論文 / 著書情報
Article / Book Information

論題(和文)	パラメータ空間のクラスタ化による固有声話者適応化の改良
Title(English)	
著者(和文)	丹治 秀太朗, 篠田 浩一, 古井貞熙, オルテガ アントニオ
Authors(English)	Shutaro Tanji, Koichi Shinoda, SADAOKI FURUI, antonio ortega
出典(和文)	日本音響学会2008年春季講演論文集, Vol. , No. 2-10-11, pp. 91-94
Citation(English)	, Vol. , No. 2-10-11, pp. 91-94
発行日 / Pub. date	2008, 3

パラメータ空間のクラスタ化による固有声話者適応化の改良*

◎丹治秀太朗, 篠田浩一, 古井貞熙 (東工大), Antonio Ortega(USC)

1 はじめに

話者適応化技術は, 音声認識システムにおいて新しい利用者の声質に合わせた適応を行い, 認識性能を改善させる技術である. 主な手法として, MAP(Maximum A Posterior Probability) 法 [1], MLLR(Maximum Likelihood Linear Regression) 法 [2], 固有声 (Eigenvoice) 法 [3] が広く知られている. 本稿ではより少量の発声での認識性能改善を目的とし, Eigenvoice 法の改良に取り組む。

Eigenvoice 法の改良を行った従来の手法の一つとして, Segmental Eigenvoice 法 [4] があげられる. この手法は, HMM における分布のクラスタリングによりパラメータ空間を分割し, それぞれの空間に対し別々に主成分分析 (PCA) を行うことでより高精度な eigenvoice の作成を目指している.

しかしこの手法で用いられているクラスタリング手法は, 一つの分布が一つのクラスタにしか属することを許されず, 各分布間に様々な相関関係が存在すると考えられるパラメータ空間のクラスタリングには適切でない可能性がある. さらに, 他のどの分布からも極端に距離の離れた分布が存在し, それに対しても対処すべきであると考えられる.

本稿ではこれら 2 つの問題点を解決するため, 分布が複数のクラスタに属することを許したソフトクラスタリングを行い, さらに極端に距離の離れた分布の情報を, eigenvoice 作成に用いるパラメータ空間の共分散行列から取り除く手法を提案する.

2 Eigenvoice 法

2.1 学習段階

Eigenvoice 法では事前知識として, 大規模音声データベースから不特定話者 HMM(SIHMM) の他に, 多数の特定話者 HMM(SDHMM) を作成する. そしてそれぞれの SDHMM に対し, 1 つの SDHMM に含まれる全ての分布の平均ベクトルを並べたベクトルをその話者の話者ベクトルとする. これら話者ベクトルで張られるパラメータ空間は話者変動を表していると考えられる. 話者 p の話者ベクトル $\boldsymbol{x}_p$ と, 全 N 話者の話者ベクトルで張られるパラメータ空間の共分散行列 C は以下の式で表される.

$$\boldsymbol{x}_p = (\mu_{p,1}^t, \cdots, \mu_{p,m}^t, \cdots, \mu_{p,M}^t)^t \quad (1)$$

$$C = \frac{1}{N}\sum_{p=1}^{N}(\boldsymbol{x}_p - \bar{\boldsymbol{x}})(\boldsymbol{x}_p - \bar{\boldsymbol{x}})^t \quad (2)$$

ここで, M は分布の総数であり, $\mu_{p,m}$ は話者 p の m 番目の分布の平均ベクトル, $\bar{\boldsymbol{x}}$ は話者ベクトルの全話者の平均, t は転置を表す. この共分散行列に対し主成分分析 (PCA) を行い, 得られる寄与率の高い固有ベクトルを eigenvoice とする.

2.2 適応段階

新しい話者に適応する際, この話者の HMM の平均ベクトルは eigenvoice $(\boldsymbol{e}_k, k = 1, \cdots, K)$ の線形結合で表されると考える. 線形結合係数を ω_k で表し, 適応後の平均ベクトルを $\hat{\boldsymbol{x}}$ とすると,

$$\hat{\boldsymbol{x}} = \sum_{k=1}^{K}\omega_k \boldsymbol{e}_k + \bar{\boldsymbol{x}} \quad (3)$$

となる. この線形結合係数は, MLED(Maximum Likelihood Eigen Decomposition) 法 [3] により新しい話者の少量の発声から最ゆう推定される.

