国立国語研究所学術情報リポジトリ

Transposed Convolution-based Articulatory-to-Acoustic Conversion using Real-Time MRI Data

メタデータ	言語: jpn 出版者: 公開日: 2022-01-07 キーワード (Ja): キーワード (En): 作成者: 丹治, 涼, 大村, 英史, 澤田, 隼, 桂田, 浩一, TANJI, Ryo, SAWADA, Shun メールアドレス: 所属:
URL	https://doi.org/10.15084/00003487

転置畳み込みニューラルネットワークを用いた rtMRI データからの調音-音響変換

丹治 涼，大村 英史，澤田 隼，桂田 浩一（東京理科大学理工学部）

Transposed Convolution-based Articulatory-to-Acoustic Conversion using Real-Time MRI Data

Ryo Tanji, Hidefumi Ohmura, Shun Sawada, Kouichi Katsurada (Tokyo University of Science)

要旨

本稿では，rtMRI データから音響特徴量を生成するための深層学習モデルを提案する。調音器官全体を高解像度で記録できる rtMRI は，調音データから音響特徴量を生成するための元データとして有用であると考えられるが，フレームレートが比較的低いという問題がある。そこで我々は，転置畳み込みネットワークを用いて時間軸方向に超解像処理を行う方法を提案する。標準的な畳み込みニューラルネットワークが畳み込みによって主に画像の近隣情報を圧縮するのに対して，転置畳み込みネットワークではこの逆の操作を行うことにより，画像の解像度を向上させる。本手法ではこの超解像処理を rtMRI データの時間方向に適用することによって，rtMRI データの時間解像度を向上させる。メルケプストラム歪みと PESQ を評価尺度として用いた実験の結果，転置畳み込みネットワークは正確な音響特徴量の生成に有効であることがわかった。また，超解像処理の倍率を上げることで，PESQ のスコアが向上することも確認した。

１．はじめに

発声時の舌や口唇，顎などの動作を調音運動という。調音‐音響変換は収録された調音運動のデータから音響特徴を生成する研究分野である。一般的には何らかの機械学習器を用いて調音運動データと音響特徴量の間の統計的関係をモデル化することが多い(Richmond et. al. 2015)が，特に近年では深層学習の発展に伴い，DNN (deep neural network)を用いた手法が数多く提案されて始めている(Katsurada and Richmond 2020, Taguchi and Kaburagi 2018, Csapó et. al. 2017, Gonzalez et. al. 2017, Liu et. al. 2018)。

調音運動の測定方法としては EMA (electromagnetic articulography) (Schönle et. al. 1987)，超音波画像(Akgul et. al. 1998)，PMA (permanent magnetic articulography) (Fagan et. al. 2008)，sEMG (surface electromyography) (Hermens et. el. 1999)，口唇動画像(Akbari et. al. 2018)，rtMRI(real time magnetic resonance imaging) (Narayanan et. al. 2014)等がしばしば用いられる。中でも rtMRI は高い解像度で調音器官全体を収録できることから，発声分析等での活用が始まっている(Ramanarayanan et. al. 2018, Toutios et. al. 2019)。調音‐音響変換においても，超音波画像と比べて rtMRI からの変換の方が高い精度が得られる(Csapó 2020)ことが示されており，rtMRI は調音‐音響変換のための元データとしてもその有用性が期待できる。しかし rtMRI はその収録装置の制限により，他の調音運動収録機器と比較して高いサンプリングレートでの収録が困難であるという問題がある。音声合成ではフレーム間隔が長いと合成音の品質が低下する (Kitamura et. al. 1985, Miyashita and Morise 2018)ことから，何らかの方法によってサンプリングレートを向上することが rtMRI からの調音‐音響変換の性能向上には必要であると考えられる。

そこで本稿では転置畳み込み（逆畳み込み）ネットワーク(Dumoulin and Visin 2016)によって精度の良い調音‐音響変換を行う方法を提案する。転置畳み込みネットワークは一般的

