社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－SLP－74（36）
2008 / 12 / 10

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

鈴木 雅之† 喬 宇† 峯松 信明† 広瀬 啓吉††

† 東京大学大学院工学系研究科 〒113-8656 東京都文京区本郷 7-3-1
†† 東京大学大学院情報理工学系研究科 〒113-0033 東京都文京区本郷 7-3-1
E-mail: †{suzuki,qiao,mine,hirose}@gavo.t.u-tokyo.ac.jp

あらまし 本稿では，不特定話者音声認識における音響特徴量として，アフィン変換不変性を持つ局所特徴量 (Localized Affine Invariant Features; LAIF) を提案する．LAIF は，ケプストラムベクトル時系列から直接計算することができる特徴量である．話者の違いは，ケプストラムベクトルに対するアフィン変換で近似できることから，ケプストラムベクトルから抽出した LAIF は話者の違いにおよそ不変となる．そのため LAIF を用いれば，話者正規化や話者適応のための学習データがまったく得られない状況でも，話者性に頑健な音声認識を実現することができる．我々は，不特定話者の日本語孤立単語音声認識に LAIF を用いる実験を行った．実験の結果，LAIF を MFCC や ΔMFCC と結合して用いることにより，不特定話者音声認識の認識率を向上させることができた．特に，学習データと評価データで性別のミスマッチがある場合，MFCC+ΔMFCC+LAIF は，MFCC+ΔMFCC と比較して 37%のエラー削減率を実現した．

キーワード 音響特徴量，不特定話者音声認識，アフィン変換，局所的特徴，話者不変

Speech recognition using localized affine invariant features

Masayuki SUZUKI†, Yu QIAO†, Nobuaki MINEMATSU†, and Keikichi HIROSE††

† Grad. School of Engineering, Univ. of Tokyo 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
†† Grad. School of Info. Sci. and Tech., Univ. of Tokyo 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan
E-mail: †{suzuki,qiao,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract This paper proposes localized affine invariant features (LAIFs) for speaker-independent automatic speech recognition. The LAIFs can be calculated directly from data sequences. As speaker variations can be approximated well by affine transform in a cepstral space, the LAIFs can provide robust features with respect to those variations. This fact inspires us to expect that the use of the LAIFs should improve the recognition performance especially when no training data is available for speaker normalization or adaptation. To verify this expectation, we apply LAIFs for isolated word recognition. The experimental results show that the combination of LAIFs with MFCC or MFCC+ΔMFCC can lead to higher performances than MFCC or MFCC+ΔMFCC only. Especially in mismatched conditions, MFCC+ΔMFCC+LAIFs can reduce the error rates by 37% when compared to MFCC+ΔMFCC only.

Key words Acoustic features, Speaker independent ASR, Affine transform, Localized features, Speaker invariance

1. はじめに

音声信号には，言語情報と話者情報が同時に符号化されているため，同一の言語情報を持つカテゴリに属する音声であっても，話者の違いによって信号の性質が大きく異なる．そのため，音声から言語情報のみを抜き出すタスクである音声認識では，話者情報が言語情報に対する歪みとなって，認識率を低下させてしまう．実際，不特定話者 (Speaker Independent; SI) 音声認識の誤り率は，特定話者 (Speaker Dependent; SD) 音声認識の誤り率の約 2 倍となることが知られている [1].

このような背景から従来，SI 音声認識の認識率を SD 音声認識の認識率に近づけるための手法が数多く行く提案されてきている．これらの手法の主なものは，大きく 2 種類に分類できる．1 つ目は入力音声を正規化して標準パターンに合わせるもの，2 つ目は逆に音響モデルを入力音声に合わせて適応するものである．1 つ目の入力音声の正規化としては，声道長正

