非負値行列因子分解に基づく唇動画像からの音声生成*

☆真坂健太, 相原龍, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

唇の動きから発話内容を読み取る技術はリップリーディング（読唇）と呼ばれ，聴覚・言語障害者のコミュニケーション手段の一つとして用いられている．本研究では，非負値行列因子分解（Non-negative Matrix Factorization：NMF）を用いて，唇動画像からそれに対応する発話音声を生成する．同時に収録した発話映像と音声からそれぞれ唇情報と音声情報を抽出し，それぞれを基底の集合である辞書として学習する．本研究では発話映像を撮影する際，ハイスピードカメラを用いて音声と同じフレームレートで撮影した．このとき，二つの辞書行列は同一時系列であり，パラレルなデータである．入力された無音声の映像から抽出された唇情報は，NMF により少数の基底の線形和で表される．唇辞書行列から選ばれた基底を対応する音声辞書の基底と取り換えることで，音声の基底の線形和として音声が出力される．

従来，音声認識や声質変換といった音声における信号処理は，音響的な特徴量にのみ着目して研究されてきた．しかし，人間は発話内容を理解する際，様々な情報を統合的に利用している．音声が聞き取りにくい場合，発話者の顔，特に唇の動きに注目して発話内容を理解しようとし，逆に唇の動きと音声が一致しない場合，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もある．これは，McGurk effect（マガーク効果）と呼ばれ，音韻知覚が音声の聴覚情報のみで決まるのではなく，唇の動きといった視覚情報からも影響を受けることが報告されている [1].

また，音声認識技術の発展により，スマートフォンでの音声による文書作成，音声認識に対応したカーナビゲーションシステムなど，さまざまな音声認識技術がコンピュータへの新しいインターフェースとして実用化されてきているものの，現在の音声認識技術には雑音の大きい環境下では認識性能が著しく低下してしまう問題がある．リップリーディングは雑音に影響されることがないため，雑音環境下で頑強に発話認識を行うための手法の一つとして，音声情報に唇動画情報を併用して認識を行うマルチモーダル音声認識が注目され，研究が進められている．

一方で，リップリーディングは聴覚障害者のコミュニケーション手段の一つとして期待されてきた．情報技術の福祉分野への応用も近年進んでおり，画像認識技術の応用による手話認識 [2]，文章読み上げシステム [3]，無喉頭音声変換 [4]，構音障害者のための声質変換 [5] など，その応用領域は幅広い．文献 [6] では，Active Appearance Model (AAM) の C パラメータを用いた顔方位変動に対応したリップリーディングを提案し，構音障害者のためのマルチモーダル音声認識を行った．現在，日本だけでも約 3 万 4 千人の言語・聴覚障害者がいることからも，このようなリップリーディングの福祉分野への応用もニーズが高まっている．

そこで，本稿では従来，雑音除去 [7] や超解像 [8] に用いられてきた Sparse Coding の代表的な手法である NMF [10] を用いて，無音声の発話動画から対応する発話音声へ変換する手法を提案する．NMF では，入力信号は辞書行列に含まれる少量の基底の線形和で表現される．無音声の唇動画が入力されると，事前に学習した唇情報の基底集合である辞書行列から，基底とその重みを推定する．推定された基底を対応する音声情報の辞書行列の基底と入れ替えることで，入力唇動画は音声基底の線形和として変換される．事前に学習を必要とするものの，変換に際しテキスト情報は用いず，唇の動きのみから発話音声へと変換する．

この技術により，声帯結節，喉頭がん，ポリープといった喉頭疾患に伴う音声障害者のコミュニケーション支援につながる．さらに音声が欠落した映像からの発話復元や，騒音環境下でのコミュニケーションツールなど，音声によるコミュニケーションが困難な状況において様々な形で応用できると考えられる．

