ボトルネック特徴量を用いた感情音声認識の検討 *

☆ 向原 康平, サクティ サクリアニ, 吉野 幸一郎, ニュービッグ グラム, 中村 哲 (奈良先端大)

1 はじめに

話者感情の揺らぎによる入力音声への影響はモデルとのミスマッチを引き起こし，音声認識精度を低下させる要因となる [1]．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に着目し，ボトルネック特徴量を用いることを提案する．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)，2 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う．また，2 種類のボトルネック特徴量を組み合わせる逐次ボトルネック計算手法と並列ボトルネック計算手法それぞれから新しい特徴量を生成した．その中でも並列ボトルネック計算手法から抽出した特徴量は最も良い認識結果を示し，感情音声の認識精度向上が確認できた．

2 特徴量変換手法

2.1 線形判別分析

音声認識では各時間フレーム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て用い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．音声認識において線形判別分析 (LDA:Linear Discriminant Analysis) 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になることが知られている．

2.2 非線形特徴量変換手法

非線形特徴量変換手法に多層パーセプトロン (MLP:Multi Layer Perceptron) から抽出した特徴量を用いて GMM を再学習する複合アプローチがある．MLP を特徴量変換器として用いることで，入力特徴量 $\boldsymbol{x}_t$ を正規化された MLP の出力 $\boldsymbol{y}_t$=$\boldsymbol{\Psi}(\boldsymbol{x_t})$ に変換する．この $\boldsymbol{\Psi}(\boldsymbol{x_t})$ 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 $\boldsymbol{x_t}$ の出力分布 $P(\boldsymbol{x_t}|\, q_t)$ は，MLP による変換規則に従い $P(\boldsymbol{\Psi}(\boldsymbol{x_t})|\, q_t)$ として表現される．MLP 特徴量を用いる複合アプローチでは，識別問題を解くように学習した MLP を用いて特徴量変換を行い，変換特徴量を GMM-HMM 音声認識の入力とする．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が，DNN においてボトルネック構造の中間層から抽出したボトルネック特徴量 [2] を変換特徴量として用いることも可能である．

3 提案手法

3.1 DNN-BNF

DNN ボトルネック特徴量（DNN-BNF）を用いたアプローチについて説明を行う．DNN は多層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) に分割される．pre-training は教師なし学習によって初期値を与える操作であるが，本実験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構成する．pre-training 完了後，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 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 HMM 状態数だけユニットを持っており，入力特徴量の HMM 状態を識別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，抽出されるボトルネック特徴量を用いて GMM の再学習を行う．本研究で用いたボトルネック構造はオートエンコーダのように，入力特徴量から出力を表現するために必要な特徴量を抽出することが期待される．

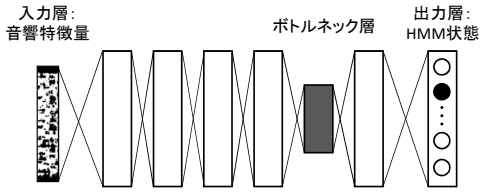

Fig. 1 ボトルネック構造ディープニューラルネットワーク (DNN-BNF)

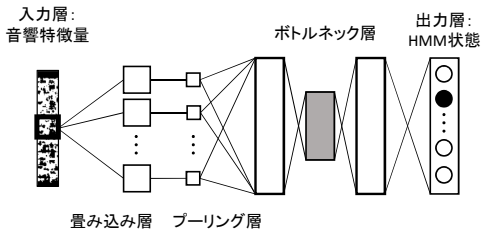

Fig. 2 ボトルネック構造畳み込みニューラルネットワーク (CNN-BNF)

3.2 CNN-BNF

CNN ボトルネック特徴量（CNN-BNF）を用いたアプローチについて説明を行う．CNN は音声の局所的な特徴量抽出を担う畳み込み層と普遍的な特徴の抽出により微小な変動への対応を行うプーリング層を交互に繰り返す多層ニューラルネットワークである [3]．Fig.2 で示すように，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．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．