2.3 Segmental Eigenvoice 法 [4]

Segmental Eigenvoice 法は, eigenvoice の高精度化を目指した手法である. あらかじめ SIHMM の各分布に対しクラスタリングを行い, パラメータ空間の分割を行う. このクラスタリングは, 分布間の距離, 分布間の音韻性の関係, 特徴量の種類の違い, また, それぞれの組合せにより行われる. そしてそれぞれの空間に対し別々の共分散行列を計算し, 別々に eigenvoice を得る. 適応の際は, 得られた特徴ベクトルのパラメータをそれぞれの空間に分割し, それぞれのクラスタごとの eigenvoice の線形結合により適応後のパラメータを推定する.

各分布間には様々な相関関係が存在すると考えられる. しかし, この手法で用いられているクラスタリングでは一つの分布が一つのクラスタにしか属することを許されず、クラスタリングにより分けられた分布同士は自動的に無相関とみなされてしまう. また, 分布の集合に対し極端に距離の離れた分布が存在し, それらを PCA の行う空間に含めると eigenvoice の精度が落ちるという問題も存在する.

---

*Improvement of eigenvoice-based speaker adaptation by parameter space clustering
By Shutaro Tanji, Koichi Shinoda, Sadaoki Furui (Tokyo Institute of Technology), and Antonio Ortega (Univercity of Southern California)

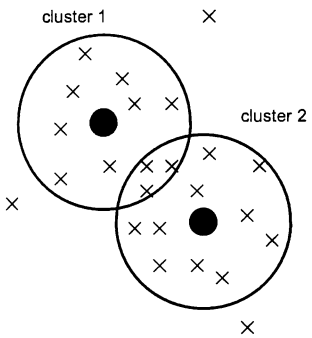

Fig. 1　ソフトクラスタリング

3　Eigenvoice 法の改良

従来法のもつ問題点に対処するため, クラスタリングの際に分布が複数のクラスタに属することを許したクラスタリング手法を提案する. 本稿ではこれをソフトクラスタリングと呼ぶ. さらに, 極端に距離の離れた分布の情報をパラメータ空間の共分散行列から取り除いた上で eigenvoice を作成する.

3.1　ソフトクラスタリング

SIHMM の各分布を用いて, 分布が複数のクラスタに属することを許したクラスタリングを行う. まず分布の集合に対し, k-means 法 [5] を用いてクラスタリングを行う. 距離の尺度としては KLD(Kullback-Leibler Divergence) を用いる. ただし KLD は非対称な値をとるため, 双方の値の平均値を距離として用いる. すなわち, P から Q の KLD を $D_{\mathrm{KL}}(P||Q)$ と表すとき, PQ 間の距離 $D(P,Q)$ を以下のように定義する.

$$D(P,Q) = \frac{D_{\mathrm{KL}}(P||Q) + D_{\mathrm{KL}}(Q||P)}{2} \qquad (4)$$

そして, これで得られた各クラスタの重心から全ての分布への距離を計算し, 閾値以下の値を取る分布をそのクラスタの新たな要素とする. また, これによりどのクラスタにも含まれなかった分布は, 他の分布の集合から極端に距離が離れていると考えられるので, eigenvoice 作成の際にはこの分布の情報は用いないことにする.

2 つのクラスタにクラスタリングを行う場合の概念図を Fig.1 に示す. ×印が各分布のベクトルを表した点, ●印が各クラスタの重心, 円が各クラスタの重心から閾値までの距離を表している. 2 つの円に含まれた分布がソフトクラスタリングにより複数のクラスタに属することになった分布で, どの円にも属さない分布が eigenvoice 作成に用いるパラメータ空間から取り除く分布である.

3.2　ソフトクラスタリングを用いた Eigenvoice 法の実装

PCA を行うパラメータ空間の共分散行列の値を変更することで、ソフトクラスタリングにより得られたクラスタ情報を反映した eigenvoice を作成する.

