MKL による構音障害者の音声特徴量評価*

☆高塚智敬, 滝口哲也, 有木康雄 (神戸大), 李義昭 (追手門大)

1 はじめに

近年，音声認識技術は飛躍的に進歩してきた．様々な環境や場面での活用が期待され，カーナビゲーションの操作や会議音声の議事録化など，様々な分野に応用されている．最近では，高齢者や子どもなどの成人とは発話スタイルの異なる人も対象とした音声認識も実現され，多くの人にとって利用する機会が増えている．しかし一方で，音声認識をより必要としている言語障害者などを対象とするものは非常に少ない．文献 [1][2] では，言語障害者を対象とした特徴量抽出や音響モデルの適応，構築を行っているが，そうした研究は盛んに行われていないのが現状である．

言語障害の原因の一つとして，脳性麻痺による発声時の不随意運動が知られている．脳性麻痺患者は意図的な動作を行う場合や緊張状態にある時に筋肉の制御が難しくなり，不随意運動を伴うことがある．この運動障害によって構音が困難となり，言語障害が生じる [3]．本研究における音声認識は，脳性麻痺により構音機能に障害を持つ障害者（構音障害者）の音声を対象としている．

構音障害者の発話スタイルは健常者と大きく異なる上，同じ構音障害者であっても症状によって個人差があり，音声認識は非常に困難である．そこで我々は，認識精度改善のために，画像分類における特徴統合の有効性を示した Nilsback らの研究 [4] に着目した．

本稿では，構音障害者の音素認識を SVM (Support Vector Machine) による音声特徴量の識別によって行う．その際，音声特徴量である MFCC (Mel-Frequency Cepstral Coefficients) の各次元毎にカーネル関数を設定し，MKL (Multiple Kernel Learnig) による重み付け統合を行うことで，音声認識に対する各特徴次元の有効性を評価する．

2 特徴量の評価方法

2.1 提案手法の概要

通常の SVM は単一のカーネル関数を用いて特徴量を高次元空間へ射影することで識別関数を求めるが，MFCC の次元中には識別に有効な次元とそうでない次元が含まれている可能性が考えられる．そこで本手法では，SVM による音素境界の学習を行う際に，MKL によって分類に最適な各特徴次元の重みを求めることで，認識精度が向上することを確認する．また，このとき得られる各次元に対する重みを，分類に対する有効性の尺度とする．提案手法の概要を Fig.1 に示す．

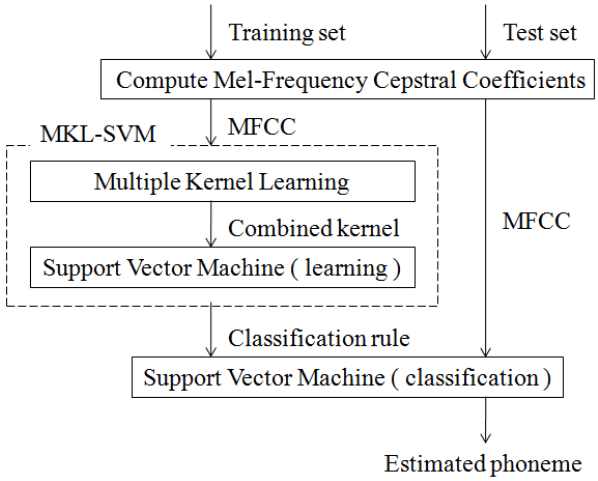

Fig. 1 Our proposal method

2.2 MKL-SVM による識別と特徴量の重み付け

MKL は複数のサブカーネルを線形結合したカーネルを作成することで，より複雑な非線形空間を作成する手法である．つまり，結合カーネル K は以下のように表現される．

$$K(x,x') = \sum_{i=1}^{N} \beta_i k_i(x,x') \ with \ \ \beta_i \geq 0, \ \ \sum_{i=1}^{N} \beta_i = 1$$

β_i は i 番目のサブカーネル関数 k_i の重みである．

MKL による重みの学習は，SVM の枠組みで解く方法が一般的で，MKL-SVM と呼ばれることもある．SVM の枠組みにおける MKL を最適化するための双対問題は，以下のような Min-Max 問題となる．

$$\begin{aligned} S_i(\alpha) &= \frac{1}{2}\sum_{j,k=1}^{M} \alpha_j \alpha_k y_j y_k k_i(x,x') - \sum_{j=1}^{M} \alpha_j \\ \max_{\beta} \min_{\alpha} & \quad \sum_{i=1}^{N} \beta_i S_i(\alpha) \\ w.r.t & \quad \alpha \in \mathbb{R}^{\mathbb{M}}, \ \beta \in \mathbb{R}^{\mathbb{N}} \\ s.t. & \quad \begin{cases} 0 \leq \alpha_j \leq C, \ 0 \leq \beta_i \\ \sum_{j=1}^{M} \alpha_j y_j = 0, \ \sum_{i=1}^{N} \beta_i = 1 \end{cases} \end{aligned}$$

*Estimation of optimal dysarthric speech features using Multiple Kernel Learning
By Norihiro Takatsuka, Tetsuya Takiguchi, Yasuo Ariki (Kobe Univ.), Ichao Li (Otemon Univ.)

