2024年度後期　修士論文

深層学習による口唇音声変換に関する研究

A Study on the Conversion from Lip-Video to Speech Using a Deep Learning Technique

2025年月日

九州大学芸術工学府音響設計コース

2DS23095M

南 汰翼

MINAMI Taisuke

研究指導教員　鏑木 時彦　教授

概要

目 次

1　序論

1.1　背景

音声は基本的なコミュニケーションの手段であり、人と人とのコミュニケーションの場面において、重要な役割を果たしている。音声は、肺からの呼気流による声帯の振動が音源波を生成し、声道特性に伴ったフィルタリングと口唇からの放射特性に従って生成される。これより、音声の生成には音源を作り出す声帯やその制御のための喉頭、舌や口唇といった調音器官の働きが重要となる。しかし、癌などの重い病気で喉頭を摘出した場合、音源波を生成することができなくなるため、これまで通り発声を行うことが不可能になってしまう。このようなコミュニケーション機能の喪失に対し、現在でも電気式人工喉頭や食道発声、シャント発声といった代用音声手法が存在する。電気式人工喉頭では、専用の発振器を顎下に当てて振動を加えることにより、それを音源とした発声を行う。発振器を用意すれば容易に発声することが可能であるが、生成される音声のピッチが発振器の振動に依存してしまうため、抑揚のない単調な音声になってしまう。食道発声では、まず口や鼻から食道内に空気を取り込み、その空気を逆流させることで食道入口部の粘膜を振動させることによって発声する。電気式人工喉頭と違って道具を必要とせず、ピッチも本人が調節できるが、その習得に長期間の訓練を要する。シャント発声では、手術によって気管と食道を繋ぐ管を設ける。これにより、息を吐き出す際に設けられた喉の穴を手で塞ぐことによって肺からの呼気流が食道に流れる。そのため、食道発声と同様に粘膜の振動を音源とし、発声することが可能となる。習得は容易であり、比較的自然に話すことが可能となるが、設けられた管を交換するための定期的な手術が必要となる。このように、現在用いられている代用音声手法にはそれぞれデメリットが存在する。

そのため、本研究ではビデオカメラで撮影した口唇の動きから音声合成を行うことによる、新たな代用音声手法を検討する。本来、音声は声帯の振動や声道の形状に依存して生成されるものであり、口唇の動きのみから音声波形を直接推定することは困難である。そこで、近年画像や自然言語処理、音声といった分野において成果を上げている深層学習を使用し、データ駆動型の方法で口唇の動きと音声の間の関係を学習することで推定を行う。これにより、従来の代用音声手法よりも自然性の高い音声を、訓練や定期的な手術の必要なく提供することを目指す。

これまでの動画音声合成は英語が中心に検討が進んでおり、近年では YouTube 上のデータを収集、処理することによって構築した大規模データセット [1, 2] を用いることで、大規模で表現力の高いモデルが構築可能となっている。これにより、従来行われてきた教師あり学習のみならず、動画と音声の関係性を自己教師あり学習（Self Supervised Learning; SSL）によって学習し、そのモデルを Visual Speech Recognition（VSR）や動画音声合成に FineTuning するアプローチが提案され、その有効性が示されている。自己教師あり学習モデルにもいくつかの種類があり、近年多くの研究で応用例のある AVHuBERT [3] は、動画・音声の入力領域においてマスクされた区間の予測と、予測対象の更新を繰り返して学習を進めていくモデルである。予測対象の更新は 5 回行われ、1 回目は音声波形から計算される MFCC をクラスタリングした結果を利用するが、2 回目以降はモデルの中間特徴量をクラスタリングした結果を新たな予測対象

に設定する。更新のたびに再度モデルをランダム初期化して再学習するが、その予測対象の複雑さが増していくことによって学習を促進するようなメカニズムとなっている。また、これに類似したVATLM [4] は、動画と音声のみならずテキストも加えた学習によって、精度改善を達成した。その他、Studentと Teacherという二つのネットワークを利用し、Teacherから出力される特徴量をStudentがマスクされた入力から予測することによって学習を進めるRAVEn [5] やAV-data2vec [6]、RAVEnの改善版として提案されたBRAVEn [7] など、多くのモデルが提案され続けている。

近年の動画音声合成やVSRでは、こういったSSLモデルを動画からの特徴抽出器として活用しつつ、さらなる工夫によって精度改善を達成している。動画音声合成について、 [8] ではメルスペクトログラムに加えてテキストを予測対象としたマルチタスク学習の有効性が示され、その後音声SSLモデルであるHuBERT[9] から得られた離散特徴量を用いる手法も提案された [10]。音声SSL離散特徴量を音声波形までの中間特徴量として扱う手法は他にも提案されており、動画から離散特徴量のみを推定して音声波形に変換する手法 [11] や、AVHuBERTから得られる動画音声SSL離散特徴量から音声SSL離散特徴量を推定するネットワークを導入した手法 [12]、モデルの内部で予測した離散特徴量を再度加算した上で、その後のメルスペクトログラムの予測に利用する手法 [13] などが提案されている。

一方VSRについて、 [14] では音声認識を利用して言語情報を格納したメモリを用意し、メモリと動画特徴量の間でアテンションをとることによって、ネットワーク内部で言語情報との関連を考慮する構成を提案した。また、 [15] ではAVHuBERTが動画あるいは音声のどちらを入力とした場合でもクロスモーダルな特徴量を返すことに着目し、音声認識デコーダに組み合わせるAVHuBERTのFew-shot Learning、Zero-shot Learningによる転移学習を検討した。加えて、同様に音声認識デコーダを転移学習するアプローチであるが、事前学習済みモデルの重みを固定し、動画特徴量から音声認識モデルの中間特徴量を予測するネットワークのみを新たに学習することで、両者を合併するようなアプローチ [16] も提案されている。さらに、静止画像と音声から動画を合成するネットワークを構築し、音声認識用のデータセットを用いてVSRの学習データを大量に合成するデータ拡張手法 [17] や、事前学習済みの音声認識モデルによって教師なしデータにラベリングを行うデータ拡張手法 [18]、10万時間分の教師ありデータを新たに増強した研究 [19] など、大規模な学習データを確保することで精度改善を達成した例も報告されている。

上記の研究は英語データを用いたものであったが、VSRにおいては英語以外の言語に焦点を当てた研究や、多言語対応モデルの構築も検討されている。 [20] ではRAVEnを利用し、英語に加えてスペイン語、イタリア語、ポルトガル語など計6種類の言語が含まれるデータセット [21, 22, 23] を用いて多言語モデルの構築を検討した。結果として、教師ありデータの少ない英語以外の言語に対する、多言語モデルの有効性が明らかとなった。また、 [24] では英語データで学習されたAVHuBERTを用いつつ、特定の言語ごとに構築した音声認識モデルのデコーダを転移学習することで、特定言語ごとにモデルを構築するアプローチを提案した。さらに、[25] では音声認識モデルであるWhisperを利用し、教師なしデータへのラベリングによるデー

タ拡張を行うことで、上記二つのアプローチを超える精度を達成した。

本研究では、世界的に見て日本語での動画音声合成の検討例が少ないこともあり、文献 [26, 27] で収録された日本語データを用いて研究を行うこととした。日本語データは英語ほど大規模なデータが存在しないが、予備実験として AVHuBERT の FineTuning を検討したところ、スクラッチで構築したモデルと比較して、より高い精度を示すことが明らかとなった。これは、多言語対応を検討した先行研究の傾向にも一致する結果であり、日本語においても有効なモデルだと考えられる。しかしながら、それでも依然として合成音声の品質は低く、自然音声に迫る合成音は実現されていないことが課題である。

