多項式カーネルを利用した歌声と朗読音声の識別特徴の分析 *

◎大石康智 (名大・情報科学), 後藤真孝 (産総研), 伊藤克亘 (法政大・情報科学), 武田一哉 (名大・情報科学)

1 はじめに

音声認識システムが扱うことのできる発話の対象は，人間が日常のコミュニケーションに用いる多様な発話様式の中の，ごく一部（読み上げや講演など）に限定されている [1]. 対象を拡大するために，多様な環境や発話様式，個人性に対応するための技術が必要である．本研究の目的は，感情音声や歌声のような発話様式の違いを特徴付ける物理的あるいは信号的性質を明らかにすることである．そこでまず，通常の話し声（読み上げや講演音声など）との違いを聞き分けやすい歌声を研究対象に取り上げる．大石ら [2] は，聴取実験から 2 秒の歌声・朗読音声に対して人間は 100%識別が可能であり，短時間のスペクトル特徴が識別に影響することを確認した．この知見を客観的に評価するために，MFCC12 次までの係数と動的特徴 ΔMFCC を特徴尺度とし，混合ガウス分布 (GMM) で学習したところ，2 秒の評価音声信号に対して 84.7%の識別が可能であった．しかし，歌声と朗読音声の短時間スペクトルにどのような物理的な性質の違いが現れるかは明らかにされていない．

本報告では，短時間スペクトルを特徴ベクトルとして，歌声と朗読音声の識別に利用する．この特徴ベクトルを多項式カーネルによって高次元の特徴空間に写像し，SVM によって識別超平面を推定する．推定された識別関数の重みベクトルから識別に影響を与える周波数成分を特定する．

2 カーネル法による歌声と朗読音声の識別

音声信号の短時間フーリエ変換 (STFT) の出力を特徴ベクトル $\mathbf{x}$ として利用する (Fig1 は，512 点で STFT したときの対数振幅スペクトルを表す 256 次元の特徴ベクトルである)．カーネル法とは，これらの入力空間を，高次の特徴空間 $F = \{\boldsymbol{\phi}(\mathbf{x}) : \mathbf{x} \in X\}$ に写像し，分類タスクを簡略化することによって，線形分離可能な識別面を推定することである．このとき，識別面は以下の線形関数とその双対形式で表される [3].

$$f(\mathbf{x}) = \sum_{n=1}^{N} w_n \phi_n(\mathbf{x}) + b = \sum_{i=1}^{l} \alpha_i y_i \langle \boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x}) \rangle + b \tag{1}$$

ここで，w_n は線形関数における重みであり，ベクトル表示したもの $\mathbf{w}$ を重みベクトルと呼ぶ．N は特徴空間の次元数，b はバイアス，α_i は以下の最適化問題を解くときのラグランジュ乗数，$\mathbf{x}_i, y_i \in \{-1, 1\}$ はそれぞれ学習ベクトルとそのラベル，l は学習ベクトルのサンプル数，$\mathbf{x}$ は評価ベクトルを示す．したがって，決定規則は学習ベクトルと評価ベクトルの内積を用いて評価できる．この内積 $\langle \boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x}) \rangle$ を直接計算する方法が，カーネル関数と呼ばれ，本手法では，以下の d 次の多項式カーネルを用いる．

$$K(\mathbf{x}, \mathbf{z}) = \langle \boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\phi}(\mathbf{z}) \rangle = (\langle \mathbf{x} \cdot \mathbf{z} \rangle + 1)^d \tag{2}$$

最終的に $\mathbf{w}$ と b を推定するためにソフトマージン SVM[3] を利用する．すなわち以下の最適化問題を一般化ラグランジアンを利用して解くことに帰着する．ここで SVM の制約条件を緩めるスラック変数 ξ_i とパラメータ C を導入することにより，学習ベクトルがマージン制約を違反できるようにする．

$$\begin{aligned} &\text{minimize}_{\boldsymbol{\xi}, \boldsymbol{w}, b} \quad \langle \mathbf{w} \cdot \mathbf{w} \rangle + C \sum_{i=1}^{l} \xi_i^2, \\ &\text{subject to} \quad y_i(\langle \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i) \rangle + b) \geq 1 - \xi_i, i = 1, \ldots, l \end{aligned} \tag{3}$$

*Differentiating Characteristics Analysis between Singing and Speaking Voices Using Polynomial Kernel by Y. Ohishi (Nagoya Univ.), M. Goto (AIST), K. Itou (Hosei Univ.), and K. Takeda (Nagoya Univ.)