3.3 特徴量抽出手法の組み合わせ

特徴量抽出では手法を組み合わせることで，認識精度の向上が見込まれる．また，抽出されたボトルネック特徴量は組み合わせることによって単独で用いる場合よりも良い結果を示すことが知られている [4]．本研究ではさらなる認識精度向上を実現するため，特徴量抽出手法を組み合わせる．また，抽出されたボトルネック特徴量を組み合わせることを提案する．

逐次ボトルネック計算手法 (Stack-Comb)

Fig.3 で示すように抽出した CNN-BNF を DNN-BNF 変換の入力とする．最終的に DNN-BNF 変換によって抽出された特徴量を用いて GMM-HMM 音響モデルを再学習する．

---

* Exploring Bottleneck Features for Emotional Speech Recognition, by MUKAIHARA, Kohei, SAKTI, Sakriani, YOSHINO, Koichiro, NEUBIG, Graham, NAKAMURA, Satoshi (NAIST)

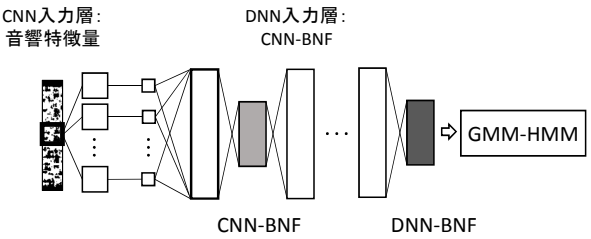

Fig. 3　DNN-CNN 逐次ボトルネック計算手法

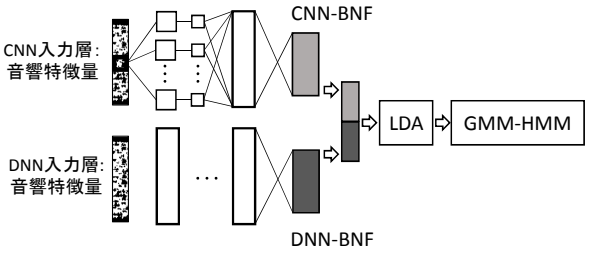

Fig. 4　DNN-CNN 並列ボトルネック計算手法

並列ボトルネック計算手法 (Parallel-Comb)

それぞれ抽出した，DNN-BNF と CNN-BNF を組み合わせて一つの特徴量ベクトルとして扱う．LDA による次元削減を施し，新しい特徴量として GMM-HMM 音響モデルを再学習する．

4　実験的評価

4.1　実験条件

本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．LDA による特徴量変換をベースラインとして，提案法である DNN-BNF 変換手法，CNN-BNF 変換手法の結果を比較検討する．また，CNN-BNF 変換を施したのちに DNN-BNF 変換を行う逐次ボトルネック計算手法，DNN-BNF と CNN-BNF を組み合わせる並列ボトルネック計算手法それぞれから特徴量を抽出し認識を行う．

本実験で使用する感情音声は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)[5] から取得した．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．また，感情強度が設定されており，同一の発話内容で 4 段階 (平静，弱，中，強) の強度の異なる音声を収録している．最も強度の低い音声は感情のない平静音声としている．

本実験の DNN，CNN の学習は Kaldi+PDNN ツールキットを用いて行い，ネットワーク構造はデフォルト設定をそのまま用いた．また，学習データは OGVC の感情音声を用いた．ベースとなる音響モデルは日本語話し言葉コーパスから学習し，言語モデル，辞書モデルは OGVC から学習した．DNN の入力はベースラインの LDA 特徴量，CNN の入力は LDA 変換を施される前の FBANK とする．LDA，ボトルネック層による次元圧縮はどちらも 40 次元への圧縮とする．

