

β -NMF を用いた唇動画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, 無音声の唇動画画像からそれに対応する発話音声を作成する. また, NMF のコスト関数として β -divergence を用いた手法を提案する.

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.

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. しかし, 音声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不一致の場合, 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. 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, 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 [6] や超解像 [7] に用いられてきた Sparse Coding の代表的な手法である NMF [8] 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, 変換に際して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.

以降, 2 章では β -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 β -divergence を用いた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.

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

$\mathbf{x}_l$ 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.

$$\mathbf{X} \approx \mathbf{W} \mathbf{H} \quad (2)$$

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

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

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 のコスト関数として β -divergence を用い、以下のように定義される。

$$d_\beta(\mathbf{x}_l, \mathbf{W}\mathbf{h}_l) + \lambda \|\mathbf{h}_l\|_1 \quad s.t. \quad \mathbf{h}_l \geq 0 \quad (4)$$

第1項は $\mathbf{x}_l$ と $\mathbf{W}\mathbf{h}_l$ の間の β -divergence である。第2項は $\mathbf{h}_l$ をスパースにするための L1 ノルム正則化項である。 β -divergence はパラメータ β によって変化し、 $\beta = 2$ のときにはユークリッド距離、 $\beta = 1$ のときには Kullback-Leibler (KL) divergence、 $\beta = 0$ のときには Itakura-Saito (IS) divergence に対応する。 β -divergence は以下の式で表される。

$$d_\beta(x, y) = \begin{cases} \frac{1}{\beta(\beta-1)}(x^\beta + (\beta-1)y^\beta - \beta xy^\beta) & \beta \in \mathbf{R} \setminus \{0, 1\} \\ x \log(x/y) - x + y & \beta = 1 \\ (x/y) - \log(x/y) - 1 & \beta = 0 \end{cases}$$

(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 [9]。

$$h_k = \hat{h}_k \left(\frac{\sum_j w_{jk} x_k \hat{x}_k^{\beta-2}}{\sum_j w_{jk} \hat{x}_k^{\beta-1}} \right)^{\gamma(\beta)} \quad (5)$$

ここで $\gamma(\beta)$ は以下のように表される。

$$\gamma(\beta) = \begin{cases} \frac{1}{2-\beta} & \beta < 1 \\ 1 & 1 \leq \beta \leq 2 \\ \frac{1}{\beta-1} & \beta > 2 \end{cases} \quad (6)$$

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。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。

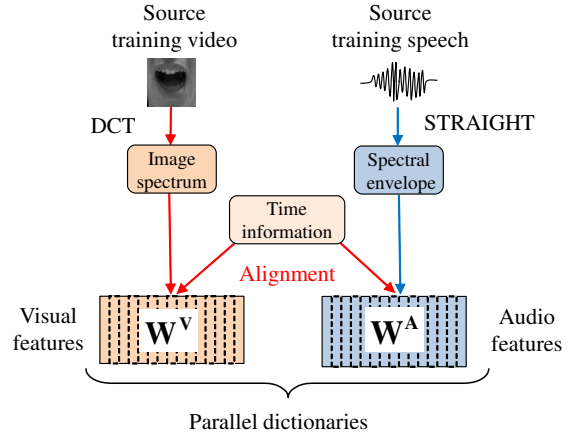


Fig. 1 Dictionary construction

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。

$$\Delta_{j,l} = \sqrt{(\mathbf{x}_l - \mathbf{w}_j)^2} \quad (7)$$

$\mathbf{x}_l$ 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。

$$\mathbf{S}_l = \text{nbest}_{\Delta_l}(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_J) \quad (8)$$

$$= \text{nbest}_{\Delta_l}(\mathbf{W}) \quad (9)$$

$\mathbf{S}_l$ 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

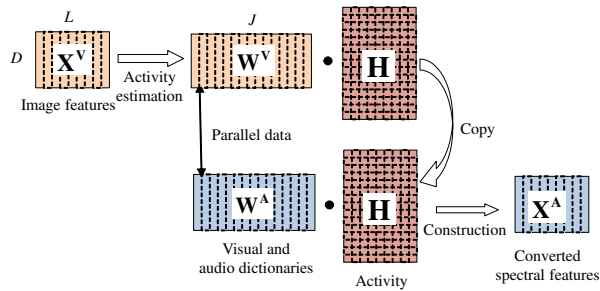


Fig. 2 Flow of conversion

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平行であるため、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連続数字発話 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。closed 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。 β -divergence のパラメータ β は 0, 1, 2 と値を変えて実験を行った。

