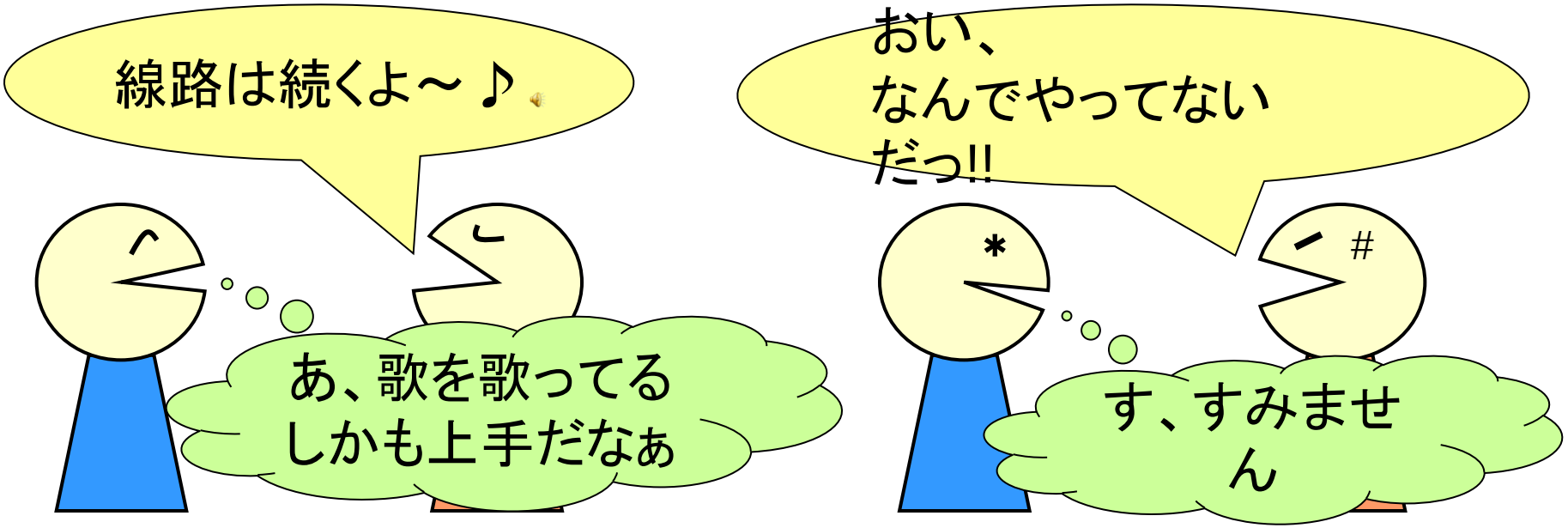
多項式カーネルを利用した歌声と朗読音声の識別特徴の分析

大石 康智[1], 後藤 真孝[2]
伊藤 克亘[3], 武田 一哉[1]

[1]名古屋大学大学院情報科学研究科
[2]産業技術総合研究所
[3]法政大学情報科学部

はじめに

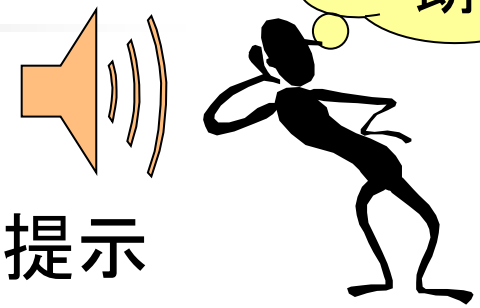
- 人間は多様な発話様式を使い分ける



- いったい音響的特徴に何が違うのか?
- 発話様式の違いを識別できる技術
  - 様々なインタフェースに応用が可能

通常の話し声との違いを聞き分けやすい歌声に着目

先行研究



- 聴取識別実験
  - 異なる長さの音声信号の提示

音声信号1s:　100%識別可

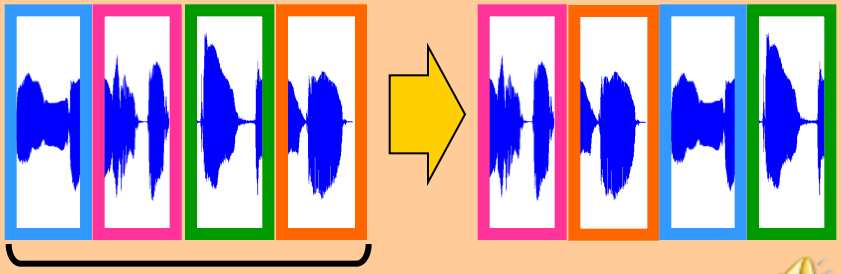
音声信号の時間構造の違い?

音声信号200ms:　70%識別可

短時間スペクトル特徴の違い?

