音声の離散特徴量を活用した End-to-End 音声翻訳における音声系列長の削減

阿久井 駿 [1]　鶴岡 慶雅 [1]
[1] 東京大学大学院 情報理工学系研究科
{akui,tsuruoka}@logos.t.u-tokyo.ac.jp

概要

本稿では，end-to-end 音声翻訳における音声系列長を動的に削減する新たな方法を提案する．一般に音声系列は非常に長いため，end-to-end 音声翻訳モデルに音声を入力する際は系列を短くする必要がある．従来手法では音素の疎密を考慮せずにフレームを等間隔で圧縮するため，非効率的である可能性がある．そこで本稿では，音声の離散特徴量を活用した教師なし音素セグメンテーションに基づいて音声系列を動的に圧縮する手法を提案する．同一セグメントに属する音声系列を平均して 1 つに圧縮してからモデルに入力することで，従来手法と同程度の翻訳精度を保ちつつ，系列長を従来手法の半分程度に削減することに成功した．

1　はじめに

音声翻訳は，ある言語の音声を別の言語のテキストに翻訳するタスクである．音声翻訳は従来，i) 音声認識，ii) テキストどうしの機械翻訳という 2 つのタスクを組み合わせることで実現されていた [1, 2]．しかしこの方法は，i) 音声認識で生じた認識誤りが機械翻訳に伝播して翻訳精度が低下する，ii) 音声のみが持つ韻律などの情報が失われる，iii) 2 つのモデルを経由するため翻訳文を生成するのに時間がかかる，といった欠点が指摘されていた [3, 4]．そこで近年では，こうしたデメリットを解消するために，書き起こし文を経由せずに翻訳文を直接生成する end-to-end 音声翻訳の研究が進められている [5, 6, 7]．

近年，テキストどうしの機械翻訳で Transformer [8] が高い翻訳精度を達成したことを受け，end-to-end 音声翻訳においても Transformer を利用した手法が提案されている [9, 10]．しかし，Transformer に音声系列を入力する際には，しばしばその計算量が問題となる．音声系列の入力には，一般的には 10 ms 程度の幅で抽出した特徴量が用いられることが多く，これは同じ内容のテキストと比べても非常に長い．Transformer の空間計算量は入力の長さの 2 乗に比例するため，長い音声系列を直接入力することは計算コストの観点から現実的ではない．

音声系列の長さを削減する既存の手法としては，畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）を利用した方法が挙げられる [9]．この手法では，音声系列に対して stride が 2 の CNN を 2 回適用することで，系列長をおよそ 4 分の 1 に削減している．しかしこの方法では，音声系列内における音素の疎密の違いによらず全フレームが等間隔で圧縮されることになるため，音素が密集している部分の情報が大きく失われる可能性がある．このほかには，音声系列とそれに対応する音素や文字とのアラインメントを学習し，隣接するフレームで同じ音素や文字を表すと推測されるものを 1 つに圧縮する方法も提案されている [11, 12, 13, 14]．これらの手法では，音素の疎密に応じて系列を動的に圧縮することができ，CNN を用いた手法より高い翻訳精度を達成している．しかし，これらの方法はいずれも音声に対応する書き起こし文や音素の情報が必要であり，文字を持たない言語に応用することが不可能である．

音声系列に対応する音素とのアラインメントを，音声のみを用いて教師なしで学習する方法も提案されており [15]，この手法を用いれば，書き起こし文を利用しなくても音素の疎密に応じて系列を動的に圧縮できることが期待される．そこで本稿では，書き起こし文を利用せず音声のみで学習した音素セグメンテーションを用いて，音声系列を動的に圧縮することを試みた．そして，従来の等間隔で圧縮する方法と比べた翻訳精度や圧縮率の差について検証した．

