Deep Neural Network を用いた音声と環境音のマルチタスク学習*

☆川西 誠司, 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

通常，音声認識における環境音は雑音とみなされ除去される対象である．しかし，音声認識において環境音を同時に認識するメリットもある．例えば，音声にどのような環境音が含まれているか認識できれば，それに応じてモデルの変更や適応がリアルタイムで可能になり，また環境音から発話者の状況推定も可能になる．そこで本稿では，Deep Neural Network (DNN) とマルチタスク学習を組み合わせた雑音環境下での音声認識と環境音の識別手法について検討する．また，音声認識において有効性が示されているボトルネック特徴量と i-vector を用いることでさらなる精度改善を行う．

2 マルチタスク DNN-HMM による音声と環境音の認識

2.1 DNN-HMM による音声認識

本研究では，音声認識で用いる音響モデルを，DNN-HMM によってモデル化する．DNN-HMM は隠れマルコフモデル (HMM：Hidden Markov Model) の各状態を DNN によって直接求めるハイブリッド型アプローチであり，従来の混合ガウス分布 (GMM：Gaussian Mixture Model) を用いた GMM-HMM によるモデル化よりも認識精度が高いことが知られている [1]．通常，DNN-HMM の学習では入力特徴量としてメルフィルタバンク対数パワー (FBANK) が利用されることが多い．

2.2 マルチタスク学習

マルチタスク学習は複数の関連するタスクについて，同時に学習を行う機械学習手法の 1 つである．マルチタスク学習では，各タスク間で素性などの情報を共有することで，個々で学習するより予測精度が向上することが期待される [2]．DNN におけるマルチタスク学習では，複数のタスクで隠れ層のパラメータを共有し，出力層のみを個々で扱う．また，各タスクごとに損失関数を用意し，それらの誤差の合計が小さくなるように学習を進めていく．Bell らは，DNN-HMM を用いた音声認識において，トライフォンの HMM 状態を求めるタスクと，モノフォンの HMM 状態を求めるタスクを DNN でマルチタスク学習する手法を提案しており [2]，マルチタスク学習が過学習を防ぐための正則化としても機能することを報告している．

2.3 マルチタスク DNN-HMM

Fig.1 に提案法であるマルチタスク DNN-HMM の概要を示す．提案法では，従来の音声認識における DNN-HMM と同様に，入力音声のフレームごとの HMM 状態を予測するタスクに，フレームごとに含まれる環境音を予測するタスクを追加する．入力特徴量と隠れ層は共有し，最後の出力層のみを各タスクでに割り振るように構築する．これにより、中間層での情報を共有することで、それぞれの認識精度が向上することを期待する．

