どのような音声離散表現が音声の再合成と継続に適するか？*

☆神藤駿介 (東大), 高道慎之介 (慶大/東大), △宮尾祐介 (東大)

1 はじめに

近年の大規模言語モデルの発展を受け，音声言語処理においても言語モデルを活用する研究が盛んになされている．Generative Spoken Language Model (GSLM; [1]) はテキスト資源を用いずに音声資源のみを学習に用いて構築される言語モデルである．GSLM はテキスト資源の少ない言語のモデリングを促進したり，音声に含まれるニュアンスや感情の情報を効果的に利用できるなど，様々な利点が期待できる．とりわけ，イントネーションや間といったテキストに現れない音声的特徴が重要となる音声対話や，音声刺激の影響が支配的な乳児期の言語獲得といった領域のモデリングに適している [2, 3].

GSLM は次の三つのモジュールを組み合わせて実現される：(1) 音声を離散表現に変換する，(2) 得られた離散表現で言語モデルを学習する，(3) 離散表現から音声を合成する．(2) で自己回帰的に離散表現の続きを生成し，それを (3) に入力して音声を合成することで，通常の言語モデルと同様に入力音声の続きを生成することが可能となる．(1) で抽出される音声の離散表現を**音声ユニット**と呼ぶ．音声ユニットは GSLM の中心的な役割を果たし，これを介することで音声のみを学習に用いた言語モデルを実現している．

音声ユニットは典型的には自己教師あり学習音声モデル [4] が出力する音声表現を量子化することによって得られる．しかし，自己教師あり学習音声モデルの出力は 20 ms 程度の短い音声に対応しており，音声ユニットの系列長が長くなる．特に Transformer ベースの言語モデルは系列長の二乗に比例して計算量が増加するため，長い系列は訓練に悪影響を及ぼす．このような課題に対処するため，音声表現を粗い単位に分節することで系列長を削減する手法が提案されている [5, 6]．神藤らは，音声表現を分節した上で量子化時のクラスタ数を増やすことで GSLM の音声言語理解の性能が向上することを実証した [7].

本研究は，神藤らの設定 [7] に倣い，どのような音声ユニットが GSLM の音声合成性能を高めるかを調査する．具体的には，音声ユニットから元の音声を合成する**音声再合成**，入力音声の続きを生成する**音声継続**の性能を評価する．実験の結果，音声言語理解の結果と同様に，分節して系列長を短くした上でクラスタ数を増やすことで性能が向上することが確認できた．これらの設定はビットレートを下げながらも十分な品質で音声を再合成でき，音声継続においてもより自然かつ多様な内容を生成できる．

2 GSLM の概要

2.1 構成要素

GSLM は以下の三つのモジュールを組み合わせて実現される．

speech2unit (s2u) 音声を離散表現（音声ユニット）に変換する．典型的には HuBERT [4] をはじめとする自己教師あり学習モデルが出力する音声表現に K-means クラスタリングを適用して実現される．

unit 言語モデル (uLM) 音声ユニットを用いて言語モデルを学習する．これにより音声の尤度を測定したり，自己回帰的に音声を生成することが可能となる．

unit2speech (u2s) 音声ユニットから音声を合成する．uLM が出力する音声ユニットに u2s を適用することで音声継続が可能となる．

3 つのモジュールを通して音声ユニット（すなわち s2u の出力）が中心的な役割を果たしており，これまでも s2u の改善を通して GSLM の性能を向上させる研究が数多くなされている [5, 6]. 神藤ら [7] は複数の s2u の設定下で uLM を構築して性能を比較し，どのような s2u が音声言語理解に適するかを調査した．本研究では，同じ設定下で u2s の性能評価を行い，どのような s2u が音声合成に適するかを調査する．

2.2 u2s の評価指標

u2s の性能を評価するタスクとして，音声ユニットから元の音声を合成する音声再合成，入力音声の続きを生成する音声継続がある．

2.2.1 音声再合成の性能評価

音声再合成は，s2u を用いて音声 $\boldsymbol{s}$ を音声ユニット系列 $\boldsymbol{u}$ に変換し，u2s を用いて $\boldsymbol{u}$ から音声 $\boldsymbol{s'}$ を合成するタスクである．音声再合成の性能は，音声ユニットの**ビットレート**および**「元音声の言語内容を復元できるか (CER)」「明瞭な音声か (UTMOS)」**という三つの観点で評価を行う．

---

*Which Types of Discrete Speech Representations are Suitable for Speech Resynthesis and Continuation? by KANDO, Shunsuke (The University of Tokyo), TAKAMICHI, Shinnosuke (Keio University/The University of Tokyo), MIYAO, Yusuke (The University of Tokyo)