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。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。

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. 3 に示す。Mel-CD は以下の式で表される。

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

mc_d^t と $\hat{mc}_d^t$ は目標音声と生成音声の MFCC 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. 3 より、最も良い

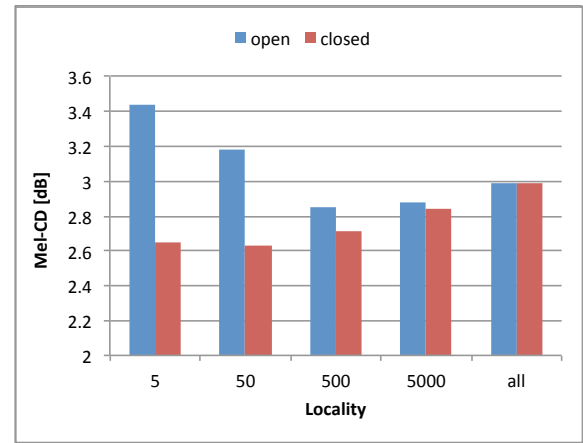


Fig. 3 Mel-cepstrum distortion

結果が得られたのは open, closed 実験で基底数がそれぞれ 500, 50 のときであった。closed 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につれて 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 に β の値を変えて行った実験結果を示す。ローカリティ制約による基底数は 5 である。open 実験, closed 実験ともに、 β を変えたときには大きな変化はみられなかった。これはローカリティ制約によって辞書に含まれる基底の数を制限したためであると考えられる。

生成音声の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、目標音声にどの程度近いかを 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

い, 2:遠い, 1:とても遠い)を行った. Fig. 5 に MOS による実験結果を示す. closed において, 客観評価実験と同様に基底の数が 50 のときに最も良い結果となった. これは open においてはより少ない基底数で音声の生成することにより, 明瞭な音声となっているためであると考えられる.

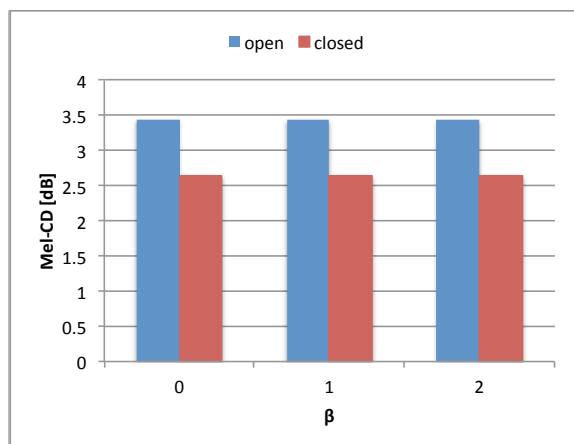


Fig. 4 Mel-cepstral distortion as a function of β

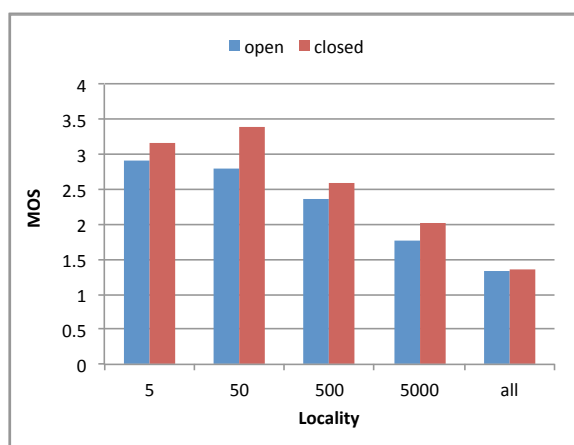


Fig. 5 MOS test as a function of the locality

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. NMF のコスト関数として β -divergence を用いた手法を適用し実験を行った. 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 Mel-CD による客観評価実験に加え, 主観評価実験を行った. β -divergence を導入する

ことにより変化は確認できなかったが, 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じることがわかった. 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. 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," *INTER-SPEECH*, 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] 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.
- [7] J. Yang, *et al.* "Image super-resolution as sparse representation of raw image patches," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, Jun. 2008.
- [8] D. D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," *Neural Information Processing System*, pp.556-562, 2001.
- [9] C. Févotte and J. Idier, "Algorithms for non-negative matrix factorization with the beta-divergence," *Neural Information Processing System*, pp.2421-2456, 2011.