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*

◎鈴木雅之, 喬宇, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

音声信号には，言語情報と話者情報が類似音響量を用いて符号化されており，同一の言語情報を有していても，話者が異なれば信号は大きく異なる．音声認識は，音声信号から言語情報のみを抜き出すタスクであるため，不特定話者の音声を認識する場合は，話者情報の違いが一種のノイズになる．そのため，一般的に不特定話者の音声認識率は，特定話者の音声認識率より低い．

このような背景から，話者の違いによる音声認識率への影響を低減するために，多くの方法が提案されてきた．これらの手法は概ね，2つに分類することができる．1つ目は入力音声を正規化して標準的な話者で作成した音響モデルに合わせるもので，声道長正規化 [1] などがそれに該当する．2つ目は音響モデルを入力音声に合わせて適応するもので，MLLR 適応 [2] などがそれに該当する．

これらの手法により，認識率の向上が数多く報告されているが，問題点も残されている．例えば，話者が次々と入れ替わるような場合，正規化・適応のためのデータが十分に得られず，効果が低くなってしまう．少ないデータで話者適応を行う技術なども近年提案されているが，それでも話者が入れ替わった直後にはデータの不足問題が生じる．

本稿では，このような問題点を解決する，話者の違いにあまり影響を受けない特徴量の利用を提案する．提案する特徴量は，ケプストラムに話者情報を取り除く処理を施すことにより得られる．ケプストラムのリフタリング処理は音源情報を取り除くために行われるが，提案手法はその話者情報版と考えることができる．

話者の違いに頑健な特徴量に関する先行研究としては，Rademacher や Irino の研究 [3, 4] がある．これらの研究では，線形の周波数ウォーピングに対する不変量が提案されている．一方我々は，アフィン変換に対して不変な局所特徴量（Localized Affine Invariant Feature; LAIF）を用いる [5, 6]．ケプストラムへのアフィン変換は，線形の周波数ウォーピングよりもより多く話者性を表現するため，LAIF の方がより話者の違いに対する頑健性が高いと考えられる．

本稿の構成は以下の通りである．第二節では，提案手法で用いる LAIF について述べる．第三節では，LAIF の不特定話者音声認識に対する効果を実験的に検証する．最後に第四節で，本稿の結論と今後の展望を述べる．

2 アフィン変換不変性を有する局所特徴量

まず，アフィン変換に不変な局所特徴量（LAIF）について述べる．$\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_T}]$ を，d 次元のケプストラムベクトル $\boldsymbol{x_t}$ の時系列データとおく．ここで，以下のような $\boldsymbol{x_t}$ に対するアフィン変換について考える．

$$\boldsymbol{x_t}' = \boldsymbol{A}\boldsymbol{x_t} + \boldsymbol{c} \tag{1}$$

ここで $\boldsymbol{A}$ は $d \times d$ の正則な定行列であり，$\boldsymbol{c}$ は d 次元の定ベクトルである．また，アフィン変換後の $\boldsymbol{X}$ を，$\boldsymbol{X'} = [\boldsymbol{x_1}', \boldsymbol{x_2}', \cdots, \boldsymbol{x_T}']$ で表すことにする．

ここで LAIF とは，ある時間フレーム t 付近における $\boldsymbol{X}$ の部分ベクトル列

$$\boldsymbol{X_{t-k_1:t+k_2}} = [\boldsymbol{x_{t-k_1}}, \boldsymbol{x_{t-k_1+1}}, \cdots, \boldsymbol{x_t}, \cdots, \boldsymbol{x_{t+k_2}}]$$

と，それにアフィン変換をかけたもの $\boldsymbol{X'_{t-k_1:t+k_2}}$ において共通の値を持つ関数，すなわち

$$F(\boldsymbol{X_{t-k_1:t+k_2}}) = F(\boldsymbol{X'_{t-k_1:t+k_2}}) \tag{2}$$

が成り立つような F の返り値である．

$F(\boldsymbol{X_{t-k_1:t+k_2}})$ の計算には，t より k_1 だけ前のフレームから k_2 だけ後のフレームまでのケプストラムベクトルが用いられる．このように，データの部分ベクトル列から新たな特徴量を計算するという考え方は，デルタ特徴量抽出の考え方と同じである．デルタ特徴量はケプストラムの回帰係数に相当する特徴量であり，音声の動的な特性を表現するのに有用であるが，通常 $\boldsymbol{\Delta}(\boldsymbol{X_{t-k:t+k}}) \neq \boldsymbol{\Delta}(\boldsymbol{X'_{t-k:t+k}})$ となるため，デルタ特徴量は LAIF ではない．