2 関連研究

2.1 Transformer を利用した End-to-end 音声翻訳

Transformer [8] は attention と呼ばれる機構を用いた encoder-decoder モデルであり，入力系列を受け取り隠れ状態にエンコードするエンコーダと，エンコーダが出力した隠れ状態と出力系列を受け取り次のステップの出力を生成するデコーダから構成される．Transformer はもともとはテキストどうしの機械翻訳のために提案されたモデルであるが，入力系列としてテキストの代わりに音声の特徴量を用いることで音声翻訳に対しても有効であることが示されている [9]．音声の特徴量には主に対数メルスペクトログラムが用いられるが，そのままでは系列長が長いため計算量が非常に大きくなり，GPU のメモリ容量内で学習を行うことが困難となる．そこで，音声系列に stride が 2 の CNN を 2 回適用し，系列長をおよそ 4 分の 1 に削減してから Transformer に入力することでこの問題に対処している．

2.2 音声の離散特徴量を利用した教師なし音素セグメンテーション

2.2.1 VQ-VAE

音声の離散的な特徴量を得る方法として，Vector-Quantized Variational Autoencoder（VQ-VAE） [16] を用いた手法が提案されている [17]．VQ-VAE は，i) エンコーダ，ii) ベクトル量子化，iii) デコーダ，の 3 つの部分から構成される．モデルの詳細な図は付録 A に記した．

エンコーダ　エンコーダはまず 1 次元の音声波形 $\boldsymbol{x} = (x_1, \ldots, x_N)$ を入力として受け取り，対数メルスペクトログラムを計算する．次に対数メルスペクトログラムを 5 層の CNN と線形変換に入力し，連続的な特徴量 $\boldsymbol{z} = (z_1, \ldots, z_T)$ を出力する．

ベクトル量子化　この部分は，コードブックと呼ばれる K 個の異なる学習可能なベクトルの集合 $\boldsymbol{e} = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K\}$ で構成される．ここでは，エンコーダが出力した特徴量 $\boldsymbol{z}$ を受け取り，各要素 z_i をコードブックの中で最も L^2 距離が近いベクトルに置換することで離散的な特徴量 $\hat{\boldsymbol{z}} = (\hat{z}_1, \ldots, \hat{z}_T)$ を得る．

$$\hat{z}_i = \underset{\hat{z}'_i \in \boldsymbol{e}}{\arg\min} \|z_i - \hat{z}'_i\|_2^2 \quad (i = 1, \ldots, T) \tag{1}$$

この操作は微分不可能であるため，誤差逆伝播の際は straight-through estimator [18] を用いて式 (2) のように勾配を近似する．ここで $\mathscr{L}$ は損失関数である．

$$\frac{\partial \mathscr{L}}{\partial z} \approx \frac{\partial \mathscr{L}}{\partial \hat{z}} \tag{2}$$

デコーダ　デコーダは離散特徴量 $\hat{\boldsymbol{z}}$ を受け取り，各フレームを 50 % の確率で隣接するフレームに置換する time-jitter 正則化 [19] を行う．その後，話者を表す埋め込みベクトルと結合し，回帰型ニューラルネットワーク（Recurrent Neural Network; RNN）を用いて元の音声波形を復元する．

モデルは，式 (3) で表される損失関数 $\mathscr{L}$ を最小化するように学習される．

$$\mathscr{L} = -\frac{1}{N}\sum_{i=1}^{N} \log p(x_i \mid \hat{\boldsymbol{z}}) + \beta \frac{1}{T}\sum_{i=1}^{T} \|z_i - \mathrm{sg}(\hat{z}_i)\|_2^2 \tag{3}$$

第 1 項は再構成誤差，第 2 項はエンコーダの出力を離散特徴量に近づける項である．ここで，$\mathrm{sg}(\cdot)$ は勾配を計算しないことを意味する．β はハイパーパラメータであり，本稿では $\beta = 0.25$ を用いた．

