Vol.2015-SLP-109 No.15
2015/12/2

感情音声認識における DNN および CNN ボトルネック特徴量の検討

向原 康平[1,a)]　サクリアニ サクティ[1]　吉野　幸一郎[1]　グラム ニュービッグ[1]　中村 哲[1]

概要：話者感情の揺らぎは音声へ影響を与え，音声認識システムにおいてモデルとのミスマッチを発生させ認識精度を悪化させる．本研究では，DNN ボトルネック特徴量および CNN ボトルネック特徴量を用いることを提案し，感情音声認識精度の改善を図る．ボトルネック構造のニューラルネットワークによって特徴量変換を施したボトルネック特徴量は，入力音声の変動に対して頑健な音響特徴量を抽出できることが示されている．ボトルネック特徴量とは，中間層のユニット数を少なくしたボトルネック構造の多層ニューラルネットワークから抽出する特徴量である．ボトルネック特徴量は特徴量強調が行われ，感情音声のゆらぎに左右されない音素の本質的な成分を抽出されていることが期待されている．本実験では感情音声に対してボトルネック特徴量変換を行い，それぞれの特徴量で GMM-HMM 音響モデルを再学習する．この時のボトルネック音響モデルの感情音声に対する精度向上を確認する．また他の特徴量変換手法と組み合わせることで認識精度の向上を図る．DNN，CNN ボトルネック特徴量を用いた認識精度はそれぞれのベースラインと比較し，認識精度の改善が確認できた．また他の特徴量変換手法の組み合わせにより認識精度の向上を確認できた．

キーワード：感情音声認識，ボトルネック特徴量，ディープニューラルネットワーク，畳み込みニューラルネットワーク，特徴量変換手法

Exploring CNN and DNN Bottleneck Features for Emotional Speech Recognition

Kohei Mukaihara[1,a)]　Sakriani Sakti[1]　Koichiro Yoshino[1]　Graham Neubig[1]
Satoshi Nakamura[1]

Abstract: Emotion influences the speech and degrades. Therefore emotional speech degrades ASR quality due to the mismatch between input speech and the acoustic model. In this study, we focus on feature transformation methods to solve this mismatch. We propose a tandem approach using DNN bottleneck features and CNN bottleneck features for emotional speech recognition. The bottleneck features are made by a deep neural network hidden layer that has a smaller number of nodes than other layers. We hypothesize that bottleneck sturucture can extract features and bottleneck features represent essential features of phonemes. By using bottleneck features for emotional speech recognition, we confirm that results improve results compared with other feature transformation methods. In addition, we combine the proposed methods and other feature transformation methods to improve emotional speech recognition.

Keywords: Emotional speech recognition, Bottleneck features, Deep neural network，Convolutional Neural Network，Feature transformation

1. はじめに

音声認識は技術の進歩に伴い普及が進み様々な場面で使用されるようになってきたが，入力音声とモデルのミス

[1] 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan.
[a)] mukaihara.kohei.me4@is.naist.jp

マッチによりうまく認識できない場面がある [1]．このような認識精度低下を引き起こす要因の一つとして，話者感情の揺らぎがある．感情は音声に影響を与えモデルとのミスマッチを生じさせる [2]．通常の音声認識システムでは平静状態での入力を想定しており，感情音声に対応した音声認識システムは少ない．

従来，感情音声認識では感情の変動に対してモデル適応手法が用いられてきた．音響特徴量に対して適応学習する手法 [3] や発話内容に関して適応学習する手法 [4] である．適応学習手法は適応のターゲットとなる感情と入力音声の感情が一致している場合には認識精度の向上が見込めるが，一致しない場合は認識精度の向上は見込めない．また感情に関して適応モデルを作成する場合，適応モデルの数は定義する感情の個数必要になるという問題点がある．

