

音素 PCA を用いた残響下における音声特徴量抽出

朴 玄信[†] 滝口 哲也[†] 有木 康雄[†]

[†] 神戸大学工学研究科 〒 657-8501 兵庫県神戸市灘区六甲台町 1-1

E-mail: †silentbattle@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

あらまし 本稿では, PCA (Principal Component Analysis) に基づく新しい固有音素部分空間法を用いた残響にロバストな特徴量抽出法を提案する. 雑音にロバストな特徴量抽出手法は多く提案されているが, 加法的あるいは乗法的雑音を完璧に抑圧するのは難しい問題として残されている. 特に非定常雑音を推定し, 抑圧するのは難しい. 音声認識分野において最も使われている特徴量は MFCC である. MFCC は対数メル周波数フィルタバンク出力に DCT を適用することにより得られる特徴量である. DCT の代わりに PCA を用いることで, 観測データの主な音韻成分は低次特徴量空間へ射影され, 加法的あるいは乗法的雑音は高次特徴量空間へ射影される. 本論文では PCA 手法と部分空間法をベースに音素 PCA 手法を提案する. 各音素データに対して PCA を行い, 得られた各音素部分空間をマージした固有音素部分空間を新しい特徴量空間とする手法である. この空間上のベクトルを音素ベクトルとする. 評価実験により, 提案手法の特徴量は従来手法より残響時間が長くなるほど有効であることが確認できた.

キーワード 主成分分析, 部分空間法, 音素ベクトル, 残響抑制, 単語認識

Feature Extraction Using Phoneme PCA for Reverberant Speech Recognition

Hyunsin PARK[†], Tetsuya TAKIGUCHI[†], and Yasuo ARIKI[†]

[†] Graduate School of Engineering, Kobe University Rokkodaicho 1-1, Nada-ku, Kobe, Hyogo, 657-8501 Japan

E-mail: †silentbattle@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract In this paper, we propose a novel subspace approach for robust speech feature extraction in reverberant environments using PCA. While much research for robust speech feature extraction has been done, it remains difficult to completely remove additive or convolutional noise. Particularly, it is difficult to estimate and suppress nonstationary noise. The most commonly used speech feature for speech recognition is MFCC that is computed applying DCT to the mel-scale filter bank output. Using PCA instead of DCT, where the main speech element is projected onto low-order features while the additive or convolutional noise is projected onto high-order features, speech recognition is improved. We propose Phoneme-PCA that is applying PCA to specific phoneme data set and merging subspaces representing unique feature in each phoneme. The performance evaluation shows that the proposed method provides a higher noise reduction and speech recognition accuracy compared with conventional methods.

Key words PCA, Subspace method, Phoneme vector, Suppression of reverberation, Word recognition

1. はじめに

近年, 音声認識システムはクリーン音響環境において優れた性能を示している. しかし, ユーザがマイクロフォンから遠い位置で発話すると, 加法的雑音や乗法的雑音の影響で認識率は急激に低下する.

この問題を解決するために, 多くの研究が行われてい

る [1] [2] [3] [4] [5] [6] [7]. 例えば, CMS (Cepstral Mean Subtraction) や RASTA は電話回線や短時間インパルス応答をもつマイクロフォンにおいて, その効果が確かめられている. しかし, インパルス応答時間が音響分析窓長よりはるかに長い残響下で観測された音声データに対しては十分な性能が得られない.

その他, PCA や KPCA (Kernel PCA) [8] をベースにした

観測データ内の特定情報だけを強調する手法もこれまでに提案されている [9] [10] [11] . 特定話者のクリーン音声に音韻性と話者性だけが含まれていて、話者内で大きい分散（低次）をもつ部分空間を音韻部分空間、小さい分散（高次）をもつ部分空間を話者部分空間と仮定すると、この二つの空間は無相関である。直交基底ベクトルをもとめる PCA (Principal Component Analysis) を用い、観測データを音韻部分空間へ射影することで、音韻性を強調することができる。

本研究では、PCA 手法と部分空間法をベースに残響下音声認識における新しい音声特徴量抽出法を提案する。各音素データに対し PCA を行い、音素性を強調する部分空間を作成した後、各音素部分空間をマージし、この全音素部分空間へ射影することで各音素空間の相関情報も取り入れることができる。本研究ではこの全音素部分空間を作成する手法を音素 PCA 法と呼ぶことにする。