2.2.2 VQ-VAE による音素セグメンテーション

VQ-VAE で得られる離散的な特徴量について，隣接するフレームで異なるエントリが選ばれている部分を音素の境界とみなし，教師なしでセグメンテーションを行う方法が提案されている [15]．しかしこの方法では，実際の音素と比べて細かく分割されすぎてしまう問題がある．そこで，全体のセグメント数に制約を設けることで，より音素境界に近いセグメンテーションが可能である．具体的には，連続特徴量とエントリ間の L^2 距離の 2 乗の和に，ペナルティ項として全体のセグメント数を加えた関数 $\mathscr{C}$ を最小化するように離散特徴量 $\hat{\boldsymbol{z}}$ を選択する．

$$\begin{aligned} \hat{z}_1, \ldots, \hat{z}_T &= \underset{\hat{z}'_1, \ldots, \hat{z}'_T}{\arg\min}\, \mathscr{C}(\hat{z}'_1, \ldots, \hat{z}'_T) \\ \mathscr{C}(\hat{z}'_1, \ldots, \hat{z}'_T) &= \sum_{i=1}^{T} \|z_i - \hat{z}'_i\|_2^2 + \lambda\, \mathrm{seg}(\hat{z}'_1, \ldots, \hat{z}'_T) \end{aligned} \tag{4}$$

ここで $\mathrm{seg}(\hat{z}'_1, \ldots, \hat{z}'_T)$ は系列 $(\hat{z}'_1, \ldots, \hat{z}'_T)$ に含まれるセグメント数を表す．また，λ はハイパーパラメータであり，本稿では $\lambda = 3$ を用いた．この最小化問題は，$\alpha_t = \min_{\hat{z}'_1, \ldots, \hat{z}'_t} \mathscr{C}(\hat{z}'_1, \ldots, \hat{z}'_t)$ とおくと，動的計画法を用いて式 (5) のように解くことができる．

$$\alpha_t = \begin{cases} 0 & (t = 0) \\ \min\limits_{j=1}^{t} \left(\alpha_{t-j} + \min\limits_{k=1}^{K} \|z_i - \boldsymbol{e}_k\|_2^2 + \lambda \right) & (t = 1, \ldots, T) \end{cases} \tag{5}$$

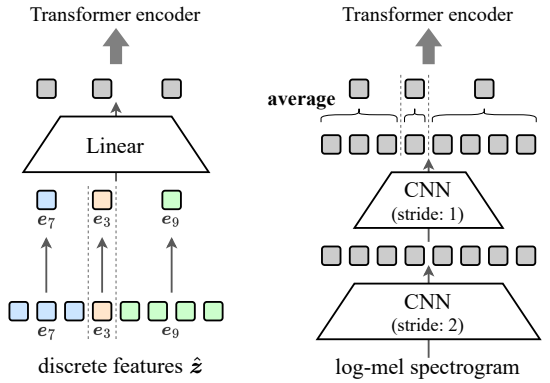

図 1　セグメンテーションを用いて音声系列を圧縮する概略図．(a) VQ-VAE の離散特徴量をそのまま用いる方法（**discrete**）と，(b) 同じセグメントに属する特徴量を平均する方法（**averaged**）の 2 通りを提案した．

3　提案手法

3.1　音声系列の圧縮

本稿では，end-to-end 音声翻訳モデルに入力する音声系列に 2.2 節で述べたセグメンテーションを適用し，同じセグメントに割り当てられたフレームを 1 つに圧縮することで，音声系列長を動的に削減することを試みた．フレームの圧縮には，図 1 に示す 2 通りの方法を考えた．

Discrete　この方法では，音声系列として各セグメントに割り当てられた VQ-VAE の離散特徴量をそのまま用いる．各離散特徴量に線形変換を施し，Transformer のエンコーダに入力する．

Averaged　この方法ではまず，2.1 節で述べた従来手法と同様に，音声から抽出した対数メルスペクトログラムに 2 層の CNN を適用する．ただし，本手法では 2 層目の CNN の stride は 1 とする．こうして得られた特徴量の各フレームをセグメンテーションと照らし合わせ，同じセグメントに属するフレームの特徴量を平均して 1 つのベクトルに統合する．

3.2　セグメントの長さに関する情報の付加

3.1 節で述べた方法では，各セグメントがその長さによらず 1 つのベクトルに圧縮されるため，セグメントの長さに関する情報が失われてしまう．そこで，圧縮後の各ベクトルの末尾にセグメントの元のフレーム数を追加し，セグメントの長さに関する情報を補完することを試みた．

