社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

識別的アプローチによる分布間距離推定の検討とその言語識別への応用

柏木　陽佑†　張　聡穎†　齋藤　大輔††　峯松　信明†

† 東京大学大学院工学系研究科，東京都
†† 東京大学大学院情報理工学系研究科，東京都
E-mail: †{kashiwagi,zhangcy,dsk_saito,mine}@gavo.t.u-tokyo.ac.jp

あらまし　本稿では，識別的なアプローチによる分布間距離の推定法について提案し，言語識別タスクへの応用を検討する．分布間距離はパターン認識において広く用いられている．分布間距離は確率密度関数に対する汎関数として定義されるため，尤度計算の関係上特徴量の分布形状を仮定する必要があった．しかし，仮定した分布形状と真の分布とのミスマッチがある場合，イベント間の正しい分布間距離とはズレが生じる．提案手法では，分布間距離をベイズの定理を用いることで，識別モデルを利用して推定する．この際，ニューラルネットワークに代表される，分布の形状を仮定しないモデルを識別モデルとして導入することにより，明示的に分布形状を仮定せずに分布間距離を推定することが可能となる．また，分布の形状を仮定しない性質上，特にデータ量が少ない場合，ニューラルネットワークの学習と特徴量空間全体における積分が困難となる．そこで，ニューラルネットワークのモデル適応手法とサンプリングによって観測データの偏りに起因する影響を取り除いた．観測データ量が少ないことが想定される言語識別タスクへの応用を試み，実験により提案法の有効性を確認した．この推定法を利用して得られた分布間距離を構造的特徴として用いることにより，従来の I-vector を用いた手法に対して認識率の改善が確認できた．
キーワード　ニューラルネットワーク，分布間距離，言語識別，音響モデル

A study on discriminative approach for estimation of the divergence between distributions and its application to language identification

Yosuke KASHIWAGI†, Congying ZHANG†, Daisuke SAITO††, and Nobuaki MINEMATSU†

† Graduate School of Engineering, The University of Tokyo, 7-3-1, Hongo, Bukyo-ku, Tokyo 113-0033, Japan
†† Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1, Hongo, Bukyo-ku, Tokyo 113-0033, Japan
E-mail: †{kashiwagi,zhangcy,dsk_saito,mine}@gavo.t.u-tokyo.ac.jp

Abstract　In this paper, we propose a method for estimating the statistical divergence between probability distributions by a discriminative approach and the use of language identification tasks. Statistical divergence is widely used in pattern recognition. For statistical divergence between which is defined as functional for the probability density function, it was claimed to be assume the shape of the distribution because of the likelihood calculation. However, if there is mismatch between the assumed distribution shape and the true distribution, the divergence estimation between events would be not correct. In our proposed method, by using Bayes' theorem distribution divergence, we estimate the statistical divergence with the discriminative model. Such as neural network models, which is not assume the shape of the distribution, discriminative model is possible to estimate the statistical divergence without assuming explicit distribution shape. On the other hand, when the amount of data is small, it is difficult to integrate the function in the feature space and train neural networks. To mitigate this problem, we use model adaptation method for neural networks and sampling approach to integrate the function. We use this approach to language identification task using statistical distribution as speech structure features. The experimental evaluation results shows our approach can improve the performance than conventional I-vector based approach.
Key words　neural network，divergence between distributions，language identification, acoustic model

1. はじめに

Kullback-Leibler 距離 (KL 距離) に代表される分布間距離は，情報量とも呼ばれ，パターン認識において重要な位置を占める．例えば，ガウス混合モデル (Gaussian Mixture Model ; GMM) に代表される隠れ変数を持つモデルの学習手法の一つである EM アルゴリズムは KL 距離を最小化する基準でパラメータのアップデートを行う．近年，音声認識の音響モデルにおいても主流となっているニューラルネットワークにおいても，KL 距離の変化を抑える基準を入れたモデル適応手法が効果的であるとの報告がなされている [1]．また，モデルの学習基準のみならず，特徴量として分布間距離を利用する構造的表象も提案されている [2]．これは，特徴量の分布をガウス分布として仮定した場合の分布間距離はアフィン変換不変であるため，ケプストラム空間におけるアフィン変換で表現することが可能である話者性の違いに対して不変である性質を利用したものである．

