PLSA による構音障害者の音素体系構築の検討*

◎高塚智敬, 滝口哲也, 有木康雄 (神戸大・工), 李義昭 (追手門大), 中林稔堯 (神戸大・発達)

1 はじめに

近年，音声認識技術はその発展に伴い，様々な環境下や場面での利用が期待されている．例えば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で既に応用されている．これらの多くは健常者を対象としており，文献 [1, 2] では，構音障害者音声を対象とした特徴量抽出や音響モデル適応，構築を行っているが，言語障害者などの障害者を対象としたものは非常に少ない．

言語障害の原因の一つとして，脳性麻痺が考えられる．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．この運動障害の一つとして，正しく構音できない場合がある [3]．本稿では，脳性麻痺により構音機能に障害を持つ被験者を対象に音声認識実験を行う．

これまで構音障害者の音声認識は健常者の音素体系を基に行われているが，両者の発声方法は異なり，音素体系が一致しない．そこで本研究では，PLSA (Probabilistic Latent Semantic Analysis) によって音素モデルを自動生成し，それによって音声認識を行う手法を検討する．

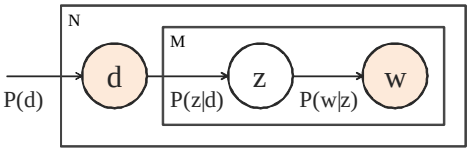

Fig. 1 PLSA graphical model

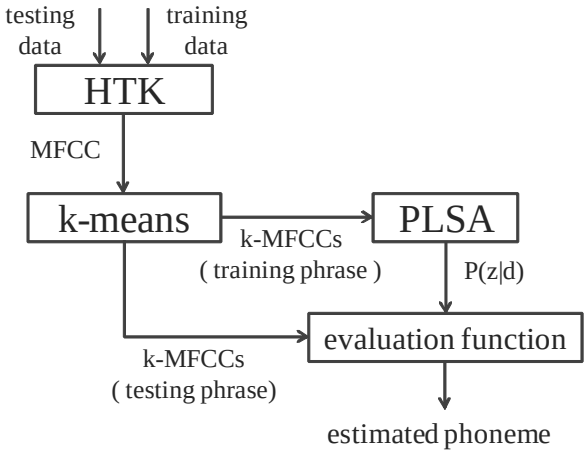

Fig. 2 Our proposal method

2 PLSA による音素推定

PLSA[4] はテキストモデリングの一手法であり，文書を潜在トピックの混合によって表現する．すなわち，文書 d における単語 w の発生確率を，潜在トピック z を用いて，

$$P(w|d) = \sum_{z\in Z} P(w|z)P(z|d) \quad (1)$$

のように表現する．PLSA のグラフィカルモデルを Fig.1 に示す．パラメータは EM アルゴリズムによって学習する．つまり，尤度の期待値を計算する E-step と 求めた期待値を最大化する各パラメータを推定する M-step を繰り返して最尤推定値を求める．

E-step

$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'\in Z} P(z')P(d|z')P(w|z')} \quad (2)$$

M-step

$$P(w|z) \propto \sum_{d\in D} n(d,w)p(z|d,w) \quad (3)$$

$$P(d|z) \propto \sum_{w\in W} n(d,w)p(z|d,w) \quad (4)$$

$$P(z) \propto \sum_{d\in D}\sum_{w\in W} n(d,w)p(z|d,w) \quad (5)$$

本研究におけるシステムの概略を Fig.2 に示す．まず，音声データから HTK[5] を用いて MFCC を抽出し，MFCC の頻度行列はスパースになりすぎるため，あらかじめ k-means を用いてクラスタリングしておく (Fig.3)．以下，このクラスタリングされた MFCC を k-MFCCs と呼ぶ．文書 d を発話，単語 w を k-MFCCs，潜在トピック z を音素として 各発話に含まれる k-MFCCs の頻度行列に対して PLSA を行い，各音素における k-MFCCs の発生確率 $P(w|z)$，発話毎の潜在トピック重み $P(z|d)$ を求めることにより音素を推定する．

*A study on recognition of dysarthric speech using phonemic system based on PLSA
By Norihiro Takatsuka, Tetsuya Takiguchi, Yasuo Ariki (Kobe Univ.), I Chao Li (Otemon Univ.) and Toshitaka Nakabayashi (Kobe Univ.)

