雑音環境下音声認識のためのディープニューラルネットワークを用いた識別的区分線形変換*

柏木　陽佑[†a)]　齋藤　大輔[††]　峯松　信明[†]　広瀬　啓吉[††]

Discriminative Piecewise Linear Transformation Based on Deep Neural Networks for Noise Robust Automatic Speech Recognition*

Yosuke KASHIWAGI[†a)], Daisuke SAITO[††], Nobuaki MINEMATSU[†], and Keikichi HIROSE[††]

あらまし　本論文では，ディープニューラルネットワークを用いた区分的線形変換による統計的特徴量強調の拡張を提案する．本提案手法の目的は，雑音環境下音声認識を想定した特徴量領域における雑音除去を目的とし，観測された音声特徴量から対応する静音環境下での音声特徴量の再現を行うことである．その際，ニューラルネットワークを用いて，観測された雑音環境下の音声特徴量より，ガウス混合分布でクラスタリングされた静音環境下における音声特徴量の領域を識別する．その後，各領域に対応する線形変換をニューラルネットワークにより得られる事後確率を重みとして足し合わせることで静音環境下での音声特徴量を推定する．これによって，ニューラルネットワークのもつ高い識別性能と，従来の生成モデルに基づく特徴量マッピング手法のもつ高い汎化性能の融合を狙う．Aurora-2 データベースを用いた連続音声認識実験により，提案手法は従来の区分線形変換法の一つである Stereo-based Piecewise LInear Compensation for Environments (SPLICE) と比較して，雑音が既知の条件では 53.72%単語誤り率を削減することができた．更に，ニューラルネットワークを回帰モデルとして用いたオートエンコーダと比較した場合，雑音環境が未知な条件で 26.96%の単語誤り率の削減が可能となった．

キーワード　音声認識，耐雑音性，特徴量強調，ディープラーニング，ニューラルネットワーク

1. まえがき

自動音声認識は，スマートフォンの普及とともにインタフェースの一つとして以前とは比較にならないほど身近なものとなってきた．しかし，雑音の影響が小さな静音環境においては高い認識性能を示す一方，雑音環境下では未だにその性能は十分とは言えない．そのため，ハンズフリーな音声認識システムの構築等を考えた場合，耐雑音性を高めることは非常に重要な課題と言える [1]．

† 東京大学大学院工学系研究科，東京都
Graduate School of Engineering, The University of Tokyo, 7–3–1 Hongo, Bukyo-ku, Tokyo, 113–0033 Japan
†† 東京大学大学院情報理工学系研究科，東京都
Graduate School of Information Science and Technology, The University of Tokyo, 7–3–1 Hongo, Bukyo-ku, Tokyo, 113–0033 Japan
a) E-mail: kashiwagi@gavo.t.u-tokyo.ac.jp
* 本論文は，学生論文特集秀逸論文である．
DOI:10.14923/transinfj.2015PDP0009

特徴量強調は耐雑音性を確保するためのフロントエンド処理の一つであり，Stereo-based Piecewise LInear Compensation for Environments (SPLICE) [2], [3] や Vector Taylor Series (VTS) [4]，Stereo-based Stochastic Mapping (SSM) [5] などの区分的線形変換をベースとした手法や，ニューラルネットワークを用いた Denoising AutoEncoder (DAE) [6]，事例ベースの特徴量強調手法 [7] などが提案されている．これらは，波形ドメインではなく，音声特徴量のドメインにおいて，観測された雑音環境下における音声特徴量から対応する静音環境下の音声特徴量を再現する．

SPLICE に代表される区分的線形変換法は，モデルを二つの段階に分けて考えることができる．まず，雑音環境下における音声特徴量空間をガウス混合モデル (Gaussian Mixture Model; GMM) でモデル化し，観測された音声特徴量の各フレームに対する，GMM の各要素分布からの事後確率を計算する．次に，得られた事後確率を重みとして，観測された雑音環境下にお

ける音声特徴量からの線形変換の足し合わせにより静音環境下における音声特徴量を推定する．

GMM は特徴量空間をマハラノビス距離を用いて確率的にクラスタリングすることに相当する．したがって，音声特徴量の GMM の各要素分布からの寄与率（事後確率）によって重み付けを行うことは，特徴量空間を確率的に領域分割することと等価である．SPLICE などの区分的線形変換は，この分割された各領域ごとでは入出力の対応が線形性を有すること（局所線形性）を仮定している．

局所線形性の仮定を考えた場合，理想的には静音環境下での特徴量空間の領域分割と雑音環境下での特徴量空間の領域分割が一致していることが望ましい．しかし，一般に雑音環境下においては，重畳されている雑音の影響で特徴量空間が縮退してしまう．そのため，例えば雑音が大きな環境では，音声特徴量が雑音によりマスクされてしまい入力特徴量空間を GMM でモデル化すると，各要素分布は雑音の種類や大きさに対応してモデル化される可能性が残る．雑音の種類による特徴量への影響は非線形であると考えられるため，SPLICE のように雑音環境下における領域分割を基準にした場合，局所線形性の仮定が不適切となる場合が考えられる．

そこで，この問題解決へのアプローチとして，REgularized piecewise linear mapping with DIscriminative region weighting And Long-span features (REDIAL) が提案されている [8], [9]．REDIAL は線形判別分析 (Linear Discriminant Analysis; LDA) により，領域分割が，静音環境下における音声特徴量の領域分割とより一致するような空間に特徴量を射影し，GMM によりモデル化する．しかし，LDA は線形変換であるため，雑音環境下における音声特徴量と静音環境における音声特徴量の複雑な関係を適切にモデル化できていない可能性がある．