以降，2 章では NMF について述べ．3 章で唇情報からの音声生成法について説明する．4 章で評価実験とその結果を示し，5 章で本稿をまとめる．

2 非負値行列因子分解

スパースコーディングの考え方において，与えられた信号は少量の学習サンプルや基底の線形結合で表現される．

$$\mathbf{x}_l \approx \sum_{j=1}^{J} \mathbf{w}_j h_{j,l} = \mathbf{W}\mathbf{h}_l \qquad (1)$$

$\mathbf{x}_l$ は観測信号の l 番目のフレームにおける D 次元の特徴量ベクトルを表す．$\mathbf{w}_j$ は j 番目の学習サンプル，あるいは基底を表し，$h_{j,l}$ はその結合重みを表す．本

*Speech Production from Lip Images based on Non-negative Matrix Factorization. by Kenta Masaka, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

手法では学習サンプルそのものを基底 $\mathbf{w}_j$ とする．基底を並べた行列 $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_J]$ は “辞書” と呼び，重みを並べたベクトル $\mathbf{h}_l = [h_{1,l} \dots h_{J,l}]^T$ は “アクティビティ” と呼ぶ．このアクティビティベクトル $\mathbf{h}_l$ がスパースであるとき，観測信号は重みが非ゼロである少量の基底ベクトルのみで表現されることになる．フレーム毎の特徴量ベクトルを並べて表現すると式 (1) は二つの行列の内積で表される．

$$\mathbf{X} \approx \mathbf{WH} \tag{2}$$

$$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_L], \quad \mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_L]. \tag{3}$$

ここで L はフレーム数を表す．本手法では，アクティビティ行列の推定にスパースコーディングの代表的手法である NMF を用いる．NMF ではコスト関数として Kullback-Leibler (KL) divergence を用いる．

$$d(\mathbf{x}_l, \mathbf{W}\mathbf{h}_l) + ||\lambda . * \mathbf{h}_l||_1 \qquad s.t. \quad \mathbf{h}_l \geq 0 \tag{4}$$

第 1 項は KL divergence である．第 2 項は $\mathbf{H}$ をスパースにするための L1 ノルム正規化項である．$.*$ は要素ごとの掛け算を表す．$\lambda^T = [\lambda_1 \dots \lambda_J]$ を調節することで，辞書内のサンプル毎に定義することができる．本稿ではスパース制約重み $[\lambda_1 \dots \lambda_J]$ を 1 に設定した．(4) 式を最小にするように以下の更新式に従いアクティビティ行列 $\mathbf{H}$ が推定される．

$$\mathbf{h}_l \leftarrow \mathbf{h}_l \frac{\mathbf{W}\mathbf{x}_l / \mathbf{W}\mathbf{h}_l}{\mathbf{1} + \lambda} \tag{5}$$

3 NMF による音声生成

3.1 辞書構成法

Fig. 1 は画像辞書，音声辞書の構成法を示したものである．本研究では，ハイスピードカメラの映像から抽出したフレーム画像を用いることで，音声と同一フレームレートの画像特徴量を得る．画像特徴量の抽出は，まずフレーム画像から唇部分を切り出した後，DCT（Discrete Cosine Transform）を行う．つづいて，得られた DCT 画像に対してジグザグスキャンを行い，低次 200 次元のみを取り出す．さらに NMF の非負制約を満たすため，負値を取らないように底上げしたものを画像特徴量とする．本研究では，音声の抽出・再合成に音声変換合成方式 STRAIGHT を使用している [9]．STRAIGHT は音声合成や声質変換で広く使われている分析合成手法である．音声辞書の構築には各発話ごとに STRAIGHT スペクトルを並べたものを音声辞書とする．

