

End-to-End 音声合成を用いた 単語単位 End-to-End 音声認識のデータ拡張

上乃 聖^{1,a)} 三村 正人¹ 坂井 信輔¹ 河原 達也¹

概要: 単語単位 End-to-End 音声認識は簡潔な構造で非常に高速な認識ができ、高い性能を達成している。しかし、単語単位音声認識モデルには、未知語を登録・認識できない問題と、テキストのみを用いた学習ができないという問題がある。一方で End-to-End 音声合成も近年研究されており、人間の音声に近い自然性を達成している。そこで本研究では、End-to-End 音声合成を用いた音声認識のデータ拡張を提案する。音声合成は通常単一話者で訓練されるが、音声認識には多様性のあるデータを必要とする。そこで、音声合成を多数話者の音声を出力できるように拡張し、音声合成による学習データ生成をより汎用的なものになることを目指す。音声合成は話者情報を符号化して、多数話者のコーパスから学習し、認識したいドメインのテキストから音声合成する。これらの合成音声と人間が発話した自然音声を組み合わせて注意機構を用いたエンコーダデコーダモデルによる単語単位音声認識モデルの学習を行う。実験により多数話者音声合成を用いたモデルはベースラインモデルや単一話者音声合成を用いたモデルよりも大きな改善が見られた。

1. はじめに

End-to-End 音声認識は音響特徴量を直接記号系列に変換するシステムであり、非常に簡潔な構造で構築が容易である。End-to-End 音声認識の実現方法として、Connectionist Temporal Classification (CTC) を用いた手法 [1] や、RNN トランジェューサや注意機構モデルを用いた sequence-to-sequence (seq2seq) モデル [2] などが挙げられる。これらの手法は HMM などの潜在状態遷移モデルを必要とせずに音響特徴量を記号系列に変換することができる。End-to-End 音声認識の出力単位に関しては、音響特徴量から単語系列を直接出力する単語単位音声認識モデル [3] が外部デコーダなどを必要としないため、特に高速な認識を実現できる。我々は注意機構モデルを用いた単語音声認識モデルが従来の DNN-HMM ハイブリッドモデルに比べて、非常に速いデコードで、単語誤り率を相対的に 25.3%改善することを示した [4]。

しかし、単語単位音声認識モデルには音素単位モデルや文字単位モデルに比べて、いくつか問題が存在する。その中でも特に重大な問題は、訓練中に出現しない単語を認識できず、またそれを訓練後に追加することができない点で

ある。さらに単語単位音声認識モデルは音声とその書き起こしのデータを多量に必要とする。ドメインへの適応を行う際に、適応先のデータは多量には手に入らない場合も多く、適応先のデータのテキストデータが使用可能であったとしても語彙が異なるため、完全には活用することはできない。

これらの単語単位認識モデルの問題を解決するために、ターゲットドメインのテキストデータから音響特徴量を End-to-End 音声合成により生成する方法を研究している [5] [6]。End-to-End 音声合成は近年研究されており [7] [8]、従来の音声合成システムに比べて非常に簡潔で訓練も容易である。その上、いくつかの研究では人間の発話に近い自然さを達成したと報告している [8]。これにより生成される特徴量を自然音声による特徴量とともに単語単位音声認識モデルの学習に用いる。自然音声の訓練データ中に出現しない単語について合成音声で増やすことで新たな単語追加も可能になる。一方で、音声合成は通常単一話者のデータを用いて学習されており、多様性がない。この多様性のなさは音声認識においては非常に問題となる。本研究では、End-to-End 音声合成の枠組みに話者埋め込みを追加することで、音声認識の学習に適した多数話者の音声データを生成できるようにする。

¹ 京都大学情報学研究科
Graduate School of Informatics, Kyoto University, Sakyo-ku,
Kyoto 606-8501, Japan

^{a)} ueno@sap.ist.i.kyoto-u.ac.jp

2. End-to-End 音声認識と End-to-End 音声合成

2.1 注意機構モデル

注意機構を用いたモデルはエンコーダとデコーダの2つのネットワークから構成される。エンコーダでは LSTM を用いて音響特徴量系列を分散表現にする。デコーダではエンコードされた系列表現と出力記号表現との関連性を考慮して出力記号系列を生成する。本研究ではエンコーダに複数層の双方向 LSTM を用い、デコーダには1層の単方向 LSTM、注意機構の計算は [9] をもとに行う。デコーダに LSTM を用いることで前の記号列をもとに次の記号列を予測する。これは言語モデルの構造が注意機構モデルは含まれているとみなすことができる。損失関数は予測記号系列と正解記号系列とのクロスエントロピーを用いる。

