CNN-LSTM 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，無音声の唇動画像からその動画像に対応する音声へと変換する手法を提案する．この手法は，画像特徴量が持つ情報を他の特徴量が持つ別の情報へと変化させる試みである．一般的に，音韻の知覚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といった視覚情報も影響していることが McGurk らによって報告されている [1]．また，郊外における車の騒音や，電車における他者の会話が混在しているような雑音環境下では音声が聞き取りづらいという問題が考えられる．この問題に対処するために，発話者の顔や唇の動きから発話内容を理解しようとすることは有効であるとされていたり，唇の動きと音声の対応がとれていない，もしくは異なる内容の発話であった場合，唇の動きの影響を強く受け，発話内容を誤認識する恐れがあることも知られている．

一般的に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，雑音環境下における言語理解や発話障害者のコミュニケーション支援，音声の一部が欠落している動画の音声復元といった様々な応用が期待できる．

本研究では，2 通りのアプローチが考えられる．1 つ目は，Lip-reading と text-to-speech(TTS) により音声を生成する方法である．この手法では，入力された唇の動きから Lip-reading によりテキスト情報を認識し，その認識により生成されたテキストから，TTS を用いて音声を生成する．近年の Lip-reading による認識 [2] や，TTS の技術の発展を考えると，このアプローチが効果的であるように思われるが，Lip-reading が誤認識を起こしたとき，意図したものとは全く異なる内容が発話される可能性があるという欠点がある．また，テキスト情報も付与しなければいけないので，唇動画像と音声の 2 つの情報だけで音声生成システムを構築することが出来ない．もう 1 つのアプローチが，発話している無音の唇動画像からテキスト情報を介することなく直接音声へと変換する方法である．この手法では，推定に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ので，本稿では，こちらのアプローチを採用する．

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る．これまでに様々なモデルが用いられてきたが [3, 4, 5]，近年では，Deep Neural Network(DNN) を用いた声質変換 [6] が高い変換精度を持つことから広く利用されており，入力話者のスペクトル特徴量を誤差逆伝搬法 (Back Propagation) により重みを更新することで，出力話者のものへと変換している．他にも，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 [7]，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 [8, 9]．ここでも，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．

本稿では，DNN を用いた声質変換の応用として，画像情報の特徴を抽出する Convolutional Neural Network(CNN) と，連続時間情報を記憶し，回帰する Long Short Term Memory(LSTM) を組み合わせた新たな DNN モデル [10] を用いた唇画像から音声への変換を提案する．本手法では，無音声の唇画像から STRAIGHT [11] を用いて抽出されたスペクトル特徴量を推定し，音声を得る．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．

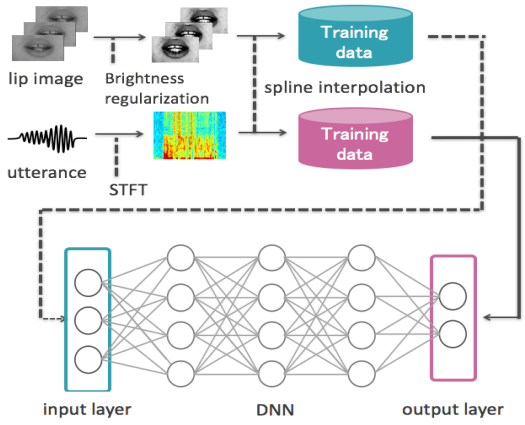

Fig. 1　Speech conversion system

*Lip image to speech conversion using Convolutional Neural Network and Long Short Term Memory, by Daiki Ito (Kobe univ.), Tetsuya Takiguchi (Kobe univ./ JST PRESTO), Yasuo Ariki (Kobe univ.)

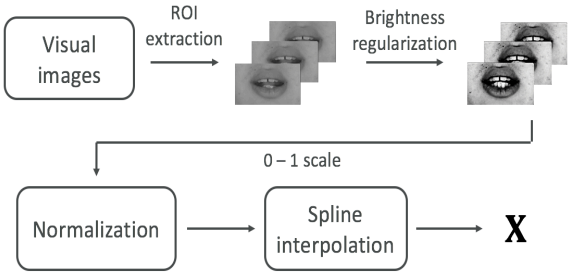

Fig. 2　Flow of visual feature extraction

以下，第 2 章で本稿の提案手法を説明し，第 3 章で評価実験とその結果について示し，第 4 章で本稿をまとめる．

2　提案手法

2.1　音声変換システム

Fig. 1 は本手法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 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．

まず，画像特徴量抽出の流れを Fig. 2 に示す．はじめに，顔画像から対象領域 (Region of Interest : ROI)，つまり唇画像領域のみを抽出し，画像の輝度値を輝度値頻度分布の平坦化によりヒストグラムの偏りをなくす．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 に正規化する．また，音声特徴量のサンプリング周波数とのフレーム間同期をとるために，画像特徴量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，非周期性指標の 3 つの特徴量を抽出した．本稿において，スペクトルが話者性を表す指標であることから，スペクトル特徴量の変換にタスクを限定し，F0 と非周期性指標は抽出された特徴量をそのまま用いるものとする．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．