2 章は提案手法のベースになる PCA による残響抑制法について概説した後、3 章で提案手法である音素 PCA について説明し、4 章で評価実験を行い提案手法の有効性を確かめる。最後に 5 章でまとめとして今後の課題について述べる。

2. PCA による残響抑制

一般的に観測されたひずみ音声のスペクトル $X_n(\omega)$ は次のようにクリーン音声と乗法性雑音の積で表される。

$$X_n(\omega) = S_n(\omega) \cdot H_n(\omega) \quad (1)$$

ここで $S_n(\omega)$ と $H_n(\omega)$ は周波数 ω , n 番目フレーム (n 番目分析窓) における、クリーン音声短時間スペクトルと乗法性雑音 (音響伝達関数) を表す。しかし、残響時間が分析窓長より長い環境で観測されるスペクトルは式 (2) のように、式 (1) の右辺に一定時間過去の音声信号に対する反響音が加算される。

$$X_n(\omega) = S_n(\omega) \cdot H_n(\omega) + \sum_{d=1}^N S_{n-d}(\omega) \cdot H_{n-d}(\omega) \quad (2)$$

ここで S_{n-d} と H_{n-d} は $(n-d)$ のフレームにおけるクリーン音声信号と乗法性雑音であり、 N はインパルス応答の長さを表す。

式 (2) の両辺に対数演算を行い、右辺を式 3 のようにクリーン音声項とそれ以外の項に分ける。

$$X_{\log-n}(\omega) = S_{\log-n}(\omega) + \log \left\{ H_n(\omega) + \frac{\sum_{d=1}^N S_{n-d}(\omega) \cdot H_{n-d}(\omega)}{S_n(\omega)} \right\} \quad (3)$$

ここで、 $X_{\log-n}(\omega)$ と $S_{\log-n}(\omega)$ は n 番目フレームにおける観測信号の対数スペクトルとクリーン音声信号の対数スペクトルである。

次に、PCA を用いたフィルタリング手法は次式のように観測信号と射影行列の内積となる。

$$\hat{S} = V^T X_{\log} \quad (4)$$

フィルタ V は、クリーン音声の相関行列 (共分散行列) の固有値展開によって得られた固有ベクトルから構成される。これらの固有ベクトルは、大きい順でソートされた固有値に対応する固有ベクトルであり、音韻性を表す次元が L 次元だとすると次のように表現できる。

$$V = [v^{(1)}, v^{(2)}, \dots, v^{(L)}] \quad (5)$$

\hat{S} と X_{\log} は縦ベクトルである。 $\hat{S} \in \mathbb{R}^L$, $X_{\log} \in \mathbb{R}^P$, $V \in \mathbb{R}^{P \times L}$ そして $L < P$. クリーン音声データ全体に対し PCA を適用すると、音韻性は分散が大きい (エネルギーが強い) 固有ベクトルで表され、雑音性は分散が小さい (エネルギーが弱い) 固有ベクトルで表される。また、音韻性と雑音性を表す固有ベクトル間は相関が低い。その結果、観測データの音韻性を表す成分は部分空間 V へ射影され、雑音成分はその補空間 $[v^{(L+1)}, v^{(L+2)}, \dots, v^{(P)}]$ へ射影される。これは、PCA により得られる新しい軸は互いに直交であるためである。式 (3) の右辺の第 2 項は残響時間が長くなると第 1 項との相関が低いと仮定できる。その時式 (4) のように、観測信号からの現在フレームの音韻情報を強調することができる。

音声認識システムでは *MFCC* (Mel-Frequency Cepstrum Coefficient) がよく用いられる。これは対数メル周波数フィルタバンク出力 (*FBANK*) に離散コサイン変換 (Discrete Cosine Transform: *DCT*) を適用した特徴量である。*MFCC* の低次元は時間領域におけるスペクトルの包絡を表す。*FBANK* に *DCT* を適用代わりに、PCA による射影変換を行った特徴量が残響にロバストであることが [11] で確認されている。

