transformer を用いたオンライン会議音声の欠落回復
Recovery of Packet Loss in Online Meeting Audio Using Transformer

稲村直哉 (Naoya Inamura)*

法政大学 情報科学部 コンピュータ科学科
naoya.inamura.3u@stu.hosei.ac.jp

Abstract

When packet loss occurs in streaming audio, the audio contained in the lost packet is missing. The missing part of the audio must be recovered on the receiving side due to communication specifications. When recovering the missing part of the audio, it is typically restored by predicting the missing part from the preceding and succeeding waveforms. However, this method only uses waveform information and cannot utilize linguistic information. In this paper, missing part interpolation for streaming audio was performed using both waveform information and linguistic information by employing audio inpainting and speech recognition. A Word Error Rate (WER) of 0.142% was achieved. This result is 0.015 points higher than the WER of speech recognition without audio inpainting, demonstrating the importance of using both waveform and linguistic information.

1 はじめに

インターネットの発展により、電話回線を使う固定電話から、インターネット回線を使う IP 電話へと置き換わりつつある。IP 電話は携帯での音声通話だけでなく、オンライン会議やリモートワークなど、生活の様々なところで使用されている。近年では、新型コロナウイルスの流行もあり、オンライン会議などの需要はさらに増加している。IP 電話とは、音声を IP パケットに変換してインターネットを介して通信する仕組みの電話である。仕組み上、IP 電話の音声品質はインターネット回線の状況に左右される。たとえば、インターネットの回線状況が悪く、パケットロスが発生することにより、音声の欠落が起こる。この音声の欠落は、会話の理解困難やユーザのストレスにつながる。インターネットを介した通話では、遅延を抑え、スムーズで中断のないユーザ体験が必要となる。そのため、通信ではベストエフォートなプロトコルを用いる。たとえば、zoom や teams などのオンライン会議サービスではトランスポート層に UDP と RTP を用いている [8][1]。UDP はデータの送受信前に正式な接続を試みずに転送を行うため高速な通信が可能となっている。しかし、パケットロスの検知や損失したパケットの再送は行えない。RTP はヘッダーにシーケンス番号やタイムスタンプなどのデータを保持している。この 2 つから、パケットロスが発生した際には、データの再送ではなく、データ受信側でパケットの回復を行う。

データ受信側でパケットの損失を回復することをパケット損失回復という。これにより損失したパケットによる欠落を埋めて、音声認識への影響を減らす。リアルタイムな音声通話では、低遅延でパケット損失の回復を行う必要がある。計算コストが低く簡単に実装できる手法に、ノイズ置換や直前パケットの繰り返しなどがある。ノイズ置換とは、音声の欠落部分にノイズを挿入する手法である。特にホワイトノイズを挿入することで可理解性と主観的な品質が向上 [2] する。繰り返しは、パケットの損失部分の直前のフレームを繰り返し挿入することで欠落を埋める手法である。近年では、ニューラルネットワークを用いた回復方法の研究が盛んである。この手法ではパケット損失の直前の数フレームから損失部分を予測し欠落部分に挿入することで回復をおこなう。TFGAN-PLC[5] では、時間領域の波形と周波数領域の複素スペクトルを生成対向ネットワークを用いて訓練し、損失部分の生成を行う。tPLCnet[6] では、全結合の畳み込み層で波形を訓練することで、欠落部分の波形を予測する。

これらの方法では、波形やスペクトログラムなどの音声情報を用いて波形の回復を行うが、文脈などの言語情報を用いることができない。音声を正確に回復するためには、言語情報が必須である。言語情報を用いるためには波形を文字にする必要がある。欠落のある波形を音声情報と言語情報を用いて回復するためには、音声情報を用いて波形を回復したあと、文字に変換し言語情報を用いて回復を行い、再度波形に変換しなければならない。

本論文では、音声情報を用いた波形の回復とその後の言語情報を用いた回復までに焦点を当てる。音声情報を用いて波形の欠落回復を行ったあと、回復しきれていない部分を音声認識を行うことで言語情報を用いて回復する。欠落のあるストリーミング音声に対し、音声情報と言語情報を用いた回復処理を行い、処理結果として出力される文章を比較することで、2 つの情報を用いた回復の有用性を示す。

2 先行研究

トランスフォーマーモデル [4] は自然言語処理分野のタスクで卓越した性能を発揮している。しかし通常のトランスフォーマーベースのモデルはストリーミング音声を処理することが困