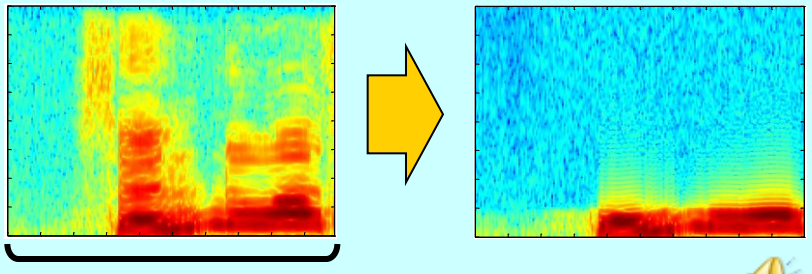
  - 信号処理により特徴変形させた音声信号の提示

時間構造の変形



→ 歌声の識別率29.4%低下

短時間スペクトル特徴の変形



→ 歌声の識別率13.1%低下

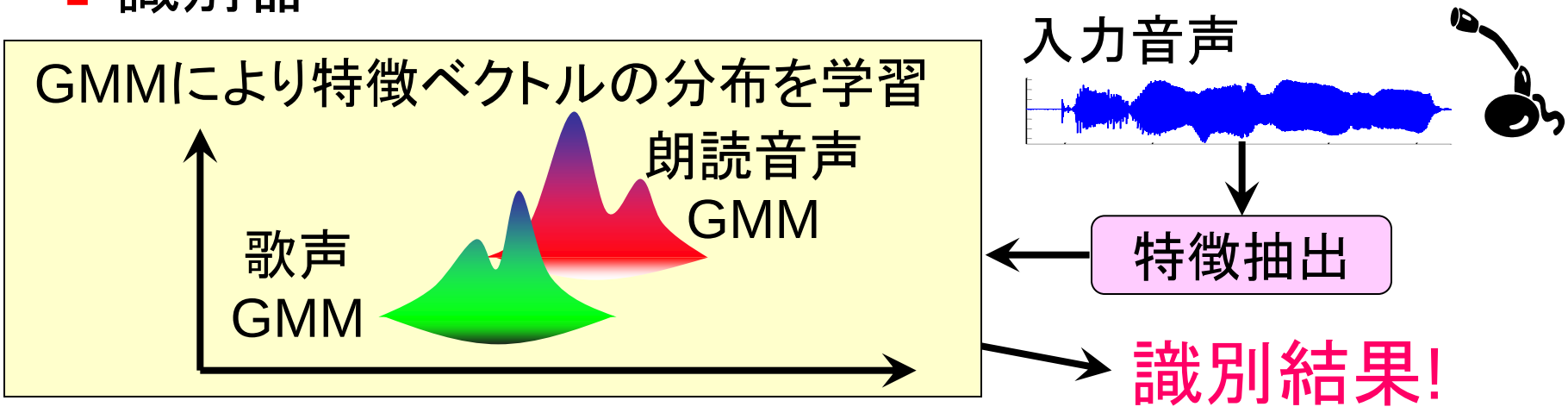
先行研究

- 自動識別実験

時間構造の特徴抽出
・F0の時間変化 ΔF0

短時間スペクトルの特徴抽出
・MFCC, ΔMFCC

- 識別器



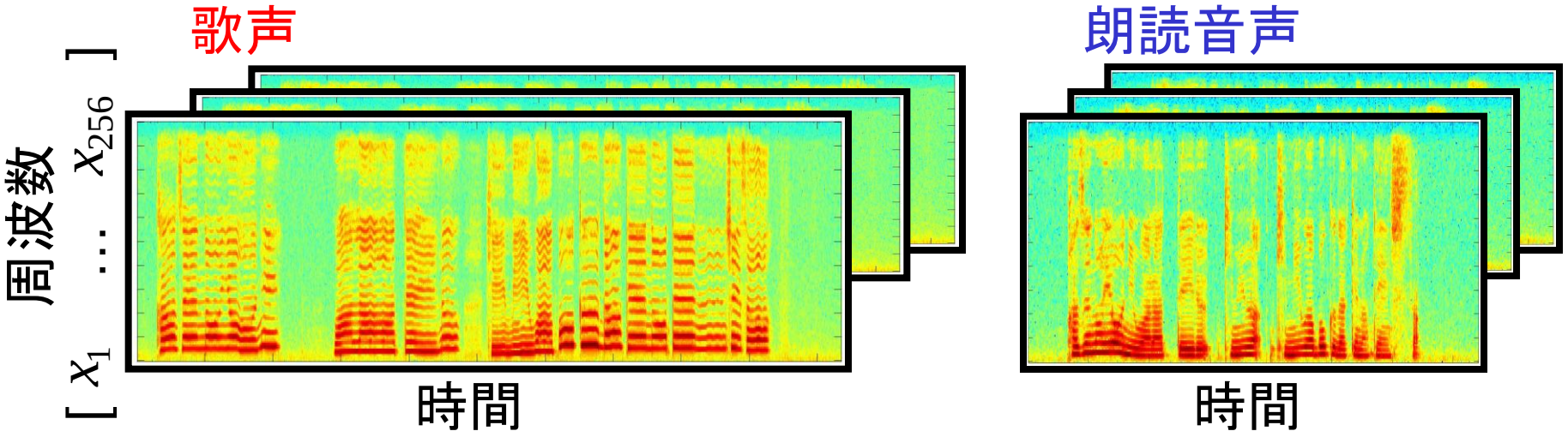
入力音声信号2s:　識別率87.3%

短時間スペクトル，F0の時間変化に
どのような違いがあるのかは明らかでない

本研究の目的

- 歌声と朗読音声の短時間スペクトルの違い
  - *Singing formant* ⇒ オペラ歌手に観測される
  - 素人の歌声と話し声の短時間スペクトルの違いは?
    - どの帯域が顕著に変化し，識別に影響するのか?