このように，多岐に渡って利用されている分布間距離であるが，これは観測特徴量の分布形状と密接な関係を持つ．一般に，分布間距離は確率密度関数に対する汎関数として定義されるため，分布間距離の推定は観測特徴量の分布の形状に対する仮定が必要となる．この分布に対する仮定が真の分布と乖離している場合，分布間距離の推定を誤る原因となるが，一般にはガウス分布等の妥当な仮定を置いて計算することが多い．

しかし，観測特徴量の真の分布は複雑な形状をしていると考えられる．そのため，単純なガウス分布ではなく，複雑な形状を仮定することで分布間距離の推定精度が向上することが考えられる．例えば，分布形状を GMM で仮定し，モデル適応を利用して分布間距離を推定する手法も提案されている [3]．しかし，変換前と後で対応するインデクスの分布形状はあくまで正規分布を仮定するため，変換によって分布形状が変化する場合妥当な近似とは言えない．

一方，音声認識において近年，多層ニューラルネットワーク (Deep Neural Network ; DNN) に代表される識別モデルが，従来の GMM をベースとした生成モデルに対して高い精度を示すことが示されている [4,5]．これは，DNN に代表される識別モデルは GMM などの生成モデルと比較して特徴量の分布に対する仮定が弱いため，柔軟なモデル化が可能であることが理由の一つと考えられる．

この識別モデルを利用した分布間距離の推定法として，Heigold らの対数線形モデルを用いた生成モデルを介する手法がある [6]．これは，対数線形モデルの性質を利用し，識別モデルのパラメータからガウス分布のパラメータを推定することで，分布間距離を計算する．しかし，対数線形モデルのパラメータからでは，ガウス分布の平均と分散パラメータが一意に決定することができないため，分散パラメータを各ラベルに対して共有する等の仮定を置く必要がある．これを用いて DNN の最終層として用いられる softmax 関数から分布間距離を計算することで効率的に DNN を学習する手法も Li らにより提案されている [7].

しかし，識別モデルの分布を明示的に仮定する必要のない性質を利用して分布間距離を推定する場合，生成モデルのパラメータを経由せずに直接推定することが理想的である．そこで，本稿では，分布間距離を識別モデルにより推定する手法を提案する．従来，生成モデルベースにより推定していた尤度をベイズ則を用いることによって識別モデルから推定する．

なお，GMM を仮定して分布間距離を推定した場合と同様に，特に識別モデルとして DNN を採用した場合，モデルの学習には大量のデータが必要となる．そこで，言語識別タスクを例に，観測特徴量を用いた DNN の発話適応を併用した分布間距離の推定法を提案する．これにより，少量のデータにおいても頑健にモデルの構築が可能となる．

本稿の構成は，第 2 節で，関連研究について述べ，第 3 節で，各分布間距離を識別モデルを利用して推定する手法について述べる．その後，第 4 節でこれらを観測データ量が少量であることが予想される言語識別タスクへ応用する手法について述べ，第 5 節で言語識別実験によりその性能を評価する．最後に第 6 節でまとめる．

2. 関連研究

本節では，GMM ベースの分布間距離推定法である Chang らの手法と，識別モデルを用いた分布間距離の推定法である Heigold らの手法を用いて DNN から分布間距離を推定する Li らの手法について紹介する．

2.1 特徴量分布に GMM を仮定した分布間距離推定

Chang らは，ガウス分布ではなく，GMM を仮定した場合の分布間距離推定法を提案している [3]．Bhattacharyya 距離は次式によって定義される．

$$BD(a,b) = -\ln \int \sqrt{p(\boldsymbol{x}|y=a)p(\boldsymbol{x}|y=b)}d\boldsymbol{x} \qquad (1)$$