一方，近年は深層ニューラルネットワーク（Deep Neural Network; DNN）[10] の発展に伴い，ニューラルネットワークを特徴量強調に利用した手法も提案されている．それらのうちの一つである，DAE はニューラルネットワークを回帰モデルとして用い，観測された音声特徴量から対応する静音環境下の音声特徴量を非線形かつ直接的に推定する．更に，DAE を多層にした Deep Denoising AutoEncoder (DDAE) は特に雑音環境が既知の条件において，高い精度で静音環境における音声特徴量を推定することが報告されている [6]．しかし，雑音環境が未知の場合では性能が低下することも分かっており，特定の環境に特化した強調となる傾向がある [11]．これは，ニューラルネットワークによる複雑な非線形変換によって，雑音環境下の音声特徴量空間を綿密にモデル化してしまうためと考えられる．

そこで，本提案手法はニューラルネットワークにより，静音環境における領域分割によって付与される領域ラベルを観測された雑音環境下の音声特徴量から識別する [12], [13]．その後，従来の区分的線形変換法と同様に，各領域ごとの線形変換を事後確率による重みづけで足し合わせ，対応する静音環境下における音声特徴量を推定する．提案法では，ニューラルネットワークを回帰モデルとしてではなく，識別モデルとして扱うことで，ニューラルネットワークの高い識別性と雑音に対する汎化性の高い区分線形変換の両立をはかる．

本論文の構成を示す．まず，**2.** で従来手法である SPLICE，REDIAL，DAE について説明を行い，**3.** で提案手法について述べる．その後，**4.** で実験により提案手法の有効性を示す．最後に **5.** でまとめる．

2. 関連研究

本章では，区分的線形変換手法である SPLICE とその拡張である REDIAL について説明を行う．また，ニューラルネットワークを用いた代表的な雑音抑圧手法である DAE についても言及する．

2.1 SPLICE

SPLICE は特徴量強調手法の一つであり，区分的線形変換によって，観測された時刻 $t \in T$ における雑音環境下における音声特徴量 $\boldsymbol{y}_t$ から，それに対応する静音環境下における音声特徴量 $\boldsymbol{x}_t$ を式 (1) で推定する．

$$\hat{\boldsymbol{x}}_t = \sum_{i=1}^{I} p(i|\boldsymbol{y}_t)\boldsymbol{A}_i \begin{bmatrix} 1 \\ \boldsymbol{y}_t \end{bmatrix} \tag{1}$$

ここで，$i \in I$ は混合のインデクスであり，$\boldsymbol{A}_i$ は各混合に対応する線形変換行列である．

SPLICE では，雑音環境下における音声特徴量の確率密度関数 $p(\boldsymbol{y})$ を GMM でモデル化する．混合 i における平均 $\boldsymbol{\mu}_i^{\boldsymbol{y}}$，分散 $\boldsymbol{\sigma}_i^{\boldsymbol{y}}$ のガウス分布を $\mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_i^{\boldsymbol{y}}, \boldsymbol{\sigma}_i^{\boldsymbol{y}})$ とすると，GMM からの特徴量 $\boldsymbol{y}_t$ の出力確率は，各ガウス分布の重みを $\pi_i^{\boldsymbol{y}}$ とした場合

$$p(\boldsymbol{y}_t) = \sum_{i=1}^{I} \pi_i^{\boldsymbol{y}} \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_i^{\boldsymbol{y}}, \boldsymbol{\sigma}_i^{\boldsymbol{y}}) \tag{2}$$

となる．これを用いて，雑音環境下の音声特徴量の各フレームに対する GMM の要素分布の寄与率を事後確率 $p(i|\boldsymbol{y}_t)$ として計算する．

$$p(i|\boldsymbol{y}_t) = \frac{p(\boldsymbol{y}_t|i)p(i)}{\sum_{i'=1}^{I} p(\boldsymbol{y}_t|i')p(i')} \tag{3}$$

式 (2) のように，SPLICE では雑音環境下の音声特徴量のみを用いて領域分割に用いる GMM の学習を行う．しかし，雑音環境においては重畳されている雑音の影響により音声の特徴量空間が縮退する．そのため，雑音環境下の音声特徴量で学習した GMM による領域分割では，各要素分布が雑音の種類や大きさに対応してしまうことが考えられる．理想的には静音環境下での特徴量空間の領域分割と雑音環境下での特徴量空間の領域分割が一致していることが望ましいと予想される．

そこで，MFCC によって構築された音声特徴量空間において，静音環境下，若しくは雑音環境下のそれぞれで，GMM によるモデル化と事後確率による重み付けを行った場合の比較を行った．表 1 は Aurora-2 [14], [15] の既知雑音条件における評価セットであるセット A を用いた単語誤り率である．SPLICE (oracle) は静音環境下におけるモデル化と観測された雑音環境下音声に対応する静音環境下の音声を領域分割にのみ用いたものである．SPLICE (oracle) の場合，

$$\begin{aligned}\hat{\boldsymbol{x}}_t &= \sum_{i^*=1}^{I^*} p(i^*|\boldsymbol{x}_t)\boldsymbol{A}_{i^*}\begin{bmatrix}1\\ \boldsymbol{y}_t\end{bmatrix}\\ p(i^*|\boldsymbol{x}_t) &= \frac{p(\boldsymbol{x}_t|i^*)p(i^*)}{\sum_{i^{*\prime}=1}^{I^*} p(\boldsymbol{x}_t|i^{*\prime})p(i^{*\prime})}\end{aligned} \tag{4}$$