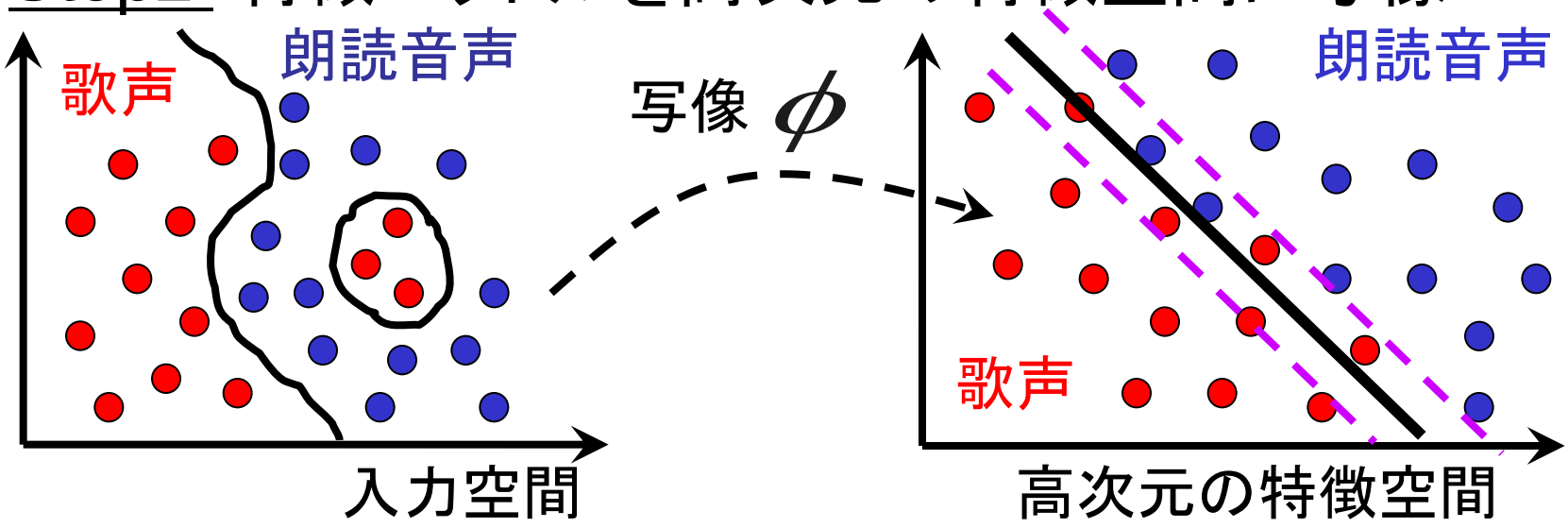
Step1. 短時間スペクトルを特徴ベクトルとして利用



512点(標本化周波数16kHzの場合32ms)のFFT→256次元ベクトル

特徴ベクトルの入力空間と高次元の特徴空間

Step2. 特徴ベクトルを高次元の特徴空間に写像



Step3. ソフトマージンSVMによって識別面の推定

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \phi(\mathbf{x}) \rangle + b$$

重みベクトル $\mathbf{w}$ と
バイアス b の推定

多項式カーネルの利用

$$\langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle = \left(\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle + 1 \right)^d$$

識別関数の重みベクトルに着目

Step4. 識別に貢献する周波数成分の特定

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \phi(\mathbf{x}) \rangle + b = \sum_{n=1}^{N} w_n \phi_n(\mathbf{x}) + b$$

0に近い　識別に影響を与えない

識別関数の重みベクトルに着目

Step4. 識別に貢献する周波数成分の特定

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \phi(\mathbf{x}) \rangle + b = \sum_{n=1}^{N} w_n \phi_n(\mathbf{x}) + b$$

w_n の大きさにより次元の貢献度がわかる

重みベクトル $\mathbf{w}$ の計算

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

α_i : ラグランジュの未定乗数

$\mathbf{x}_i$: 学習データ

y_i : 学習データのクラスラベル

識別関数の重みベクトルに着目

Step4. 識別に貢献する周波数成分の特定

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \phi(\mathbf{x}) \rangle + b = \sum_{n=1}^{N} w_n \phi_n(\mathbf{x}) + b$$

w_n の大きさにより次元の貢献度がわかる

重みベクトル $\mathbf{w}$ の計算

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

α_i : ラグランジュの未定乗数

$\mathbf{x}_i$: 学習データ

y_i : 学習データのクラスラベル

$\phi(\mathbf{x}_i)$ は? $(d = 2$の場合$)$

$$\langle \phi(\mathbf{x}) \cdot \phi(\mathbf{z}) \rangle = (\langle \mathbf{x} \cdot \mathbf{z} \rangle + 1)^2 \quad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_{256} \end{pmatrix} \quad \mathbf{z} = \begin{pmatrix} z_1 \\ \vdots \\ z_{256} \end{pmatrix}$$

$$= (x_1 z_1 + x_2 z_2 + \ldots + 1)^2$$

$$= (x_1^2 z_1^2 + \ldots + x_{256}^2 z_{256}^2 + 2x_1 z_1 x_2 z_2 + \ldots$$

$$+ 2x_{255} z_{255} x_{256} z_{256} + 2x_1 z_1 + \ldots + 2x_{256} z_{256} + 1)$$

識別関数の重みベクトルに着目

Step4. 識別に貢献する周波数成分の特定

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \phi(\mathbf{x}) \rangle + b = \sum_{n=1}^{N} w_n \phi_n(\mathbf{x}) + b$$

w_n の大きさにより次元の貢献度がわかる

重みベクトル $\mathbf{w}$ の計算

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

α_i : ラグランジュの未定乗数

$\mathbf{x}_i$: 学習データ

y_i : 学習データのクラスラベル

$\phi(\mathbf{x}_i)$ /

$\langle \phi(\mathbf{x}) \cdot$

$= (x_1 z_1$

$= (x_1^2 z_$

$+ 2x_{255}$

$\phi(\mathbf{x})$ の特徴空間

2次の項	$x_1^2, \ldots, x_{256}^2$
2次の項	$\sqrt{2}x_1x_2, \ldots, \sqrt{2}x_{255}x_{256}$
1次の項	$\sqrt{2}x_1, \ldots, \sqrt{2}x_{256}$
定数項	1

w_n が大きい ⇒ $\phi_n(\mathbf{x})$ を構成する特徴ベクトル $(x_1, \ldots, x_{256})$ の要素が識別に貢献する

評価実験

- AISTハミングデータベース
  - 特別な歌唱訓練を受けていない日本人歌唱者
    (男性 37名，女性 38名)
  - “RWC音楽データベース:ポピュラー音楽”から抜粋した25曲
  - 歌唱の出だし部分と一番盛り上がる主題の部分の二箇所を，歌唱した歌声と歌詞を読み上げた朗読音声
- 特徴抽出