ここで，a, b は音響イベントのラベルである．これは，音素や，音素状態ラベル等の特徴量との対応が取れるものであれば用いることができる．各ラベルに対応する特徴量の分布を正規分布 $p(\boldsymbol{x}|y) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_y, \Sigma_y)$ と仮定した場合，Bhattacharyya 距離は

$$\begin{aligned} BD(a,b) &= \frac{1}{8}(\boldsymbol{\mu}^{(a)} - \boldsymbol{\mu}^{(b)})^\top \Sigma^{-1}(\boldsymbol{\mu}^{(a)} - \boldsymbol{\mu}^{(b)}) \\ &\quad + \frac{1}{2}\ln\left(\frac{\det\Sigma}{\sqrt{\det\Sigma^{(a)}\det\Sigma^{(b)}}}\right) \qquad (2) \\ \Sigma &= \frac{\Sigma^{(a)} + \Sigma^{(b)}}{2} \end{aligned}$$

として正規分布のパラメータから推定することができる．従来はこれを用いることで観測特徴量により正規分布のパラメータを推定することを経由し，Bhattacharyya 距離を推定することが一般的であった．

Li らは特徴量分布を GMM に置き換えた場合，

$$\begin{aligned} BD(a,b) &= \sum_j^J BD(a_j, b_j) \\ &= -\sum_j^J \ln \int \sqrt{w_j^{(a)} p(\boldsymbol{x}|y=a)}\sqrt{w_j^{(b)} p(\boldsymbol{x}|y=b)}d\boldsymbol{x} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{8}\sum_{j}^{J}(\boldsymbol{\mu}_j^a - \boldsymbol{\mu}_j^b)^\top \Sigma^{-1}(\boldsymbol{\mu}_j^{(a)} - \boldsymbol{\mu}_j^{(b)}) \\
&+\frac{1}{2}\sum_{j}^{J}\ln\left(\frac{\det\Sigma}{\sqrt{\det\Sigma_j^{(a)}\det\Sigma_j^{(b)}}}\right) \\
&-\frac{1}{2}\sum_{j}^{J}\ln(w_j^{(a)}w_j^{(b)}) \\
\Sigma &= \frac{\Sigma_j^{(a)} + \Sigma_j^{(b)}}{2}
\end{aligned} \tag{3}
$$

として計算している．各 GMM はあらかじめ学習された Universal-Background Model (UBM) を適応したモデルを初期値として GMM を学習している．そのため，GMM 同士のインデクスの対応が取れており，インデクスの対応する正規分布間の Bhattacharyya 距離が重要であるという仮定に基づく．

2.2 出力分布基準に基づく分布間距離推定

識別モデルを用いた分布間距離の計算として，Heigold らの手法がある．これは，対数線形モデルの性質を利用して正規分布のパラメータを推定することで，分布間距離を推定することができる．対数線形モデルは，入力を $\boldsymbol{x}$ とすると

$$
p(y|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})}\exp(\sum_i \lambda_{yi} f_i(\boldsymbol{x})) \tag{4}
$$

として出力を y を推定することができる．ここで，i は次数である．ニューラルネットワークなどで用いられる softmax 関数は，

$$
\begin{aligned}
f_1(\boldsymbol{x}) &= [\boldsymbol{x}^\top, 1]^\top \\
\lambda_{y1} &= [\boldsymbol{W}_y, b_y] \\
\lambda_{yi} &= 0 \quad (i \geqq 2)
\end{aligned} \tag{5}
$$

とした場合と等価となる．各ラベルに対応する分布が正規分布であると仮定した場合，この対数線形モデルのパラメータから正規分布のパラメータへ対応づけを行うことができる．クラスに依存しない半正定値行列 Σ を与えた場合，各正規分布のパラメータは

$$
\Sigma_y = \Sigma \tag{6}
$$

$$
\boldsymbol{\mu} = \Sigma^{-1}[\boldsymbol{W}_y, b_y] \tag{7}
$$

となる．

これを用いて Li らは，DNN を学習する際に，大きなネットワークにより推定される正規分布と学習する DNN との KL 距離が，変化しない基準を導入して学習を行った．これによって認識性能の向上が実現されている．しかし，このアプローチは，分散パラメータの共有化は音声認識において効果的な仮定ではあるものの，分散と平均との間に相互依存が存在し，パラメータを一意に決定することができない．また，正規分布を仮定するため，実際の分布の形が正規分布と大きく異なっている場合，分布間距離の推定精度が大きく低下すると考えられる．