本研究では感情音声を入力としたときに発生するミスマッチを解消する手段として特徴量変換手法に着目し，ボトルネック特徴量を用いることを提案する [5][6]．ボトルネック特徴量は多層ニューラルネットワークの中間層のユニット数を少なくしたボトルネック構造のネットワークから抽出される．ボトルネック構造の中間層から抽出する特徴量は入力特徴量を次元圧縮し，様々な入力音声の中から普遍的な特徴を抽出していると考えられる．今回は感情音声データを学習に用いてネットワークを構築するため，感情による影響の少ない特徴量の抽出が期待される．

本実験では深層ニューラルネットワーク（DNN: Deep Neural Network）と畳み込みニューラルネットワーク (CNN: Convolutional Neural Network) の二種類のニューラルネットワークからボトルネック特徴量を抽出し，複合アプローチによってそれぞれの特徴量を用いて音声認識を行い比較・検討する．また線形判別分析 (LDA: Leniar Discriminant Analysis) や特徴空間最尤線形回帰 (fMLLR: feature-space Maximum Liklihood Linear Regression) などの特徴量変換手法を組み合わせることで，さらなる精度の向上を図る．

2. GMM/HMM 音声認識

2.1 音声認識の定式化

音声認識は入力音声 X が与えられたときにその単語列 W を求める問題であり，下式のように表現できる [7]．

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W|X) \tag{1}$$

ここで $P(W|X)$ をベイズ則に基づいて書き換えたとき下式で示すように 2 項の積が最大となる W を求める問題として定式化される．

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W) P(X|W) \tag{2}$$

このとき $P(W)$ は単語列が生成される事前確率を表し，$P(X|W)$ は単語列 W から入力音声が生成される確率を表す．それぞれの確率モデルはそれぞれに定式化ができる．$P(W)$ は言語モデルと呼ばれ，単語列は時系列に探索される性質から単語 n-gram によってモデル化される．このとき言語モデルは音声認識のタスクドメインに合致するように学習データを用意する必要がある．$P(X|W)$ は音響モデルと呼ばれ，音素状態ごとに音響特徴量の分布を混合ガウス分布（GMM: Gaussian Mixture Model）で表現する隠れマルコフモデル（HMM: Hidden Marcov Model) によってモデル化される．また，実際の音響モデルでは音響特徴量から直接的に単語列のモデル化を行うわけではなく，単語と音素列の関係を定義した発音辞書を用いて音素列と単語のモデル化を行う．つまり，音響モデルにおいても音声認識システムが使われる状況に合致した学習データを用意する必要がある．音響モデルにおいて入力音響特徴量とモデルの間にミスマッチが生じた場合，頑健な特徴量への変換やモデル適応などの手法が必要となる．

2.2 線形判別分析

音声認識では各時間フレームごとの音響特徴量を用いて認識を行う．このとき各時間フレームは前後のフレームの影響を受けるため，隣り合ったフレームの特徴量を連結させることで有効な特徴量になる．しかし単純なフレームの連結は次元数の増加につながるため，次元圧縮を施してから用いられることが多い．音声認識において LDA は次元圧縮を行う特徴量変換手法として用いられ，雑音や残響音に対して頑健な特徴量を出力することが知られている [8]．

ここで LDA による特徴量変換手法について説明する．入力となる特徴量ベクトルは，時間フレーム t に対して前後 k フレームの特徴量を時系列に連結して作成する．LDA では，d 次元の特徴量ベクトル $\boldsymbol{x}_t$ を d' 次元の特徴量ベクトル $\boldsymbol{y}_t$ への変換する行列 $\boldsymbol{W}$ を得ることで次元削減を行う．このとき，特徴量ベクトル $\boldsymbol{x}_t$ は時間フレーム t に対する HMM 状態ラベルと対応付けられ，このデータをもとにクラス内分散 $\boldsymbol{S}_W$, クラス間分散 $\boldsymbol{S}_B$ の計算を行う．ここでクラス内分散が小さく，クラス間分散の大きくなるような d' 次元の $\boldsymbol{y}_t$ を抽出することが LDA による次元削減となる．$\boldsymbol{y}_t$ を得るため，$d \times d'$ の変換行列 $\boldsymbol{W}$ を考えると，以下の目的関数を定義することができる．

$$\hat{W} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \frac{|\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_W \boldsymbol{W}|}{|\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_B \boldsymbol{W}|} \tag{3}$$

