6V-03

ビデオ通話におけるニューラルネットワークを利用した話者変換の検討

齋藤　優貴†　　能勢　隆†　　篠崎　隆宏‡　　伊藤　彰則†
†東北大学大学院工学研究科　　‡東京工業大学大学院総合理工学研究科

1　はじめに

近年，Skype や FaceTime，LINE といったアプリケーションにおけるビデオ通話が日常生活に浸透してきている．ビデオ通話では通常，お互いの顔を見ることと声を聞くことによってコミュニケーションをとる．声だけでなく顔の表情なども相手に伝えることによって，円滑なコミュニケーションをとることが可能になる．しかし一方では，匿名性は損なわれてしまうため，ビデオ通話をインターネット上で用いることは危険性を孕んでいる．

匿名性を保護しつつビデオ通話を利用する方法として，顔や声といった話者性を変換する手法が考えられる．話者性を他人 (例えば有名人) に変換することで，匿名性を保ったビデオ通話を利用することが可能になる．匿名性が保護されることによって，より気軽にビデオ通話を利用することができることや，表情や口の動きを伝えることで，電話と比較してもより円滑で楽しいコミュニケーションをとれることがメリットとして挙げられる．

本研究では，動画像に含まれる顔画像および音声の個人性の変換 (話者変換) による匿名性の高いビデオ通話システムの構築を目的とする．音声については既に声質変換の研究が行われているが [1][2]，顔画像の変換の技術については未だ十分に検討が行われていない．そこで，本稿では顔画像の個人性の変換についてニューラルネットワーク (NN) に基づく手法を提案し，その有効性を示す．

2　顔画像における話者変換

本稿では，音声における声質変換のアプローチに基づいて顔画像変換を行う．声質変換では，あらかじめ元話者と目標話者が同一文を発話したパラレルデータを用いて，元話者の音声特徴量を目標話者のものへ変換するモデルを学習する．提案法では音声特徴量の代わりに画像特徴量を用いて学習および変換を行う．学習・変換には文献 [3] と同様に NN に基づく手法を利用する．顔画像変換は以下の流れで行う．

1. 顔画像特徴量を抽出する．
2. 音声特徴量により元話者と目標話者の顔画像特徴量のアラインメントを取る．
3. 対応付けられた顔画像特徴量により NN を学習し変換を行う．

A study on speaker conversion using neural networks for video chatting
†Yuki SAITO　†Takashi NOSE　‡Takahiro SHINOZAKI　†Akinori ITO
†Graduate School of Engineering, Tohoku University
‡Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

2.1　顔画像特徴量の抽出

本研究では，元話者および目標話者の顔画像特徴量として，文献 [4] と同様に動画像から得られる顔画像フレームの輝度値に対して主成分分析 (PCA) を行い次元削減したものを用いる．PCA は分散共分散行列に基づく手法を用い，元話者と目標話者の顔画像フレームを合わせ，それらに共通の主成分軸を求め削減に使用する．

2.2　音声特徴量を利用した顔画像特徴量のアラインメント

後述する NN の学習のため，あらかじめ元話者と目標話者が同じ文を発話した動画像を収録し，それから抽出した顔画像特徴量に対してアラインメントをとる必要がある．しかし，予備実験により直接顔画像特徴量により動的時間伸縮 (DTW) を行った場合には元話者・目標話者間の口唇の同期が適切にとれないことがわかった．そこで本研究では音声特徴量によりアラインメントをとり，そのフレーム間の対応付けを顔画像特徴量にも用いた．DTW 処理の流れについては図 1 に，実験条件については表 1 に示した．

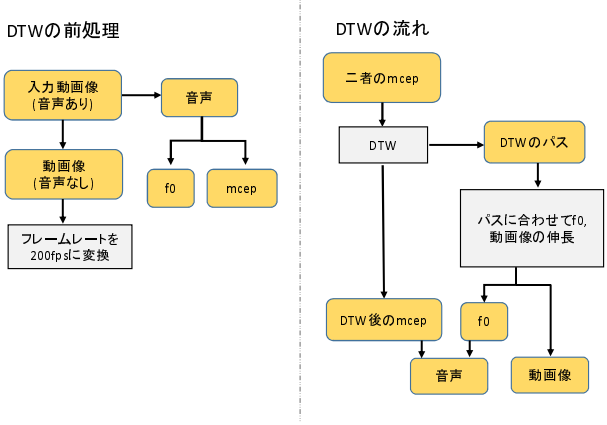

図 1: 処理の流れ

図 1 において，f0 は基本周波数を，mcep はメルケプストラムを表す．また，今回は音声のメルケプストラムに対して DTW を行ったため，メルケプストラムのフレームシフト 5 ms(200 fps) に従って動画像のフレームレートを 30 fps から 200 fps に変換した．変換には線形補間を用いた．DTW の処理を行った後の動画像を同時に再生し，口の開きが一致しており，また音声・画像ともに発話の開始・終了のタイミングも一致しており，アラインメントがとれていることを確認した．