ビットレート [bps]　音声ユニットのエントロピーと，単位時間あたりの音声ユニットの個数の平均値を掛け合わせることで算出する．

CER　元音声の内容を復元できるかを評価するため，$\boldsymbol{s}$ の書き起こしテキストと $\boldsymbol{s'}$ を音声認識した結果のテキストとの Character Error Rate (CER) を測定する．

UTMOS　出力音声の自然さを測定するため，本研究では UTMOS [8] を用いる．UTMOS は音声の自然さに関する Mean Opinion Score (MOS) を予測するモデルである．

2.2.2　音声継続の性能評価

音声継続の性能は**「自然な内容を生成できるか (PPL)」「多様な内容を生成できるか (VERT)」**という二つの観点で評価を行う．音声継続の結果からなるデータセット $S = \{\boldsymbol{s}_i\}_{i=1}^n$ に音声認識をかけ，書き起こしデータセット $T = \{\boldsymbol{t}_i\}_{i=1}^n$ を得る．PPL および VERT は T のもとで以下のように算出される．

PPL (Perplexity)　データセット T のパープレキシティ (PPL) は以下のように算出される．

$$\exp\left(-\frac{1}{n_{word}}\sum_{i=1}^{n}\log p(\boldsymbol{t}_i)\right)$$

ここで，n_{word} はデータセット中の単語の総数を表し，確率 $p(\boldsymbol{t}_i)$ は任意のテキスト言語モデルを用いて算出する．言語モデルは流暢な文に低いパープレキシティを与えることから，生成文の自然さをパープレキシティで定量化できる．

VERT (diVERsiTy)　VERT は self-BLEU と auto-BLEU の調和平均として定義される [1]．

self-BLEU は以下のように定義され，低ければ低いほど各生成文の間で重複が少なく，多様な文が生成されていることを表す．

$$\left(\prod_{i=1}^{n}\mathrm{BLEU}(\boldsymbol{t}_i, T\setminus\{\boldsymbol{t}_i\})\right)^{\frac{1}{n}}$$

特に温度パラメータの低い uLM は “fin, fin, fin,” のような繰り返しを生成しやすいが，self-BLEU ではこの問題を捉えにくい．そこで，各文内の繰り返し数を定量化する指標として以下に定義する auto-BLEU を用いる．

$$\left(\prod_{i=1}^{n}\frac{\sum_s \mathbf{1}[s\in(\mathrm{NG}(\boldsymbol{t}_i)\setminus\{s\})]}{|\mathrm{NG}(\boldsymbol{t}_i)|}\right)^{\frac{1}{n}}$$

ここで，NG(·) はテキスト中の n-gram のリストを返す関数であり，$\mathbf{1}[\cdot]$ は入力が真のときに 1 を返す指示関数である．

PPL と VERT の間のトレードオフについて

Lakhotia ら [1] が指摘しているように，uLM の温度パラメータによって PPL と VERT の間にはトレードオフが生じる．温度が低いときは似たような内容や繰り返しが生成されやすいため，PPL は低く，VERT は高くなりやすい．一方，温度が高いときはランダムな内容が生成されやすいため，PPL は高く，VERT は低くなりやすい．したがって，複数の温度パラメータでチューニングを行う必要がある．

3　検証対象の s2u の設定

本研究では神藤ら [7] の設定に則り，音声ユニットの分節幅 N とクラスタ数 K を定めて s2u を行い，両者の影響を調査する．まず HuBERT [4] の出力を N ms 幅で分節し平均プーリングをかける．HuBERT のフレームシフト幅は 20 ms であるため，N は 20 の倍数となる．本研究では $N = \{20, 40, 80, 120, 160, 200, 240, 280\}$ の 8 つの値を用いた．$N/20$ は平均プーリングされるフレーム数を表し，$N = 20$ は 1 フレームに対応する．K については $2^7 = 128$ から $2^{14} = 16384$ までの 8 つの値を用いた．以上，合計 $8\times 8 = 64$ 個の設定で s2u を行なった．

4　実験設定

s2u の K-means モデルは LibriSpeech [9] の訓練データの 100 時間 clean サブセットを用いて学習した．uLM は自己回帰型の Transformer 言語モデルである OPT [10] を採用し，LibriSpeech の全ての訓練データ (960 時間) を用いて学習した．s2u は text2mel モデルとして Tacotron2 [11] を，ボコーダーとして Parallel WaveGAN [12] を採用し，LJSpeech [13] を用いて訓練した．合成音声の書き起こしは `whisper-large-v3`[1] を用いた．