これを満たす $\hat{W}$ を解析的に求めることで変換行列を得る．

2.3 fMLLR

音響モデルと入力音声のミスマッチ解消の手段として，モデル適応手法がある．最尤線形回帰 (MLLR: Maximum Likelihood Linear Regression) は，モデル適応手法の一つで，HMM/GMM 音声認識における話者適応の代表的な手

法である [9]．MLLR は音響モデルの GMM における平均ベクトル $\boldsymbol{\mu}$=(μ_1,..., μ_n) の線形変換によってモデルを更新する．式 (4) に平均ベクトル $\boldsymbol{\mu}$ 関するアフィン変換の式を示す．

$$\hat{\boldsymbol{\mu}} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} \tag{4}$$

$\boldsymbol{A}$ は $n \times n$ の行列であり，$\boldsymbol{b}$ は次元数 n のベクトルである．MLLR は教師データとして与えられる話者の音声に対して尤度が最大になるようにこの $\boldsymbol{A}$, $\boldsymbol{b}$ を推定する．話者適応の手法として，平均ベクトルの変換のみでなく GMM の共分散に対しても変換をかける fMLLR 手法がある．この手法はモデルの更新にとどまらず，特徴量空間全体の座標変換ととらえることができるため，特徴量変換手法として考えられている．

3. 提案手法

本研究では特徴量変換手法に着目し，感情音声に対して認識精度の向上を図る．本章では複合アプローチに基づき，ボトルネック層を持つ多層パセープトロン (MLP: Multi Layer Perceptron）から得られたボトルネック特徴量を用いた特徴量変換手法を提案する．提案手法である DNN ボトルネック特徴量，CNN ボトルネック特徴量を構成する要素についてそれぞれ説明を行い，GMM/HMM 音響モデルとボトルネック特徴量の複合アプローチによる特徴量変換手法について説明する．

3.1 ボトルネック特徴

3.1.1 DNN ボトルネック特徴量

DNN は多層のニューラルネットワークによって構成されるが，本手法では図 1 の通り中間層の一部を他のユニット数よりも小さくするボトルネック構造を用いる．DNN による学習は事前学習 (pre-training) と微調整 (fine-tuning) に分けられる．pre-training では，入力層から順にボトルネック中間層に向けてオートエンコーダを用いて教師なし学習を行い，初期値を与える．pre-training が終了すると，ボトルネック中間層から HMM 状態を表現している出力層までをつなげて fine-tuning を行う．fine-tuning には誤差逆伝搬による教師あり学習を行う．この時，出力層は状態数の数だけユニットを持っており，入力音声に対応する音素を分類するように学習が行われる．完成したネットワークは特徴量変換のために用いられ，変換されたボトルネック特徴量を用いて GMM の再学習を行う．

3.1.2 CNN ボトルネック特徴量

CNN は，音声の局所的な特徴量抽出を担う畳み込み層と，細かいズレを問題にしないようにぼかし効果を加えるプーリング層を交互に繰り返す構造の多層ニューラルネットワークである．図 2 に示す通り，畳み込み層，プーリング層を交互に繰り返しその後全結合の多層ニューラルネットワークを連結させる．

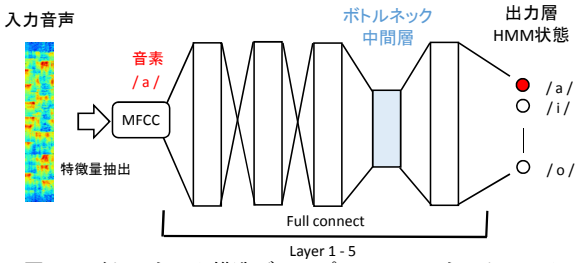


図 1 ボトルネック構造ディープニューラルネットワーク

Fig. 1 The bottleneck structure of deep neural network.

