顔画像特徴量を用いた統計的手法による F0 推定*

☆羅里奈, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

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．

我々は，最尤変換による VTSC 手法を提案し，唇動画像からの音声生成を行った [9]．変換に用いた動画像は 29.97fps であり，男性 1 名の連続文章発話となっている．この文献で，統計的手法を用いることで，無音声の動画像から発話音声を生成することができた．しかし，より自然な音声を合成するには，F0 (Fundamental frequency) も重要な要素であり，自然な抑揚を実現するためには，唇の動きをより精細に捉える必要があるという課題ができた．よって，ハイスピードカメラで撮影した 500fps の無音声動画像を用いて F0 推定を行う．ハイスピードカメラを用いることでより細かい口元の振動を捉えることができ，F0 推定に関しても有効な結果を示せると考えられる．

本稿では，ハイスピードカメラで収録した動画像に対し GMM による F0 推定を行う．まず，500fps の無音声ハイスピード画像から画像特徴量を生成する，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，無音声な動画像から自然な音声を得るためには，唇の動きの流れを捉える必要があるため，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．また，ハイスピードカメラで撮影した動画像を用いるので，画像データのフレームレートの同期を取れた上で唇の動きの流れを捉えることができる．提案手法では，無音声の動画像から F0 を推定し，連続数字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，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 [13]，これは，F0 に関しての推定を行っていない．

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

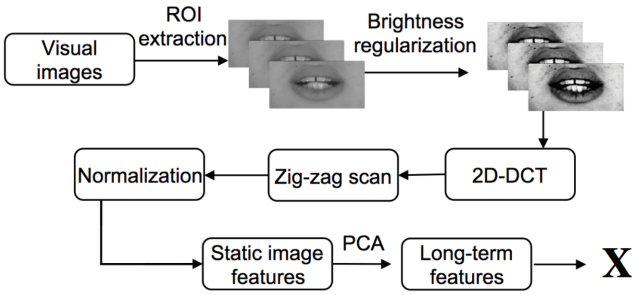

Fig. 1 Flow of the visual feature extraction.

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 (Region 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

*F0 estimation Based on Statistical methods using Facial Image Features. by Rina Ra, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

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．今回用いる動画像のフレームレートに合わせて，音声のフレームレートも 2ms で分析を行うことで，画像と音声の同期をとった．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 [14] を用いて抽出した．本稿では，非周期成分については考慮しない．F0 推定では，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．

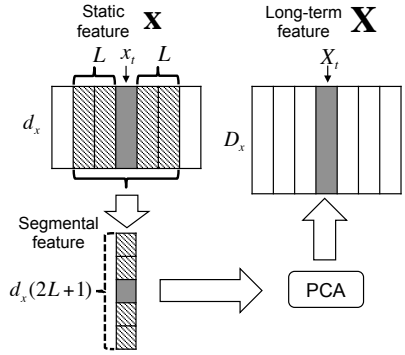

Fig. 2　Flow of the construction of long-term image features.

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．

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

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 $p(\mathbf{Y}|\mathbf{X}, \hat{\mathbf{m}}, \boldsymbol{\Theta}^{(z)})$ 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，

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

と書ける．ここで，

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

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

である．これより，変換特徴量 $\hat{\mathbf{y}}$ は，

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

で表される．

3　実験結果

3.1　実験条件

本稿では，0 から 1，〇/maru/を含む数字発話 157 文をハイスピードカメラで収録した．Table 1 に収録した連続数字発話の桁数と発話数を示す．収録した 157 発話のうち 4 発話をテストデータとした．closed 実験では，テストデータを含む 157 発話全てを用いて学習データを構築し，open 実験ではテストデータを除いた 153 発話を用いて学習データを構築した．収録は男性 1 名の被験者について正面で撮影した，撮影機器は，MEMRECAM GX-1 であり，フレームレートは 500fps を使用した．動画像のフレームレートの比較には，500fps で収録した動画像から 30fps となるように間引いたものを使用した．その際，音声とのフレームレートの差を埋めるためにスプライン補間を適用した．元画像の全体のサイズは 640×480 ピクセルであり，唇領域の解像度は，100×150，対象領域を抽出し 30×45 ピクセルにリサイズする．DCT 静的画像特徴量の次元数は 50 次元であり，セグメント特徴量は PCA により 100 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 48kHz で，フレームシフトは 2ms である．各サンプルは STRAIGHT [14] 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．F0 推定においては，スペクトル分析で得られた F0 に対数をとる．そして，動的特徴量を計算し結合した 2 次元の特徴量を用いる．

今回，F0 推定の評価基準として，平均二乗誤差 (Root Mean Square Error:RSME) を用いる．ここで，y_i^{tar} と y_i^{conv} はそれぞれ i 番目のおけるターゲット，変換の対数 F0 である．

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．

Table 1　Number of digit strings

length of digits	number of data
1	10
2	25
3	30
4	33
5	29
6	5
7	23
8	2
total	157

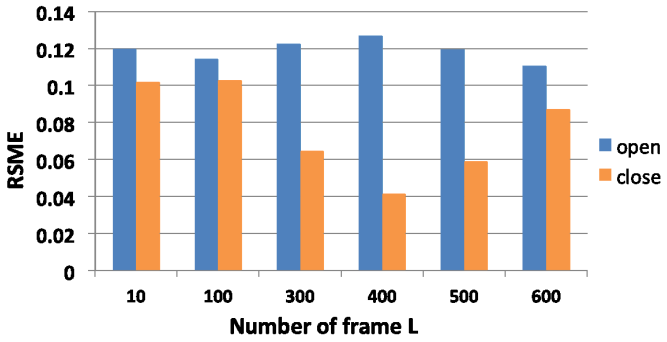

Fig. 3　RMSE as a function of number of PCA frames.