ただし，α_j はラグランジュ係数，y_j はクラスを表す変数，C は SVM のスラック変数である．

本研究では， MKL-SVM を用いて音素識別を行う．MFCC の各次元毎にサブカーネルを定義することで，学習データの分類に最適な各次元の重みを学習させる．

3 音素認識実験による評価

提案手法を評価するために特定話者での切り出し音素の識別実験を行った．また，MKL による重み付けの識別に対する有効性を評価するために，MKL-SVM を用いた実験のほかに，通常の単一カーネル SVM を用いた実験も行った．

3.1 実験条件

評価実験の認識対象は 5 つの母音とし，評価音素の各フレームに対して 5 クラス分類を行った．実験条件を Table.1 に示す．また，SVM による 5 クラスの分類は one-versus-one 法を用いて行った．パラメータは，提案手法における認識に最適な値を実験的に決定し，比較実験においても同様の値を用いた．カーネルには，Varma らの研究 [5] において最も性能が良かったとされるガウシアンカーネルを採用した．

3.2 MKL-SVM による音素認識

認識率は 32.1%であった．また，MKL によって求めた各特徴次元の重みを Fig.2 に示す．横軸の 1 から 13 までが MFCC の 13 次元で，14 から 26 が Δ MFCC の 13 次元である．これよりまず，Δ MFCC よりも MFCC の方が圧倒的に認識に寄与していることが分かる．次に，13 次元目と 26 次元目に注目すると，両次元ともにその重みが 0 となっている．これらの次元はそれぞれ MFCC のエネルギーとその Δ の特徴量であり，この結果は音声のエネルギーに関する特徴は認識に一切寄与しなかったことを示している．

Table 1 Experiment condition

認識対象音素	5 母音: /a/,/i/,/u/,/e/,/o/
評価データベース	構音障害者（男性）1 名による 210 単語発話
学習資料	単語からの切り出し音素（各音素あたり 200 個）
評価資料	単語からの切り出し音素（各音素当たり 20 個）
特徴量	13 次元 MFCC + 13 次元Δ MFCC
カーネル	RBF(Gaussian)

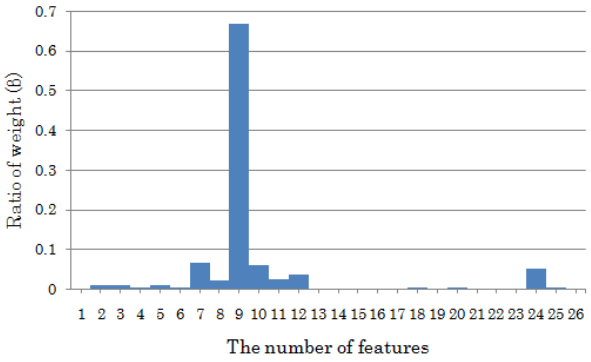

Fig. 2 The weight of feature dimensions

3.3 Single Kernel SVM による音素認識

同じ実験条件の下で単一カーネル SVM による音素認識実験を行い，提案手法との比較を行った．認識率は 25.8%で，提案手法が認識に有効な特徴次元を適切に選定できていることを示せた．

4 おわりに

MKL-SVM によって構音障害者の音声特徴量評価を行う手法を提案した．実験により，構音障害者の音声認識において，特徴次元毎に認識への貢献度が異なることが確認できた．今後はデータ数を増やした実験を行い，音声認識に対してより適当な MKL による特徴次元の重みを探索すると同時に，one-versus-rest 型の SVM を用いることで，各音素の識別に有効な特徴量の探索を行っていく．また，重み付けされた特徴量による連続音声認識の可能性について検討する．

参考文献

[1] 中村 他, “発話障害者音声を対象にした健常者音響モデルの適応と検証,” 音講論 (秋), 109-110, 2005.

[2] H. Matsumasa *et al.*, “Integration of Metamodel and Acoustic Model for Speech Recognition,” Interspeech2008, 2234-2237, 2008.

[3] S. Terry Canale *et al.*, “キャンベル整形外科手術書第 4 巻,” エルゼビア・ジャパン, 2004.

[4] M. Nilsback, A. Zisserman, “Automated flower classification over a large number of classes,” Proc. of Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing, 2008.

[5] M. Varma, D. Ray, “Learning the discriminative power-invariance trade-off,” In Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, 2007.