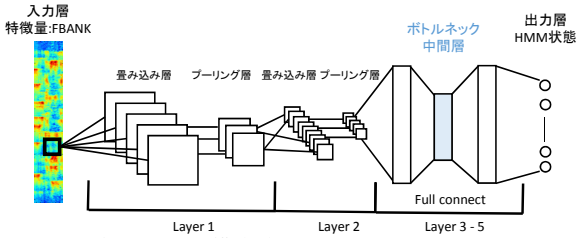


図 2 ボトルネック構造畳み込みニューラルネットワーク

Fig. 2 The bottleneck structure of convolutional neural network.

CNN は局所的な歪みに対しても頑健性を持つため，感情の変化による影響を受けにくくなることが期待される．本実験ではボトルネック構造の CNN を考える．入力からの数層は畳み込み層・プーリング層を繰り返す構造を持ち，全結合のニューラルネットワークを接続する．この時，中間層の一部が他のユニットよりも少ないボトルネック構造になっている．

3.2 GMM/HMM 音響モデル・ボトルネック特徴量による複合アプローチ

本節では複合アプローチによる特徴量変換について説明する．複合アプローチは [10]，MLP を用いた非線形特徴量変換手法とみなすことができる．複合アプローチでは，音響特徴量ベクトル $\boldsymbol{x}_t$ は MLP の出力ベクトルを正規化した $\boldsymbol{\Psi}(\boldsymbol{y_t})$ に変換される．この $\boldsymbol{\Psi}(\boldsymbol{y_t})$ の出力分布は GMM によりモデル化され，HMM 状態 q_t の尤度計算を可能にする．元の音響特徴量ベクトル $\boldsymbol{x_t}$ の出力分布 $P(\boldsymbol{x_t}|q_t)$ は，MLP による変換規則に従い $P(\boldsymbol{\Psi}(\boldsymbol{y_t})|q_t)$ として表現される．複合アプローチは，識別問題を解くように学習したネットワークを特徴量変換器としてとらえ，変換特徴量を HMM/GMM 音声認識の入力として用いる手法である．複合アプローチでは，変換特徴量として出力層における MLP 特徴量が考えられてきた [11] が，異なる構造としてボトルネック構造となっている中間層をボトルネック特徴量として用いることも可能である [12]．本手法においては DNN，CNN それぞれで音素識別問題を解くように学習した．ニューラルネットワークの枠組みとして考えると，少ない中間層から音素識別問題を解くことになるため，ボ

トルネックになっている中間層には識別に有効な特徴量が抽出されていることが期待される．また複合アプローチの利点として，GMM/HMM 音響モデルで用いられてきた次元圧縮や話者適応などの手法をそのまま用いることができる．本研究では，ボトルネック特徴量音響モデルに従来の特徴量変換手法を適用し，認識精度向上を図る．

4. 実験

4.1 実験設定

実験では，感情音声に対してボトルネック特徴量変換手法を用いて音声認識を行う．また他の特徴量変換手法とを組み合わせた手法との比較検討を行う．提案法である DNN ボトルネック特徴量 (DNN-BNF) 変換手法，CNN ボトルネック特徴量 (CNN-BNF) 変換手法に対してそれぞれ異なる設定で実験を行い，結果を比較検討する．今回は音声認識器に Kaldi tool kit[13] を用いて学習，テストを行う．音響モデルは日本語話し言葉コーパス (CSJ: Corpus of Spontaneous Japanese)，約 45 時間を用いて学習した．DNN-BNF は MFCC+Δ+ΔΔ，CNN-BNF は FBANK をそれぞれ特徴量として HMM/GMM 音響モデルを学習する．また感情音声として感情評定値付きオンラインゲーム音声チャットコーパス (OGVC: Online gaming voice chat corpus with emotional label)[14] を用いた．OGVC には受容（ACC），怒り（ANG），期待（ANT），嫌悪（DIS），恐怖（FEA），喜び（JOY），悲しみ（SAD），驚き（SUR）の 8 種類の感情ラベルが定義されている．それぞれの感情発話が約 20 種類用意されており，その発話内容をプロの俳優男女 2 名ずつが読み上げた音声を収録している．感情には強度が設定されており，感情を含まない平静状態（level 0）から弱（level 1），中（level 2），強（level 3）と感情強度を上げて同じ文章を収録する．一名につき 664 発話，計 2656 発話が収録されており，今回の実験では，言語モデルは OGVC のデータのみから学習する．また男女のペアでデータを半分に分け，一方をテストデータ，もう一方を学習データとして用いる．テストは感情と強度でそれぞれ分類し，各テストデータは男女合わせて 40 発話で構成される．学習データはテストデータに用いない 1328 発話で構成される．

