社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Sparse Codingを用いた唇情報からの音声変換

相原　龍†　高島　遼一†　滝口　哲也††　有木　康雄††

† 神戸大学システム情報学研究科　〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
†† 神戸大学自然科学系先端融合研究環　〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: †{aihara,takashima}@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

あらまし　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，Sparse Coding を用いて，唇動画像から対応する発話音声へテキスト情報なしで変換を行う．事前に音声を含んだ発話映像から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．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，Sparse Coding 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．本稿では，唇情報から識別可能と考えられる母音について変換を行った．

キーワード　Sparse Coding，音声変換，リップリーディング，唇情報，音声変換，障害者支援

Sparse Coding-Based Voice Conversion from Lip Information

Ryo AIHARA†, Ryoichi TAKASHIMA†, Tetsuya TAKIGUCHI††, and Yasuo ARIKI††

† Graduate School of System Informatics, Kobe University 1-1 Rokkodai, Nada, Kobe 657-8501, Japan
†† Organization of Advanced Science and Technology, Kobe University 1-1 Rokkodai, Nada, Kobe 657-8501, Japan
E-mail: †{aihara,takashima}@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract　A technology to recognize speech content from lip motion is called visual speech recognition (VSR). VSR is an important communication method for people who have a handicap with hearing or speaking. In this paper, we propose a sparse-coding-based voice conversion method using lip motion without text information. Lip information and voices are extracted from videos, where they are used to construct lip dictionary and voice dictionary. Input lip information is represented by a linear combination of a small number of bases in the lip dictionary. The bases are replaced to coordinate bases in the voice dictionary, and they are recomposed to voice information. In this paper, we conducted vowel conversion because vowels are able to recognize from lip information.

Key words　Sparse Coding, Voice Conversion, Lipreading, Lip information

1. まえがき

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，もっぱら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].

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，近年研究が進んできた．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 [2]．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 [3]，文章読み上げシステム [4]，無喉頭音声変換 [5]，構音障害者のための声質変換 [6] など，その応用領域は幅広い．文献 [7] では，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 [8] や超解像 [9] に用いられてきた Sparse Coding 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．Sparse Coding 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．唇情報の抽出には，Haar-like 特徴量を用いた AdaBoost 手法 [10] による唇検出を行い，特定した唇領域に対して Discrete Cosine Transform (DCT) を行う．音声情報としては，音声分析合成 STRAIGHT [11] を用いて基本周波数・スペクトル包絡・非周期成分を抽出して用いる．

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．本稿では，唇情報のみから比較的判別可能と考えられる母音について，唇情報から音声への変換実験を行った．

以降，2 章で提案手法の流れを説明する．3 章で唇情報の抽出について，4 章で音声情報の抽出，再合成について述べる．5 章で Sparse Coding による変換手法について詳細を説明する．6 章で評価実験とその結果を示し，7 章で本稿をまとめる．

2. 提案手法の流れ

図 1 に提案手法の流れを示す．本手法には，学習段階と変換段階の 2 つの段階が存在する．

まず学習段階では，発話動画を画像情報と音声情報に分離する．画像情報は，Haar-like 特徴量を用いた AdaBoost 法 [10] による顔画像領域抽出を行う．その後，抽出した顔画像領域の下半分に対して，同様の手法で唇領域抽出を行う．これは，唇領域の誤検出を防ぐための処理である．抽出した唇画像は画面内のサイズ変動を受けないよう，縦横の比率を一定にしたままリサイズする．リサイズした画像に対し，グレイスケール化し，その領域に対して 2 次元 DCT を適用することで唇特徴量を得る．本研究では，16 × 16 ピクセルの 2 次元データに対し DCT を行い，得られた 256 次元から低次 128 次元を取り出して唇情報として用いる．各発話から得られた唇情報を結合したものを