ここで，以下に示す $F(\boldsymbol{X_{t-k_1:t+k_2}})$ は LAIF を返す．

$$F(\boldsymbol{X_{t-k_1:t+k_2}}) = \sqrt{(\boldsymbol{\mu_a} - \boldsymbol{\mu_b})^T (\boldsymbol{\Sigma_a} + \boldsymbol{\Sigma_b})^{-1} (\boldsymbol{\mu_a} - \boldsymbol{\mu_b})} \tag{3}$$

ただし，$\boldsymbol{\mu}$ は平均ベクトルを，$\boldsymbol{\Sigma}$ は分散共分散行列を表す．添字 a はフレーム t より前の時間フレーム $[t-k_1, \cdots, t-1]$ を，添字 b はフレーム t 以後の時

*Speech recognition using localized affine invariant features. by M.Suzuki, Y.Qiao, N.Minematsu, and K.Hirose (The University of Tokyo)

間フレーム $[t, \cdots, t+k_2]$ を表す．$\boldsymbol{\mu_a}$ および $\boldsymbol{\Sigma_a}$ は，以下の式で ML 推定できる．

$$\boldsymbol{\mu_a} = \frac{1}{k_1} \sum_{\tau=t-k_1}^{t-1} \boldsymbol{x_\tau} \tag{4}$$

$$\boldsymbol{\Sigma_a} = \frac{1}{k_1} \sum_{\tau=t-k_1}^{t-1} (\boldsymbol{x_\tau} - \boldsymbol{\mu_a})(\boldsymbol{x_\tau} - \boldsymbol{\mu_a})^T \tag{5}$$

$\boldsymbol{\mu_b}$ および $\boldsymbol{\Sigma_b}$ も同様に推定できる．

時系列データとして LAIF を抽出するためには，t を一つずつずらしながら式 (3) を計算していけばよい．ここで，$t = n$ の場合，F は，$\boldsymbol{X_{n-k_1:n+k_2}}$ に対するアフィン変換に不変となる．また，$t = m$ の場合は，F は，$\boldsymbol{X_{m-k_1:m+k_2}}$ に対するアフィン変換に不変となる．ここで，$\boldsymbol{X_{n-k_1:n+k_2}}$ に対するアフィン変換と，$\boldsymbol{X_{m-k_1:m+k_2}}$ に対するアフィン変換は，同一のアフィン変換でなくてもよい．すなわち LAIF は，ケプストラムの時系列データすべてに対する単一のアフィン変換のみに不変なのではなく，局所的な部分におけるアフィン変換に不変な特徴量となる．話者の違いを近似するアフィン変換は音素の種類によって異なるため，これはより現実に則した性質であるといえる．

2.1 アフィン変換不変性の証明

LAIF のアフィン変換不変性を証明する．平均ベクトル $\boldsymbol{\mu_a}$ や分散共分散行列 $\boldsymbol{\Sigma_a}$ を用いて，アフィン変換後の平均ベクトル $\boldsymbol{\mu'_a}$ と分散共分散行列 $\boldsymbol{\Sigma'_a}$ を表すと，

$$\boldsymbol{\mu_a} = \boldsymbol{A\mu_a} + \boldsymbol{c} \tag{6}$$

$$\boldsymbol{\Sigma'_a} = \boldsymbol{A\Sigma_a A}^T \tag{7}$$

となることから，

$$\begin{aligned}
&F(\boldsymbol{X'_{t-k_1:t+k_2}}) \\
&= \sqrt{(\boldsymbol{\mu'}_b - \boldsymbol{\mu'}_a)^T (\boldsymbol{\Sigma'_a} + \boldsymbol{\Sigma'_b})^{-1} (\boldsymbol{\mu'}_b - \boldsymbol{\mu'}_a)} \\
&= \sqrt{(\boldsymbol{A\mu_b} - \boldsymbol{A\mu_a})^T (\boldsymbol{A(\Sigma_a + \Sigma_b)A}^T)^{-1} (\boldsymbol{A\mu_b} - \boldsymbol{A\mu_a})} \\
&= \sqrt{(\boldsymbol{\mu_b} - \boldsymbol{\mu_a})^T \boldsymbol{A}^T (\boldsymbol{A}^T)^{-1} (\boldsymbol{\Sigma_a} + \boldsymbol{\Sigma_b})^{-1} \boldsymbol{A}^{-1} \boldsymbol{A} (\boldsymbol{\mu_b} - \boldsymbol{\mu_a})} \\
&= F(\boldsymbol{X_{t-k_1:t+k_2}}) \quad \square
\end{aligned} \tag{8}$$