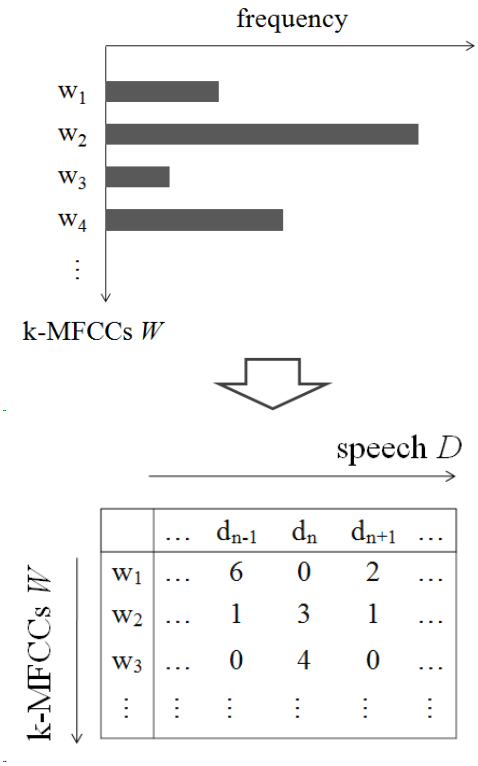

Fig. 3　Co-occurrence table based on k-MFCCs

3　母音認識実験

以上のようなモデルを用い，初期実験として，母音のみを PLSA を用いて学習し，母音のみを含む単語を認識した．学習データは各母音をそれぞれ 10 回発話した 50 個のデータを用意した．また，音声データのサンプリング周波数は 16 kHz，フレーム窓長は 25 msec，フレーム周期は 10 msec であり，特徴量は 12 次元 MFCC ＋Δ MFCC を用いた．各発話の潜在トピック重み $P(z|d)$ を Table 1 に示す．表より，各トピックがひとつの母音を表していることが分かる．テストデータには学習データと同様のデータを 10 個用意した．テストデータ d^* の MFCC を，学習時に求めた k-MFCCs の内，最も距離が近いものに分類し，頻度行列 $n(w|d^*)$ を求める．各テストデータ d^* において，

$$k^* = \arg\max_k \sum_{w \in W} n(w|d^*) P(w|z_k) \qquad (6)$$

(6) 式より母音 z_{k^*} を認識結果とした．このとき，認識率は 100 % であった．この結果は，母音のみであれば，PLSA によって音素を推定することが可能であると示している．

Table 1　$P(z|d)$

	d_1	d_2	d_3	d_4	d_5
z_1	0	1	0	0	0
z_2	0	0	0	0	1
z_3	1	0	0	0	0
z_4	0	0	1	0	0
z_5	0	0	0	1	0

4　おわりに

PLSA による音素体系を用いた構音障害者の音声認識の検討を行った．初期実験により，発話が母音のみであれば，MFCC の発生確率としての母音のモデルが生成可能であることが示せた．

今後は，子音を含めた音響モデルの推定や LDA[6] による音素体系構築を検討していく．

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] T. Hofmann, “Probabilistic Latent Semantic Analysis,” In Proc. Uncertainty in Artificial Intelligence, 1999.
[5] “HTK (Hidden Markov Model Toolkit),” http://htk.eng.cam.af.uk/
[6] David M. Blei, “Latent Dirichlet Allocation,” Journal of Machine Learning Research, 993-1022, 2003.