Vol.2021-NL-251 No.28
Vol.2021-SLP-139 No.28
2021/12/3

音声認識のデータ拡張のための合成音声の周波数スペクトログラム強調

上乃 聖[1,a)]　河原 達也[1]

概要：End-to-End 音声認識が高い精度を達成しつつあるが，大量のデータを必要とする．その問題の解決のために音声合成を用いて音声認識の訓練データを生成することを検討してきた．音声合成においては，通常テキストから対数メルフィルタバンクを作る text-to-mel ネットワークを用いた後に，メルスペクトログラムを音声波形に変換するボコーダを用いて，音声波形を生成する．それを再びメルスペクトログラムに変換し，音声認識の訓練データとして用いる．ボコーダには合成音声と自然音声の差異を埋める効果があるが，この波形生成に非常に時間がかかるという問題がある．そこで本研究ではボコーダを用いずに周波数スペクトログラム上で直接強調を行うネットワークを提案する．提案手法では，生成されたメルスペクトログラムだけでなく，音声合成のタスクで利用可能な音素情報も用いる．評価実験から，提案手法がボコーダを用いるよりも少ない処理時間で拡張の効果が高いことを示し，また，音素情報の利用が改善に重要であることを示した．

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，従来の DNN-HMM ハイブリッドモデルと比較して，高い精度を達成しつつある．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 トランデューサ [2] や注意機構モデルを用いた sequence-to-sequence (seq2seq) モデル [3]，Transformer を用いたモデル [4], [5] などが挙げられる．しかし，これらのモデルの訓練には多くの音声とその書き起こしのペアデータを必要とし，特に特定ドメインの音声認識を行いたい場合には問題となる．

一方で，対象ドメインに適合した音声のないテキストのみのデータは多く入手できることが多いため，テキストのみのデータを活用して訓練データの不足を解決できる可能性がある．本研究では，このような音声合成を用いた音声認識のデータ拡張に焦点を当てる．この枠組みでは音声合成を用いて，テキストのみのデータから音声データを生成し，それを自然音声と合わせて音声認識の訓練データとし，データ拡張 [6], [7], [8], [9], [10], [11], [12] やドメイン適応 [13], [14], [15], [16] などに適用する．その際には Tacotron 2 [17] などのテキスト系列からメルスペクトログラムを合成する text-to-mel ネットワークが用いられる．メルスペクトログラムを合成した後に音声認識の訓練データとして用いる際には，主に二つのアプローチがある．一つは，そのまま後処理なしにメルスペクトログラムを直接，音声認識の訓練データとして用いる方法 [6], [7], [8], [9], [10], [11], [12] である．もう一つの方法はボコーダを用いて，メルスペクトログラムから音声波形に変換し，それを再びメルスペクトログラムに変換し，音声認識の訓練データに用いるものである [6], [7], [11], [12], [15], [16]．ボコーダを用いる利点として，音声波形を介することで音声認識と音声合成を独立に設計でき，また，音声データの品質が向上して拡張の効果が高い．この方法では，ボコーダは合成後の後処理としてみなすことができる．

しかし，音声認識の訓練データのために多量のデータを音声波形に変換する必要があるため，合成の全体として余分に時間がかかる．本研究では，ボコーダを使わずに合成されたメルスペクトログラムを直接強調するネットワークを提案する．提案手法は，合成音声と自然音声の差異を埋めることを目的とし，ボコーダの代わりに用いる．周波数スペクトログラム上で強調を行うことで，音声波形上で処理を行うボコーダと比較し，生成を高速に行うことができる．また，音声合成のタスク上取得可能な音声のテキスト情報 (音素情報) を用いて強調のさらなる改善を行う．

