Kernel PCA を用いた残響下におけるロバスト特徴量抽出の検討

滝口哲也[†] 有木康雄[†]

本稿では，Kernel PCA（Principal Component Analysis）を用いた雑音や残響にロバストな特徴量抽出手法について検討する．画像の分野では，Kernel PCA は雑音除去などに優れた性能を発揮することが知られている（Mika ら (1999)）．たとえば，画質の粗い画像を高画質な画像へと変換を行う Super-Resolution なども提案されている（Kim ら (2004)）．一方，これまで Kernel PCA を音声特徴量抽出において適用した研究も行われてきている．たとえば Lima ら (2004, 2005) では，クリーン音声に対して MFCC 計算後，音声のスペクトル包絡成分に対応する低次ケプストラムに対して，Kernel PCA を適用した結果が報告されている．これまで，様々な雑音除去手法について研究が行われてきているが，残響や非定常雑音などを完全に除去することは困難であり，まだ課題が残されている．従来の雑音除去手法の多くは，スペクトル領域において演算が行われ，認識時には，(対数スペクトルに対し）離散コサイン変換が適用され，特徴量としてケプストラムが使われる．本稿では，残響音声に対してロバストな特徴量抽出法として，離散コサイン変換の代わりに Kernel PCA を用いた手法を提案し，残響下音声認識により，その有効性を示す．

A Study on Robust Feature Extraction Using Kernel PCA in Reverberant Environments

Tetsuya Takiguchi[†] and Yasuo Ariki[†]

We investigate robust feature extraction using kernel PCA (Principal Component Analysis). Kernel PCA has been suggested for various image processing tasks requiring an image model such as, e.g., denoising (Mika, et al. (1999)). Image denoising is the task of constructing a noise-free image from a noisy input image. From the point of view of a kernel PCA, image denoising can be also regarded as the same problem as image super-resolution (Kim, et al. (2004)). Also, an approach for feature extraction in speech recognition systems using kernel PCA has been proposed (Lima, et al. (2004, 2005)), where kernel PCA was applied to the low-dimension cepstrums. Much research for noise-robust feature extraction has been done, but it is difficult to remove the reverberation or non-stationary noise. The most commonly used noise-removal techniques are based on the spectral-domain operation, and then for the speech recognition MFCC (Mel Frequency Cepstral Coefficients) are computed, where DCT (Discrete Cosine Transformation) is applied to the mel-scale filter bank output. In this paper, we propose robust feature extraction based on kernel PCA instead of DCT. Its effectiveness is confirmed by word recognition experiments on reverberant speech.

1. はじめに

現在，会議などの書き起こし，ロボットとの対話など，ハンズフリーでの音声認識機能を使用するタスクに関する要求が，多く存在する．しかしながら，現状のシステムではユーザがマイクロフォンから離れて発話すると，入力音声は周囲雑音および残響の影響を受けて認識性能が劣化してしまう．またデスクトップマイクロフォンやピンマイクロフォンを用いた場合でも，ユーザが横を向くと音響伝達特性の影響により音声がひずみ，認識性能が劣化する場合がある．

従来，音声の伝達経路による影響に対処する方法として，ケプストラム平均減算法（Cepstrum Mean Subtraction: CMS）が使われている．この手法は，たとえば電話回線の影響などのように，伝達特性のインパルス応答が比較的短い場合には有効であるが，マイクロフォンから離れて発話した際には，残響の影響を受けて十分な性能が得られない．残響成分を除去する方法として，複数のマイクロフォンを利用し，逆フィルタを設計して観測信号から残響成分を除去する方法[5),6)] などが提案されているが，音響伝達特性のインパルス応答が，最小位相とならない場合があり逆フィルタの設計は難しい．また使用環境下においてコスト

† 神戸大学工学部
Faculty of Engineering, Kobe University

や物理的な配置状況により，複数のマイクロフォンを設置できない場合がある．シングルマイクロフォンによる残響除去手法として，短時間分析窓と長時間分析窓を組み合わせて残響を除去する手法[7]，調波構造に基づく逆フィルタ設計法[8]，パワートラジェクトリー残響モデルによる残響除去[9] などが提案されている．これらの手法の多くは，スペクトル領域において演算が行われ，音声認識を実行する際には，通常ケプストラム分析が用いられている．ケプストラム分析では，対数スペクトルに離散コサイン変換が適用される．その後，音声のスペクトル包絡成分に対応する低次のケプストラムが抽出され，音声認識の特徴量として使われる．

