には式 (4) を用いる．式 (4) は時刻 t で考慮する $K_t + 2$ 個の文字埋め込み $\{\boldsymbol{e}_k | \boldsymbol{e} \in \mathbb{R}^D\}_{k=1}^{K_t+2}$ と，認識モデルの埋め込み $\hat{\boldsymbol{e}}_t = MLP_{\phi,z_t}(\boldsymbol{r}_\phi(t)) \in \mathbb{R}^D$ の類似度 $\sigma(\boldsymbol{e}, \hat{\boldsymbol{e}}_t) = exp\left(\frac{-||\boldsymbol{e}-\hat{\boldsymbol{e}}_t||^2}{2}\right)$ で $\boldsymbol{\pi}_t$ に重み付けする．これにより，モデルが対応確率を上げるために両者を近づけるように学習するので，$\boldsymbol{e}$ に $\hat{\boldsymbol{e}}_t$ の情報が直接，含まれるように促す狙いがある．

$$\begin{aligned} &Categorical(z_t \mid \boldsymbol{\pi}_t, K_t, \boldsymbol{x}_{\leq t}, z_{<t}) \\ &\quad = Categorical(c_{(1)_t}, ..., c_{(K+2)_t}), \\ &c_{(i)_t} = \frac{\pi_{(i)_t}\sigma(\boldsymbol{e}_{(i)}, \hat{\boldsymbol{e}}_t)}{\sum_{j=1}^{K+2}\pi_{(j)_t}\sigma(\boldsymbol{e}_{(j)}, \hat{\boldsymbol{e}}_t)}, \quad i = 1, 2, ..., K+2 \end{aligned} \tag{4}$$

認識モデル内の DNN の構成は付録 A.1 に示す．

3.2 生成モデル

提案手法の生成モデルは認識モデルが挙げた文字の候補を評価する役割をもつ．その際に実際の音声データとモデルが持つ事前知識が参照される．生成モデルは式 (5) に表される．ここで式 (5) - ① は認識モデルの立てた文字候補の予測のしやすさをもとに評価値として算出する．そして式 (5) - ② は音声データの再構成誤差をもとに評価値を算出する．ベイズ推論の観点でいえば前者は事前分布，後者は尤度関数であるため，以後そのように呼ぶ．

$$p_\theta(\boldsymbol{x}_{1:T}, z_{1:T}) = \prod_{t=1}^{T} \overbrace{p_\theta(\boldsymbol{x}_t \mid z_t)}^{②} \overbrace{p_\theta(\boldsymbol{z}_t \mid z_{1:t-1}) p_\theta(z_1)}^{①} \tag{5}$$

また，ここで事前分布 (5) - ② における過去の文字列への依存関係は，認識モデルと同様に IndRNN [15] の内部状態 $\boldsymbol{r}_\theta(t)$ が捉える．なお，(5) - ② は機能的には認識モデルを正則化する項であるが，形式的には文字列上の言語モデルとみなせる．

生成モデルの時刻 $t = 1$ の事前分布は，式 (6) に示したように指定する．

$$\begin{aligned} p_\theta(z_1) &= p_\theta(z_t \mid \boldsymbol{\pi}_1, K_1) \times p_\theta(\boldsymbol{\pi}_1 \mid K_1) \times p_\theta(K_1) \\ &= Categorical(z_t \mid \boldsymbol{\pi}_1) \\ &\quad \times Dirichlet\,(\boldsymbol{\pi}_1 \mid 1/(K_1 + 2)) \\ &\quad \times Geometric(K_1 \mid 0.03) \end{aligned} \tag{6}$$

ここで，$p_\theta(K_1)$ は，図 1 で示した世界の言語の音素数が平均 34.93 であるという事前知識のみをもとに，自然数上で平均が既知の時に最も不確実な分布，すなわちエントロピー最大化分布の *Geometric* 分布を，弱情報事前分布として指定している．同分布はより小さい自然数にほど，確率質量が集中する．つまり，提案手法においては，認識モデルがより少ない文字の種類数をもとに文字の候補を立てた場合に，生成モデルがそれをより高く評価する．この節約性により，認識モデルが過度な値を設定することを抑えつつ，いわば音素らしいスケールのなかで適切な種類数を推定させることが可能になる．