以上により式 (3) が LAIF であることが証明された．

式 (3) 以外にも，LAIF は数多く存在する [5]．ただし，本稿では，特に断りのない限り式 (3) を LAIF として使用することにする．

2.2 特徴量マルチストリーム化

LAIF はアフィン変換に不変である．ここでアフィン変換は話者の違いなどを近似する変換であるので，LAIF を抽出する処理は，ケプストラムから話者情報を取り除くような処理になっているといえる．しかしながらアフィン変換は，話者情報と同時に，言語情報の一部も近似している．そのため，LAIF は話者の違いに頑健であると同時に，言語情報の識別能力まで低くなってしまう．アフィン変換不変性というものは，音声認識というタスクにとっては不変性が強すぎるのである [7]．そこで，話者の違いなどのみに不変で，言語情報には不変にならないように，適切な制約条件を導入することを考える．

ここで，ケプストラムベクトル $\boldsymbol{x}$ に対する話者変換を表すアフィン変換 $\boldsymbol{Ax} + \boldsymbol{c}$ の，$\boldsymbol{A}$ に注目する．話者変換を周波数ウォーピングと仮定すると，$\boldsymbol{A}$ はおおよそ帯行列のような形になることが知られている [8, 9]．この帯行列の幅は，周波数ウォーピングを大きくかければかけるほど広くなっていく．

このような帯行列 $\boldsymbol{A}$ のみに不変となるような制約条件を課すため，特徴量マルチストリーム化を導入する．この手法は，我々の先行研究である音響不変構造を用いた音声認識に関する検討の中で，既に有効性が確認されている [7]．以下，特徴量マルチストリーム化の説明を行う．

ケプストラムベクトル $\boldsymbol{x}$ を 2 つの部分ベクトル $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}$ に分けることを考える．これらの部分ベクトルを用いてそれぞれで LAIF を抽出した場合，各 LAIF は以下のそれぞれのアフィン変換に対して不変となる．

$$\boldsymbol{x'}^{(1)} = \boldsymbol{A}^{(1)} \boldsymbol{x}^{(1)} + \boldsymbol{c}^{(1)} \tag{9}$$

$$\boldsymbol{x'}^{(2)} = \boldsymbol{A}^{(2)} \boldsymbol{x}^{(2)} + \boldsymbol{c}^{(2)} \tag{10}$$

両式をまとめると下記のようになる．

$$\begin{pmatrix} \boldsymbol{x'}^{(1)} \\ \boldsymbol{x'}^{(2)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}^{(1)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}^{(2)} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}^{(1)} \\ \boldsymbol{x}^{(2)} \end{pmatrix} + \begin{pmatrix} \boldsymbol{c}^{(1)} \\ \boldsymbol{c}^{(2)} \end{pmatrix} \tag{11}$$

このように特徴量を分割することで，行列 $\boldsymbol{A}$ の右上・左下要素を $\boldsymbol{0}$ にすることが可能となる．$\boldsymbol{A}$ を幅 s の帯行列にするには，以下のように，隣接する s 個の特徴量をまとめて一つのストリームとし，次元を一つずつずらしながら複数のストリームに分割すればよい．

