最尤変換による唇動画像からの音声生成*

☆羅里奈, 相原龍, 滝口哲也, 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対応する音声を変換する手法 (Visual to Speech Conversion: VTSC) を提案する．音韻知覚は聴覚情報を含む音声からだけでなく，発話者の唇や顔の動きから得られる視覚情報からも影響を受けることが McGurk らによって報告されている [1]．さらに，雑音環境下のような音声が聞き取りにくい状況において，発話者の顔，特に唇の動きから発話内容を理解しようとすることや，唇の動きと音声が一致していない場合に，唇の動きに影響されて発話内容を誤って理解してしまうことがあることも知られている．一般的に，動画像のみから得られる言語情報は音声発話に比べて少ないため，VTSC は困難なタスクであると考えられるが，この技術により，音声障害者のコミュニケーション支援，音声が欠落した映像からの発話復元など，様々な応用が考えられる．

本タスクにおいては，二つのアプローチが考えられる．一つは，リップリーディングと TTS (Text-To-Speech synthesis) を組み合わせるものである．このアプローチでは，入力された唇の動きからリップリーディングを用いてテキスト情報を認識したのち，推定されたテキストから TTS によって音声を生成する．もう一つのアプローチは，入力される唇の動きからテキスト情報を明示的に認識せずに直接音声へと変換するものである．近年のリップリーディング [2] や TTS [3] の技術の発展を考慮すると，前者のアプローチも有効であると考えられるが，リップリーディングが認識誤りを起こした場合，出力される音声の言語情報は入力と大幅に異なったものとなることに加え，リップリーディングと TTS の構築には大量の学習データが必要になるという欠点もある．従って，本稿では後者のアプローチを採用し，この明示的にテキスト情報を認識しないアプローチを VTSC と呼ぶことにする．

声質変換は VTSC と極めて近いタスクであり，発話音声の音韻性などの言語情報を保ちながら，話者性などの非言語情報を変換する技術である．声質変換には様々なモデルが用いられてきたが [4, 5, 6, 7]，それらには入力発話から明示的にテキスト情報を認識しないという共通点がある．その中でも，混合正規分布モデル (Gaussian Mixture Model: GMM) は，その柔軟性と高い変換精度により広く利用されており [4]，入力話者と出力話者のスペクトル特徴量を GMM によって近似し，出力話者のスペクトル特徴量の期待値を考慮することで変換を行っている．変換パラメータは，学習データの最小二乗法または尤度最大化基準 (Maximum Likelihood: ML) を用いて推定されることが一般的である [8]．本稿では，GMM を用いた声質変換 [8] をベースとして，最尤推定に基づく新しい VTSC を提案する．結合された画像特徴量と音声特徴量を，GMM で近似し，入力した画像特徴量は最尤推定を用いて音声特徴量へと変換される．声質変換では短時間のスペクトル特徴量を用いるが，画像データのフレームレートは音声データより小さく，画像データに含まれる情報は音声データに比べて少ないため，短時間特徴量は VTSC には適さない．従って，本稿では，複数のフレームを考慮した長時間画像特徴量を用いる．提案手法では，無音声の動画像からスペクトル特徴量と F0 (Fundamental frequency) を独立に推定し，連続文章発話データベースを用いて，客観評価により評価実験を行った．

関連研究としては，VTSC の逆問題である音声からの口唇動作生成を挙げることができ，隠れマルコフモデルを用いた認識手法が広く研究されている [9]．その他にも，難聴障害者のための支援技術として，ニューラルネットワークを用いて口唇動作生成を行った例や [10]，GMM 声質変換 [8] を口唇動作生成に適用した例もある [11]．非負値行列因子分解を用いた唇動画像からの音声生成も提案されているが [12]，これは，数字発話のような限定されたタスクにおいてのみ有効性が示されている．

以降，2 章では，提案手法について述べる．3 章では，評価実験とその結果を示し，4 章で本稿をまとめる．

2 提案手法

2.1 特徴量構成法

Fig. 1 に画像特徴量抽出の流れを示す．まず，視覚画像から対象領域 (Reagion of Interest: ROI) を抽出した後，画像の輝度値を輝度値頻度分布の平坦化によって正規化をする．次に，画像に対して 2 次元離散コサイン変換 (2-dimensional Discrete Cosine Transform: 2D- DCT) を行った後，ジグザグスキャ

*Visual-to-Speech Conversion Based on Maximum Likelihood Estimation. by Rina Ra, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

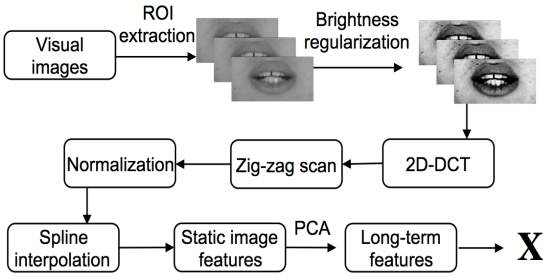

Fig. 1　Flow of the visual feature extraction.

