話者行列の重みづけによる少量適応発話における DNN の話者適応法*

○太刀岡 勇気, 成田 知宏 (三菱電機・情報総研)

1 はじめに

Deep Neural Network (DNN) を音声認識の音響モデルに利用することで、音声認識性能が向上する。一方で、音響モデルを対象の話者に適応化することで、認識性能を向上させることができることが知られている。以前主流であった Gaussian Mixture Model (GMM) では最尤基準に基づき、モデルパラメータを適応化させる方法が広く使われている [1]。ところが、DNN ではこのような理論的に最適な適応法が確立されていないため、さまざまな適応法が提案されている。中でも、モデルの一部を話者ごとに切り替え、切り替えた部分もしくは切り替えた後にモデル全体を、通常のモデル学習と同じ誤差逆伝搬の手順で適応化させる方法が主流である [2, 3]。

文献 [2] の方法は i-vector のような補助特徴量を必要とし、計算量が多い。またその補助特徴量の精度により適応の精度が大きく異なる。文献 [3] の方法は適応データがある程度多い場合には有効であるが、通例多くの適応データを利用することは難しい。本報では、[3] の方法をベースに、[2] のように補助特徴量を使うことなく、[3] の適応データが大量に必要となるという問題を解決する話者適応法を提案する。

2 SAT-DNN

ここでは、ベースラインとしている speaker-adaptive-training-DNN (SAT-DNN) [3] について簡単に説明する。例を、図 1 に示す。SAT-DNN ではある特定の層 (SAT 層、図では第 2 層) を話者ごとに切り替える。まず学習セットにより、不特定話者で DNN 全体を最適化する。次に評価音声に対して、不特定話者の DNN を用いて音声認識結果とその時系列の HMM の状態番号であるアライメントを得る。この時のアライメントは上述のように教師なしで音声認識により得ることもできるし、予め適応データの発話内容が分かっていれば、それをもとに教師有りでも得ることができる。適応には、学習セットから求めた不特定話者の重み行列を初期値として、評価話者ごとに SAT 層を切り替え、DNN 全体の最適化を複数エポック繰り返す。このとき N_e 人の評価話者中の話者 $n_e(1 \leq n_e \leq N_e)$ に対する重み行列 (ここでは話者行列と呼ぶ) を $\boldsymbol{W}_{n_e}$、話者適応層への入力である話者適応層の下の中間層の出力を $\boldsymbol{x}_{in} \in \mathbb{R}^{D_{in}}$、話者適応層の出力を $\boldsymbol{x}_{out} \in \mathbb{R}^{D_{in}}$ とすると、$\boldsymbol{x}_{out}$ は

$$\boldsymbol{x}_{out} = \boldsymbol{W}_{n_e}\boldsymbol{x}_{in} + \boldsymbol{o}_{n_e} \tag{1}$$

で表される。図 1 では、$D_{in} = D_{out} = 3$ である。これにより GMM での SAT と同様に、DNN 全体を話者適応することができる。最後に得られた話者適応モデルにより、最終的な音声認識結果を得る。

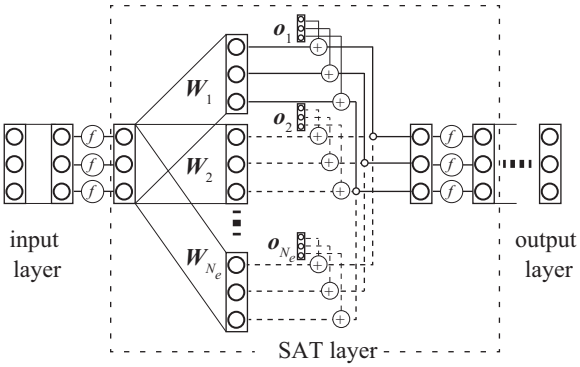

Fig. 1 An example of the conventional speaker-adaptive-training-DNN (SAT-DNN) [3].