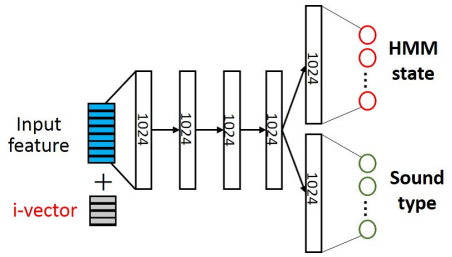

Fig. 1 マルチタスク DNN-HMM

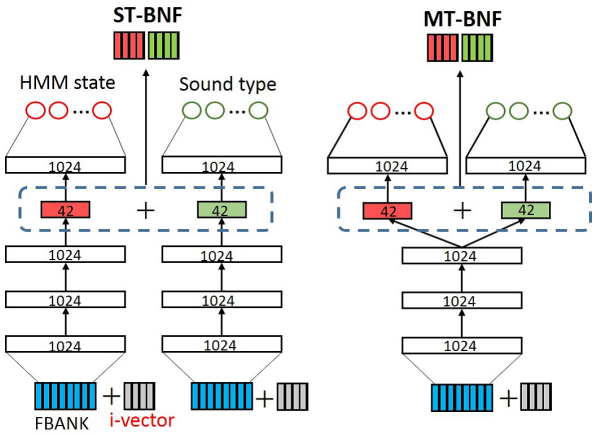

Fig. 2 シングルタスク BN-DNN とマルチタスク BN-DNN を用いたボトルネック特徴量の抽出

3 特徴量変換手法の導入

3.1 ボトルネック特徴量

ボトルネック特徴量はボトルネック層を持つ DNN (BN-DNN) から得ることができる特徴量であり，そのタスクに特に有効な特徴量を抽出することができる [3]．Fig.2 に本実験で用いるボトルネック特徴量の抽出過程を示す．1 つ目は左側に示しているシングルタスクボトルネック特徴量 (ST-BNF) である．これは音声認識と環境音認識それぞれのタスクについて，FBANK を入力としてシングルタスク BN-DNN を学習し，ボトルネック特徴量を抽出，結合したものである．2 つ目は右側の図で示しているマルチタスクボトルネック特徴量 (MT-BNF) である．これはマルチタスク BN-DNN の分岐後に各タスクでボトルネック層を設け，そこから抽出した 2 つのボトルネック特徴量を結合したものである．これらの 2 種類のボトルネック特徴量を，それぞれマルチタスク DNN-HMM の入力特徴量とした場合の認識性能の比較を行う．

3.2 i-vector

i-vector は話者情報をよく表現していると考えられていることから [4]，話者分類によく用いられ，音声認識においても話者適応としての有効性が示されている．i-vector は各発話を GMM でモデル化し，その平均ベクトルを結合した GMM スーパーベクトルに，因子分析の手法を用いることで得ることができる．本研究では，話者を識別するのではなく、環境音

*Multi-Task Deep Neural Networks for Speech and Environmental Sound Recognition, by KAWANISHI, Seiji, SAKTI, Sakriani, YOSHINO, Koichiro, NEUBIG, Graham, NAKAMURA, Satoshi (NAIST)

Table 1　評価データに対する各手法の認識性能の比較

	Speech recognition (WER:%)					Sound recognition (FER:%)				
	0db	5db	10db	20db	Ave.	0db	5db	10db	20db	Ave.
1 Single-Task with FBANK	60.9	42.2	30.3	22.8	39.0	50.3	46.4	45.1	54.8	49.2
2 Multi-Task with FBANK	61.8	42.5	31.1	23.3	39.7	42.2	44.8	44.1	53.6	46.2
3 Multi-Task with ST-BNF	59.2	41.1	29.9	21.7	38.0	31.2	31.5	34.1	48.5	36.3
4 Multi-Task with MT-BNF	62.5	44.6	32.0	23.5	40.7	42.3	43.2	44.7	54.9	46.3
5 Multi-Task with i-vector + ST-BNF(FBANK)	61.1	42.3	29.3	21.9	38.6	17.8	15.7	15.2	20.2	17.2
6 Multi-Task with i-vector + MT-BNF(FBANK)	61.4	42.2	31.3	22.4	39.3	19.7	18.4	19.2	26.5	21.0
7 Multi-Task with ST-BNF(i-vector + FBANK)	57.6	36.2	25.2	18.6	**34.4**	12.1	12.7	13.1	16.3	**13.6**
8 Multi-Task with MT-BNF(i-vector + FBANK)	60.4	40.4	28.0	21.4	37.6	11.9	12.0	12.7	18.7	13.8

を識別する目的で使用する．また，i-vector とボトルネック特徴量を組み合わせた音声認識手法も提案されており [5]，本研究でもその手法を利用する．その際，i-vector の導入タイミングの違いが認識結果に与える影響も調査するため，2 通りの方法で i-vector を使用した．1 つ目は，FBANK を入力として BN-DNN から抽出したボトルネック特徴量に i-vector を結合させ，それをマルチタスク DNN-HMM の入力とする方法である．2 つ目は，ボトルネック特徴量の抽出の際に FBANK に i-vector を結合させて BN-DNN を学習し，抽出したボトルネック特徴量をマルチタスク DNN-HMM の入力とする方法である．