4.1.1 DNN-BNF 変換手法実験条件

DNN-BNF 変換手法は図 3 (a)，図 3 (b) に示すように実験を行う．DNN-BNF 変換の前に特徴量変換を施す場合と，ボトルネック特徴量に対してさらに特徴量変換を施す場合に着目して実験を行う．これ以降，DNN-BNF を DBNF と呼ぶ．それぞれの実験結果について説明する．

LDA
: MFCC に対して LDA 特徴量変換を施し，LDA 特徴量で音響モデルの再学習を行う．特徴量変換手法のベースラインとして扱う．

DBNF
: LDA 特徴量変換を施した特徴量を入力として DNN を学習し，そのネットワークを特徴量変換器として用いボトルネック特徴量を抽出，音響モデルを再学習する．

fMLLR-DBNF
: LDA 特徴量変換の後に fMLLR によって特徴量空間の変換を施し，その特徴量を入力として DNN の学習を行う．

LDA-fMLLR
: LDA 特徴量変換を施し，テストデータの特徴量に対して尤度が高くなるように特徴量空間を変換する．

DBNF-fMLLR
: DBNF 変換を施し，テストデータの特徴量に対して尤度が高くなるように特徴量空間を変換する．

本実験では LDA の入力を MFCC 前後 4 フレームを合わせてベクトルとし，40 次元の特徴量に次元圧縮を施す．fMLLR はテストデータに対して特徴量空間の変換を行っている．この時テストデータは感情，強度ごとに分類してそれぞれでテストを行っているので，話者に加えて感情に適した空間に特徴量を写像している．DBNF の学習には Kaldi+PDNN toolkit を用いる．DNN の入力次元は LDA，fMLLR ともに 40 次元 ×11 フレームで 440 次元とする．ネットワーク構造はデフォルト設定である 6 層，1024 ユニット，ボトルネック中間層は 5 層目で 42 ユニットになっている．

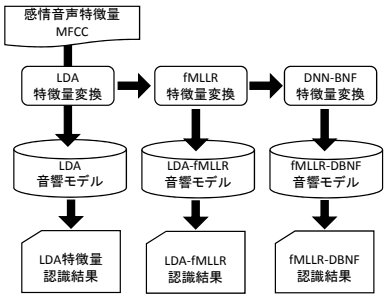
4.1.2 CNN-BNF 変換手法実験条件

CNN-BNF 変換手法は，図 4 に示すように実験を行う．本手法においては FBANK 特徴量から CNN-BNF 変換する場合と，変換した後にさらに fMLLR 特徴量変換を施した場合それぞれについて認識精度を確認する．これ以降，CNN-BNF を CBNF と呼ぶ．それぞれの実験結果について説明する．

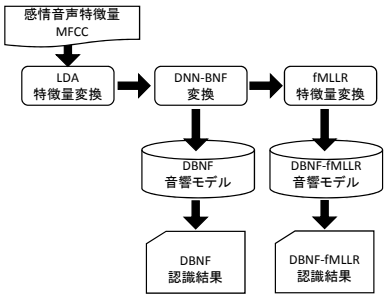
CBNF
: FBANK 特徴量を入力として CNN を学習し，そのネットワークを特徴量変換器として用いてボトルネック特徴量を抽出，音響モデルを再学習する．

CBNF-fMLLR
: CBNF 変換を施したのちに fMLLR 特徴量変換を行う．この時の変換は話者，感情に適応して行われる．

CNN-BNF の学習は Kaldi-PDNN toolkit を用いる．入力には FBANK 1320 次元を用いる．畳み込み層，プーリング層を合わせて 1 層の CNN 特徴量抽出層と考え，入力から 2 層はこの層を重ねる．その後に 4 層からなるボトルネック構造ニューラルネットワークを接続する．この時 3 層目がボトルネック特徴量となりユニット数は 42，そのほかのユニット数は 1024 とした．