として静音環境下における音声特徴量の要素分布 i^* に対する事後確率 $p(i^*|\boldsymbol{x}_t)$ を計算する．領域分割にのみ対応する静音環境下の音声特徴量を用いた SPLICE (oracle) の結果は，雑音の大きな環境でも頑健に対応する静音環境下の音声特徴量を推定できている．一方，これと比較した場合，通常の SPLICE の認識性能は特に雑音の大きな環境下において低下している．これは，雑音が大きくなった場合，特徴量空間の縮退により，要素分布が静音環境下における領域分割と異なる意味をもつためと考えられる．

表 1　雑音環境下における音声特徴量を用いて領域分割を行った場合の SPLICE と，正解の静音環境下における音声特徴量を用い領域分割を行った場合の SPLICE (oracle) の単語誤り率 (%)
Table 1　Word error rate (WER) for SPLICE and SPLICE (oracle) (%).

	SPLICE	SPLICE (oracle)
clean	0.57	0.69
SNR 20	1.08	0.76
SNR 15	1.99	0.70
SNR 10	4.65	0.77
SNR 5	16.76	0.93
SNR 0	49.96	0.95
SNR −5	81.46	1.13
Average	14.89	0.82

2.2 REDIAL

静音環境下の音声特徴量をモデル化した GMM の混合インデクスを $k \in K$ としたとき，学習データセットを $\{\{p(k|\boldsymbol{x}_t)\}_{k=1\ldots K}, \boldsymbol{d}_t\}_{t=1\ldots T}$ とする．ここで $\boldsymbol{d}_t$ は時刻 t における該当フレーム $\boldsymbol{y}_t$ とその前後 s フレームを連結した入力特徴量ベクトルであり，

$$\boldsymbol{d}_t = [\boldsymbol{y}_{t-s}^\top, \ldots, \boldsymbol{y}_{t-1}^\top, \boldsymbol{y}_t^\top, \boldsymbol{y}_{t+1}^\top, \ldots, \boldsymbol{y}_{t+s}^\top]^\top \tag{5}$$

である．学習段階では，静音環境下の特徴量空間で計算された事後確率と，観測特徴量から得られる事後確率が近くなるように，観測特徴量空間を次元圧縮する．次元圧縮行列 $\boldsymbol{L}$ は，ラベルを確率的に利用した LDA によって学習することができる．

$$\hat{\boldsymbol{L}} = \operatorname*{argmin}_{\boldsymbol{L}} \frac{\boldsymbol{L}^\top \boldsymbol{\Sigma}^w \boldsymbol{L}}{\boldsymbol{L}^\top \boldsymbol{\Sigma}^b \boldsymbol{L}} \tag{6}$$

$$\boldsymbol{\Sigma}^w = \sum_{k=1}^{K}\sum_{t=1}^{T} p(k|\boldsymbol{x}_t)\,(\boldsymbol{d}_t - \boldsymbol{\mu}_k^w)\,(\boldsymbol{d}_t - \boldsymbol{\mu}_k^w)^\top \tag{7}$$

$$\begin{aligned}\boldsymbol{\Sigma}^b = &\sum_{k=1}^{K}\left(\sum_{t=1}^{T} p(k|\boldsymbol{x}_t)\right)\\ &\times\left(\boldsymbol{\mu}_k^w - \frac{\sum_{t=1}^{T}\boldsymbol{d}_t}{T}\right)\left(\boldsymbol{\mu}_k^w - \frac{\sum_{t=1}^{T}\boldsymbol{d}_t}{T}\right)^\top\end{aligned} \tag{8}$$

$$\boldsymbol{\mu}_k^w = \frac{1}{\sum_{t=1}^{T} p(k|\boldsymbol{x}_t)}\sum_{t=1}^{T} p(k|\boldsymbol{x}_t)\boldsymbol{d}_t \tag{9}$$

次に，LDA により次元圧縮を行ったベクトル $\boldsymbol{v}_t = \hat{\boldsymbol{L}}\boldsymbol{d}_t$ を用いて K^* 混合の GMM を学習する．

$$p(\boldsymbol{v}_t) = \sum_{k^*=1}^{K^*} \pi_{k^*}^{\boldsymbol{v}} \mathcal{N}(\boldsymbol{v}_t; \boldsymbol{\mu}_{k^*}^{\boldsymbol{v}}, \boldsymbol{\sigma}_{k^*}^{\boldsymbol{v}}) \tag{10}$$

入力特徴量が得られた際の静音環境下における音声特徴量状態インデクス $k^* \in K^*$ に対する事後確率

$p(k^*|\boldsymbol{y}_t)$ を

$$p(k^*|\boldsymbol{y}_t) \simeq p(k^*|\boldsymbol{v}_t) = \frac{p(\boldsymbol{v}_t|k^*)p(k^*)}{\sum_{k^{*\prime}=1}^{K^*} p(\boldsymbol{v}_t|k^{*\prime})p(k^{*\prime})} \tag{11}$$

として計算する．最終的に得られた事後確率を重みとして用いた区分的線形変換により静音環境下における音声特徴量を

$$\hat{\boldsymbol{x}}_t = \sum_{k^*=1}^{K^*} p(k^*|\boldsymbol{y}_t)\boldsymbol{A}_{k^*}\boldsymbol{e}_t \tag{12}$$

$$\boldsymbol{e}_t = [1, \boldsymbol{y}_{t-u}^\top, \ldots, \boldsymbol{y}_{t-1}^\top, \boldsymbol{y}_t^\top, \boldsymbol{y}_{t+1}^\top, \ldots, \boldsymbol{y}_{t+u}^\top]^\top \tag{13}$$