まず, 話者ベクトルにより張られるパラメータ空間の共分散行列を用意する. これに対し, ある 2 つの分布に着目したとき, どのクラスタにおいてもこの 2 つの分布が同じクラスタ内に存在しなかった場合, 共分散行列におけるこれらの分布が交わるパラメータの要素を 0 とする. これを全ての分布のペアについて行う.

そして, どのクラスタにも含まれなかった分布については, この分布自身の分散共分散のパラメータの要素を 0 とする. この分布はどのクラスタにも属さないため, すでに全ての分布間の共分散が 0 となっているはずである. それゆえ eigenvoice 作成の際には、共分散行列からこの分布の情報が全て取り除かれることとなる.

これにより得られた共分散行列に対して PCA を行い, 寄与率の高い固有ベクトルを新しい eigenvoice とする.

4　認識実験

4.1　実験条件

データベースとして, 新聞記事読み上げ音声コーパス (JNAS)[6], 高齢者の音声認識用大規模データベース (S-JNAS)[7] を用いた. JNAS は各話者, 新聞記事を約 100 文の他に音素バランス文 50 文の計約 150 文を読み上げている. S-JNAS は各話者, 新聞記事を約 100 文の他に音素バランス文 100 文の計約 200 文を読み上げている. 実験には, JNAS の 222 話者 (男女各 111 話者), S-JNAS の 300 話者 (男女各 150 話者) の計 522 話者を学習データとして用い, JNAS の 44 話者 (男女各 22 話者) を認識データとして用いた.

まず準備段階として, 学習データを用い, monophone(音素数 43).1 混合,3 状態の SIHMM を作成した.　このとき各分布は, MFCC(12 次元)+Δ MFCC(12 次元)+Δ パワー (1 次元) の計 25 次元のパラメータをもつ.

次にこの SIHMM に含まれる全ての分布 (音素数 43 × 3 状態 = 129 分布) に対してソフトクラスタリングを行った. 本実験では k-means におけるクラスタの数を, 1, 2, 4(2 分木からさらに 2 分木にしたもの), 8(4 からさらに 2 分木にしたもの) とした. 重心からの閾値は 10, 20, ···, 100 と設定した.

SIHMM から, 学習用 522 話者それぞれのデータを用いて話者数と同じ数の SDHMM を作成した. それ

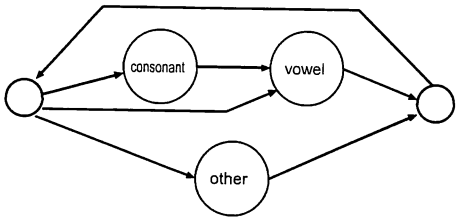

Fig. 2　音素認識で用いる文法

ぞれの SDHMM から, 全ての分布の平均ベクトルを抜き出し, それぞれの話者の話者ベクトルとした. そのときこの話者ベクトルの次元数は, 音素数 43 × 3 状態 × 25 次元の計 3225 次元である. この話者ベクトルにより張られるパラメータ空間の共分散行列を求め, そこから直接 eigenvoice を求めたものが従来の Eigenvoice 法, その共分散行列に修正を加えた後, eigenvoice を求めたものが提案手法となっている.

適応・認識には JNAS の 44 話者のデータを用い, 各話者 20 文を適応用, 50 文を認識用に用いた. 適応用 20 文のうち, 1, 3, 5, 10, 20 文をそれぞれ適応に用いて実験を行っている. 認識の際は音素認識を行い, Fig.2 のような単純な音素認識用の文法を用いた. ここで other には促音「っ」(/q/), 撥音「ん」(/N/), 無音区間 (/sp/) が含まれる. 評価基準としては音素正解精度 (phoneme accuracy) を用いた.