3.2 ローカリティ制約の導入

本手法では，アクティビティのスパース性を高めるため，アクティビティ推定時にローカリティ制約を導入する．

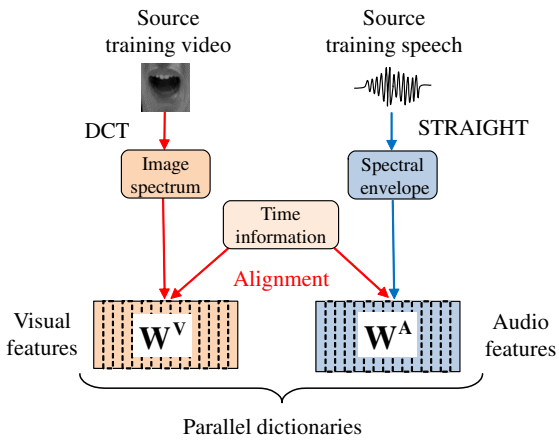

Fig. 1 Dictionary construction

$$\Delta_{j,l} = \sqrt{(\mathbf{x}_l - \mathbf{w}_j)^2} \tag{6}$$

$\mathbf{x}_l$ と $\mathbf{w}_j$ はそれぞれ，入力特徴量の l フレーム目のベクトル，画像辞書の j 番目の基底を表す．$\Delta_{j,l}$ は $\mathbf{x}_l$ と $\mathbf{w}_j$ とのユークリッド距離である．入力ベクトルに対して，$\Delta_{j,l}$ の小さいものから n 個の基底のみを用いてアクティビティを推定する．

$$\mathbf{S}_l = \mathbf{nbest}_{\Delta_l}(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_J) \tag{7}$$

$$= \mathbf{nbest}_{\Delta_l}(\mathbf{W}) \tag{8}$$

$\mathbf{S}_l$ は l 番目の入力ベクトルに対して選ばれた基底の集合である．$\mathbf{S}_l$ に対応するアクティビティにのみ初期値を与え，他のアクティビティを 0 としすることで，入力ベクトルそれぞれに対して距離の近い n 個の基底のみでアクティビティが推定される．

3.3 生成手法

Fig. 2 に，唇情報から音声情報への変換方法の概要を示す．一発話から取り出された唇特徴量を $\mathbf{X}^V$，画像辞書行列を $\mathbf{W}^V$，音声辞書行列を $\mathbf{W}^A$，求める音声特徴量を $\mathbf{X}^A$ とする．ここで D，L，J はそれぞれ唇特徴量の次元数，入力唇情報及び出力音声情報のフレーム数，唇辞書行列および音声辞書行列のフレーム数である．

変換する無音声の入力映像は，唇情報を抽出し，Fig. 2 の上段に示すように NMF を用いて唇辞書行列と係数行列に分解され，少数の基底の線形和で表される．係数行列には，入力唇情報が，辞書行列のどの基底が，どのくらいの重みで構成されるかの情報が含まれる．Fig. 2 の下段にあるように，推定されたアクティビティは音声辞書行列とかけあわされる．唇辞書行列と音声辞書行列はパラレルであるため，唇辞書行列で使われる基底と同じ基底が音声辞書行列

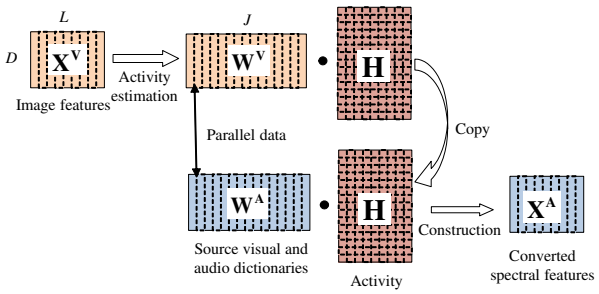

Fig. 2　Flow of conversion

から得られる．つまり，唇辞書行列の基底の線形和で表されていた入力唇情報が，対応する音声辞書行列の基底の線形和へと変換されたことになる．得られた音声情報は，STRAIGHT を用いて再合成され，無音声の発話映像が，対応する発話内容の音声へと変換される．