3. 音素 PCA による残響抑制

現在音声認識システムにおける特徴量空間は各地域の言語によることなく、*MFCC* が一般的に使われている。しかし、各地域の言語によって用いている音素の種類には差が存在する。音素モデル作成前の特徴量抽出段階で、特徴量空間に音素種類情報を取り入れることで、各地位の言語に特化した特徴量空間を作成できる。さらにこの空間は、音声データの音素性を強調する空間であるため、雑音の抑圧性能も期待できる。

2 章で、PCA による残響抑制法について概説した。PCA による従来手法はクリーン音声データに対して PCA を行うが、提案手法の音素 PCA はクリーン音声データの中の音素毎のデータに対して PCA を行なう。音声データに対して PCA を適用すると音韻成分は分散が大きい (変動が大きい) 軸に現れるが、音素データに対して PCA を適用すると音素成分は分散が小さい (変動が少ない) 軸に現れる。なぜなら、音声データは多くの音素データからなっていて、音素を区別するために有効な成分 (音韻成分) は変動の大きい成分であるが、音素データは特定音素だけのデータであるため、その音素を特定するために有効な成分 (音素成分) は変動の少ない共通成分になるからである。従って、提案手法では分散が小さい固有ベクトルを音素部分空間の基底ベクトルとする。各音素の部分空間を求めた後、これらの部分空間を原点を合わせてマージし、全音素部分空間を作る。図 1 のように、全音素部分空間上の各音素デー

たは、同一音素部分空間の原点付近に射影され、他の音素部分空間の原点から離れたところに射影される。

図2と図3は/a i d a/と発音した元のデータのスペクトルと、全音声データに対するPCAと音素/a/データに対するPCAを適用し得られた射影行列でフィルタリングした後、逆変換で得られたスペクトルである。横軸が時間、縦軸が周波数で、元の周波数は256次元、射影された空間は50次元になっている。/a i d a/と発音した元のデータのパワースペクトルが図2の左図で、音声データに対するPCAで得られた空間へ射影した値を逆変換したパワースペクトルが右図である。PCAにより元のデータのフォルマントが強調されていることが分かる。図3では音素/a/データに対してPCAを行い得られた分散の大きい固有ベクトル(左図)と分散の小さい固有ベクトル(右図)で構成されたフィルタによるスペクトルである。左図より右図の/a/のフォルマントの形が元の形と似ていることがわかる。分散が小さい(エネルギーが弱い)軸へ射影されたため、パワースペクトルの値は低い音素/a/のフォルマントの形は残っている。

全音素部分空間は各音素部分空間をマージした空間であるため、定義した音素数に比例して大きくなる。従って本稿では得られた音素ベクトルに対し次元削減の意味で、もう一回PCAを適用する。従って、提案手法では2回のフィルタリングが行われることになる。2回のフィルタリング時に用いる射影行列を V_p と V_c とする。 V_p は各音素データに対しPCAを行い得られた各音素部分空間を統合した空間の基底ベクトルである。 V_c は V_p により大きくなった次元を削減するための射影行列である。

本論文では、音素PCAを行い得られた全音素部分空間を固有音素部分空間(Eigen Phoneme Subspace: EPS)、この固有音素部分空間へ射影された特徴ベクトルを音素ベクトル(Phoneme Vector: PV)と呼ぶことにする。

3.1 音素ベクトル

まず、 i 番目音素の平均ベクトルを $M_{p,i} \in \mathbb{R}^P$ とし、音素部分空間 $V_{p,i} \in \mathbb{R}^{P \times K}$ を次のように定義する。

$$V_{p,i} = [v_{p,i}^1, v_{p,i}^2, \dots, v_{p,i}^K] \quad (6)$$

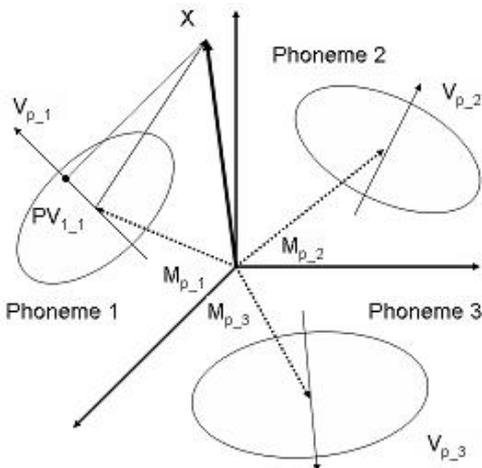


図1 Observation space and phoneme subspaces.