として推定する．ただし，$\boldsymbol{A}_{k^*}$ は，要素分布 k^* に対応する線形変換行列であり，$\boldsymbol{e}_t$ は当該フレームと，その前後 u フレームを連結した拡張行列である．

$$\boldsymbol{e}_t = [1, \boldsymbol{y}_{t-u}^\top, \ldots, \boldsymbol{y}_{t-1}^\top, \boldsymbol{y}_t^\top, \boldsymbol{y}_{t+1}^\top, \ldots, \boldsymbol{y}_{t+u}^\top]^\top \tag{14}$$

なお，$\boldsymbol{A}_{k^*}$ は重み付き最小 2 乗誤差基準で学習する．

$$\hat{\boldsymbol{A}}_{k^*} = \underset{\boldsymbol{A}_{k^*}}{\operatorname{argmin}} \sum_{t=1}^{T} p(k|\boldsymbol{y}_t)||\boldsymbol{x}_t - \boldsymbol{A}_{k^*}\boldsymbol{e}_t||^2 \tag{15}$$

これは解析解を得ることができ，$\boldsymbol{A}_{k^*}$ は，

$$\hat{\boldsymbol{A}}_{k^*} = \boldsymbol{X}\boldsymbol{P}\boldsymbol{E}^\top(\boldsymbol{E}\boldsymbol{P}\boldsymbol{E}^\top)^{-1} \tag{16}$$

として計算することができ，ここで，特徴量の次元数を D とすると，$\boldsymbol{X} \in \mathcal{R}^{D\times T}$ と $\boldsymbol{E} \in \mathcal{R}^{(D(2u+1)+1)\times T}$ はそれぞれ出力と入力特徴量の拡張ベクトルを並べたデータ行列，$\boldsymbol{P} \in \mathcal{R}^{T\times T}$ は $p(k^*|\boldsymbol{y}_t)$ を対角に並べた行列である．なお，$\boldsymbol{A}_{k^*}$ は非常に大きな行列になるため，学習の際に正則化を導入する．

REDIAL は LDA によって雑音の影響を低減することで，雑音の大きな環境における観測特徴量の領域分割と，それに対応する静音特徴量の領域分割のミスマッチを減らし，静音特徴量の推定精度を向上させる．しかし，LDA はあくまで線形変換であるため，静音環境下における特徴量空間と雑音環境下における特徴量空間の非線形な対応を適切にモデル化できているとは言えない．

2.3 DAE

DAE はニューラルネットワークを用いて雑音環境下における音声特徴量から静音環境下における音声特徴量を直接的に推定する手法である．DDAE は DAE を多層にしたものであり，時刻 t の静音環境下における音声特徴量 $\boldsymbol{x}_t$ を

$$\hat{\boldsymbol{x}}_t = \boldsymbol{U}\boldsymbol{h}^{(n)}(\boldsymbol{d}_t) + \boldsymbol{c} \tag{17}$$

$$\boldsymbol{h}^{(n)}(\boldsymbol{d}_t) = \boldsymbol{\sigma}(\boldsymbol{W}^{(n)}\boldsymbol{h}^{(n-1)}(\boldsymbol{d}_t) + \boldsymbol{b}^{(n)}) \tag{18}$$

$$\boldsymbol{h}^{(1)}(\boldsymbol{d}_t) = \boldsymbol{\sigma}(\boldsymbol{W}^{(1)}\boldsymbol{d}_t + \boldsymbol{b}^{(1)}) \tag{19}$$

として推定する．ここで，$n \in N$ は中間層のインデクスであり，$\boldsymbol{U}$, $\boldsymbol{W}^{(n)}$ は重み行列，$\boldsymbol{c}$, $\boldsymbol{b}^{(n)}$ はバイアス項である．また，$\boldsymbol{h}^{(n)}$ は中間層の出力を表す．

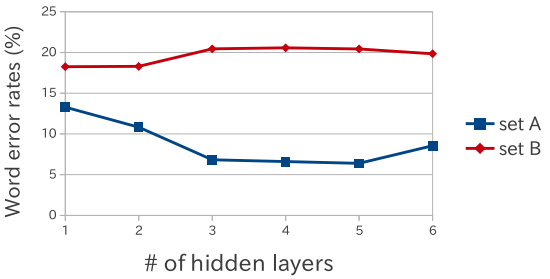

図 1　Aurora-2 データベースにおける，隠れ層の数を変化させた場合の DDAE の単語誤り率の変化
Fig. 1　The performance of DDAE with different numbers of hidden layers in Aurora-2 dataset.

図 1 は中間層の層数 N を変化した際の Aurora-2 における単語誤り率 (%) を示したものである．中間層の層数以外の詳細な実験条件は **4.** の実験と共通であるため，そちらを参照して頂きたい．セット A とセット B はそれぞれ既知雑音条件と未知雑音条件のテストセットである．なお，ニューラルネットワークの各隠れ層のノード数は予備実験により 1024 で統一した．既知雑音条件では層の数を増やすと単語誤り率が減少するが，未知雑音環境では逆に単語誤り率がわずかながら増加し，既知雑音の場合との差が大きくなる．これは，ニューラルネットワークを回帰モデルとして利用した場合，その複雑な非線形変換によって，雑音環境下の音声特徴量空間を綿密にモデル化してしまうためと考えられる．

