Vol.2020-SLP-133 No.12
2020/10/23

音声の破損により失った文字情報を復元する音声認識

東 佑樹[1,a)]　Sakriani Sakti[1,2,b)]　中村 哲[1,2,c)]

概要：音声認識は近年飛躍的に進歩を遂げ，日常的に接する多くのシステムで実用化されている基盤技術である．実生活において音声認識システムは多様な外部環境での使用が想定される．入力音声はしばしば突発的な雑音によってマスクされ，認識率が低下する可能性がある．本稿では，外部環境の影響により一部が破損した音声の認識誤りを BERT，VQ-VAE それぞれを用いて低減させる手法を提案し，検証を行った．この結果，BERT による手法，VQ-VAE による手法の両方において音声認識システムの改善が確認できた．

1. はじめに

音声認識は近年「人間と同程度」の性能を発揮するほどめざましい発展を遂げ [1]，スマートスピーカーなど，普段の生活で触れる多くのシステムの基盤技術となっている．実生活において，音声認識システムは空港や街中など多様な環境での使用が想定されるため，環境に左右されずに認識精度を高い品質に保つことが重要となる．

一般に，音声認識は入力対象の音声以外の雑音の影響を受けやすく，雑音の多い環境下で認識率が低下するという問題がある．音声認識の学習に使用するデータセットに雑音に関する情報が含まれていれば，性能の低下を抑えられる可能性がある．しかし，多様な環境下においては重畳される外部雑音の性質をあらかじめ把握しておくことは困難である．そのため，多くの研究 [2], [3] では外部雑音を低減し，認識対象の音声信号を強調するモデルが提案されてきた．

しかしながら，突発的で大きな雑音がシステムに入り込んだ場合，入力音声はそのような支配的な雑音によって破損し，ノイズ削減のアルゴリズムを用いても元の音響情報を復元することは難しくなる．そのような雑音に対し頑健な音声認識システムを構築できれば，実環境における音声認識の品質を向上することができることが期待できる．

[1] 奈良先端科学技術大学院大学
NAIST, Takayama-cho, Ikoma, Nara 630-0192, Japan
[2] 理化学研究所 革新知能統合研究センター
RIKEN AIP, Chuo-ku, Tokyo 103-0027, Japan
[a)] azuma.yuki.ax1@is.naist.jp
[b)] ssakti@is.naist.jp
[c)] s-nakamura@is.naist.jp

そこで本稿では，音声認識システムに欠落した情報を復元する機構を加えることによって，支配的で急進的な雑音の影響をどの程度低減できるのか検証を行った．復元するシステムとして，BERT[4] を用いた後処理，VQ-VAE[5] を用いた前処理の 2 種類の手法を提案した．その結果，両手法において，雑音により一部が破損した音声の認識精度を改善できることを確認した．

2. 先行研究

破損した音声を用いて，音声認識の性能の低下を抑える研究はいくつかのアプローチで行われている．遠藤ら [6] は低帯域幅によるパケットロスが原因で音声信号の一部が欠落してしまう問題に対し，ミッシングフィーチャー理論（以下，MFT）[7] を適用した．MFT は音声認識の雑音に対する頑健性を獲得する手法の一つであり，欠損データを置換値で埋め合わせたり，無視することを指す．また，Srinivasan ら [8] は入力音声の一部をマスクし，欠損させた箇所に対応する画像を入力に加えたマルチモーダル音声認識システムを構築することで，音声信号から消失した単語情報の復元を行った．本研究では，音声認識のモデルに変更を加えず，適切な前処理や後処理を施すことで認識性能の改善を図る．