ンを用いて 1D-DCT 係数ベクトルを得る．得られた 1D-DCT 係数ベクトルに対して，Z-score による正規化を行う．また，音声特徴量とのサンプリング周波数のギャップを埋めるために，画像特徴量に対しスプライン補間を適用する．以上の処理により画像データに対する静的特徴量が得られる．

さらに，唇の動きを精細に捉えるため，複数フレームを考慮した長時間特徴量を求める．Fig. 2 に長時間特徴量を抽出する流れを示す．まず，d_x 次 静的画像特徴量ベクトル $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ から，$d_x(2L-1)$ 次元のセグメント特徴量を求める．ここで，T はフレームの総数である．セグメント特徴量に主成分分析 (Principal Component Analysis: PCA) を用いることで，D_x 次元の，複数フレームを考慮した画像特徴量ベクトル $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}$ が得られる．

音声特徴量に関しては，スペクトル特徴量や F0，非周期成分を STRAIGHT [13] を用いて抽出した．本稿では，スペクトル包絡と F0 は画像特徴量からそれぞれ独立に推定される．また，非周期成分については考慮しない．スペクトル推定では，STRAIGHT により抽出されたスペクトルから d_y 次元のメルケプストラムと同次元数の動的特徴量を計算し，それらを結合することで，出力音声特徴量 $\mathbf{Y} = [\mathbf{y}^\mathsf{T} \Delta\mathbf{y}^\mathsf{T}]^\mathsf{T}$ とする．F0 推定では，メルケプストラムと同様，静的特徴量と動的特徴量を結合した $\mathbf{Y}$ を F0 特徴量とする．また，変換において，連続した音声特徴量を推定するために，静的特徴量と動的特徴量間の関係を考慮するトラジェクトリモデルを用いる．

2.2　最尤変換

画像特徴量と音声特徴量の同時確率は平均ベクトル $\boldsymbol{\mu}$ と分散行列 $\boldsymbol{\Sigma}$ をパラメータとする多変量ガウス分布 $\mathcal{N}(.; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ を用いてモデル化される．モデルの学習において，画像特徴量 $\mathbf{X}$ と音声特徴量 $\mathbf{Y}$ を連結させた結合ベクトル $\mathbf{Z} = [\mathbf{X}^\mathsf{T}\mathbf{Y}^\mathsf{T}]^\mathsf{T}$ を用いる．確率 $p(\mathbf{Z})$ は GMM によりモデル化され，次のように表される．

$$p(\mathbf{Z}|\boldsymbol{\Theta}^{(z)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\mathbf{Z}; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}) \tag{1}$$

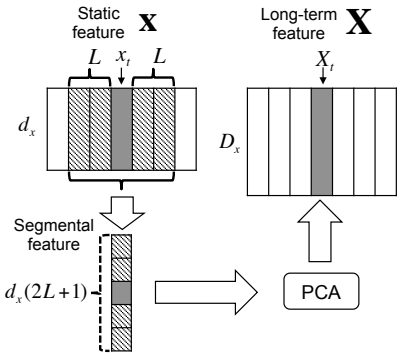

Fig. 2　Flow of the construction of long-term image features.