2.2 単語単位音声認識モデル

単語単位音声認識モデルは音響特徴量から単語系列を直接出力するモデルであり、外部機構の処理を一切用いずに非常に簡潔で高速な認識が可能となる。しかし、このモデルの学習には非常に多くの音声と書き起こしのペアデータが必要となる。この問題は文字単位モデルとのマルチタスク学習を行うことで緩和できる [10]。しかし、サブワード単位のモデルと異なり、新たな語彙を追加することができない。その上、新たなドメインに対して容易に用意できるテキストのみのデータを活用することができない。語彙が一致していないと、[11] のように外部の言語モデルと組み合わせることは容易でない。

2.3 End-to-End 音声合成

End-to-End 音声合成は文字系列や音素系列から音声を生成する。多くのモジュールをからなり、人手がかかる従来の音声合成に比べて、非常に簡潔な構造となっている。近年では End-to-End 音声合成は自然音声に近い MOS (Mean Opinion Score) を獲得している [8]。

本研究では Tacotron 2 [8] モデルをベースに用いる。Tacotron 2 は注意機構を持つエンコーダデコーダモデルと WaveNet を用いたボコーダで構成される。エンコーダデコーダでは音響特徴量やボコーダに用いるパラメータを音素や文字系列から生成する。ボコーダではそれらの特徴量から音声波形に変換する。本研究では音声認識の学習のために音声波形ではなく音響特徴量のみが必要なためボコーダは用いない。エンコーダでは文字系列から文字埋め込み、3層の畳み込み層と1層の双方向 LSTM を経て分散表現を得る。デコーダでは注意機構を用いて、各デコーダのステップで一度に5フレーム分の音響特徴量を生成する。

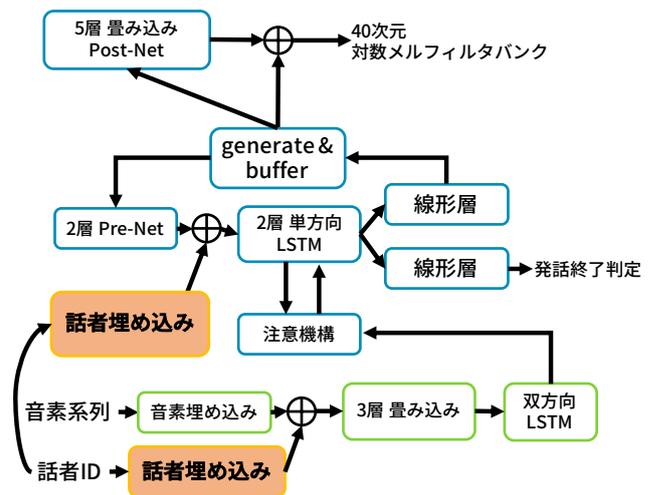


図1 多数話者音声合成の概念図。エンコーダでは畳み込みの入力は音素埋め込みと話者埋め込みとの足し合わせにし、デコーダでは話者埋め込みは pre-net との和をとり、LSTM の前状態とする。

3. 提案手法

3.1 単語単位音声認識モデル学習のための End-to-End 音声合成の利用

テキストデータを用いて単語単位音声認識を新たなドメインに適応するために、End-to-End 音声合成を利用し、訓練データを拡張する手法を提案する。まず、音声認識を行う対象のターゲットドメインのテキストを集める。それらのテキストの情報を End-to-End 音声合成に与えて、音響特徴量を生成する。その合成された音響特徴量と対応する単語系列を自然音声のコーパスに加えて、単語単位音声認識モデルを学習する。この手法は任意の文から学習データを生成でき、語彙を増やして、単語単位音声認識モデルにおける言語モデル機構の改善をすることができる。また言語モデルの語彙が一致ようになるため、shallow fusion [11] を適用することもできる。