本稿では，離散コサイン変換よりも，ノイズロバストな特徴量抽出法として，Kernel PCA（Principal Component Analysis）を検討する．画像の分野では，Kernel PCA は雑音除去などに優れた性能を発揮することが知られている[1]．たとえば，画質の粗い画像を高画質な画像へと変換を行う Super-Resolution などが提案されている[2]．一方，これまで Kernel PCA を音声特徴量抽出において適用した研究も行われてきており，たとえば文献 3) では，クリーン音声に対して MFCC 計算後，音声のスペクトル包絡成分に対応する低次ケプストラムに対して Kernel PCA を適用した結果が報告されているが，ノイズロバスト性に関しては報告がされていない．本稿では，残響音声に対してロバストな特徴量抽出法として，MFCC における DCT（Discrete Cosine Transformation）の代わりに，Kernel PCA を用いた手法を提案し，残響下音声認識においてその有効性を報告する．

2. Kernel PCA による特徴抽出

2.1 提案手法

現在の音声認識システムでは，音声特徴量として MFCC（Mel Frequency Cepstral Coefficient）が広く用いられている．MFCC では，メル尺度フィルタバンクの短時間対数エネルギー出力系列に対して，離散コサイン変換（Discrete Cosine Transformation: DCT）が適用され，ケプストラムが得られる．さらに音声のスペクトル包絡成分に対応する低次ケプストラムのみを抽出し，音声認識における特徴量として使用される．

これまでの多くの雑音除去手法は，スペクトル領域において演算が行われ，認識時には（対数スペクトルに対し）離散コサイン変換が適用され，ケプストラムが求められる．そこで本稿では，よりノイズロバストな特徴量抽出法として，離散コサイン変換の代わりに Kernel PCA を用いた手法を検討する（**図 1**）．スペクトル上で PCA することで，エネルギーの強い主な音声成分は低次に集まり，雑音・残響成分は高次に集まる．この結果，PCA により雑音・残響除去が行われると期待できる．さらに，Kernel PCA では非線形写像を用いて高次元空間への写像を行うことにより，高精度な主成分の抽出が期待できる．

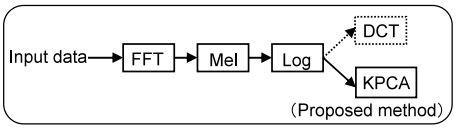

図 1 Kernel PCA による特徴抽出
Fig. 1 Feature extraction using Kernel PCA.

文献 3)，4) では，クリーン音声に対して MFCC 計算後，低次ケプストラムに対して Kernel PCA を適用している（つまり DCT による次元圧縮後に Kernel PCA を適用している）．一方，本手法では DCT を行わずに，対数スペクトルに対して Kernel PCA を適用することにより，より有効なスペクトル情報の抽出を試みる．

2.2 残響の抑圧

短時間分析によって得られたフレーム n，周波数 ω の観測音声を $X_n(\omega)$，クリーン音声を $S_n(\omega)$，雑音を $N_n(\omega)$ とすると，観測信号は以下のように表現される．

$$X_n(\omega) = S_n(\omega) + N_n(\omega) \tag{1}$$

ここで，観測信号 X に対して PCA を適用すると，

- クリーン音声 S の主なエネルギーは D 個の主な固有値に集中する，
- それ以外の固有値に対応する主な成分は，雑音である，

と期待できる．ここで，主な D 個の固有値に対応する固有ベクトルを $\mathbf{V} = [\mathbf{v}^{(1)}, \cdots, \mathbf{v}^{(D)}]$ とすると，この $\mathbf{V}$ を用いて以下のようなフィルタを考える．

$$\hat{S} = VX \tag{2}$$

このフィルタは，$\mathbf{V}$ 以外の部分空間内に存在する $N(\omega)$ の成分と直交関係にあるため，このフィルタリングにより雑音 $N(\omega)$ を抑圧することができる．また，主軸 $\mathbf{V}$ の計算をクリーンな音声信号のみから行えば，クリーン音声の構造のみを考慮したフィルタを作成することができる．

本稿では，上記フィルタリングを残響音声に適用する．短時間分析によって得られたフレーム n，周波数 ω の残響音声を $X_n(\omega)$，クリーン音声を $S_n(\omega)$ とす

る．ここで，残響音声を短時間分析の窓長内における影響と，窓長外からの影響の和として，以下のように近似的に表現する．

$$X_n(\omega) \approx S_n(\omega)\cdot H_0(\omega)+\sum_{d=1} S_{n-d}(\omega)\cdot H_d(\omega) \quad (3)$$