3. 識別的アプローチによる分布間距離の推定

分布間距離の計算は，分布の形を仮定する必要があるが，近年の DNN に代表される識別モデルの台頭を鑑みた場合，分布の形状は複雑なものであると想定される．そのため，分布の形状を明示的に仮定せずに分布間距離を推定することが望まれる．本節では，構造的表象などに用いられる BD 距離を従来の生成モデルをベースとしたものではなく，識別モデルを用いて推定する手法を提案する．

さて，Bhattacharyya 距離を明示的に分布の形状を仮定せず，識別モデルを用いて推定したい．ここで，我々は識別モデルにより事後確率 $p(y=a|\boldsymbol{x}), p(y=b|\boldsymbol{x})$ を直接計算することが可能である．そこで，式 (1) をベイズ則を用いることにより，

$$
\begin{aligned}
BD(a,b) &= -\ln\int p(\boldsymbol{x})\sqrt{p(y=a|\boldsymbol{x})p(y=b|\boldsymbol{x})}d\boldsymbol{x} \\
&+\frac{1}{2}\ln p(y=a) \\
&+\frac{1}{2}\ln p(y=b)
\end{aligned} \tag{8}
$$

とすることで，事後確率を用いて計算することが可能となる．さらに，事後確率を DNN などの分布の形状を明示的に仮定しない識別モデルにより計算する．これにより陽にそれぞれ音響イベントにおける特徴量の分布の形を決定することなく，Bhattacharyya 距離の計算が可能となる．

データ集合を $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L\}, \boldsymbol{Y} = \{y_1, y_2, \ldots, y_L\}$ とすると，

$$
\begin{aligned}
BD(a,b) &= -\ln\frac{1}{L}\sum_l \sqrt{p(y_l=a|\boldsymbol{x}_l,\theta)p(y_l=b|\boldsymbol{x}_l,\theta)} \\
&+\frac{1}{2}\ln\frac{1}{L}\sum_l p(y_l=a) \\
&+\frac{1}{2}\ln\frac{1}{L}\sum_l p(y_l=b)
\end{aligned} \tag{9}
$$

として計算することができる．なお，ここで θ は識別モデルのパラメータであり，例えば，以下のクロスエントロピー基準により学習することができる．

$$
\hat{\theta} = \underset{\theta}{\operatorname{argmin}} -\sum_l p(\hat{y}_l|\boldsymbol{x}_l,\theta)\ln q(y_l|\boldsymbol{x}_l,\theta) \tag{10}
$$

4. 言語識別への応用

識別的アプローチによって推定された Bhattacharyya 距離の言語識別への応用を考えた場合，数発話のような少量のデータから推定する必要が往々にして生じる．しかし，データが少量の場合，その発話に対応する DNN を学習することは困難である．識別モデルの性質上，特徴量空間をそれなりの精度で埋める必要がある．そのため，本来有意に発生するであろうラベルが，その少量のデータ中に発生しなかった場合，式 (9) による Bhattacharyya 距離の計算は大きく誤ることが予想される．これは音響イベントを音素ラベル等で考えた場合，発話内容によっては発声されない音素があることは容易に想像がつく．また，仮に DNN が正しく学習された場合においても，観測特徴量は発話内容によって偏りが生じるため，事後分布の全積分自体が誤ると想定される．そこで，これらの問題を回避するために，DNN の適応とサンプリング法を導入する．あらかじめ大量のデータによりグローバルな DNN を学習しておき，これを入