1.2 目的

本研究の目的は、動画音声合成によって得られる合成音声の品質を向上させることである。近年高い精度を達成した手法 [10] では、AVHuBERT の利用および、メルスペクトログラムと音声 SSL 離散特徴量を利用したマルチタスク学習が採用されている。その他にも近年高い精度を達成したモデルは存在 [11, 12, 13] するが、手法 [10] が採用しているマルチタスク学習の有効性は、テキストを用いた先行研究 [8] でも同様に示されている。これより、このアプローチが現状特に有効そうだと判断し、この手法をもとにさらなる改善を狙う形で研究を進めることとした。この手法では、動画を入力としてメルスペクトログラムと音声 SSL 離散特徴量を推定し、これら両方を Multi-input Vocoder に入力することで音声波形へと変換する。しかし、動画と音声の間には、同様の口の動きであっても声道形状の違いによって生じる発話内容の曖昧さや、話者によるパターンの多様さが存在すると考え、推定を動画のみに依存した先行研究の手法ではこういった側面への対処が難しいと考えた。これに対して本研究では、音声 SSL モデルである HuBERT を利用した動画音声合成モデルを提案し、合成音声の推定残差を HuBERT を利用した後処理によって軽減することで、合成音声の品質改善を狙った。HuBERT は、音声波形を畳み込み層を通すことによってダウンサンプリングしつつ特徴量に変換し、ここでマスクをかけた上で Transformer 層を通す。そして、マスクされたフレームにおける予測対象を推定する、Masked Prediction を行うことで学習する。大規模な音声データを用いてこの自己教師あり学習を行うことで、音声のコンテキスト自体をデータそのものから学習することが可能であり、音声認識において有効性が確認されている。本研究では、大規模日本語音声データで事前学習済みの HuBERT を活用し、動画音声合成モデルにおいて生じる推定残差を、音声自体のコンテキストを考慮する形で補うようなアプローチを検討した。

1.3 本論文の構成

2　音声信号処理

3　深層学習

4　音声 SSL モデルを活用した動画音声合成の検討について

4.1　音声合成法

提案手法の構築手順は 3 段階に分かれる。ネットワークの概要を図に示す。一段階目では、動画を入力として、メルスペクトログラムと HuBERT 離散特徴量に加え、HuBERT 中間特徴量を推定するネットワークを学習する。ここで、HuBERT 離散特徴量は HuBERT Transformer 層から得られる特徴量をクラスタリングによって離散化した結果、HuBERT 中間特徴量は HuBERT における畳み込み層出力のことを指す。第一段階では、AVHuBERT を動画からの特徴抽出に利用した。これにより、動画の空間情報は完全に圧縮され、768 次元の一次元系列となる。その後、事前学習済みの話者認識モデル [28] によって音声波形から得られる 256 次元の話者 Embedding を、各フレームでチャンネル方向に結合する。これによって特徴量は 1024 次元に拡張され、全結合層によって再度 768 次元に圧縮する。その後、畳み込み層と全結合層からなるデコーダを通すことによって、話者 Embedding を結合した特徴量に対する変換を施した。これにより、特にメルスペクトログラムにおいて話者性が正しく反映されることを狙った。動画の見た目から話者性を判断できる可能性も考えられたが、本研究では念の為入力することとした。デコーダは残差結合を利用したブロック単位で構成され、各ブロックに 2 層の畳み込み層を設けた。各畳み込み層のチャンネル数は 768、カーネルサイズは 3 であり、3 ブロック積み重ねた。最後に、全ブロックを通した出力を全結合層を通すことによって、最終的な予測対象を得た。

二段階目では、一段階目に学習されたネットワークの重みを固定した状態で HuBERT 中間特徴量を推定し、それを入力として HuBERT Transformer 層の学習を行った。ここでは、メルスペクトログラムと HuBERT 離散特徴量を推定対象とする。HuBERT Transformer 層出力に対し、一段階目と同様に話者 Embedding を結合し、デコーダを通すことで、予測値を得た。HuBERT の自己教師あり学習においてマスクが適用されるのは畳み込み層出力であるため、その後の Transformer 層を本研究のタスクに Fine Tuning することにより、AVHuBERT の予測結果における推定残差の軽減を狙った。また、一段階目のネットワークの予測値を入力として再び予測を行うという構造は、アンサンブル手法におけるスタッキングに相当し、この意味での汎化性能向上による予測精度の改善にも期待した。実験においては HuBERT の事前学習済み重みを読み込む場合と、ランダム初期化する場合の両条件を検討し、事前学習済みモデルである HuBERT の転移学習における有効性と、アンサンブルモデルとすることの有効性の両面を評価できるようにした。

三段階目では、二段階目までに学習されたネットワークの重みを固定した状態で、AVHuBERT の最終出力特徴量と、HuBERT Transformer 層から得られる最終出力特徴量の二つを結合し、それらを入力として再びメルスペクトログラムと HuBERT 離散特徴量の予測を行うネットワークを学習した。ここでは、最終出力特徴量はどちらも予測対象に関連したものとなっており、これを元にした予測それぞれが推定残差を持つことを考慮して、両方を利用した予測を改めて行うことで単一特徴量への依存を解消し、汎化性能の向上を狙った。はじめに二つの最終出力特徴量をチャンネル方向に結合することで、1536 次元の特徴量を獲得し、これに対して全結合

表 4.1: 利用したデータセットの文章数

	学習	検証	テスト
動画音声データセット Hi-Fi-Captain JVS	1600	200	212

層を施すことで再度 768 次元に圧縮した。その後、4 層の Transformer 層を通すことで系列全体を考慮した特徴抽出を行い、一段階目および二段階目と同一のデコーダによって予測値を得た。ここで、各 Transformer 層のヘッド数は 12 とした。

以上のモデルにより、動画からメルスペクトログラムと HuBERT 離散特徴量が推定可能となる。その後、先行研究 [10] に基づく Multi-input Vocoder を用い、メルスペクトログラムと HuBERT 離散特徴量を入力として音声波形に変換することで、最終的な合成音声を得た。Multi-input Vocoder は HiFi-GAN [29] をベースとしたモデルであり、HuBERT 離散特徴量についてはインデックスからベクトルへの変換を行なっている。実験では各段階における予測値を入力として音声波形を合成することで、その比較を行った。

最後に、今回ベースラインとする先行研究に基づいたマルチタスク学習手法は、本研究における第一段階で、HuBERT 中間特徴量を推定しないものに当たる。

4.2 実験方法

4.2.1 利用したデータセット

動画音声データセットには、男女二人ずつから収録した合計 4 人分のデータセット [26, 27] を用いた。これは ATR 音素バランス文 [30] から構成され、全話者共通で A から H セットを学習データ、I セットを検証データ、J セットをテストデータとして利用した。各分割ごとの文章数を表 4.1 に示す。

Multi-input Vocoder の学習に利用する音声データセットには、Hi-Fi-Captain（日本語話者二名分） [31] と JVS（parallel100 と nonpara30） [32] を利用した。ボコーダの学習時にはサンプル全体から 1 秒分をランダムにサンプリングして用いるが、元データには話し声のない無音区間が一定存在しており、これはボコーダの学習に望ましくない。これに対して、無音区間のトリミング（-40 dBFS 未満かつ 500 ms 継続する区間を 100 ms までカット）を適用した。Hi-Fi-Captain は train-parallel および train-non-parallel を学習データ、val を検証データ、eval をテストデータとして分割した。各分割ごとの文章数を表 4.1 に示す。JVS には話者に対して 1 から 100 まで番号が割り振られており、本実験では 1 から 80 番の話者を学習データ、81 番から 90 番の話者を検証データ、91 番から 100 番までの話者をテストデータとした。各分割ごとの文章数を表 4.1 に示す。