(a) DNN-BNF 変換前の特徴量変換の適応



(b) DNN-BNF 変換後の特徴量変換の適応

図 3　DNN-BNF 変換手法
Fig. 3　DNN-BNF feature transformation

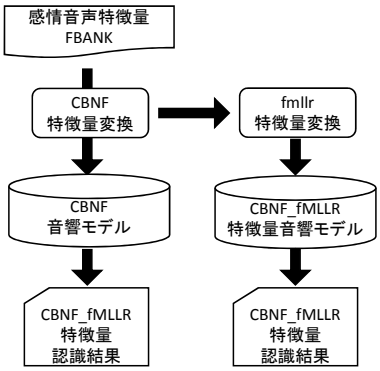


図 4　CNN-BNF 変換手法
Fig. 4　CNN-BNF feature transformation

4.2　実験結果および検討

4.2.1　DNN-BNF 変換手法

感情ごとの単語誤り率 (WER: word error rate) の平均を表 1 に示し，感情強度ごとにすべての感情音声の WER を平均した結果を図 5 に示す．ベースラインである LDA 特徴量変換に対して，DBNF 変換によって平均で約 4.6%の認識精度の向上が確認できた．また fMLLR 特徴量変換と組み合わせることで，認識精度の改善も確認できた．DBNF と比較すると，DBNF-fMLLR では約 4.1%の改善．fMLLR-DBNF では約 8%の改善を確認できた．fMLLR 変換は LDA に施した場合と比べると効果があまり大きくなく，変換をしてから識別学習を行うほうが認識精度を向上させることが確認できた．感情強度ごとに実験結果を確認すると，感情強度が高くなるにしたがって全体の認識精度は低下する．各手法ごとに確認すると，fMLLR-DBNF が感情に対して頑健な特徴量になっていることが確認できる．また単体の DBNF と比較して，fMLLR 特徴量変換を合わせた場合の精度向上も確認できた．今回の fMLLR は話者と感情の両方に適応しているため，感情音声に対しても影響を受けにくい特徴量になっていると考えられる．

表 1　DNN-BNF を用いた感情ごとの音声認識結果
Table 1　Emotional speech recognition results using DNN-BNF

	LDA	DBNF	fMLLR-DBNF	LDA-fMLLR	DBNF-fMLLR
ACC	23.33	15.40	14.29	14.92	13.81
ANG	24.02	20.34	12.26	17.28	15.20
ANT	15.14	16.56	10.13	10.24	13.94
DIS	21.89	35.74	22.29	25.10	32.93
FEA	37.34	26.79	22.78	32.07	27.43
JOY	23.87	17.27	8.11	13.36	12.91
SAD	27.08	30.27	17.53	22.67	23.04
SUR	58.43	31.27	22.09	32.77	27.90
AVE	28.89	24.21	16.19	21.05	20.09