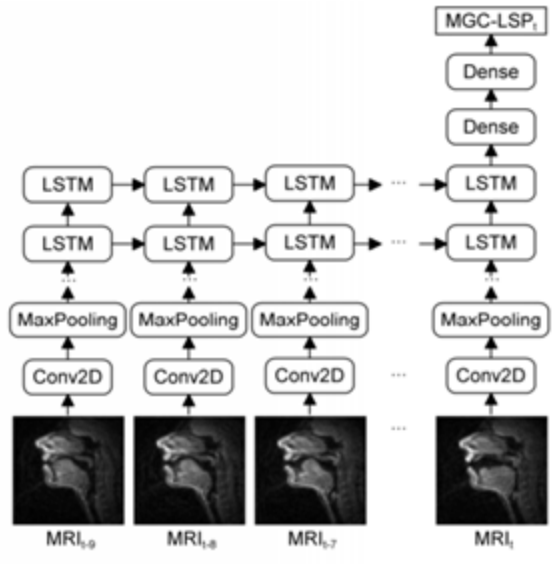

図 1 CNN-LSTM のネットワーク構成(Csapó 2020)

な畳み込みと逆の操作を行うニューラルネットワークであり，画像の解像度を上げる超解像処理等でしばしば用いられる(Long et. al. 2015, Radford et. al. 2015, Dong et. al. 2016)。近年では音声の分野においても音声強調等で用いられ始めており(Pascual et. al. 2017)，調音 - 音響変換においても効果的であることが予想できる。本稿では転置畳み込みネットワークを用いた調音 - 音響変換の精度を検証するため，日本語 rtMRI データセット(Maekawa 2019)を用いて変換性能を評価する。また，ストライド幅を変えることによって超解像処理の倍率を変化させ，その変化が調音 - 音響変換に与える影響を検証する。

２．関連研究

Csapó は rtMRI データからの調音 - 音響変換を，(1)全結合ニューラルネットワーク，(2)CNN，(3)CNN-LSTM，の 3 種類のネットワークで実装し，(3)の CNN-LSTM の精度が最も良いことを示した(Csapó 2020)。CNN-LSTM モデルでは CNN で実装された深層画像特徴抽出器と LSTM で実装された時系列データ生成器を組み合わせることで調音 - 音響変換を実現している。このネットワークは Csapó らが以前に提案した超音波画像からの調音 - 音響変換器(Csapó et. al. 2017)と同じ構造の物であり，rtMRI データからの変換においても有効であることを示している。

CNN-LSTM のネットワーク構造の詳細を図 1 に示す。CNN は 3 層の畳み込み層（フィルタサイズ 3×3，ストライド 1×1，フィルタ数 8，16，32）から構成され，それぞれの層の後にプーリング層（フィルタサイズ 2×2，ストライド 2×2）が設置されている。CNN の出力は 2 層からなる LSTM（ユニット数 512）に送られ，最後に 2 層の全結合層を通して 40 次元の音響特徴量が生成される。全ての層に適用される活性化関数は ReLU である。

Csapó は声道形状を表す低次元パラメータである MGC-LSP を音響特徴量としてネットワークで推定しており，それを高次元の声道形状パラメータであるメルケプストラムに変換した上でオリジナル音声のメルケプストラムとの差（メルケプストラム歪み）を評価している。その結果，CNN-LSTM モデルが最もメルケプストラム歪みが小さいことを示しており，時系列情報を処理する LSTM の有効性を示している。また，先行研究である超音波画像から生成されたメルケプストラムと比較して，rtMRI データから生成されたメルケプストラムの歪みが小さいことも示しており，rtMRI からの調音 - 音響変換の有効性を示している。