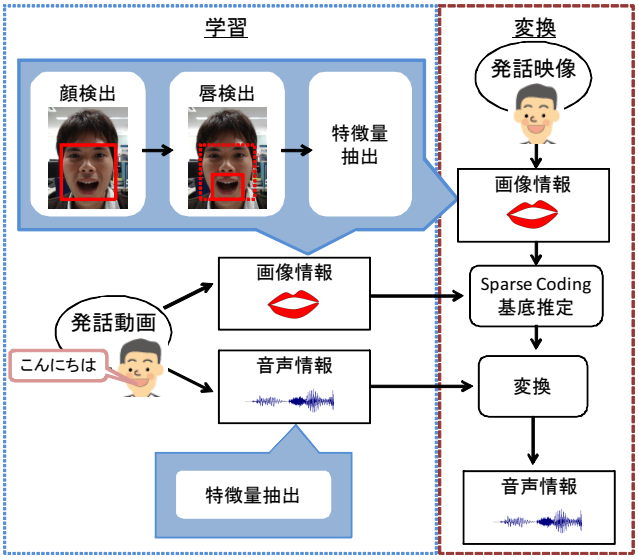

図 1　提案手法の流れ
Fig. 1　Proposed method flow

唇辞書行列として用いる．

音声情報は音声分析合成 STRAIGHT [11] で，スペクトル包絡，基本周波数，非周期成分に分離する．複数発話に対して上記の工程を行い，得られた特徴量を画像・音声ごと連結して，それぞれの基底の集合である唇辞書行列と音声辞書行列を得る．2 つの辞書行列は同一時系列のパラレルデータとなるよう，音声辞書行列を唇辞書行列のフレームレートに合わせてリサンプリングする．

変換段階では，音声のない発話映像のみが入力される．発話映像から学習段階と同様にして唇特徴量を得る．入力唇特徴量は Sparse Coding により，唇辞書行列から推定された基底の線形和で表現される．これにより入力唇情報は，唇辞書行列と係数ベクトルの積で表現される．係数ベクトルはスパース性を持っており，入力唇情報は唇辞書行列に含まれる少数の基底で表現される．

ここで得られた係数ベクトルを，音声辞書行列と掛け合わせる．2 つの辞書行列は同一時系列のパラレルデータであるため，音声辞書行列から唇辞書行列と共通の基底が選ばれることで，音声基底の線形和が出力され，入力された唇の動きと対応する音声情報が得られる．こうして選ばれた音声情報は再び音声分析合成 STRAIGHT で再合成され，音声として出力される．

3. 唇情報の抽出

画像内より発話対象者の顔領域と唇領域を抽出するアルゴリズムとして，Viola と Jones の提案する手法 [10] を用いた．この手法は，3 つの大きな流れに分けられる．

1. 積分画像を元に Haar-like 特徴量 [12] を抽出する．
2. AdaBoost 法を用いて上述の Haar-like 特徴量を元に作成される弱識別器から，目標の識別に有効なものを選出し，それらを線形結合することで強識別器を構築する．また，ここでの弱識別器とは，単純な閾値計算により入力が目標と一致するかしないかをバイナリ値で出力するものである．

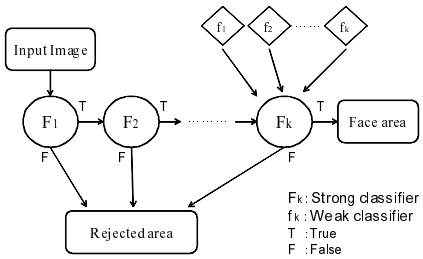

図 2　識別器のカスケード
Fig. 2　Cascade of classifier

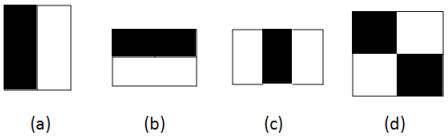

図 3　Haar-like 特徴量
Fig. 3　Haar-like feature

3. 弱識別器の結合数が少なく，演算は高速だが認識性能の低い弱識別器と，結合数が多く，演算は遅いが認識精度の高い強識別器を図 2 のようにカスケード状に繋ぐ．そうすることで，明らかに抽出目標でない領域は演算が高速な初期段階で除去され，目標に類似した領域は，精度の高い後段の識別器で詳細な判定を行う．