音声継続においては，LJSpeech の検証データのうち 6 秒以上の音声のはじめの 3 秒に対応する音声ユニットを uLM に入力し，`<eos>` トークンが出力されるまで生成を行なった．PPL は `Llama-3.1-8B`[2] を用いて測定した．$t \in \{0.3, 0.4, \ldots, 1.2\}$ の 10 個の温度パラメータで PPL (P_t) と VERT (V_t) を測定した．また，LJSpeech の検証データの書き起こしテキストの PPL (P_o) と VERT (V_o) を oracle とした．各評価値を min-max 正規化した上で $(P_t, V_t), (P_o, V_o)$ 間のユークリッド距離を測定し，それが最小となる温度を最適設定とみなし，当該 (N, K) における評価値とした．

[1] `huggingface.co/openai/whisper-large-v3`
[2] `huggingface.co/meta-llama/Llama-3.1-8B`

Table 1　各 (N, K) におけるビットレート．

N\K	2^7	2^8	2^9	2^{10}	2^{11}	2^{12}	2^{13}	2^{14}
280	31.6	36.8	41.7	46.1	50.5	54.7	58.0	61.2
240	36.7	42.6	48.3	53.6	58.7	63.7	67.9	71.8
200	44.3	50.8	57.6	63.9	69.9	76.0	81.4	86.5
160	54.4	63.2	71.6	78.8	86.8	94.3	101.5	108.5
120	69.9	81.9	92.6	103.0	112.8	123.4	133.4	143.1
80	96.5	116.1	130.3	145.4	159.4	175.9	192.4	208.7
40	152.3	181.0	211.8	237.7	265.1	299.6	334.9	372.8
20	194.3	237.2	282.3	326.0	372.5	431.3	497.4	576.7

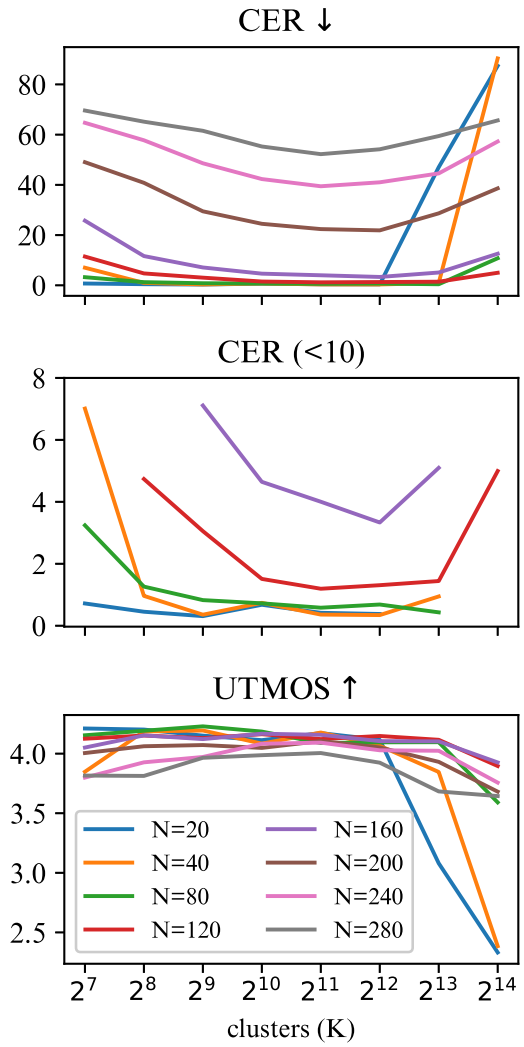

Fig. 1　音声再合成の評価結果．中段に CER が 10 より小さい設定の結果を拡大した図を示す．

5　結果と議論

5.1　音声再合成

CER と UTMOS の結果を Fig. 1 に示す．$(N, K) = (20, 2^{13}), (20, 2^{14}), (40, 2^{14})$ の性能は CER, UTMOS 共に例外的に非常に悪い．これらのケースはそもそも u2s の学習が収束しておらず，モデルの最適化によって改善すると考えられる．それ以外のケースについて，CER は $N \geq 200$ では大きく劣化することが観察された．CER が 10 より小さいケースに着目すると，$N = 20$ が安定して高い性能を示す一方，$N \geq 40$ においては K の値によって性能に差が出ることが分かった．その中でも $N = 120, 160$ においては性能が劣化する一方，$N = 40, 80$ における $2^{10} \leq K \leq 2^{13}$ の性能は $N = 20$ とほぼ差がない．UTMOS は $N \geq 200$ において性能の劣化が観察されるが，それ以外の設定においては概して大きな差はないことが分かった．

全体として，$N = 20$ と比べて $N = 40, 80$ の性能が CER, UTMOS 共に遜色ないことが分かる．Table 1 に示すように，ビットレートは N が大きくなるにつれて小さくなるため，より大きな N で性能が維持されることは望ましい結果であると言える．

