雑音環境下音声認識のための長時間セグメント特徴量に関する検討*

☆正木大介, 鈴木雅之, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

音声認識の性能は背景雑音などによって大幅に低下することが知られている．この問題に対するアプローチの 1 つとして，雑音に頑健な音響特徴量を用いて音声認識を行うという手法がある．中でも線形判別分析（LDA）などの識別学習によって得られる識別的な特徴量は，雑音に頑健な特徴量であることが知られている [1, 2].

従来の LDA による特徴量は，前後数フレームずつからなる合計 10 フレーム程度の時間セグメントの特徴量の連結ベクトルから抽出されることが多い．しかしその一方で，前後 10 フレームずつというさらに長い時間セグメントから抽出された長時間動的特徴量も雑音に頑健な特徴量であることが知られている [3]．したがって LDA による特徴量抽出においても，従来より長い時間セグメントを用いることは有効であると考えられ，従来の LDA において長い時間セグメントから特徴量抽出を行うと性能が低下してしまうのは，推定すべきパラメータ数の増大によって過学習を起こしていることが原因である可能性がある．

そこで本稿では，従来よりも長い時間セグメントを用いた LDA による特徴量抽出の手法を提案する．LDA による特徴量抽出を行う際に，LDA の正則化および特徴量の次元分割（マルチストリーム化）を行うことにより，近接 11 フレームを用いるよりも，近接 31 フレームを用いて特徴量抽出を行ったほうが雑音環境下における音声認識精度が向上することを示した．

2 LDA による特徴量抽出

各時間フレームにおける MFCC などの短時間音響特徴量は，近隣のフレームから大きく影響を受けているはずである．したがって各時間フレームに対して，その近接フレームを含めた時間セグメントが有する短時間特徴量系列は，音声認識に有効な特徴量であると考えられる．しかし複数フレームの特徴量系列は，そのまま各フレームに対する特徴量として用いるには次元数が大きすぎるため，主成分分析（PCA）や線形判別分析（LDA）などを用いて次元圧縮されてから用いられることが多い．本節では，LDA を用いた特徴量抽出の従来手法 [1] について説明する．

各時間フレーム t に対し，その前後 k フレームずつを含めた時間セグメントが有する d 次元の特徴量 $\boldsymbol{c}$ の時系列データ $[\boldsymbol{c}_{t-k}, \cdots, \boldsymbol{c}_{t+k}]$ を連結して，新たに $D = d(2k+1)$ 次元の特徴ベクトル $\boldsymbol{x}_t = [\boldsymbol{c}_{t-k}{}^\top, \cdots, \boldsymbol{c}_{t+k}{}^\top]^\top$ を作り，これを各フレーム t に対する特徴ベクトルとする．また，学習データに強制アライメントをかけることによって得られた各フレーム t に対する HMM の状態ラベルを l_t とする．こうして，学習データの各フレームに対し，特徴ベクトル $\boldsymbol{x}_t$ と教師ラベル l_t が対になったデータが得られる．

全クラス数が L，総データ数が n，各クラス l のデータ数が n_l であり，全データ集合を $\boldsymbol{\mathcal{X}}$，クラス l のデータ集合を $\boldsymbol{\mathcal{X}}_l$，データ全体の平均ベクトルを $\bar{\boldsymbol{x}}$，各クラスの平均ベクトルを $\bar{\boldsymbol{x}}_l$ と表記すると，クラス内分散 $\boldsymbol{S}_\mathrm{W}$，およびクラス間分散 $\boldsymbol{S}_\mathrm{B}$ はそれぞれ，

$$\boldsymbol{S}_\mathrm{W} = \frac{1}{n}\sum_{l=1}^{L}\sum_{\boldsymbol{x}_t\in\boldsymbol{\mathcal{X}}_l}(\boldsymbol{x}_t-\bar{\boldsymbol{x}}_l)(\boldsymbol{x}_t-\bar{\boldsymbol{x}}_l)^\top \quad (1)$$

$$\boldsymbol{S}_\mathrm{B} = \frac{1}{n}\sum_{l=1}^{L} n_l(\bar{\boldsymbol{x}}_l-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_l-\bar{\boldsymbol{x}})^\top \quad (2)$$