4　評価実験

4.1　実験条件

本稿では，発話映像として CENSREC-1-AV データベースに含まれる数字発話 26 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 26 発話のうち 6 発話をテストデータとした．close 実験ではテストデータを含む 26 発話全てを用いて辞書を構築し，open 実験ではテストデータを除いた 20 発話で辞書を構築した．収録は男性 1 名の被験者について正面，カメラからの距離 65cm，撮影機器は MEMRECAM GX-1 で，唇領域の解像度は 130 × 80，フレームレートは 1000fps を使用した．

Table 1　Number of digit strings

length of digits	number of data
2	9
3	7
4	10
total	26

画像特徴量は，唇領域を抽出した後 DCT を行って得た低周波成分 200 次元と，その前後 2 フレームずつを加えた 1000 次元を用いた．ハイスピードカメラを用いて収録した唇画像の例を Fig. 3 に示す．音声特徴量は，唇動画収録と同時に収録した音声を用いる．特徴量として STRAIGHT スペクトル 513 次元を用いた．サンプリング周波数は 8 kHz，フレームシフトは 1ms である．

Fig. 3　Lip images

4.2　実験結果・考察

本手法における目標音声と生成音声の Mel-CD（Mel-cepstrum Distortion）を Fig. 4 に示す．Mel-CD は以下の式で表される．

$$\text{Mel-CD[dB]} = 10/\ln 10\sqrt{2\sum_{d=1}^{24}(mc_d^t - \hat{mc}_d^t)^2} \quad (9)$$

mc_d^t と $\hat{mc}_d^t$ は目標音声とと生成音声の d 次元目の係数である．横軸の数字はローカリティ制約で選択する基底数を示している．Fig. 4 より，結果が一番良かったのは open，close 実験で基底数がそれぞれ 500，50 のときであった．close 実験において基底数が増えるにつれて Mel-CD が大きくなるのは，不要な基底まで選択されてしまい，不明瞭な音声になっているからだと考えられる．open 実験において基底数が小さくなるつれて Mel-CD が大きくなるのは，ローカリティ制約による基底の選択誤りの影響が大きいと考えられる．

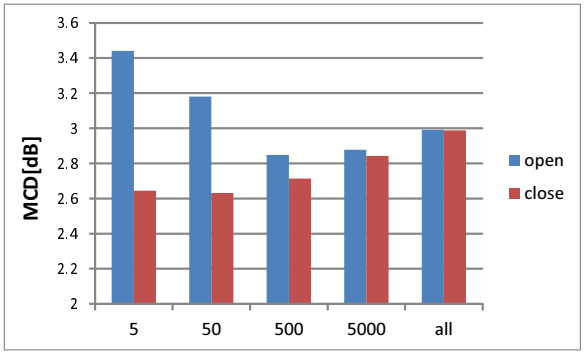

Fig. 4　Mel-cepstrum distortion

生成音声を評価するために主観評価実験を行った．成人男性 7 人を対象に，目標音声にどれぐらい近いかを MOS (Mean Opinion Score) 評価基準に基づく 5 段階評価 (5:とても近い，4:近い，3:どちらともいえない，2:遠い，1:とても遠い) を行った．また，発話認識実験も行った．この実験では数字発話を聞いてどの数字を言っているかを書き取ってもらい，正解率を算出している．

Fig. 5 に MOS による実験結果示す．close において，客観評価実験と同様に基底の数が 50 の時に一番良い結果となった．open においてはより少ない基底数で音声を生成することにより，明確な音声となっているからだと考えられる．Fig. 6 に書き取りテスト

による認識結果を示す．close 実験において，ローカリティ制約導入時には認識結果が 60% を超えており，open 実験でも 50% を超える認識結果が得られた．一方，ローカリティ制約を導入しない場合には，どちらの実験でもほとんど発話内容が認識できないという結果が得られた．