* 指導教員：伊藤克亘 教授

難である。そこで遅延と自己注意の計算量を削減することでストリーミング音声に対応できるようにする。AM-TFR[7] はス

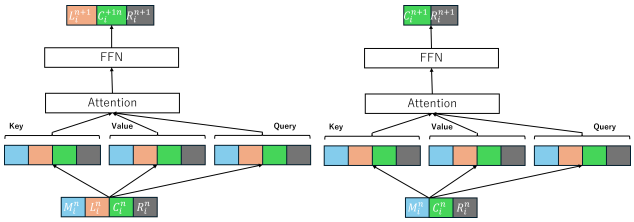

図 1　AM-TFR と emformer の比較

トリーミング音声に対応したトランスフォーマーである。図 1 左図が AM-TFR の構造を示している。まず入力特徴ベクトルのシーケンスを重ならないように分割する。分割したセグメントの一つである C_i^n とその左右のセグメント L_i^n と R_i^n を合わせて $X_i^n = [L_i^n, C_i^n, R_i^n]$ をつくる。第 i セグメントで X_i^n とメモリバンクを受け取り、X_{n+1}^n と m_i^n を出力する。m_i^n はメモリバンクに挿入される。最終的に C_i が連結しエンコーダに渡される。

AM-TRF では左側コンテキストが毎ステップで計算されるが、以前のセンターブロックとオーバーラップしている。emformer では、X_i^n の左側セグメントのかわりにメモリバンクを用いることで計算量を削減した。また、メモリバンクは将来のセグメントのために過去のコンテキスト情報を保持している。emformer ではメモリバンクを同じ層ではなく下位層から取得することで、書く emformer 層がシーケンス全体を並列に処理できるようになり、GPU の計算資源を最大限活用できるようになる。

emoformer の評価にはアルゴリズム的遅延という指標が用いられている。emformer はブロック処理ベースでデコードを行う。その際の遅延はブロックの最も左のフレームと最も右のフレームで異なるものになる。そこで、センターブロック内のすべてのフレームの平均をアルゴリズム的遅延として定義している。これは、先読みコンテキスト遅延にセンターブロック遅延を 0.5 倍して足した値である。emformer は 80ms のアルゴリズム的遅延でクリーンな音声の認識した場合の WER が 0.0344 となっている。

3　音声回復手法

3.1　処理フロー

提案手法はオーディオインペインティングによる欠落回復層と音声認識層で構成する。本論文で提案する手法は、ストリーミング音声を入力とする。ストリーミング音声から波形をチャンクとして切り出す。切り出した波形はメルスペクトログラムに変換され、オーディオインペインティングで欠落部分の回復を行う。次に、欠落回復したメルスペクトログラムを音声認識層に入力し、デコーダで推論を行いテキストを出力する。チャンク内の右側コンテキストを次のチャンクに書き加え、そのチャンクに対して処理を行う。入力波形を最後まで処理し、音声認識結果を出力する。