時刻 $t > 1$ における生成モデルの事前分布は，式 (7) に示したように K_t を除いて認識モデルとともに学習される．式 (7) の *Dirichlet* 分布と *Categorical* 分布は，RNN の内部状態 $\boldsymbol{r}_\phi(t)$ を入力に，認識モデルで導入したパラメーター化手法の式 (3) と式 (4) によりパラメーターが算出される．K_t の事前分布は，全時刻において時刻 $t = 1$ と同一の分布を仮定する．

$$\begin{aligned} p_\theta(z_t \mid z_{1:t-1}) &= p_\theta(z_t \mid \boldsymbol{\pi}_t, K_t, z_{1:t-1}) \\ &\quad \times p_\theta(\boldsymbol{\pi}_t \mid K_t, z_{1:t-1}) \\ &\quad \times p_\theta(K_t \mid \boldsymbol{x}_{\le t}, z_{1:t-1}) \\ &= Categorical(z_t \mid \boldsymbol{\pi}_t, K_t, z_{1:t-1}) \\ &\quad \times Dirichlet(\boldsymbol{\pi}_t \mid K_t, z_{1:t-1}) \\ &\quad \times Geometric(K_1 \mid 0.03) \end{aligned} \tag{7}$$

尤度関数には，式 (8) に示すように，ガウス分布 $\mathcal{N}$ を指定する．ここで $\mu_\theta(.)$ と $\sigma^2_\theta(.)$ の算出にはそれぞれ適当な DNN が用いられる．

$$p_\theta(\boldsymbol{x}_t \mid z_t) = \mathcal{N}(\boldsymbol{x}_t \mid \mu_\theta(z_t), \sigma^2_\theta(z_t)\boldsymbol{I}) \tag{8}$$

生成モデル内の DNN の構成は付録 A.2 に示す．

3.3 学習・推論アルゴリズム

本研究で提案するモデルは，時系列性，階層性，離散性を有する複雑な潜在構造を学習・推論する必要がある．これに対して，本研究では NASMC [16] という，逐次モンテカルロ法と深層学習を組み合わせた学習・推論アルゴリズムの応用を提案する．

NASMC による推論では，重点抽出法を基礎とする式 (9) をもとに，その再帰的関係を利用して，認識モデルと生成モデルを通じて漸進的に重要度重み $w_T^{(n)}, n = 1, 2, ..., N$ を算出する．

$$\begin{aligned} w_T^{(n)} &= \frac{p_\theta(\boldsymbol{x}_{1:T}, z_{1:T}^{(n)}) / q_\phi(z_{1:T}^{(n)} \mid \boldsymbol{x}_{1:T})}{\Sigma_{m=1}^{N} p_\theta(\boldsymbol{x}_{1:T}, z_{1:T}^{(m)}) / q_\phi(z_{1:T}^{(m)} \mid \boldsymbol{x}_{1:T})} \\ &\propto w_{T-1}^{(n)} \frac{p(\boldsymbol{x}_t \mid z_T^{(n)}) p(z_T^{(n)} \mid z_{1:T-1}^{(n)})}{q(z_T^{(n)} \mid z_{1:T-1}^{(n)}, \boldsymbol{x}_{1:T})} \end{aligned} \tag{9}$$

結果として，$\{w_T^{(n)}\}_{n=1}^N$ と $\{z_{1:T}^{(n)}\}_{n=1}^N$ を合わせて，式 (10) に示す形で $z_{1:T}$ の事後分布の近似分布がデータ表現として得られる．なお，$\delta(.)$ は デルタ関数で，N を増やすほどに，より良い近似が達成される．

$$\sum_{n=1}^{N} w_T^{(n)} \delta(z_{1:T} - z_{1:T}^{(n)}) \approx p(z_{1:T} \mid \boldsymbol{x}_{1:T}) \tag{10}$$

認識モデルと生成モデルの学習においても，$w_T^{(n)}$ を用いて，両モデルの学習パラメーターが，それぞれ，式 (11) と式 (12) により個別の目的関数を最適化する形で同時に更新される．

$$\begin{aligned} &\nabla_\phi KL[p(z_{1:T} \mid \boldsymbol{x}_{1:T}) || q_\phi(z_{1:T} \mid \boldsymbol{x}_{1:T})] \\ &\quad \approx -\Sigma_{t=1}^{T} \Sigma_n^N w_t^{(n)} \nabla_\phi \log q_\phi(z_t \mid \boldsymbol{x}_{\le t}, z_{1:t-1}) \end{aligned} \tag{11}$$

$$\begin{aligned} &\nabla_\theta \log p_\theta(\boldsymbol{x}_{1:T}) \\ &\quad \approx \Sigma_{t=1}^{T} \Sigma_n^N w_t^{(n)} \nabla_\theta \log p_\theta(\boldsymbol{x}_t, z_t \mid z_{1:t-1}) p_\theta(z_1) \end{aligned} \tag{12}$$