と計算できる．ここで D 次元の特徴量 $\boldsymbol{x}_t$ から，クラス内分散 $\boldsymbol{S}_\mathrm{W}$ ができるだけ小さく，クラス間分散 $\boldsymbol{S}_\mathrm{B}$ ができるだけ大きくなるような d' 次元の特徴量 $\boldsymbol{y}_t$ を抽出するための処理が LDA による次元圧縮手法である．$\boldsymbol{x}_t$ から $\boldsymbol{y}_t$ への変換が，$D\times D'$ の変換行列 $\boldsymbol{W}$ を用いて

$$\boldsymbol{y}_t = \boldsymbol{W}^\top\boldsymbol{x}_t \quad (3)$$

と表されるとすると，分散共分散行列 $\boldsymbol{S}$ は変換後の特徴量空間において $\boldsymbol{W}^\top\boldsymbol{S}\boldsymbol{W}$ となるため，

$$\operatorname*{argmin}_{\boldsymbol{W}}\frac{|\boldsymbol{W}^\top\boldsymbol{S}_\mathrm{W}\boldsymbol{W}|}{|\boldsymbol{W}^\top\boldsymbol{S}_\mathrm{B}\boldsymbol{W}|} \quad (4)$$

を解くことによって変換行列 $\boldsymbol{W}$ が得られる．この問題は解析的に解くことができ，

$$\boldsymbol{W} = \mathrm{eig}(\boldsymbol{S}_\mathrm{W}^{-1}\boldsymbol{S}_\mathrm{B}, D') \quad (5)$$

となる．ただし，$\mathrm{eig}(\boldsymbol{A}, D')$ は，D 次元の正方行列 $\boldsymbol{A}$ の固有ベクトルを，対応する固有値が大きい順に D' 個並べてできる行列を返す関数であるとする．こうして得られた LDA の変換行列 W を用いて (3) 式の特徴量変換を行うことにより，クラス情報をラベリングされた D 次元の特徴量 $\boldsymbol{x}_t$ から，D' 次元に圧縮された識別的な特徴量 $\boldsymbol{y}_t$ を得ることができる．

---

* An experimental study on long-term segmental features for noise robust speech recognition. by D. Masaki, M. Suzuki, N. Minematsu, and K. Hirose (The University of Tokyo)

3 提案手法

前節で説明した従来の LDA による特徴量抽出においては，抽出に用いるセグメントの時間幅 k の値を大きくすると，LDA の学習において推定すべきパラメータの数に対して学習データの量が不足してしまい，過学習を起こしやすいと考えられる．そこで本節では，従来よりも長い時間セグメントから過学習を起こさずに，雑音により頑健な特徴量を抽出するための LDA の手法として，LDA の 2 次正則化および特徴量のマルチストリーム化の 2 つの手法を提案し，両手法について説明を行う．

3.1 LDA の 2 次正則化

(4) 式で表される従来の LDA における $\boldsymbol{W}$ の最適化の式に対し，下式のように $\boldsymbol{W}$ の 2 乗ノルム $||\boldsymbol{W}||^2 = \boldsymbol{W}^\top \boldsymbol{W}$ を正則化項として付加することにより，$\boldsymbol{W}$ の各成分が大きくなりすぎないよう制約がかかり，過学習の防止をはかることができる．

$$\underset{\boldsymbol{W}}{\operatorname{argmin}} \left(\frac{|\boldsymbol{W}^\top \boldsymbol{S}_{\mathrm{W}} \boldsymbol{W}|}{|\boldsymbol{W}^\top \boldsymbol{S}_{\mathrm{B}} \boldsymbol{W}|} + \lambda ||\boldsymbol{W}||^2 \right) \tag{6}$$

ここで正則化パラメータ λ は 0 以上の値をとる．上式も解析的に解くことができ，

$$\boldsymbol{W} = \mathrm{eig}((\boldsymbol{S}_{\mathrm{W}} + \lambda \boldsymbol{I})^{-1} \boldsymbol{S}_{\mathrm{B}}, D'') \tag{7}$$