3.1　Haar-like 特徴量

Haar-like 特徴量とは，図 3 に示すような 4 種類の矩形において，矩形領域内の白色領域の平均輝度値から黒色領域の平均輝度値を引いた特徴量のことである．この 4 種類の矩形において，縦横比，スケール，位置を変化させたものを複数作り，それらを AdaBoost で弱識別器として学習させる．それらを線形結合することで強識別器を構築し，領域を検出する．領域内の輝度値の平均を効率よく求めるために，次項で述べる積分画像が使用される．

3.2　Integral Image

上述の Haar-like 特徴量の演算を高速に行う手法として，Viola らが提案した Integral Image (積分画像) がある [10]．ある画像において $I(x,y)$ を座標 (x,y) における輝度値と定義すると，座標 (x,y) における積分画像 $ii(x,y)$ は以下の式（1）で求められる．

$$ii(x,y) = \sum_{x^{'} \leqq x, y^{'} \leqq y} I(x^{'}, y^{'}) \tag{1}$$

また次の 2 つの式を用いることで，元の画像から一回の走査で計算が可能となる．

$$s(x,y) = s(x,y-1) + I(x,y) \tag{2}$$

$$ii(x,y) = ii(x-1,y) + s(x,y) \tag{3}$$

ただし，

$$s(x,-1) = 0, ii(-1,y) = 0$$

式（2）から解るように，$s(x,y)$ は y 軸方向への輝度値の累積和である．この累積和を式（3）が示すように x 軸方向に漸化

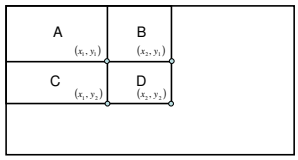

図 4　積分画像の計算法
Fig. 4　Computation of integral image

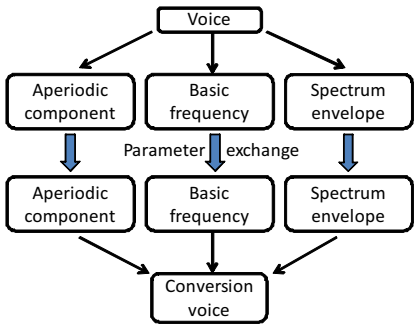

図 5　音声分析合成 STRAIGHT
Fig. 5　STRAIGHT

的に足し合わせることで，原点から座標 (x,y) までの輝度値の累積和が得られる．以上の式から，積分画像の座標 (x,y) の値は，4 点 $(0,0),(0,y),(x,0),(x,y)$ により囲まれる領域の輝度値の総和として捉えることができる．積分画像を用いることで，矩形領域の輝度値の総和を図 4，式（4）で表されるように，4 点の値から求めることが可能となり，上述した Haar-like 特徴量を高速に求めることができる．

$$\sum_{(x,y) \in X} I(x,y) = ii(x1,y1) + ii(x2,y2) - ii(x2,y1) - ii(x1,y2) \tag{4}$$

4.　音声情報の抽出・再合成

4.1　音声分析合成 STRAIGHT

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 [11]．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．

図 5 は STRAIGHT の概念を説明したものである．入力音声を基本周波数（Basic frequency），スペクトル包絡（Spectrum envelope），非周期成分（Aperiodic component）に分離し，変換を加えた後，合成する．本研究では，各発話ごと STRAIGHT で得られた基本周波数，スペクトル包絡，非周期性成分を垂直に連結し，全てのフレームについて水平に並べたものを音声辞書行列として用いる．

4.2　画像と音声のフレームレート問題

動画像のフレームレートは 15fps(67ms)，30fps(33ms) がよく利用され，また動画像の特徴量は，切り出された静止画像ごとに計算されるため，画像特徴量のサンプリングレートは 15Hz，30Hz となる．しかし，音声特徴量のサンプリングレートは 12000Hz と，対応がとれない．

