マルチモーダル声質変換を用いた読唇から発話への試み

A study on multi-modal voice conversion using visual information

澤田耕平 [1*]　川嶋大義 [1]　竹原正矩 [1]　田村哲嗣 [1]　速水悟 [1]

Kohei Sawada[1], Daiki Kawashima[1], Masanori Takehara[1], Satoshi Tamura[1], and Satoru Hayamizu[1]

[1] 岐阜大学
[1]Gifu University

Abstract: Voice Conversion (VC) is a technique to convert speech data of source speaker into ones of target speaker. VC is expected in various applications, and one of them is for handicapped people who cannot speak speech due to laryngectomy. VC has been investigated and statistical VC is used for various purposes. Conventional VC uses acoustic features, however, the audio-only VC has suffered from the degradation in noisy or real environments. This paper proposes an Audio-Visual VC (AVVC) method integrating conventional VC and lip-reading technologies especially visual feature extraction. Experiments were conducted to evaluate our AVVC scheme comparing with audio-only VC, using noisy data. The results show that AVVC can improve the performance even in noisy environments, by using audio-visual features. It is also found that visual VC, only using visual information, is also successful.

1　はじめに

声質変換 (Voice Conversion: VC) とは，言語情報を保持したまま非言語情報を変換する技術である．本技術は，発話障害者の会話支援法として期待されている．VC は，入力された音声信号に対して変換処理を施す枠組みである．音声とテキストの両方を入力として VC をする試みもある．さらには，テキストを入力して音声を出力するテキスト音声合成の研究もされている．このように，VC はさらなる人間のコミュニケーション能力の拡大につながる．

読唇は口の動きから言語情報を読み取ることである．しかし，口唇画像から言語情報を直接読み取ることは困難である．そこで，その途中段階として，画像情報から発話を生成することを考える．

本稿では，音声情報に加え，口唇画像情報を用いたマルチモーダル VC を提案する．混合正規分布 (Gaussian Mixture Model) に基づく VC [1] には，主に雑音下で変換音声の品質が劣化する問題がある．一方で，音声認識において音声に加えて口唇画像の特徴量を統合することで，雑音による音声認識率の低下を抑えるマルチモーダル音声認識の研究がされている [2]. そこで，本稿では雑音下で VC を実現するため，従来の VC [1]に口唇画像を統合したマルチモーダル VC によって変換精度の向上を図る．画像情報は音声情報の質とは独立であるため，雑音下での VC の他，音声の質が低い話者に対しても変換性能を維持できることが期待される．また，本技術を用いれば，限られた条件ではあるが読唇による発話の生成も期待される．

本稿は次のような構成になっている．まず，2 章で本稿で用いた GMM に基づく VC について説明する．次に，我々が提案するマルチモーダル VC について 3 章で述べる．4 章では，音声のみの VC とマルチモーダル VC の比較の実験，及び口唇画像からの発話生成を試みる．最後に，本稿のまとめを 5 章で述べる．

* 連絡先：岐阜大学大学院工学研究科応用情報学専攻
〒501-1193　岐阜県岐阜市柳戸 1-1
E-mail: kouhei@asr.info.gifu-u.ac.jp

2　統計的声質変換

本章では，音声のみの従来の統計的声質変換 [1] の概要を述べる．まず，訓練データを用いて変換モデルを構築する．次に，学習したモデルに基づいて入力特徴量から出力特徴量を推定する．そして，推定された出力特徴量を合成して変換音声を生成する．図 1 に本章で述べる VC の流れを示す．

2.1　学習部

まず，元話者と目標話者の同一内容発話の音声データから音響特徴量を抽出する(**T1**, **T3**). ここで，元

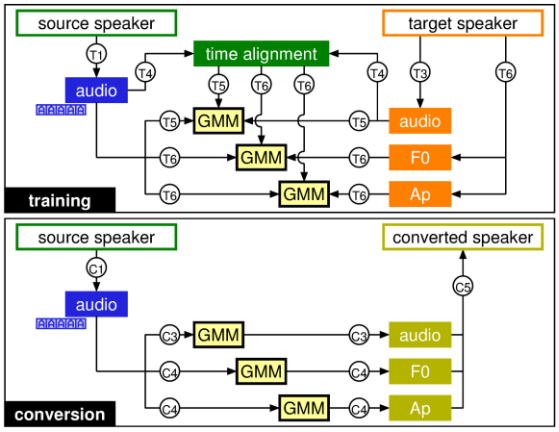

図 1: 音声のみを用いた VC (従来法)