4.2.2 データの前処理

動画データは60 FPSで収録されたものをffmpegにより25 FPSに変換して用いた。その後、手法 [33] により動画に対してランドマーク検出を適用した。このランドマークを利用することで口元のみを切り取り、画像サイズを（96, 96）にリサイズした上で、グレースケールに変換した。加えて、画像に対する正規化および標準化を適用した。全体として、今回は事前学習済みのAVHuBERTの転移学習を行うため、そこでの前処理に合わせている。学習時は、ランダムクロップ、左右反転、Time Masking（一時停止）をデータ拡張として適用した。ランダムクロップは、（96, 96）で与えられる画像から（88, 88）をランダムに切り取る処理である。検証およびテスト時は、必ず画像中央を切り取るよう実装した。左右反転はランダムクロップ後に適用しており、50%の確率で左右が反転されるよう実装した。Time Maskingは、連続する画像の時間平均値を利用することによって、一時停止させるような効果を与えるデータ拡張手法である。動画1秒あたり0から0.5秒の間でランダムに停止区間を定め、その区間における動画の時間方向平均値を計算し、区間内のすべてのフレームをこの平均値で置換した。

音声データは48 kHzで収録されたものを16 kHzにダウンサンプリングして用いた。それから、窓長25 msのハニング窓を用いて、シフト幅10 msでSTFTを適用することでフレームレート100 Hzのスペクトログラムに変換した。さらに、振幅スペクトログラムに対して80次のメルフィルタバンクを適用し、80次のメルスペクトログラムを得た上で対数スケールに変換した。話者Embeddingの取得には事前学習済みモデル [28] を利用し、学習データから100個ランダムサンプリングして計算した平均値を利用した。

HuBERTは、HuggingFaceに公開されているReazonSpeechというデータセットによって学習されたモデル [34, 35] を利用した。ReazonSpeechは約19000時間の日本語音声からなるデータセットであり、日本語音声のコンテキストを大量のデータから学習したモデルになっていることが予想される。本研究のアプローチでは、日本語音声に関する事前知識を豊富に有するモデルが適していると考え、このモデルを選択した。動画からの予測対象となるHuBERT離散特徴量は、k-means法によるクラスタリング（クラスタ数100）をHuBERT Transformer層の8層目出力に適用することで得た。8層目を選択した理由は、HuBERTのレイヤーごとの特徴量について、音素のOne-hotベクトルおよび単語のOne-hotベクトルとの相関を、Canonical Correlation Analysis（CCA）によって調べた先行研究 [36] より、8層目出力がそのどちらとも相関が高く言語的な情報に近いと判断したからである。k-means法の学習には動画音声データセットにおける学習用データを利用し、これによって動画音声データおよび外部の音声データについてクラスタリングを適用した。

4.2.3 学習方法

一段階目について、損失関数はメルスペクトログラムのMAE Loss L_{mel} とHuBERT離散特徴量のCross Entropy Loss L_{ssl^d} 、HuBERT中間特徴量のMAE Loss L_{ssl^i} の重み付け和とし

た。それぞれの重み係数を $\lambda_{mel}, \lambda_{ssl^d}, \lambda_{ssl^i}$ とすると、

$$L = \lambda_{mel} * L_{mel} + \lambda_{ssl^d} * L_{ssl^d} + \lambda_{ssl^i} * L_{ssl^i} \tag{4.1}$$

となる。第一段階では $\lambda_{mel} = 1.0, \lambda_{ssl^i} = 1.0$ と固定し、λ_{ssl^d} のみ 0.0001 から 1.0 まで 10 倍刻みで 5 段階試してチューニングを行った。最適化手法には AdamW [37] を利用し、$\beta_1 = 0.9$、$\beta_2 = 0.98$、$\lambda = 0.01$ とした。学習率は 1.0×10^{-3} から開始し、Warmup Scheduler によってその値を変化させた。バッチサイズはメモリの都合上 4 としたが、学習の安定のため Gradient Accumulation によって各イテレーションにおける勾配を累積させ、8 イテレーションに一回重みを更新するようにした。そのため、実質的にはバッチサイズ 32 となる。モデルに入力する動画の秒数は 10 秒を上限とし、それを超える場合はランダムにトリミング、それに満たない場合はゼロパディングした。勾配のノルムは 3.0 を上限としてクリッピングすることで、過度に大きくなることを防止した。最大エポック数は 50 とし、10 エポック連続して検証データに対する損失が小さくならない場合には、学習を中断するようにした。また、学習終了時には検証データに対する損失が最も小さかったエポックにおけるチェックポイントを保存し、これをテストデータに対する評価に用いた。

第二段階について、損失関数はメルスペクトログラムの MAE Loss と HuBERT 離散特徴量の Cross Entropy Loss の重み付け和とした。これは式 (4.1) において、$\lambda_{mel} = 1.0, \lambda_{ssl^i} = 0.0$ と固定した場合に相当する。第一段階と同様に、λ_{ssl^d} のみ 0.0001 から 1.0 まで 10 倍刻みで 5 段階試してチューニングを行った。最適化手法には AdamW を利用し、$\beta_1 = 0.9$、$\beta_2 = 0.98$、$\lambda = 0.01$ とした。学習率は 5.0×10^{-4} から開始し、Warmup Scheduler によってその値を変化させた。学習率について第一段階と異なるが、これは値を半減させることによって学習を安定させることができたためである。その他のパラメータは、第一段階における値と同じである。

第三段階について、学習率のみ第一段階と同様に 1.0×10^{-3} としたが、それ以外は第二段階と全く同じである。

Multi-input Vocoder の学習について、学習は動画音声データセットではなく、外部データである Hi-Fi-Captain と JVS を用いた。はじめに Hi-Fi-Captain のみを用いて学習させ、その後学習済みモデルを JVS によって再学習した。Hi-Fi-Captain は男女一人ずつの文章数が豊富なデータセットであるため、高品質なモデルを構築可能であった。しかし、学習できる話者数が少ない分、学習外話者に対する合成音声の品質が低かった。そのため、一人当たりの文章数は 100 文章程度と少ないながらも、100 人分の話者からなる JVS を利用して再学習することによって、学習外話者に対する合成音声の品質を向上させた。最適化手法には AdamW を利用し、$\beta_1 = 0.8$、$\beta_2 = 0.99$、$\lambda = 0.01$ とした。学習率は 2.0×10^{-4} から開始し、指数関数的にその値を変化させた。バッチサイズは 16 とし、ここでは Gradient Accumulation は利用しなかった。モデルへの入力は 1 秒を上限とし、それを超える場合はランダムにトリミング、それに満たない場合はゼロパディングした。勾配のノルムは 3.0 を上限としてクリッピングすることで、過度に大きくなることを防止した。最大エポック数は 30 とし、ここでは Early Stopping は適用しなかった。また、学習終了時には検証データに対する損失（メルスペクトログラムに対する

L1 Loss）が最も小さかったエポックにおけるチェックポイントを保存し、これをテストデータに対する評価に用いた。

GPU は NVIDIA RTX A4000 を利用し、計算の高速化のため Automatic Mixed Precision を適用した。

4.2.4 比較手法

比較手法は、以下の五つである。

1. AVHuBERT によるメルスペクトログラムと HuBERT 離散特徴量のマルチタスク学習手法
2. 提案手法第二段階で、事前学習されていない Transformer 層を用いる手法
3. 提案手法第三段階で、事前学習されていない Transformer 層を用いる手法
4. 提案手法第二段階で、事前学習済みの Transformer 層（HuBERT Transformer 層）を用いる手法
5. 提案手法第三段階で、事前学習済みの Transformer 層（HuBERT Transformer 層）を用いる手法

手法 1 が先行研究において有効性が確認された手法であり、今回の実験においてベースラインとなる。これに対する改善案として、手法 2 から手法 5 が提案手法である。手法 2 および手法 3 と、手法 4 および手法 5 を比較することで、提案手法において事前学習済みモデルを用いることに意味があるかを調べられるようにした。

