LipNet 構造を用いた唇画像から音声への変換*

☆伊藤大貴 (神戸大), 滝口哲也 (神戸大/JST さきがけ), 有木康雄 (神戸大)

1 はじめに

本稿では，音声情報が含まれていない唇動画像からその唇の動きに対応した音声へと変換する手法を提案する．この手法は，画像特徴量の情報を異なる特徴量の情報へと変化させる試みである．一般に，音韻情報を知覚するには，聴覚情報を含んだ音声だけではなく，発話者の顔や唇の動きなどの視覚情報も影響していることが McGurk らにより報告されている [1].

唇動画像から得られる言語情報は，音声発話と比べて非常に少ないため，本研究は困難なタスクであると考えられるが，音声の一部が欠落しているような動画の音声復元，雑音環境下における音声情報の理解や発話障害者のためのコミュニケーションツールの開発といった様々な応用が期待できる．

本研究のアプローチには，Lipreading と text-to-speech(TTS) を組み合わせた手法も考えられるが，誤って認識した場合に意図したものと異なる内容が発話される可能性があることから，本稿では，発話している無音の唇動画像からテキスト情報を介さずに直接音声へと変換する方法を用いる．この手法では，必要な情報は唇画像と音声の 2 種類のみであり，テキスト情報を用いて認識する必要がないので，誤認識することなく期待する発話が得られると考えられる．

テキスト情報を用いない類似研究として，声質変換が挙げられる．声質変換とは，入力した音声の音韻性などの言語情報は保ったまま，話者性などの非言語情報を変換する技術であり，近年では Deep Neural Network(DNN) を用いた声質変換が高い変換精度を持つことから広く利用されている．

また，非負値行列因子分解を用いた唇動画像からの数字発話による音声生成において有効性が示されていたり [3]，難聴障害者のコミュニケーション支援技術として，本タスクとは逆問題にあたる音声から口唇動作の生成が関連研究として挙げられる [4]．ここでは，隠れマルコフモデル (Hidden Markov Model: HMM) や DNN などの様々なモデルが適用されており，有効性が示されている．

唇画像から音声への変換方法として，これまでに GMM [5] や DNN を用いた変換方法を提案してきたが，変換音声はまだまだ不十分であった．そこで本稿では，DNN モデルの改良方法として，LipNet [6] 構造を用いた唇画像から音声への変換を提案する．また，従来手法においては，STRAIGHT [7] を用いて抽出される特徴量のうちスペクトル特徴量のみを変換していたが，本稿では基本周波数も変換の対象とし，より精度の高い音声変換システムの構築を目指す．以上の方法を用いて得られた特徴量と音声に対して評価実験を行い，本手法の有効性を検討する．

以下，第 2 章で従来の音声変換システムについて示し，第 3 章で提案手法，第 4 章で評価実験とその結果について示し，第 5 章で本稿をまとめる．

2 従来手法

筆者らはこれまでに，Fig. 1 に表される DNN モデルを用いたスペクトル推定を行ってきた．DNN のモデルパラメータは，正解スペクトルと推定スペクトルの差を客観的に評価するメルケプストラム歪みが最もよくなるように調整していたが，音声は客観的な指標だけで判断することは難しいので，従来の DNN モデルが精度が良いモデルであるとは言い難かった．

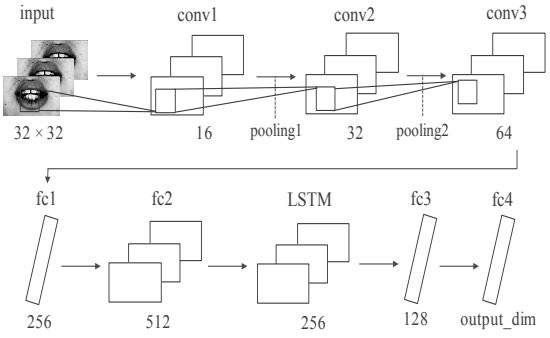

Fig. 1 Conventional DNN model

3 提案手法

3.1 音声変換システム

Fig. 2 は本稿の音声変換システムの概要である．本手法では，画像と音声を抽出し前処理を行うことでそれぞれ DNN の入力，教師データとしている．以下に画像と音声特徴量の抽出方法を述べる．

画像特徴量抽出方法は以下の通りである．顔画像から対象領域 (Region of Interest : ROI)，つまり唇

*Lip image to speech conversion using LipNet, by Daiki Ito (Kobe univ.), Tetsuya Takiguchi (Kobe univ./ JST PRESTO), Yasuo Ariki (Kobe univ.)