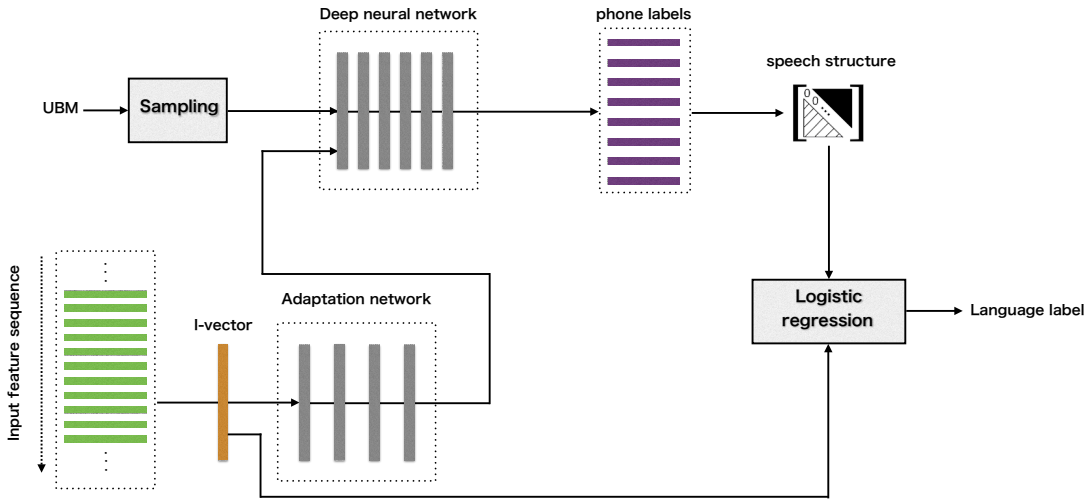

図 1　Proposed language identification system.

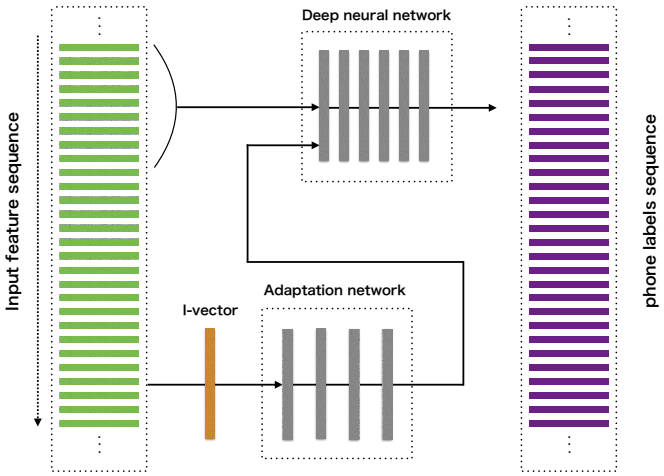

図 2　SAT-DNN.

力発話によって適応することで少量のデータでも頑健に DNN パラメータの推定を行うことができる．また，サンプリングにより発話内容に依存せずに特徴量空間全体からデータ集合を生成することが可能となる．

言語識別システムの概要を図に示す．前提条件として多言語のデータには音素ラベルが付与されておらず，音素ラベルが付与されているデータは一部の言語のみである．まず，複数言語のデータを用いて学習された I-vector モデルにより，観測発話に対する I-vector を計算する．その後，I-vector を用いることで，観測特徴量から音素を識別するニューラルネットワークの発話適応を行う．なお，適応に用いるベースとなるニューラルネットワークは，音素ラベルが存在する 1 つの言語のデータのみを用いて学習したものを用いる．その後，適応されたニューラルネットワークを用いてサンプリングにより音素間の分布間距離を推定し，構造的表象を構築する．これを特徴量として I-vector と連結しロジスティック回帰により言語識別を行う．

4.1　DNN の適応

DNN の発話適応は，I-vector 自体が言語識別に用いられる特徴量であるため，Miao らの I-vector を用いた適応法である SAT-DNN を利用する [8]．この手法は，通常の DNN の入力に I-vector からのニューラルネット出力を足し合わせることによりモデル適応を行う．SAT-DNN のモデル構造を図 2 に示す．SAT-DNN は MFCC などの特徴量 $\boldsymbol{o}_t$ を入力として，ラベル $\boldsymbol{y}_t$ を

$$p(\boldsymbol{y}_t) = g(\boldsymbol{a}_t) \tag{11}$$

$$\boldsymbol{a}_t = \boldsymbol{o}_t + f(\boldsymbol{i}_s) \tag{12}$$

として識別する．ただし，$\boldsymbol{i}_s$ は入力話者を表現する I-vector である．また，$f(\cdot), g(\cdot)$ はそれぞれ，発話適応用の DNN，SAT モデルの DNN を表す．