4.2　従来法と提案手法との比較実験

まず, SIHMM, MLLR 法, 従来の Eigenvoice 法, Segmental Eigenvoice 法, 提案手法の比較実験を行った. Eigenvoice 法, Segmental Eigenvoice 法, 提案手法で用いた固有ベクトルの数はどれも 50 としている. これは, 従来の Eigenvoice 法において寄与率が 80 %に達するときの数となっている. Segmental Eigenvoice 法におけるクラスタリングは, 提案手法との比較のため, SIHMM に含まれる全ての分布に対し, 分布間距離に KLD を用いた k-means 法により行われている. Segmental Eigenvoice 法, 提案手法ともクラスタ数は 2 としている. 提案手法のクラスタリングで用いた閾値は,10〜100 のうち事後的に選んだ最適値 60 を用いている.

認識結果を Fig.3 に示す. 横軸が適応に用いた発声数, 縦軸が音素正解精度となっており, Original が従来の Eigenvoice 法, Segmental が Sengental Eigenvoice 法, Soft-clustering が提案手法となっている.

結果を見ると, この条件下では Segmental Eigenvoice 法の効果は見られなかった. これは, クラスタリングにより分けられた分布間に有益な情報があったにも拘らず無相関とみなされ, その情報を失ったことが原因と考えられる. MLLR 法も, 適応発声数が少ないときは効果があまり見られず, 3 発声以下のときは

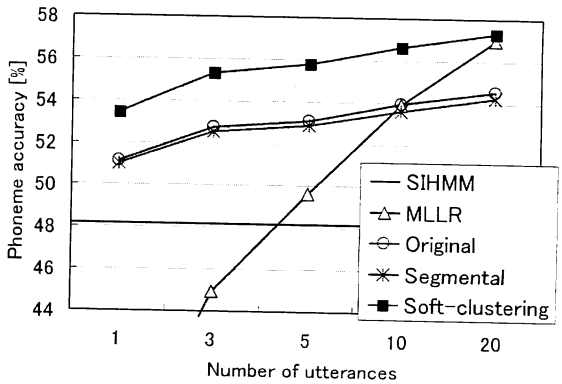

Fig. 3　従来法と提案手法との比較実験

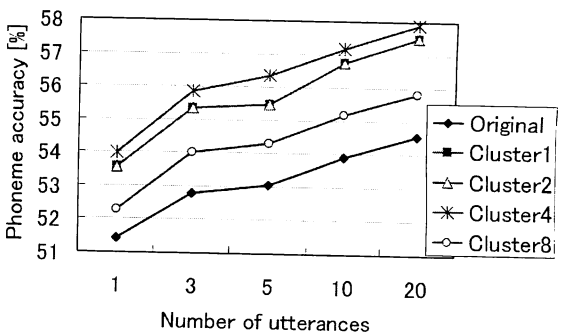

Fig. 4　クラスタ数の違いによる比較実験

適応前の SIHMM に劣る結果となった. それに対し提案手法では, 従来の Eigenvoice 法と比べ高い認識性能を示した. この結果より, eigenvoice 自体の精度が向上していることが確認できる.

4.3　クラスタ数の違いによる比較実験

次に提案手法において, クラスタ数の違いによる認識精度の比較実験を行った. 用いる固有ベクトルの数は先の実験と同様 50 とした. クラスタ数は 1, 2, 4, 8 とし、閾値は各クラスタにおいて最適の値とした. 認識結果を Fig.4 に示す. 横軸が適応に用いた発声数, 縦軸が音素正解精度となっており, Original は従来の Eigenvoice 法, Cluster1, 2, 4, 8 はそれぞれクラスタ数を表している.

従来の Eigenvoice 法から比べると, どのクラスタ数のときも認識性能が向上した. 特にクラスタ数が 4 のとき, 最も良い認識結果となった. また, クラスタ数 1 のときと 2 のときは全く同じ結果となった. これは, クラスタ数 2 のときの最適な閾値がかなり大きく, 1 つのクラスタがもう 1 つのクラスタに包含されてしまい, 結果的にクラスタ数 1 のときと全く同じになったためである. 逆にクラスタ数 8 のときは他のクラスタ数のときより悪い結果となった. これ以上クラスタ数を増やしすぎると効果が得られなくなると考えられる.