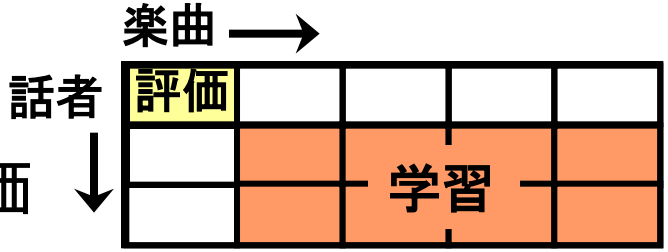
音声信号 → 短時間フーリエ変換 → 特徴ベクトルの正規化

標本化周波数16kHz

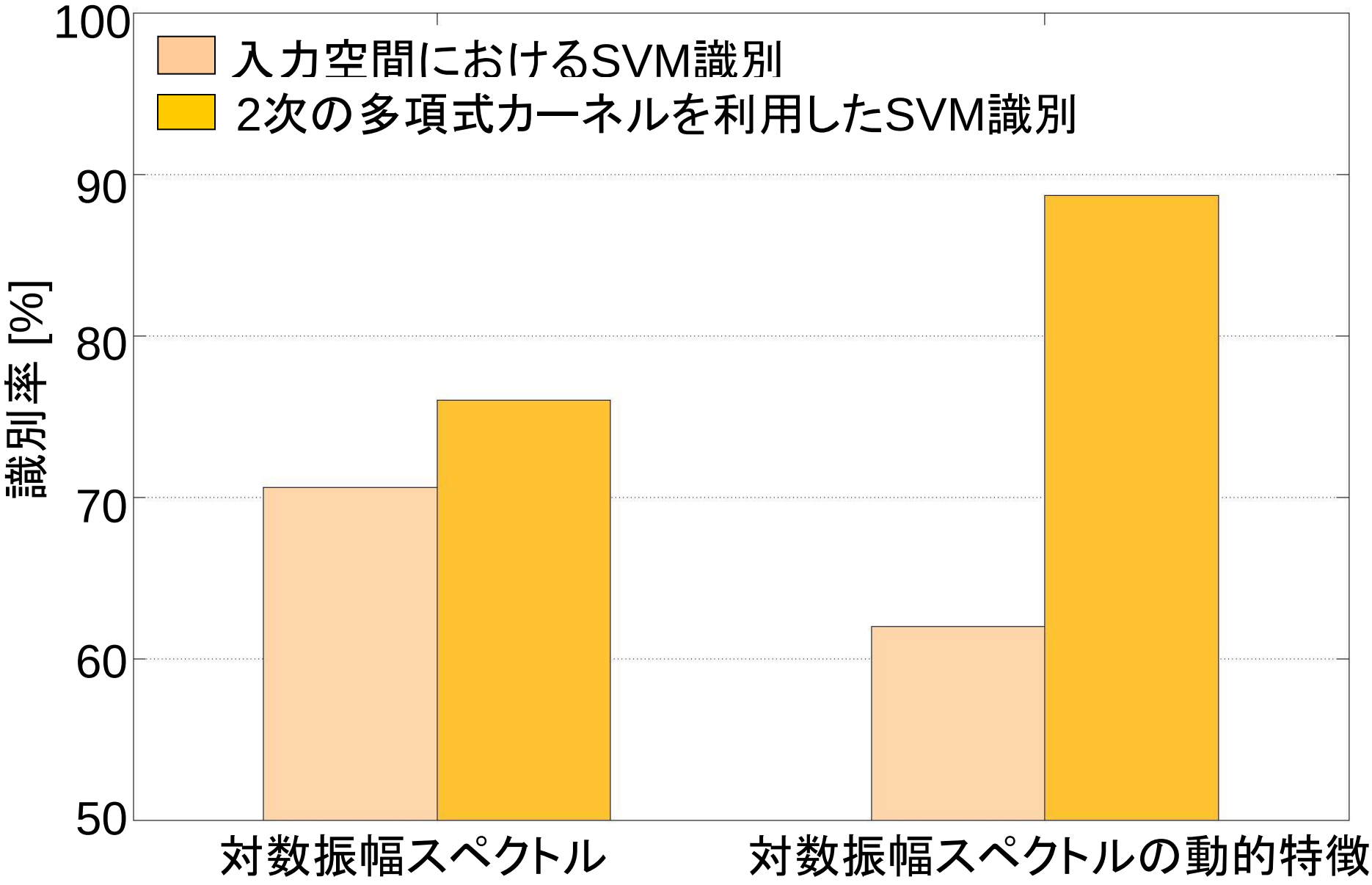
ハミング窓: 窓幅512点(32ms)
フレームシフト: 160点(10ms)