学習の際は，まず，グローバルな DNN を学習する．その後，これを $g(\cdot)$ の暫定値とし，$g(\cdot)$ のパラメータを固定したまま back-propagation により $f(\cdot)$ を学習する．その後，逆に $f(\cdot)$ のパラメータを固定して，$g(\cdot)$ のパラメータを更新することで，SAT モデルが学習される．この枠組みにより学習することにより，認識時は，各入力発話に対応する I-vector を入力することで入力層の対してバイアス適応の形で発話適応が可能となる．

なお，当然ながら DNN の学習には音素ラベル，もしくは音素状態ラベルが必要となる．不特定多数の言語に対して共通の音素ラベルを用意することは困難であるため，DNN は音素ラベルの存在する 1 言語の音声のみによって学習を行う．

4.2　発話適応 DNN を用いた Bhattacharyya 距離の計算

式 (9) を発話適応により得られた DNN モデルパラメータ θ_c により計算することで，安定した事後確率の計算が可能となる．しかし，観測発話のデータ数が少ないという問題は依然として残る．そのため，データの偏りによって，特徴量空間全体に対する全積分を計算する必要のある分布間距離推定は計算することができない．．そこで，話者非依存な分布，もしくはそれを基に適応を行った分布を用いて，特徴量空間をサンプリングすることによりこれを回避する．式 (9) をベースとして，分布間距離を

$$BD(a,b) = \ln \frac{1}{N} \sum_l \sqrt{p(y_n = a|\boldsymbol{x}_n, \theta_c) p(y_n = b|\boldsymbol{x}_n, \theta_c)} + \frac{1}{2} \ln \frac{1}{L} \sum_l p(y_l = a) + \frac{1}{2} \ln \frac{1}{L} \sum_l p(y_l = b) \quad (13)$$

として推定する．ただし，$\{\boldsymbol{x}_n\}$ は DNN の学習に用いたデータと同じ言語のデータによって構築された UBM，もしくは UBM を基に発話適応を行った GMM より得られるサンプル集合である．

$$\boldsymbol{x}_n \sim \mathrm{GMM}(\phi_{global}) \quad (14)$$

or

$$\boldsymbol{x}_n \sim \mathrm{GMM}(\phi_c) \quad (15)$$

また，事前分布 $p(a), p(b)$ は，グローバルな DNN を学習した際の学習データ中の出現頻度により近似する．これは，話者により音素の発生頻度が大きく異なることはないという仮定に基づく．言語によって音素の発生頻度自体は変化すると考えらえるが，DNN は英語音声のみで学習しているため，アンカーモデルとして機能すると考えられるため，事前分布 $p(a), p(b)$ は全ての話者で英語の出現頻度を共通として用いた．なお，これは DNN を音響モデルとして利用する場合に用いられる近似と同じである．

$$p(a) = \frac{1}{L} \sum_l p(y_l = a) \quad (16)$$

$$p(b) = \frac{1}{L} \sum_l p(y_l = b)$$

ただし，サンプリングにより計算する都合上，周辺分布と事前分布との誤差により推定された Bhattacharyya 距離が負の値となる場合があるため，注意が必要となる．

4.3 構造的表象と言語識別

提案手法により推定した分布間距離によって構築した構造的表象と I-vector を連結したものを特徴量とし，ロジスティック回帰により言語を識別する．音声の構造的表象は，分布間距離を特徴量として利用する手法の一つである．性別や年齢などの話者性の変換は，ケプストラム空間上におけるアフィン変換と対応している．構造的表象は音響イベントの分布間距離を用いて定義されるケプストラム空間上の構造体である．各音響イベントの分布をガウス分布によりモデル化することで分布間距離はアフィン変換不変となる．そのため，構造的表象を特徴量として用いることで，話者性などに起因するバラツキに対して頑健なシステムを構築することが可能となる．