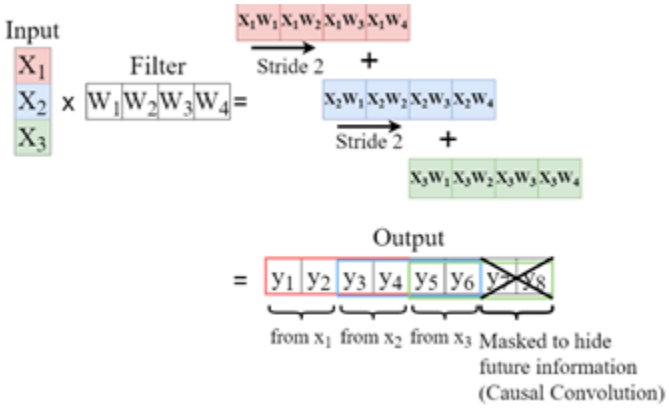

図 2 転置畳み込みネットワーク

３．転置畳み込みネットワークを用いた調音 - 音響変換

調音運動から音響特徴を生成する最もシンプルな方法は，フレームごとに音響特徴量を生成する方法である。この方法を採った場合，生成される音響特徴量の時間方向の解像度は調音運動データの時間解像度と同一になる。このため rtMRI の時間解像度が十分でない場合，CNN-LSTM の生成する音響特徴のフレーム時間間隔は大きな値となる。フレーム時間間隔が大きいとメルケプストラムから生成される音声の品質低下を招くため，何らかの改善策を検討することが望ましい。その一つの方法が時間方向に超解像処理のような手法を導入することである。時間方向の超解像処理でフレーム間隔を補うことによって，合成音の品質を改善できる可能性がある。そこで我々は転置畳み込みネットワークを用いた時間方向の超解像処理によって音響特徴量のフレーム時間間隔の短縮を目指す。

転置畳み込みネットワークは，一般的な畳み込みニューラルネットワークの解像度削減のプロセスを逆方向に行うネットワークである。逆向方向の処理によって解像度が増大するため，超解像処理が実現できる。図 2 に転置畳み込みネットワークの例を示す。この図では入力ベクトルのサイズが 3 で，フィルタのサイズが 4，ストライドが 2 となっている。転置畳み込みの処理は次の通りである。まず，入力データの各要素とフィルタベクトルの積が計算され，一つの要素につき一つのベクトルが生成される。次に，得られた各ベクトルをストライドの大きさだけシフトしながら加算し，一つのベクトルを形成する。最後に，生成されたベクトルの末尾からフィルタサイズとストライドの差の大きさの要素が削除され，得られたベクトルが転置畳み込み処理後のベクトルとなる。以上の処理により最終的に得られるベクトルの大きさは，入力ベクトルにストライドサイズを積算した大きさとなる。ベクトルの加算の過程で近接する過去の入力情報を現在の出力に反映できるため，転置畳み込みネットワークは近接する時系列データの関係をモデル化していると捉えることもできる。

本稿で提案する rtMRI 動画像からの調音 - 音響変換のネットワーク構成を図 3 に示す。ネットワークは基本的に Csapó のモデル(3)の CNN と LSTM の間に 1 層の畳み込み層および max-pooling 層と転置畳み込みネットワークを挿入した形になっている。畳み込み層と max-pooling 層の追加により，rtMRI 動画像と音響特徴の間のより複雑な対応関係を学習できるようにし，また転置畳み込みネットワークに送られるデータの次元圧縮も実現している。これに加えて，提案するネットワークでは LSTM の後の 2 層の全結合層をなくして，代わりに 1 層の線形変換層を追加しており，ネットワークから直接メルケプストラムを出力する形にしている。これらの工夫によって本手法では精度の向上とネットワーク全体のパラメータ数の削減を実現している。