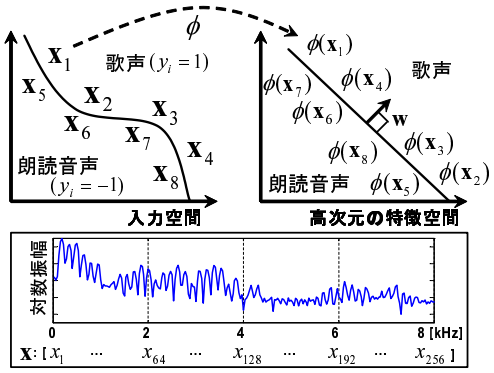

Fig. 1 特徴ベクトルの入力空間と高次元の特徴空間: 音声信号を STFT して得られる短時間スペクトルを特徴ベクトル $\mathbf{x}$ と考え，歌声 (ラベル $y_i = 1$) と朗読音声 (ラベル $y_i = -1$) の識別に利用する．$\boldsymbol{\phi}$ によって写像した高次元の特徴空間において，線形関数による識別面を推定する．

重みベクトル w の構成

重みベクトル $\mathbf{w}$ は，式 (1) の関係から以下のように計算される．ここで $\mathbf{w}$ は特徴空間での超平面の法線ベクトルである．

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \boldsymbol{\phi}(\mathbf{x}_i) \tag{4}$$

$\boldsymbol{\phi}(\mathbf{x})$ は，例えば $\mathbf{x}$ が 256 次元ベクトルで，式 (2) において $d = 2$ の 2 次の多項式カーネルを用いた場合，33153 次元のベクトルになり，以下のように表現される．

$$\begin{aligned} \boldsymbol{\phi}(\mathbf{x}) = (&x_1^2, \ldots, x_{256}^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_{255}x_{256}, \\ &\sqrt{2}x_1, \ldots, \sqrt{2}x_{256}, 1) \end{aligned} \tag{5}$$

ここで，$x_i x_j (i \neq j)$ の組み合わせは 32640 通り，x_i^2, x_i がそれぞれ 256 通り，定数項が 1 通りとなる．

3 評価実験

3.1 歌声データベース

産業技術総合研究所 (AIST) によって収録された歌声研究用音楽データベース「AIST ハミングデータベース」[4] の一部である，日本人歌唱者 75 名分（男性 37 名，女性 38 名）の音声データを使用した．特別な歌唱訓練を受けていない一般的な歌唱者が，*“RWC Music Database: Popular Music”* から抜粋した合計 25 曲の歌唱の出だしの部分と一番代表的な盛り上がる主題の部分の二カ所を、うろ覚えの状態で歌った歌声と当該部分の歌詞を読み上げた朗読音声を用いた．1 名あたり計 100 サンプル (歌声: 50 サンプル，朗読音声: 50 サンプル) となる．

3.2 識別器の学習と評価方法

話者，楽曲ともにすべてオープンデータで学習と評価を行うために，話者を 3 グループ，楽曲を 5 グループに分け，15 回のクロスバリデーションを行う．

識別器の学習では，まず音声信号を標本化周波数 16kHz，窓関数として窓幅 512 点のハミング窓を用いた STFT を高速フーリエ変換 (FFT) によって計算する．その際，FFT のフレームを 160 点ずつシフトするため，フレームシフト時間は 10ms となる．このフレームシフトをすべての処理の時間単位とする．計算された短時間スペクトルを 256 次元の特徴ベクトルとし，歌声と朗読音声の識別面を SVM により推定するための学習データとする．計算量を考慮して，歌声，朗読音声ともに学習に利用する特徴ベクトルを今回はランダムに 3000 個までに削減した．また，SVM を学習する際の数値的な問題を防ぐために，全学習データの平均と分散を利用して特徴ベクトルを正規化した．