さて，提案手法により識別的に推定された分布間距離を用いて構造的表象を構築する．前述のように I-vector を用いることによって DNN のモデル適応を行うため，I-vector を計算できさえすれば観測発話中に全音素が出現する必要なく，入力発話から式 (13) により全音素間の Bhattacharyya 距離を計算することができる．従って，DNN により推定する音素数を K とすると，最終的に観測発話から得られる Bhattacharyya 距離は自身との距離を除けば，常に $K(K-1)/2$ 次元となり，それぞれの次元の持つ意味は必ず対応が取れている．ガウス分布により各音素の分布をモデル化した場合，構造的表象はアフィン変換に不変であり，話者性の変化に対して頑健である．これは特徴量の分布の形状が，変換による前後で共にガウス分布であるという仮定に基づく．しかし，DNN により特徴量分布の形を陽に決定せずに Bhattacharyya 距離を式 (13) を用いて計算した場合，DNN が精確に特徴量空間の対応を学習できたと仮定すると，あらゆる全単射である変換に対して不変となると考えられる．そのため，理想的には全ての発話に対して Bhattacharyya 距離は等しくなる．

しかし，言語によって音素が異なるため，DNN は音素ラベルの存在する言語のみにより学習する．この際，言語によって I-vector が学習に用いた言語のもつ I-vector と比較して大きく異なる場合，DNN が正しく適応できない．これによって適応後の DNN が想定する入力特徴量の分布が実際の分布から大きく崩れること想定される．この適応後のモデルを用いてサンプリングにより得た事後確率で構造的表象を構築した場合，本来観測特徴量に対して不変である構造が，逆に発話によっては異なる構造に変換される．つまり，学習に用いた言語をアンカーとした特徴量が生成されることが期待できる．

表 1　The number of utterances of each language in train/test set.

Language	Test set	Train set
Arabic	240	906
Bengali	240	200
Chinese	1194	3577
English	720	5739
Farsi	240	717
German	240	970
Hindustani	720	1135
Japanese	240	2050
Korean	240	1655
Russian	480	440
Spanish	720	3007
Tamil	480	1237
Thai	240	200
Vietnamese	480	707

5. 言語識別実験

LRE07 データベースをベースとして言語識別実験により提案法の有効性を評価した．I-vector の学習は，LRE03，LRE05，LRE07 の学習セットを用いて行った．各言語の学習データ，評価データ中の発話数を表 1 に示す．学習データ中の言語は英語が最も多いため，DNN の学習データは英語のデータである WSJ を用いて学習した．また，サンプリングに用いる UBM も WSJ を用いて学習した．評価は LRE07 のテストセットを用いており，識別対象は学習データ中の言語数と同じ 14 言語である．なお，言語クローズドセットであるため，評価セット中に未知の言語は存在しない．

DNN の入力特徴量は MFCC12 次元と C0 の計 13 次元を当

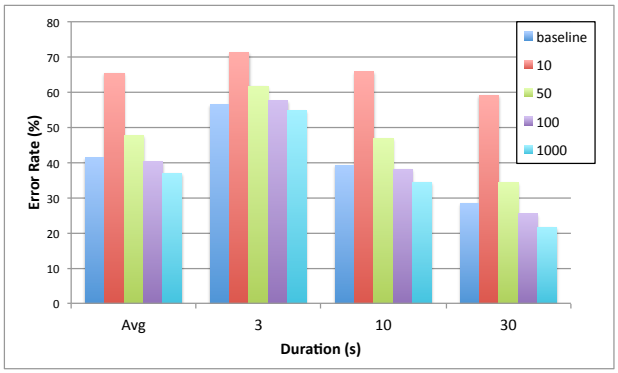

図 3　Experimental results (Error rate %)

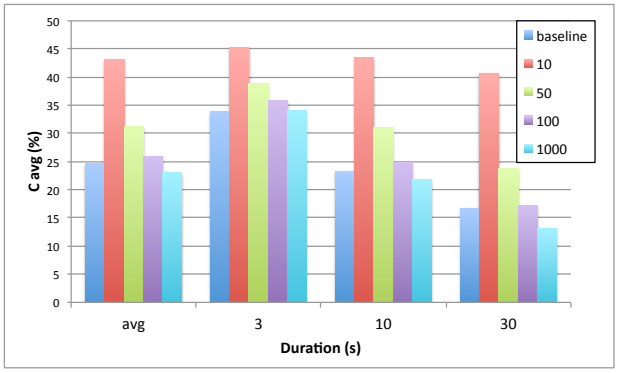

図 4　Experimental results (C_{avg} %)