3. 提案手法

本章では，提案手法である BERT による後処理，VQ-VAE による前処理の詳細について述べる．

3.1 BERT による後処理

BERT[4] は Bidirectional Encoder Representations from

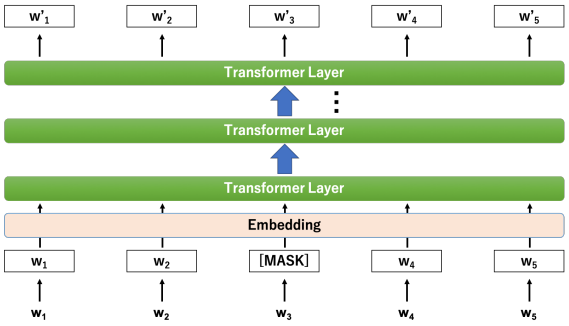

図 1　BERT のアーキテクチャ

図 2　BERT による後処理の概略

Transformers の略で，感情分析や文書分類など様々なタスクで高い性能を獲得した汎用的な言語処理モデルである．BERT は事前学習モデルであり，Masked Language Model（以下，MLM）と Next Sentence Prediction という 2 つの事前学習タスクを解くことで広域的な文脈情報を獲得している．MLM では入力文の一部の単語を [MASK] で置き換え，前後の文脈から元の単語を予測させる，というタスクを解く．BERT のアーキテクチャを図 **1** に示す．本稿では，音声認識の認識誤りの修復タスクにおいても有効であると仮定して，これを後処理の機構として用いた．以下，本手法を手法 1 と呼び，その概略を図 **2** に示す．

3.2　VQ-VAE による前処理

VQ-VAE[5] は Vector Quantised-Variational AutoEncoder の略で，生成モデルの 1 種である．Posterior Collapse と呼ばれる，潜在変数が強力なデコーダーにより無視されるという問題をベクトル量子化により解決することで，高品質な画像やビデオ，音声のサンプリングを可能にした．VQ-VAE のアーキテクチャを図 **3** に示す．本稿では，VQ-VAE が突発的な雑音により破損した音声情報を復元する能力があると仮定し，入力音声の前処理の機構として用いた．具体的には，VQ-VAE には一部をホワイトノイズに置き換えた音声を入力データとし，ノイズに置き換える前の音声を再構成するように学習させる．その後，音声認識に学習済みの VQ-VAE による合成音声を入力させ，認識

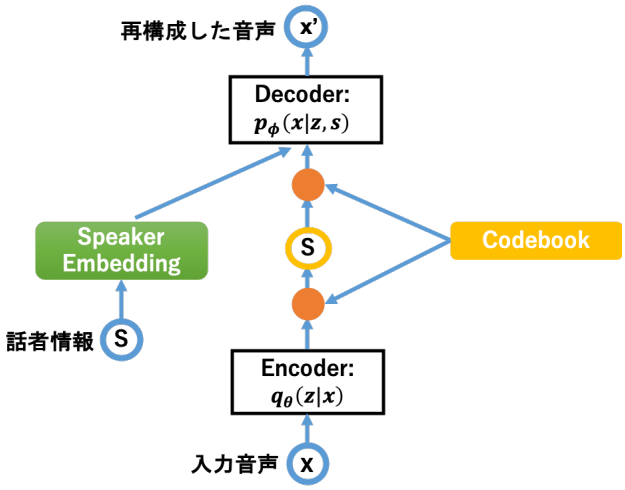

図 3　VQ-VAE のアーキテクチャ

誤り率に改善が見られたか検証を行う．以下，本手法を手法 2 と呼び，その概略を図 **4** に示す．

4.　実験

本章では，前項で述べた 2 種類の提案手法に対し，それぞれの認識精度を検証するための実験設定について述べる．

4.1　データセット

データセットは，単一女性話者による英語音声である LJSpeech（約 24 時間，1 万 3000 文）を用いた．本実験では，支配的な雑音により，音声情報がかき消される状況を再現するため，入力音声の 5%をランダムに選び，ホワイトノイズで置き換える．この操作をデータセット全体に対し 5 回独立に行い，データを拡張した．以下，ホワイトノイズで置き換える前のデータを Clean，置き換えて拡張したデータを Missing と呼ぶこととする．