4 実験評価

本実験は，提案手法の各文字とその埋め込み表現が，分節的単位の言語学的情報を捉えられているか，従来手法の VQVAE との比較を通じて評価した．

4.1 実験設定

訓練データセット Zero Resource Speech Challenge 2017（ZeroSpeech2017）[17] の英語（方言不明）のデータセットを利用した．これには，69 話者による合計約 45 時間の読み上げ音声で構成されている．

データ前処理 16kHz の各音声信号に，STFT（窓幅 400，シフト幅 160）をかけ，対数メル尺度（メル・バンド数 40）に変換した．

評価手法 本実験では次の 2 つの評価手法を用いた．以下のうち，前者では獲得文字の埋め込み表現が，後者では獲得文字それ自体が用いられた．

1. ABX 音素識別誤り率（ABX）[18]：獲得表現が各音素の弁別特徴を捉えているか，異なる音素間で表現の類似度をもとに評価する尺度．話者不変性の評価のために音素の標本が話者内と話者間から抽出される場合で計測される．本実験では埋め込みのコサイン類似度を利用した．ZeroSpeech2017 のテストデータを利用した．
2. 補正相互情報量（AMI）[19]：人手の書き起こしテキストと，各手法独自の文字体系で自動で書き起こされた疑似テキストの「対応の良さ」を，情報理論に基づき定量評価する尺度．評価データに TIMIT [20]（米国英語）の SI/SX 文および Lee らの文字体系（39 種類）[21] を用いた．

提案手法の実験設定 学習の詳細は B に示す．評価時には $N = 1024$ に固定した．ABX の算出にはベクトル表現を必要とするため，本実験では $\bar{e}_{1:T} = \sum_{n=1}^{N} w_{1:T}^{(n)} e_{1:T}^{(n)}$ を採用した．一方で，AMI の文字 z の出現頻度は各 $w_t^{(n)}$ で重み付けして測定した．

従来手法の実験設定 先行研究 [12] により公開されている VQ-VAE の実装[1)]をベースに，提案手法に合わせてデコーダーを対数メルスペクトログラムの再構築誤差で学習するようにした．

話者埋め込みの利用 本実験で比較する両手法には L2 正規化した同一の事前学習済みの話者埋め込み ECAPA-TDNN[2)] [22] を文字埋め込みに連結させる形でデコーダーの補助入力として利用した．

1） https://github.com/bshall/ZeroSpeech
2） https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

4.2 実験結果とその考察

表 1 ZeroSpeech 2017 英語のテストデータにおける ABX 音素識別誤り率

	文字の種類数	1 秒の文脈 話者内	1 秒の文脈 話者間	10 秒の文脈 話者内	10 秒の文脈 話者間	120 秒の文脈 話者内	120 秒の文脈 話者外
提案手法	44	17.50	24.71	17.68	24.68	18.29	25.13
VQVAE	32	17.60	24.16	17.37	24.17	17.35	24.34
	128	**15.36**	22.39	15.48	22.24	15.50	22.43
	512	15.66	22.50	15.61	22.22	15.46	22.28
	2048	15.63	**22.28**	**15.26**	**21.89**	**15.43**	**21.93**

表 2 TIMIT コーパスの人手書き起こしと各手法独自の文字体系に基づく自動書き起こしの補正相互情報量 (AMI)

	文字の種類数	AMI ([0, 1] ↑)
提案手法	44	**0.3163**
VQVAE	32	0.3064
	128	0.2645
	512	0.2606
	2048	0.2493

表 1 に示すように，ABX では提案手法が従来手法を上回ることはなかった．しかし表 2 により，AMI では提案手法の優位性が示された．

ABX では埋め込みの特徴量で評価していることから，従来手法のほうが埋め込みにより言語学的に有意義な情報を捕捉していることが示唆される．

AMI では直接，獲得した文字を人手で創られた文字体系を参照して評価している．その AMI における優位性から，提案手法で獲得される文字体系が，より「人間の文字らしい」ということが示唆される．

従来手法は文字の種類数が 2048 に設定されたときに ABX で最高性能を記録している．しかしこの数は英語諸方言の音素数の 39~45 [6] や図 1 の分布からも大きく逸脱する．一方で提案手法は，比較的妥当な数，最大 44 と推定している．

5 おわりに

本研究では，機械が「音素らしい」文字体系を創り出せるようにするために，深層ベイズの枠組みから，世界の言語の報告データに依拠する文字の種類数の弱情報事前分布を導入した教師なし学習の手法を提案した．従来手法との比較実験により，提案手法が人手でつくられた文字体系とより強く対応する文字体系を構築したことを示した．今後はモデルを改良しつつ，大規模なデータや多様な言語での評価，より高度な自然言語処理への展開を行う．

謝辞

本研究は, 国立研究開発法人産業技術総合研究所事業の令和 5 年度覚醒プロジェクトの助成を受けたものです。

参考文献

[1] Gary F. Simons and M. Paul Lewis. The world's languages in crisis: A 20-year update. In Elena Mihas, Bernard Perley, Gabriel Rei-Doval, and Kathleen Wheatley, editors, **Studies in Language Companion Series**, Vol. 142, pp. 3–20. John Benjamins Publishing Company, Amsterdam, 2013.