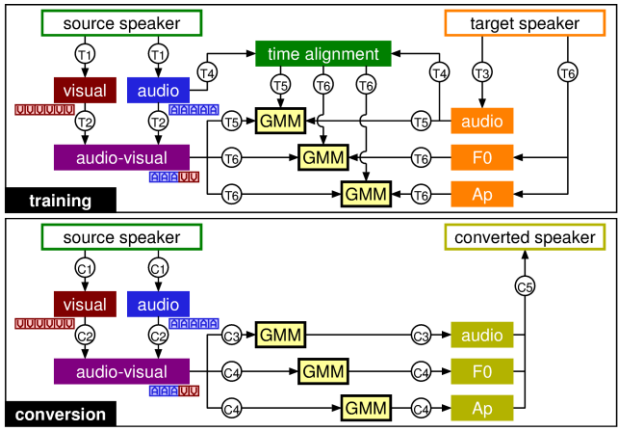

図 2: マルチモーダル VC (提案手法)

話者の特徴量を$\boldsymbol{X}_t$，目標話者の特徴量を$\boldsymbol{Y}_t$とする．ただし，tはフレーム番号を表す．メルケプストラムは入・出力音響特徴量としてよく用いられており，本稿でもそれによった．次に，動的時間伸縮により$\boldsymbol{X}_t$と$\boldsymbol{Y}_t$間の音素単位のフレームアライメントを取得する(**T4**)．このアライメント結果は以降のモデル学習時に使用される．これらの特徴量を訓練データとし，結合確率密度$p(\boldsymbol{X}_t, \boldsymbol{Y}_t)$を GMM によりモデル化する(**T5**)．GMM の学習は式(1)に従う．

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \prod_{t} p(\boldsymbol{X}_t, \boldsymbol{Y}_t|\lambda) \tag{1}$$

ここで，λは GMM のモデルパラメータである．最後に，音声を生成するのに必要となる F0 と非周期成分(aperiodic: Ap)についても目標話者から抽出し，GMM によってモデル化をする(**T6**)．

2.2 変換部

まず，入力音声から元話者の特徴量系列$\boldsymbol{X}= [\boldsymbol{X}_1^\top, \dots, \boldsymbol{X}_T^\top]^\top$を算出する(**C1**)．ただし，$\top$は転置を表す．次に 2.1 で得られた GMM を用いて，出力特徴量を最尤推定法に基づき推定する(**C3**)．目標話者の特徴量系列を$\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \dots, \boldsymbol{Y}_T^\top]^\top$とすると，出力特徴量系列$\hat{\boldsymbol{Y}}$は式(2)で決定される．

$$\hat{\boldsymbol{Y}} = \underset{\boldsymbol{Y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \lambda) \tag{2}$$

式(2)は，式(1)から求められる．最後に，目標話者の F0 及び非周期成分も同様にして算出する(**C4**, **C5**)．

3 マルチモーダル VC

本章では，2 章で述べた音声のみの VC に，口唇の画像から算出した画像特徴量を加えたマルチモーダル VC の手法を述べる．音声認識や音声区間検出(Voice Activity Detection: VAD) では，音響特徴量に加えて画像特徴量を用いたバイモーダル処理の研究がされている[2, 3]．画像特徴量は音響特徴量に比べて雑音の影響を受けにくいため，雑音環境下での性能向上が期待される．我々は，VC に画像特徴量を適用することで，雑音下での精度向上を図る．また，実環境での VC の性能向上による，人間のコミュニケーション能力の拡大を図る．

3.1 概要

我々が提案するマルチモーダル VC は基本的には 2 章で説明した音声のみの VC と同じ工程である．異なる点は，元話者から画像特徴量を抽出して GMM の学習に使用している点である．目標話者については，画像特徴量は使用していない．図 2 に，提案する VC の学習と変換の概略を示す．モデル学習は以下の流れで行われる．

- **T1.** 元話者から音響特徴量と画像特徴量を抽出する．
- **T2.** **T1.** で得られた音響特徴量と画像特徴量を結合する．
- **T3.** 目標話者から音響特徴量を抽出する．
- **T4.** **T1.** と **T3.** で得られた元話者と目標話者の音響特徴量を用いてフレームアライメントを調整する．
- **T5.** **T2.** と **T3.** で得られた特徴量，及び **T4.** でのフレームアライメント結果を用いて GMM を構築する．
- **T6.** 同様にして F0 及び非周期成分推定用の GMM を構築する．