[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. 関連研究

本研究では，音声認識の訓練データ拡張のために音声合成を用いる．本章では，本研究で用いる音声合成の枠組みとデータ拡張に適用した研究について紹介する．

2.1 Text-to-mel ネットワークとボコーダ

音声合成を実現する上で，text-to-mel ネットワークとボコーダの二つのネットワークに分けて構成するアプローチが広く用いられている．text-to-mel ネットワークでは，音素 (あるいは文字) 系列を入力とし，メルスペクトログラムを出力とする Tacotron 2 [17] や Transformer を用いたモデル [18] などが提案されている．これらのモデルは従来の統計的なモデルと比較して，簡潔なモデルで高い品質の音声を生成することができる．また，近年，FastSpeech-1,2 [19], [20] や Parallel Tacotron-1,2 [21], [22] といった，高速に合成ができる非自己回帰モデルが提案されている．本研究では FastSpeech 2 をもとにしたモデルを用いる．FastSpeech 2 では各入力音素に対してメルスペクトログラムの長さを variance adaptor と呼ばれるモジュールで予測する．エンコーダによって埋め込めれた音素情報を，予測された長さ分だけ拡張を行い，デコーダでメルスペクトログラムの予測を行う．variance adaptor では各音素のメルスペクトログラムに対応する長さの予測に加えて，F0 やパワーといった音響情報も予測する．

text-to-mel ネットワークにより生成されたメルスペクトログラムから音声波形をボコーダを用いて生成する．ボコーダでは LPCNet [23] や MelGAN [24] などが速い推論が可能なため，用いられている．しかし，メルスペクトログラムと比較して音声波形はサンプル数が多く，非常に長い系列長になるため，推論には依然として時間がかかる．

2.2 音声合成を用いた音声認識のデータ生成

先行研究において，合成された音声を効率的に用いるためのさまざまな手法が検討されている．三村ら [13] は合成音声を用いた音声認識の学習時にエンコーダを更新しないことを提案した．Wang ら [9], [10] は自然音声と合成音声の一貫性を保つための正則化について検討を行なった．Zheng ら [15] は音声認識モデルの未知語学習時に音声認識のデコーダの正則化のための損失関数を導入した．Fazel ら [12] は音声合成の複数スタイルの獲得，データ拡張時の合成音声の使用法，音声認識のエンコーダのパラメータ固定，パラメータの正則化などを用いて，複数段階にわたる訓練法の提案を行なった．Chen ら [8] は GAN を用いて合成音声の音響的な多様性を増やす手法を導入した．倉田ら [16] は合成音声と自然音声のミスマッチを緩和するために音声認識のエンコーダの前に mapping network を導入した．

3. 提案手法

3.1 ベースラインモデル

音声合成の text-to-mel ネットワークには，複数話者の出力ができるネットワークを用いる．これらは広く使われており，単一話者のモデルよりも拡張の効果が高い [6], [7], [12], [14]．複数話者の text-to-mel モデルには，話者 ID を用いたモデル [14]，VAE を用いたモデル [6]，事前学習した話者識別のモデルを用いるシステム [12]，global style token の利用 [7] などが挙げられるが，本研究では話者 ID を音素とともに音声合成の入力とするモデルを用いる．

3.2 周波数スペクトログラム強調

通常の音声合成のタスクでは，ボコーダは人間が聞き，評価できるようにするために用いられる．一方で，音声認識の訓練データ拡張では，ボコーダは音声認識と音声合成で用いるメルスペクトログラムの設定の差異を両モデルを変更することなく埋めるために用いられている．また，ボコーダは自然音声と合成音声の差異を埋め，訓練データ生成時に音声認識性能の改善をすることができる．実際にボコーダを使うことでメルスペクトログラムがぼやけていたものが改善されたことを確認した．しかし，ボコーダの推論と，さらに音声波形から再びメルスペクトログラムに変換する必要があるため，最終的にデータ準備に長い時間を必要とする．

本研究では，ボコーダに代わる推論の速いネットワークの提案を行う．提案ネットワークでは合成されたメルスペクトログラムから音声波形ではなく，メルスペクトログラムを出力する mel-to-mel ネットワークを用いる．合成されたメルスペクトログラムを直接強調することで，推論を速くする．スペクトログラム上で処理を行う従来の音声強調では，雑音を含むスペクトログラムに対して適用され，書き起こしの情報などは通常入手が難しいため，用いることができない．しかし，音声強調の性能は音素情報を用いることで改善することが知られており [25], [26]，音声合成のタスクでは音素情報が獲得可能であるため，本研究では埋め込まれた音素情報を用いる．

図 1 に提案手法の概要を示す．本手法では，まず FastSpeech 2 の学習を行う．学習後は FastSpeech 2 のパラメータを更新はしない．提案ネットワークは Transformer をもとにしたネットワークで，周波数軸上で強調を行うため，FastSpeech 2 で予測されたメルスペクトログラムと音素情報の入力のために variance adaptor の出力を用いる．また，出力部には残差ブロックを用いる．損失関数には正解メルスペクトログラムと予測メルスペクトログラムの L1 損失を用いる．FastSpeech 2 でも同様に L1 損失を用いるが，FastSpeech 2 ではテキストからメルスペクトログラム