式 (2) のようなフィルタリングを行うには，式 (1) において，

- $S(\omega)$ と $N(\omega)$ は互いに無相関である．

と仮定している．式 (3) における第 1 項と第 2 項の相関性であるが，短時間分析の窓長がある程度短いもの（本実験では 32 msec）であると，第 2 項には，$n-1$ フレーム以前の信号に対する反射音が多く含まれるので，両者の相関は低下すると期待できる．そこで，短時間分析の窓長よりも遅れて到達する反射音を雑音として扱い，式 (2) のフィルタリングにより残響の抑圧を行うことが考えられる．しかし，未知環境下に対して，あらかじめ H_0 を知ることはできず，H_0 を含む第 1 項のデータの主軸を計算するのは困難である．したがって，本稿では対数変換を行い，第 1 項から H_0 を除去し，クリーン音声 S だけの項になるようにする．

$$\log X_n(\omega) = \log S_n(\omega) + \log\left\{H_0(\omega)+\frac{\sum_{d=1} S_{n-d}(\omega)\cdot H_d(\omega)}{S_n(\omega)}\right\} \quad (4)$$

主軸 $\mathbf{V}$ の計算には，あらかじめ音響モデルの学習に使用するクリーン音声の対数スペクトルを使うことができるので，式 (2) のフィルタリングにより式 (4) の第 2 項の残響成分の抑圧を行うことができる．さらに，Kernel PCA では非線形写像を用いて高次元空間への写像を行い，高次元空間にて PCA を行うので，より高精度な主成分の抽出が期待できる．本稿では，非線形写像による残響抑圧の効果について残響下音声認識実験により示す．

2.3 Kernel PCA[10)]

PCA（Principal Component Analysis）の目的は，データの本質的な構造を残しながら次元数を削減することにある．ただし，PCA はデータが非線形な構造を持つとき，有効に動作しない．そこで Kernel PCA では，非線形写像関数 $\mathbf{\Phi}$ を用いて，元の次元よりもはるかに大きな次元へ写像を行い，(データが線形表現可能な）高次元空間において PCA が行われる．

Kernel PCA の特徴の 1 つは，対象に対する事前知識をカーネル関数の形で表現することにある．クリーン音声の構造をカーネル関数で表現することができれば，雑音・残響音声を特徴空間へ写像した際に，クリーン音声の構造に適合した特徴量が得られ，そのほかの雑音成分が除去されると期待できる．

ここで，d 次元観測ベクトルを $\mathbf{x}_j$（j はフレーム番号）とすると，共分散行列 C は，

$$C = \frac{1}{N}\sum_{j=1}^{N} \bar{\mathbf{\Phi}}(\mathbf{x}_j)\bar{\mathbf{\Phi}}(\mathbf{x}_j)^T \quad (5)$$

$$\bar{\mathbf{\Phi}}(\mathbf{x}_j) = \mathbf{\Phi}(\mathbf{x}_j) - \frac{1}{N}\sum_{j=1}^{N} \mathbf{\Phi}(\mathbf{x}_j) \quad (6)$$

となる（N は全フレーム数）．C の固有値を λ，固有ベクトルを $\mathbf{v}$ とおくと，

$$\lambda\mathbf{v} = C\mathbf{v} \quad (7)$$

$$\lambda(\bar{\mathbf{\Phi}}(\mathbf{x}_k)\cdot\mathbf{v}) = (\bar{\mathbf{\Phi}}(\mathbf{x}_k)\cdot C\mathbf{v}), \quad k=1,\ldots,N \quad (8)$$

が得られる．また $\mathbf{v}$ は以下のようにサンプル点の線形結合で表現できる．

$$\mathbf{v} = \sum_{i=1}^{N} \alpha_i \bar{\mathbf{\Phi}}(\mathbf{x}_i) \quad (9)$$

式 (5) と (9) を式 (8) に代入すると左辺は，

$$\lambda(\bar{\mathbf{\Phi}}(\mathbf{x}_k)\cdot\mathbf{v}) = \lambda\sum_i \alpha_i \bar{\mathbf{\Phi}}(\mathbf{x}_k)\cdot\bar{\mathbf{\Phi}}(\mathbf{x}_i) = \lambda\sum_i \alpha_i \bar{K}_{ki} \quad (10)$$

となる．ここで，

$$\bar{K}_{ki} = \bar{\mathbf{\Phi}}(\mathbf{x}_k)\cdot\bar{\mathbf{\Phi}}(\mathbf{x}_i) \quad (11)$$