これらの方法で抽出した画像特徴量から，音声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．学習して得られた DNN に，オープンの画像特徴量を入力し，音声特徴量を推定する音声変換システムを実現した．

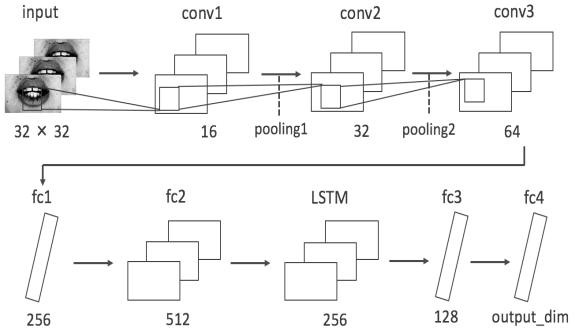

Fig. 3　Proposed DNN architecture

2.2　DNN model

Fig. 3 は本手法で用いる DNN モデルの概要である．本稿で提案している DNN モデルの入力として，画像特徴量 $\mathbf{X}$，教師データには音声特徴量 $\mathbf{Y}$ を用いて，モデル学習し，音声特徴量を推定する．DNN モデルは，CNN から LSTM まで連続しており，CNN と LSTM で別々に学習するようなモデル構造にはなっておらず，勾配の更新は CNN-LSTM モデルを通して行われる．一般に，画像認識において CNN による成果が多く挙げられるので，唇画像から口の開き方や唇の形を分析し，唇の形の特徴を捉えることが出来ると考えられる．また，時系列データに対する予測に有効なモデルである LSTM を用いることで，日本語における母音，子音の連続して一つの音節を表す言語情報を保持したまま回帰出来るようなモデル構造であると考えられる．

以上の理由から本稿において，画像特徴量 $\mathbf{X}$ から音声特徴量 $\mathbf{Y}$ を推定するために，CNN-LSTM モデルを用いている．モデルについて具体的に説明する．カーネルフィルタサイズがそれぞれ 16，32，64 である畳み込み層が 3 つ連続し，カーネルフィルタの大きさはすべて 3 × 3 である．次にノードの数がそれぞれ 512，256 の全結合層が 2 つ続き，その後，隠れ層の数が 256 の LSTM 層が用いられる．最後にノードの数が 128 の全結合層を経て，音声特徴量が Fig.3 の fc4 において得られる構成になっている．

また，第 3 章の評価実験において，CNN や LSTM の層数やノード数を変更した場合との比較を行い，提案手法の DNN モデルの有効性を示している．

3　評価実験

3.1　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 [12] 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 [13] の音

素バランス文 503 文が使用されており，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の評価をする検証データ，結果を考察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 文として用いた．

収録されている動画像データは，唇から鼻先の範囲が映されている．画像のフレームレートは 1/29.97s であり，音声特徴量とのフレーム間の同期をとるためにスプライン補間を適用した．元画像全体のサイズは 720 × 480 ピクセルであり，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 32 × 32 ピクセルにリサイズして用いる．

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 とし，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，スペクトルの変換には STRAIGHT スペクトル 1025 次元から推定され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特徴量は，平均 0 分散 1 に正規化して用いている．

客観評価では，第 2 章における提案手法における DNN モデルの有効性を示すために，比較対象として，以下の 2 つのモデルを用いる．1 つ目は，LSTM による時系列データ予測の有効性を示すために，LSTM を用いず畳み込み層のみで DNN モデルを形成したものを比較に用いる．具体的には，カーネルフィルタサイズがそれぞれ 16，32，64 である畳み込み層が 3 つ連続し，カーネルフィルタの大きさはすべて 3 × 3 で，次にノードの数が 256 の全結合層から音声特徴量が得られるようなモデルである．2 つ目は，CNN の層数の変化による有効性を示すために，畳み込み層を 1 層のみ用いた場合を考える．カーネルフィルタサイズが 32，フィルタの大きさは 3 × 3 であるような畳み込み層の後に，ノードの数がそれぞれ，512，256 の全結合層が続き，その後隠れ層が 256 の LSTM 層，最後にノードの数が 128 の全結合層が用いられたモデルである．

また，提案手法と比較に用いる DNN モデルのどちらに対しても，畳み込み層は Batch Normalization を使用している．活性化関数には，LSTM 層の前までは ReLU 関数，LSTM 層ではシグモイド関数，最後の全結合層では線形関数を用いた．最適化アルゴリズムは Adam を用いた．