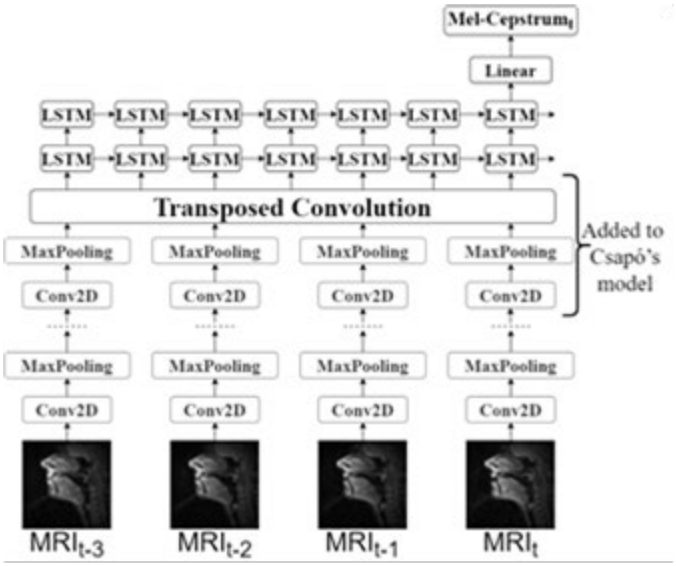

図 3 提案するネットワーク（CNN-TC-LSTM）

４．実験設定

４.１　データセット

本研究では日本語 rtMRI データセット(Maekawa 2019)を用いて調音 - 音響変換の性能を評価する。このデータセットは 1 モーラの日本語 103 個，代表的な 2 モーラの日本語の組み合わせ 676 個，および 100 モーラの発話を含んでいる。rtMRI 動画像の解像度は 256×256 ピクセルで，1 ピクセルが 1mm に相当する。時間解像度は約 13.72fps で，音声のサンプリングレートは 44.1kHz である。実験では，動画像から調音器官全体が含まれる 150×150 ピクセルの領域を抽出して利用した。また，音声は 16kHz にダウンサンプリングした。音声には MRI 機器から発生する高いレベルのノイズが含まれていたため，合成音の品質を向上するために事前にスペクトルサブトラクション法(Boll 1979)によるノイズ削除を行った。なお, スペクトルサブトラクション法における係数は一般的に 1.2～1.4 程度に設定されるが，本データセットではノイズレベルが非常に高かったため 4.0 に設定した。実験ではデータセット内から男性話者 2 名，女性話者 2 名を選択して学習及び評価を行った。

４.２　画像の前処理

RtMRI 動画像は多くのノイズを含むため，本研究では non-local means デノイジング(Buades et. al. 2005)を用いてノイズ除去を行った。Non-local means フィルタは対象画素と周辺画素の類似性を重みとして重み付き平均を取るフィルタであり，エッジの保存性が良いという特徴がある。こうした特徴が他のノイズ除去法と比べて rtMRI 動画像からの調音器官の境界抽出等に適していると考え，本研究ではこの手法を用いることにした。

４.３　音響特徴の抽出

音響特徴量の抽出には World(Morise 2016)のモジュールである Harvest，CheapTrick，D4C を用いた。それぞれ基本周波数の抽出，メルケプストラムの抽出，および非周期性指標の抽出に用いている。実験ではこのうちメルケプストラムのみを教師データとして用い，その他の音響特徴量（基本周波数および非周期性指標）は合成音の生成の際にそのまま用いることにした。

表 1 各ネットワークのメルケプストラム歪み(dB)

Stride size (frame shift)	CNN-LSTM	CNN-TC-LSTM (-1)	CNN-TC-LSTM
1 (72.5 ms)	7.06	6.16	6.18
2 (36.3 ms)		6.14	6.13
4 (18.1 ms)		**6.12**	6.12
8 (9.1 ms)		6.15	**6.08**

４.４　実験の詳細と比較対象のネットワーク構成

実験では話者 1 名のデータのうち 80%を訓練データとして用い，残りのうち 10%ずつをそれぞれ検証用データ，評価用データとした。入力の MRI 動画像は平均 0，分散 1 に正規化した。CNN と転置畳み込みネットワークの重みの初期値は He らの初期設定(He et. al. 2015)と同様にした。LSTM の重みは一様分布とした。最適化手法には Adam を用い，学習率は 0.0001 の固定値とした。学習回数は 100 エポックとした。損失関数には，本研究の評価手法でも用いるメルケプストラム歪み(Kubichek 1993)を用いた。転置畳み込みネットワークのフィルタサイズは 6 とし，フィルタ数は 128 とした。

評価したネットワークは 2 節で説明したモデル（CNN-TC-LSTM）である。このモデルでは転置畳み込みネットワークの有無に加えて CNN の層の数と全結合層の数が Csapó のモデル(3)のネットワークと異なる。そこで転置畳み込みネットワークのみの有効性を検証するため，転置畳み込みネットワークの有無以外の部分が Csapó のモデル(3)と同様であるネットワーク（CNN-TC-LSTM (-1)）を用意して，同時に評価を行った。転置畳み込みネットワークのストライドサイズの影響を検証するため，ストライドが 1，2，4，8 の場合を比較した。さらに，Csapó のモデル(3)の出力をメルケプストラムにしたネットワーク（CNN-LSTM）を用意してベースラインとすることにした。