End-to-End 音声認識の学習にテキストデータを用いる研究はいくつか存在する。Renduchintala ら [12] は音響特徴量に変換せずに、テキストデータに特殊なエンコードを行うことで End-to-End 音声認識の学習を行った。Sriram ら [13] はあらかじめ学習した言語モデルを End-to-End 音声認識の学習に用いることで収束を早くし、新たなモデルへの適応を行った。また Tjandra らは、End-to-End 音声認識と音声合成との組み合わせた Speech Chain というモデルを提案している [14], [15]。本研究では直接的で効率的に単語単位音声認識モデルの改善を行うことを目指す。

3.2 複数話者 End-to-End 音声合成

音声合成は通常一人の話者のみで学習される。つまり、合成音声には多様性がない。一方で音声認識には話者の多

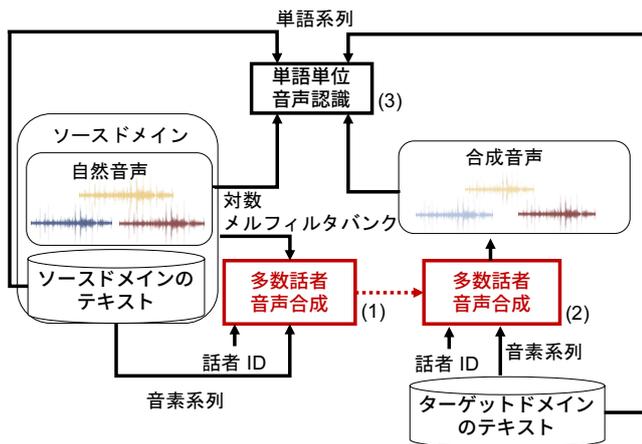


図 2 多数話者音声合成を利用したデータ生成の処理の流れ。(1) 多数話者音声合成を学習。(2) 音響特徴量を生成。(3) 自然音声と合成音声を用いて単語単位音声認識モデルを学習。

様性が必要である。

多様性のある音声の生成のために、多くの話者の音声が入録された大規模コーパスから訓練できるように音声合成を設計する。複数話者の音声合成はいくつか提案されている。[7]ではエンコーダ、デコーダ、ポコーダに話者埋め込みを用いている。また Jia ら [16] は d-vector[17] などのような固定長の埋め込みを用いている。本研究では、[7]を参考に、Tacotron 2 の枠組みに話者埋め込みを追加する(図 1)。エンコーダでは softsign 関数により非線形化した話者埋め込みを畳み込み層の出力を足し合わせる。デコーダでも softsign 関数により非線形化した後に、pre-net の出力に足し合わせる。複数話者のコーパスを用いた音声合成の学習は単一話者コーパスによる学習よりも難しい。実際に予備実験ではランダムに初期化したモデルで学習した場合は収束しなかった。本研究では、はじめに単一話者のコーパスで学習し、そのパラメータを用いて複数話者のコーパスで学習する。そのため、話者埋め込みと他の出力を結合せず、足し合わせる方式とした。

図 2 に複数話者の音声合成を用いたデータ生成のフローを示す。まず、ベースラインのコーパスを用いて複数話者の音声合成の学習を行う。この際には話者埋め込みを使用する。次に、音響特徴量を適応先ドメインのテキストから生成する。ランダムに話者 ID を選ぶことで同じ文章から多様性のある音声を生成できる。最後に自然音声と合成音声を両方を用いて単語単位音声認識モデルを学習する。

4. 評価実験

4.1 データセット

本研究では『日本語話し言葉コーパス』(CSJ)を用いる。CSJはCSJ-APSとCSJ-SPSの二つのサブコーパスで構成されている。CSJ-APSは学会講演を収録したコーパスで、訓練データは247.9時間、986名の話者(男性:809名、女

性:177名)のデータで構成される。CSJ-SPSは3つのテーマでスピーチを行った模擬講演コーパスで、訓練データは281時間、1074話者(男性:799名、女性:905名)のデータで構成される。それぞれのサブコーパスでテストセットが提供されており、本研究ではテストセット1(CSJ-APS)とテストセット3(CSJ-SPS)を使用する。語彙には2回以上出現した単語と〈sos〉、〈eos〉、〈UNK〉といった特殊なラベルを使用する。語彙サイズはAPSでは19,146、SPSでは24,286、APSとSPSを合わせたものでは34,331となる。APSとSPS内で共有している単語は11,446単語存在する。