とした．また右辺は，

$$\begin{aligned}
&\bar{\mathbf{\Phi}}(\mathbf{x}_k)\cdot C\mathbf{v} \\
&= \bar{\mathbf{\Phi}}(\mathbf{x}_k)\cdot\frac{1}{N}\sum_j \bar{\mathbf{\Phi}}(\mathbf{x}_j)\bar{\mathbf{\Phi}}(\mathbf{x}_j)^T\sum_i \alpha_i\bar{\mathbf{\Phi}}(\mathbf{x}_i) \\
&= \bar{\mathbf{\Phi}}(\mathbf{x}_k)\cdot\frac{1}{N}\sum_i \alpha_i\left\{\sum_j \bar{\mathbf{\Phi}}(\mathbf{x}_j)\bar{\mathbf{\Phi}}(\mathbf{x}_j)^T\bar{\mathbf{\Phi}}(\mathbf{x}_i)\right\} \\
&= \frac{1}{N}\sum_i \alpha_i\left[\bar{\mathbf{\Phi}}(\mathbf{x}_k)\cdot\left\{\sum_j \bar{\mathbf{\Phi}}(\mathbf{x}_j)\bar{\mathbf{\Phi}}(\mathbf{x}_j)^T\bar{\mathbf{\Phi}}(\mathbf{x}_i)\right\}\right] \\
&= \frac{1}{N}\sum_i \alpha_i \cdot\sum_j \left\{\bar{\mathbf{\Phi}}(\mathbf{x}_k)\cdot\bar{\mathbf{\Phi}}(\mathbf{x}_j)\right\}\left\{\bar{\mathbf{\Phi}}(\mathbf{x}_j)\cdot\bar{\mathbf{\Phi}}(\mathbf{x}_i)\right\} \\
&= \frac{1}{N}\sum_i \alpha_i\sum_j \bar{K}_{kj}\bar{K}_{ji} \quad (12)
\end{aligned}$$

となる．したがって，式 (10) と (12) より，

$$N\lambda\alpha = \bar{\mathbf{K}}\alpha$$
$$\hat{\lambda}\alpha = \bar{\mathbf{K}}\alpha \tag{13}$$

となり，最終的に $\bar{\mathbf{K}}$ の固有値問題に帰着することになる（α は $\bar{\mathbf{K}}$ の固有ベクトルとなる）．ここで $N\lambda$ を $\hat{\lambda}$ とし，また $\bar{K}_{ki}$ を要素とする行列を $\bar{\mathbf{K}}$ とした．ただし，以下に示すように $\bar{K}_{ij}$ は K_{ij} から計算することが可能である．

$$\begin{aligned}\bar{K}_{ij} &= \bar{\mathbf{\Phi}}(\mathbf{x}_i)\cdot\bar{\mathbf{\Phi}}(\mathbf{x}_j)\\ &= (\mathbf{\Phi}(\mathbf{x}_i) - \frac{1}{N}\sum_{m=1}^{N}\mathbf{\Phi}(\mathbf{x}_m))\\ &\quad\cdot(\mathbf{\Phi}(\mathbf{x}_j) - \frac{1}{N}\sum_{n=1}^{N}\mathbf{\Phi}(\mathbf{x}_n))\\ &= \mathbf{\Phi}(\mathbf{x}_i)\cdot\mathbf{\Phi}(\mathbf{x}_j) - \frac{1}{N}\sum_{m=1}\mathbf{\Phi}(\mathbf{x}_m)\cdot\mathbf{\Phi}(\mathbf{x}_j)\\ &\quad -\frac{1}{N}\sum_{n=1}\mathbf{\Phi}(\mathbf{x}_n)\cdot\mathbf{\Phi}(\mathbf{x}_i)\\ &\quad +\frac{1}{N^2}\sum_{m,n=1}\mathbf{\Phi}(\mathbf{x}_m)\cdot\mathbf{\Phi}(\mathbf{x}_n)\\ &= K_{ij} - \frac{1}{N}\sum_{m=1}1_{im}K_{mj} - \frac{1}{N}\sum_{n=1}K_{in}1_{nj}\\ &\quad +\frac{1}{N^2}\sum_{m,n=1}1_{im}K_{mn}1_{nj}\end{aligned} \tag{14}$$

$$K_{ij} = K(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{\Phi}(\mathbf{x}_i)\cdot\mathbf{\Phi}(\mathbf{x}_j) \tag{15}$$
$$1_{ij} = 1 \quad \text{for all } i,j \tag{16}$$

よって，$\bar{K}_{ij}$ の行列表現は次式で与えられる．

$$\bar{\mathbf{K}} = \mathbf{K} - \mathbf{1}_N\mathbf{K} - \mathbf{K}\mathbf{1}_N + \mathbf{1}_N\mathbf{K}\mathbf{1}_N \tag{17}$$

$\mathbf{1}_N$ はすべての要素が $1/N$ である $N\times N$ 行列である．ここで，式 (15) の計算だが，これは元の入力空間のデータを非線形写像関数 $\mathbf{\Phi}$ を用いて高次元空間への写像を行い，その後，内積の計算を行うことになる．しかしながら，そのような計算は，計算量が膨大となり実際には困難である．そこで，カーネル法では，式 (24) のような多項式カーネルなどを用いて，非線形写像関数 $\mathbf{\Phi}$ の具体的な形を知る必要はなく，元の入力空間のデータのみから高次元空間における内積の計算が行われる．