全学習データの平均と分散を利用

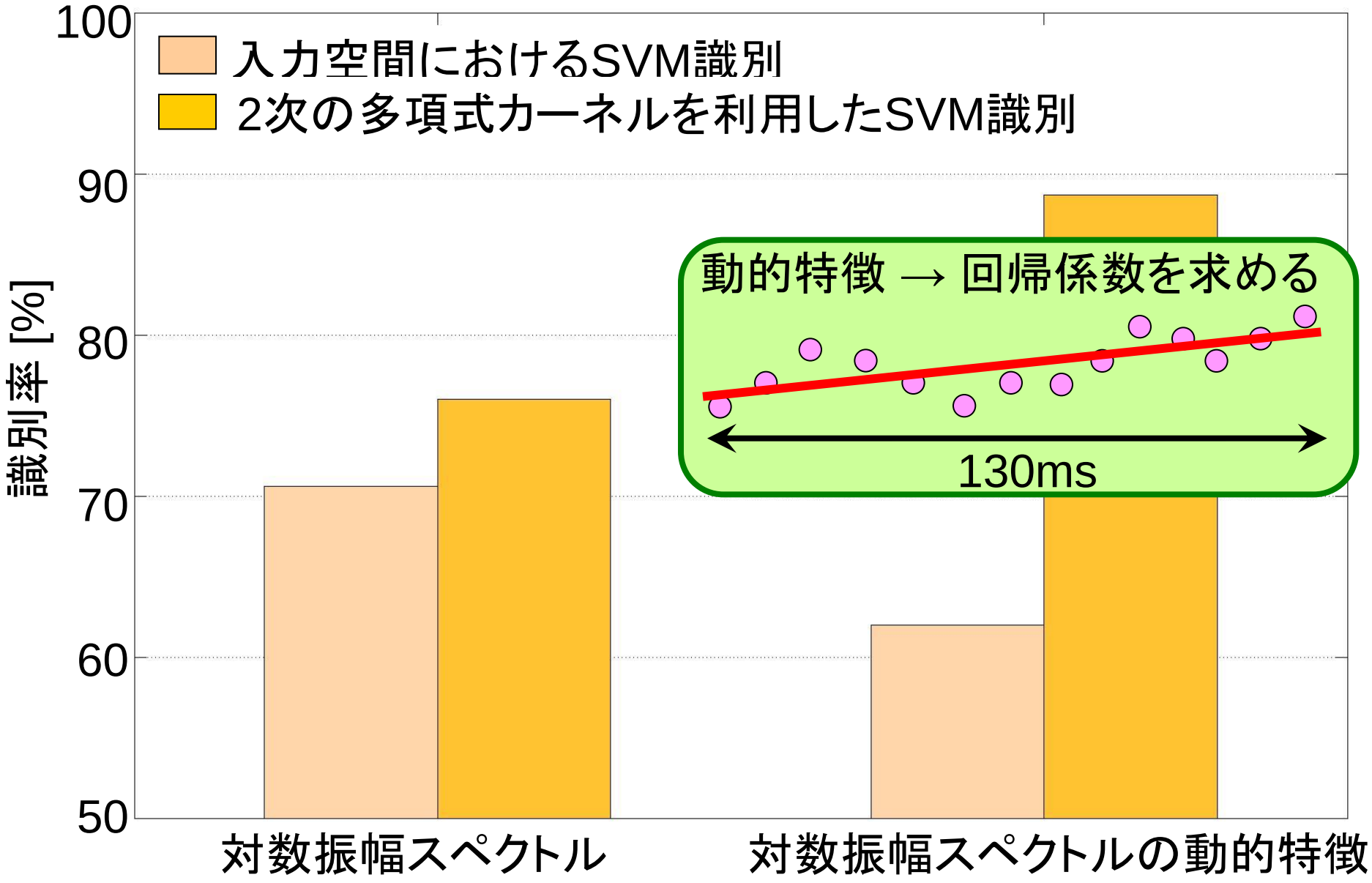
- 評価
  - 2秒の音声信号の識別率
  - 15回のクロスバリデーションによる評価