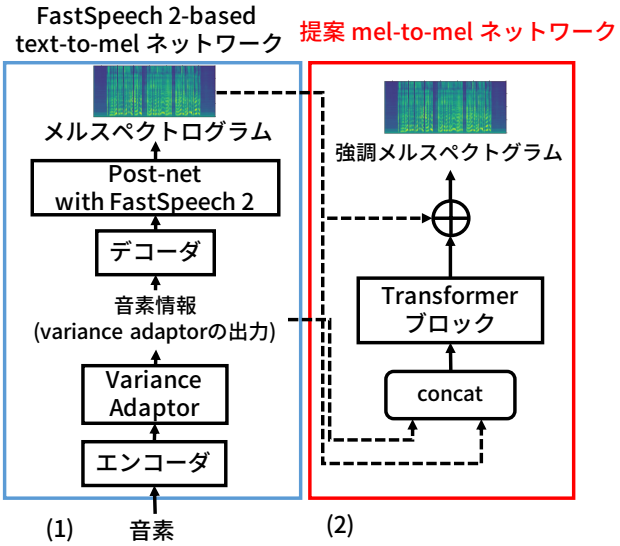

図 1 提案手法の概念図．(1) FastSpeech 2 (2) 強調ネットワーク．提案手法では合成されたメルスペクトログラムと音素情報 (variance adaptor の出力) を用いる．

の予測と同時に，継続長や F0，パワーの予測も行う複雑な学習を行う必要がある．一方で，提案ネットワークは予測されたメルスペクトログラムから最適化を行うため，より効率的に L1 損失を最小化することを期待する．また，音素情報の入力のために FastSpeech 2 内の variance adaptor の出力を用いる．通常，音素とメルスペクトログラムの系列長は異なるが，FastSpeech 2 の variance adaptor で長さを推定して，拡張するため，variance adaptor の出力である音素情報の埋め込みは系列長がメルスペクトログラムと同じになる．

ボコーダを用いる場合は，音声波形を介することでメルスペクトログラムの設定を音声認識に用いる設定に変更できるが，提案手法では，出力されたメルスペクトログラムの設定自体の変更はしない[*1]．近年の Transformer を用いたモデルは CNN を用いて，複数フレームを 1 フレームとして統合する機構があるため，音声認識と音声合成のメルスペクトログラムの設定の差異は埋めることができると考えられる．

4. 評価実験

4.1 データセット

本研究では英語と日本語のドメイン適応の実験を行う．英語の音声合成と音声認識のモデルの訓練には LibriTTS [27] と LibriSpeech [28] を用いる．LibriTTS は LibriSpeech のサブコーパスで音声合成のタスク用に設計されたものである．LibriTTS の音声はダウンサンプルを行い，16kHz のサンプリングレートの音声波形を用いる．LibriTTS と LibriSpeech では train-clean-100 のサブセットを用いる．

[*1] ただし，メルスペクトログラムの周波数ビン数は同じである必要があるため，本研究では 80 次元を用いる．

LibriTTS では，247 話者 (男性: 123 名，女性: 124 名)，約 53.8 時間の音声データが含まれ，LibriSpeech では約 100 時間のデータが含まれる．

音声合成タスクに際して，単語系列は 85 個の音素系列に変換を行う．variance adaptor の学習のための正解継続長は CTC モデルを用いて強制アライメントを行うことで獲得する．F0 情報は WORLD [29] により獲得する．

音声認識タスクでは，単語系列は 10,000 個の BPE(byte-pair-encoding) のサブワード系列に変換する．この実験では音声合成による訓練データ拡張は，読み上げ音声から話し言葉へのドメイン適応を目的とする．話し言葉の対象ドメインは TED-LIUM release-2 コーパス (以下 TED-LIUM 2) [30] を用いる．TED-LIUM 2 は 211 時間の 91,967 発話の講演音声であり，本研究では書き起こしのみを用い，音声合成によるデータ拡張を行う．言語モデル統合のために公式の TED-LIUM 2 の言語モデルのためのテキストデータを用いる．