ここで，$\boldsymbol{\mu}_m^{(z)}$ と $\boldsymbol{\Sigma}_m^{(z)}$ は，

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix} \tag{2}$$

である．パラメータ $\boldsymbol{\mu}_m^{(x)}$ と $\boldsymbol{\Sigma}_m^{(xx)}$，$\boldsymbol{\mu}_m^{(y)}$ と $\boldsymbol{\Sigma}_m^{(yy)}$ はそれぞれ画像特徴量と音声特徴量のガウス分布のものである．α_m は m 番目のガウス分布に対する重みである．$\boldsymbol{\Sigma}_m^{(xy)} (= \boldsymbol{\Sigma}_m^{(yx)^\mathsf{T}})$ は観測データ $\mathbf{X}$ と $\mathbf{Y}$ に対する共分散行列であり，$\boldsymbol{\Theta}^z$ はすべての m に対して α_m, $\boldsymbol{\mu}_m^{(x)}$, $\boldsymbol{\mu}_m^{(y)}$, $\boldsymbol{\Sigma}_m^{(xx)}$, $\boldsymbol{\Sigma}_m^{(yy)}$, $\boldsymbol{\Sigma}_m^{(xy)}$ を含む GMM のパラメータ集合とする．M はガウス混合分布の総数である．

変換段階では，入力 $\mathbf{X}$ が与えられた時の $\mathbf{Y}$ の確率を考える．

$$\begin{aligned} & p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Theta}^{(z)}) \\ &= \sum_{all\mathbf{m}} p(\mathbf{m}|\mathbf{X}, \boldsymbol{\Theta}^{(z)}) p(\mathbf{Y}|\mathbf{X}, \mathbf{m}, \boldsymbol{\Theta}^{(z)}) \\ &= \prod_{t=1}^{T} \sum_{m_t=1}^{M} p(m_t|\mathbf{X}_t, \boldsymbol{\Theta}^{(z)}) p(\mathbf{Y}_t|\mathbf{X}_t, m_t, \boldsymbol{\Theta}^{(z)}) \end{aligned} \tag{3}$$

ここで，$\mathbf{m} = \{m_1, m_2, \cdots, m_T\}$ は分布系列である．また，式 (3) の右辺の確率は次のように表せる．

$$p(m_t|\mathbf{X}_t, \boldsymbol{\Theta}^{(z)}) = \frac{\alpha_m \mathcal{N}(\mathbf{X}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_{n=1}^{M} \alpha_n \mathcal{N}(\mathbf{X}_t; \boldsymbol{\mu}_n^{(x)}, \boldsymbol{\Sigma}_n^{(xx)})} \tag{4}$$

$$p(\mathbf{Y}_t|\mathbf{X}_t, m_t, \boldsymbol{\Theta}^{(z)}) = \mathcal{N}(\mathbf{Y}_t; \mathbf{E}_{m,t}^{(y|x)}, \mathbf{D}_m^{(y|x)}) \tag{5}$$

ここで，

$$\mathbf{E}_{m,t}^{(y|x)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} (\boldsymbol{\Sigma}_m^{(xx)})^{-1} (\mathbf{X}_t - \boldsymbol{\mu}_m^{(x)}) \tag{6}$$

$$\mathbf{D}_m^{(y|x)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} (\boldsymbol{\Sigma}_m^{(xx)})^{-1} \boldsymbol{\Sigma}_m^{(xy)} \tag{7}$$

である．変換特徴量 $\hat{\mathbf{y}}$ は式 (3) の対数尤度関数を最大化することで得られる．まず，分布系列 $\mathbf{m}$ は出力

確率 $p(\mathbf{Y}|\mathbf{X}, \hat{\mathbf{m}}, \mathbf{\Theta}^{(z)})$ を最大化する準最適な分布系列 $\hat{\mathbf{m}}$ で近似される．従って，尤度関数の対数は，

$$\begin{aligned}&\log p(\mathbf{Y}|\mathbf{X}, \hat{\mathbf{m}}, \mathbf{\Theta}^{(z)}) \\ &= -\frac{1}{2}\mathbf{Y}^{\mathsf{T}}{\mathbf{D}^{(y|x)}_{\hat{\mathbf{m}}}}^{-1}\mathbf{Y} + \mathbf{Y}^{\mathsf{T}}{\mathbf{D}^{(y|x)}_{\hat{\mathbf{m}}}}^{-1}\mathbf{E}^{(y|x)}_{\hat{\mathbf{m}}} + K\end{aligned} \tag{8}$$

と書ける．ここで，

$$\mathbf{E}^{(y|x)}_{\hat{\mathbf{m}}} = [\mathbf{E}^{(y|x)}_{\hat{m}_1,1}, \mathbf{E}^{(y|x)}_{\hat{m}_2,2}, \cdots, \mathbf{E}^{(y|x)}_{\hat{m}_T,T}] \tag{9}$$

$$\mathbf{D}^{(y|x)}_{\hat{\mathbf{m}}} = \mathrm{diag}[\mathbf{D}^{(y|x)}_{\hat{m}_1,1}, \mathbf{D}^{(y|x)}_{\hat{m}_2,2}, \cdots, \mathbf{D}^{(y|x)}_{\hat{m}_T,T}]. \tag{10}$$