規化（Vocal Tract Length Normalization; VTLN）[2] がよく用いられている．話者の声道長の違いは，音声スペクトルに対する周波数ウォーピングで近似することができる．VTLN は，入力音声から話者の声道長推定し，周波数ウォーピングによって標準的な話者のスペクトルに変換を行うものである．2 つ目の音響モデル適応としては，例えば最尤線形回帰（Maximum Likelihood Liniear Regression; MLLR）に基づく手法 [3] があげられる．話者の違いは，音声スペクトル情報を持つ特徴量であるケプストラム $\boldsymbol{x}$ のアフィン変換 $\boldsymbol{Ax}+\boldsymbol{c}$ として近似することができる．MLLR は，入力音声から話者変換パラメータ $\boldsymbol{A},\boldsymbol{c}$ を推定し，これを用いて音響モデルのパラメータを変換するものである．

これらの手法により，SI 音声認識の認識率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/変換パラメータの推定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ことがある [4]．そのため，非常に高速な話者適応技術なども近年提案されているが [5]，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，話者に不変な特徴量の利用を提案する．話者不変量を利用することで，正規化や適応といった処理を一切行わないまま，SI 音声認識の性能を改善することができる．話者不変量を用いた音声認識は，正規化のために事前に準備するデータがまったく必要ない話者正規化手法ということができる．

我々の提案手法では，話者の違いを近似するアフィン変換に対して不変な局所特徴量（Localized Affine Invariant Feature; LAIF）[6] を用いる．ケプストラムへのアフィン変換は話者の違いを近似しているため，LAIF は話者の違いにおよそ不変になる．また LAIF は，既存のスペクトル特徴量（例えば MFCC など）などに簡単な計算を行うことで抽出する特徴量であるため，既存の話者正規化や話者適応といった手法と組み合わせて用いることが容易である．これにより，適応用のデータがないときの認識率改善に加え，データが得られたときには既存の話者正規化/適応技術を用いてよりよい認識率を得ることができるようになる．

話者の違いに不変な特徴量に関する先行研究としては，Mertins や Irino の研究がある [7]〜[9]．これらの研究は，線形の周波数ウォーピングに対する不変量（Warping Invariant Feature; WIF）を扱っているため，話者性の一部のみに不変なものを扱っていることになる．我々の提案手法である LAIF は，話者性の表現としてより一般的なアフィン変換に対する不変性を保証するため，WIF より LAIF の方が話者の違いへの頑健性が高いと考えられる．

LAIF は，我々が提案してきたアフィン変換不変性を持つ音響不変構造（Invariant Structure Representation; ISR）を用いた音声認識 [10]〜[12] とも異なる．ISR は，音声を動的にセグメントし，時間的に離れているものも含めすべての音声イベント間距離を特徴量として用いている．LAIF は，音声を幅を固定してセグメントし，さらに時間的に隣接した音声イベント間の距離のみに注目することにより，各時間フレームごとに一つの特徴量を抽出する．これにより，HMM をはじめとする既存の音声認識手法で用いる特徴量として，LAIF を導入することが可能になる．そのため，ISR では実現が困難であった大語彙連続音声を実現することができる．

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

この節では，アフィン変換に不変な局所特徴量 LAIF について述べる．まず，ケプストラムの時系列データから LAIF を抽出する計算方法の説明と，そのアフィン変換不変性の証明を行う．

$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T]$ 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．このようなケプストラムに対するアフィン変換は，話者の違いや録音機器の違いなどを近似することが知られている．

LAIF は，$\boldsymbol{X}$ のある時間フレーム t 付近における部分ベクトル列 $\boldsymbol{X}_{t-k_1:t+k_2}=[\boldsymbol{x}_{t-k_1},\boldsymbol{x}_{t-k_1+1},\cdots,\boldsymbol{x}_t,\cdots,\boldsymbol{x}_{t+k_2}]$ 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

$$F(\boldsymbol{X}_{t-k_1:t+k_2})=F(\boldsymbol{X}'_{t-k_1:t+k_2}) \quad t=1,\cdots,T \tag{2}$$