4　実験条件

本実験では，マルチタスク DNN-HMM を用いることで，環境音を含む音声認識と，含まれる環境音の認識を同時に行う．音声データは日本語話し言葉コーパスから 20000 発話，1000 発話をそれぞれ学習用，評価用として用意した．また，Web サイト [6] から収集した人混み雑音や拍手音，アラーム音など 10 種類の環境音を，1 発話につき 1 種類だけ発話の開始から終了まで重畳することで，人工的に雑音環境下の音声データを作成した．評価データは SN 比が 0db，5db，10db，20db となるように環境音を重畳した場合の 4 セット用意した．

実験で用いる各種 DNN の構造は Fig.1 と Fig.2 に示している通りである．初期 GMM-HMM は，メル周波数ケプストラム係数に線形判別分析と最尤線形変換を施した特徴量を入力としている．その学習には環境音を重畳した 20000 発話の学習用データを使用した．評価方法については，音声認識タスクではマルチタスク DNN-HMM を用いて音声認識を行い，その単語誤り率を算出した．環境音認識タスクでは，マルチタスク DNN-HMM によって予測された環境音が，正解ラベルと一致しているかをフレームごとに調べ，誤り率を算出した．

5　実験結果

Table1 に，評価データの SN 比ごとに認識実験を行った結果を示す．ベースラインとなる FBANK を用いたシングルタスク DNN-HMM の認識結果を 1 行目，マルチタスク DNN-HMM の認識結果を 2 行目に示す．それぞれの結果から，音声認識においては提案法により 0.7%ほど精度が低下したが，環境音認識においては 3 %の精度向上が確認できた．

Table1 の 3 行目と 4 行目にボトルネック特徴量を用いた場合の認識結果を示す．ST-BNF を用いた場合，各認識精度は FBANK を用いた場合より向上したが，MT-BNF を用いた場合は低下した．このことから，ST-BNF はマルチタスク DNN-HNN の入力特徴量として有効であるが，MT-BNF は有効でないことが明らかになった．

Table1 の 5 行目から 8 行目に，ボトルネック特徴量と i-vector を組み合わせた場合の認識結果を示す．5，6 行目はマルチタス DNN-HMM の入力特徴量として i-vector を加えた場合，7，8 行目はボトルネック特徴量の抽出の際に，BN-DNN の入力特徴量として加えた場合である．結果から，i-vector の導入により環境音認識精度の大幅な改善が確認できた．特に，7 行目の ST-BNF の抽出の際に i-vector を用いた場合に最も優れた性能を示しており，ベースラインから音声認識精度は 4.6 %，環境音認識精度は 35.6 %の改善が確認できた．これらの結果から，i-vector を環境音適応として利用することで認識性能の向上が可能になること，i-vector は特徴量抽出の際の入力特徴量として利用することが有効であることも明らかになった．

6　まとめ

本稿では，DNN とマルチタスク学習を組み合わせた音声認識と環境音認識手法であるマルチタスク DNN-HMM を提案した．実験結果から，マルチタスク DNN-HMM では環境音の認識精度の改善しかできなかったが，ボトルネック特徴量や i-vector を組み合わせることで，音声認識においても改善できることが確認できた．今後は各種ハイパーパラメータの最適化や，マルチタスク学習を特徴量抽出に効果的に応用する方法を検討する．

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] F. Seide *et al.*, *Proc. INTERSPEECH*, pp. 437-440, 2011.

[2] P. Bell and S. Renals, *Proc. INTERSPEECH*, pp. 4290-4294, 2015

[3] J. Gehring *et al.*, *Proc. ICASSP*, pp. 3377-3381, 2011.

[4] N. Dehak *et al.*, *IEEE trans. Speech Audio Process* , Vol.19, No.4, pp. 788-798, 2013.

[5] P. Cardinal *et al.*, *Proc. INTERSPEECH*, pp. 2867-2871, 2015.

[6] ”soundsnap”, http://www.soundsnap.com, (参照 2016-01-09)