[2] David M. Eberhard, Gary F. Simons, and Charles D. Fennig. **Ethnologue: Languages of the World**. SIL International, Dallas, 26 edition, 2023.

[3] Steven Bird and Dean Yibarbuk. Centering the Speech Community. In Yvette Graham and Matthew Purver, editors, **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 826–839, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[4] Ewan Dunbar, Nicolas Hamilakis, and Emmanuel Dupoux. Self-supervised language learning from raw audio: Lessons from the Zero Resource Speech Challenge. **IEEE J. Sel. Top. Signal Process.**, pp. 1–16, 2022.

[5] Emmanuel Dupoux. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. **Cognition**, Vol. 173, pp. 43–59, April 2018.

[6] Steven Moran and Daniel McCloy, editors. **Phoible 2.0**. Max Planck Institute for the Science of Human History, Jena, 2019.

[7] Chia-ying 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 (Volume 1: Long Papers)**, pp. 40–49, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[8] Lucas Ondel, Lukaš Burget, and Jan Černocký. Variational Inference for Acoustic Unit Discovery. **Procedia Computer Science**, Vol. 81, pp. 80–86, January 2016.

[9] Michael Heck, Sakriani Sakti, and Satoshi Nakamura. Dirichlet Process Mixture of Mixtures Model for Unsupervised Subword Modeling. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, Vol. 26, No. 11, pp. 2027–2042, November 2018.

[10] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. In **Proceedings of the 31st International Conference on Neural Information Processing Systems**, pp. 6309–6318, Red Hook, NY, USA, 2017. Curran Associates Inc.

[11] Andros Tjandra, Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and Satoshi Nakamura. VQVAE Unsupervised Unit Discovery and Multi-Scale Code2Spec Inverter for Zerospeech Challenge 2019. In **Interspeech 2019**, pp. 1118–1122. ISCA, September 2019.

[12] Benjamin van Niekerk, Leanne Nortje, and Herman Kamper. Vector-Quantized Neural Networks for Acoustic Unit Discovery in the ZeroSpeech 2020 Challenge. In **Interspeech 2020**, pp. 4836–4840. ISCA, October 2020.

[13] Ewan Dunbar, Julien Karadayi, Mathieu Bernard, Xuan-Nga Cao, Robin Algayres, Lucas Ondel, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. The Zero Resource Speech Challenge 2020: Discovering Discrete Subword and Word Units. In **Interspeech 2020**, pp. 4831–4835. ISCA, October 2020.

[14] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes, May 2014.

[15] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN. In **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 5457–5466, June 2018.

[16] Shixiang (Shane) Gu, Zoubin Ghahramani, and Richard E Turner. Neural Adaptive Sequential Monte Carlo. In **Advances in Neural Information Processing Systems**, Vol. 28. Curran Associates, Inc., 2015.

[17] Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. The zero resource speech challenge 2017. In **2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)**, pp. 323–330, Okinawa, December 2017. IEEE.

[18] Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux. Evaluating speech features with the minimal-pair ABX task: analysis of the classical MFC/PLP pipeline. In **Interspeech 2013**, pp. 1781–1785. ISCA, August 2013.

[19] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. **Journal of Machine Learning Research**, Vol. 11, No. 95, pp. 2837–2854, 2010.

[20] Garofolo, John S., Lamel, Lori F., Fisher, William M., Pallett, David S., Dahlgren, Nancy L., Zue, Victor, and Fiscus, Jonathan G. TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1993.

[21] K.-F. Lee and H.-W. Hon. Speaker-independent phone recognition using hidden Markov models. **IEEE Trans. Acoust., Speech, Signal Processing**, Vol. 37, No. 11, pp. 1641–1648, November 1989.

[22] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In **Interspeech 2020**, pp. 3830–3834, October 2020.