が成り立つような F のことである．アフィン変換は話者の違いや録音機器の違いを近似しているため，このような F は話者の違いや録音機器の違いに不変になる．

$F(\boldsymbol{X}_{t-k_1:t+k_2})$ 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データが用いられる．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．ここでデルタ特徴量 $\boldsymbol{\Delta}(\boldsymbol{X}_{t-k_1:t+k_2})$ とは，$k=k_1=k_2$ として以下の式で抽出される特徴量である．

$$\boldsymbol{\Delta}(\boldsymbol{X}_{t-k:t+k})=\frac{\sum_{\tau=1}^{k}\tau\left(\boldsymbol{x}_{t+\tau}-\boldsymbol{x}_{t-\tau}\right)}{2\sum_{\tau=1}^{k}\tau^2} \tag{3}$$

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．しかし，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 である．

$$F(\boldsymbol{X}_{t-k_1:t+k_2})=\sqrt{(\boldsymbol{\mu}_b-\boldsymbol{\mu}_a)^T(\boldsymbol{\Sigma}_a+\boldsymbol{\Sigma}_b)^{-1}(\boldsymbol{\mu}_b-\boldsymbol{\mu}_a)} \tag{4}$$

ただし，μ は平均ベクトルを，Σ は分散共分散行列を表す．添字 a はフレーム t より前の部分ベクトル列 $[t-k_1,\cdots,t-1]$ を，添字 b はフレーム t 以後の部分ベクトル列 $[t,\cdots,t+k_2]$ を表す．すなわち，μ_a および Σ_a は以下のように推定できる．

$$\mu_a = \frac{1}{k_1}\sum_{\tau=t-k_1}^{t-1} x_\tau \tag{5}$$

$$\Sigma_a = \frac{1}{k_1}\sum_{\tau=t-k_1}^{t-1} (x_\tau-\mu_a)(x_\tau-\mu_a)^T \tag{6}$$

この式は平均や共分散行列の ML 推定値を与える．μ_b および Σ_b も同様に推定できる．

LAIF のアフィン変換不変性を証明する．平均ベクトル μ_a や分散共分散行列 Σ_a を用いて，アフィン変換後の平均ベクトル μ'_a と分散共分散行列 Σ'_a を表すと，

$$\mu_a = A\mu_a + c \tag{7}$$

$$\Sigma'_a = A\Sigma_a A^T \tag{8}$$

となることから，

$$\begin{aligned} &F(X'_{t-k_1:t+k_2}) \\ &= \sqrt{(\mu'_b-\mu'_a)^T(\Sigma'_a+\Sigma'_b)^{-1}(\mu'_b-\mu'_a)} \\ &= \sqrt{(A\mu_b-A\mu_a)^T(A(\Sigma_a+\Sigma_b)A^T)^{-1}(A\mu_b-A\mu_a)} \\ &= \sqrt{(\mu_b-\mu_a)^TA^T(A^T)^{-1}(\Sigma_a+\Sigma_b)^{-1}A^{-1}A(\mu_b-\mu_a)} \\ &= F(X_{t-k_1:t+k_2}). \quad \square \end{aligned} \tag{9}$$

以上により式 (4) が LAIF であることが証明された．

式 (4) 以外にも，LAIF は数多く存在する [6]．例えば，$X_{i-k_1:i+k_2}$ の各ケプストラムベクトルに一定の重みをかける，すなわち $WX_{i-k_1:i+k_2}$ (W は対角行列) を使って式 (4) を計算したものも LAIF である．重みをつけた場合でも，式 (7,8) が成立するため，アフィン不変性の証明は式 (9) とまったく同様である．ただし，本稿では，特に断りのない限り式 (4) を LAIF として使用することにする．

時系列データとして $F(X_{t-k_1:t+k_2})$ を抽出するためには，t を一つずつずらしながら式 (4) を計算していけばよい．ここで，$t=n$ の場合，F は，$X_{n-k_1:n+k_2}$ 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，$X_{m-k_1:m+k_2}$ に対するアフィン変換に不変となる．ここで，$X_{n-k_1:n+k_2}$ に対するアフィン変換と，$X_{m-k_1:m+k_2}$ 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と考えられるため，これはより現実に則した性質であるといえる．

2.1 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が，同時に，言語情報の一部も表現してしまう．そのため，LAIF は話者の違いに頑健であると同時に，単語弁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 [11]．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 x に対する話者変換を表すアフィン変換 $Ax+c$ の，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，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 [13], [14]．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．[14] では，帯行列の幅は 2 で近似できるとしている．