となる．ただし，I は D 次元の単位行列を表している．

3.2 特徴量のマルチストリーム化

これまでに説明した手法では，音響特徴量のすべての次元の成分をひとくくりにして近接フレームの特徴量と連結し，LDA をかけていた．しかし MFCC などのケプストラム特徴量においては，声道長の違いが MFCC に対する帯行列による線形変換で表されることが知られているように [4]，第 1 次元と第 2 次元のように次元が隣合う成分どうしは相関が高いが，第 1 次元と第 5 次元のように次元が離れている成分どうしは相関が低いと考えられる．そこで，d 次元の特徴量を以下のように s 次元ずつのストリームに分割してから LDA を行うことを考える．

$$\begin{aligned}
&\textbf{\textit{stream}}\ 1 : (c^{(1)}, c^{(2)}, \cdots, c^{(s)}) \\
&\textbf{\textit{stream}}\ 2 : (c^{(2)}, c^{(3)}, \cdots, c^{(s+1)}) \\
&\quad \vdots \\
&\textbf{\textit{stream}}\ d-s+1 : (c^{(d-s+1)}, c^{(d-s+2)}, \cdots, c^{(d)})
\end{aligned}$$

つまり，$d \times (2k+1)$ 次元の特徴量時系列データを，$s \times (2k+1)$ 次元の $d-s+1$ 個のストリームに分割し，各ストリームごとに LDA を施して d'' 個ずつの特徴量を抽出していき，それらを連結することにより，$D'' = d''(d-s+1)$ 次元の特徴ベクトルが得られる．

こうして特徴量のマルチストリーム化を行うことにより，次元の離れた成分どうしの相関を無視し，次元の近い成分どうしの相関のみに着目することができる．マルチストリーム化を行わない手法は $s = d$ の場合に相当し，また文献 [2] で用いられている手法は $s = 1$ の場合と同様である．以下，s のことはブロックサイズと呼ぶことにする．マルチストリーム化は，構造的表象を用いた音声認識 [5] や，アフィン変換不変性を有する局所特徴量を用いた音声認識 [6] などにおいてもその有効性が示されている．

4 実験

4.1 実験条件

AURORA-2 データベース [7] を用いた雑音環境下連続数字認識実験を行い，従来より長時間のセグメントを用いて LDA を行う際，LDA の 2 次正則化および特徴量マルチストリーム化の両手法を用いたときの認識率の変化を評価した．なお，LDA による特徴量抽出に用いるセグメントの時間幅を決定するパラメータ k の値は，従来手法の場合を $k = 5$（合計 11 フレーム），今回の提案手法の場合を $k = 15$（合計 31 フレーム）とした．

HMM 音声認識に用いる特徴量は，ケプストラム平均正規化（CMN）[8] をかけた MFCC12 次元とそのエネルギー項，およびその Δ, $\Delta\Delta$ をあわせた合計 $(12+1) \times 3 = 39$ 次元の特徴ベクトルと，近接 11 フレーム，あるいは 31 フレーム分の CMN をかけた MFCC12 次元の連結ベクトルに（6）式の 2 次正則化 LDA を施して得られた特徴ベクトルとの結合ベクトルを用いた．ただし，MFCC の抽出に用いる窓の種類やシフト長などのパラメータは，すべて AURORA-2 の標準として用いられている値を採用した．また，LDA に用いる教師ラベルとしては，CMN をかけた MFCC とそのエネルギー項，およびその Δ, $\Delta\Delta$ をあわせた合計 39 次元の特徴量を用いて，クリーン音声のみの学習データとして学習した音響モデルによる強制アライメントによって得られた HMM の状態ラベルを用いた．さらに，マルチストリーム化におけるブロックサイズ s の値については，マルチストリーム化を行わない場合（$s = 12$），および $s = 1, 2, 3$ の 3 通りの値でマルチストリーム化を行う場合の合計 4 通りの条件を設定した．なお，LDA によって得られる特徴ベクトル次元数については，マルチストリーム化を行わない場合は 12 次元，マルチストリーム化を