日本語での実験では『日本語話し言葉コーパス』 (CSJ) を用いる．CSJ は CSJ-APS と CSJ-SPS の 2 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サブコーパスで，訓練データは 247.9 時間のデータで構成される．CSJ-SPS は日常生活に関する 3 つのテーマでスピーチを行った模擬講演サブコーパスで，訓練データは 281 時間のデータで構成される．本研究では CSJ-SPS のペアデータを用いて音声合成と音声認識の学習を行う．音声合成の学習には単語系列を 33 個の音素系列に変換し，音声認識の学習には 10,000 個の BPE 系列に変換する．音声合成器を用いて CSJ-APS の 151,627 発話について合成を行い，音声認識のドメイン適応を行う．評価にはテストセット 1(CSJ-APS ドメイン) を使用する．

4.2 システム構成

4.2.1 FastSpeech2 と提案手法

Text-to-mel ネットワークに FastSpeech 2 を用いる．エンコーダには 6 層，隠れ層 384 次元，feed-forward network(FFN)1,536 次元，4 つの head を持つ Transformer block で構成する．variance adaptor は 2 層の CNN，ReLU の活性化関数，layer normalization を用い，継続長，F0，パワーを予測する．variance adaptor の出力を用いて，エンコーダと同構成の Transformer block と 5 層の CNN で構成された post-net を用いて窓幅 50ms，シフト幅 12.5ms，80 次元のメルスペクトログラムの生成を行う．複数話者の音声合成モデルのために，話者 ID による話者埋め込みをエンコーダとデコーダに適用する．学習時には 4,000 イテレーションに対して，warmup を適用する．

提案する強調ネットワークは 6 層，隠れ層 384 次元，FFN1,536 次元，4 つの head を持つ Transformer block で構成し，FastSpeech 2 により予測されたメルスペクトログ

表 1 TED-LIUM 2 の dev セットと test セットの単語誤り率 (%) と音声合成部の生成時間.

手法	dev	test	生成時間
ベースライン: train-clean-100	30.19	27.60	–
拡張モデル: train-clean-100 ＋ TED-LIUM 2 (合成音声)			
メルスペクトログラムそのまま	17.12	17.79	1×
ボコーダによる波形生成	16.71	16.76	2.75×
提案手法	**16.54**	**16.62**	**1.26**×
oracle: TED-LIUM 2 (自然音声)	9.28	8.56	–

表 2 提案手法における音素情報の効果.

手法	dev	test
音素情報あり (F0，パワーあり)	**16.54**	**16.62**
音素情報あり (F0，パワーなし)	**16.52**	16.88
音素情報なし	16.89	17.16

ラムと variance adaptor の出力を用いて，FastSpeech 2 と同様の設定のメルスペクトログラムを生成する．

比較手法のために，VocGAN ボコーダ [31] を用いてメルスペクトログラムから音声波形に変換を行う手法についても実験を行う．オープンソースのコード[*2]を用いて実装し，text-to-mel モデルの学習で使用したデータを用いて，16kHz の音声の波形の変換のために up-sampling rate を 5,5,2,2,2 に変更し，学習した．

4.2.2 Transformer 音声認識モデル

音声認識モデルの入力には (自然音声の場合) 窓幅 25ms，シフト幅 10ms，80 次元のメルスペクトログラムを用いる．音声認識モデルは 2 層の CNN を用いてスーパーフレームを作成する (各 CNN は 2 フレーム毎に 1 フレームのスーパーフレームを作成し，エンコーダの入力時には 1 フレーム毎に，メルスペクトログラムの 4 フレームが含まれる)．エンコーダには 12 層，隠れ層 384 次元，FFN1,536 次元，4 つの head を持つ Conformer [5] で構成する．訓練中にはラベルスムージング [32]，SpecAugment [33]，CTC 損失とのマルチタスク学習を適用する．Conformer の学習中は 25,000 回の更新に対して warmup を適用する．合成音声を用いた学習時には，1 バッチ内の自然音声と合成音声の割合は調整しない．デコード時はビーム幅 10 とし，4 層，512 次元の隠れ層を持つ単方向 LSTM による言語モデルを用いたビームサーチを行う．

4.3 認識実験結果

表 1 に TED-LIUM 2 の dev と test に対する単語誤り率 (%) と音声合成全体のデータ生成にかかった時間を示す．ボコーダを用いた生成時間は，音声波形から再びメルスペクトログラムに変換する時間を含む．合成音声を用いない