$\boldsymbol{stream}\ 1 : (x^{(1)}, x^{(2)}, \cdots, x^{(s)})$

$\boldsymbol{stream}\ 2 : (x^{(2)}, x^{(3)}, \cdots, x^{(s+1)})$

$\vdots$

$\boldsymbol{stream}\ d-s+1 : (x^{(d-s+1)}, x^{(d-s+2)}, \cdots, x^{(d)})$

このように重複を含めた形で分割して各ストリームでそれぞれ LAIF 抽出を行うことにより，$\boldsymbol{A}$ を帯行

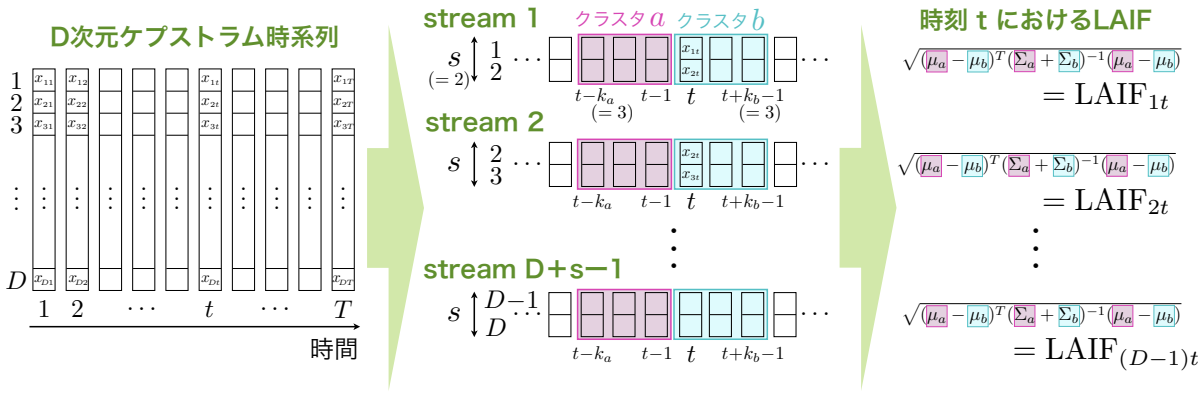

Fig. 1　特徴量次元分割を導入した LAIF の抽出

列の形式にするのと同等の制約条件をかけることができる．ここで s の値は，各ストリームにおける特徴量ブロックの大きさを表しており，これをブロックサイズと呼ぶ．ブロックサイズ s の値が小さいと不変性に強い制約をかけることになり，逆に s が大きいと不変性の制約条件を弱くすることになる．

特徴量マルチストリーム化を導入した場合の LAIF の抽出過程を，Fig. 1 にまとめる．特徴量マルチストリーム化により，$\boldsymbol{F}(\boldsymbol{X}_{\boldsymbol{t-k_1:t+k_2}})$ は $d-s+1$ 次元のベクトルとなる．

3　実験

本節では，不特定話者音声認識に対する効果を実験的に検証する．それに先立ち，LAIF を抽出する際の各種パラメータの設定値を予備実験により決定した．

LAIF を抽出する際の窓の幅を決めるパラメータである k_1, k_2 としては，$k_1 = k_2 + 1 = 16$ を用いる．ケプストラムを抽出する際の分析窓の窓長は 25msec，シフト長は 10msec に設定しているため，1 つのクラスタは 175msec 程度の時間幅を持つことになる．これは，音声に含まれる変調周波数成分のうち 4Hz 付近に最も言語情報が含まれているという Kanedera らの結果におおよそ対応がとれる [10]．また，特徴量マルチストリーム化におけるブロックサイズ s は，1，2，3 を用いる．

実験に用いる特徴量としては，LAIF は MFCC のようなスペクトル特徴量とは異なる音声の特徴を捉えていると考えられるため，MFCC と LAIF の結合ベクトルを用いることにする．具体的には，Table 1 に示した 5 種類それぞれについて認識実験を行う．

3.1　データベース

実験には，東北大松下音声単語データベースを用いる [11]．このデータベースは，男声話者 30 名と女声話者 30 名による，日本語 212 単語の孤立発声音が収録されている．また，音声は 12bit/16kHz でサンプリングされている．実験では，これを 16bit/16kHz に再サンプリングした音声ファイルを用いた．

Table 1　使用する特徴量

特徴量の種類 (次元数)
MFCC(12) + ΔMFCC(12)
MFCC(12) + ΔMFCC(12) + ΔΔMFCC(12)
MFCC(12) + ΔMFCC(12) + $\mathrm{LAIF}_{s=1}$(12)
MFCC(12) + ΔMFCC(12) + $\mathrm{LAIF}_{s=2}$(11)
MFCC(12) + ΔMFCC(12) + $\mathrm{LAIF}_{s=3}$(10)

これらの音声データに，窓幅 25msec，シフト長 10msec のハミング窓をかけ，$1-0.97z^{-1}$ をかけて高域強調を行い，24 次元のメルフィルタバンク出力を DCT して 12 次元の MFCC を抽出した．さらに MFCC から，ΔMFCC，ΔΔMFCC，及び LAIF を抽出し，それらを結合することにより特徴量ベクトルの時系列を作成した．

3.2　不特定話者音声認識

HMM による音声認識を行ない，LAIF の不特定話者音声認識に対する有効性を評価する実験を行った．HMM は単語単位で作成し，1 単語につき 25 状態の left-to-right 型 HMM，各状態の出力確率分布としては対角共分散の 4 混合正規分布を用いた．通常の不特定話者音声認識タスク（Matched condition）として，学習話者を男声 15 名／女声 15 名，評価話者を男声 15 名／女声 15 名としたタスクと，学習データと評価データのミスマッチを大きくした条件の実験として，学習話者を男声 30 名，評価話者を女声 30 名とする実験，および学習話者を女声 30 名，評価話者を男声 30 名とする，計 3 種類の実験を行なった．