である．これより，変換特徴量 $\hat{\mathbf{y}}$ は，

$$\hat{\mathbf{y}} = (\mathbf{W}^{\mathsf{T}}{\mathbf{D}^{(y|x)}_{\hat{\mathbf{m}}}}^{-1}\mathbf{W})^{-1}\mathbf{W}^{\mathsf{T}}{\mathbf{D}^{(y|x)}_{\hat{\mathbf{m}}}}^{-1}\mathbf{E}^{(y|x)}_{\hat{\mathbf{m}}} \tag{11}$$

で表される．

3 実験結果

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 [14] を用いて評価実験を行った．収録されている音素バランス 503 文のうち．300 文を学習データとし，学習に用いていない 50 文をテストデータとした．

動画像データは唇から鼻先の範囲を映している．画像のフレームレートは $1/29.97[s]$ であり，音声特徴量とのフレームレートの差を埋めるためスプライン補間を適用した．元画像の全体のサイズは 720×480 ピクセルであり，対象領域を抽出し 40×20 ピクセルにリサイズする．DCT 静的画像特徴量の次元数は 25 次元であり，セグメント特徴量は PCA により 50 次元に圧縮している．

音声発話データのサンプリング周波数は 16kHz で，フレームシフトは 5ms である．各サンプルは STRAIGHT [13] によって分析することで，スペクトル特徴量と F0，非周期成分が抽出される．スペクトル推定においては，STRAIGHT スペクトルから推定されたメルケプストラム，動的特徴量，パワーを結合した 50 次元の特徴量を用いる．

スペクトル推定の評価基準として，以下の式で定義されるメルケプストラム歪み (Mel-cepstrum Distortion: MelCD) を用いる．

$$MelCD = (10/\log 10)\sqrt{2\sum_{d}^{25}(mc_d^{conv} - mc_d^{tar})^2} \tag{12}$$

ここで，mc_d^{conv} と mc_d^{tar} はそれぞれ d 次における変換，ターゲットのメルケプストラムである．

F0 推定においては，F0 特徴量の次元数は静的特徴量に動的特徴量を結合させた 2 次元である．推定されたスペクトルと F0 は非周期成分を考慮せず，STRAIGHT を用いて合成される．GMM の混合数は $\{2, 4, 8, 16, 32, 64, 128\}$ の中から実験的に最適なものを選択する．F0 推定では二乗平均平方根 (Root Mean Square Error: RSME) を用いて評価した．

3.2 実験結果と考察

まず，スペクトル推定において長時間特徴量の比較を行った．Fig. 3 にその結果を示す．Static+delta は静的 DCT 特徴量と動的特徴量を結合したもの，PCA は，複数フレームを考慮した長時間特徴量を表し，L は Fig. 2 で説明されている．図より，長時間特徴量の有効性が示された．

さらに，学習データ量による変換精度の違いを比較し，Fig. 4 にその結果を示す．図より，学習データ量に比例して変換精度が向上することがわかる．

F0 推定の結果を，Fig. 5 に示す．スペクトル推定と同様，F0 推定においても長時間特徴量の有効性が示されている．Fig. 6, 7 に目標スペクトル包絡と変換スペクトル包絡の例を示す．

4 まとめ

本稿では，画像特徴量を用いたスペクトルと F0 推定のための統計的手法を提案した．提案手法によって，音声情報が欠落した画像から，発話音声を再構築することができる．スペクトル包絡と F0 はそれぞれ画像特徴量と結合し，独立した GMM によってモデル化され，目標の音声特徴量は最尤推定によって得られる．音声特徴量と比較してフレームレートの小さい画像特徴量から唇の動きを精細に捉えるために，複数フレームを考慮した長時間画像特徴量を用いた．今後，データベースを拡張した上で，より効果的な画像特徴量との比較を行っていく．

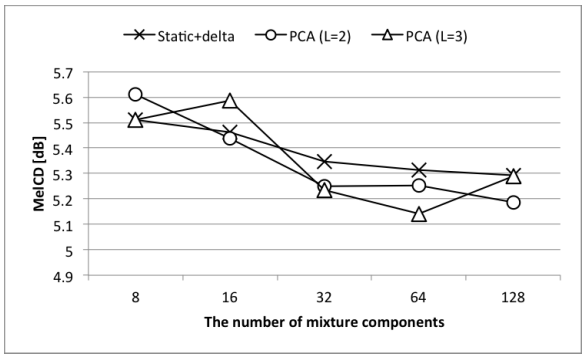

Fig. 3 MelCD as a function of the number of mixture components using 300 training sentences.