該フレームとその前後 5 フレームの計 11 フレーム連結したものを用いた．DNN は各層の初期値を RBM により学習し，中間層の層数は 6 層，話者適応に用いるネットワークは中間層が 4 層に設定した．また各層のノード数は全て 1024 としている．サンプリングに用いる UBM の混合数は 1024 とし，サンプリングにより出力するデータは各発話毎に 10, 50, 100, 1000 サンプルの 4 通りを用意した．

ベースラインとサンプリング数毎の提案手法の識別実験結果を図 3，4 に示す．ベースラインは I-vector のみで識別したものであり，提案法はそれぞれのサンプリング数から計算された分布間距離により得られる構造的表象を I-vector と連結したものを用いたものである．なお，C_{avg} は LRE タスクで採用されている評価指標である．

$$C_{avg} = \frac{1}{N_L} \sum_{L_T} \left\{ \begin{array}{l} C_{Miss} P_{Target} P_{Miss}(L_T) \\ + \sum_{L_N} C_{FA} P_{Non-Target} \\ \quad \times P_{FA}(L_T, L_N) \\ + C_{FA} P_{Out-of-Set} P_{FA}(L_T, L_O) \end{array} \right\} \quad (17)$$

LRE07 の場合，$C_{Miss} = C_{FA} = 1.0$ であり，$P_{Target} = 0.5$ である．なお，言語クローズドセットを用いた評価であるため，$P_{Out-of-Set} = 0$ としている．また，$P_{Non-Target}$ は

$$P_{Non-Target} = (1 - P_{Target} - P_{Out-of-Set})/(N_L - 1) \quad (18)$$

として計算することができる．また，Duration は観測発話の秒数を表しており，秒数が短いほど難しいタスクとなる．

結果により，サンプリング数が少ない場合はベースラインと比較して認識誤りが増加してしまう．これは，サンプリング数が少ない場合，特徴量空間全体に対しサンプルの偏りが生じてしまうため，分布間距離の推定が誤るためと考えられる．しかし，サンプリング数が 100 程度あれば提案法によって認識誤りを低下することが可能となる．特に，サンプリング数が 1000 の場合，ベースラインのシステムと比較して 10.85%の認識誤り率の削減が可能となった．

6. ま と め

本稿では，識別モデルを利用した分布間距離推定法を提案した．また，実際の利用としてはデータ数が限られた場面での分布間距離の推定が必要となる．そのため，言語識別タスクを挙げ，分布間距離である Bhattacharyya 距離を少量のデータから識別的に推定する手法について述べた．言語識別実験により，ロジスティック回帰の入力特徴量として I-vector に Bhattacharyya 距離により得られた構造的表象を連結したもので認識を行ったところ，I-vector 単体と比較し認識率の向上が確認できた．

文　献

[1] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, “Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition,” Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference onIEEE, pp.7893–7897 2013.

[2] N. Minematsu, “Yet another acoustic representation of speech sounds,” Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP’04). IEEE International Conference on, vol.1IEEE, pp.I–585 2004.

[3] C.H. You, K.A. Lee, and H. Li, “Gmm-svm kernel with a bhattacharyya-based distance for speaker recognition,” Audio, Speech, and Language Processing, IEEE Transactions on, vol.18, no.6, pp.1300–1312, 2010.

[4] G.E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” Neural computation, vol.18, no.7, pp.1527–1554, 2006.

[5] A.-r. Mohamed, G.E. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” Audio, Speech, and Language Processing, IEEE Transactions on, vol.20, no.1, pp.14–22, 2012.

[6] G. Heigold, H. Ney, P. Lehnen, T. Gass, and R. Schlüter, “Equivalence of generative and log-linear models,” Audio, Speech, and Language Processing, IEEE Transactions on, vol.19, no.5, pp.1138–1148, 2011.

[7] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, “Learning small-size dnn with output-distribution-based criteria,” Proc. Interspeech, pp.1910–1914, 2014.

[8] Y. Miao, L. Jiang, H. Zhang, and F. Metze, “Improvements to speaker adaptive training of deep neural networks,” 2014.