3 話者行列の重みづけによる SAT-DNN

2 節に述べた従来法の問題点としては、5 節の実験で示すように、評価話者ごとに SAT 層の話者行列を全て切り替えるため、高い適応化効果が得られる反面、適応データ量が少ない場合には、学習がうまくいかないという課題がある。そこで本報では、一般的に一人当たりある程度の規模の音声データが得られやすいと考えられる、学習時に得られた学習話者に対する話者行列を有効に活用することを考える。SAT-DNN の学習時に、学習話者それぞれに対する話者行列が求まっているが、従来法ではそれが全く使われていない。実際、GMM では fMLLR 適応時に、認識時の特徴量変換行列を学習時の特徴量変換行列の重み付き和で表す手法 [4] が知られており、我々も学習データが少量の場合に、当該手法が有効であることを確認している [5, 6]。ここでもそれと同様に、学習時の話者行列の重み付き和で、認識時の話者行列を表す方法を提案する。

3.1 話者行列の適応

上述のように提案法では、学習時に学習話者 N_t 人から得ておいた N_t 個の話者行列 $\boldsymbol{W}_1 \ldots \boldsymbol{W}_{n_t} \ldots \boldsymbol{W}_{N_t}$

*Robust DNN speaker adaptation for a few adaptation utterances by weighting speaker matrices. by TACHIOKA, Yuuki and NARITA, Tomohiro (Mitsubishi Electric Corporation)

を元にそれらの重みパラメータを推定する。話者行列は、文献 [2, 3] に記載の方法で求めることができる。図 2 には提案の重みづけ SAT-DNN を示している。ここで、話者適応層の出力を $\boldsymbol{x}_{out}$ は

$$\boldsymbol{x}_{out} = \frac{1}{N_t}\sum_{n_t=1}^{N_t} w_{n_t}\boldsymbol{W}_{n_t}\boldsymbol{x}_{in} \tag{2}$$

のように各評価話者に対して N_t 個の重み w_{n_t} を推定する方法と

$$\boldsymbol{x}_{out} = \frac{1}{N_t}\sum_{n_t=1}^{N_t} \boldsymbol{w}_{n_t} \odot (\boldsymbol{W}_{n_t}\boldsymbol{x}_{in}) \tag{3}$$

のように各話者、各次元の重み $\boldsymbol{w}_{n_t}$ を推定する方法の 2 通り考えられる。ただしここで $\odot$ はベクトルの要素ごとの積である。

またこれらでは畳み込みネットワーク (CNN) における 'average pooling' のように各話者行列の出力の平均を取っているが、同じく CNN の 'max pooling' のように最大を取る方法も考えられる。例えば式 (2) は以下のようにもできる。

$$\boldsymbol{x}_{out} = \max_{n_t}\left[w_1\boldsymbol{W}_1\boldsymbol{x}_{in}, \ldots, w_{n_t}\boldsymbol{W}_{n_t}\boldsymbol{x}_{in}, \ldots\right] \tag{4}$$

ここで max は行ごとに対応する行ベクトルの内の最大の要素を返すとする。式 (3) も同様であり、以下予備的な実験で性能が高かった平均による方法で代表するが、すべて同じように最大を取る方法にも代えられる。

提案法では、このようにして DNN を話者適応する。$\boldsymbol{x}_{in} \in \mathbb{R}^{D_{in}}$、$\boldsymbol{x}_{out} \in \mathbb{R}^{D_{out}}$ とすると、文献 [3] は $\boldsymbol{W}_{n_e} \in \mathbb{R}^{D_{in}\times D_{out}}$ を適応していたので、$D_{in} \times D_{out}$ の数のパラメータの適応が必要だったのが、式 (2) を用いる場合には N_t 個、式 (3) を用いる場合には $N_t \times D_{out}$ 個のパラメータの適応に減らせる。学習話者数が膨大でないコーパスに対しては $N_t \ll D_{in}$ なので、適応すべきパラメータ数が大幅に削減でき、適応データが少ない場合の頑健性を向上させることができる。CSJ のような学習話者数が膨大なコーパスに対しては $N_t \ll D_{in}$ が成立しないので、パラメータ数は必ずしも減らない。これの問題は、4 節の話者行列のクラスタリングにより解決できる。

学習時には SAT 層で話者行列の行列ベクトル積の計算量が N_t 倍に増えるが、認識時には重み係数と話者行列の積を事前に計算しておけば、1 つの行列ベクトル積になるので、話者適応を行わない DNN と計算量・メモリ量は等しくなる。