4.2　音声認識モデル

音声認識のモデルには，隠れマルコフモデルやディープニューラルネットワークなどを用いた手法が知られているが，本研究では Attention 機構付きの Encoder-Decoder モデル [9], [10] を用いた．具体的にはエンコーダに三層の bidirectional LSTM（Hidden size: 256 × 3），デコーダに一層の unidirectional LSTM（Hidden size: 512），Attention 機構（Hidden size: 256）で構成されるモデルを使用した．また，手法 1 において，音声認識の出力はマスクをかけた後 BERT へ入力するため，BERT と同じ語彙サイズ，認識

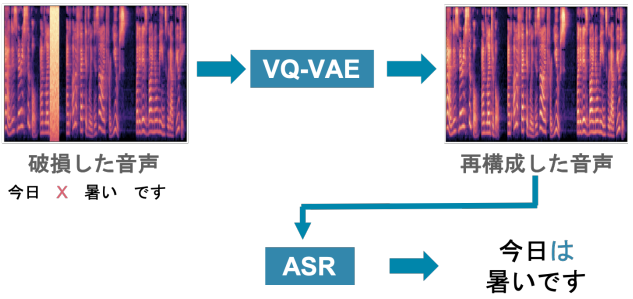

図 4　VQVAE による前処理の概略

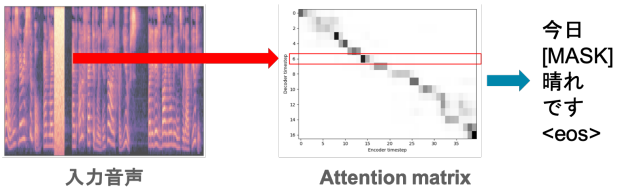

図 5　Attention matrix から認識誤り位置を予測する例

単位にする必要がある．そこで，語彙サイズは 30522，認識単位は WordPiece[11] を用いた．

4.3　手法 1：BERT による後処理

本手法では，BERT は学習済みの BERT-Base（Transformer Block×12，Hidden size: 768）を使用した．音声認識の Attention matrix は入力音声の各フレームが出力テキストの各単語の生成にどれだけ寄与したかを示すと考えられる．そこで，入力音声の中で雑音に置き換わった箇所に対応する出力単語を図 5 のように Attention matrix から推測し，該当単語を [MASK] で置き換え，BERT を用いて単語の復元を試みた．以下，この置き換え手法を Attention based と呼ぶ．また，BERT による音声認識誤りの復元能力のみを検証するために，認識誤りの箇所と誤った単語数を正解テキストを参照して正確に [MASK] に置き換えたテキストを用意して BERT へ入力し，Attention based と性能を比較した．以下，正解テキストを参照する置き換え手法を Reference based と呼ぶ．

4.4　手法 2：VQ-VAE による前処理

本手法では，モデルは話者情報を考慮した Conditional VQ-VAE[12] を使用した．VQ-VAE は入力，再構成音声ともに Clean に設定したもの，入力を Clean と Missing，再構成音声を Clean に設定した 2 つのモデルを学習させた．以下，前者により Clean を再構成した音声を Recon（Clean），後者により Missing を再構成した音声を Recon（Missing）と呼ぶことにする．Recon（Missing）はホワイトノイズによる音声情報の欠落が改善されると期待ができるが，再構成された音声は合成音声であるため，Clean のみで学習させた音声認識では，入力音声が学習データに存在しない未知の音声波形となる可能性があり，性能が低下するおそれがある．そこで，Clean のみを学習データに設定したものと，Clean と Recon（Clean）を学習データに設定した 2 つの音声認識モデルを学習させ，本手法の有用性を検証した．