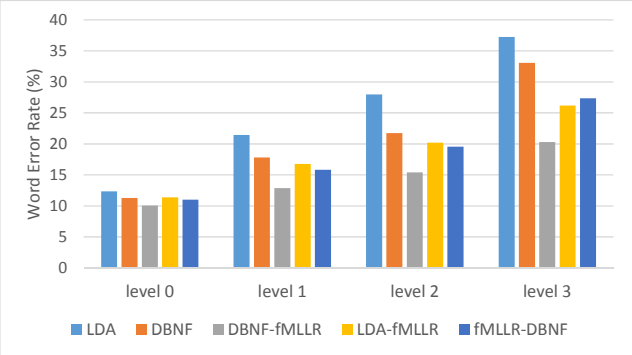


図 5　感情強度ごとの認識結果
Fig. 5　DNN-BNF feature transformation on each emotion level

4.2.2　CNN-BNF 変換手法

感情ごとの WER の平均を表 2 に示し，感情強度ごとにすべての感情音声の WER を平均した結果を図 6 に示す．CBNF 変換手法による結果は，LDA を入力として用いて学習する DBNF とその fMLLR 変換を行った CBNF-fMLLR と比較する．CBNF を DBNF と比較した場合，平均の WER で 1.4%ほど精度が劣っている．また fMLLR 変換による認識精度の向上は確認できるものの，DBNF-fMLLR 変換と比較すると 3%ほど精度が劣っている結果になった．しかし感情ごとに確認すると，DNN-BNF 全体で認識精度の低下がみられていた DIS の感情において認識精度が勝る結果となった．このことから，CNN-BNF は DNN-BNF とは異なる特徴量抽出を可能にしているということが確認

表 2　DNN-BNF と CNN-BNF の音声認識結果の比較
Table 2　CNN-BNF emotional speech recognition results comparing with DNN-BNF

	DBNF	DBNF-fMLLR	CBNF	CBNF-fMLLR
ACC	15.40	13.81	18.73	20.95
ANG	20.34	15.20	20.10	15.69
ANT	16.56	13.94	18.52	13.07
DIS	35.74	32.93	30.32	27.71
FEA	26.79	27.43	25.95	27.43
JOY	17.27	12.91	21.77	17.87
SAD	30.27	23.04	31.5	25.37
SUR	31.27	27.90	38.39	35.77
AVE	24.21	20.09	25.66	22.98

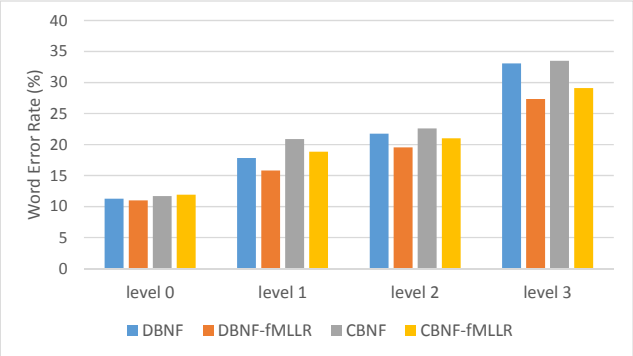


図 6　DNN-BNF と CNN-BNF の感情強度ごとの認識結果比較
Fig. 6　CNN-BNF results comparing with DNN-BNF on each emotion level

できた．また認識精度に関しては，特徴量変換を施した特徴量の入力を行っていないため，さらに認識精度が向上する可能性がある．

5. まとめ

本研究では感情音声認識に対して，ボトルネック特徴量変換手法による精度改善を図った．具体的には，DNN-BNF 変換手法，CNN-BNF 変換手法をそれぞれ提案し，感情音声の認識精度を確認した．また感情音声に対してはボトルネック特徴量変換手法に加えて，LDA，fMLLR の特徴量変換手法を合わせて用いることで，さらなる認識精度の向上を確認した．この時，fMLLR 特徴量をボトルネック特徴量変換の入力とする fMLLR-DBNF 手法は，ベースラインと比較して約 12%の精度向上を確認できた．CNN-BNF 変換手法は DNN-BNF と比較したとき精度が劣るが，DNN-BNF とは異なる特徴量を抽出していることが確認できた．今後は CBNF に特徴量変換を施した入力を用いることで，DNN-BNF の精度に近づけることができるのか確認を行う．また，今回は感情音声認識に対して頑健な特徴量への変換を考えたが，今後は認識の際に感情情報を用いる適応手法についても検討する．