Table 2　認識実験の結果．M，Δ，ΔΔ，L はそれぞれ MFCC，ΔMFCC，ΔΔMFCC，LAIF を表す．また，s はマルチストリーム構造化のブロックサイズを表す．

特徴量の種類	M+Δ	M+Δ+ΔΔ	M+Δ+$L_{s=1}$	M+Δ+$L_{s=2}$	M+Δ+$L_{s=3}$
Matched condition	99.47%	99.42%	99.51%	99.39%	99.29%
Training: 男声 – Testing: 女声	82.79%	86.88%	88.35%	89.27%	90.06%
Training: 女声 – Testing: 男声	85.34%	89.67%	89.88%	90.70%	90.83%

実験結果を Table2 に示す．結果，MFCC+ΔMFCC 単独で認識を行うよりも，LAIF を付け加えた方がより不特定話者に対する頑健性が高くなることがわかる．特に，学習データと評価データに性別のミスマッチがある場合，MFCC+ΔMFCC に対し $\mathrm{LAIF}_{s=3}$ を結合することによるエラー削減率は 40.04%となった．MFCC+Δ+ΔΔ と比べても，エラー削減率は 18.51%となった．また，ブロックサイズ s が小さいほど，ミスマッチがない条件では認識率がより良く，ブロックサイズ s が大きいほど，ミスマッチがある条件では認識率がより良くなる傾向が得られた．この結果は，ブロックサイズが小さいほど，アフィン変換不変性への制約条件が大きくなるという性質と対応がとれる．

4　まとめ

本稿では，アフィン変換不変性を持つ局所的特徴量 LAIF を不特定話者音声認識のために用いることを提案した．ケプストラムのアフィン変換は話者の違いを近似するため，LAIF は話者の違いに対しておよそ不変となり，不特定話者音声認識に効果があると考えられる．LAIF を使って不特定話者音声認識実験を行った結果，MFCC+ΔMFCC に対して，$k_1 = k_2 + 1 = 16$, $s = 3$ で抽出した LAIF を加えることで，ミスマッチ条件下において 40.04%の誤り削減率を得た．

LAIF は，話者不変性を持ち，しかも話者正規化や話者適応などと異なり非常に簡単に計算できるという性質から，不特定話者音声認識以外にもさまざまな応用が考えられる．例えば，パラ言語情報の識別や，発音教育システムなどへの応用が考えられる．

参考文献

[1] E.Eide and H.Gish, “A parametric approach to vocal tract length normalization,” *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, vol.1, pp.346–348, 1996.

[2] C.J. Leggetter and P.C. Woodland, “Mazimum likelihood speaker adaptation of continuous density hidden Markov models,” *Computer Speech and Language*, Vol. 9, pp. 171–185, 1995.

[3] J. Rademacher, *et.al.* “Improved warping-invariant features for automatic speech recognition,” *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 1499–1502 2006.

[4] T. Irino, *et.al.* “Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform,” *Speech Communication*, vol.22, pp. 181–203 2002.

[5] Y. Qiao, *et.al.* “Affine invariant features and its application to speech recognition,” *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 2009 (accepted).

[6] 鈴木雅之 他, “アフィン変換不変性を有する局所的特徴量を用いた音声認識,” 信学技報, SP2008-12, pp.209-214, 2008.

[7] S. Asakawa, *et.al.* “Multi-stream parameterization for structural speech recognition,” *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 4097–4100, 2008.

[8] M. Pitz, *et.al.* “Vocal tract normalization equals linear transformation in cepstral space,” *IEEE Trans. Speech and Audio Processing*, vol.13, pp.930–944, 2005.

[9] 江森正 他, “音声認識のための高速最ゆう推定を用いた声道長正規化,” 電子情報通信学会論文誌 , vol.J83–D–II, no.11, pp.2108–2117, 2000.

[10] N. Kanedera, *et.al.* “On the relative importance of various components of the modulation spectrum for automatic speech recognition,” *Speech Communication*, vol. 28, no. 1, pp. 4355, 1999.

[11] 牧野正三 他, “東北大-松下単語音声データベース,” 日本音響学会誌, vol.48, no.12, pp.899–905, 1992.