Table 1　振幅，パワー，対数振幅スペクトルを特徴ベクトルとしたときの識別結果: 対数振幅による識別率が最も高い．

	d	C	歌声	朗読音声	総合
	1	1	68.5%	58.2%	63.4%
振幅	2	10^{-2}	65.7%	81.6%	73.7%
スペクトル	3	10^{-4}	64.9%	80.2%	72.6%
	1	10^{-1}	22.7%	98.8%	60.8%
パワー	2	10^{-2}	36.3%	95.4%	65.9%
スペクトル	3	10^{-4}	30.7%	98.2%	64.5%
	1	10^{-1}	78.7%	62.4%	**70.6%**
対数振幅	2	1	79.9%	72.0%	**76.0%**
スペクトル	3	1	80.5%	72.0%	**76.3%**

Table 2　対数振幅スペクトルの動的特徴 256 次元ベクトルを利用したときの識別結果: 170ms にわたる動的特徴を利用したとき最も高い識別率 88.7%が得られた．

Δ 時間幅	d	C	歌声	朗読音声	総合
50ms	1	1	70.3%	46.8%	58.5%
(前後 2 点の	2	1	**91.8%**	**78.0%**	**84.9%**
回帰係数)	3	1	0%	99.4%	49.7%
90ms	1	10^{-1}	68.3%	56.0%	62.1%
(前後 4 点の	2	1	**92.2%**	**84.8%**	**88.5%**
回帰係数)	3	1	7.0%	87.8%	47.4%
130ms	1	10^{-1}	57.0%	64.0%	60.5%
(前後 6 点の	2	1	**86.7%**	**84.8%**	**85.8%**
回帰係数)	3	10^{-4}	98.6%	61.8%	80.2%
170ms	1	10^{-1}	71.5%	52.6%	62.0%
(前後 8 点の	2	1	**90.8%**	**86.6%**	**88.7%**
回帰係数)	3	10^{-6}	99.4%	15.8%	57.6%
210ms	1	1	67.0%	57.8%	62.4%
(前後 10 点の	2	1	**86.9%**	**81.8%**	**84.4%**
回帰係数)	3	10^{-6}	99.6%	17.0%	58.3%

評価音声信号は，従来法と比較するために音声信号長を発声開始から 2 秒に固定する．学習と同様に周波数分析を行い，学習データの平均，分散を用いて正規化する．算出された特徴ベクトル系列と学習した SVM のパラメータを用いて，歌声・朗読音声の 2 クラス分類を行い，分類された特徴ベクトルの多いクラスを識別結果とする．

3.3　実験結果

Table1 は，振幅，パワー，対数振幅スペクトルを特徴ベクトルとしたときの識別結果である．$d = 1$ は，入力空間での SVM による識別である．$d = 2,3$ は 2 次，3 次の多項式カーネルを利用した SVM による識別である．各特徴空間において，SVM の制約条件を緩めるパラメータ C を 10^{-6}〜10 の間で変化させたときの最も高い識別率と，その C の値を示す．対数振幅を特徴ベクトルとしたときに，総合的に最も高い識別率が得られた．

Table2 は，対数振幅の各周波数成分の動的特徴を特徴ベクトルとした識別結果である．動的特徴は，ある時間幅にわたる回帰係数で表す．特に，170ms にわたる対数振幅の動的特徴で 2 次の多項式カーネルを利用したとき，識別率は最も高い 88.7%であった．対数振幅よりもその動的特徴を利用した方が識別性能が向上したことから，多数の被験者に基づく”歌声らしさ”，発声様式の違いは特に音声信号のダイナミクスに大きく現れると考えられる．

Table3 は，従来の MFCC による識別率と本手法の最も性能が高い条件下での識別率との比較を示す．本手法は，歌声と朗読音声の識別率間の差が小さく，従来法に比べて性能が改善された．これは，短時間スペクトルそのものを特徴ベクトルとしたことによって MFCC の 12 次までの係数では表現しきれない特徴を考慮できたとともに，多項式カーネルを適用することにより，周波数ビン間の相関関係を考慮した高次の特徴空間において識別を行ったためであると考えられる．さらにどの周波数ビンに基づく特徴空間の要素が識別面に影響を与えるかについて次節で検討する．