変換部は以下の処理に従う．

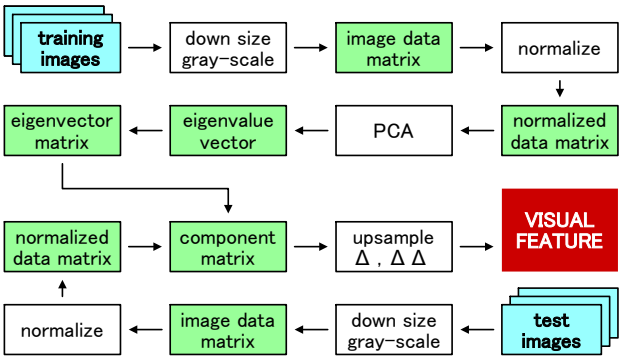

図 3: 画像特徴量抽出

C1. 元話者から音響特徴量と画像特徴量を抽出する．

C2. **C1.** で得られた音響特徴量と画像特徴量を結合する．

C3. **T5.** で構築した GMM を用いて，**C2.** で得られた特徴量から変換特徴量を推定する．

C4. **T6.** で構築した GMM を用いて，**C2.** で得られた特徴量から F0 と非周期成分を推定する．

C5. **C3.** と **C4.** で推定した特徴量を用いて変換音声を生成する．

3.2 特徴量抽出

3.2.1 音響特徴量

元話者と目標話者にそれぞれ異なるスペクトル分析を施す．元話者には高速フーリエ変換 (Fast Fourier Transform: FFT) を用いる．目標話者には高精度な分析合成系 STRAIGHT [4] を用いて分析する．元話者には簡易なスペクトル分析を用い，一方で目標話者には高品質な分析法を適用することで，VC の演算処理を削減しながら変換音声の品質を保つことができる．この分析により，入力メルケプストラム$\boldsymbol{X}_{At}$と出力メルケプストラム$\boldsymbol{Y}_t$が算出される．

3.2.2 画像特徴量

我々が提案するマルチモーダル VC では，画像特徴量として CENSREC-1-AV のベースライン[5]に基づいて算出される固有唇 (eigenlip) を用いる．図 3 に画像特徴量抽出の概略を示す．固有唇を得るために，訓練データの口唇画像に対して主成分分析 (Principal Component Analysis: PCA) を適用する．算出した固有唇を用いて，各口唇画像の固有値を取得する．多くの場合，画像のフレームレートの方が音声のそれより小さいため，特徴量間のフレームの同期が取れるように画像特徴量を時間方向に補間する．同時に，動的特徴量も算出する．以上の処理で，入力画像特徴量ベクトル$\boldsymbol{X}_{Vt}$が求められる．

表 1: 実験条件

データ	タスク	連続数字
	元話者	男性 1 名 (収録音声)
	目標話者	男性 1 名 (CENSREC-1-AV)
	訓練データ数	66
	テストデータ数	10
音声	サンプリング	16000 [Hz]
	フレームサイズ	5 [msec]
	フレームシフト	5 [msec]
画像	フレームレート	30 [fps] (元話者) 29.97 [fps] (目標話者)
	画像サイズ	40×27 (1080 次元)
	訓練データ数	4620
GMM	混合数	16 (F0), 32 (その他)

表 2: 各特徴量の構成要素

	A	**AV**	**V**
Audio	25	15	
Visual		10	25

3.2.3 音響・画像特徴量

音響・画像特徴量は音響特徴量と画像特徴量を用いて式(3)により算出される．

$$\boldsymbol{X}_t = \left(\boldsymbol{X}_{At}{}^{\top}, \boldsymbol{X}_{Vt}{}^{\top}\right)^{\top} \tag{3}$$

式(3)はマルチモーダル音声認識や VAD ではしばしば用いられる．しかし，VC では，次元数が大きすぎると出力特徴量の推定が困難となる．

そこで，我々は音響特徴量・画像特徴量いずれも低次元成分が高次元成分よりも重要であることに着目した．本稿で用いる音響・画像特徴量は入力音響特徴量と入力画像特徴量の低次元成分からそれぞれ抽出して結合される．具体的な統合手法は後述する．

4 評価実験

3 章で述べたマルチモーダル VC の評価実験を行った．本実験の目的は，マルチモーダル VC の性能評価，及び口唇画像からの発話生成である．

4.1 実験条件

表 1 に実験条件を示す．目標話者のデータはマルチモーダル音声認識評価用コーパス CENSREC-1-AV [5]から 1 名を選択した．CENSREC-1-AV は 42 名分のデータを含んでいるが，発話内容は話者ごとに異なる．そこで，パラレルデータを得るために，デー