4　実験

4.1　実験設定

提案した音声系列長の削減方法の有効性を検証するため，3.1 節で述べた 2 種類の手法（**discrete**, **averaged**）を，CNN のみを用いて系列を圧縮する従来手法（**baseline**）[9] と比較する実験を行った．提案手法で利用するセグメンテーションには，VQ-VAE で得られる離散特徴量をそのまま用いたセグメンテーション（**vq**）と，セグメント数に応じたペナルティを追加したセグメンテーション（**dp**）の 2 種類を用いた．また，3.2 節で述べた方法で圧縮後の特徴量に各セグメントの長さを付加した場合，実験結果にどのような違いが現れるかを比較した．翻訳精度の評価には BLEU [20] を用いた．

4.1.1　VQ-VAE

音声翻訳に用いるデータセットには話者の情報が含まれていないため，VQ-VAE の学習には話者アノテーション付きの英語コーパスである Buckeye [21] に含まれる約 18 時間分の音声を使用した．モデルには 2.2 節で述べた VQ-VAE モデルを使用した．学習の詳細は先行研究 [17] に従い，学習データからサンプルした 320 ms の音声系列 52 個を 1 つのバッチとしてモデルを学習した．パラメータの最適化には Adam [22] を使用し，学習率は 4×10^{-4} で初期化した後 300,000 ステップと 400,000 ステップで半分に減衰するように設定した．また，コードブックの最適化には指数移動平均を用いた方法 [16] を使用した．モデルは合計で 500,000 ステップ学習した．

4.1.2　End-to-end 音声翻訳

End-to-end 音声翻訳のデータセットには，MuST-C [23] の英語–ドイツ語コーパスを用いた．このデータセットには，学習データ 250,942 文，検証データ 1,415 文，テストデータ 2,580 文が含まれる．音声の特徴量には 10 ms ごとに 25 ms 幅の窓で抽出した 80 次元の対数メルスペクトログラムを使用した．また，テキストは SentencePiece [24] を用いてサブワードに分割した．モデルの学習時は，音声特徴量の系列長が 2,000 フレームを超える文を除外し，各バッチに含まれる音声系列のフレーム数の合計が最大で 20,000 となるようにバッチ化した．パラメータの最適化には Adam [22] を使用し，8 バッチ

表 1 MuST-C 英語–ドイツ語コーパスのテストデータを用いた各手法の BLEU スコア．手法名のかっこ内はセグメンテーションの方法を表す．また，normal は圧縮後の特徴量に各セグメントの長さを付加しなかった場合の結果，concat は付加した場合の結果である．

手法名	BLEU	
	normal	concat
baseline	19.15	–
discrete（vq）	9.34	11.12
discrete（dp）	6.41	9.09
averaged（vq）	18.72	18.01
averaged（dp）	**19.17**	18.56

ごとにパラメータを更新した．また，学習率は最初の 10,000 ステップで 0 から 2×10^{-3} まで線形に増加させ，その後はステップ数の平方根に反比例するように減衰させた．モデルは合計で 100,000 ステップ学習した．その他のハイパーパラメータは付録 B に記した．

4.2 実験結果

MuST-C 英語–ドイツ語コーパスのテストデータにおける各手法の BLEU スコアを表 1 に示す．離散特徴量をそのまま用いた手法（discrete）は従来手法と比べて大きく翻訳精度が下がる結果となった．Discrete 内で比較すると，セグメンテーションには dp よりも vq を用いたほうが翻訳精度は高かった．また，セグメントの長さに関する情報を付加することで BLEU スコアに 2 ポイント前後の向上が見られた．一方，セグメンテーションに基づいて特徴量を平均した場合（averaged）は従来手法と同程度の翻訳精度が維持されるか従来手法よりやや劣る結果となった．Averaged 内で比較すると，セグメンテーションには vq よりも dp を用いたほうが翻訳精度は高く，この方法では従来手法と比べて BLEU スコアで 0.02 ポイントの改善が見られた．この手法では，セグメントの長さの情報を付加することによる翻訳精度の向上は見られなかった．