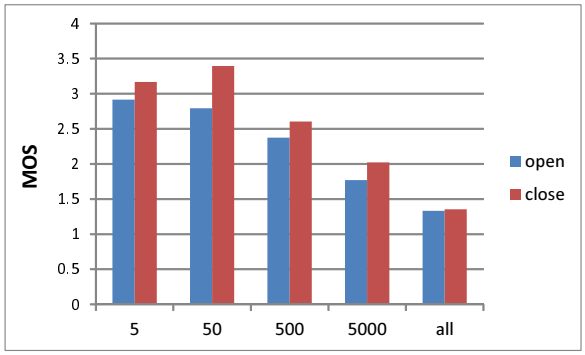

Fig. 5　Mean opinion score for subjective evaluations

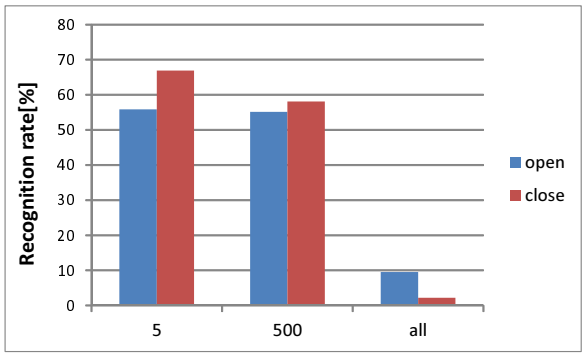

Fig. 6　Recognition rate for subjective evaluation

5　おわりに

本稿では，NMF を用いて無音声の発話映像から唇情報を抽出し，対応する発話音声へ変換を行った．音声のある動画から抽出した唇情報と音声情報を，基底の集合である辞書として用意し，入力した唇情報を唇辞書の基底の線形和で表現する．唇辞書の基底を対応する音声辞書の基底と取り替えることで，音声へと変換した．今回は目標音声と生成音声の MCD による客観評価実験に加え，主観評価実験を行った．ローカリティを導入することにより，不要な基底を用いずに済むため，より明瞭な音声が生成できることがわかった．今後はより自然な音声生成を目指すため，時間制約項などの導入を検討する．

謝辞　本研究の一部は，電気通信普及財団の助成を受け実施したものである．

参考文献

[1] H. McGurk, J. MacDonald,“Hearing lips and seeing voices,” Nature 264(5588), pp.746-748, 1976.

[2] J. Lin *et al.*, “Capturing human hand motion in image sequences,” IEEE Motion and Video Computing Workshop, pp. 99–104, 2002.

[3] M. K. Bashar *et al.*, “Unsupervised Texture Segmentation via Wavelet-based Locally Orderless Images (WLOIs) and SOM,” 6th IASTED International Conference COMPUTER GRAPHICS AND IMAGING, pp. 279–284, 2003.

[4] K. Nakamura *et al.*, “Speaking Aid System for Total Laryngectomees Using Voice Conversion of Body Transmitted Artificial Speech,” INTERSPEECH, pp. 1395–1398, 2006.

[5] 相原龍, 高島遼一, 滝口哲也, 有木康雄, “非負値行列因子分解による構音障害者の声質変換”, 日本音響学会 2012 年秋季研究発表会, 3-2-5 , pp. 331-334, 2012.

[6] C. Miyamoto *et al.* , “Multimodal Speech Recognition of a Person with Articulation Disorders Using AAM and MAF,” 2010 IEEE International Workshop on Multimedia Signal Processing (MMSP’10), pp. 517-520, 2010.

[7] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries, ” IEEE Trans. Image Process., vol. 15, pp. 3736-3745, Dec. 2006.

[8] J. Yang, *et al.* “Image super-resolution as sparse representation of raw image patches,” in Proc. IEEE Conf. Comput. Vision Pattern Recognition , Jun. 2008.

[9] H. Kawahara *et al.*, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequencybased F0 extraction: possible role of a repetitive structure in sounds,” Speech Communication, Vol. 27, No. 3–4, pp. 187– 207, 1999.

[10] D. D. Lee and H.S. Seung, “Algorithms for non-negative matrix factorization,” Neural Information Processing System, pp.556–562, 2001.