3.2 オフセット適応も行った場合

3.1 節では、学習話者の話者行列に対する重みを推定したが、これに加えて、もしくはこれに代えて、オフセットを推定することも考えられる。このときも同様に、

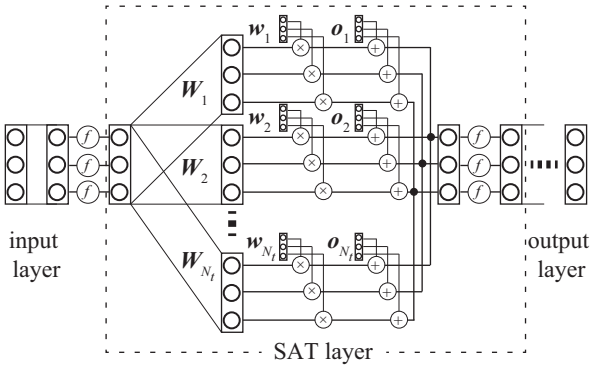

Fig. 2 An example of the proposed weighted SAT-DNN.

$$\boldsymbol{x}_{out} = \frac{1}{N_t}\sum_{n_t=1}^{N_t} (w_{n_t}\boldsymbol{W}_{n_t}\boldsymbol{x}_{in} + o_{n_t}) \tag{5}$$

のように各話者に対して、1 次元のオフセット o_n を推定する方法と

$$\boldsymbol{x}_{out} = \frac{1}{N_t}\sum_{n_t=1}^{N_t} (\boldsymbol{w}_{n_t} \odot (\boldsymbol{W}_{n_t}\boldsymbol{x}_{in}) + \boldsymbol{o}_{n_t}) \tag{6}$$

のように各話者に対して、$\boldsymbol{x}_{out}$ と同じ次元のオフセット $\boldsymbol{o}_n$ を推定する方法の 2 通り考えられる。重みパラメータ数は 3.1 節での場合と同じである。

4 学習話者をクラスタリングした話者クラスターの話者行列の重みを求める方法

3 節の方法は学習話者数 N_t が多くなり、$N_t > D_{in}$ となると、適応すべきパラメータ数が従来の SAT-DNN よりも増えてしまうという課題がある。これは不特定話者音声認識のために使う CSJ のような大規模音声コーパスを学習に使う際には問題となる。そこであらかじめ学習話者に対する N_t 個の話者行列を何等かの方法でクラスタリングして、M クラスの話者行列で代表する方法が考えられる。すなわち、学習が終了した段階で、N_t 個の話者行列をクラスタリングにより M 個の行列に減らしておくのである。この時 $M < N_t$ なので、パラメータ数を 3 節の方法から、さらに減らすことができる。クラスタリング手段としては、$\boldsymbol{W}_{n_t}$ 間の距離に基づく k-means クラスタリングや、$\boldsymbol{W}_{n_t}$ をベクトル化して $D_{in} \times D_{out}$ 行、N_t 列の行列に対してスペクトルクラスタリングを行う方法が考えられる。ここでは、クラスタリングされたクラスターのセントロイドを話者行列 $\boldsymbol{W}'_1 \ldots \boldsymbol{W}'_M$ として用いると、式 (7) のように、式 (5),(6) と同様

にして話者適応層を適応することができる。

$$\boldsymbol{x}_{out} = \frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{w}_m \odot (\boldsymbol{W}'_m \boldsymbol{x}_{in}) + \boldsymbol{o}_m) \quad (7)$$

5 騒音下音声認識実験 (第 3 回 CHiME チャレンジ)

5.1 実験条件

第 3 回 CHiME チャレンジ [7] において、提案手法の有効性を確認した。これは、発話が「ウォールストリート・ジャーナル」から採られている中語彙の騒音下音声認識タスクである。実データ（「Real」）およびシミュレーションデータ（「Sim」）の 2 種類のデータがある。それぞれのデータは、「バス」、「カフェ」、「歩行者天国」および「通り」の 4 環境からなる。以下に示す WER は、4 つの環境の平均 WER である。学習セットは、実データとシミュレーションデータそれぞれで、4 および 83 話者による 1,600 と 7,138 発話からなる。開発（dt）、および評価（et）セットは、実データとシミュレーションデータともに 4 話者によるそれぞれ 1,640 と 1,320 発話からなる。本報では、SN 比最大化ビームフォーマー [8] により強調された音声で評価する。音響特徴量は、0 次から 22 次のフィルターバンク (fbank) 特徴量とその Δ と ΔΔ 特徴量を使っている。