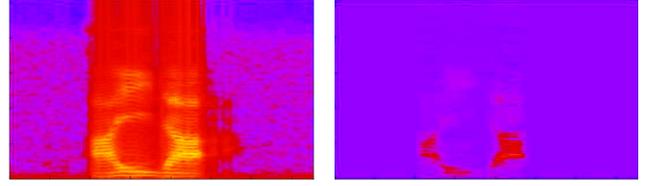


図2 Speech spectrogram of original and by PCA filter of the clean Japanese utterance /a i d a/.

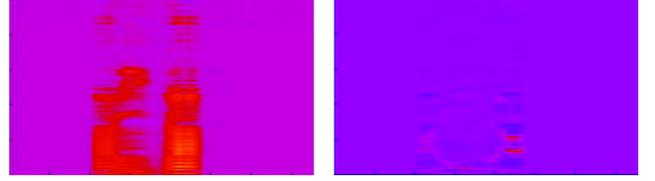


図3 Speech spectrogram by major and minor /a/-PCA filter of the clean Japanese utterance /a i d a/.

ここで P は元の特徴量空間の次元、 K は各音素部分空間の次元、 $v_{p,i}^j \in \mathbb{R}^P$ ($j = 1, 2, \dots, K$)は小さい順でソートされた固有値 K 個に対応する固有ベクトルで、縦ベクトルになっている。これは i 番目音素部分空間の基底ベクトルといえる。

続いて、マージした全音素部分空間 $V_p \in \mathbb{R}^{P \times (K \times N)}$ と全音素平均ベクトル $M_p \in \mathbb{R}^{P \times N}$ は次のようになる。

$$V_p = [V_{p,1}, V_{p,2}, \dots, V_{p,N}] \quad (7)$$

$$M_p = [M_{p,1}, M_{p,2}, \dots, M_{p,N}]$$

ここで N は音素種類の数を表し、特徴空間設計時に任意に変更が可能である。

観測信号 $X_{log} \in \mathbb{R}^P$ を縦ベクトルとして、次のように固有音素部分空間へ射影する。

$$PV_1 = \begin{bmatrix} PV_{1,1} \\ PV_{1,2} \\ \vdots \\ PV_{1,N} \end{bmatrix} = \begin{bmatrix} V_{p,1}^T [X_{log} - M_{p,1}] \\ V_{p,2}^T [X_{log} - M_{p,2}] \\ \vdots \\ V_{p,N}^T [X_{log} - M_{p,N}] \end{bmatrix} = \begin{bmatrix} V_{p,1}^T X_{log} - V_{p,1}^T M_{p,1} \\ V_{p,2}^T X_{log} - V_{p,2}^T M_{p,2} \\ \vdots \\ V_{p,N}^T X_{log} - V_{p,N}^T M_{p,N} \end{bmatrix} = V_p^T X_{log} - C_p \quad (8)$$

ここで、 $C_p = [C_{p,1}^T, C_{p,2}^T, \dots, C_{p,N}^T]^T$ 、 $C_{p,i} = V_{p,i}^T M_{p,i}$ 。 $PV_1 \in \mathbb{R}^{N \times K}$ は固有音素部分空間における音素ベクトルである。

3.2 音素ベクトルの次元削減

前節で V_p 、 C_p 、 PV_1 を定義した。この時 PV_1 の次元数は

$N \times K$ であり、次元数が高い問題が残されている．そこで、クリーン音声からの特徴量 PV_1 に対してもう一回 PCA を行い、次元圧縮を試みる．次元圧縮のための PCA で得られた射影行列を V_c とすると、低次元空間へ射影された新しい特徴量 PV_2 は次式のように表される．

$$PV_2 = V_c^T PV_1 \quad (9)$$

ここで、 $V_c = [v_c^1, v_c^2, \dots, v_c^{K'}]$ 、 $v_c^j \in \mathbb{R}^P$ ($j = 1, 2, \dots, K'$) は大きい順にソートされた K' 個の固有値に対応する固有ベクトルで、縦ベクトルである．次元圧縮の意味での PCA であるため、音素 PCA ステップ 1 の時とは逆に、分散が大きい順で射影行列を作成している． K' の値を調整しながら実験することで、最適値を見つけることができる．