4.3 考察

Discrete の翻訳精度が従来手法と比べて大きく劣っていることから，音声翻訳の入力に離散特徴量をそのまま用いるだけでは，翻訳に必要な音声の情報が大きく欠落してしまうと考えられる．セグメントの元の長さを補うことで翻訳精度は向上したものの，依然として従来手法の翻訳精度を大きく下回っているため，離散特徴量を音声翻訳に活用するためにはさらなる工夫が必要である．

表 2 MuST-C 英語–ドイツ語コーパスにおける，セグメンテーションの種類ごとの音声系列長の平均圧縮率．

セグメンテーション	平均圧縮率（%）
baseline	25.06
vq	35.09
dp	**13.21**

Averaged では，最も BLEU スコアが高かった dp セグメンテーションを用いた方法でも従来手法と同程度の翻訳精度にとどまった．しかし，表 2 に示した通り，dp セグメンテーションを用いた方法では音声系列長を従来手法のおよそ半分に圧縮できている．このことから，提案手法は従来手法の翻訳精度を維持しながら音声の系列長を大きく削減できるという点で有効であると考えられる．

5 おわりに

本稿では，end-to-end 音声翻訳において音声系列長を動的に削減するため，VQ-VAE を利用した教師なし音素セグメンテーションに基づいて音声系列を圧縮する手法を提案した．実験の結果，同一セグメントに属する音声系列を平均して 1 つに圧縮してからモデルに入力することで，従来手法と同等の翻訳精度を保ちながら系列長を従来手法の半分程度にまで削減することに成功した．

今後の課題としては，以下の 3 点が挙げられる．

教師ありセグメンテーション手法との比較　本実験では，ベースラインとして CNN を用いてフレームを等間隔で圧縮する方法のみを用いた．今後は教師ありセグメンテーションを使用した従来手法との比較実験も行い，翻訳精度や圧縮率，計算量などの違いについて検証する必要がある．

音声翻訳に適した離散表現の獲得　本実験では，音声翻訳の入力に離散特徴量を用いた場合，翻訳精度が大きく下がる結果になった．離散特徴量を音声翻訳に活用するためには，より音声翻訳に適した離散表現を得る方法を模索する必要がある．

低リソースな条件下での実験　提案手法のメリットの 1 つに，音声の書き起こし文が必要でないため文字を持たない言語に応用できることが挙げられる．そうした言語では，本実験のように大規模な学習データが集められない可能性が高い．そこで，低リソースな条件下でも本手法が有効であるかを実験して検証する必要がある．

参考文献

[1] Fred Stentiford and M.G. Steer. Machine translation of speech. **British Telecom Technology Journal**, Vol. 6, No. 2, pp. 116–123, 1988.

[2] Hermann Ney. Speech translation: Coupling of recognition and translation. In **1999 IEEE International Conference on Acoustics, Speech, and Signal Processing**, Vol. 1, pp. 517–520, 1999.

[3] Nicholas Ruiz and Marcello Federico. Assessing the impact of speech recognition errors on machine translation quality. In **Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track**, pp. 261–274, 2014.

[4] Matthias Sperber and Matthias Paulik. Speech Translation and the End-to-End Promise: Taking Stock of Where We Are. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7409–7421, 2020.

[5] Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. **arXiv:1612.01744 [cs]**, 2016.

[6] Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. In **Interspeech 2017**, pp. 2625–2629, 2017.

[7] Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. End-to-End Automatic Speech Translation of Audiobooks. In **2018 IEEE International Conference on Acoustics, Speech and Signal Processing**, pp. 6224–6228, 2018.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In **Proceedings of the 31st International Conference on Neural Information Processing Systems**, pp. 6000–6010, 2017.

[9] Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. Fairseq S2T: Fast Speech-to-Text Modeling with Fairseq. In **Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations**, pp. 33–39, 2020.

[10] Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. ESPnet-ST: All-in-One Speech Translation Toolkit. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, pp. 302–311, 2020.

[11] Elizabeth Salesky, Matthias Sperber, and Alan W Black. Exploring Phoneme-Level Speech Representations for End-to-End Speech Translation. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 1835–1841, 2019.

[12] Elizabeth Salesky and Alan W Black. Phone Features Improve Speech Translation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2388–2397, 2020.