そこで，学習・変換の際は音声のフレーム数を動画に合わせてリサンプリングを行う．変換して得られた音声特徴量は，フレーム間を 3 次スプライン補間 [14] して内挿することで音声本来のフレーム数に戻す．

5. Sparse Coding による変換

5.1 スパース生成モデル

Sparse Coding とは, 画像を複数のパッチに分解し，それぞれのパッチを基底画像群の線形和で近似する画像生成モデルである．Sparse Coding は, 画像処理において，雑音除去 [8]，ぼかし取り [15]，画像補完 [16]，超解像 [9] など様々な応用がなされている．

$\sqrt{p} \times \sqrt{p}$ ピクセルのパッチを，列ごとに縦にならべたベクトル $f \in \mathrm{R}^{\mathrm{p}}$ を考える．ここで，基底画像の集合である辞書 $D \in \mathrm{R}^{\mathrm{p}\times\mathrm{M}}$ を定義する．冗長性を持たせるため，$M > p$ とする．Sparse Coding では，パッチ画像 $f \in \mathrm{R}^{\mathrm{p}}$ を辞書 $D \in \mathrm{R}^{\mathrm{p}\times\mathrm{M}}$ に含まれる M 個の基底画像の線形結合で表す．このときの係数ベクトルを $a \in \mathrm{R}^{\mathrm{M}}$ とすると，次の式が成り立つ [17].

$$f = Da \tag{5}$$

ここで，D は既知で固定されているものと仮定すると，a は次式で表すことができる．

$$\hat{a} = \arg\min_{a} \|a\|_0 \ \ sub.to \ \ Da \approx f \tag{6}$$

ここで，$\|a\|_0$ は Sa 内の非ゼロ要素の個数を表す．式 (6) により得られた a は $\|\hat{a}\|_0 \ll p$ となり，非常にスパースなベクトルである．これはすべての入力信号は辞書からの少数の列の線形和で表すことができるという考えによるものである．

このモデルの $Da \approx f$ という表現をより正確に，またより明確にするために近似誤差として $\epsilon \in \mathrm{R}$ を導入する．この ϵ を用いて，$Da \approx f$ を $\|Da - f\|_2^2 \leqq \epsilon$ と書き換える．また，どの程度のスパース性を要求するのか定義するために $L \in \mathrm{R}$ を導入し，L 個以下の非ゼロ要素を用いたスパース表現のための a を $\|\hat{a}\|_0 \ \leqq L \ll n$ と表す．この ϵ, L, D の三つの変数によって表されるスパース性をもった信号を，(ϵ, L, D)-スパース信号とする．

ここで，f が (ϵ, L, D)-スパース信号に属すると仮定する．この画像パッチ f に対する a を次のように求める．

$$\hat{a} = \arg\min_{a} \|a\|_0 \ \ sub.to \ \ \|Da - f\|_2^2 \leqq \epsilon \tag{7}$$

f は (ϵ, L, D)-スパース信号に属しているため，スパース係数ベクトル $\hat{a}$ により表現することが可能である．この式 (7) を，μ を用いて次のように書き換える．

$$\hat{a} = \arg\min_{a}\{\|D\alpha - f\|_2^2 + \mu\|a\|_0\} \tag{8}$$

適切な μ を選んだ場合，式 (7) と式 (8) は等価である．

一般に，この問題は解くことが非常に難しいとされているが，BP/MP アルゴリズム (Matching and Basis Pursuit algorithms) を用いることにより効率的に近似解を求めることが可能である [18] [19]．近年ではより正確な近似解を求めることができる手法も提案されている [20]．ここでは α を求めるために，シンプルで効率的なことから OMP(Orthogonal Matching Pursuit) [19] という手法を用いる．

5.2 Orthogonal Matching Pursuit

OMP は信号を任意の辞書行列の単純な線形和で表すことのできる，再帰的なアルゴリズムである．OMP は Mallat と Zhang によって提案された Matching Pursuit (MP) を改良したもので，この改良によって有限フレームの辞書において，有限回の再帰で収束することが保証された．k 回目のモデルにおいて，辞書行列 D に含まれる基底ベクトルを x_i，対応する係数を a_i とすると，入力ベクトル f は以下のように表せる．

$$f = \sum_{n=1}^{k} a_n^k x_n + R_k f \ \ with \ \langle R_k f, x_n = 0, n = 1, ..., k\rangle \tag{9}$$