5.　結果及び考察

手法 1 による結果を表 1 に，手法 2 による結果を表 2 に示す．

表 1 において，Attention based により認識誤り箇所を [MASK] に置き換えた結果，単語誤り率（以下，WER）はおよそ 0.8%増加した．もし Attention based による誤り箇所の推定が大きく失敗しているなら，WER はさらに大きく増大するはずである．しかしながら，実際には WER に大きな変化は見られなかったため，誤り箇所の推定はうまく機能していると考えられる．一方，Reference based の置き換え手法では WER が約 22.5%にまで減少し，Attention based と比較し 7.5%ほどの差が生じた．認識誤りには挿入，削除，置換誤りの 3 種類が存在するが，Attention based では誤認識単語の位置のみを推定しているため，置換誤りにのみ対処していると考えられる．さらに，Reference based は音声認識の誤り箇所全てに対処しているため，人為的に置き換えたホワイトノイズに由来しない誤り箇所も含めて正しく置き換えが施され，Attention based と比較し大きく WER が減少したものと考えられる．また，いずれの置き換え手法に対しても，BERT を適用させた結果，およそ 2.5-5%程度の WER の改善が見られた．しかしながら，BERT による推定によって正しく復元できた単語は冠詞やコンマ，ピリオドなどの，任意の文章に対し高頻度に出現するものが大半であり，名詞や動詞，副詞といった，その文章を特徴づける単語の多くは文章に無関係な単語に置き換わるという結果となった．そのため，言語情報のみを用いて BERT によって入力音声固有の欠損情報を復元することは困難であり，更なる精度向上のためには入力音声などの追加の情報が必要であると考えられる．

また，表 2 において，Clean+Recon（Clean）を用いて学習を行った音声認識では，Recon（Missing）の WER が Missing と比較し 6%程度改善した．これにより，VQ-VAE を用いることで，破損した音声から元の情報を復元し，音声認識の誤認識を低減させることができることが示された．また，Clean のみを用いて学習を行った音声認識と比較すると，入力音声が Clean の時の WER が 3%ほど改善することも確認できた．このことから，学習データに VQ-VAE

表 1 手法 1：BERT による実験結果

後処理	入力	出力例	WER(%)
	(Answer)	even so severe a critic as mister wakefield states that a stranger to the scene	
	Clean	even so severe a critic as mister wakefield states that its stranger to the scene .	16.798
	Missing	even so ##rre ##ls as mister wakefield ' s states that its stranger to the scene .	29.412
Masked Text (Attention based)	Missing	even so [MASK] [MASK] as mister wakefield ' s states that its stranger to the scene .	30.288
Missing+BERT (Attention based)	Missing	even so far strange as mister wakefield ' s states that its stranger to the scene .	27.722
Masked Text (Reference based)	Missing	even so [MASK] [MASK] [MASK] as mister wakefield states that [MASK] stranger to the scene	22.455
Missing+BERT (Reference based)	Missing	even so far a , as mister wakefield states that a stranger to the scene	16.932

表 2 手法 2：VQ-VAE による実験結果

学習データ	入力	出力例	WER(%)
	(Answer)	even so severe a critic as mister wakefield states that a stranger to the scene	
Clean	Clean	even so severe a critic as mister wakefield states that its stranger to the scene .	16.798
	Missing	even so ##rre ##ls as mister wakefield ' s states that its stranger to the scene .	29.412
	Recon (Clean)	even so severe a critic as a rec ##al field states that its stranger to the scene ,	22.844
	Recon (Missing)	even so powerless para ##ed as mis ##ess lin field states , that a stranger to the scene ,	33.731
Clean + Recon (Clean)	Clean	even so severe a credit as mister wakefield states , that a stranger to the scene .	13.697
	Missing	even so short credit as mister wakefield states , that a stranger to the scene .	30.933
	Recon (Clean)	even so severe credit as mister wakefield states that a stranger to the scene .	16.216
	Recon (Missing)	even so severe a printed as mister wakefield states , that of a stranger to the scene .	24.070

による再構成音声を加えることにより，音声認識の頑健性が向上する可能性が示唆された．一方で，Clean のみを用いて学習を行った音声認識では，Clean ，Missing ともに再構成音声による WER の改善は見られず，むしろ悪化する結果となった．VQ-VAE の出力は合成音声であり，入力の自然音声に近いが，一部が歪んだ音声となる．そのため，そのような歪みが音声認識によって未知の信号として扱われ，性能の劣化を引き起こしたと考えられる．