次に，固有値を $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_N$ とし，それに対応する固有ベクトルを $\alpha^{(1)},\cdots,\alpha^{(N)}$ とした際，

$$\mathbf{v}^{(l)}\cdot\mathbf{v}^{(l)} = 1, \quad \text{for all } l = p,\cdots,N \tag{18}$$

を満たすように，α を正規化する（p 番目の固有値が，正の固有値の中で一番小さい値とする）．式 (9) と (13) より，式 (18) は

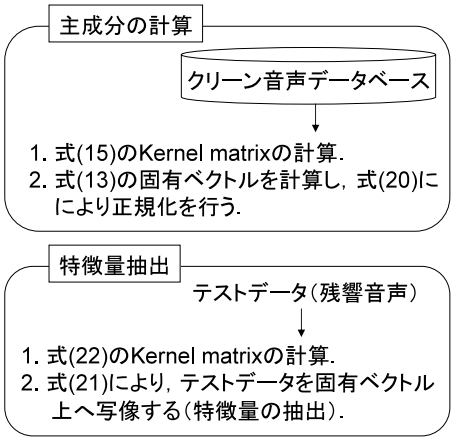

図 2 Kernel PCA の計算手順
Fig. 2 Procedure of Kernel PCA.

$$\begin{aligned}1 &= \sum_{i,j}^{N}\alpha_i^{(l)}\alpha_j^{(l)}(\mathbf{\Phi}(\mathbf{x}_i)\cdot\mathbf{\Phi}(\mathbf{x}_j))\\ &= \sum_{i,j}^{N}\alpha_i^{(l)}\alpha_j^{(l)}K_{ij}\\ &= (\alpha^{(l)}\cdot\bar{\mathbf{K}}\alpha^{(l)})\\ &= \hat{\lambda}_l(\alpha^{(l)}\cdot\alpha^{(l)})\end{aligned} \tag{19}$$

となる．よって，$\bar{\mathbf{K}}$ の固有ベクトル α に対して次式のように正規化を行う．

$$\hat{\alpha}^{(l)} = \frac{\alpha^{(l)}}{\sqrt{\hat{\lambda}_l}} \tag{20}$$

次に，高次元空間において主成分を抽出するために，テストデータ $\mathbf{y}$ の高次元における値 $\bar{\mathbf{\Phi}}(\mathbf{y})$ を，固有ベクトル $\mathbf{v}^{(l)}$ 上に写像する．

$$\begin{aligned}(\mathbf{v}^{(l)}\cdot\bar{\mathbf{\Phi}}(\mathbf{y})) &= \sum_{i=1}^{N}\hat{\alpha}_i^{(l)}(\bar{\mathbf{\Phi}}(\mathbf{x}_i)\cdot\bar{\mathbf{\Phi}}(\mathbf{y}))\\ &= \sum_{i=1}^{N}\hat{\alpha}_i^{(l)}\bar{K}^{test}(\mathbf{x}_i,\mathbf{y})\end{aligned} \tag{21}$$

ここで，同様に $\bar{K}^{test}$ も K^{test} から求めることができる．

$$\begin{aligned}\bar{K}_{ij}^{test} &= \left(\mathbf{\Phi}(\mathbf{y}_i) - \frac{1}{N}\sum_{m=1}^{N}\mathbf{\Phi}(\mathbf{x}_m)\right)\\ &\quad\cdot\left(\mathbf{\Phi}(\mathbf{x}_j) - \frac{1}{N}\sum_{n=1}^{N}\mathbf{\Phi}(\mathbf{x}_n)\right)\end{aligned} \tag{22}$$

$$\bar{\mathbf{K}}^{test} = \mathbf{K}^{test} - \mathbf{1}'_N\mathbf{K} - \mathbf{K}^{test}\mathbf{1}_N + \mathbf{1}'_N\mathbf{K}\mathbf{1}_N \tag{23}$$

テストデータ $\mathbf{y}$ のフレーム数が L の場合，$\mathbf{1}'_N$ は要素がすべて $1/N$ の $L\times N$ 行列となる．**図 2** に，

Kernel PCA の計算手順の概要を示す．

3. 認識実験

3.1 実験条件

評価用データとして残響音声を使用し，提案手法の有効性を検討する．残響音声の作成には，RWCP 実環境音声・音響データベース[11] より残響時間 470 ms のインパルス応答を使用した．マイクロフォンまでの距離は約 2 m，部屋の大きさは約 6.7 m × 4.2 m である．**図 3** にクリーン音声と残響音声の波形とスペクトルグラムを示す．カーネル行列の計算に使用したクリーン音声データのフレーム数は，$N = 2{,}500$ とした．これは音響モデルの学習データ（2,620 単語）からランダムに選択した．本実験では，カーネル行列の計算の際に次式の多項式カーネルを使用した．