Table 1　$k = 5$（合計 11 フレーム）のとき

Test set	Set A	Set B	Set C
s=12, 正則化なし	64.74%	64.24%	70.94%
s=12, 正則化あり	68.86% ($\lambda = 10^5$)	69.14% ($\lambda = 10^5$)	72.36% ($\lambda = 10^3$)
s=1, 正則化なし	66.09%	65.51%	70.39%
s=1, 正則化あり	68.15% ($\lambda = 0.1$)	67.66% ($\lambda = 0.1$)	72.76% ($\lambda = 0.1$)
s=2, 正則化なし	69.51%	69.12%	73.64%
s=2, 正則化あり	**69.51%** ($\lambda = 0$)	69.12% ($\lambda = 0$)	**73.64%** ($\lambda = 0$)
s=3, 正則化なし	65.79%	65.79%	70.83%
s=3, 正則化あり	68.81% ($\lambda = 10$)	**69.55%** ($\lambda = 10$)	72.72% ($\lambda = 10$)

Table 2　$k = 15$（合計 31 フレーム）のとき

Test set	Set A	Set B	Set C
s=12, 正則化なし	61.60%	63.27%	68.75%
s=12, 正則化あり	67.65% ($\lambda = 10^5$)	67.05% ($\lambda = 10^4$)	72.51% ($\lambda = 10^5$)
s=1, 正則化なし	66.15%	65.86%	71.44%
s=1, 正則化あり	67.33% ($\lambda = 10^2$)	67.16% ($\lambda = 1$)	71.67% ($\lambda = 1$)
s=2, 正則化なし	66.57%	67.04%	72.08%
s=2, 正則化あり	**70.34%** ($\lambda = 0.1$)	**70.67%** ($\lambda = 0.1$)	**74.58%** ($\lambda = 0.1$)
s=3, 正則化なし	67.11%	67.43%	71.37%
s=3, 正則化あり	69.48% ($\lambda = 10$)	70.00% ($\lambda = 10$)	73.13% ($\lambda = 10$)

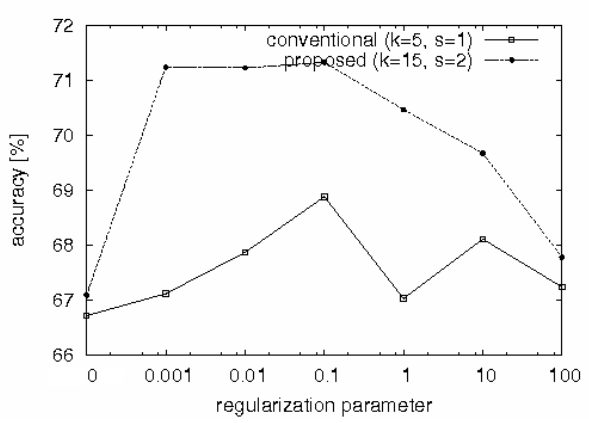

Fig. 1　従来手法および提案手法における正則化パラメータを変化させたときの認識率の変化

行う場合は各ストリームから 1 次元ずつの合計 $13 - s$ 次元とした．したがって，認識に用いる特徴量の次元数は，マルチストリーム化を行わない場合は合計 $39 + 12 = 51$ 次元，マルチストリーム化を行う場合は合計 $52 - s$ 次元である．以上で説明した 2 通りの時間セグメント長，および 4 通りのマルチストリーム化条件を組み合わせた合計 8 通りの条件で特徴量抽出を行い，AURORA-2 を用いた HMM 音声認識実験を行った．ただし，HMM の各種パラメータは，すべて AOURORA-2 の complex backend に準拠した値を用い，音響モデルの学習にはクリーン音声のみからなる学習データを用いた．評価セットにはセット A，B，C の 3 つがあり，セット A と B にはそれぞれ 4 種類ずつ，セット C には 2 種類の背景雑音が，それぞれ 20–0 dB の 5 段階の SNR で重畳されており，各セットの認識率はセットごとの全条件の認識率の平均値とした．セット A は雑音環境クローズド，セット B は雑音環境オープンであるが，本実験ではクリーン音声のみでを学習行ったため，とくに差異はない．また，セット C は背景雑音に加えてチャネルノイズも含まれている．