このような帯行列 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 [11]．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 x を 2 つの部分ベクトル $x^{(1)}, x^{(2)}$ 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．

$$x'^{(1)} = A^{(1)}x^{(1)} + c^{(1)} \tag{10}$$

$$x'^{(2)} = A^{(2)}x^{(2)} + c^{(2)} \tag{11}$$

両式をまとめると下記のようになる．

$$\begin{pmatrix} x'^{(1)} \\ x'^{(2)} \end{pmatrix} = \begin{pmatrix} A^{(1)} & \mathbf{0} \\ \mathbf{0} & A^{(2)} \end{pmatrix} \begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix} + \begin{pmatrix} c^{(1)} \\ c^{(2)} \end{pmatrix} \tag{12}$$

このように特徴量を分割することで，行列 A の右上・左下要素を $\mathbf{0}$ にすることが可能となる．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．

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

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，A を帯行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，図 1 にまとめる．特徴量マルチストリーム化により，$F(X_{t-k_1:t+k_2})$ は $d-s+1$ 次元のベクトル $[F^{(1)}(X_{i-k_1:i+k_2}),\cdots,F^{(d-s+1)}(X_{i-k_1:i+k_2}))]^T$ となる．

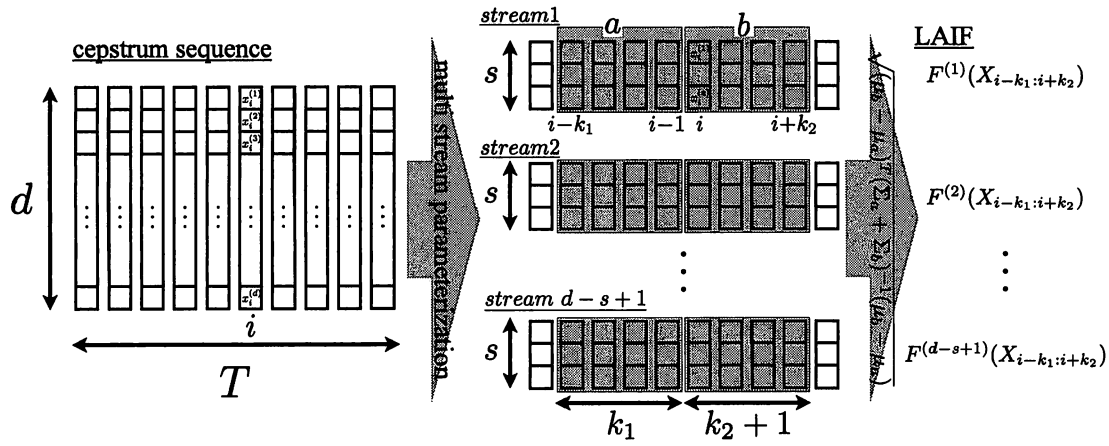


図 1　特徴量次元分割を導入した LAIF の抽出

Fig. 1　Calculation of LAIFs with multi stream parameterization

2.2 LAIF と他の特徴量との関係

マルチストリーム化において，ブロックサイズ $s=1$ の場合，LAIF は $\boldsymbol{A}$ を対角行列とした場合の $\boldsymbol{Ax}+\boldsymbol{c}$ のみに不変になる．このとき，一つのストリームから式 (4) で抽出する LAIF は以下のように計算することができる．

$$f_t(X_{i-k_1:i+k_2}) = \sqrt{(\mu_a-\mu_b)^T(\sigma_a^2+\sigma_b^2)^{-1}(\mu_a-\mu_b)} = \frac{|\mu_a-\mu_b|}{\sqrt{\sigma_a^2+\sigma_b^2}} \quad (13)$$

ここで，ブロックサイズ s は 1 であるので，μ や σ はスカラーである．

さらに，$|\mu_a-\mu_b|$ は，$k=k_1=k_2$ とするとさらに以下のように書き下せる．

$$|\mu_a-\mu_b| = \left|\sum_{\tau=1}^{k} w_\tau (x_{i+\tau-1}-x_{i-\tau})\right| \quad (14)$$