[*2] https://github.com/rishikksh20/VocGAN

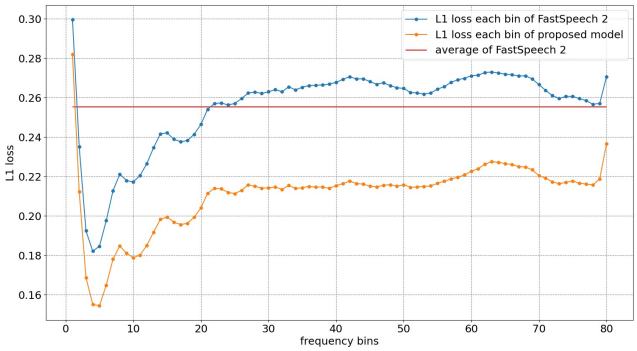

図 2 学習済みの FastSpeech 2(青線) と提案手法 (黄線) の L1 損失と FastSpeech 2 における平均値．低い値のビンは低周波を示す．損失は訓練データからランダムな 1000 サンプルで計算．

場合は LibriSpeech と TED-LIUM 2 のドメインミスマッチがあるため，単語誤り率はかなり悪い．音声合成を用いて，合成音声を用いることで，後処理は行わなくても，dev セットと test セットで相対で 43.3%，35.5%の単語誤り率の改善が見られた．ボコーダを用いることで各評価セットに対して相対で 44.6%，39.3%とさらに改善が見られた．提案手法では，簡潔なネットワークでボコーダよりも速い生成時間で高い結果を得られた (45.2%，39.8%の改善)．

表 2 に音素情報と F0，パワー情報を埋め込んだ variance adaptor の出力を用いた効果を示す．音素情報を用いないモデルでは，FastSpeech 2 により予測されたメルスペクトログラムのみを入力に用いる．このモデルでは音声認識の改善は限定的で，ボコーダを用いた場合と比較して，性能は悪くなった．一方で，F0，パワーの音響情報を用いなくてもほぼ性能に変化はなかった．これらの結果から，音素情報の埋め込みが音声認識の性能の改善に重要であることを確認できた．

図 2 に FastSpeech 2 と提案手法の各周波数ビンに対する L1 損失の値を示す．図から，高周波 (高い値のビン) の損失の値は低周波 (低い値のビン) と比較して高い．これは低周波の方がパワーの値が大きく，音声合成モデルにとっては学習しやすいためと考えられる．損失値から，高周波にも自然音声と合成音声の差があり，高周波を強調した場合にはこれらの差異が埋められたため，音声認識の性能が改善したと考えられる．また，提案手法により FastSpeech 2 と比較して，全ての周波数ビンにおいて L1 損失が改善していることがわかる．

図 3 に強調ネットワークによる強調例を示す．例 1 の赤枠では調波構造がより明確に出ていることがわかる．例 2 の黒枠では F0 のずれがなくなり，より自然になっていることが確認できる．FastSpeech 2 では，1 フレーム毎に予測を行うため，前後のメルスペクトログラムの結果を考慮できないためこのようなずれが発生するが，提案した強調ネットワークでは前後を考慮できるため，より自然な音声になったと考えられる．