5.2　音声継続

評価対象の音声の質を担保するため，音声再合成における CER が 20 以下かつ UTMOS が 4 以上の設定に絞って評価を行なった．結果を Table 2 に示す．全体として粗い分節単位の方が良い性能を示しており，特に VERT の値は同一の K において $N = 20$ よりも他の単位の方が良い傾向にある．個別に見ると，$(N, K) = (80, 2^{11}), (80, 2^{12}), (120, 2^{13}), (160, 2^{13})$ は PPL, VERT 共に $N = 20$ のどの設定よりも高い性能を示している．また，$N = 20$ のときには K の値が小さい方が性能が良く，その他の N においては K の値が大きい方が性能が良くなる傾向が観察された．以上の傾向は音声言語理解の性能を比較している先行研究と一貫しており [7]，GSLM の枠組全体を通して音声ユニットの単位を粗くしながらクラスタ数を大きくすることの利点が示唆された．

音声再合成と音声継続の結果を総合すると，$(N, K) = (80, 2^{11}), (80, 2^{12})$ が本研究で調査した中では最適な設定であると言える．これらの設定は $N = 20$ 設定よりもビットレートを下げつつ，CER と UTMOS は高いスコアを維持し，PPL と VERT はより良いスコアを記録している．

Table 2　音声継続の結果．上段は PPL↓・下段は VERT↓ を表す．CER と UTMOS の条件を満たさない設定は “-” で表す．$N = 200, 240, 280$ および $K = 2^{14}$ については全ての設定で条件を満たさなかった．$N = 20$ のどの設定よりも性能が高い設定を太字で示す．oracle の値は PPL=52.3, VERT=10.5 である．

$N \backslash K$	2^7	2^8	2^9	2^{10}	2^{11}	2^{12}	2^{13}
160	- -	392.4 19.1	290.7 22.0	247.9 21.5	216.5 19.8	248.4 15.9	**189.7** **17.4**
120	397.3 17.6	289.7 21.0	252.3 20.8	232.7 20.9	220.4 17.2	229.3 15.5	**137.9** **19.6**
80	312.1 22.9	278.7 19.0	230.0 17.5	224.9 17.1	**206.3** **18.1**	**191.0** **19.5**	238.8 15.3
40	- -	205.2 23.3	254.2 16.2	187.0 24.3	217.2 23.0	245.4 15.7	- -
20	210.0 25.7	220.4 25.1	241.4 21.7	237.4 21.0	220.4 22.2	252.5 22.3	- -

6　おわりに

本研究では，GSLM において s2u における分節幅・クラスタ数が u2s の性能にどのような影響を与えるかを調査した．uLM の音声言語理解の性能を調査した先行研究と一貫し，粗い単位に分節した上でクラスタ数を増やすことの利点が示唆された．本研究で得られた知見を応用することで，言語内容を深く理解し，一貫性のある生成が可能な言語モデルを，音声資源のみを用いて構築できることが期待される．

謝辞　本研究は，JST ACT-X (JPMJAX24C9) の支援を受けたものである．

参考文献

[1] K. Lakhotia, *et al.* “On generative spoken language modeling from raw audio.” TACL, Vol. 9, pp. 1336–1354, 2021.

[2] T. A. Nguyen, *et al.* “Generative Spoken Dialogue Language Modeling.” TACL, Vol. 11, pp. 250-266, 2023.

[3] M. Lavechin, *et al.* “Modeling early phonetic acquisition from child-centered audio data.” Cognition, Vol. 256, 105734, 2024.

[4] W.-N. Hsu, *et al.* “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units.” TASLP, Vol. 29, pp. 3451-3460, 2021.

[5] R. Algayres, *et al.* “Generative Spoken Language Model based on continuous word-sized audio tokens.” EMNLP 2023.

[6] A. Baade, *et al.* “SyllableLM: Learning Coarse Semantic Units for Speech Language Models.” ICLR 2025.

[7] S. Kando, *et al.* “Exploring the Effect of Segmentation and Vocabulary Size on Speech Tokenization for Speech Language Models.” Interspeech 2025. (to appear)

[8] T. Saeki, *et al.* “UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022.” Interspeech 2022.

[9] V. Panayotov, *et al.* “Librispeech: An ASR corpus based on public domain audio books.” ICASSP 2015.

[10] S. Zhang, *et al.* “OPT: open pre-trained transformer language models.” 2022.

[11] J. Shen, *et al.* “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions.” ICASSP 2018.

[12] R. Yamamoto, *et al.* “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram.” ICASSP 2020.

[13] K. Ito and L. Johnson, “The LJ Speech Dataset.” `https://keithito.com/LJ-Speech-Dataset/`, 2017.