図 2、図 3 は提案する音声回復手法の構造を示している。チャンクごとに区切られた音声は波形からメルスペクトログラ

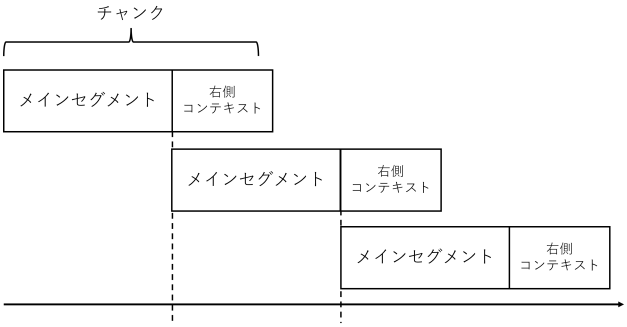

図 2　ストリーミング音声の処理フロー

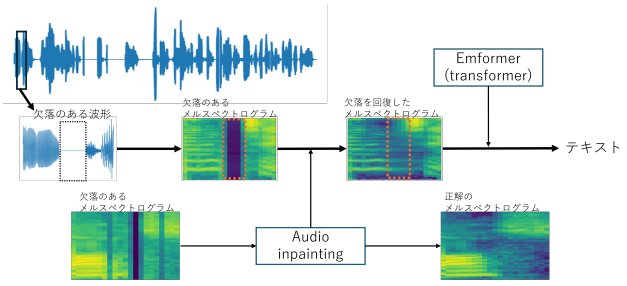

図 3　音声回復手順

ムに変換され、言語モデルを使って処理されるテキストとして出力される。

3.1.1　欠落回復層

この層では、訓練済みモデルを使用して、チャンクごとに区切られた音声の欠落回復を行う。音声の欠けている部分は音声認識では無音として処理される。欠落回復層を取り入れ、欠けていた部分の音声を回復することで、音声認識層で認識できるようになる。まず、入力されたサンプリングレート 16kHz のストリーミング音声は、メインセグメントと右側コンテキストを合わせた音声波形チャンクに区切られる。チャンクの長さは 200ms である。チャンクに区切った音声波形をメルスペクトログラムに変換する。変換時のフレームサイズは 400 点、フィルタバンク数は 80、シフト幅は 160 点である。そのメルスペクトログラムをインペインティングモデルに入力し、欠落部分の回復を行い、メルスペクトログラムを出力する。出力されたメルスペクトログラムを音声認識層に渡す。

3.1.2　音声認識層

この層では emformer[3] を使用した音声認識推論を行う。音声認識を用いることで、音声回復層で回復しきれなかった波形情報を、前後の文脈から得られる言語情報によって回復する。まず波形回復層から欠落部分を回復したメルスペクトログラムをチャンクとして受けとる。過去のチャンクの情報を記憶してあるメモリフレームをチャンクの左側に加える。デコーダにメモリフレームを加えたチャンクを入力してトークン列を推論する。最後にモデルが推論したトークン列をテキストに変換して出力する。また、チャンク内の右側コンテキストの情報は次のチャンクに入力される。

3.2　オーディオインペインティングモデル

このモデルは、200ms のメルスペクトログラムの欠落部分を修復できるように訓練を行った。図 4 はオーディオインペ

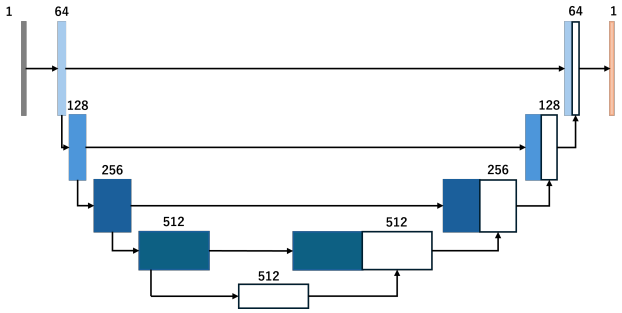

図 4　UNet のモデル構造

インティングモデルの構造を示したものである。入力として欠落のある音声のメルスペクトログラムを使用する。入力はエンコーダを通してダウンサンプリングされる。各ダウンサンプリングで畳み込み層と maxpooling 層がある。各層で次元数を増やしながら特徴マップを半分にする。ダウンサンプリングは合計で 4 回行い、最終的なボトムレイヤーは 512 チャンネルになる。次にボトムレイヤーをアップサンプリングする。このとき、エンコーダの出力をスキップ接続で結びつけることで多重解像度の情報を保持しながら学習が行われる。アップサンプリングは合計で 4 回行う、最終的には 64 チャンネルとして出力される。最後に出力層を通して 1 チャンネルになる。出力はメルスペクトログラムである。

入力に使用するメルスペクトログラムには、データセットの音声ファイルを 200ms に区切ったものを変換し使用する。200ms に足りなかった場合は 0 埋めし、200ms にする。区切り時間は音声認識層のチャンク時間に合わせて 200ms としている。変換時のフレームサイズは 25ms、フィルタバンク数は 80、ホップ長は 10ms とした。

また、学習に使用する損失関数は平均二乗誤差とした。

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

ここで n はデータ数、y_i は i 番目の正解値、$\hat{y}_i$ は i 番目の予測値である。torch で学習することで、GPU を使用し学習を行った。epoch 数は 30 でバッヂサイズは 32 である。また、学習率は 0.001 とした。

4　データセットと評価

4.1　データセット

データセットには学習用のデータとして INTERSPEECH Deep Packet Loss Concealment challenge 2022 のデータセットを使用した。このデータセットは約 10 秒の欠落のない音声と欠落のある音声、欠落位置を示したメタデータで構成される。欠落のない音声は libribox Community Podcast オーディオクリップを使用している。また、欠落のない音声は teams で実際に観測したパケットロスのトレースを欠落のない音声に重ねることで作成されている。テスト用に 4000 発話と検証用に 1000 発話を用いた。