4. 認識実験

4.1 実験データ

残響下の日本語 500 単語の認識実験を行い、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 [12] より残響インパルス応答を用いた．残響インパルス応答とクリーン音声を畳み込み演算することで残響音声を得られる．残響インパルス応答の例を表 1 に示す．

表 1 Reverberant conditions

Time (ms)	Room
50	Anechoic room
120	Echo room(cylinder)
300	Echo room(panel)
380	Echo room(cylinder)
470	Tatami - floored room (S)
600	Tatami - floored room (L)
1300	Echo room(panel)

4.2 実験条件

音声波形はハミング窓を用い短時間離散フーリエ変換を行うことでスペクトルに変換される．音響分析条件の詳細を表 2 に示す．スペクトルに対する対数メルフィルタバンク出力 32 次元 ($FBANK$) をベース特徴量とする．これらの音響分析と特徴量抽出には HTK-toolkit [13] を用いた．

表 2 Analysis conditions

Sampling bits	16 bits
Sampling rate	12 kHz
Window width	32 ms
Window shift	8 ms

射影行列 V 、 V_p 、 V_c と平均ベクトル M_p は主成分分析により求まる．クリーンの音声データから得られた特徴量 $FBANK$ を用いてこれらは計算される．その時、計算に用いるフレーム数と特徴量の次元数を表 3 に示す．本実験では 4 つの特徴量 ($MFCC$ 、 PCA 、 PV_1 、 PV_2) を用い、比較実験を行う． $MFCC$ は音声認識分野で広く使われている特徴量で、 $FBANK$ に離

表 3 The number of frames used to calculate filter and feature dimension

Filter (Feature)	Frames	Dimension
($FBANK$)	—	32
DCT ($MFCC$)	—	16
V (PCA : Feature by PCA)	3000	16
V_p (PV_1 : PV on EPS)	100×54	5×54
V_c (PV_2 : PV on compressed EPS)	3000	Adjusting

散コサイン変換 (DCT) を行うことで得られる．又、 PCA はベース手法である PCA による残響抑制手法による特徴量で、射影行列 V から計算できる．最後に PV_1 と PV_2 は提案手法である音素 PCA を用いて得られた固有音素部分空間 (EPS) における特徴量で、射影行列 V_p と V_c から求まる．本実験では、日本語の音素数は 54 (N) に、 V_p における各音素部分空間の次元数は 5 (K) に固定する．各音素部分空間射影行列は、音素ごと 100 フレームを用いて作成するが、フレーム数が 100 より少ない音素に対しては、全フレームを用い射影行列を計算する．

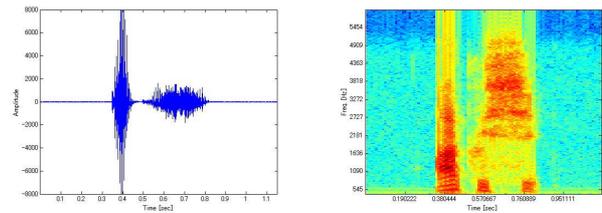


図 4 Original speech: the speech wave form and spectrogram of the Japanese utterance /akusyu/.

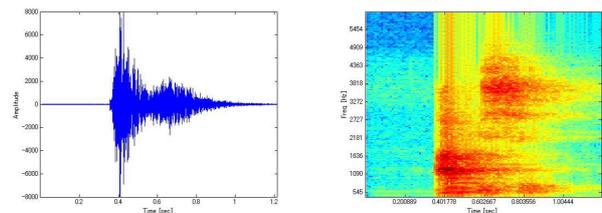


図 5 Reverberant speech (reverberation time = 0.47 sec): the speech wave form and spectrogram of the Japanese utterance /akusyu/.