4.2.5 客観評価

合成音声の客観評価には、二種類の指標を用いた。一つ目は、音声認識の結果から算出した Word Error Rate（WER）である。Whisper [38] を用いて音声認識を行い、出力されるテキストに対して MeCab を用いて分かち書きを行った上で uiwer というライブラリを用いて算出した。値は 0 から 100 であり、この値が低いほど音声認識の誤りが少ないため、より聞き取りやすい音声であると判断できる。二つ目は、話者 Embedding から計算したコサイン類似度である。話者 Embedding の計算は、モデルへの入力値を計算したものと同様の話者認識モデル [28] を利用し、対象音声と原音声のペアでコサイン類似度を計算した。今回構築するモデルは 4 人の話者に対応するモデルとなるため、原音声に似た声質の合成音声が得られているかをこの指標で評価した。値はから 1 であり、高いほど原音声と類似した合成音声だと判断できる。

4.2.6　主観評価

4.3　結果

4.3.1　客観評価

まず、損失関数 (4.1) の重み係数 λ_{ssl^d} を変化させた時の客観評価指標の結果を表 4.2 に示す。各手法ごとに 0.0001 から 1.0 まで 10 倍刻みで 5 段階検討し、各手法の客観指標ごとに最も優れた値を下線で示している。また、これ以降の比較のために、最適なチューニングだと考えられる λ_{ssl^d} の値を選択しており、それを選択フラグの列で示している。True が入っている行が、最適値として選択されたことを意味する。

手法 1 では、λ_{ssl^d} の値が 0.001 および 0.01 の場合が最も WER が低く、さらに 0.01 の場合に話者類似度が最も高いため、0.01 が最適であると判断した。また、0.01 よりも重み係数の値を大きくすると、WER と話者類似度がともに悪化する傾向が見られた。手法 1 における検証データに対する損失曲線を図 4.1 に示す。横軸がエポック数、縦軸が損失の値を表す。損失の値は各エポックにおける平均値である。判例は λ_{ssl^d} の値であり、これに応じて線の色を分けている。また、丸いマーカーはテストデータの合成に利用したチェックポイントのエポック数に対応させてプロットしている（チェックポイントの選択においてはエポックごとの平均値ではなく、イテレーションごとの損失を用いているため正確な損失の値は異なる）。図 4.1b を見ると、λ_{ssl^d} の値を大きくするに従って、学習初期における L_{ssl^d} の下がり方が急峻になり、その後増加傾向に転じていることが分かる。これに対応して、λ_{ssl^d} が大きい場合は L において L_{ssl^d} が支配的となるため、結果として検証データに対する L の値も早いうちから増加傾向に転じ、L_{mel} の値が低下しきらないまま学習が中断されていることが分かる（λ_{ssl^d} が小さい場合、その後 L_{mel} がさらに小さくなる傾向にあるため）。客観評価指標をもとに選択した最適な λ_{ssl^d} の値は 0.01 であるが、この場合は L_{ssl^d} の値が安定して減少傾向にあり、結果として L_{mel} および L_{ssl^d} の両方がバランスよく下がるような損失曲線となっていることが分かる。しかし、L_{ssl^d} の値について、λ_{ssl^d} が 1.0 の場合に得られる最小値までは下がり切っていないことから、改善の余地が示唆された。

手法 2 では、λ_{ssl^d} の値が 0.1 の場合に最も WER が低く、0.01 の場合に最も話者類似度が高くなった。今回は特に、λ_{ssl^d} の値が 0.1 の場合にベースラインで選択された最適なケースと比較して WER が 9.7%低下しており、話者類似度についても等しく 0.845 となっていることから、0.1 が最適だと判断した。また、λ_{ssl^d} の値を大きくするに従って WER が下がっていき、1.0 まで上げると話者類似度が顕著に低下する傾向が確認された。手法 3 から手法 5 について、最適値の選択理由は手法 2 と同様である。また、λ_{ssl^d} と評価指標の関係についても、同様の傾向が得られた。手法 2 から手法 5 について、検証データに対する損失曲線を図 4.2 から図 4.5 にそれぞれ示す。これらを見ると、手法 1 と同様に λ_{ssl^d} の値を大きくするに従って、学習初期における L_{ssl^d} の下がり方が急峻になり、その後増加傾向に転じていることが分かる。また、最適とした λ_{ssl^d} が 0.1 のケースでは、マーカーで示すテストデータの合成に使用したチェックポイントの存在するエポック以降、L_{mel} が横ばいになっていることが分かる。これより、手法 1 と同

様に L_{mel} および L_{ssl^d} の両方がバランスよく下がるような λ_{ssl^d} であるとき、客観評価指標において最良の値が得られると考えられる。加えて、L_{ssl^d} の値について、λ_{ssl^d} が 1.0 の場合に得られる最小値までは下がり切っていないことから、手法 1 と同様に改善の余地が示唆された。

次に、最適なチューニングをした場合における手法ごとの比較を行った結果を表 4.3 に示す。分析合成は、原音声から計算した特徴量を入力として、Multi-input Vocoder で逆変換した合成音声である。手法 1 から手法 5 については、表 4.2 において選択フラグを True としたものを掲載している。また、分析合成、原音声を除いた合成音声の中で、最も優れた値を下線で示している。これより、WER、話者類似度ともに手法 3 が最も優れていることが分かる。特に、今回のベースラインである手法 1 と比較すると、同程度の話者類似度を達成しつつ WER が 9.7% 低下していることから、話者性を同程度の品質で反映しつつ、より聞き取りやすい音声が合成できたと考えられる。また、手法 2 と手法 3 を比較すると、AVHuBERT の最終出力特徴量と、HuBERT Transformer 層から得られる最終出力特徴量の両方を利用したアンサンブル手法により、WER が 0.4%低下し、話者類似度が 0.003 増加していることから、わずかに改善していることが分かる。一方、事前学習済み重みを用いることの効果について、手法 2 と手法 4 を比較すると、手法 2 の方が WER が 0.9%、話者類似度が 0.12 高いことが分かる。これより、本研究では HuBERT の事前学習済み重みを初期値とした転移学習が有効である可能性を期待したが、この仮説は外れていたと言える。事前学習重みを初期値とした場合の方が、そうでない場合と比較して悪い局所解に収束したのだと考えられる。また、手法 4 と手法 5 を比較すると、手法 4 の方が WER が 0.2%、話者類似度が 0.11 高いことから、アンサンブル手法が改善につながっていないことが分かる。これより、提案したアンサンブル手法は必ずしも改善につながるわけではないと言える。