phoneme	state	phoneme	state
a:	2	e:	2
i:	2	o:	2
silB	1	silE	3
q	3		

Table 1　どのクラスタにも属さなかった分布

phoneme	state	phoneme	state
o:	3	u:	2
w	2	w	3
gy	3	ry	3
sh	2	N	3
gy	3	ry	3
silE	2	sp	2
sp	3		

Table 2　一部のクラスタにのみ属した分布

ここでクラスタ数ごとの違いを考察する．クラスタ数1のときは，分布全体の重心からの距離が閾値より大きい分布をクラスタから取り除くのみである．よって，従来のEigenvoice法とクラスタ数1の認識性能の差は，極端に距離の離れた分布を取り除いた効果であると考えられ，クラスタ数1とそれ以上のクラスタ数との差は，分布が複数のクラスタに属することを許した効果であると考えられる．ここから，この手法における性能向上は極端に距離の離れた分布を取り除く効果が大きいと考えられる．

それぞれのクラスタリング結果を見ると，長母音の状態2，/q/の状態3，/silB/の状態1，2，/silE/の状態2，3，/sp/の状態2，3がどのクラスタにも属さないという場合が多かった．例として，Table 1，Table 2にクラスタ数4のときのクラスタリング結果を簡単に示す．ここに示されていない分布は全て同じクラスタに含まれている．

長母音の状態2に関しては，伸ばした母音の中央の状態にあたるため，HMMの特徴量における差分の項が非常に小さな値となり，そのため他の分布の集合より距離が遠く離れていた．また/q/，/silB/，/silE/，/sp/に関しては，音響的な特徴自体があまり含まれていない音素であると考えられるため，eigenvoice作成の際には取り除いたほうが良い結果になると考えられる．どちらの場合も話者の違いがあまり表れない音素状態なので，適応の際は取り除いたほうがeigenvoiceの精度向上につながり，結果として認識性能が向上したと考えられる．

5　まとめ

本稿ではEigenvoice法のさらなる改善のため，分布が複数のクラスタに属することを許したクラスタリング手法を提案し，さらに極端に距離の離れた分布の情報を，eigenvoice作成に用いるパラメータ空間の共分散行列から取り除く手法を提案した．比較実験から，従来のEigenvoice法より認識性能の向上が確認され，eigenvoiceの高精度化が達成された．またこの手法においては，複数のクラスタに属することを許したクラスタリングによる効果よりも，分布の集合から極端に距離の離れた分布を取り除いた効果が大きいことがわかった．実際，話者による違いがあまり表れない分布が取り除かれることが多く，それによりeigenvoiceの精度がより向上したと考えられる．

今後は，クラスタを用いずに分布間の距離を用いて共分散行列の要素を直接修正する手法が考えられる．さらに，混合数の増加やtriphoneにするなどパラメータの次元数が増加したとき，この手法により共分散行列を疎な行列に変えることで計算量の削減の効果が期待できると考えられる．

参考文献

[1] J.L. Gauvain *et al.*, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains," IEEE Trans. Speech Audio Prosess., vol.2, no.2, pp.291-298, 1994.

[2] C.J. Leggetter *et al.*, "Maximum likelihood linear regression for speaker adaptation of continuous-density hidden markov models," Computer Speech and Language, vol.9, pp.171-185, 1995.

[3] R. Kuhn *et al.*, "Rapid speaker adaptation in eigenvoice space," IEEE Trans. Speech Audio Process., vol.8, no.6, pp.695-707, 2000.

[4] Y. Tsao *et al.*, "Sengental eigenvoice with delicate eigenspace for improved speaker adaptation," IEEE Trans. Speech Audio Process., vol.13, no.3, pp.399-411, 2005.

[5] J.B. MacQueen, "Some methods for classification and analysis of multivariate observations, *Proseedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*", Berkeley, University of California Press, 1967.

[6] JNAS(新聞記事読み上げ音声コーパス), http://www.milab.is.tsukuba.ac.jp/jnas/.

[7] S-JNAS(高齢者の音声認識用大規模データベース), http://db.ciair.coe.nagoya-u.ac.jp/dbciair/koureisha_files/.