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^p \qquad (24)$$

音声のサンプリング周波数は 12 kHz，窓幅は 32 ms，窓シフトは 8 ms とした．タスクは語彙 1,000 単語として，テストデータは男性話者 3 人が対象語彙を 1 回発声したものである．音響モデルは特定話者 HMM（54 音素 HMM）を使用した．HMM は 3 状態 3 ループ，各状態が 4 混合ガウス分布（対角共分散行列）とした．CMS 適用後の MFCC + ΔMFCC 32 次元での評価データに対する認識率は 63.9%（ベースライン）である．提案手法では，メルフィルタバンク出力 32 次元に対し Kernel PCA を適用した．得られた値を基本係数とし，基本係数+Δ 係数を音声認識の特徴量とした．

3.2 実験結果

図 4 に多項式カーネル次数 $p = 1$ の認識結果を示す（式 (24) において $p = 1$）．Baseline は，CMS 適用後の MFCC + ΔMFCC（32 次元）の結果である．DCT の代わりに Kernel PCA を適用することにより，主成分 16 個で 75.0%まで認識率が改善された．DCT よりも Kernel PCA の方が，ノイズロバストな特徴量抽出法であるといえる．また，主成分 32 個で 72.0%の認識率となり，このケースでは，次元数を増やしても，認識率の改善は得られなかった．認識時のパラメータは，主成分 16 個のときは基本係数 16 次元 + Δ 係数 16 次元となり，全体で 32 次元の特徴量となっている．**図 5** に多項式カーネル次数 $p = 2$ の認識結果を示す．平均認識率は，主成分 16 個で 76.8%となり，$p = 1$ と比べると，1.8%の認識率の改善が得られた．主成分 32 個でも 76.6%の認識率が得られており，speaker3 では，主成分 32 個の方が，16 個よりも高い認識率となった．

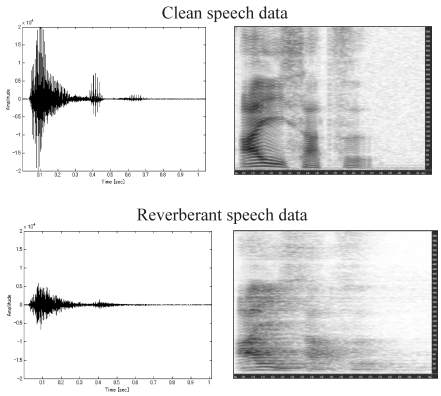

図 3 クリーン音声と残響音声（残響時間 470 ms）の波形データとスペクトルグラム．/a i sa tsu/
Fig. 3 Clean speech and reverberant speech (reverberation time = 470 ms): the speech waveform and spectrogram of the Japanese utterance /a i sa tsu/.

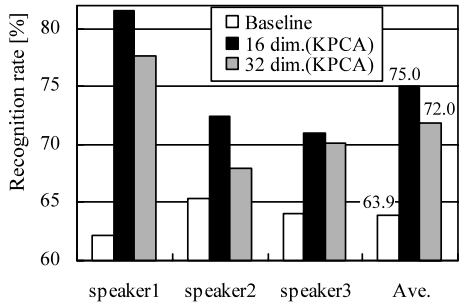

図 4 提案手法による残響音声認識率（多項式カーネル次数 $p = 1$）
Fig. 4 Recognition rates for the reverberant speech by the proposed method. ($p = 1$)

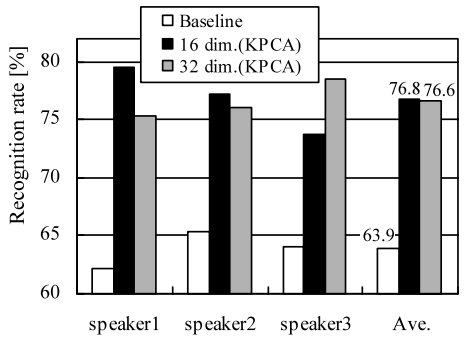

図 5 提案手法による残響音声認識率（多項式カーネル次数 $p = 2$）
Fig. 5 Recognition rates for the reverberant speech by the proposed method. ($p = 2$)

多項式カーネル次数 $p = 1$ の場合，普通の PCA にほぼ対応するが，実際に PCA を適用した結果，3 人の平均認識率は PCA 16 次元で 75%となった．これは図 4 に示されている Kernel PCA の結果と同等の結果となっている．したがって，普通の PCA と図 5

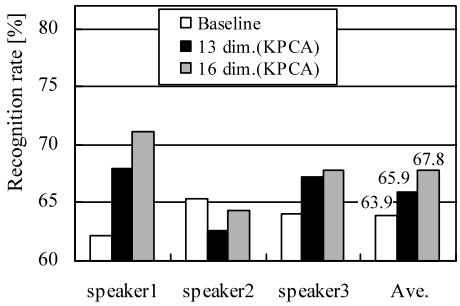

図 6 MFCC（Baseline）に Kernel PCA を適用した結果（多項式カーネル次数 $p = 1$）
Fig. 6 Recognition rates for the reverberant speech, where the kernel PCA is applied to MFCC (Baseline). ($p = 1$)

表 1 フレーム数と認識率の関係
Table 1 Relation between the size of frames and the recognition accuracy.