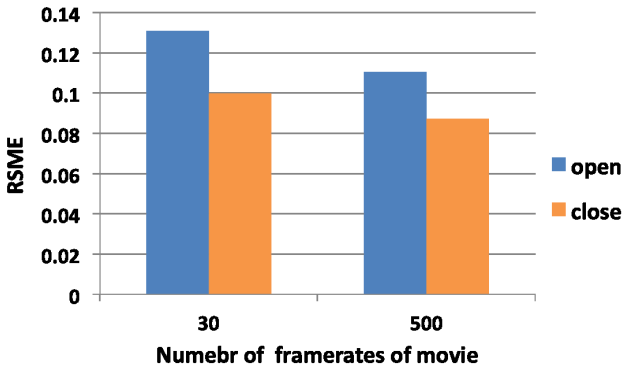

Fig. 4　RMSE as a function of number of frame rates of video.

3.2　実験結果と考察

まず，長時間特徴量の比較を行った．Fig. 3 にその結果を示す．L は Fig. 2 で説明されており，それぞれの open 実験 (青) と closed 実験 (橙) 結果を示してある．図より，close 実験においては，400 フレーム，つまり 800ms 分の情報を加味した特徴量がいい結果となった，対して，open 実験では，各フレーム数の大差がみられず，600 フレーム，つまり 1.2 秒分の情報を加味した特徴量がいい結果となった．

さらに，動画像のフレームレートでの違いを比較し，Fig. 4 にその結果を示す．図より，フレームレート 500fps の特徴量が良い結果となった．

目標 F0 波形と変換 F0 波形の比較結果を，Fig. 5, Fig. 6 に示す．青線が目標波形，赤線が変換結果であり，Fig. 5 が close 実験，Fig. 6 が open 実験結果となっている．close 実験の結果は，ほぼ目標波形と一致しているが，open 実験結果では，大きく外れた値が見られた．

4　まとめ

本稿では，統計的手法を用いたハイスピード画像特徴量からの F0 推定を行った．ハイスピードカメラを用いることで，従来の低フレームレートのカメラ

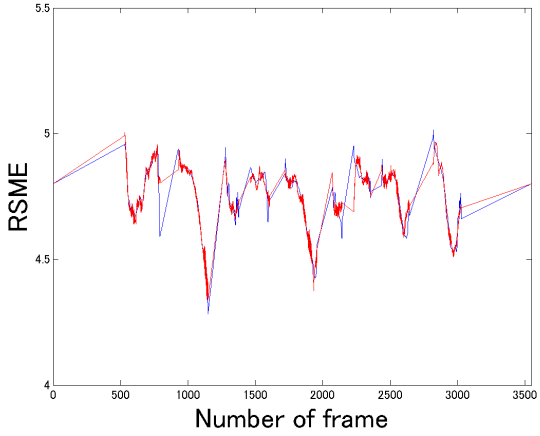

Fig. 5　Plot of F0 in closed experiments.

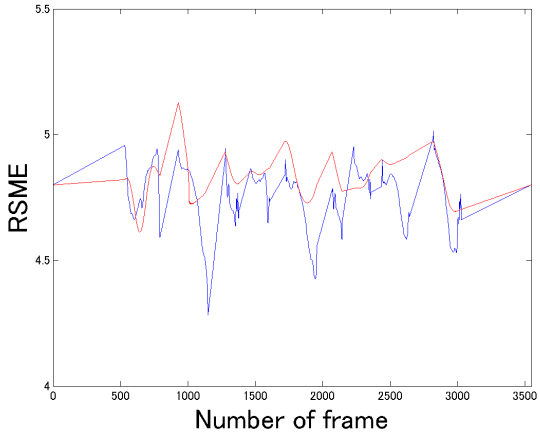

Fig. 6　Plot of F0 in open experiments.

で収録したものよりも，唇の精細な動きを捉えられ，良い結果を示した．これにより，無音声の唇動画像からより自然な声を生成することができる．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 F0 特徴量は最尤推定によって得られる．音声特徴量と同等のフレームレートの画像特徴量から唇の動きの流れ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，ハイスピードカメラによるスペクトル包絡の推定を行い音声を作り上げる．

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] 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] R. Ra *et al.*, “Visual-to-speech conversion based on maximum likelihood estimation,” MVA2017, pp. 488-491, 2017. 5.

[10] 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.

[11] 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.

[12] 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.

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

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