2.3　顔画像特徴量による NN の学習と変換

NN の学習は文献 [3] と同様に入力層から順に学習した RBM を積み上げることでプレトレーニングを行い，

それに対してバックプロパゲーションによるファインチューニングを行う．入力層には元話者の顔画像特徴量を，出力層には目標話者の顔画像特徴量を与えNNの学習を行った．なお，使用する顔画像特徴量は系列の最大値と最小値を利用して0〜1の値をとるように正規化を行った．変換時は入力層に元話者の顔画像特徴量を与えることで出力層にて変換された特徴量を得る．

3 実験

3.1 実験条件

顔画像には，200x200のサイズのグレースケール画像を用いた．実際に用いた顔画像を図2に示す．これをPCAにより100次元に圧縮し顔画像特徴量とする．学習・評価用データとしてATR音素バランス文503文からそれぞれ異なる1文を選び元話者・目標話者がこれらを発話した動画像を用いた．

図 2: 顔画像データ

表 1: DTW の実験条件

音声データ	フォーマット	PCM
	サンプリング周波数	16kHz
	量子化ビット数	16bit
画像データ	フレームレート	30 → 200fps
	サイズ	200x200
mcep	次元数	25
	窓	Hamming
	フレーム幅	25ms
	フレームシフト	5ms

3.2 実験結果

NNの適切なユニット数・中間層の数を調べるため，客観評価実験を行った．図3に変換前後の元話者・目標話者の顔画像特徴量間の平均ユークリッド距離を示す．ユニット数と中間層の数については，(ユニット数)*(中間層の数)のように示した．今回の実験では，全ての中間層においてユニット数は一定とした．図3から，変換前の顔画像特徴量と比較するとすべての場合で平均ユークリッド距離が大幅に小さくなっており，変換された元話者の顔画像特徴量は目標話者に近づいていることが示された．100*3の場合に平均ユークリッド距離が最小となっているが，1000*1の場合を除き，条件による違いは比較的小さかった．100*3のNNを用いた場合の顔画像の例を図4に示す．図から，元の画像に比べ平滑化の影響が見られるが，概ね目標話者に変換されていることがわかる．

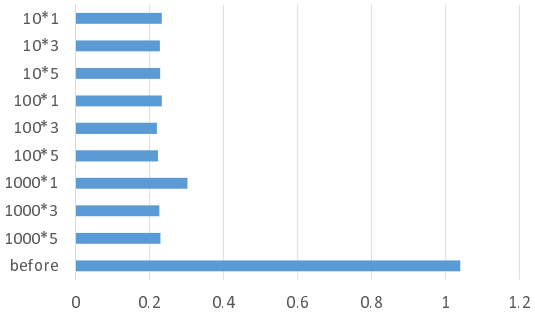

図 3: 平均ユークリッド距離

図 4: 特徴量変換後の顔画像

4 まとめ

本稿では，顔画像の輝度値に対してPCAを適用して次元圧縮を行い，それを特徴量としてNNで学習することで顔画像における話者性の変換が可能であることを示した．今後の予定としては，更に高品質な顔画像変換のために，顔画像特徴量の抽出手法や変換手法の検討を行っていく予定である．

参考文献

[1] Kain and Macon “Spectral voice conversion for text-to-speech synthesis,” in Proc. ICASSP, pp.285-288, 1998.

[2] Toda *et al.* “Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectry,” IEEE Trans. ASLP, Vol. 15, No. 8, pp.2222-2235, 2007.

[3] 伊藤 他 “話者特徴量入力を付与したデノイジングオートエンコーダによるクロスリンガル声質変換”, 情報処理学会技術報告, Vol.2014-SLP-104, 2014.

[4] 酒向 他 “ピクセルベースアプローチによるHMMに基づいた唇動画像の生成”, 電子情報通信学会技術報告, Vol.PRMU99-157, pp.55-60, 1999.