4.2 システム構成

4.2.1 単語単位音声認識モデル

単語単位音声認識モデルに入力する音響特徴量として40次元の対数メルフィルタバンクを用いる。この音響特徴量にフレームスタッキング[18]を適用し、オーバーラップのない3フレーム分の音響特徴量を単語単位音声認識モデルの入力とする。エンコーダは5層の320次元の隠れ層を持つ双方向LSTMで構成する。また、ドロップアウトを0.2に設定し、各双方向LSTMに適用する。注意機構を用いたデコーダは、1層の320次元の隠れ層を持つ単方向LSTMで構成し、その後出力単語数分の単語数のノードを持つsoftmaxの出力層となる。最適化アルゴリズムはAdam[19]を用い、Gradient Clippingの閾値を5.0とした。正則化のためにラベルスムージング[20]を用いる。また、認識時のビーム幅は4とした。shallow fusionを行うための言語モデルとして、3層の256次元の隠れ層を持つ単方向LSTMを用いる。LSTMの処理前に、各単語は512次元の分散表現にする。これらはPyTorchを用いて実装されている[21]。

4.2.2 多数話者音声合成

オリジナルのTacotron 2では入力は文字系列で、出力は80次元のメルスペクトログラムであり、損失関数はメルスペクトログラムとの平均二乗誤差である。しかし、本研究では入力を音素系列とし、出力を40次元の対数メルフィルタバンク、損失関数をL1 lossとする。この40次元の対数フィルタバンクは直接単語単位音声認識モデルの入力として用いられる。

テキストの形態素解析と読み付与にはMecabを用いた。ポーズ、単語境界、文末を含む33個の音素を用いる。エンコーダは512次元の音素埋め込み、各層に512次元のフィルタを持つ3層の畳み込み層と256次元の話者埋め込み、256次元の隠れ層を持つ双方向LSTMで構成される。また注意機構の計算は単語単位音声認識モデルと同様に[9]を用いる。注意機構の重みの計算にはデコーダのLSTMの状態、エンコーダの出力系列、位置情報を128次元にマッピングして計算する。位置情報は32次元の畳み込み層を用いて計算される。予測のための潜在表現を2層の256次

表 1 CSJ-APS と CSJ-SPS テストセットにおける単語誤り率 (%).
 本表ではベースラインは APS. 適応先ドメインは SPS.

	ベースライン (APS)	適応先 (SPS)	+ 言語モデル統合
ベースライン 自然音声 + 適応先 自然音声 (oracle)	10.35	9.06	9.00
ベースライン 自然音声	12.30	19.22	18.84
ベースライン 自然音声 + 適応先 合成音声 (単一話者)	11.89	14.64	14.16
ベースライン 自然音声 + 適応先 合成音声 (多数話者)	11.43	13.37	13.27

表 2 CSJ-APS と CSJ-SPS テストセットにおける単語誤り率 (%).
 本表ではベースラインは SPS. 適応先ドメインは APS.

	ベースライン (SPS)	適応先 (APS)	+ 言語モデル統合
ベースライン 自然音声 + 適応先 自然音声 (oracle)	9.06	10.35	10.24
ベースライン 自然音声	9.69	23.30	23.14
ベースライン 自然音声 + 適応先 合成音声 (単一話者)	9.86	18.74	18.24
ベースライン 自然音声 + 適応先 合成音声 (多数話者)	9.36	16.68	15.94

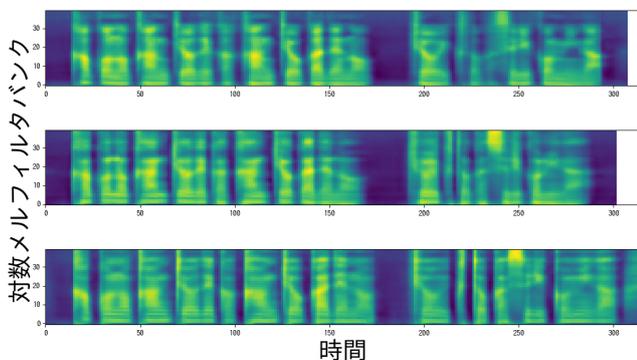


図 3 合成した対数メルフィルタバンクの例. 話者 ID はランダムに設定し, 言い直しが入った文章である “kako no kaNkyo: de ka kaNkyo: ka de no, kyo:iku to ka surikomi ga” (「過去の環境でか環境下での教育とか刷り込みが」) を入力とした.