4.2　実験結果

手法ごとの音声認識結果を単語誤り率 (%) で表し，感情ごとに平均した結果を Table 1 に示す．また，感情強度，弱，中，強ごとの平均認識結果と全体の平均を Fig.5 に示す．ベースラインである LDA 特徴量変換とそれぞれ比較すると DNN-BNF では平均で約 4%，CNN-BNF では平均で約 3%の認識精度改善が確認できた．また，DNN-BNF と CNN-BNF を比較すると平均の認識精度では DNN-BNF の方が認識精度が向上しているが，認識結果を感情ごとに確認すると

Table 1　感情ごとの認識結果:Word Error Rate(%)

	LDA	DNN-BNF	CNN-BNF	Stack-Comb	Parallel-Comb
ACC	23.33	15.40	18.73	17.78	11.11
ANG	24.02	20.34	20.10	18.13	10.66
ANT	15.14	16.56	18.52	22.98	13.29
DIS	21.89	35.74	30.32	28.92	19.28
FEA	37.34	26.79	25.95	36.08	21.73
JOY	23.87	17.27	21.77	14.71	5.15
SAD	27.08	30.27	31.50	28.18	15.81
SUR	58.43	31.27	38.39	43.82	28.84
AVE	28.89	24.21	25.66	28.13	16.97

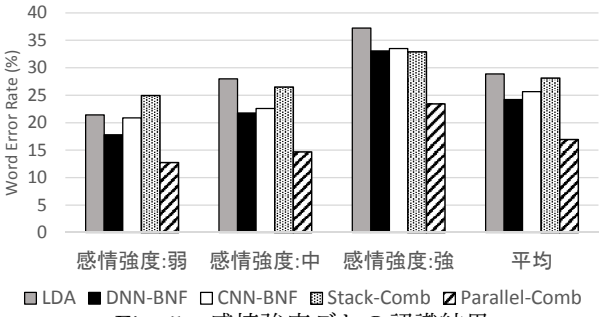

Fig. 5　感情強度ごとの認識結果

必ずしも DNN-BNF の方が良い結果ではない．感情によっては CNN-BNF の結果が良いこともあり，異なる特徴量を抽出していると考えられる．CNN-BNF を DNN-BNF の入力として用いる逐次ボトルネック計算手法は平均で確認すると DNN-BNF，CNN-BNF 単独の場合よりも認識精度が悪くなるという結果になった．この結果から特徴量変換手法の単純な組み合わせは精度を向上につながらないことを確認した．DNN-BNF と CNN-BNF を連結させて，一つのベクトルとして扱い次元圧縮を施した並列ボトルネック計算手法は最も良い結果を示し，単独で用いる場合に比べて約 7%以上の認識精度改善を確認できた．ボトルネックから抽出された特徴量は HMM 状態それぞれを表現するために必要な特徴量と考えられ，それぞれ単独で用いた時よりも組み合わせて用いたほうがより有効な特徴量として機能することが分かった．

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．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向上を確認した．実験結果から DNN-BNF，CNN-BNF は異なる特徴量を抽出していることが示唆され，それらを組み合わせた並列ボトルネック計算手法が最も認識精度を向上させた．

謝辞

本研究の一部は，(独）情報通信研究機構の委託研究「知識・言語グリッドに基づくアジア医療交流支援システムの研究開発」および JSPS 科研費 24240032 および 26870371 の助成を受け実施した．

参考文献

[1] 門谷信愛希 他. 電子情報通信学会技術研究報告. Vol. 100, No. 522, pp. 43-48, dec 2000.

[2] Jonas Gehring et al., In Proc. ICASSP, pp. 3377-3381. 2013.

[3] Alexander Waibel et al., IEEE TASLP, Volume 37, No.3, pp. 328. - 339

[4] D. Yu et al., Proc. Interspeech 2011, pp.237 - 240, 2011

[5] 有本泰子 他. 情報処理学会研究報告. MUS,[音楽情報科学], Vol. 74, pp. 133-138,feb 2008.