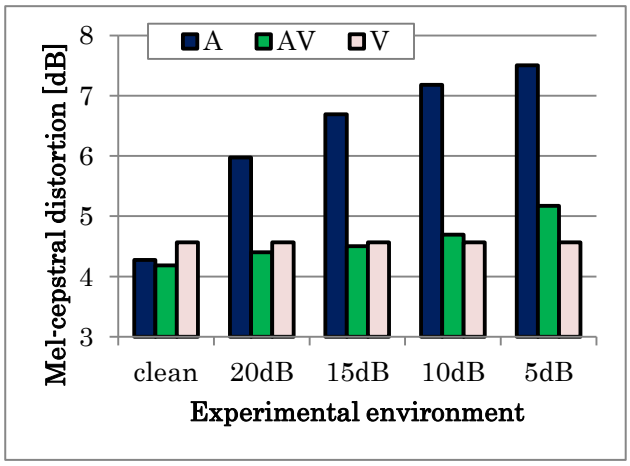

図 4: 各 VC によるメルケプストラム歪み

タベース上の任意の話者と同一内容の発話をクリーン環境で収録し，この音声を元話者音声とした．また，雑音下の性能をみるために，白色雑音 (SNR20dB~0dB) を元話者の音響特徴量に重畳した．

元話者の音響特徴量にはメルケプストラム 25 次元を用いた．元話者の画像特徴量は 10 次元 (static) とその Δ，Δ Δから成る計 30 次元の特徴量を用いた．表 2 に本実験で用いた 3 種類の特徴量を示す．**A** は音声のみの VC (従来法) である． **AV** は 15 次元の音響特徴量と 10 次元の画像特徴量から成る音響･画像特徴量である． **V** は画像特徴量 25 次元のみで構成される. これが, 読唇から発話の生成に相当する. 次元数を 25 次元に統一するために, 元の画像特徴量 30 次元から static を 5 次元分取り除いている.

4.2 評価尺度

客観評価の指標としてメルケプストラム歪み [dB] を使用した．この数値は，目標特徴量と変換特徴量間の歪みを表しており，数値が小さいほど変換音声が目標音声に近いことを示す.

4.3 実験結果

図 4 に 3 つの環境下での音声のみの VC，マルチモーダル VC，及び画像のみ VC の結果を示す. ただし，画像特徴量のみから成る **V** は雑音の影響を受けないため，すべての実験環境で同じ結果となる.

図 4 より, クリーン環境において, 音声のみの VC (**A**) とマルチモーダル VC (**AV**) が同程度の性能 (4.2dB 程度)となった．また，画像のみの VC (**V**) は 4.5dB 程度であり, 他の VC と比較して若干精度は劣るが, 発話の生成に成功した. 雑音環境においては, 音声のみの VC の性能は大幅に劣化している．一方で，雑音が比較的小さい場合は **AV** が，大きい場合は **V** が最も性能が高い．これは，雑音の影響を受けないという画像特徴量の利点が効果的に働いた結果であるといえる．雑音のレベルに合わせて適当な特徴量を選択することで，性能の改善が可能となる.

5 むすび

本稿では，画像情報から音声を生成することを目指し，統計的声質変換 (VC) に画像特徴量を適用したマルチモーダル VC を提案した．実験では，マルチモーダル VC の性能の評価，及び画像のみからの VC を検証した．クリーン環境では，提案手法の変換音声の品質が音声のみの VC のそれと同等となった．雑音環境下では，音声のみの VC と比較して提案手法の方が環境の変化に頑健であった．特に，雑音が大きい場合は，画像情報のみを用いて VC をした方が精度が高いことを確認した． また，画像情報だけから発話を生成することに成功した.

謝辞

本研究の一部は A-STEP (Adaptable and Seamless Technology transfer Program through target–driven R&D)の助成による。また，本研究で使用した声質変換のプログラムをご提供していただいた奈良先端科学技術大学院大学情報学研究科の戸田智基准教授に深く感謝する.

参考文献

[1] T.Toda et al., “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” IEEE Trans. Audio, Speech and Language, vol. 15, no. 8, pp. 2222-2225, 2007.

[2] G.Potamianos et al., “Discriminative training of HMM stream exponents for audio-visual speech recognition,” in Proc. ICASSP ’98, pp. 3733-3736.

[3] C.Ichi et al., “Real-time audio-visual voice activity detection for speech recognition in noisy environments,” in Proc. AVSP 2010, pp. 81-84.

[4] H.Kawahara et al., “Restructuring speech representation using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol. 27, pp. 187-207, 1999.

[5] S.Tamura et al., “CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition,” in Proc. AVSP 2010, pp. 85-88.