元の線形層を持つ pre-net に用いる. この pre-net の出力と 256 次元の話者埋め込み, 注意機構により計算されたエンコード情報を足し合わせた値を 1024 次元の隠れ層を持つ 2 層の単方向 LSTM に与える. LSTM の出力を線形層を用いて, 5 フレームごとに音響特徴量を出力する.

本研究では, 単一話者の JSUT コーパス [22] を用いて初期モデルを学習する. JSUT コーパスは 1 人の女性話者による 7607 文の読み上げ音声を収録したもので, 10 時間の音声が入力されている. この単一話者のモデルは比較システムとしても用いる. その後多数話者コーパスを用いてファインチューニングを行う. 多数話者のモデルを用いて音響特徴量を生成する際には, 与えられたテキストに対して, ランダムに話者 ID を選ぶ.

4.3 結果

2 つのサブコーパスの中から 1 つをベースラインとし, もう 1 つを適応先ドメインとして設定する. ベースライン

のデータでは単語単位音声認識モデルと複数話者の音声合成を学習する. 適応先ドメインでは書き起こしデータのみを用いて適応を行う. 適応先ドメインのテキストから音声データを生成し, 拡張したデータを用いて単語単位音声認識モデルの再学習を行う.

図 3 に複数話者モデルにより生成された対数メルフィルタバンクを示す. CSJ-SPS により複数話者モデルを学習し, CSJ-APS のテキストから 3 話者分の音響特徴量を生成している. これらの音響特徴量の長さやスペクトルの特徴が異なり, 複数話者モデルが多様性のある音声を出力できていることがわかる.

表 1 と表 2 に, CSJ-APS と CSJ-SPS のテストセットに対して単語誤り率 (WER) を評価した結果を示す. 表 1 ではベースラインは CSJ-APS, 適応先ドメインは CSJ-SPS とし, 213 時間の合成音声を用いた. 未知語率はベースラインの APS モデルでは 3.53% であるが, SPS のテキストと合成音声を用いることで 1.21% まで減少した. 表 2 ではベースラインは CSJ-SPS, 適応先ドメインは CSJ-APS としている. 合成音声は 209 時間となり, 未知語率は 4.28% から 0.85% に減少した. 4.1 節で述べた音声の長さや異なるのは, End-to-End 音声合成が与えられたテキストから音声の長さを推定するためである.

適応先の評価では, ベースラインモデルの単語誤り率 (WER) が非常に高いが, 単一話者のモデルを用いた音声合成によるデータ拡張を行うことで大きな改善が見られ, ベースラインモデルで出現しなかった単語が認識できた. また, 複数話者モデルによりさらに適応先ドメインで大きく改善し, 元のドメインに対しても改善が見られた. これは同じ量のデータを生成しても単一話者よりも多様性があるためである. しかし, 1 つの文から 2 人分の音声を生成しても, WER は改善しなかった. また自然音声 (oracle)

との差は大きく、合成音声は自然音声ほどの多様性はないといえる。

単語単位音声認識モデルに加えて、shallow fusion による外部言語モデルを用いることができる。言語モデルはAPSとSPSを両方用いて学習した、元のドメインのみのベースラインモデルに適用する際は語彙を合わせている。この外部言語モデル統合は全てのモデルについて効果的である。しかし、ベースラインモデルとの統合では提案手法の性能に大きく及ばない。この結果は言語モデル統合のみではデータ拡張ほどは効果が得られないことを示している。

5. おわりに

本研究では単語単位音声認識モデルのためのデータ拡張を行うために、多数話者 End-to-End 音声合成モデルを用いることを提案した。多数話者の大規模コーパスで学習することで、多様性のある音声を合成できた。この拡張方法はドメイン適応において大きな改善を示した。さらに、shallow fusion による外部言語モデルとの統合により改善を得られた。