フレーム数	1,500	2,000	2,500	3,000
認識率 [%]	78.0	76.5	78.5	75.8

に示された Kernel PCA ($p = 2$) の場合の結果を比較すると，1.8%の改善が得られたことになる．

次に，Baseline の MFCC 基本係数 16 次元に対して，Kernel PCA を適用した結果を**図 6** に示す[3]（DCT 出力に対して，スペクトル包絡成分に対応する低次ケプストラムを抽出し，その後 Kernel PCA を適用する）．Baseline（63.9%）と比べて，3.9%認識率が改善されている．一方，提案手法（図 5）と比べると 9%近くの差がある．多項式カーネル次数を $p = 1$ とした場合，Kernel PCA は線形変換に近い形となるので，提案手法（フィルタバンク係数に対して $p = 1$ の Kernel PCA を適用）と，直交変換である DCT 後の MFCC に $p = 1$ の Kernel PCA を適用することは原理的にほぼ同じであると考えられる．したがって，ここでの認識率の差は，MFCC の時点で高次の係数を捨てていることが原因であると考えられる．

表 1 に，カーネル行列の計算に使用したクリーン音声のフレーム数と認識率の関係を示す（テスト speaker3，多項式カーネル次数 $p = 2$ の結果）．表 1 の結果より，$N = 2{,}500$ 以上増やしても認識率の改善は得られないので，本実験では $N = 2{,}500$ で十分であるといえる．また，**表 2** に以下に示すシグモイドカーネルを使用した際の認識率を示す（テスト speaker3，δ=0.01）．

$$K(\mathbf{x}, \mathbf{y}) = tanh(a\mathbf{x} \cdot \mathbf{y} - \delta) \qquad (25)$$

多項式カーネルの結果と比較して，認識率は低下しているのが分かる．また，パラメータが a と δ と 2 つあり，最適なパラメータの調整が難しいといえる．そのほか，ガウスカーネルもあるが，本実験において

表 2 シグモイドカーネルを使用した際の認識率 [%]
Table 2 Recognition rates with the sigmoid function.

次元数	16	24	32
a=0.0001	58.8	60.7	61.7
a=0.00005	71.6	69.7	68.3
a=0.00001	73.0	71.3	72.6
a=0.000005	71.6	72.7	73.4

表 3 主成分の計算に不特定話者の音声を使用した場合の認識率．() 内は特定話者の音声を使用した場合の認識率
Table 3 Recognition rates using speaker independent data. (*) shows the recognition rates using speaker dependent data.

次元数	16	24	32
$p = 1$	70.7 (71.0)	72.9 (74.0)	72.2 (70.1)
$p = 2$	72.0 (73.7)	73.7 (74.8)	74.4 (78.5)
$p = 3$	72.0 (75.6)	73.3 (74.1)	73.3 (76.1)

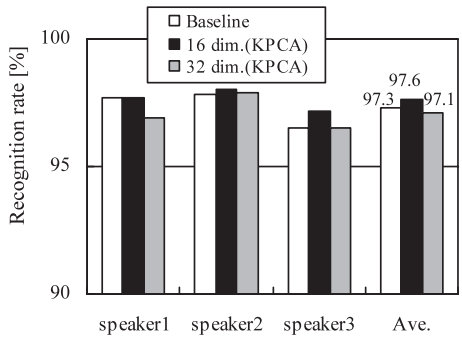

図 7 クリーン音声に対する認識結果（多項式カーネル次数 $p = 2$）
Fig. 7 Recognition rates for the clean speech by the proposed method. ($p = 2$)

認識率の改善は得られなかった．

次に，図 2 における主成分の計算の際に使用しているクリーン音声データベースを，特定話者から不特定話者に変更した際の性能変化を調べる．式 (13) の $\bar{\mathbf{K}}$ の計算に，ASJ データベース男性話者 25 人から各話者 100 フレームずつ合計 2,500 フレームを使用して，実験を行った．ここで，残響抑圧フィルタの精度を考察する意味で，主成分の計算のみに不特定話者の音声データを使用し，音響モデルの作成用データには，特定話者音声を使用する．**表 3** にテスト speaker3 の結果を示す．多項式カーネル次数 $p = 1$，2，3 に対して実験を行ったところ，不特定話者の音声を用いることにより，平均で 1.5%程度，精度が劣化しているのが分かる．これは，様々な話者の性質が主成分の計算に影響していると考えられる．