3. DNN に基づく領域分割を用いた区分線形変換

前章で示したとおり，SPLICE に代表される区分的線形変換をベースとした特徴量強調手法は，局所線形性の仮定により，静音環境における特徴量空間の領域

分割と雑音環境における特徴量空間の領域分割の対応がとれていることが望ましい．しかし，雑音の大きな環境では，雑音の影響により音声の特徴量空間が縮退するため，GMM による特徴量空間のクラスタリングでは，雑音の種類や大きさに各要素分布が対応することが想定される．したがって，REDIAL のように観測特徴量空間をモデル化する際に識別的な基準を導入することが効果的だと考えられる．しかし，雑音の影響は非線形であると予想されるため，LDA のような線形変換をベースとしたモデルでは十分とは言えない．

一方，ニューラルネットワークを回帰モデルとして用い，観測特徴量から静音環境下の音声特徴量を推定するモデルは，高い認識性能を示すものの，未知雑音条件と既知雑音条件における単語誤り率には大きな差が存在する．これは，ニューラルネットワークのもつ非線形性により，雑音環境下の音声特徴量空間を綿密にモデル化してしまうことで，特定の環境に特化した傾向となるためだと考えられる．

そこで，提案手法では，観測特徴量から，それに対応する静音環境下における音声特徴量に対して最も高い寄与率をもつ要素分布をニューラルネットワークにより識別し，得られた事後確率を重みとした区分的線形変換により静音環境下の音声特徴量を推定する．これによって，ニューラルネットワークのもつ非線形性により，観測された雑音環境下における特徴量から，対応する静音環境下の特徴量の領域分割を高い精度で再現することが可能となる．また，ニューラルネットワーク自体は回帰モデルではなく領域の識別モデルとして機能するため，特徴量変換そのものは各要素分布である正規分布のもつ汎化性により，未知雑音環境下においても頑健に静音環境下の音声特徴量を推定することが可能となる．

今，時刻 t における D 次元の静音環境下における音声特徴量 $\boldsymbol{x}_t$ とそれに対応する雑音環境下における音声特徴量 $\boldsymbol{y}_t$ のパラレルデータ $\{(\boldsymbol{x}_t, \boldsymbol{y}_t)\}$ を考える．図 2 に学習段階の流れを示す．まず，静音環境下における音声特徴量の確率密度関数 $p(\boldsymbol{x})$ を GMM で学習する．

$$\begin{aligned} p(\boldsymbol{x}_t) &= \sum_{k=1}^{K} p(k)p(\boldsymbol{x}_t|k) \\ p(\boldsymbol{x}_t|k) &= \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_k^{\boldsymbol{x}}, \boldsymbol{\sigma}_k^{\boldsymbol{x}}) \\ p(k) &= \pi_k^{\boldsymbol{x}} \end{aligned} \tag{20}$$

これを用いて，GMM の各要素分布 k に対する事後確率を，

$$p(k|\boldsymbol{x}_t) = \frac{p(\boldsymbol{x}_t|k)p(k)}{\sum_{k'=1}^{K} p(\boldsymbol{x}_t|k')p(k')} \tag{21}$$

と表すことができる．

一方，認識時には，静音環境下における音声特徴量 $\boldsymbol{x}_t$ は観測できないため，観測された雑音環境下における音声特徴量 $\boldsymbol{y}_t$ から，それに対応する静音環境下における音声特徴量の，各要素分布に対する事後確率 $p(k|\boldsymbol{y}_t)$ を推定する必要がある．そこで，静音環境下における音声特徴量に対して寄与率の最も大きい要素分布を観測された音声特徴量から識別するニューラルネットワークを学習する．

$$\begin{aligned} p(k|\boldsymbol{y}_t) &\simeq p(k|\boldsymbol{d}_t) = \mathrm{softmax}_k(\boldsymbol{V}h^{(n)}(\boldsymbol{d}_t) + \boldsymbol{c}) \\ h^{(n)}(\boldsymbol{d}_t) &= \boldsymbol{\sigma}(\boldsymbol{W}^{(n)}h^{(n-1)}(\boldsymbol{d}_t) + \boldsymbol{b}^{(n)}) \\ h^{(1)}(\boldsymbol{d}_t) &= \boldsymbol{\sigma}(\boldsymbol{W}^{(1)}\boldsymbol{d}_t + \boldsymbol{b}^{(1)}) \end{aligned} \tag{22}$$

ここで，$\boldsymbol{\sigma}$ はベクターシグモイド関数であり，$\boldsymbol{V}$，$\boldsymbol{W}^{(n)}$ と $\boldsymbol{c}$，$\boldsymbol{b}^{(n)}$ はニューラルネットワークの重みとバイアスのパラメータ，$\boldsymbol{h}^{(n)}(\boldsymbol{y})$ は n 番目の隠れ層の出力ベクトルである．なお，事前学習として各層の初期値を制約付きボルツマンマシン (Restricted Boltzmann Machine; RBM) で学習した [10]．以上により，観測特徴量から静音環境下における音声特徴量に対する要素分布の事後確率 $p(k|\boldsymbol{y}_t)$ を推定することが可能となる．

評価段階では，ニューラルネットワークにより得られる事後確率 $p(k|\boldsymbol{y}_t)$ を重みとして，区分的線形変換によって静音環境下における音声特徴量 $\boldsymbol{x}_t$ を推定する (図 3)．

$$\hat{\boldsymbol{x}}_t = \sum_{k=1}^{K} p(k|\boldsymbol{y}_t)\boldsymbol{A}_k\boldsymbol{e}_t \tag{23}$$

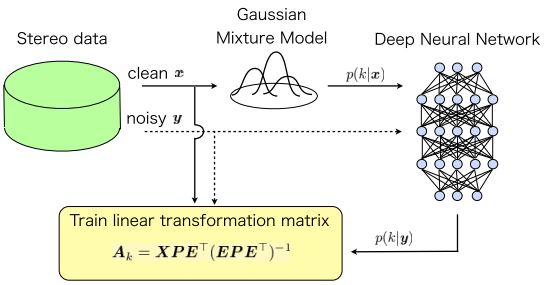

図 2　提案手法の学習時の流れ
Fig. 2　The training phase of our proposed method.