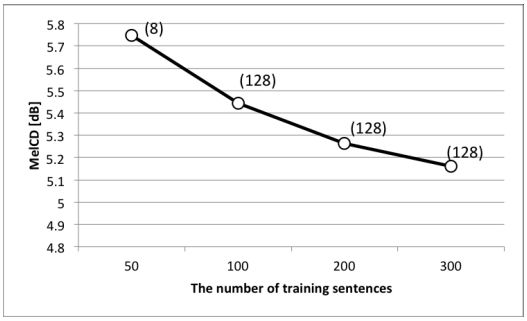

Fig. 4　MelCD as a function of the number of training sentences using PCA ($L = 3$). The numbers within parentheses indicate the optimum number of mixture components.

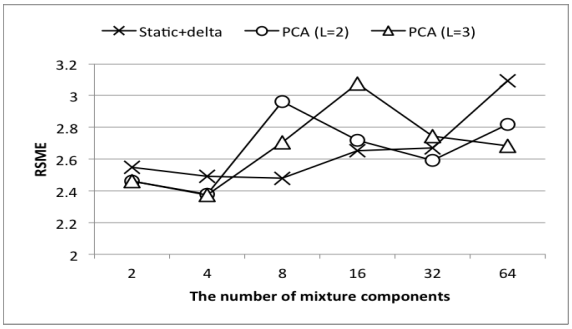

Fig. 5　RMSE as a function of the number of mixture components using 300 training sentences.

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] Y. M. Assael *et al.*, "Lipnet: Sentence-level lipreading," arXiv:1611.01599, 2016.

[3] A. van den Oord *et al.*, "Wavenet: A generative model for raw audio," CoRR, vol. abs/1609.03499, 2016.

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

[5] C. Ling-Hui *et al.*, "Joint spectral distribution modeling using restricted boltzmann machines for voice conversion," in Proc. Interspeech, pp. 3052—3056, 2013.

[6] R. Aihara *et al.*, "Multiple non-negative matrix factorization for many-to-many voice conversion," IEEE/ACM Trans. on Audio, Speech, and Language Processing, vol. 24, no. 7, pp. 1175–1184, 2016.

[7] 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.

[8] 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.

[9] 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.

[10] F. Lavagetto, "Converting speech into lip movements: a multimedia telephone for hard of hearing people," IEEE Trans. on Rehabilitation Engineering, vol. 3, no. 1, pp. 90–102, 1995.

[11] X. Zhuang *et al.*, "A minimum converted trajectory error (MCTE) approach to high quality speech-to-lips conversion," in Proc. INTERSPEECH, pp. 1736–1739, 2010.

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

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

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

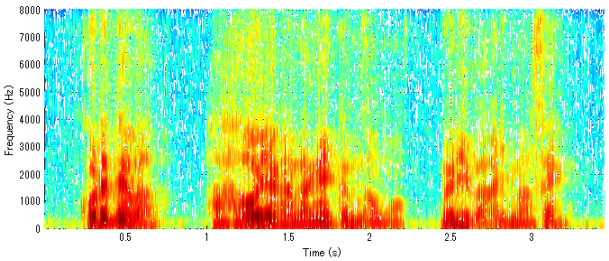

Fig. 6　An example audio target spectrogram.

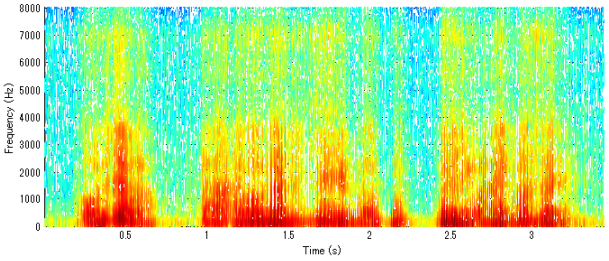

Fig. 7　An example of an estimated audio spectrogram using PCA ($L = 3$) and 300 training sentences.