最後に，**図 7** にクリーン音声に対する提案手法の認識率を示す．Baseline では 97.3%，Kernel PCA では主成分 16 個で 97.6%となり，クリーン環境下では同

程度の結果が得られている．したがって本手法では，雑音成分のあるなしに関係なく，比較的安定した認識率を得ることができる．

4. おわりに

残響音声に対してロバストな特徴量抽出法として，MFCC における DCT の代わりに，Kernel PCA を用いた音声特徴量抽出法について検討した．残響下音声認識の結果（残響時間 470 msec），Baseline 63.9%の単語認識率に対して，提案手法により 76.8%まで認識率が改善された．また，普通の PCA と比較すると，1.8%程度の改善が得られた．今後は，様々な雑音除去手法との統合，より適切なカーネル関数の設定方法などの検討を行い，考察を続けていく．

参考文献

1) Mika, S., Scholkopf, B., Smola, A.J., Muller, K.-R., Scholz, M. and Ratsch, G.: Kernel PCA and de-noising in feature spaces, *Advances in Neural Information Processing Systems 11*, Kearns, M.S., Solla, S.A. and Cohn, D.A. (Eds), pp.536–542, MIT Press (1999).
2) Kim, K.I., Franz, M.O. and Schölkopf, B.: Kernel Hebbian Algorithm for Single-Frame Super-Resolution, *Statistical Learning in Computer Vision* (*SLCV 2004*), pp.135–149 (2004).
3) Lima, A., Zen, H., Nankaku, Y., Miyajima, C., Tokuda, K. and Kitamura, T.: On the Use of Kernel PCA for Feature Extraction in Speech Recognition, *IEICE Trans. Inf. & Syst.*, Vol.E87-D, No.12, pp.2802–2811 (2004).
4) Lima, A., Zen, H., Nankaku, Y., Tokuda, K., Kitamura, T. and Resende, F.G.: Applying Sparse KPCA for Feature Extraction in Speech Recognition, *IEICE Trans. Inf. & Syst.*, Vol.E88-D, No.3, pp.401–409 (2005).
5) Miyoshi, M. and Kaneda, Y.: Inverse Filtering of room acoustics, *IEEE Trans. ASSP*, Vol.36, pp.145–152 (1988).
6) Wang, H. and Itakura, F.: An Approach of Dereverberation using Multi-Microphone Sub-Band Envelope Estimation, *ICASSP*, pp.953–956 (1991).
7) Avendano, C., Tivrewala, S. and Hermansky, H.: Multiresolution channel normalization for ASR in reverberant environments, *Eurospeech*, pp.1107–1110 (1997).
8) 中谷智広，三好正人，木下慶介：調波構造に基づくモノラル音声信号のブラインド残響除去，電子情報通信学会論文誌 D-II, Vol.J88-D-II, No.3, pp.509–520 (2005).
9) 竹居 翼，松本 弘，山本一公：短時間スペクトル系列残響モデルの付加雑音下での推定と音声認識による評価，日本音響学会秋季講演論文集，3-7-23, pp.147–148 (2005).
10) Schölkopf, B., Smola, A. and Müller, K.-R.: Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, Vol.10, pp.1299–1319 (1998).
11) Nakamura, S., Hiyane, K., Asano, F., Nishiura, T. and Yamada, T.: Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition, *Proc. International Conference on Language Resources and Evaluation*, Vol.2, pp.965–968 (2000).

(平成 17 年 10 月 17 日受付)
(平成 18 年 4 月 4 日採録)

滝口 哲也（正会員）

平成 6 年岡山理科大学理学部応用数学科卒業．平成 8 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．平成 11 年奈良先端科学技術大学院大学博士後期課程修了．平成 11 年日本アイ・ビー・エム東京基礎研究所．平成 16 年神戸大学工学部講師．博士（工学）．ロバスト音声認識，マイクロフォンアレー等の研究に従事．日本音響学会，電子情報通信学会，IEEE 各会員．

有木 康雄（正会員）

昭和 49 年京都大学工学部情報工学科卒業．昭和 51 年同大学大学院修士課程修了．昭和 54 年同大学院博士課程修了．昭和 55 年京都大学工学部情報工学科助手．平成 2 年龍谷大学理工学部電子情報学科助教授，平成 4 年同教授．平成 15 年神戸大学工学部教授．工学博士．昭和 62〜平成 2 年エディンバラ大学客員研究員．画像処理，音声情報処理に従事．日本音響学会，人工知能学会，画像電子学会，IEEE 各会員．