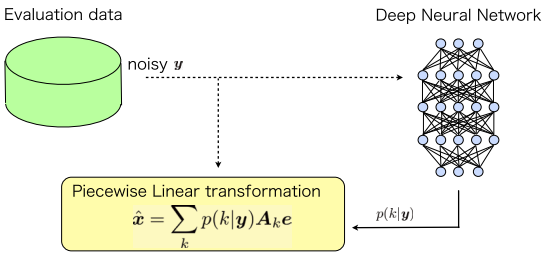

図 3　提案手法の認識時の流れ
Fig. 3　The testing phase of our proposed method.

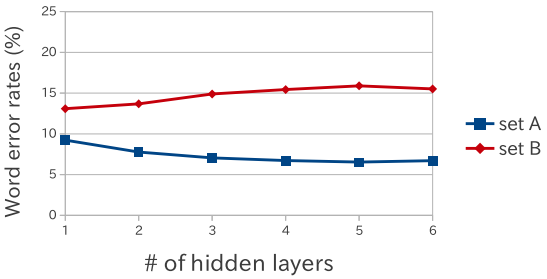

図 4　Aurora-2 データベースにおける，隠れ層の数を変化させた場合の提案手法の単語誤り率の変化
Fig. 4　The performance of the proposed method with different numbers of hidden layers in Aurora-2 dataset.

なお，$\boldsymbol{e}_t$ は式 (14) にあるとおり，時刻 t を中心に数フレーム分の特徴量を連結したベクトルである．

4. 実　　験

提案手法の有効性を Aurora-2 による連続数字読み上げ認識実験により評価した．音響モデルの学習と評価は complex backend と呼ばれる Aurora-2 の評価に標準的に用いられている設定を利用した [16]．データは幾つかの環境の雑音が重畳された雑音環境下における音声とそれに対応する静音環境下における音声が含まれている．学習データは 8,440 発話あり，本実験ではこれを全て音響モデルと雑音抑圧手法の学習に用いた．認識に用いる音響モデルは隠れマルコフモデル (Hidden Markov Model; HMM) を利用した．各 HMM は単語モデルであり，単語は 0 から 9 の数字（0 は 2 通りの読み方がある）から成る．各単語は 18 状態の HMM であり各状態は 20 混合の GMM であり，無音区間は 4 状態 HMM で各状態が 36 混合の GMM で構成される．また，ネットワーク文法を用いて認識を行う．静音環境下における音声のみで学習したもの (clean condition) と雑音環境下における音声も含むデータで学習したもの (multi condition) の 2 通りで比較した．なお，雑音抑圧処理を行ったデータに対して認識を行う際は，multi condition のモデルを学習する際のデータにも同様の雑音抑圧処理を行っている．

評価データセットは 3 種類 (A, B, C) があり，セット A とセット B はそれぞれ 28,028 発話，セット C は 14,014 発話から構成されている．セット A は学習時と同じ雑音のみを含む既知雑音条件，セット B は学習時と異なる雑音を含む未知雑音条件，セット C は雑音は未知，既知共に含むが，伝達関数が異なる．特徴量として MFCC とパワー，その 1 次，2 次微分の計 39 次元を用いた．本実験では，異なる音声対雑音比（SNR）のデータを用いて雑音抑圧を行うため，雑音の正規化に有効であると想定されるパワー項を用いている．また，ニューラルネットワークの学習は KALDI [17] を利用した．ニューラルネットワークを学習する際の誤差逆伝播法においては，学習データ 8,440 発話のうち 844 発話を開発セットとした．また，REDIAL と同様に線形変換の学習では正則化を導入した．

まず，ニューラルネットワークの層の数による単語誤り率の違いを図 4 に示す．入力に用いる特徴量 $\boldsymbol{d}_t$，$\boldsymbol{e}_t$ は当該フレームと，その前後 3 フレーム ($s = u = 3$) の計 7 フレームを用いた．あらかじめ予備実験により，ニューラルネットワークの各層のノード数は 1024 に設定し，静音環境下における音声特徴量空間をモデル化する際の GMM の混合数は $K = 1024$ に設定した．層の数を増やすと既知雑音条件では 5 層までは単語誤り率が単調に減少するが，未知雑音条件では DDAE と同様に効果が表れない．しかし，未知雑音条件と既知雑音条件で，層数 $N = 5$ のときに DDAE の場合は 14.05 ポイントの誤り率の差があったが，提案手法では，8.91 ポイントまで減少することが確認できた．これは，ガウス分布のもつ雑音に対する汎化性能を効果的に利用できていると考えられる．

次に，SPLICE，REDIAL，DDAE を行った場合の単語誤り率 (word error rate; WER) の比較を表 2 に示す．提案手法のパラメータは図 4 と同様のものを用い，中間層の数は $N = 3$ とした．SPLICE の雑音環境下における音声特徴量をモデル化する際の GMM の混合数 I と，REDIAL の静音環境下と雑音環境下における音声特徴量をモデル化する際の GMM の混合数 K，K^* は予備実験により共に 1024 に設定した．また，REDIAL の LDA による次元圧縮後の特徴量 $\boldsymbol{v}_t$ の次元数は 64 としている．ニューラルネットワークの