[13] Marco Gaido, Mauro Cettolo, Matteo Negri, and Marco Turchi. CTC-based Compression for Direct Speech Translation. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 690–696, 2021.

[14] Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. Speechformer: Reducing Information Loss in Direct Speech Translation. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 1698–1706, 2021.

[15] Herman Kamper and Benjamin van Niekerk. Towards Unsupervised Phone and Word Segmentation Using Self-Supervised Vector-Quantized Neural Networks. In **Interspeech 2021**, pp. 1539–1543, 2021.

[16] 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, 2017.

[17] 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, 2020.

[18] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. **arXiv:1308.3432 [cs]**, 2013.

[19] Jan Chorowski, Ron J. Weiss, Samy Bengio, and Aaron van den Oord. Unsupervised Speech Representation Learning Using WaveNet Autoencoders. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, Vol. 27, No. 12, pp. 2041–2053, 2019.

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A Method for Automatic Evaluation of Machine Translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[21] Mark A. Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. **Speech Communication**, Vol. 45, No. 1, pp. 89–95, 2005.

[22] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In **3rd International Conference on Learning Representations**, 2015.

[23] Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: A multilingual corpus for end-to-end speech translation. **Computer Speech & Language**, Vol. 66, p. 101155, 2021.

[24] Taku Kudo. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 66–75, 2018.

A　VQ-VAE モデルの構成

2.2 節で述べた VQ-VAE モデルの詳細な構成を図 2 に示す．モデルの構成やパラメータはすべて先行研究 [17] に従った．

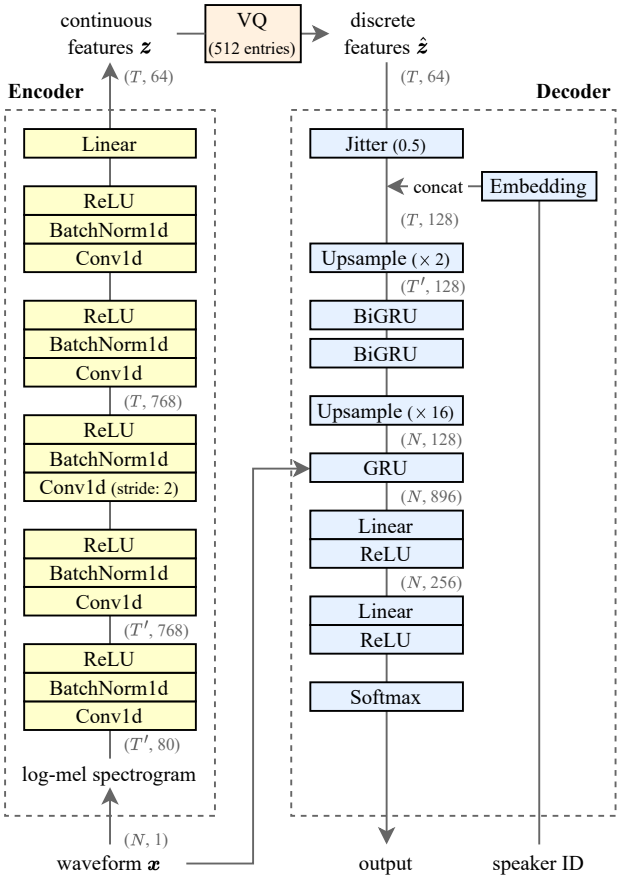

図 2　VQ-VAE モデルの概略図 [17]．図中の $(N, 1)$ のような数値は各特徴量の長さおよび次元数を表す．N は音声波形の長さ，T' は音声から抽出した対数メルスペクトログラムの長さ，T は離散特徴量の長さである．

B　ハイパーパラメータ

本実験で使用した end-to-end 音声翻訳モデルのハイパーパラメータを表 3 に示す．

表 3　End-to-end 音声翻訳モデルのハイパーパラメータ．

パラメータ名	値
エンコーダの層数	12
デコーダの層数	6
attention head 数	4
Transformer の潜在変数の次元数	256
フィードフォワードネットワークの中間層の次元数	2048
ドロップアウト	0.1