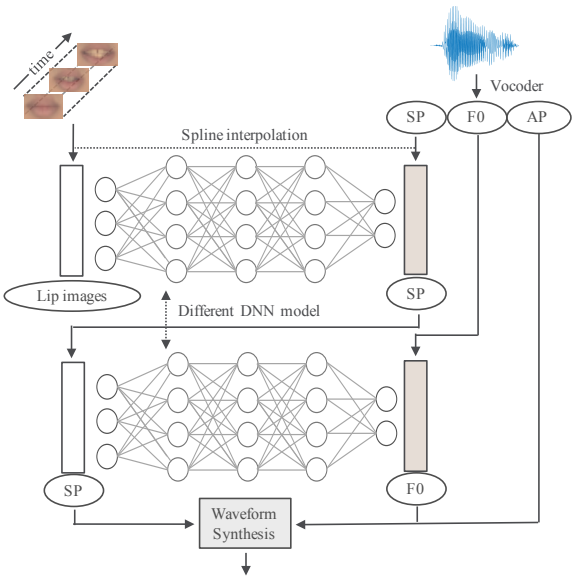

Fig. 2　Flow of proposed speech conversion system

画像領域のみを抽出する．得られた唇画像の取りうる値の範囲は 0 〜 255 であるので，すべての要素に対して 255 で割ることで取りうる値の範囲を 0 〜 1 に正規化している．さらに，音声特徴量とのフレーム間同期をとるためにスプライン補間を適用することで，画像特徴量 $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_T]$ が得られる．

音声特徴量は，STRAIGHT を用いてスペクトル，F0，非周期性指標を抽出した．STRAIGHT によって抽出するスペクトルは高次元であるため，学習には低次元で扱えるメルケプストラム 40 次元 $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_T]$ に変換して用いる．F0 については，抽出されたものを，log スケールに変換して学習に用い，非周期性指標に関してはそのまま用いるものとする．

これらの方法で抽出した画像特徴量からスペクトル特徴量への非線形関数を DNN のパラメータとして学習する．画像特徴量からスペクトル特徴量を得る時に，LipNet 構造のネットワークを用いることで唇特徴の分析精度を上げ，スペクトルへと回帰する．この学習して得られた LipNet 構造型の DNN に，オープンの画像特徴量を入力すると，スペクトル特徴量を推定することが出来る．

続いて，学習時には正解のスペクトル特徴量から F0 特徴量への非線形関数を画像からスペクトル特徴量とは異なる DNN モデルのパラメータとして学習する．推定には上記の画像特徴量から推定されたスペクトル特徴量を入力すると，F0 を推定することが出来る．スペクトルから F0 の推定には，隠れ層 256 の LSTM2 層を用いてパラメータ学習を行った．

3.2　LipNet

LipNet [6] は，end-to-end なモデルで文単位の読唇術を可能にした DNN モデルである．本稿では LipNet の出力層以外の部分を用いることで DNN 回帰モデルを生成した．Fig. 3 に本稿で用いる DNN モデルの概要を示す．

LipNet 構造の DNN モデルへの入力は画像特徴量 $\mathbf{X}$，教師データにはスペクトル特徴量 $\mathbf{Y}$ を用いてモデル学習し，スペクトル特徴量を推定する．入力画像は RGB3 チャネルの画像を用いるので，CNN は 3 次元構造の STCNN(Spatiotemporal convolutional neural networks) を用いており，CNN により学習されたパラメータ調整に用いる Pooling 層や，Dropout 層も 3 次元構造のネットワークになるように調整されている．

LipNet は読唇術に用いられており，一般に読唇術とは，唇画像から何と発話したのかを認識するものであったので，主に STCNN が唇画像の特徴を解析するために優れたモデルであると考えられる．さらに，双方向 GRU を用いることで，単語レベルではなく文章レベルの発話を認識するのに頑健なネットワークを構成していると考えられるので，本稿においては LipNet 構造のネットワークを用いている．

LipNet 構造モデルの構成について説明する．カーネルフィルタサイズがそれぞれ 32，64，96 である 3 次元畳み込み層が 3 つ連続し，カーネルフィルタの大きさはすべてそれぞれ，3 × 5 × 5，3 × 5 × 5，3 × 3 × 3 である．また，それぞれの畳み込み層は Pooling 層により層を縮小している．次に隠れ層の数が 256 の双方向 GRU 層が 2 層用いられる．活性化関数として線形関数を用いることで，出力層からスペクトル特徴量が Fig.3 の Output で得られる構成になっている．

4　評価実験

4.1　実験条件

男性 1 名の文章発話音声とその動画像が含まれる M2TINIT [8] を用いて評価実験を行った．収録には，ATR 研究用日本語音声データベースセット [9] の音素バランス文 503 文が使用されている．DNN の学習に用いる学習データ，DNN の各 epoch におけるモデル性能を評価する検証データ，結果を確認するための学習に用いていないテストデータをそれぞれ，460 文，20 文，23 文として使用した．