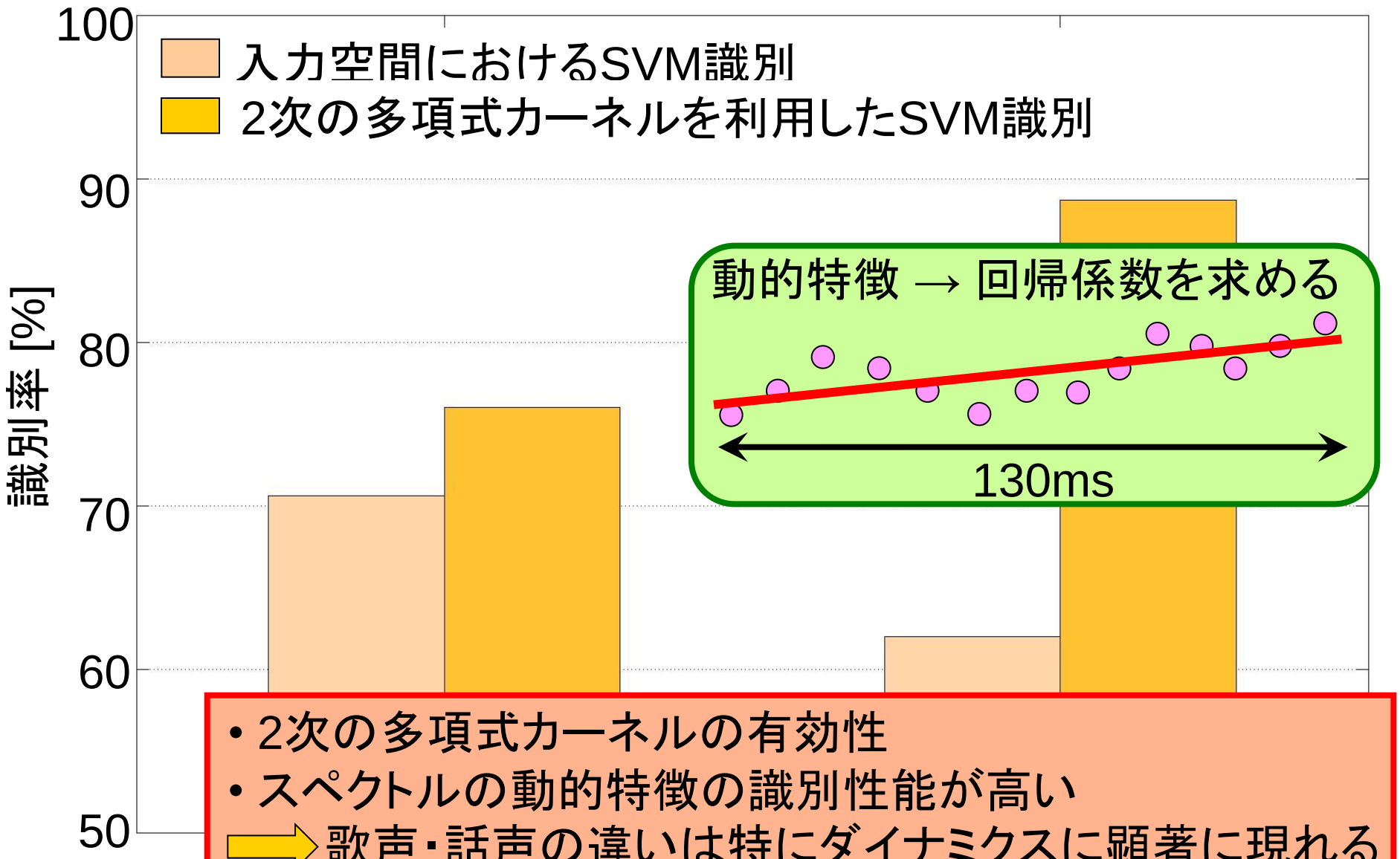
識別性能


100
90
80
70
60
50
識別率 [%]
入力空間におけるSVM識別
2次の多項式カーネルを利用したSVM識別
対数振幅スペクトル
対数振幅スペクトルの動的特徴

識別性能


100
90
80
70
60
50
識別率 [%]
入力空間におけるSVM識別
2次の多項式カーネルを利用したSVM識別
動的特徴 → 回帰係数を求める
130ms
対数振幅スペクトル
対数振幅スペクトルの動的特徴