評価用のデータセットとして LJSpeech データセットを用いる。このデータセットは 1~10 秒の英語の音声クリップと、そのテキストデータで構成される。音声データは一人の女性話者が複数のノンフィクション書籍の抜粋を朗読して作成されている。評価には 300 発話を用いた

4.2　評価方法

提案手法の評価には文字誤り率（WER）を用いる。WER とは、システムがどれだけ正確に入力された音声をテキストにできるかを評価する指標であり、

$$WER = \frac{S + D + I}{N}$$

という式で表される。S は誤置換の数、D は削除した数、I は誤挿入した数、N は正解の単語数である。欠落のある音声に提案手法で音声認識した結果、欠落のない音声を emformer で認識した結果、欠落のある音声を emformer で認識した結果の WER の比較を行う。

4.3　評価結果

表 1 は WER で欠落のない音声を emformer で認識した結果、欠落のある音声を emformer で認識した結果、提案手法で音声認識した結果を比較している。

表 1　WER での音声認識評価

	欠落のない音声	欠落のある音声	提案手法
WER	0.0775	0.167	0.142

欠落のある音声を認識した結果の WER は 0.167 であるのに対して、欠落のある音声を提案手法で認識した結果の WER は 0.142 であった。WER を比較し、音声欠落回復層があることで音声認識精度が向上することがわかる。

また、音声認識結果の例を示す。

欠落なし音声＋ emformer
it is of the first importance that the letter used should be fine in form

欠落のある音声＋ emformer
it is of the first importance that (　　) letter used should be fought and in form

欠落のある音声＋提案手法
it is of the first importance that the letter used should be fine in form

下線は音声認識誤り、カッコは音声認識されなかった単語を表す。欠落のない音声を emformer で音声認識した結果では、8 単語目に the という単語が発話されている。この発話は図 5 の左のメルスペクトログラム 2.1 秒から 2.2 秒あたりである。図 5 の中央のメルスペクトログラムを見るとその時間の音声が欠落していることが分かる。そのため、欠落のある音声を emformer で音声認識した結果では、この単語は欠落していて、

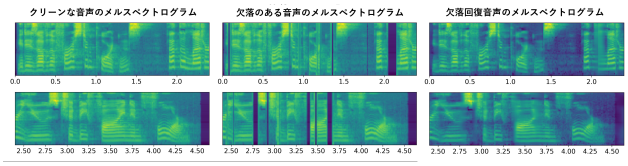

図 5　音声認識結果例

認識されていない。欠落のある音声に対して欠落回復を行うことで、図 5 の右のメルスペクトログラムのようになる。2.1 秒から 2.2 秒あたりを見ると低周波数成分の倍音構造が復元できている。そのため、提案手法では認識できた。

また、2.8 秒から 3.8 秒あたりで should be fine という単語が発話されている。図 5 の中央のメルスペクトログラムを見ると複数個所に欠落が発生していることが分かる。そのため、欠落のある音声を emformer で認識した結果は fought and という単語が認識されている。欠落回復を行うことで、それぞれの欠落部分のメルスペクトログラムの周波数成分を回復できていることから、正しい音声を認識した。

5　考察

前セクションで示した例は、オーディオインペインティングした結果正しく音声認識されたものである。オーディオインペインティングした後の認識結果が正解とことなっている場合には 2 つのパターンが存在した。欠落が埋まったが認識結果が正しくなかったパターン、オーディオインペインティングを行うことで認識結果が間違ってしまったパターンの 2 つである。これらの理由について考察を述べる。

まず、欠落が埋まったが正しく認識できなかったパターンである。図 6 は左が欠落のないメルスペクトログラム、中央が

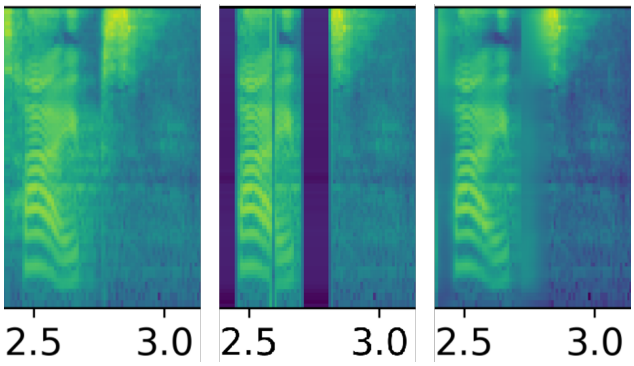

図 6　欠落が埋まったが正しく認識できなかった例

欠落のあるメルスペクトログラム、右が欠落回復処理を行ったメルスペクトログラムである。2.5 秒から 3.0 秒付近で types という単語が発話されている。中央の欠落のある音声では、pipes という単語で認識されている。欠落回復を行うことで右のメルスペクトログラムになる。元のメルスペクトログラムと比べ、高周波数成分がぼやけ、低周波数成分が増えていることがわかる。その結果、欠落が回復しても fight という誤った単語が認識された。このように、欠落部分の周波数成分が回復しきれていないことが原因で、欠落が埋まったが正しく認識できない問題が発生する。