3.4　重みベクトル w の分析

識別の手がかりとなる周波数帯域を特定するために，識別面を決定付ける重みベクトル $\mathbf{w}$ を分析する．重みベクトルの各要素は，式 (4)，式 (5) から任意の周波数ビンの積で構成される．Fig2 は，2 次の多項式カーネルを利用したときの 33153 要素の重みに対して，近接する 500Hz の帯域内に含まれる重みの平均値を算出し，16 × 16 のカラーマップで可視化したものである．色の濃い部分ほど重みが大きい．1 次の項では帯域によって重みの大きさに差が生じた．対数振幅では 1.5kHz 以下の低域，動的特徴では，2，4，6kHz 付近の帯域の重みが大きい．2 次の項では，対数振幅は，2kHz 以下の低域，1kHz 以下の低域と 2kHz 以上の帯域との組合せ，2〜4kHz の帯域内で構成される項の重みが大きい．動的特徴では，さらに 5〜8kHz の帯域内で構成される項の重みも大きい．したがって，歌声と朗読音声の短時間スペクトル，またはその動的特徴の違いは，特定の帯域に大きく現れることが考えられる．また帯域間の相関関係にも違いがあることが，多項式カーネルによる非線形写像を行ったことから明らかとなった．

Table 3　従来法 (MFCC の分布を GMM で学習した識別器) と対数振幅スペクトルを利用した本手法の識別結果との比較

特徴ベクトル	歌声	朗読音声	総合
従来法 (MFCC)	66.2%	78.7%	72.5%
対数振幅スペクトル	80.5%	72.0%	**76.3%**
従来法 (ΔMFCC)	72.0%	96.0%	84.0%
対数振幅スペクトルの動的特徴	90.8%	86.6%	**88.7%**
従来法 (MFCC+ΔMFCC)	76.1%	92.9%	84.5%
対数振幅スペクトル + 動的特徴	94.2%	80.8%	**87.5%**

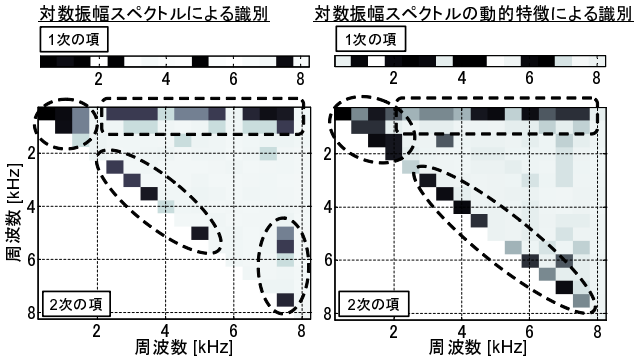

Fig. 2　識別の手がかりとなる周波数帯域の分析結果: 2 次の多項式カーネルを利用したとき，33153 要素の重みが式 (4) から計算される．色が濃い部分ほど，重みが大きいことを示す．

4　まとめ

歌声と朗読音声の短時間スペクトルを特徴ベクトルに利用し，多項式カーネルによって写像した高次元特徴空間において SVM による自動識別実験を行った．推定された SVM の重みベクトルから，歌声と朗読音声の短時間スペクトルの違いは，特定の帯域に大きく現れていることが推測される．また多項式カーネルによる非線形写像によって，帯域間の相関関係を考慮することができ，従来法よりも識別性能が改善されたと考えられる．特定の帯域の重みが大きいことが，歌声，朗読音声の発声におけるどのような物理的な性質と対応するかについてはさらなる考察が必要である．

参考文献

[1] S. Furui, M. Nakamura, T. Ichiba and K. Iwano, ”Analysis and recognition of spontaneous speech using Corpus of Spontaneous Japanese”, Speech Communication, vol.47, pp.208-219, 2005.

[2] 大石 康智, 後藤 真孝, 伊藤 克亘, 武田 一哉, “スペクトル包絡と基本周波数の時間変化を利用した歌声と朗読音声の識別,” 情報処理学会論文誌, Vol.47, No.6, pp.1822-1830, 2006.

[3] B. Scholkopf and A. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond , MIT Press, 2002.

[4] 後藤 真孝, 西村 拓一, “AIST ハミングデータベース: 歌声研究用音楽データベース,” 情処研報音楽情報科学, Vol.2005, No.82, pp.7-12, 2005.