参考文献

- [1] Graves, A., Fernandez, S., Gomez, F. and Schmidhuber, J.: Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks, *Proc ACM*, pp. 369–376 (2006).
- [2] Battenberg, E., Chen, J., Child, R., Coates, A., Gaur, Yi Li, Y., Liu, H., Satheesh, S., Sriram, A. and Zhu, Z.: Exploring neural transducers for end-to-end speech recognition, *Proc. ASRU*, pp. 206–213 (2017).
- [3] Sak, H., Senior, A., Rao, K., Irsoy, O., Graves, A., Beaufays, F. and Schalkwyk, J.: Learning acoustic frame labeling for speech recognition with recurrent neural networks, *Proc. ICASSP*, pp. 4280–4284 (2015).
- [4] Mimura, M., Sakai, S. and Kawahara, T.: Forward-backward attention decoder, *Proc. INTERSPEECH*, pp. 2232–2236 (2018).
- [5] Mimura, M., Ueno, S., Inaguma, H., Sakai, S. and Kawahara, T.: Leveraging sequence-to-sequence speech synthesis for enhancing acoustic-to-word speech recognition, *Proc. SLT* (2018).
- [6] 上乃聖, 三村正人, 河原達也: End-to-End 音声合成を用いた単語単位 End-to-End 音声認識の訓練データ拡張, 日本音響学会秋季研究発表会講演論文集,(2018).
- [7] Gibiansky, A., Arik, S., Diamos, G., Miller, J., Peng, K., Ping, W., Raiman, J. and Zhou, Y.: Deep voice 2: Multi-speaker neural text-to-speech, *Proc. NIPS*, pp. 2962–2970 (2017).
- [8] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. et al.: Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions, *Proc. INTERSPEECH*, pp. 4779–4783 (2017).
- [9] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y.: Attention-Based Models for Speech Recognition, *Proc. NIPS*, pp. 577–585 (2015).
- [10] Ueno, S., Inaguma, H., Mimura, M. and Kawahara, T.: Acoustic-to-word attention-based model complemented with character-level CTC-based model, *Proc. ICASSP*, pp. 5804–5808 (2018).
- [11] Kannan, A., Wu, Y., Nguyen, P., Sainath, T. N., Chen, Z. and Prabhavalkar, R.: An analysis of incorporating an external language model into a sequence-to-sequence model, *Proc. ICASSP, IEEE*, pp. 5824–5828 (2018).
- [12] Renduchintala, A., Ding, S., Wiesner, M. and Watanabe, S.: Multi-Modal Data Augmentation for End-to-end ASR, *Proc. INTERSPEECH*, pp. 2394–2398 (online), DOI: 10.21437/INTERSPEECH.2018-2456 (2018).
- [13] Sriram, A., Jun, H., Satheesh, S. and Coates, A.: Cold Fusion: Training Seq2Seq Models Together with Language Models, *Proc. INTERSPEECH*, pp. 387–391 (online), DOI: 10.21437/INTERSPEECH.2018-1392 (2018).
- [14] Tjandra, A., Sakti, S. and Nakamura, S.: Listening while speaking: Speech chain by deep learning, *Proc. ASRU*, pp. 301–308 (2017).
- [15] Tjandra, A., Sakti, S. and Nakamura, S.: Machine Speech Chain with One-shot Speaker Adaptation, *Proc. INTERSPEECH*, pp. 887–891 (online), DOI: 10.21437/INTERSPEECH.2018-1558 (2018).
- [16] Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I. L. et al.: Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis, *arXiv preprint, 1806.04558* (2018).
- [17] Heigold, G., Moreno, I., Bengio, S. and Shazeer, N.: End-to-end text-dependent speaker verification, *Proc. ICASSP*, pp. 5115–5119 (2016).
- [18] Sak, H., Senior, A., Rao, K. and Beaufays, F.: Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition, *INTERSPEECH*, pp. 1468–1472 (2015).
- [19] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *arXiv preprint, 1412.6980*, pp. 1–15 (2014).
- [20] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z.: Rethinking the inception architecture for computer vision, *Proc. CVPR*, pp. 2818–2826 (2016).
- [21] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. and Lerer, A.: Automatic differentiation in PyTorch, *NIPS-W* (2017).
- [22] Sonobe, R., Takamichi, S. and Saruwatari, H.: JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis, *arXiv preprint, 1711.00354* (2017).