ここで，a_n^k における上付き文字の k はこれらの係数が繰り返し変化することを示している．$R_k f$ は k 回目のモデリングにおける誤差である．

OMP のアルゴリズムを図 6 に示す．本研究では，終了条件として $\delta = 1.0e^{-10}$ とした．このアルゴリズムにおいて，入力信号のフレーム数を N とすると，N 回目の繰り返し時には最も良い N 個の基底ベクトルが辞書行列から選ばれている．したがって，辞書行列のフレーム数を M とすると，あらゆる入力ベクトル f の基底推定の繰り返し回数が，辞書のフレーム数 M を超えることはなく，必ず有限回で収束する [19].

Initialization:

$f_0 = 0$, $R_0 f = f$, $x_0 = 0$, $a_0^0 = 0$, $k = 0$

(I) Compute $\{\langle R_k f, x_n\rangle; x_n \in D\}$.

(II) Find $x_{n_{k+1}} \in D$ such that

$|\langle R_k f, x_{n_{k+1}}\rangle| \geqq \alpha \sup_j |\langle R_k f, x_j\rangle|, \ 0 < \alpha \leqq 1$

(III) If $|\langle R_k f, x_{n_{k+1}}\rangle| < \delta$, $(\delta > 0)$ then stop.

(IV) Compute $\{b_n^k\}_{n=1}^k$, such that,

$x_{k+1} = \sum_{n=1}^{k} b_n^k x_n + \gamma_k$

and

$\langle \gamma_k, x_n\rangle = 0, \ n = 1, ..., k.$

(V) Set, $a_{k+1}^{k+1} = \alpha_k = \|\gamma_k\|^{\perp 2}\langle R_k f, x_{k+1}\rangle$,

$a_n^{k+1} = a_n^k - \alpha_k b_n^k, n = 1, ..., k,$

and update the model,

$f_{k+1} = \sum_{n=1}^{k+1} a_n^{k+1} x_n$

$R_{k+1} f = f - f_{k+1}$

(VI) Set $k \leftarrow k + 1$, and repeat (I)-(VI)

図 6　Orthogonal Matching Pursuit アルゴリズム
Fig. 6　Orthogonal Matching Pursuit Algorithm

5.3 唇情報から音声情報への変換

図 7 に，唇情報から音声情報への変換方法の概要を示す．

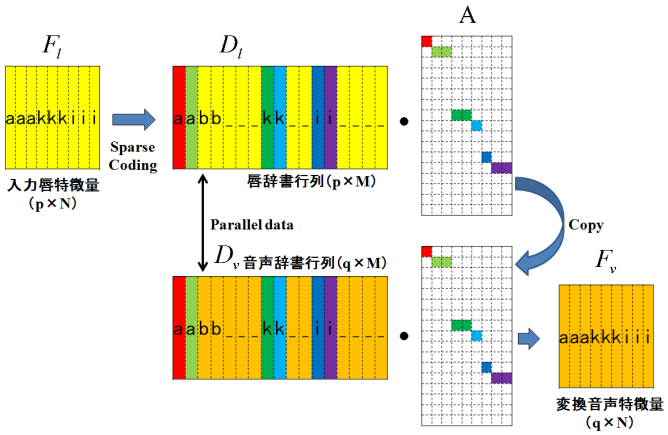

図 7　変 換 方 法

Fig. 7　Approach of conversion

一発話から取り出された唇特徴量を $F_l \in \mathrm{R}^{\mathrm{p}\times\mathrm{N}}$，唇辞書行列を $D_l \in \mathrm{R}^{\mathrm{p}\times\mathrm{M}}$，係数行列を $A \in \mathrm{R}^{\mathrm{M}\times\mathrm{N}}$，音声辞書行列を $D_v \in \mathrm{R}^{\mathrm{q}\times\mathrm{N}}$，求める音声特徴量を $X_v \in \mathrm{R}^{\mathrm{q}\times\mathrm{N}}$ とする．ここで p，q，N，M はそれぞれ唇特徴量の次元数，音声特徴量の次元数，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．