表 2　提案手法と従来手法との単語誤り率 (%) の比較 (%)
Table 2　Performance comparison among our proposal and conventional methods (WER %).

	clean condition (WER. %)				multi condition (WER. %)			
	set A	set B	set C	Ave.	set A	set B	set C	Ave.
Baseline	48.93	55.80	39.23	47.98	10.57	**11.89**	14.33	12.27
SPLICE	14.89	19.31	21.59	18.60	9.20	14.50	15.22	12.97
REDIAL	16.70	20.59	21.14	19.48	8.98	13.26	**12.45**	11.56
DDAE	**6.39**	20.44	17.20	14.68	5.97	18.50	14.67	13.04
Proposed	7.04	**14.93**	**15.54**	**12.51**	**5.64**	15.20	13.29	**11.38**

表 3　音響モデルを静音環境下の音声特徴量で学習した場合の，DDAE の雑音の種類，大きさごとの単語誤り率 (%)
Table 3　The WER of DDAE in each noisy condition (clean condition).

	closed-noise condition (Set A)					open-noise condition (Set B)				
	Subway	Babble	Car	Exhibition	Ave.	Restaurant	Street	Airport	Station	Ave.
Clean	0.58	0.60	0.81	0.46	0.61	0.58	0.60	0.81	0.46	0.61
SNR 20	1.29	0.85	0.78	0.96	0.97	1.11	2.21	1.82	1.14	1.57
SNR 15	1.54	1.45	1.16	1.79	1.49	2.21	6.32	3.91	3.46	3.98
SNR 10	2.27	3.02	2.00	2.81	2.53	6.05	18.02	10.11	8.67	10.71
SNR 5	4.64	8.01	4.47	6.70	5.96	18.33	40.39	25.95	24.38	27.26
SNR 0	14.77	31.29	18.85	19.10	21.00	51.30	69.35	58.75	55.29	58.67
SNR −5	46.58	72.64	63.11	49.27	57.90	97.91	90.93	96.42	86.52	92.95
Average	4.90	8.92	5.45	6.27	6.39	15.80	27.26	20.11	18.59	20.44

表 4　音響モデルを静音環境下の音声特徴量で学習した場合の，提案手法の雑音の種類，大きさごとの単語誤り率 (%)
Table 4　The WER of the proposed method in each noisy condition (clean condition).

	closed-noise condition (Set A)					open-noise condition (Set B)				
	Subway	Babble	Car	Exhibition	Ave.	Restaurant	Street	Airport	Station	Ave.
Clean	0.49	0.57	0.60	0.65	0.58	0.49	0.57	0.60	0.65	0.58
SNR 20	1.01	0.91	0.66	0.86	0.86	0.71	1.45	1.04	1.05	1.06
SNR 15	1.72	1.21	1.22	1.57	1.43	1.35	2.96	2.00	1.85	2.04
SNR 10	2.36	2.60	2.12	2.81	2.47	3.22	9.98	5.58	4.72	5.88
SNR 5	5.25	7.62	6.80	7.41	6.77	10.96	29.29	17.36	16.17	18.45
SNR 0	16.40	31.59	26.04	20.77	23.70	35.40	62.36	43.48	47.73	47.24
SNR −5	49.86	72.13	69.31	54.46	61.44	78.94	88.09	83.84	84.11	83.75
Average	5.35	8.79	7.37	6.68	7.04	10.33	21.21	13.89	14.30	14.93

パラメータは，DDAE は中間層の数を $N = 5$ とした．

音響モデルを静音環境下における音声のみで学習した場合，提案手法が最も良い結果となった．特に，SPLICE と比較した場合，既知雑音条件では 53.72%，未知雑音条件においては 18.54%の誤り削減率を得ることができている．また，音響モデルを雑音環境下における音声を含むデータで学習した場合(注1)も，SPLICE と比較して未知雑音環境においては，4.82%と誤り率が上昇してしまっているが，既知雑音条件では 38.70%の誤り削減率を得た．なお，REDIAL が set B において高い性能を示している．これは，REDIAL が線形変換行列により次元圧縮を行った空間で対角共分散の GMM を再び構築するため，擬似的に全角共分散の GMM を構築することが可能であることが原因であると考えられる．

更に，従来手法で最も単語誤り率が低かった DDAE と提案手法の既知雑音条件，未知雑音条件における単語誤り率を，雑音の種類と信号雑音比（SNR）別に表 3 と表 4 に示す．提案手法と DDAE のパラメータ設定は共に表 2 と同様のものを用い，音響モデルは静音環境下における音声のみで学習したものを用いた．既知雑音条件では，雑音が大きい場合，提案手法と比較したとき DDAE は 9.23%誤り率が低い．しかし，未知雑音条件においては，雑音の大きな場合でも

(注1)：Baseline が最も単語誤り率が低いが，これは Aurora-2 のデータセット特有の問題と考えられ，Droppo らの実験 [2] においても SPLICE により multiconditon における set B の性能が低下している．

提案手法が有効であり，DDAE と比較して平均として26.96%の誤り削減率を得ることができた．これによって，静音環境下における音声特徴量空間をモデル化した際のガウス分布による汎化性能が効果的であることが確認できた．

5. むすび

本論文では，ニューラルネットワークによる，観測音声特徴量からそれに対応する静音環境下における音声特徴量の所属する要素分布の識別と，それにより得られる事後確率を重みとして用いた区分線形変換法を提案した．実験的にも未知雑音条件，既知雑音条件共にSPLICE と比較して雑音が既知の条件では 53.72%単語誤り率を削減することができた．また，ニューラルネットワークを回帰モデルとして用いる DAE と比較した場合，ニューラルネットワークのもつ非線形性によって実現される複雑なモデル化と，静音環境下における音声特徴量のモデルとして用いたガウス分布のもつ汎化性能を組み合わせることで，既知雑音条件で低い誤り率を維持しつつ，雑音環境が未知な条件で26.96%の単語誤り率の削減が可能となった．

文　献

[1] M.J.F. Gales, “Model-based approaches to handling uncertainty,” in Robust Speech Recognition of Uncertain or Missing Data, pp.101–125, Springer, 2011.