4.3.2 主観評価

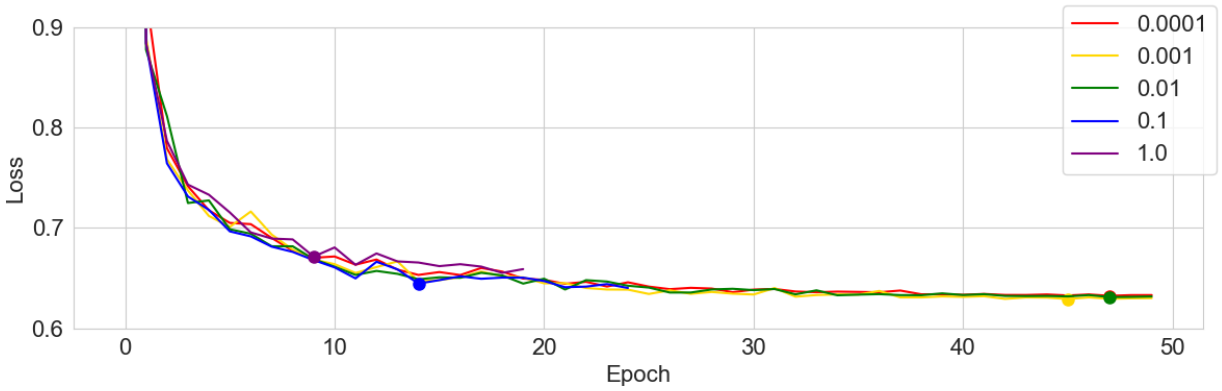
4.4 まとめ

表 4.2: 損失関数の重み係数による客観評価指標の比較

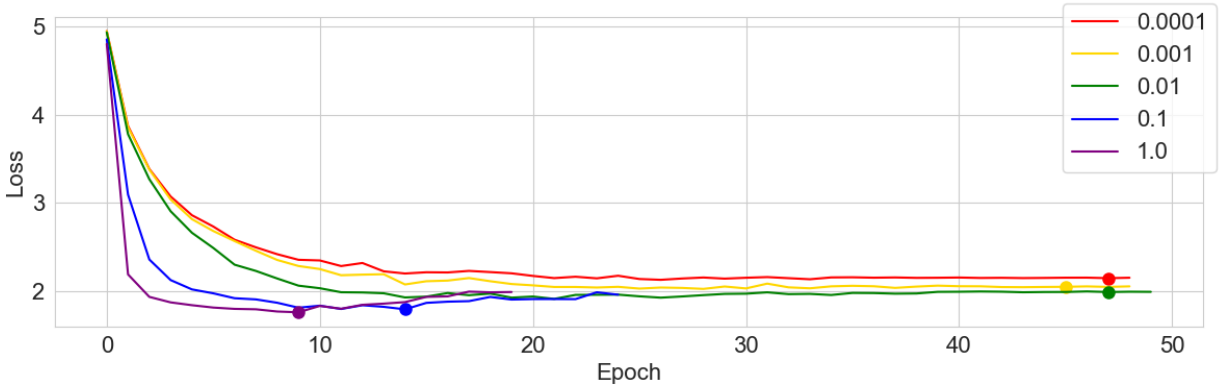
手法	詳細	λ_{ssl^d}	WER [%]	話者類似度	選択フラグ	L_{mel}	L_{ssl^d}	エポック数
1	ベースライン	0.0001	56.3	0.844		0.632	2.147	47
1	ベースライン	0.001	55.4	0.844		0.629	2.049	45
1	ベースライン	0.01	55.4	0.845	True	0.631	1.990	47
1	ベースライン	0.1	59.4	0.783		0.645	1.793	14
1	ベースライン	1.0	61.9	0.717		0.672	1.760	9
2	事前学習なし Transformer	0.0001	59.1	0.858		0.630	2.521	46
2	事前学習なし Transformer	0.001	57.9	0.855		0.631	2.382	45
2	事前学習なし Transformer	0.01	58.5	0.858		0.631	2.104	48
2	事前学習なし Transformer	0.1	46.1	0.845	True	0.620	1.942	32
2	事前学習なし Transformer	1.0	50.4	0.719		0.651	1.713	9
3	事前学習なし Transformer・アンサンブル	0.0001	59.4	0.856		0.632	2.584	22
3	事前学習なし Transformer・アンサンブル	0.001	57.4	0.852		0.633	2.448	22
3	事前学習なし Transformer・アンサンブル	0.01	56.4	0.862		0.630	2.077	45
3	事前学習なし Transformer・アンサンブル	0.1	45.7	0.848	True	0.622	2.041	18
3	事前学習なし Transformer・アンサンブル	1.0	47.1	0.770		0.634	1.735	14
4	事前学習あり Transformer	0.0001	58.9	0.849		0.632	2.593	20
4	事前学習あり Transformer	0.001	57.5	0.846		0.633	2.496	24
4	事前学習あり Transformer	0.01	58.0	0.854		0.630	2.106	30
4	事前学習あり Transformer	0.1	47.0	0.833	True	0.622	1.929	17
4	事前学習あり Transformer	1.0	48.9	0.738		0.648	1.746	10
5	事前学習あり Transformer・アンサンブル	0.0001	57.6	0.851		0.632	2.547	22
5	事前学習あり Transformer・アンサンブル	0.001	58.1	0.851		0.635	2.449	27
5	事前学習あり Transformer・アンサンブル	0.01	58.2	0.861		0.632	2.100	44
5	事前学習あり Transformer・アンサンブル	0.1	46.8	0.822	True	0.633	1.981	6
5	事前学習あり Transformer・アンサンブル	1.0	49.5	0.767		0.637	1.786	16

表 4.3: 最適なチューニングをした場合における手法ごとの比較

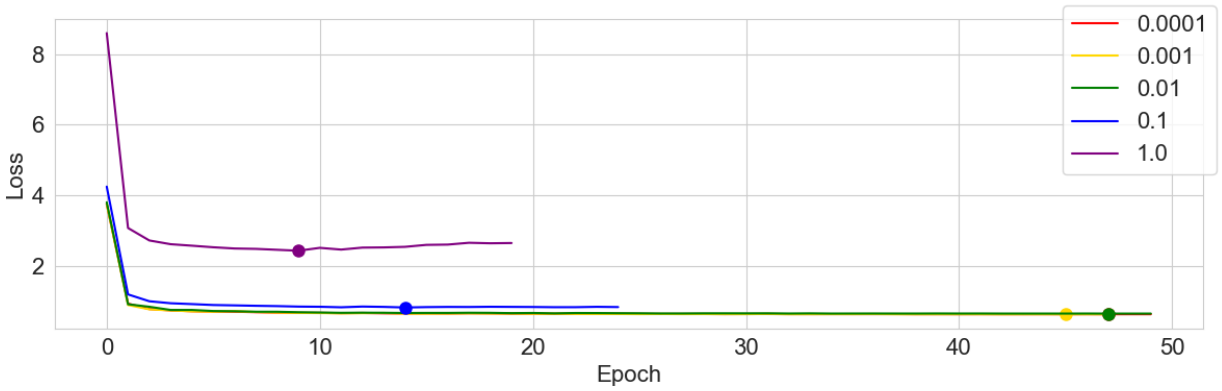
手法	詳細	WER [%]	話者類似度
1	ベースライン	55.4	0.845
2	事前学習なし Transformer	46.1	0.845
3	事前学習なし Transformer・アンサンブル	45.7	0.848
4	事前学習あり Transformer	47.0	0.833
5	事前学習あり Transformer・アンサンブル	46.8	0.822
6	分析合成	4.7	0.926
7	原音声	4.5	1.000