４.５　評価指標

生成したメルケプストラムの正確さを客観的に評価するためにメルケプストラム歪み(Kubichek 1993)を評価指標として用いた。これに加えて，生成した音声の音質を評価するために PESQ スコア(Rix et. al. 2001)を用いた。PESQ は，生成したメルケプストラムから合成された音声全体を用いて算出される。この指標を用いた理由は，メルケプストラム歪みが同じであっても音質が大きく異なる場合があるためである。メルケプストラムはフレームごとに算出されるため，メルケプストラム歪みもフレームごとに算出される。このためフレームとフレームの間の合成音の品質はメルケプストラム歪みでは評価できない。これを補うために，音声全体の品質を評価できる PESQ スコアを評価指標として用いることにした。

５．実験結果と考察

生成されたメルケプストラムと正解のメルケプストラムのメルケプストラム歪みを表 1 に示す。表に示す通り，CNN-TC-LSTM モデルは CNN-LSTM モデルと比較してメルケプストラム歪みを 0.81dB 低減できていることが分かる。一方で，結果が示すようにストライドサイズはメルケプストラム歪みに大きな影響を与えていないことも確認できる。CNN-TC-LSTM と CNN-TC-LSTM(-1)を比較すると，CNN-TC-LSTM の方が若干メルケプストラム歪みを低減できているがほとんど同様の結果であることが分かる。CNN-TC-LSTM と CNN-TC-LSTM(-1)の違いは，CNN-TC-LSTM の方が畳み込み層と max-pooling 層が 1 層多く，一方で 2 層の全結合層が削除されて線形変換層が追加されていることである。これらの変更はメルケプストラムの推定に大きな影響を及ぼしていないことが分かる。

表 2 各ネットワークの PESQ スコア

Stride size (frame shift)	CNN-LSTM	CNN-TC-LSTM (-1)	CNN-TC-LSTM
1 (72.5 ms)	0.83	0.82	0.80
2 (36.3 ms)		1.47	1.42
4 (18.1 ms)		1.91	1.90
8 (9.1 ms)		**2.01**	**2.02**

表 3 各ネットワークのパラメータ数

Stride size	CNN-LSTM	CNN-TC-LSTM (-1)	CNN-TC-LSTM
1	24.9 M	11.4 M	5.4 M
2		19.3 M	7.4 M
4		35.2 M	11.4 M
8		67.1 M	19.3 M

ストライドサイズが 1 のとき，CNN-LSTM と CNN-TC-LSTM(-1)の唯一の違いは転置畳み込み層の有無である。転置畳み込み層ではフィルタリングとベクトルの加算の処理だけしか行われていないにもかかわらず，メルケプストラム歪みはこれら二つのネットワークで大きく異なる。フィルタリングに相当する処理はネットワーク内の他の層でも行われているため，この差はベクトルの加算によって生じたと考えられる。3 節でも述べたように，加算の処理は近接した時系列データの関係をモデル化しているという特徴がある。CNN-LSTM モデルでは LSTM ネットワークで時系列情報を扱っているが，転置畳み込みネットワークでは過去数フレーム分のデータの関係を明示的に現在の出力に反映している。我々はこの差がメルケプストラム推定の大きな差に繋がったと考えている。そこでこれを検証するため，CNN-LSTM モデルに含まれる 2DCNN を，近接する時系列データを直接扱う 3DCNN に変更して実験を行った。その結果，メルケプストラム歪みは 5.87dB となり，CNN-TC-LSTM モデルに近い結果を得ることができた。この結果から，時間的に連続する数フレームの rtMRI 画像を入力として低レイヤーで処理することが精度の良いメルケプストラムの推定に重要であると分かった。