ただし w_τ は τ によらず $1/k$ である．ここで，w_τ の値を変えることは，各ケプストラムベクトルに重みをかけることに相当する．そのため，w_τ に任意の定数を与えても先に示したように式 (4) は LAIF となる．そこで，$w_\tau=\tau/(2\sum_{\tau=1}^{k}\tau^2)$ とおく．すると，式 (14) はデルタ特徴量の計算式 (3) とインデックスが一つずれていることと絶対値をとっていること除き同じものになる．

これをふまえ式 (13) に戻る．すると，ブロックサイズ $s=1$ のときの LAIF は，おおまかにいって，分散を正規化して絶対値をとったデルタ特徴量に等しいことがわかる．

次に，ブロックサイズ s が 2 以上の場合を考える．今回我々は，ケプストラムから LAIF を抽出しているが，ケプストラムとスペクトルは直交変換の関係にあるので，スペクトル領域でも同様に LAIF を抽出することができる [12]．そのため，LAIF は一種の時間-周波数領域特徴量 (Spectro-Temporal Features; STF) ということができる．STF は，近年盛んに研究が行われている．例えば Muroi らは，パワースペクトルにある時間幅と帯域幅をもった 2 次元の窓をかけ，得られたパターンを識別学習して用いることにより話者性に頑健な音声認識を実現している [15]．LAIF も，ある時間幅と帯域幅をもった 2 次元の窓の出力から特徴量を計算している点は Muroi らの手法と同じである．しかし，そこにアフィン変換不変性という数学的基盤を持っていることが，大きく異なる点となっている．話者性を表すアフィン変換 $\boldsymbol{Ax}+\boldsymbol{c}$ の $\boldsymbol{A}$ は，幾何学的には回転性が強いことが知られており [16]，回転に影響を受けない LAIF を用いることは理にかなっている．

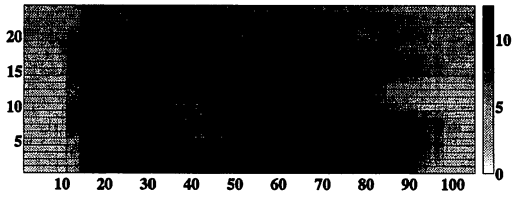
3. 実　　験

この節では，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．

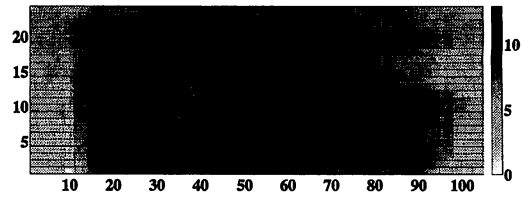
LAIF を抽出する際の窓の幅を決めるパラメータである k_1, k_2 としては，サンプリング周波数が 16kHz の場合 $k_1=k_2+1=16$ を用いる．k_1+1 や k_2 が 16 のときに最も成績がよいというのは，音声に含まれる変調周波数成分のうち，4Hz 付近に最も言語情報が含まれているという Kanedera らの結果におおよそ対応がとれる [17]．特徴量マルチストリーム化におけるブロックサイズ s は，1 または 2 を用いる．

このように抽出した LAIF と，メルフィルタバンク出力や MFCC と比較した図を図 2 に示す．図 2 では，成人男声が発声した/aiueo/という音声と，それを straight [18] を用いて声道長を約 0.7 倍にするような周波数ウォーピングをかけた音声それぞれから各特徴量を抽出している．また LAIF は，ブロックサイズ $s=2$ の場合のものと，次元分割を行わない $s=12$ 場合の 2 種類をのせている．

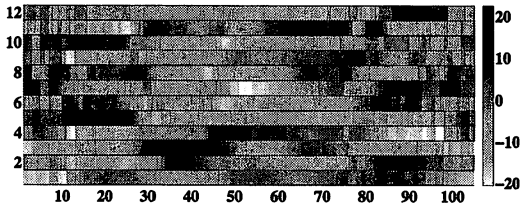
音声認識実験においては，LAIF は MFCC のようなスペクトル特徴量とは異なる音声の特徴を捉えていると考えられるため，MFCC と LAIF の結合ベクトルを用いることにする．また，ΔMFCC の抽出における窓の幅 k は，Hidden-Markov-Toolkit(HTK) [1] のデフォルト値である 2 を用いているため，