識別性能



- 2次の多項式カーネルの有効性
- スペクトルの動的特徴の識別性能が高い

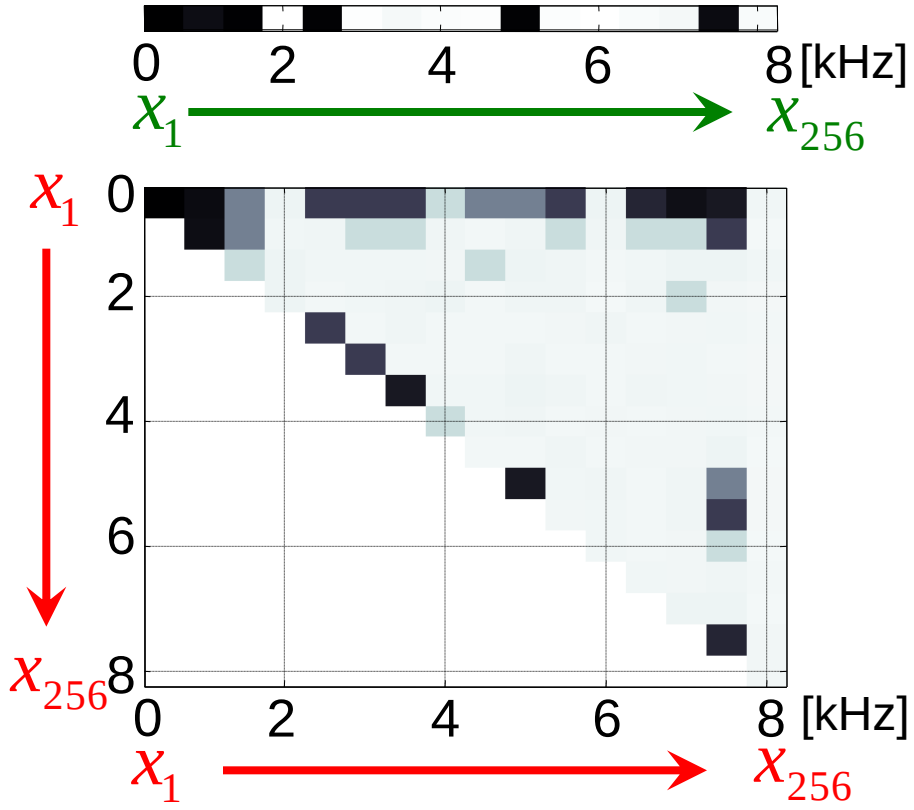
⇒ 歌声・話声の違いは特にダイナミクスに顕著に現れる

識別に貢献する周波数成分

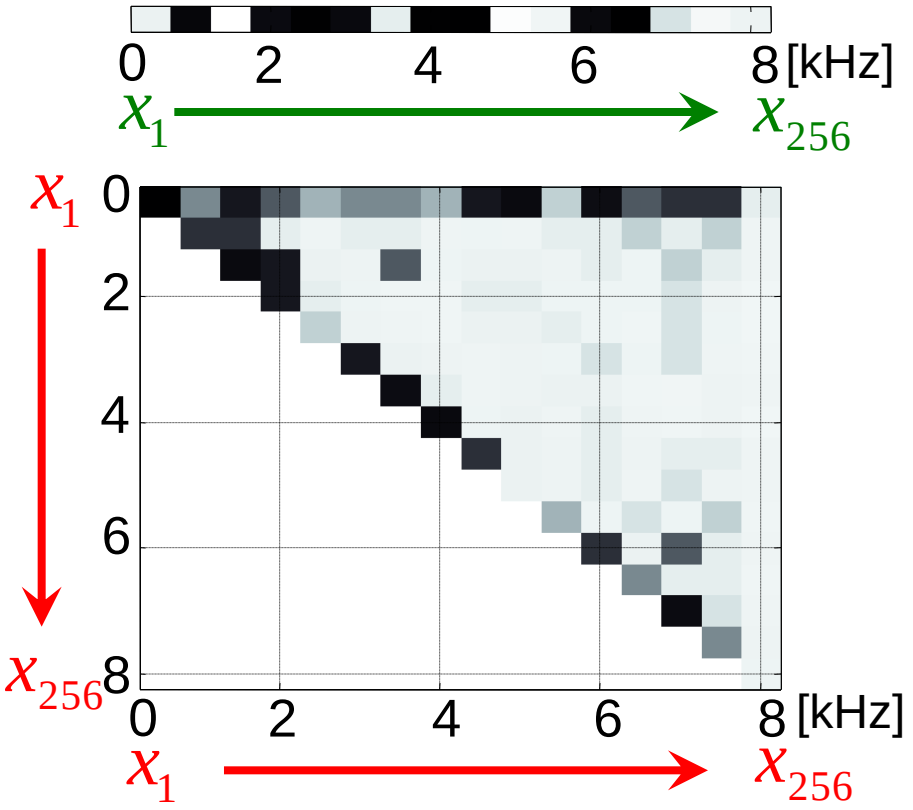
2次の多項式カーネルの利用

$\mathbf{w}$	$w_1, \ldots, w_{256}$	$w_{257}, \ldots, w_{512}$	$w_{513}, \ldots, w_{33152}$
$\phi(\mathbf{x})$	$\sqrt{2}x_1, \ldots, \sqrt{2}x_{256}$	$x_1^2, \ldots, x_{256}^2$	$\sqrt{2}x_1x_2, \ldots, \sqrt{2}x_{255}x_{256}$