(a) メルスペクトログラムの MAE Loss（式 (4.1) の L_{mel}）

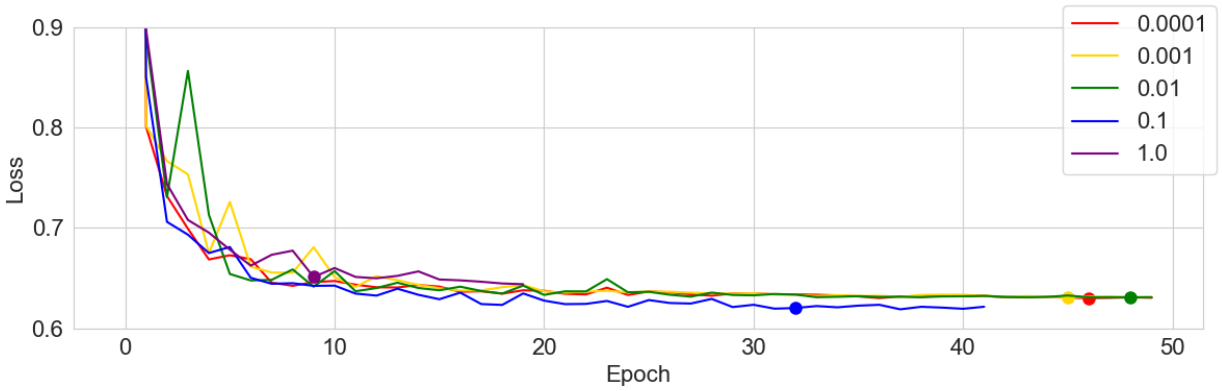


(b) HuBERT 離散特徴量の Cross Entropy Loss（式 (4.1) の L_{ssl^d}）

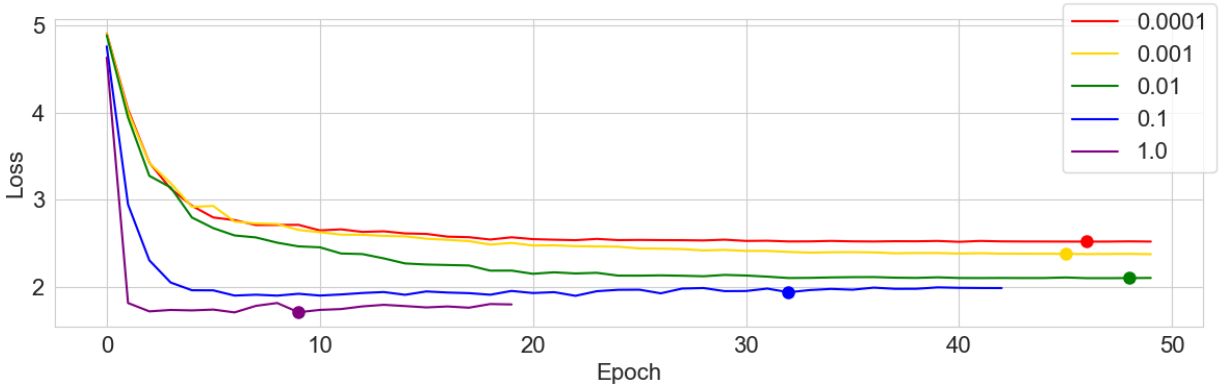


(c) 損失の合計値（式 (4.1) の L）

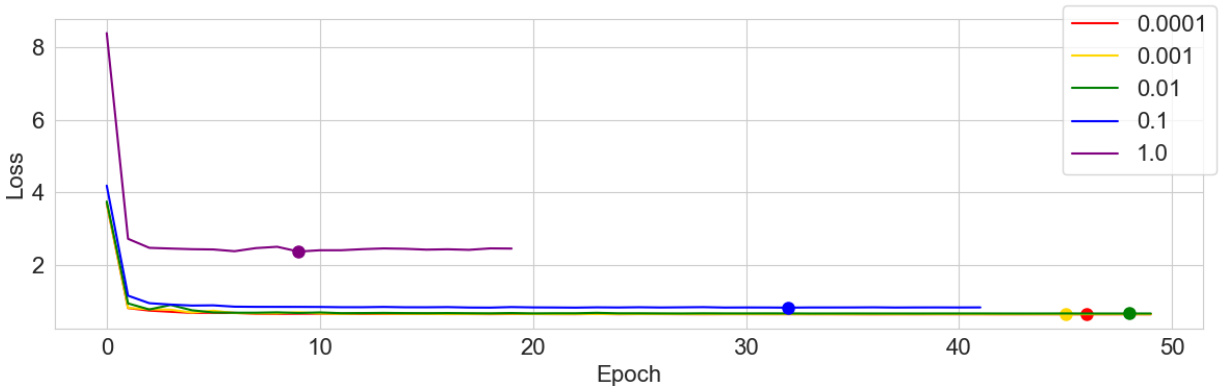
図 4.1: 手法 1 における検証データに対する損失曲線



(a) メルスペクトログラムの MAE Loss（式 (4.1) の L_{mel}）

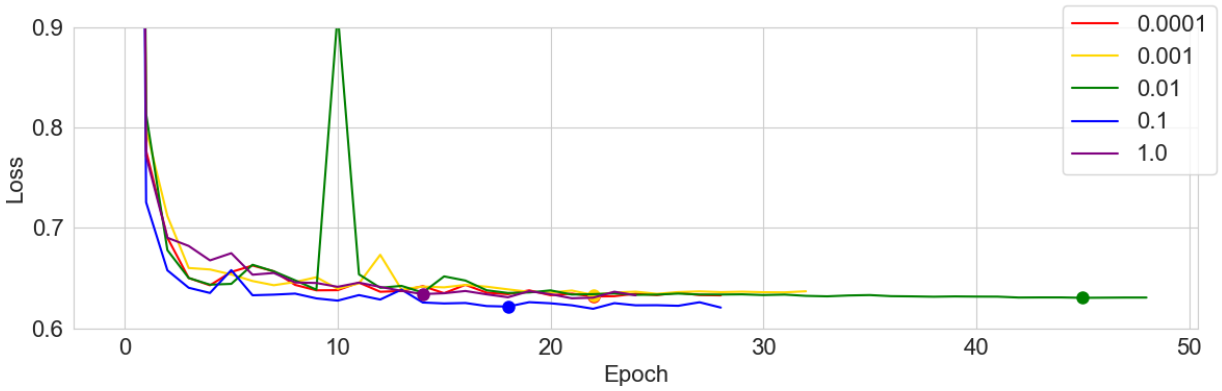


(b) HuBERT 離散特徴量の Cross Entropy Loss（式 (4.1) の L_{ssl^d}）

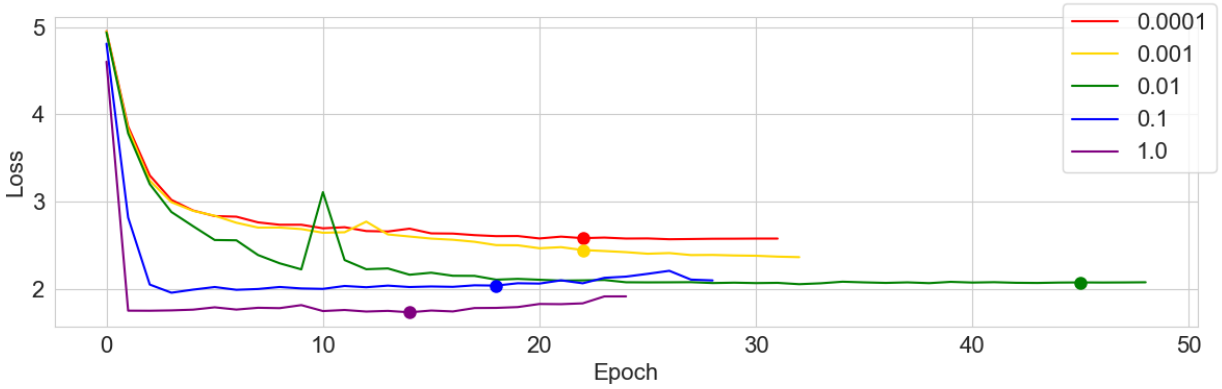


(c) 損失の合計値（式 (4.1) の L）

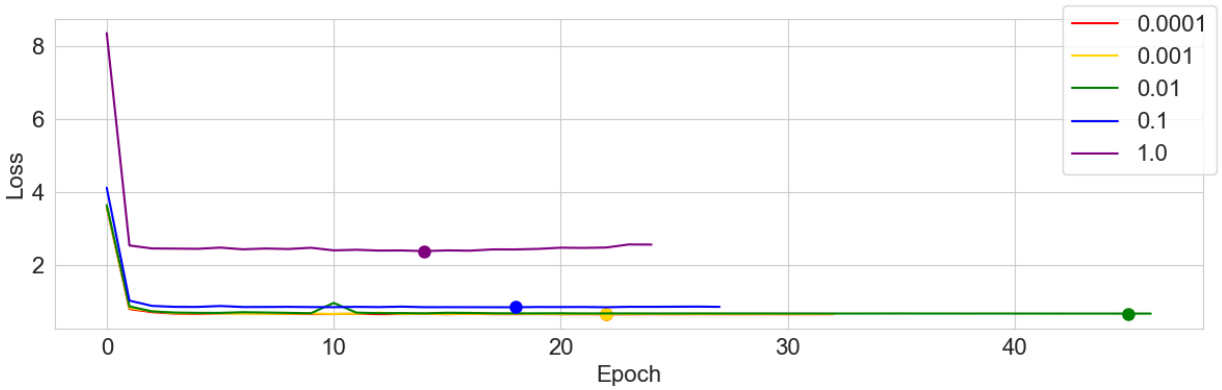
図 4.2: 手法 2 における検証データに対する損失曲線



(a) メルスペクトログラムの MAE Loss（式 (4.1) の L_{mel}）

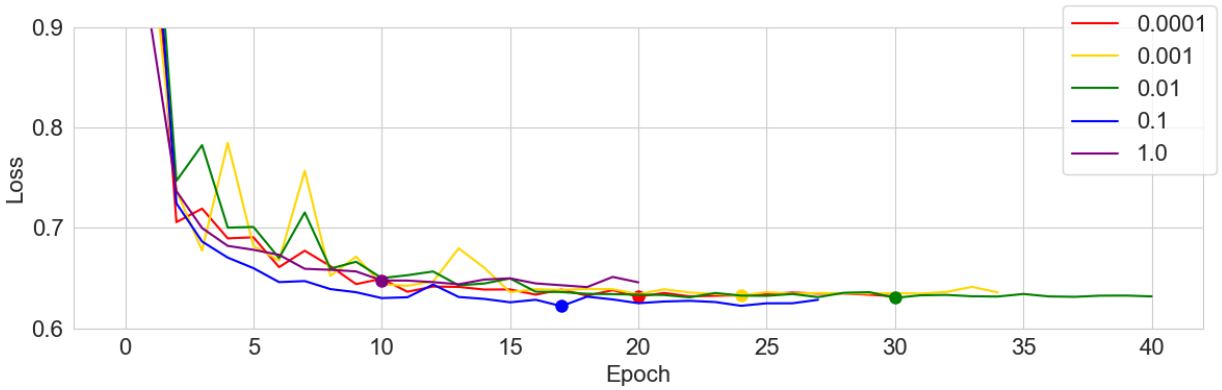


(b) HuBERT 離散特徴量の Cross Entropy Loss（式 (4.1) の L_{ssl^d}）

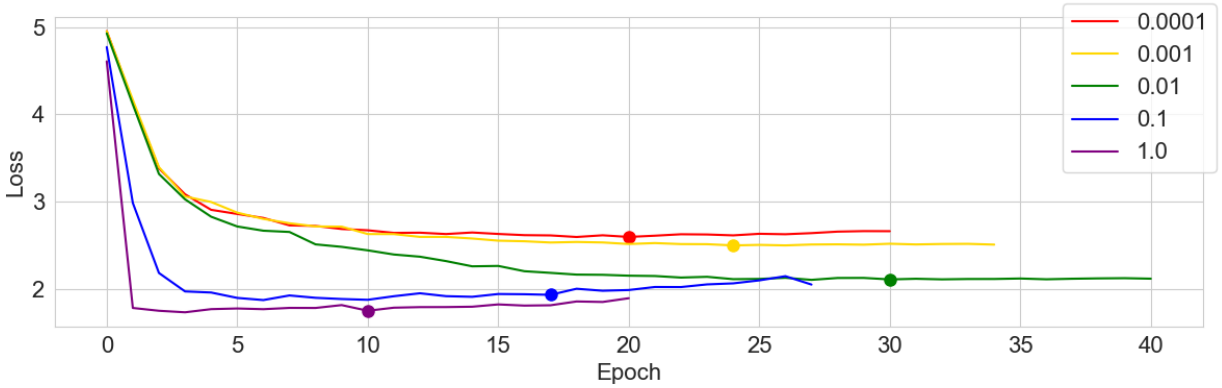


(c) 損失の合計値（式 (4.1) の L）

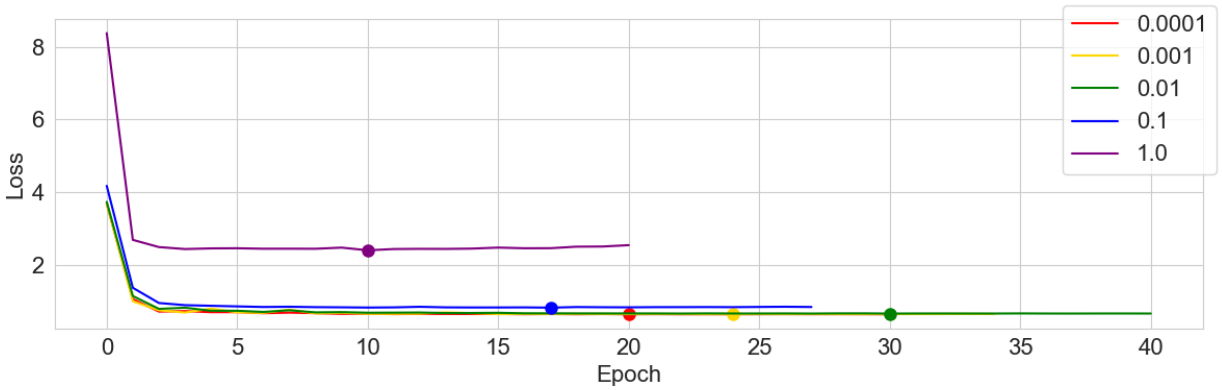
図 4.3: 手法 3 における検証データに対する損失曲線



(a) メルスペクトログラムの MAE Loss（式 (4.1) の L_{mel}）

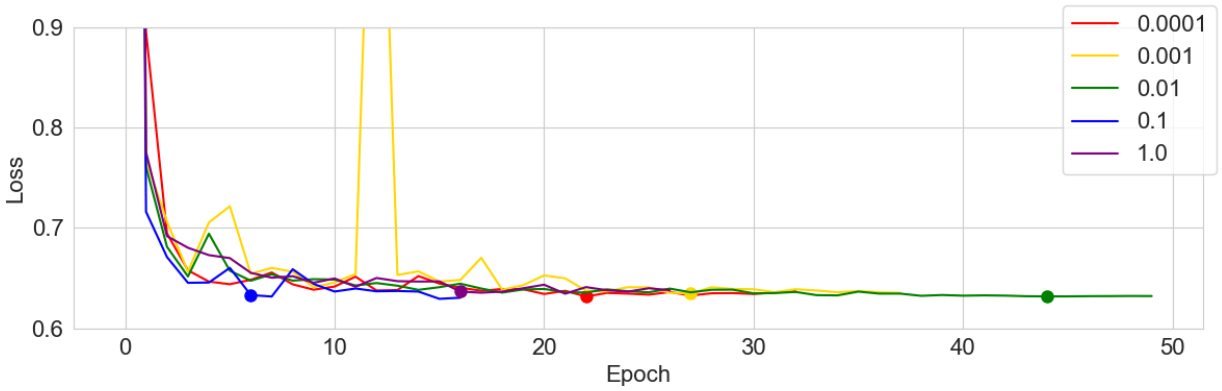


(b) HuBERT 離散特徴量の Cross Entropy Loss（式 (4.1) の L_{ssl^d}）

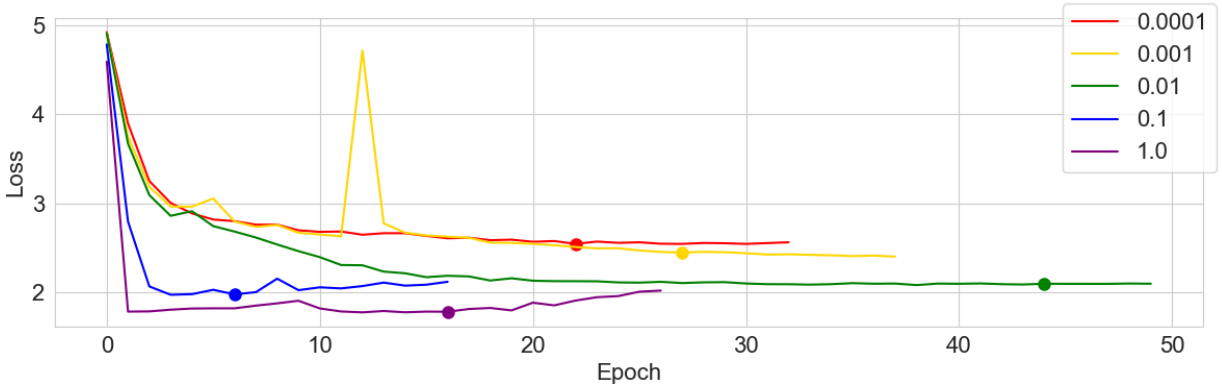


(c) 損失の合計値（式 (4.1) の L）

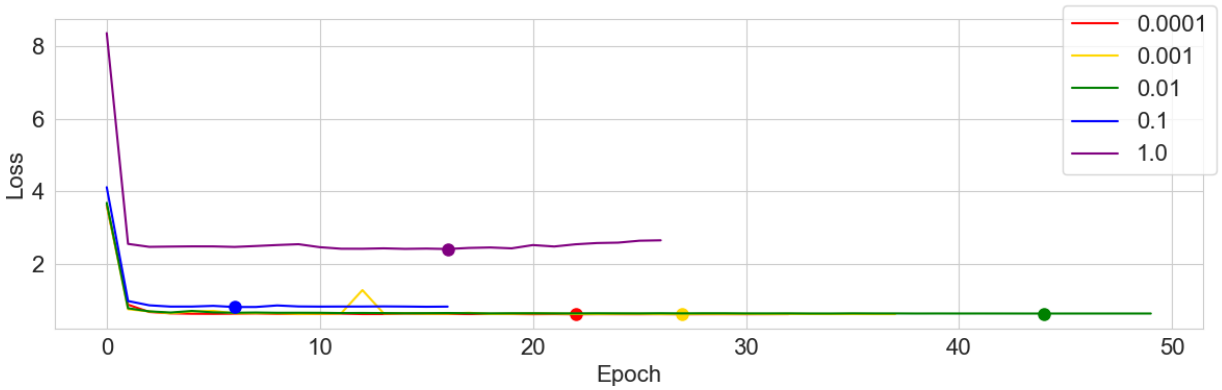
図 4.4: 手法 4 における検証データに対する損失曲線



(a) メルスペクトログラムの MAE Loss（式 (4.1) の L_{mel}）



(b) HuBERT 離散特徴量の Cross Entropy Loss（式 (4.1) の L_{ssl^d}）



(c) 損失の合計値（式 (4.1) の L）

図 4.5: 手法 5 における検証データに対する損失曲線

5　結論

謝辞

参考文献

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

[2] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[3] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022.

[4] Qiushi Zhu, Long Zhou, Ziqiang Zhang, Shujie Liu, Binxing Jiao, Jie Zhang, Lirong Dai, Daxin Jiang, Jinyu Li, and Furu Wei. Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. *IEEE Transactions on Multimedia*, 2023.

[5] Alexandros Haliassos, Pingchuan Ma, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Jointly learning visual and auditory speech representations from raw data. *arXiv preprint arXiv:2212.06246*, 2022.

[6] Jiachen Lian, Alexei Baevski, Wei-Ning Hsu, and Michael Auli. Av-data2vec: Self-supervised learning of audio-visual speech representations with contextualized target representations. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023.

[7] Alexandros Haliassos, Andreas Zinonos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Braven: Improving self-supervised pre-training for visual and auditory speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11431–11435. IEEE, 2024.