A　提案手法の DNN 構成

提案手法は Pytorch (version: 2.0.1+cu118) で実装された．以下を構成する層は，いずれも基本的には Pytorch のデフォルトのハイパーパラメーター設定である．

A.1　認識モデルの DNN 構成

1. Causal 1D-CNN
   - {入力: 40, 出力: 512, カーネルの大きさ：4, 畳み込み幅：2}
   - 活性化関数：ReLU
2. MLP (I)
   - {入力: 512, 出力: 512, 中間層の数：4}
   - 活性化関数：ReLU
3. IndRNN [15]
   - {入力: 64+512, 出力: 256, レイヤー数：2}
   - 活性化関数：ReLU
4. MLP (II)
   - {入力: 512, 出力: 512×3, 中間層の数：1}
   - 活性化関数：ReLU
5. 式 (2) の各分布のパラメーターに写像する線形層

A.2　生成モデルの DNN 構成

1. 文字埋め込み
   - {次元数：64, 初期値：平均 0，標準偏差 0.01 の正規分布からのサンプル}
2. IndRNN [15]
   - {入力: 64, 出力: 256, レイヤー数：2}
   - 活性化関数：ReLU
3. MLP (I)
   - {入力: 256, 出力: 512×2, 中間層の数：1}
   - 活性化関数：ReLU
4. 式 (7) の各分布のパラメーターに写像する線形層
5. MLP (II)
   - {入力: 256, 出力: 512×3, 中間層の数：1}
   - 活性化関数：ReLU
6. Transposed 1D-CNN
   - {入力: 64 × 2, 出力: 512, カーネルの大きさ：2, 畳み込み幅：1}
   - 活性化関数：ReLU
7. MLP (III)
   - {入力: 192, 出力: 64, 中間層の数：2}
   - 活性化関数：ReLU
8. MLP (IV)
   - {入力: 64×2, 出力: 40×2 × 2, 中間層の数：2}
   - 活性化関数：ReLU
9. 式 (8) の分布のパラメーターに写像する線形層

B　学習時の実験設定

各モデルの最適化には Pytorch 実装 Adam を学習率 0.004 で使用した．バッチサイズは 64，候補数 N は 16 とした．検証データでの損失が収束したところで訓練を停止させた．

C　話者埋め込みを使わず訓練させた場合の評価結果

表 3　ZeroSpeech 2017 英語のテストデータにおける ABX 音素識別誤り率

	文字の 1 秒の文脈			10 秒の文脈		120 秒の文脈	
	種類数	話者内	話者間	話者内	話者間	話者内	話者外
提案手法	128	22.5	33.87	23.17	34.04	23.06	34.04
VQVAE	32	17.82	26.14	18.50	26.32	18.54	26.67
	128	16.31	24.13	15.98	23.72	15.87	23.57
	512	**14.76**	**22.07**	**14.70**	**21.45**	**14.49**	**21.29**
	2048	15.24	23.00	15.08	22.40	15.09	22.36

表 4　TIMIT コーパスの人手書き起こしと各手法独自の文字体系に基づく自動書き起こしの補正相互情報量 (AMI)

	文字の種類数	AMI ([0, 1] ↑)
提案手法	**128**	**0.2881**
VQVAE	32	0.2844
	128	0.2603
	512	0.2216
	2048	0.2177

D　提案手法の獲得した文字体系と人手で創られた文字体系の対応

図 2 に TIMIT コーパスをもとに構築した混同行列を視覚化した．

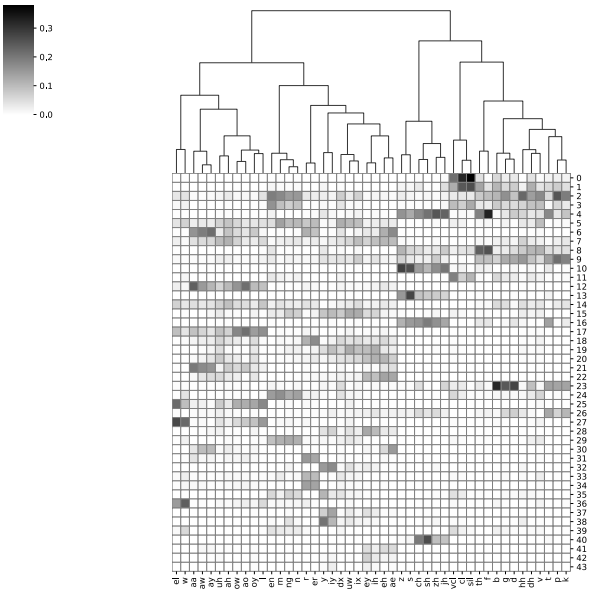

図 2　人手で創られた文字体系と提案手法が獲得した文字体系の混合行列．獲得した文字のどれがどの文字に反応する傾向にあるか示している．上部の樹形図はその反応パターンの類似度をもとに階層クラスタリングしたもの．