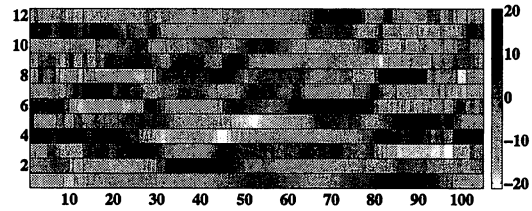
(a–1) メルフィルタバンク出力，オリジナル音声/aiueo/



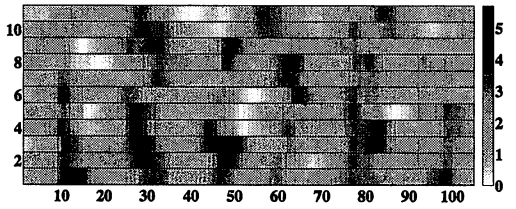
(a–2) メルフィルタバンク出力，ウォーピング後音声/aiueo/



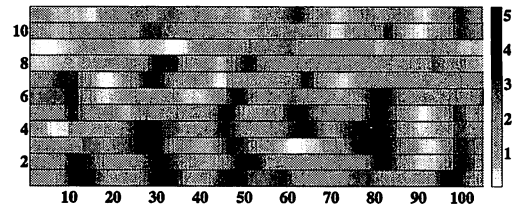
(b–1) MFCC，オリジナル音声/aiueo/



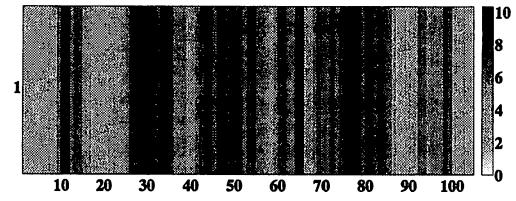
(b–2) MFCC，ウォーピング後音声/aiueo/



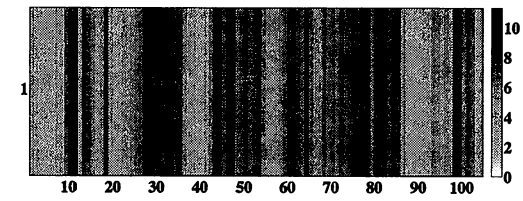
(c–1) $\mathrm{LAIF}_{s=2}$，オリジナル音声/aiueo/



(c–2) $\mathrm{LAIF}_{s=2}$，ウォーピング後音声/aiueo/



(d–1) $\mathrm{LAIF}_{s=12}$（分割なし），オリジナル音声/aiueo/



(d–2) $\mathrm{LAIF}_{s=12}$（分割なし），ウォーピング後音声/aiueo/

図 2　メルフィルタバンク出力と MFCC と LAIF の比較

Fig. 2　Comparison between outputs of mel filter bank, MFCCs and LAIFs

$\mathrm{LAIF}_{s=1}$ と ΔMFCC も，それぞれ異なる音声の特徴を捉えているものと考えられる．具体的には，従来のスペクトル特徴量のみのものも含めて表 1 に示した 6 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 [19]．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．

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，LAIF 及びデルタ特徴量を抽出した．

表 1　使用する特徴量

Table 1　Acoustic features used for the experiment

Features(# of dimension)
MFCC(12)
MFCC(12) + $\mathrm{LAIF}_{s=1}$(12)
MFCC(12) + $\mathrm{LAIF}_{s=2}$(11)
MFCC(12) + ΔMFCC(12)
MFCC(12) + ΔMFCC(12) + $\mathrm{LAIF}_{s=1}$(12)
MFCC(12) + ΔMFCC(12) + $\mathrm{LAIF}_{s=2}$(11)

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，出力分布としては対角共分散の単一正規分布を用いた．通常の不

表 2 認識実験の結果

Table 2 Recognition result. M, Δ, L denote MFCC, delta cofficients of MFCC, and LAIF. s means block size for multi stream parameterization.