[8] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip-to-speech synthesis in the wild with multi-task learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

[9] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, Vol. 29, pp. 3451–3460, 2021.

[10] Jeongsoo Choi, Minsu Kim, and Yong Man Ro. Intelligible lip-to-speech synthesis with speech units. *arXiv preprint arXiv:2305.19603*, 2023.

[11] Wei-Ning Hsu, Tal Remez, Bowen Shi, Jacob Donley, and Yossi Adi. Revise: Self-supervised speech resynthesis with visual input for universal and generalized speech regeneration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18795–18805, 2023.

[12] Neha Sahipjohn, Neil Shah, Vishal Tambrahalli, and Vineet Gandhi. Robustl2s: Speaker-specific lip-to-speech synthesis exploiting self-supervised representations. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1492–1499. IEEE, 2023.

[13] Ji-Hoon Kim, Jaehun Kim, and Joon Son Chung. Let there be sound: Reconstructing high quality speech from silent videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, pp. 2759–2767, 2024.

[14] Jeong Hun Yeo, Minsu Kim, Jeongsoo Choi, Dae Hoe Kim, and Yong Man Ro. Akvsr: Audio knowledge empowered visual speech recognition by compressing audio knowledge of a pretrained model. *IEEE Transactions on Multimedia*, 2024.

[15] Xize Cheng, Tao Jin, Linjun Li, Wang Lin, Xinyu Duan, and Zhou Zhao. Opensr: Open-modality speech recognition via maintaining multi-modality alignment. *arXiv preprint arXiv:2306.06410*, 2023.

[16] Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Haithem Boussaid, Ebtessam Almazrouei, and Merouane Debbah. Lip2vec: Efficient and robust visual speech recognition via latent-to-latent visual to audio representation mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13790–13801, 2023.

[17] Xubo Liu, Egor Lakomkin, Konstantinos Vougioukas, Pingchuan Ma, Honglie Chen, Ruiming Xie, Morrie Doulaty, Niko Moritz, Jachym Kolar, Stavros Petridis, et al. Synthvsr: Scaling up visual speech recognition with synthetic supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18806–18815, 2023.

[18] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. Auto-avsr: Audio-visual speech recognition with automatic labels. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

[19] Oscar Chang, Hank Liao, Dmitriy Serdyuk, Ankit Shahy, and Olivier Siohan. Conformer is all you need for visual speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10136–10140. IEEE, 2024.