謝辞
本研究の一部は， JSPS 科研費 26540117 および 26870371 の助成を受け実施した．

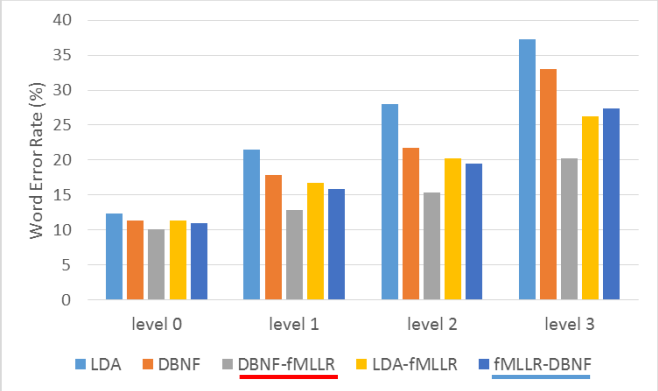
参考文献
[1] 篠田浩一, 堀貴明, 堀智織, 篠崎隆宏. 「音声認識」は今後こうなる! 情報処理学会研究報告. SLP, 音声言語情報処理, Vol. 2014, No. 2, pp. 1-6, jan 2014.
[2] 門谷信愛希, 阿曽弘具, 鈴木基之, 牧野正三. 音声に含まれる感情の判別に関する検討. 電子情報通信学会技術研究報告. SP, 音声, Vol. 100, No. 522, pp. 43-48, dec 2000.
[3] Bjorn Schuller, Jan Stadermann, and Gerhard Rigoll. Affect-robust speech recognition by dynamic emotional adaptation. In Proc. speech prosody. Citeseer, 2006.
[4] Theologos Athanaselis, Stelios Bakamidis, Ioannis Dologlou, Roddy Cowie, Ellen Douglas-Cowie, and Cate Cox. Asr for emotional speech: Clarifying the issues and enhancing performance. Neural Networks, Vol. 18, No. 4, pp. 437-444, 2005.
[5] Jonas Gehring, Yinping Miao, Florian Metze, and Alex Waibel. Extracting deep bottleneck features using stacked auto-encoders. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 3377-3381. IEEE, 2013.
[6] 高島悠樹, 中鹿亘, 滝口哲也, 有木康雄. (2015). 構音障害者音声認識のための混合正規分布に基づく音素ラベリングの検討. 電子情報通信学会技術研究報告 IEICE technical report: 信学技報, 115(100), 71-76.
[7] 西崎博光. 音声言語処理のための要素技術と音声ドキュメント処理への応用. 電子情報通信学会技術研究報告. EMM, マルチメディア情報ハイディング・エンリッチメント, Vol. 114, No. 33, pp. 11-16, 2014.
[8] 正木大介，鈴木雅之，峯松信明，広瀬啓吉 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 日本音響学会秋季講演論文集，1-1-10，pp.29-32 (2012-9)
[9] 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 D, Vol. 87, No. 2, pp.371-386, 2004.
[10] Hynek Hermansky, Daniel W Ellis, and Shantanu Sharma. Tandem connectionist feature extraction for conventional hmm systems. In Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on, Vol. 3, pp. 1635-1638. IEEE, 2000.
[11] Chen, Barry Y., Qifeng Zhu, and Nelson Morgan. "Tonotopic Multi-Layered Perceptron: A Neural Network for Learning Long-Term Temporal Features for Speech Recognition." ICASSP (1). 2005.
[12] F. Grezl and P. Fousek, "Optimizing bottleneck features for LVCSR," in Proc. ICASSP. 2008, pp. 4729-4732.
[13] D.Povey, A.Ghoshal, G.Boulianne, L.Burget, O.Glembek, N.Goel, M.Hannemann, P.Motlicek, Y.Qian, P.Schwarz, J.Silovsky, G.Semmer, K.Vesely, "The Kaldi Speech Recognition Toolkit," in Proc.ASRU, 2010.
[14] 有本泰子, 河津宏美, 大野澄雄, 飯田仁. 感情音声のコーパス構築と音響的特徴の分析: Mmorpg における音声チャットを利用した対話中に表れた感情の識別. 情報処理学会研究報告. MUS,[音楽情報科学], Vol. 74, pp. 133-138,feb 2008.

正誤表

1. ５ページ右段 図５ 感情強度ごとの認識結果

データ系列名

誤：LDA, DBNF, DBNF-fMLLR, LDA-fMLLR, fMLLR-DBNF



正：LDA, DBNF, fMLLR-DBNF, LDA-fMLLR, DBNF-fMLLR