学習段階では，3 章で述べた方法で唇領域を特定する．特定された唇領域に DCT をかけ，低次 126 次元を取り出して唇情報とする．複数の単語について抽出された唇情報の行列は水平に並べられ，唇辞書行列となる．音声は 4.1 節で述べた方法で抽出する．1 つの単語から抽出した基本周波数，スペクトル包絡，非周期成分を垂直に連結し，全発話単語のフレームを水平に並べることで音声辞書行列となる．2 つの辞書行列のフレーム数は 4.2 節で述べた方法により統一される．こうして，唇辞書行列と音声辞書行列は同一時系列データから構成され，パラレルな関係が維持される．

変換する無音声の入力映像は，唇情報を抽出し，図 7 の上段に示すように Sparse Coding 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．図 7 の下段にあるように，推定された係数行列はコピーされ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．

6. 評 価 実 験

6.1 実 験 条 件

本稿では，発話映像として 2 母音の組み合わせ 20 組（以下，発話セットと呼ぶ）をそれぞれ 2 回づつ収録した．表 1 に発話セットの内容を示す．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 40cm，撮影機器は SONY HDR-CX590V で，

図 8　検出唇領域（左上から順に，a,i,u,e,o）

Fig. 8　Upper left to lower right:a,i,u,e,o

解像度は 320 × 240，フレームレートは 30fps を使用した．収録にあたってはピッチ変動の影響を避けるため，全ての発話をできる限り同じピッチで，はっきりと発話した．実験は，発話セットの 1 回目発話を学習し，2 回目発話を変換した．

画像特徴量は，3. 章で述べた唇特徴量抽出を行い，16 × 16 ピクセルの唇領域を抽出した後 DCT を行って得た低周波成分 126 次元と，その前後 2 フレームずつを加えた 630 次元を用いた．検出された唇領域の例を図 8 に示す．音声特徴量は，音声分析合成 STRAIGHT で得られた基本周波数・スペクトル包絡・非周期成分を用いた．

表 1　発話セット

Table 1　Word list

あい	いあ	うあ	えあ	おあ
あう	いう	うい	えい	おい
あえ	いえ	うえ	えう	おえ
あお	いお	うお	えお	おう

6.2 主観評価実験

変換音声を評価するために主観評価実験を行った．成人男女 6 人を対象に，音声の音素判別と，音素ごとの MOS (Mean Opinion Score) 評価基準に基づく 5 段階評価 (5:とても聞き取りやすい，4:聞き取りやすい，3:ふつう，2: 聞き取りにくい，1:とても聞き取りにくい) を行った．

表 2 に音素判別の知覚結果を示す．「a」と「o」については，85 %，96 %と高い知覚結果を示した．そのほかの音素については，48 %，60 %，50 %と誤って知覚されることが少なからずあった．知覚率が低い音素は，「a」と誤っている場合がいずれも 10 %以上である．これらは変換の際に，本来の音素と「a」が混ざってしまったと考えられる．

図 9 に誤って知覚された音素の唇領域の例を示す．図 8 と比較すると，同じ音素であって唇の形状が異なっていることがわかる．このことから，高い知覚結果を示した「a」,「o」と比較して誤って知覚された「i」,「u」,「e」は発話による変動が大きく，また学習データが少量であったためにうまく変換できない発話があったと考えられる．

表 3 に MOS 評価実験の結果を音素知覚結果ごとに示す．ほとんどの発話が MOS 評価基準で 3 以下であることがわかり，音質が劣化していることがわかる．誤って知覚された音素は正

図 9　誤って知覚された唇領域の例（左から順に，i,u,e）
Fig. 9　Examples of lips which are perceived erroneously. (Left to right:i,u,e)