客観評価指標には，メルケプストラム歪み (Mel-cepstrum Distortion: MelCD) を用いて各モデルの精度比較をした．主観評価は成人男女 6 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 15 文に関して評価実験を行い，文章中で聞き取れた文節の個数の割合を評価した．読み上げる文章の一覧は与えておらず，スクリプトなしで聞き取れるかを評価している．

3.2 実験結果・考察

Fig. 4 に客観評価による実験結果を示す．左から，畳み込み層 1 層-LSTM の DNN モデル，畳み込み層のみの DNN モデル，提案手法の比較である．図より，畳み込み層の数はある程度増やしたほうが MelCD の精度はあがることから 1 層だけではなく 3 層ほどの Deep 構造を用いた方がよいことがわかる．また，LSTM 層を用いなかった場合と用いた場合を比較すると，本稿の提案モデルを用いたほうが MelCD の精度が向上した．しかし，どちらにおいても顕著な差とは言い難いので，CNN の層数はある程度扱い，CNN のボトルネック特徴量を取り出すなどしてから，そのボトルネック特徴量を LSTM 層へと入力するほうが精度が向上する可能性が考えられる．

次に，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．

Table 1 Correct answer rate (%) for each subject

Subject	1	2	3	4	5	6	Ave.
correct	27	26	45	41	27	24	32.6

また，Fig. 5， 6 に正解，推定された文章中の「その夫人は」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，/j/の子音の高域のスペクトルが，母音の高域でも現れていることから，口の動き，形と音声特徴量の音声が対応していない可能性が考えられる．

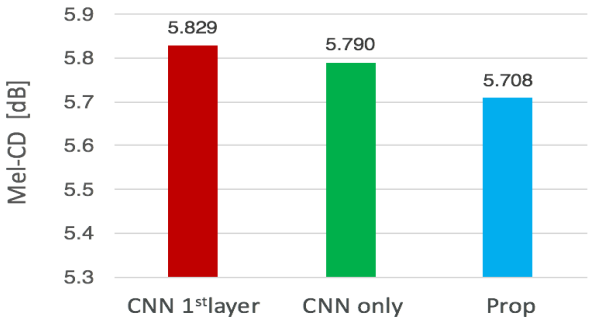

Fig. 4 MelCD of each DNN model

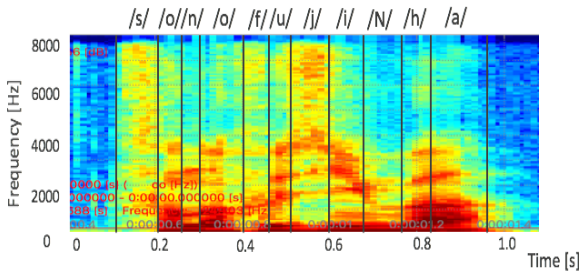

Fig. 5　Spectrogram of an original sound

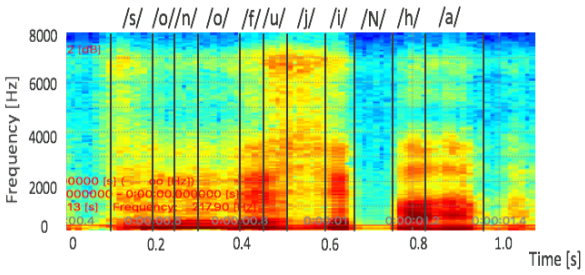

Fig. 6　Spectrogram of a proposed synthesized sound

4　おわりに

本稿では，唇画像特徴量を用いたスペクトル変換による音声生成システムを提案した．提案手法を用いることで，音声情報が欠落した画像から，テキスト情報を介さず，発話音声を復元することができる．画像特徴量分析に強い CNN と，時系列データ予測に強い LSTM を用いることで，画像と音声特徴量の静的特徴量から音声を生成することを可能にした．

今後の課題として，音素，もしくは音節ごとによるスプライン補間を用いたフレーム間同期，もしくは，画像，音声特徴量のアライメントを整えなくても学習できるような手法を考え，今後研究を進めていく．

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] Y. Stylianou *et al.*, “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, 1998.

[4] T. Toda *et al.*, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 8, pp. 2222–2235, 2007.

[5] K. Nakamura *et al.*, “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech,” Speech Communication, vol. 54, no. 1, pp. 134–146, 2012.

[6] Sun, Lifa, *et al.*, “Voice conversion using deep bidirectional long short-term memory based recurrent neural networks,” IEEE International Conference, Acoustics, Speech and Signal Processing (ICASSP), pp. 4869–4873, 2015.

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

[8] E. Yamamoto *et al.*, “Lip movement synthesis from speech based on Hidden Markov Models,” Speech Communication, vol. 25, no. 1–2, pp. 105–115, 1998.

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

[10] Sainath *et al.*, “Convolutional, long short-term memory, fully connected deep neural networks,” IEEE International Conference, Acoustics, Speech and Signal Processing (ICASSP), pp. 4580–4584, 2015.

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

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

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