次に、オーディオインペインティングを行うことで認識結果が間違ってしまうパターンである。図 7 は左が欠落のないメ

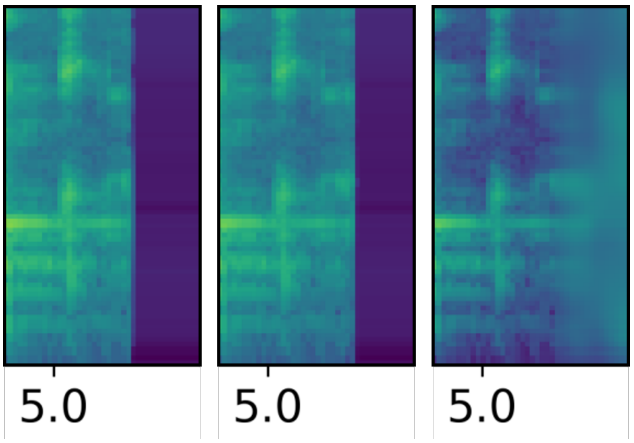

図 7　欠落回復誤りの例

ルスペクトログラム、中央が欠落のあるメルスペクトログラム、右が欠落回復処理を行ったメルスペクトログラムである。右のメルスペクトログラムの最後の部分は音声が 200ms に達していなかった際に行う 0 埋め処理を行っているため周波数成分が存在していない。右と中央のメルスペクトログラムに違いがないことから、この部分では欠落が起こってないことがわかる。しかし、欠落回復層でオーディオインペインティングを行うことで、0 埋めを行った部分に周波数成分を生成してしまう。このように、本来周波数成分がない部分に対してインペインティングモデルが周波数成分を生成してしまうため、オーディオインペインティングが原因の認識誤りが発生する。

6　おわりに

本論文では、欠落のあるストリーミング音声に対して、波形の欠落回復と音声認識を行い WER を測定した。その結果、波形の欠落回復を行わなかった音声認識の WER が 0.167 であるのに対し、波形の回復を行った音声認識の結果の WER は 0.142 であった。この結果から音声情報と言語情報の 2 つの情報を用いることが、欠落回復に有用であることがわかった。本論文で実装したのは、欠落のある波形について、音声情報を用いて波形の欠落部分を回復し、言語情報を用いて言語情報で回復できなかった部分をさらに回復するところまでである。本論文の後段処理として、波形情報を維持したまま再度波形に変換する必要がある。また、本論文では提案システムのチャンクを 200ms とした。このチャンクはストリーミング音声に対してのシステムの遅延時間である。ストリーミング音声の処理を行う上で、音声認識の精度を下げずに遅延時間を減らすことが今後の課題である。

参考文献

[1] Microsoft Corporation. Microsoft Teams の通話フロー, 2023. [アクセス日: 2024-12-24].
[2] C. Perkins, O. Hodson, and V. Hardman. A survey

of packet loss recovery techniques for streaming audio. *IEEE Network*, 12(5):40–48, 1998.

[3] Y. Shi, Y. Wang, C. Wu, C.-F. Yeh, J. Chan, F. Zhang, D. Le, and M. Seltzer. Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition, 2020.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.

[5] J. Wang, Y. Guan, C. Zheng, R. Peng, and X. Li. A temporal-spectral generative adversarial network based end-to-end packet loss concealment for wideband speech transmission. *J. Acoust. Soc. Am.*, 150(4):2577, Oct. 2021.

[6] N. L. Westhausen and B. T. Meyer. tPLCnet: Real-time Deep Packet Loss Concealment in the Time Domain Using a Short Temporal Context. In *Proc. Interspeech 2022*, pages 2903–2907, 2022.

[7] C. Wu, Y. Wang, Y. Shi, C.-F. Yeh, and F. Zhang. Streaming transformer-based acoustic models using self-attention with augmented memory, 2020.

[8] Zoom Video Commuunications, Inc. クライアント接続プロセス, 2024. [アクセス日: 2024-12-24].