表 2 に示すのはメルケプストラムから作成した合成音の PESQ スコアである。表に示す通り，CNN-TC-LSTM(-1)と CNN-TC-LSTM の各モデルはストライドサイズが 2 以上のときに CNN-LSTM の音質を上回っていることが分かる。また，表 1 のメルケプストラム歪みの評価ではストライドサイズが結果にほとんど影響しなかったのに対して，PESQ スコアはストライドサイズが大きいほど向上していることが分かる。これはメルケプストラム歪みと PESQ スコアの計算方法の違いに起因すると考えられる。メルケプストラム歪みはフレームごとに算出されるため，フレームごとのメルケプストラムの推定値が同程度に正確であればフレームごとのメルケプストラム歪みの値は変化しない。これに対して PESQ スコアは合成された音声全体に対して別途細かなフレームシフトを行って算出される。ストライドサイズを大きくすることによって超解像処理の倍率が高まり，結果としてメルケプストラム推定のフレーム間隔が短くなり，MRI 画像間の音声全体の品質が向上したために PESQ スコアが向上したと考えられる。

以上の結果から，CNN-TC-LSTM モデルと CNN-TC-LSTM(-1)モデルは CNN-LSTM モデルと比べて性能が高いことが分かった。しかし CNN-TC-LSTM モデルと CNN-TC-LSTM(-1)モデルの性能は大きく変わらないといえる。CNN-TC-LSTM モデルと CNN-TC-LSTM(-1)モデルの最大の相違点はネットワーク内のパラメータ数の差である。表 3 にネットワーク内のパラメータ数を示す。 3 節でも述べたように，転置畳み込みネットワークと

LSTM のノード数が 1 層の畳み込み層と max-pooling 層の追加によって次元圧縮され，大幅に削減される。CNN-TC-LSTM モデルは CNN-TC-LSTM(-1)モデルおよび CNN-LSTM モデルと比べてパラメータ数が少ないことから，rtMRI からの調音 - 音響変換には CNN-TC-LSTM モデルが最適であると考えられる。

６．おわりに

本稿では rtMRI 動画像から調音 - 音響変換を行うための深層学習モデルを提案した。rtMRI の時間解像度が十分でないことに対応するため，転置畳み込みネットワークを用いた超解像処理によってサンプリングレートを向上させる方法を示した。提案したモデルを，転置畳み込みネットワークを含めないモデルと比較した結果，声道形状パラメータの差異を表すメルケプストラム歪み，および音声の品質を表す PESQ スコアの双方について本稿で提案したモデルの生成する音声の品質が高いことが分かった。提案モデルはネットワークに含まれるパラメータ数という点でも軽量であることから，rtMRI からの調音 - 音響変換および音声合成において転置畳み込みネットワークを用いた本モデルは有効性が高いと考えられる。本システムでは声道形状パラメータであるメルケプストラムのみを推定しているが，今後は音声を直接合成する end-to-end システムの構築を進めたい。

謝　辞

本研究は 2021 年度科学研究費基盤研究(C) 19K12024 および 2021 年度科学研究費基盤研究(B) 20H01265 の補助により行われたものである。

文　献

P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader and B. Conrad, “Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract,” Brain and Language, pp. 26–35, 1987.

Y. Akgul, C. Kambhamettu and M. Stone, “Extraction and tracking of the tongue surface from ultrasound image sequences,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.298–303, 1998.

M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman, “Development of a (silent) speech recognition system for patients following laryngectomy,” Medical Engineering and Physics, vol. 30, no. 4, pp. 419–425, 2008.

H. J. Hermens, B. Freriks, R. Merletti, D. Stegeman, J. Blok, G. Rau, C. Disselhorst-Klug, and G. Hägg, “European recommendations for surface electromyography,” Roessingh research and development 8.2, pp.13–54, 1999.

H. Akbari, H. Arora, L. Cao, and N. Mesgarani, “LIP2AUDSPEC: Speech reconstruction from silent lip movements video,” in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing., Calgary, Canada, 2018, pp. 2516–2520.

S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis, and M. Proctor, “Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC),” The Journal of the Acoustical Society of America, vol. 136, no. 3, pp. 1307–1311, Sep 2014.