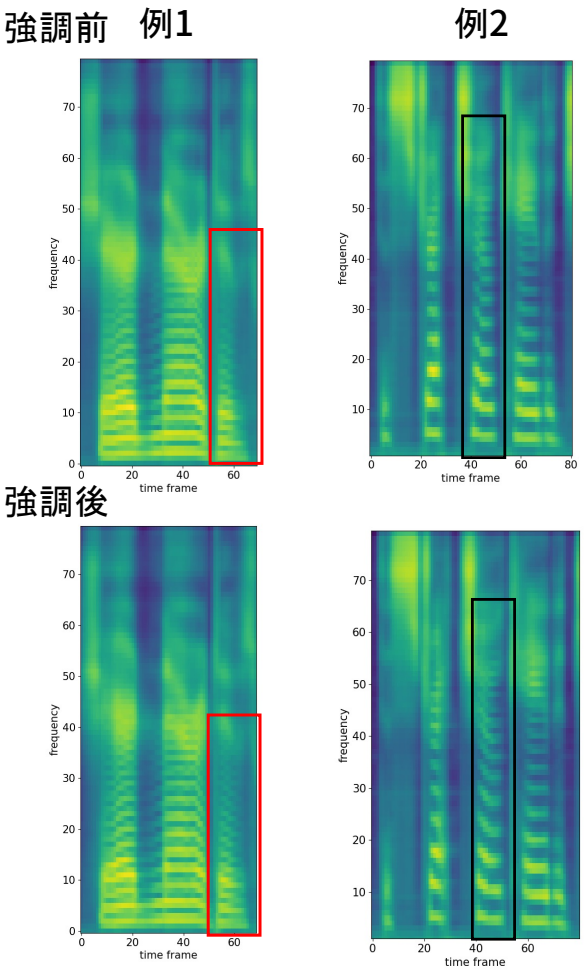

図 3　メルスペクトログラムの強調例.

表 3　『日本語話し言葉コーパス』(CSJ) 中のテストセット 1 (CSJ-APS ドメイン) の単語誤り率 (%) と合成音声の生成時間.

手法	eval1	生成時間
ベースライン: CSJ-SPS	17.09	–
拡張モデル: CSJ-SPS + CSJ-APS (合成音声)		
メルスペクトログラムそのまま	10.19	1×
ボコーダによる音声波形生成	10.09	2.03×
提案手法	**9.74**	**1.26×**
oracle: CSJ-SPS + APS (自然音声)	8.37	–

表 3 に日本語でのドメイン適応の結果を示す．後処理のないモデルでは相対で 40.4%の単語誤り率の改善が見られた．拡張モデルと CSJ-APS の音声データで学習した oracle モデルと比較すると，2 ポイント以下の差しか見られなかった．これは CSJ-SPS の話し方が講演に近く，話題以外のドメインのミスマッチが少ないためと考えられる．提案手法はボコーダと比較して生成時間を短くしつつ，性能の改善が確認できた．

5. おわりに

本研究では，音声合成を用いた音声認識のデータ拡張においてボコーダを用いずに周波数スペクトログラムを強調するネットワークを提案した．提案手法では合成されたメルスペクトログラムを直接強調することで生成時間を短縮することができる．また，提案ネットワークにはメルスペクトログラムだけでなく，音素情報を用いることができる．実験結果から，ボコーダよりも少ない生成時間でデータ拡張の効果を高めること，及び改善において音素情報が重要であることを確認した．

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, *International Conference on Machine Learning (ICML)*, 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, *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 206–213 (2017).

[3] Chorowski, J. and Jaitly, N.: Towards better decoding and language model integration in sequence to sequence models, *INTERSPEECH*, pp. 523–527 (2017).

[4] Dong, L., Xu, S. and Xu, B.: Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition, *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5884–5888 (2018).

[5] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. and Pang, R.: Conformer: Convolution-augmented Transformer for Speech Recognition, *INTERSPEECH*, pp. 5036–5040 (2020).

[6] Rosenberg, A., Zhang, Y., Ramabhadran, B., Jia, Y., Moreno, P., Wu, Y. and Wu, Z.: Speech Recognition with Augmented Synthesized Speech, *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 996–1002 (2019).

[7] Rossenbach, N., Zeyer, A., Schluter, R. and Ney, H.: Generating Synthetic Audio Data for Attention-based Speech Recognition Systems, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7064–7068 (2020).

[8] Chen, Z., Rosenberg, A., Zhang, Y., Wang, G., Ramabhadran, B. and Moreno, P. J.: Improving Speech Recognition Using GAN-Based Speech Synthesis and Contrastive Unspoken Text Selection, *INTERSPEECH*, pp. 556–560 (2020).

[9] Wang, G., Rosenberg, A., Chen, Z., Zhang, Y., Ramabhadran, B., Wu, Y. and Moreno, P.: Improving Speech Recognition Using Consistent Predictions on Synthesized Speech, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7029–7033 (online), DOI: 10.1109/ICASSP40776.2020.9053831 (2020).