図 4 と図 5 はクリーン音声と残響音声の波形とパワースペクトルを表している．残響音声のスペクトルとクリーン音声のスペクトルを比較してみると、残響音声のスペクトルの各フレームにおいて、残響時間分の過去のフレームの影響でエネルギーが強くなっていることが分かる．特に、発話後半の部分で、クリーン音声の低周波数のパワーは弱くなっているが、残響音声の低周波数のパワーには過去発話の残響成分が残っている．そのため、クリーンの音声モデルとのズレが生じ認識が難しくなる．

本実験では、4 つの比較特徴量の静的特徴量と、動的特徴

量を用いた．さらに，各特徴量には CMS を適用した．ATR の日本語音素バランス単語発話データベース (SET-A) から，特定話者の 2,620 単語クリーン発話を用い 54 個の音素モデル (HMM) を作成する．各音素 HMM はモノフォンモデルであり，各状態で 4 混合正規分布，状態数は 3 にした．評価実験は，モデル学習時と同じ特定話者で学習データと違う 500 のクリーン単語と各残響単語を用いて行った．

4.3 実験結果

最初に，特徴量 $MFCC$ ， PCA ， PV_1 を用いて残響時間を変化させながら単語認識実験を行った．その結果を図 6 に示す．横軸が残響時間，縦軸が単語認識率になっている．残響時間が短い時は，クリーンの時とほぼ変わらず高い認識率が得られた．しかし残響時間が 300ms より長くなると，全手法において認識率の低下が起こるが，提案手法の特徴量は従来手法の特徴量より，高い認識率を示している．

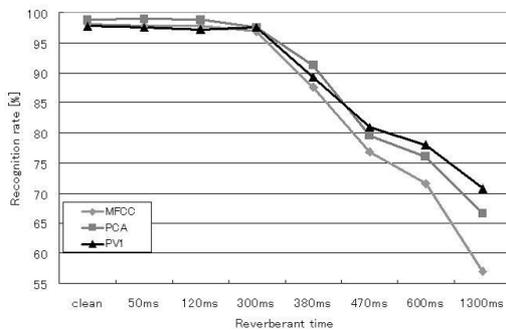


図 6 Results of word recognition as a function of reverberant time.

次に，クリーンの評価音声データに対して，特徴量 PV_2 の次元数を 16 から 128 まで調整しながら，単語認識実験を行った．その結果を図 7 に示す．右には比較対象 3 つの特徴量，左には次元数を調整した PV_2 の認識結果を示している． PV_2 の次元数が 20 の時，認識率が 98.8% で一番高い．その時， $MFCC$ ， PV_1 よりも認識率が高く， PCA と同じ認識率を示している．

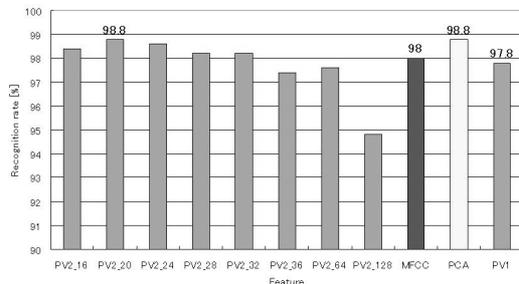


図 7 Results of word recognition in clean room.

最後に，残響時間が 470 ms と 1300 ms の時の単語認識実験を行った．その結果を図 8 と図 9 に示す．残響時間が 480 ms の時は 28 次元，1300 ms の時は 32 次元の PV_2 が一番高い認識率 (83.2%，75.8%) を示している．また，他の 3 つの特徴量に比べて，一番高い認識結果が得られた．

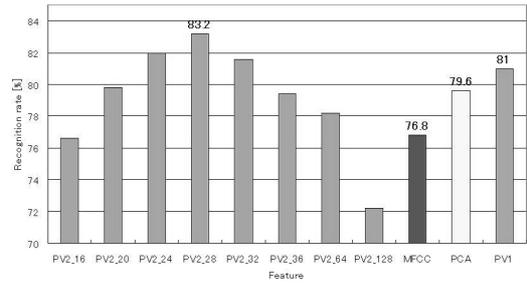


図 8 Results of word recognition in 470 ms reverberant room.

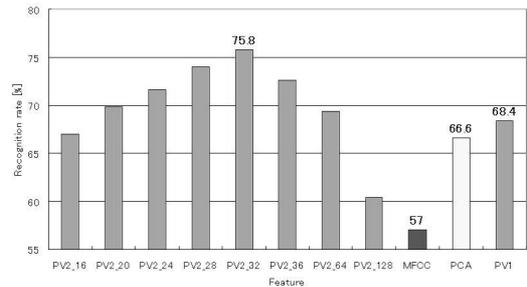


図 9 Results of word recognition in 1300 ms reverberant room.