V. Ramanarayanan, S. Tilsen, M. Proctor, J. Töger, L. Goldstein, K. S. Nayak, and S. Narayanan, “Analysis of speech production real-time MRI,” Computer Speech and Language, vol. 52, pp. 1–

22, 2018.
A. Toutios, D. Byrd, L. Goldstein, and S. Narayanan, “Advances in vocal tract imaging and analysis,” in The Routledge Handbook of Phonetics. Taylor and Francis, Jan. 2019, pp. 34–50.
K. Richmond, Z. Ling, and J. Yamagishi, “The use of articulatory movement data in speech synthesis applications: An overview—application of articulatory movements using machine learning algorithms—,” Acoustical Science and Technology 36.6, pp. 467–477, 2015.
K. Katsurada and K. Richmond, “Speaker-Independent Mel-Cepstrum Estimation from Articulator Movements Using D-Vector Input,” in Proc. INTERSPEECH, 2020, pp. 3176–3180.
F. Taguchi and T. Kaburagi, “Articulatory-to-speech conversion using bi-directional long short-term memory,” in Proc. INTERSPEECH, Hyderabad, India, 2018, pp. 2499–2503.
T. G. Csapó, T. Grósz, G. Gosztolya, L. Tóth, and A. Markó, “DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface,” in Proc. INTERSPEECH, Stockholm, Sweden, 2017, pp. 3672–3676.
J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, “Direct Speech Reconstruction from Articulatory Sensor Data by Machine Learning,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, pp. 2362–2374, Dec. 2017.
Z. C. Liu, Z. H. Ling, and L. R. Dai, “Articulatory-to-acoustic conversion using BLSTM-RNNs with augmented input representation,” Speech Communication, vol. 99, pp.161–172, 2018.
T. G. Csapó, “Speaker Dependent Articulatory-to-Acoustic Mapping Using Real-Time MRI of the Vocal Tract,” in Proc. INTERSPEECH, 2020, pp. 2722–2726.
T. Kitamura, S. Imai, C. Furuichi, and T. Kobayashi, “Speech analysis-synthesis system and quality of synthesized speech using mel-cepstrum,” Transactions of the Institute of Electronics and Communication Engineers of Japan. A, vol. 68, pp. 957–964, Sep. 1985.
G. Miyashita and M. Morise, “Influence of frame shift in speech parameters on sound quality by high–quality speech analysis/synthesis system,” IEICE Technical Report. vol. 117, no. 393, SP2017–72, pp. 35–38, Jan. 2018.
V. Dumoulin and F. Visin, “A guide to convolution arithmetic for deep learning,” arXiv preprint arXiv:1603.07285, 2016.
J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015, pp. 3431–3440.
A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” arXiv preprint arXiv:1511.06434, 2015. 2, 3.
C. Dong, C. C. Loy, and X. Tang, “Accelerating the Super-Resolution Convolutional Neural Network,” in Proc. European Conference on Computer Vision (ECCV), Amsterdam, Netherlands, 2016, pp 391–407.
S. Pascual, A. Bonafonte, and J. Serr, “Segan: Speech enhancement generative adversarial network,” in Proc. INTERSPEECH, 2017, pp. 3642–3646.
E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, “Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans,” The Journal of the Acoustical Society of America, pp. 1791–1794, 2006.

K. Maekawa, “A real-time MRI study of Japanese moraic nasal in utterance-final position,” in Proc. International Congress of Phonetic Sciences (ICPhS), 2019, pp. 1987–1991.

S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Transactions on acoustics, speech, and signal processing, Vol.27, pp.113–120, 1979.

A. Buades, B. Coll, and J. M. Morel, “A non-local algorithm for image denoising,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA 2005, pp. 60–65.

M. Morise, F. Yokomori, and K. Ozawa, “WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications,” IEICE Transactions on Information and Systems, vol. 99, pp. 1877–1884, 2016.

K. He, X. Zhang, S. Ren, and J. Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification,” IEEE International Conference on Computer Vision (ICCV), pp. 1026–1034, 2015.

R. F. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing., Victoria, Canada, 1993, pp. 125–128.

A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, USA, 2001, pp. 749–752.