[2] J. Droppo, L. Deng, and A. Acero, “Evaluation of the SPLICE algorithm on the Aurora2 database,” INTERSPEECH, vol.1, pp.217–220, 2001.

[3] J. Droppo, L. Deng, and A. Acero, “Evaluation of splice on the aurora 2 and 3 tasks,” INTERSPEECH, vol.1, pp.29–32, 2002.

[4] J. Li, M.L. Seltzer, and Y. Gong, “Improvements to vts feature enhancement,” 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4677–4680, 2012.

[5] M. Afify, X. Cui, and Y. Gao, “Stereo-based stochastic mapping for robust speech recognition,” IEEE Trans. Audio Speech Language Process., vol.17, no.7, pp.1325–1334, 2009.

[6] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” Proc. 25th International Conference on Machine Learning, ACM, pp.1096–1103, 2008.

[7] J.F. Gemmeke, T. Virtanen, and A. Hurmalainen, “Exemplar-based sparse representations for noise robust automatic speech recognition,” IEEE Trans. Audio Speech Language Process., vol.19, no.7, pp.2067–2080, 2011.

[8] M. Suzuki, T. Yoshioka, S. Watanabe, N. Minematsu, and K. Hirose, “MFCC enhancement using joint corrupted and noise feature space for highly non-stationary noise environments,” 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4109–4112, 2012.

[9] M. Suzuki, T. Yoshioka, S. Watanabe, N. Minematsu, and K. Hirose, “Feature enhancement with joint use of consecutive corrupted and noise feature vectors with discriminative region weighting,” IEEE Trans. Audio Speech Language Process., vol.21, no.10, pp.2172–2181, 2013.

[10] G. 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.

[11] A.L. Maas, Q.V. Le, T.M. O’Neil, O. Vinyals, P. Nguyen, and A.Y. Ng, “Recurrent neural networks for noise reduction in robust ASR,” INTERSPEECH, in CD-ROM, 2012.

[12] Y. Kashiwagi, D. Saito, N. Minematsu, and K. Hirose, “Discriminative piecewise linear transformation based on deep learning for noise robust automatic speech recognition,” 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp.350–355, 2013.

[13] 柏木陽佑，齋藤大輔，広瀬啓吉，峯松信明，“Deep learning に基づくクリーン音声状態識別による雑音環境下音声認識，” 音響秋季講論集，pp.9–12, 2013.

[14] http://aurora.hsnr.de/aurora-2.html

[15] H.-G. Hirsch and D. Pearce, “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW), 2000.

[16] D. Pearce and A. Gunawardana, “Aurora 2.0 speech recognition in noise: Update 2. Complex backend definition for aurora 2.0,” 2002 [Online]. http://icslp2002.colorado.edu/special_sessions/aurora

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The kaldi speech recognition toolkit,” IEEE 2011 Workshop on Automatic Speech Recognition and Understanding CFP11SRW-USB, Dec. 2011.

（平成 27 年 6 月 1 日受付，10 月 5 日再受付，
12 月 3 日早期公開）

柏木　陽佑

2013 年東京大学大学院情報理工学系研究科修士課程修了．修士 (情報理工学)．現在，同大大学院工学系研究科博士課程に在籍．2014 年より日本学術振興会特別研究員 DC2．音声認識，特に音響モデルに関する研究に従事．日本音響学会会員．

齋藤　大輔　（正員）

2011 年東京大学大学院工学系研究科博士課程修了．博士 (工学)．2010–2011 年日本学術振興会特別研究員 DC2．現在，同大学院情報理工学系研究科助教．音声合成，音声変換，音声分析，音声認識の研究に従事．ISCA Interspeech Best Student Paper Awards，日本音響学会独創研究奨励賞板倉記念などを受賞．日本音響学会，情報処理学会，映像情報メディア学会，信号処理学会，IEEE，ISCA 各会員．

峯松　信明　（正員）

1995 年東京大学大学院工学系研究科博士課程修了．博士 (工学)．現在，同大大学院工学系研究科教授．2002–2003 年在外研究員 (KTH，スウェーデン)．科学から工学に至るまで，音声コミュニケーションに関する研究に従事．IEEE，ISCA，SLaTE，IPA，CALICO，音響学会，情報処理学会，人工知能学会，音声学会，音声言語医学会，外国語教育メディア学会各会員．

広瀬　啓吉　（正員：フェロー）

1972 年東京大学工学部電気工学科卒業．1977 年同大学院博士課程修了．工学博士．同年東京大学工学部電気工学科講師．1994 年同電子工学科教授．1996 年東京大学大学院工学系研究科電子情報工学専攻教授．1999 年同新領域創成科学研究科教授．2004 年 10 月より同情報理工学系研究科教授．1987 年米国 MIT 客員研究員．音声言語情報処理分野一般についての研究開発に従事，特に韻律に着目した研究．IEEE，米国音響学会，ISCA (Board メンバー)，情報処理学会，日本音響学会，人工知能学会，言語処理学会，信号処理学会各会員．