音素 PCA による特徴量 PV_2 は残響時間が短い時は従来手法と同等な認識性能を示している，残響時間が長い時は従来手法より認識率の改善が得られた．これらの結果より，提案手法の有効性が確認できる．

5. まとめ

残響によるひずみ音声に対して，音声認識に有効な音声特徴量抽出法を提案した．提案手法を用いることで残響時間が長くなるほど，従来手法より顕著な認識率改善が実験の結果より確認できた．しかし，提案手法には課題がいくつか残されている．まず，提案手法の特徴量 PV_2 の最適次元数は，残響時間によって異なり，残響時間が長くなるほど次元数が大きくなる．次に，各音素部分空間の次元数の調整問題もある．本論文では音素ベクトルの作成において，各音素部分空間の次元数を同じ次元数に統一したため，音素ごとに最適な次元数は考慮していない．今後は寄与率などを用い，音素ごとに最適な次元数を調整していく予定である．また，本論文では 54 個の音素を用いているがこの音素数も最適とはいえない．音素によってはデータ数が少なく，音素空間の正しい推定が難しい．さらに次元数も膨大になっている．クラスタリング手法を用い，似ている音素同士を統合することにより次元数の削減と各音素部分空間の正しい推定も試みる予定である．最後に，提案手法のステップ 2 において， PCA を用いて次元削減を行っているが， LDA などのクラス間の識別に有効な手法も検討したい．

文献

- [1] C. Avendano, S. Tivrewala, and H. Hermansky, "Multiresolution channel normalization for ASR in reverberant environments," Proc. Eurospeech1997, pp. 1107-1110, 1997.
- [2] W. Li, K. itou, K. Takeda and F. Itakura, "Two-Stage Noise Spectra Estimation and Regression Based In-Car

- Speech Recognition Using Single Distant Microphone,” Proc. ICASSP2005, pp. 533-536, 2005.
- [3] K. Kinoshita, T. Nakatani and M. Miyoshi, “Efficient Blind Dereverberation Framework for Automatic Speech Recognition,” Proc. Interspeech2005, pp. 3145-3148, 2005.
- [4] R. Vetter, N. Virag, P. Renevey and J.-M. Vesin, “Single Channel Speech Enhancement Using Principal Component Analysis and MDL Subspace Selection,” Proc. Eurospeech99, pp. 2411-2414, 1999.
- [5] S.-M. Lee, S.-H. Fang, J.-W. Hung and L.-S. Lee, “Improved MFCC Feature Extraction by PCA-Optimized Filter Bank for Speech Recognition,” Proc. ASRU2001, pp. 49-52, 2001.
- [6] F. Asano, Y. Motomura, H. Asoh and T. Matsui, “Effect of PCA Filter in Blind Source Separation,” Proc. ICA2000, pp. 57-62, 2000.
- [7] H. Hermansky and N. Morgan, “RASTA Processing of Speech,” IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2, pp. 578-589, 1994.
- [8] B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” Neural Computation, Vol. 10, pp. 1299-1319, 1998.
- [9] Y. Ariki and K. Doi, “Speaker Recognition based on Subspace Method,” Proc. ICSLP94, pp. 1859-1862, 1994.
- [10] Y. Ariki, S. Tagashira and M. Nishijima, “Speaker Recognition and Speaker Normalization by Projection to Speaker Subspace,” Proc. ICASSP96, Vol. 1, pp. 319-322, 1996.
- [11] 滝口哲也, 有木康雄, “Kernel PCA を用いた残響下における口バースト特徴量抽出の検討,” 情報処理学会論文誌, Vol.47, No.6, pp. 1767-1773, 2006.
- [12] S. Nakamura, K. Hiyane, F. Asano, T. Nishimura and T. Yamada, “Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition,” Proc. LREC2000, Vol. 2, pp. 965-968, 2000.
- [13] S. Young et. al., “The HTK Book,” Entropic Labs and Cambridge University, 1995-2002.