しく知見された音素より MOS 評価基準が低い傾向がある．このことから，正しく変換できなかった音素は音質が劣化している，あるいは複数の音素が混ざってしまっていることが考えられる．音質の劣化原因は，音声のフレームレートを画像に合わせた後，補間を行ったためと考えられる．

表 2　音素知覚結果 [%]
Table 2　Results of vowel classification

Tar. \ Percept.	a	i	u	e	o
a	**85.42**	2.08	0	4.17	6.25
i	12.50	**47.92**	18.75	4.16	1.46
u	10.42	12.50	**60.42**	6.25	1.25
e	14.58	4.17	0	**50.00**	31.25
o	2.08	2.08	0	0	**95.83**

表 3　音質評価結果
Table 3　Results of MOS test on speech quality

Tar. \ Percept.	a	i	u	e	o
a	**2.28**	1.50	0	1.00	3.33
i	1.36	**2.59**	1.88	1.50	2.20
u	1.00	2.20	**2.79**	1.50	1.43
e	2.00	3.00	0	**3.2**	2.36
o	2.00	2.00	0	0	**2.27**

7. ま と め

本稿では，Sparse Coding 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．今回は唇情報から変換が可能と考えられる母音について変換を行った．母音「a」と「o」については十分に変換できたが，他の母音については変換できない発話もみられた．また，変換音声の音質にも劣化が見られた．

本稿では 1 名話者の母音変換のみであったため，今後は複数話者の子音を含んだ発話変換を行う．そのため，今後は高感度のビデオカメラを用いて，より詳細な唇情報を得ることを目指す．動画のフレームレートが音声のフレームレートに近づけば近づくほど，変換音声の音質は向上すると考えられる．また，カメラが高感度になれば，唇の変動が詳細に取得できるために子音の変換も可能になると考えられる．

文　献

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

[2] “読唇携帯電話”, http://www.jiten.com/dicmi/docs/k20/20082s.htm

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

[4] 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.

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

[6] 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.

[7] Chikoto Miyamoto, Yuto Komai, Tetsuya Takiguchi, Yasuo Ariki, Ichao Li, “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.

[8] 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.

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

[10] P.Viola, M. Jones, “Rapid Object Detection Using Boosted Cascade of Simple Features,” In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.1-9, 2001.

[11] 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.

[12] 三田雄志，金子敏充，堀修，“顔検出に適した Joint Haar-like 特徴の提案”，画像認識・理解シンポジウム (MIRU2005)，pp.104-111，2005.

[13] 稲田佳子，肖業貴，尾田政臣，“空間周波数を用いたベクトルマッチングによる顔画像の表情認識”，電子情報通信学会技術研究報告，Vol.101, No.385, pp.25-32, 2001.

[14] E. クライツィグ, 田村義保，“技術者のための高等数学　数値解析”， 培風館，1988.

[15] M. Figueiredo and R. Nowak, “Wavelet-based image estimation: An empirical Bayes approach using Jeffreys ’ noninformative prior,” IEEE Trans. Image Process. , vol. 10, pp. 1322-1331, 2001.

[16] M. Elad *et al.*, “Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA),” J. Appl. Comput. Harmon. Anal. , vol. 19, pp. 340-358, Nov. 2005.

[17] M. Elad, M. Figueiredo, Ma. Y, “On the role of sparse and redundant representations in image processing,” Proceedings of the IEEE 98(6), pp. 972-982, Jun. 2010

[18] S. Mallat and Z. Zhang, “Matching pursuit in a time-frequency dictionary,” IEEE Trans. Signal Process. , vol. 41, no. 12, pp. 3397-3415, Dec. 1993.

[19] Y. C. Pati *et al.*, “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition,” presented at the 27th Annu. Asilomar Conf. Signals, Systems, and Computers, 1993.

[20] D. L. Donoho *et al.*, “Stable recovery of sparse overcomplete representations in the presence of noise,” IEEE Trans. Inf. Theory, vol. 52, no. 1, pp. 6-18, Jan. 2006.