[20] Andreas Zinonos, Alexandros Haliassos, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Learning cross-lingual visual speech representations. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

[21] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

[22] Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W Oard, and Matt Post. The multilingual tedx corpus for speech recognition and translation. *arXiv preprint arXiv:2102.01757*, 2021.

[23] Ya Zhao, Rui Xu, and Mingli Song. A cascade sequence-to-sequence model for chinese mandarin lip reading. In *Proceedings of the 1st ACM International Conference on Multimedia in Asia*, pp. 1–6, 2019.

[24] Minsu Kim, Jeong Hun Yeo, Jeongsoo Choi, and Yong Man Ro. Lip reading for low-resource languages by learning and combining general speech knowledge and language-specific knowledge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15359–15371, 2023.

[25] Jeong Hun Yeo, Minsu Kim, Shinji Watanabe, and Yong Man Ro. Visual speech recognition for low-resource languages with automatic labels from whisper model. *arXiv preprint arXiv:2309.08535*, 2023.

[26] 田口史郎. ”深層学習を用いたデータ駆動型調音・音声間変換に関する研究”. 九州大学大学院芸術工学府芸術工学専攻 博士論文, 2021.

[27] 江崎蓮. ”深層学習を用いた口唇動画・音声変換に関する調査”. 九州大学大学院芸術工学府芸術工学専攻 修士論文, 2022.

[28] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4879–4883. IEEE, 2018.

[29] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, Vol. 33, pp. 17022–17033, 2020.

[30] Yoshinori Sagisaka, Kazuya Takeda, M Abel, Shigeru Katagiri, Tetsuo Umeda, and Hisao Kuwabara. A large-scale japanese speech database. In *ICSLP*, pp. 1089–1092, 1990.

[31] T Okamoto, Y Shiga, and H Kawai. Hi-fi-captain: High-fidelity and high-capacity conversational speech synthesis corpus developed by nict, 2023.

[32] Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari. Jvs corpus: free japanese multi-speaker voice corpus. *arXiv preprint arXiv:1908.06248*, 2019.

[33] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pp. 1021–1030, 2017.

[34] Yukiya Hono, Kentaro Mitsui, and Kei Sawada. rinna/japanese-hubert-base.

[35] Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. Release of pre-trained models for the Japanese language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 13898–13905, 5 2024. Available at: https://arxiv.org/abs/2404.01657.

[36] Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[38] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.