収録されている動画像データは唇から鼻先の範囲が映っており，画像のフレームレートは 1/29.97s ，元

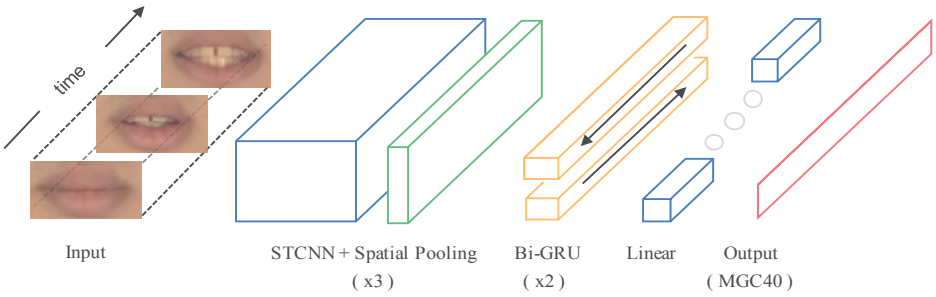

Fig. 3　LipNet architecture

画像全体のサイズは 720 × 480 ピクセルでその中から対象領域となる唇画像を抽出し，抽出された画像をさらに 60 × 30 ピクセルにリサイズして用いた．また，音声データと画像データをパラレルなデータとして用いるためにスプライン補間を適用した．

音声発話データのサンプリング周波数は 16kHz，フレームシフトは 5ms で，特徴量抽出には STRAIGHT を用いた．STRAIGHT によりスペクトル，F0，非周期性指標が抽出され，そのうちスペクトルは，STRAIGHT スペクトル 1025 次元を変換したメルケプストラム 40 次元の特徴量を用いる．このメルケプストラム，F0 特徴量は，平均 0 分散 1 に正規化している．

本実験では，提案手法の有効性を図るために，客観評価と主観評価を行う．客観評価では，スペクトル推定の評価として，メルケプストラム歪み (Mel-cepstrum Distortion: MelCD) により従来手法との比較を行った．

F0 推定の評価には，RMSE による精度検証を行う．従来手法においては，F0 推定を行っていないので，比較として，正解のスペクトルから F0 を推定した場合と，推定されたスペクトルから F0 を推定した場合を用いる．この正解のスペクトルを与える理由として，画像からスペクトルの推定がうまくいっているかどうかの指標にもなり，また正解のスペクトルから F0 が精度よく推定できれば，本稿で提案している F0 推定の有効性を示せるという考えからである．

主観評価は成人男女 7 名に対して，ヘッドホンを用いた聴取実験を行った．テストデータ 23 文の中から客観評価において結果の良かった 10 文について評価実験を行い，文章中で聞き取れた文節の個数の割合の評価と，従来手法との音声の優位性を示す一対評価を行った．

4.2　実験結果・考察

Fig. 4 に客観評価によるメルケプストラム歪みの実験結果を示す．

Fig. 4 より，提案手法の DNN モデルを用いてスペクトル推定したときのほうが精度が良いことがわかる．

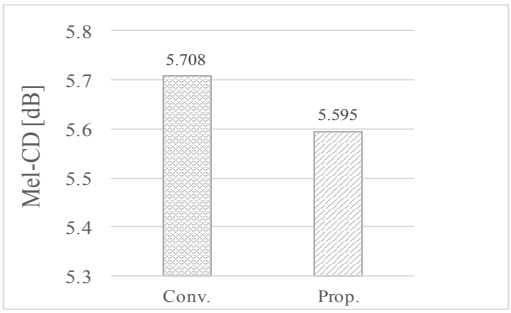

Fig. 4　Objective evaluation as melCD

次に，F0 推定による RMSE の結果を Fig. 5 に示す．Fig. 5 の左は，推定されたスペクトルから F0 を推定した場合で，右は正解のスペクトルから F0 を推定したときの RMSE の結果である．

推定されたスペクトルから F0 を推定した場合，RMSE が大きいことから，推定されたスペクトルは完全なものではないが，正解のスペクトルから F0 を推定すると，RMSE がかなり抑えられていることから，スペクトル変換がうまくいくと F0 の推定も有効性が示せると考察できる．