対数振幅スペクトルによる識別



スペクトルの動的特徴による識別



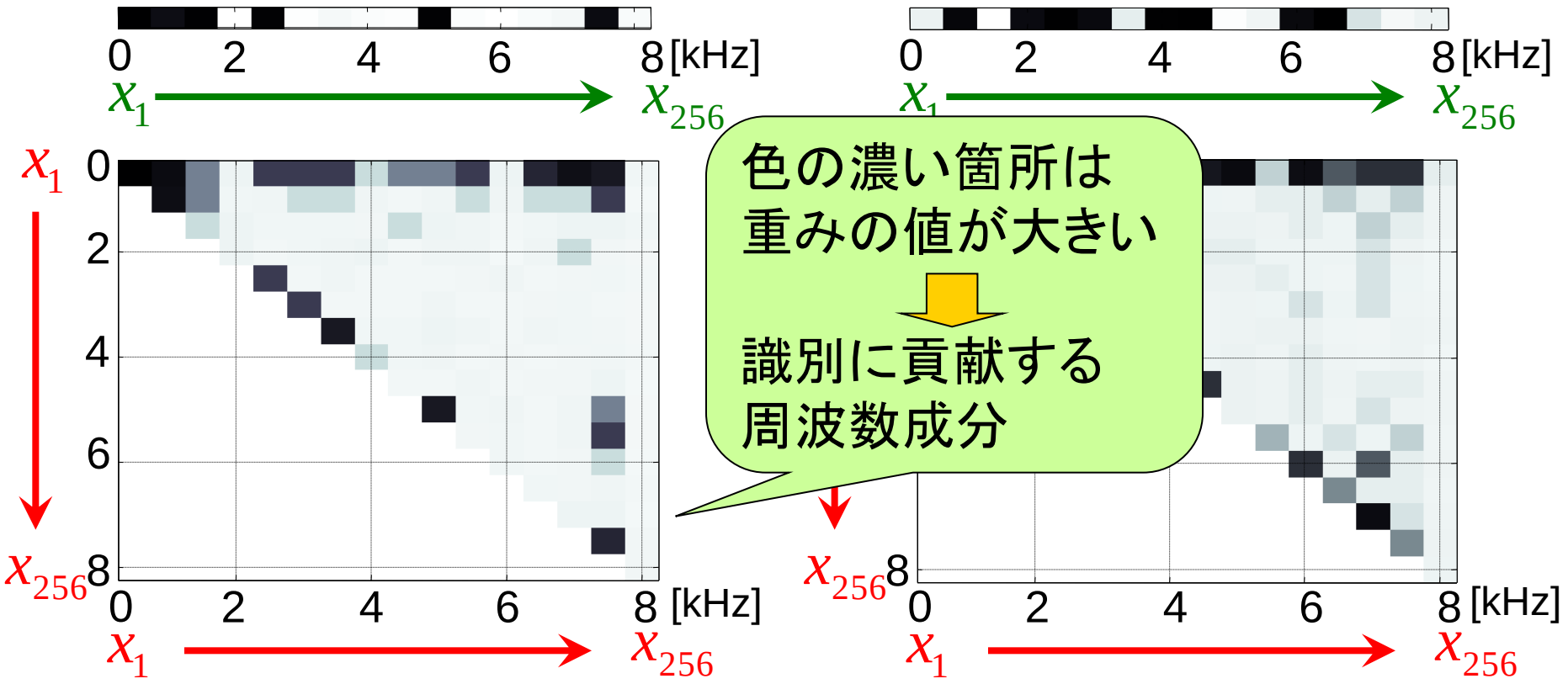
識別に貢献する周波数成分

2次の多項式カーネルの利用

$\mathbf{w}$	$w_1, \ldots, w_{256}$	$w_{257}, \ldots, w_{512}$	$w_{513}, \ldots, w_{33152}$
$\phi(\mathbf{x})$	$\sqrt{2}x_1, \ldots, \sqrt{2}x_{256}$	$x_1^2, \ldots, x_{256}^2$	$\sqrt{2}x_1x_2, \ldots, \sqrt{2}x_{255}x_{256}$

対数振幅スペクトルによる識別

スペクトルの動的特徴による識別



識別に貢献する周波数成分

2次の多項式カーネルの利用

$\mathbf{w}$	$w_1, \ldots, w_{256}$	$w_{257}, \ldots, w_{512}$	$w_{513}, \ldots, w_{33152}$
$\phi(\mathbf{x})$	$\sqrt{2}x_1, \ldots, \sqrt{2}x_{256}$	$x_1^2, \ldots, x_{256}^2$	$\sqrt{2}x_1x_2, \ldots, \sqrt{2}x_{255}x_{256}$

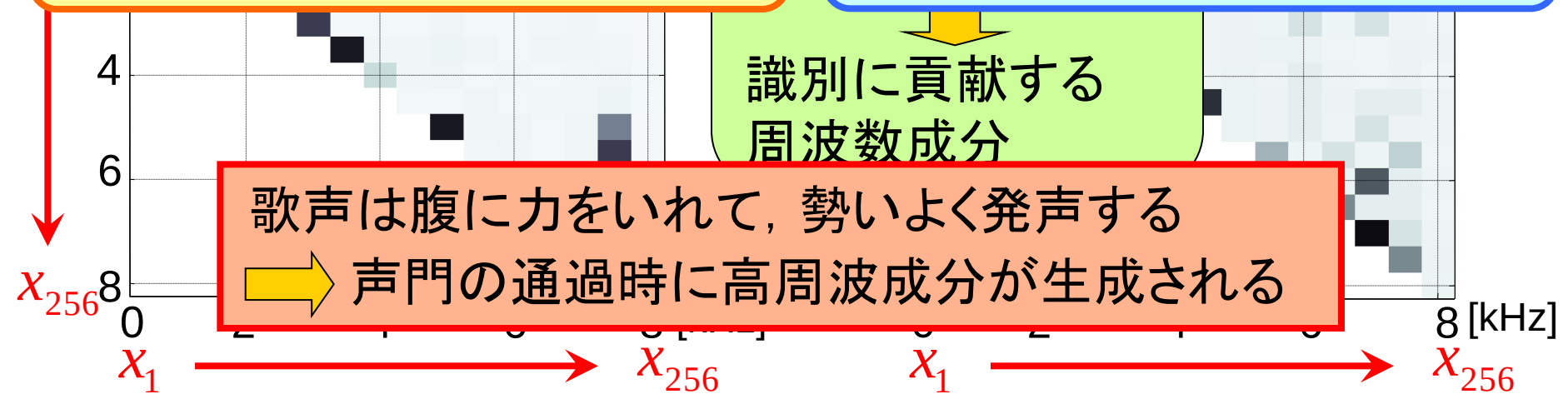
・1kHz以内の低域との組合せ項
・2～4kHzの対角に位置する項
で構成される重みの値が大きい

・1kHz以内の低域との組合せ項
・対角に位置する項
で構成される重みの値が大きい

2kHz以内の低域で構成される重みW_nの値が大きい

2, 4, 6kHz付近の成分で構成される重みW_nの値が大きい

識別に貢献する
周波数成分

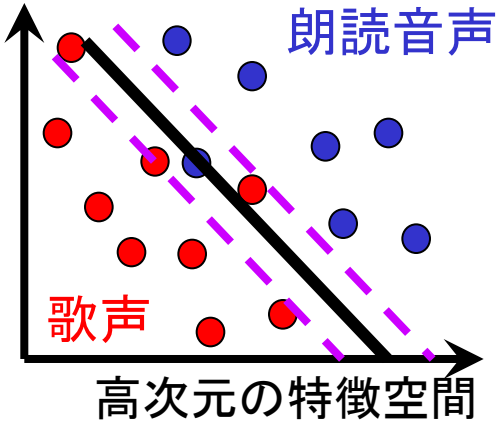


歌声は腹に力をいれて，勢いよく発声する
⇒ 声門の通過時に高周波成分が生成される

まとめと今後の展開

- 歌声と朗読音声の短時間スペクトルの違い

Step1. 短時間スペクトルそのものを特徴ベクトルに利用
Step2. カーネル法により高次元特徴空間に写像
Step3. ソフトマージンSVMによる識別
Step4. 識別関数の重みベクトルから
識別に貢献する周波数成分の特定



- スペクトルの動的特徴の識別に対する有効性
- 基本周波数が変化する低域とともに，高域の周波数への着目
  - 2, 4, 6kHz付近の周波数帯域の挙動

- 歌声・朗読音声の物理的な発声のしくみとの対応に関する考察