Method	M	M+$L_{s=1}$	M+$L_{s=2}$	M+Δ	M+Δ+$L_{s=1}$	M+Δ+$L_{s=2}$
Matched condition	98.35%	99.24%	98.88%	99.47%	99.51%	99.39%
Male training – Female testing	72.71%	83.22%	83.83%	82.79%	88.35%	89.27%
Female training – Male testing	70.59%	83.25%	83.21%	85.34%	89.88%	90.70%

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，学習話者を女声 30 名，評価話者を男声 30 名とする実験も行なった．

実験結果を表 2 に示す．結果，MFCC や 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 に対し，$LAIF_{s=2}$ を結合することによるエラー削減率は 41%，MFCC+ΔMFCC に対し $LAIF_{s=2}$ を結合することによるエラー削減率は 37%となった．また，ブロックサイズ s が 1 のときと 2 のときを比べると，ミスマッチがない条件では単語弁別能力がより高い $s=1$ の方が認識率が良く，ミスマッチがある条件では話者不変性がより高い $s=2$ の方がよいという結果になっている．

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．LAIF は，MFCC などといった既存のスペクトル特徴量に対し特徴量マルチストリーム化と簡単な計算を行うことにより容易に抽出できるものである．ケプストラムベクトルへ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，SI 音声認識に効果があると考えられる．LAIF を使って SI 音声認識実験を行った結果，LAIF+MFCC+ΔMFCC を用いることで，ミスマッチ条件下において 37%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，SI 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．

文 献

[1] S. Young *et.al.*, "The HTK Book (for HTK Version 3.4)"

[2] 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.

[3] 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.

[4] 篠田浩一, "確率モデルによる音声認識のための話者適応化技術," 電子情報通信学会論文誌, vol.J87–D–II, no.2, pp.371–386, 2004.

[5] R. Gomez, T. Toda, H. Saruwatari, K. Shikano, "Techniques in rapid unsupervised speaker adaptation based on HMM-sufficient statistics," *Speech Communication*, vol. 51, pp. 42–57 2009.

[6] Y. Qiao, M. Suzuki, N. Minematsu, "Affine invariant features and its application to speech recognition," *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 2009 (submitted).

[7] A. Mertins and J. Rademacher, "Frequency-warping invariant features for automatic speech recognition," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol.5, pp. 1025–1028, 2006.

[8] J. Rademacher, M. Wachter and A. Mertins, "Improved warping-invariant features for automatic speech recognition," *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 1499–1502 2006.

[9] T. Irino and R. D. Patterson, "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.

[10] N. Minematsu, "Mathmatical evidence of the acoustic universal structure in speech," *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 889–892 2005.

[11] S. Asakawa, N. Minematsu, K. Hirose, "Multi-stream parameterization for structural speech recognition," *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 4097–4100, 2008.

[12] 鈴木雅之, 朝川智, 喬宇, 峯松信明, 広瀬啓吉 "スペクトル特徴量を用いた音声の構造的表象に関する実験的検討," 電子情報通信学会技術報告, SP2008–32, pp.73–78, 2008.

[13] M. Pitz and H.Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech and Audio Processing*, vol.13, pp.930–944, 2005.

[14] 江森正, 篠田浩一, "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.

[15] T. Muroi, T. Takiguchi, Y. Ariki "Speaker Independent Phoneme Recognition Based on Fisher Weight Map," *International Journal of Hybrid Information Technology*, Vol. 1, No. 3, 2009.

[16] 齋藤大輔, 松浦良, 朝川智, 峯松信明, 広瀬啓吉, "ケプストラムの声道長依存性に関する幾何学的考察," 電子情報通信学会音声研究会, SP2007–128, pp.189–194, 2007.

[17] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communication*, vol. 28, no. 1, pp. 43—55, 1999.

[18] H. Kawahara, "STRAIGHT, Exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustic Science and Technology*, Vol. 27, No. 6 2006.

[19] 牧野正三, 二矢田勝行, 真船裕雄, 城戸健一, "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.