続いて聞き取れた音節の個数の割合を示す主観評価実験の結果を Table 1 に示す．本実験において，本

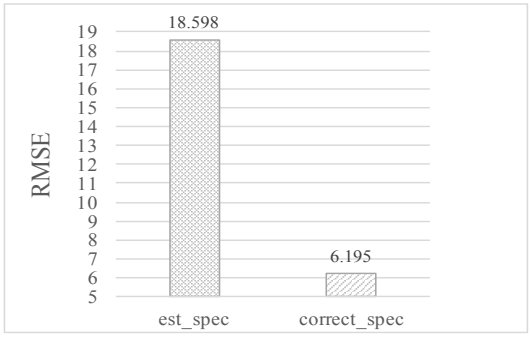

Fig. 5　RMSE-F0 using correct SP and estimated SP

来のシステムとしてはF0は推定したスペクトルから求めるべきだが，RMSEの精度を考慮して，正解のスペクトルから推定したF0を元に音声を生成した．

Table 1　Correct answer rate (%) for each subject

Subject	1	2	3	4	5	6	7	Ave.
correct	47	60	49	28	66	55	42	50.9

また，従来手法との音声の自然性に関する一対評価による結果をFig. 6に示す．一対評価から，提案手法の方が自然音であることがわかる．

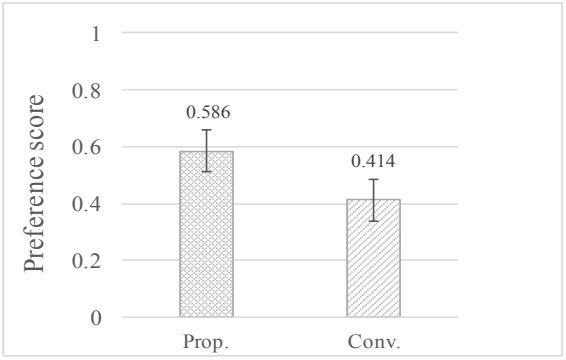

Fig. 6　Proposed vs Conventional

さらに，Fig. 7，8に正解，推定された文章中の「生まれて間もない」のそれぞれのスペクトルを示す．推定されたスペクトルでは，高周波に近づくほど分析の精度が落ちていることが確認できる．

5　おわりに

本研究では，LipNet構造を用いた唇画像特徴量からスペクトル，F0へと変換する音声変換システムを提案した．提案手法を用いることで，F0の変換まで行うことのできるシステムの構築に成功した．しかしながら，その音声はまだ完全なものとは言えないことから，今後，フレーム間同期の必要がない学習方法，唇の範囲に留まらず，顔や喉の動きまで捉えれるようなデータベースの利用による精度向上を考え，研究を進めていく．

謝辞

本研究の一部は，JST さきがけ JPMJPR15D2，JSPS科研費JP17H01995の支援を受けたものである．

参考文献

[1] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” Nature, vol. 264, no. 5588, pp. 746–748, 1976.

[2] J. S. Chung *et al.*, “Lip Reading Sentences in the Wild,” IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[3] R. Aihara *et al.*, “Lip-to-speech synthesis using locality-constraint non-negative matrix factorization,” in Proc. MLSLP, 2015.

[4] S. Taylor *et al.*, “Audio-to-Visual Speech Conversion using Deep Neural Networks,” in Proc. Interspeech, pp. 1482–1486, 2016.

[5] R. Ra *et al.*, “Visual-to-Speech Conversion Based on Maximum Likelihood Estimation,” in Proc. MVA, 2017.

[6] Y. Assael *et al.*, “LipNet : end-to-end sentence-level lipreading,” arXiv:1611.01599, 2016.

[7] H. Kawahara, “STRAIGHT, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds,” Acoustical Science and Technology, pp. 349–353, 2006.

[8] S. Sako *et al.*, “HMM-based text-to-audio-visual speech synthesis image-based approach,” in Proc. ICSLP, vol. 3, pp. 25–28, 2000.

[9] A. Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol. 9, pp. 357–363, 1990.

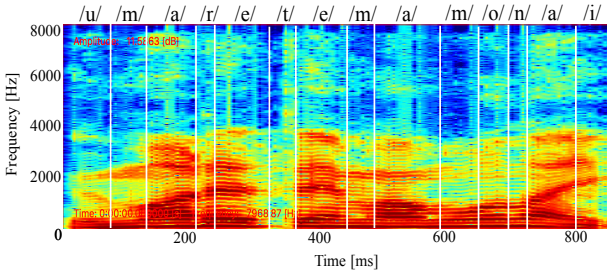

Fig. 7　Spectrogram of an original sound

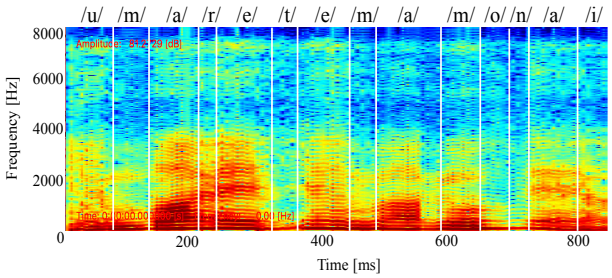

Fig. 8　Spectrogram of a proposed synthesized sound