[10] Wang, G., Rosenberg, A., Chen, Z., Zhang, Y., Ramabhadran, B. and Moreno, P. J.: SCADA: Stochastic, Consistent and Adversarial Data Augmentation to Im-

prove ASR, *INTERSPEECH*, pp. 2832–2836 (2020).
[11] Laptev, A., Korostik, R., Svischev, A., Andrusenko, A., Medennikov, I. and Rybin, S.: You Do Not Need More Data: Improving End-To-End Speech Recognition by Text-To-Speech Data Augmentation, *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* (2020).
[12] Fazel, A., Yang, W., Liu, Y., Barra-Chicote, R., Meng, Y., Maas, R. and Droppo, J.: SynthASR: Unlocking Synthetic Data for Speech Recognition, *INTERSPEECH*, pp. 896–900 (2021).
[13] Mimura, M., Ueno, S., Inaguma, H., Sakai, S. and Kawahara, T.: Leveraging sequence-to-sequence speech synthesis for enhancing acoustic-to-word speech recognition, *Workshop on Spoken Language Technology (SLT)*, pp. 477–484 (2018).
[14] Ueno, S., Mimura, M., Sakai, S. and Kawahara, T.: Multi-speaker Sequence-to-sequence Speech Synthesis for Data Augmentation in Acoustic-to-word Speech Recognition, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6161–6165 (2019).
[15] Zheng, X., Liu, Y., Gunceler, D. and Willett, D.: Using Synthetic Audio to Improve The Recognition of Out-Of-Vocabulary Words in End-To-End ASR Systems, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5659–5663 (2021).
[16] Kurata, G., Saon, G., Kingsbury, B., Haws, D. and Tüske, Z.: Improving Customization of Neural Transducers by Mitigating Acoustic Mismatch of Synthesized Audio, *INTERSPEECH*, pp. 2027–2031 (online), DOI: 10.21437/Interspeech.2021-1656 (2021).
[17] 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, *INTERSPEECH*, pp. 4779–4783 (2017).
[18] Li, N., Liu, S., Liu, Y., Zhao, S. and Liu, M.: Neural Speech Synthesis with Transformer Network, *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)* (2019).
[19] Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z. and Liu, T.-Y.: FastSpeech: Fast, Robust and Controllable Text to Speech, *Advances in Neural Information Processing Systems*, Vol. 32 (2019).
[20] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z. and Liu, T.-Y.: FastSpeech 2: Fast and High-Quality End-to-End Text to Speech, *International Conference on Learning Representations (ICRL)* (2020).
[21] Elias, I., Zen, H., Shen, J., Zhang, Y., Jia, Y., Weiss, R. and Wu, Y.: Parallel Tacotron: Non-Autoregressive and Controllable TTS, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5694–5698 (2021).
[22] Elias, I., Zen, H., Shen, J., Zhang, Y., Jia, Y., Skerry-Ryan, R. and Wu, Y.: Parallel Tacotron 2: A Non-Autoregressive Neural TTS Model with Differentiable Duration Modeling, *INTERSPEECH*, pp. 141–145 (2021).
[23] Valin, J.-M. and Skoglund, J.: LPCNet: Improving Neural Speech Synthesis Through Linear Prediction, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5891–5895 (2019).
[24] Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brébisson, A., Bengio, Y. and Courville, A. C.: MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis, *Advances in Neural Information Processing Systems*, Vol. 32 (2019).
[25] Kinoshita, K., Delcroix, M., Ogawa, A. and Nakatani, T.: Text-informed speech enhancement with deep neural networks, *INTERSPEECH*, pp. 1760–1764 (2015).
[26] Sennrich, R., Haddow, B. and Birch, A.: Neural Machine Translation of Rare Words with Subword Units, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725 (2016).
[27] Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z. and Wu, Y.: LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech, *arXiv preprint arXiv:1904.02882* (2019).
[28] Panayotov, V., Chen, G., Povey, D. and Khudanpur, S.: Librispeech: An ASR corpus based on public domain audio books, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210 (online), DOI: 10.1109/ICASSP.2015.7178964 (2015).
[29] Morise, M., Yokomori, F. and Ozawa, K.: WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications, *IEICE Transactions on Information and Systems*, Vol. E99.D, No. 7, pp. 1877–1884 (online), DOI: 10.1587/transinf.2015EDP7457 (2016).
[30] Rousseau, A., Deléglise, P. and Estève, Y.: Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pp. 3935–3939 (2014).
[31] Yang, J., Lee, J., Kim, Y., Cho, H. and Kim, I.: VocGAN: A High-Fidelity Real-time Vocoder with a Hierarchically-nested Adversarial Network, *INTERSPEECH*, pp. 200–204 (2020).
[32] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z.: Rethinking the inception architecture for computer vision, *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826 (2016).
[33] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D. and Le, Q. V.: SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition, *INTERSPEECH*, pp. 2613–2617 (2019).