(6) 式の正則化パラメータ λ をさまざまな値に設定しながら，認識率を調べていくことにより，それぞれの特徴量抽出条件における適切な λ の値を決定した．なお，$\lambda = 0$ のときは正則化を行っていない場合に相当する．

4.2　実験結果

$k = 5$（合計 11 フレーム）のときの結果を Table 1 に，$k = 15$（合計 31 フレーム）のときの結果を Table 2 に示した．なお，各 Table 中の $s = 12$ はマルチストリーム化を行っていない場合を表している．また，

正則化ありの場合では，それぞれの条件において最も認識率が高かったときの結果のみを示しており，括弧内に示したのはそのときの λ の値である．各セットにおいて最も認識率が高かったものを太字で示した．また，従来手法と提案手法の比較を行うために，両手法における正則化パラメータ λ を変化させたときの認識率の変化を Fig. 1 に示した．ここで示した認識率は，セット A，B，C すべての平均値である．なお，ここでいう従来手法とは時間セグメント長を 11 フレーム, ブロックサイズを $s = 1$ とした場合を表しており，文献 [2] で用いられている手法に相当する．また，提案手法は最も認識率の高かった 31 フレーム，$s = 2$ の場合を表している．

正則化およびマルチストリーム化のいずれか片方を用いただけでは，まだ 31 フレームよりも 11 フレームを用いたときのほうが認識率が高いが，両手法を適切に組み合わせることにより 31 フレームが 11 フレームのときの認識率を上回った．とくにブロックサイズ $s = 2$ でのマルチストリーム化の効果が高いことが分かった．この結果は，MFCC の隣り合う次元の成分どうしが強い相関を持っていることを表しているといえる．

5 まとめ

従来の LDA による特徴量抽出手法よりも長時間のセグメントを用いて，LDA によって雑音により頑健な特徴量の抽出を行うために，LDA の 2 次正則化および特徴量のマルチストリーム化の 2 つの手法を提案した．セグメント長が 31 フレームの条件において，ブロックサイズ $s = 2$ でのマルチストリーム化を施した上で，正則化パラメータ $\lambda = 0.1$ での LDA による特徴量抽出を行うことにより，文献 [2] に基づくベースライン条件 (11 フレーム，$s = 1, \lambda = 0$) に対してセット A で 13%，セット B で 15%，セット C で 14%の誤り改善率が得られた．

参考文献

[1] H. Soltau, G. Saon, and B. Kingsbury, “The IBM ATTILA speech recognition toolkit,” IEEE Workshop on *SLT*, pp. 97–102, 2010.

[2] Jeih-Weih Hung and Lin-shan Lee, “Optimization of temporal filters for constructing robust features in speech recognition,” IEEE Trans. Audio, Speech and Language Processing, Vol. 14, No. 3, pp. 808–832, 2006.

[3] T. Fukuda, O. Ichikawa, and M. Nishimura, “Short- and long-term dynamic features for robust speech recognition,” Proc. Interspeech, pp. 2262–2265, 2008.

[4] M. Pitz, and H. Ney, “Vocal tract normalization equals to linear transformation in cepstral space,” IEEE Trans. Speech and Audio Processing, Vol. 13, No. 5, pp. 930–944, 2005.

[5] S. Asakawa, N. Minematsu, and K. Hirose, “Multi-stream parameterization for structural speech recognition,” Proc. *ICASSP*, pp. 4097–4100, 2008.

[6] 鈴木雅之, 喬宇, 峯松信明, 広瀬啓吉, “アフィン変換不変性を有する局所特徴量を用いた音声認識,” 信学技報, SP2008-12, pp. 209–214, 2008.

[7] H. G. Hirsch and D. Pearce, “The AURORA ex-perimental framework for the performance evaluation of speech recognition systems under noisyconditions,” Proc. *ISCA ITRW ASR*, pp. 29–32, 2000.

[8] B. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, Journal of the Acoustical Society of America, Vol. 55, No. 6, pp. 1304–1312, 1974.