6. おわりに

本稿では，入力音声の一部が破損することによる音声認識システムの性能の劣化を抑えるため，BERT による後処理，VQ-VAE による前処理の 2 種類の手法を提案し，性能比較を行った．実験結果より，ホワイトノイズによる認識誤り箇所は，入力音声のノイズの位置をもとに Attention matrix から推定ができることを確認した．しかし，Attention matrix から推定できる誤りは置換誤りのみであり，挿入誤り，削除誤りへの対処が十分でないことが明らかになった．また，BERT により正しく修正できた単語の多くは，記号や冠詞などのデータセットに依存せず高頻度に出現する単語であり，データセット固有の単語の復元は十分に行えていないという結果になった．今後は 3 種類の認識誤り位置を推定するシステムの構築を目指し，言語情報以外のデータを BERT による単語推定に利用するモデルの構築を行う予定である．一方で，VQ-VAE を用いた前処理を行うことで，入力音声から欠落した情報を復元し，音声認識の誤り率を抑えることができることが確認できた．また，VQ-VAE により入力から再構成した音声を音声認識の学習データに加えることで，音声認識の頑健性が向上することが示唆された．

本実験では，ホワイトノイズに置き換える箇所やその割合，パワーによって音声認識の誤認識がどのように変化するのか十分に検証を行っていなかった．そのため，今後はノイズが出現する条件（位置，持続時間，パワー）の変化による誤認識の傾向の変化等について検証を行う予定である．

謝辞 本研究の一部は JSPS 科研費 JP17H06101 の助成を受けたものです．

参考文献

[1] 河原達也：音声認識技術の変遷と最先端－深層学習による End-to-End モデル－，日本音響学会誌, Vol.74, No.7, pp.381–386, 2018.
[2] Jingdong Chen, J. Benesty, Yiteng Huang and S. Doclo.: *New insights into the noise reduction Wiener filter*, IEEE Transactions on Audio, Speech, and Language Processing, vol.14, No.4, pp.1218-1234, 2006.
[3] A. Maas, Q. V. Le, T. M. Oneil, O. Vinyals, P. Nguyen, and A. Y. Ng.: *Recurrent neural networks for noise reduction in robust ASR*, in Proc. InterSpeech, pp.22–25, 2012.
[4] Jacob Devlin., Ming-Wei Chang., Kenton Lee., and Kristina Toutanova.: *BERT: Pre-training of deep bidirectional transformers for language understanding*, in Proc. NAACL: Human Language Technologies, vol.1, pp.4171–4186, 2019.
[5] A. van den Oord., O. Vinyals., et al.: *Neural discrete*

representation learning, in Advances in Neural Information Processing Systems, pp. 6306–6315, 2017.

[6] T. Endo, S. Kuroiwa, and S. Nakamura.: *Missing Feature Theory Applied to Robust Speech Recognition over IP Network*, IEICE transactions on information and systems 87(5), pp.1119-1126, 2004.

[7] P. Renevey, R. Vetter, and J. Krauss.: *Robust speech recognitionusing missing feature theory and vector quantization*, in Proc. InterSpeech, pp.1107–1110, 2001.

[8] T. Srinivasan, R. Sanabria and F. Metze.: *Looking Enhances Listening: Recovering Missing Speech Using Images*, in Proc. ICASSP, pp.6304-6308, 2020.

[9] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals.: *Listen, attend and spell: A neural network for large vocabulary conversational speech recognition*, in Proc. ICASSP, pp.4960–4964, 2016.

[10] A. Tjandra, S. Sakti, and S. Nakamura.: *Machine speech chain*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.28, pp.976–989, 2020.

[11] M. Schuster, and K. Nakajima.: *Japanese and Korean voice search*, in Proc. ICASSP, pp.5149-5152, 2012.

[12] 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*, arXiv preprint arXiv:1905.11449, 2019.