音響モデルは、学習セットにより学習し、パラメータを、開発セットの WER により調整した。表 1 に音声認識の設定を示す。DNN 学習には、Kaldi ツールキットの「nnet1」を使った。7 層の制約付きボルツマンマシンから始めて、各隠れ層にシグモイド活性化関数を持つ DNN を構築した。開発セットのクロスエントロピーの減少分が閾値以下であった場合には、学習率を初期の学習率 (0.008) から低減していく方法で学習した。9 隣接フレームの特徴量を入力し、各隠れ層あたりノードの数は 1,024 であった。

学習話者数 N_t が 83、ノード数 D_{in}, D_{out} が 1,024 なので、提案法により、推定すべきパラメータ数を 92%削減できる。ここではスペクトルクラスタリング

Table 1 Setup for the ASR systems.

Sampling freq.	16 kHz
Window length	25 ms
Window shift	10 ms
Features	0–22th fbanks + Δ + ΔΔ
HMM states	2,500 shared triphone states
# Gaussians	15,000
DNN nodes per layer	1024 nodes
DNN layer size	7 layers
Vocabulary size	5,000

Table 2 Correlation coefficients of weight parameters among adapted four speakers; F01 and F04 were female, whereas M03 and M04 were male speakers. Upper triangle shows the coefficients of Eq. (2)-type adaptation. Lower triangle shows those of Eq. (3)-type one.

	F01	F04	M03	M04
F01	1	0.806	-0.389	-0.658
F04	0.390	1	-0.439	-0.473
M03	0.117	0.147	1	0.858
M04	0.049	0.127	0.373	1

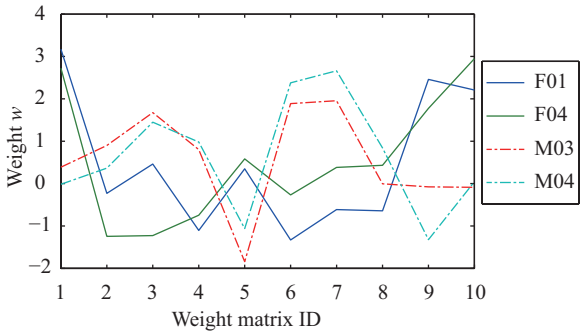

Fig. 3 Respective weights w_m in Eq. (7) per adaptation speaker for $M = 10$ weight matrices.

により、学習話者の話者行列を 10 クラスタにクラスタリングした方法および、[2] 同様話者の性別ごとに 2 クラスタとした方法、また学習話者全体を 1 クラスタとした方法の 3 通りを試した。10 クラスターの場合、推定すべきパラメータ数は従来法の 1%以下である。事前の調整により、SAT-DNN の学習は 5 エポック分とし、適応データに対する繰り返しは 1〜5 回までの内の開発セットで最も性能が高かった回数とした。

5.2 重み係数の考察

4 人の評価話者に対して、話者ごとに重み係数を 1 つ決める方法 (式 (2)) と話者・D_{out} 次元ごとに重み係数を決める方法 (式 (3)) の 2 種類の適応法を試した。重み係数がうまく学習されているか考察するため、表 2 に話者ごとの重み係数間の相関係数を示している。両方式ともに、同性間の相関係数は異性間に比べて高いことがわかる。異性間については、式 (3) の方法ではほぼ無相関、式 (2) の方法では負の相関となっている。これにより、重み係数には話者性が含まれていることが明らかとなった。図 3 に、10 の話者行列に対する重み係数を示す。同性間は同じような傾向を、異性間は異なる傾向を示していることが見て取れる。

5.3 適応発話が十分にある場合

表 3 に、式 (2) 型の適応と式 (3) 型の適応時の WER を比較している。適応なし (speaker independent: SI)

Table 3 Average WER [%] of the third CHiME challenge development set for the SAT-DNN when all utterances per adapted speaker were used for DNN adaptation. The performance of Eq. (2)-type and Eq. (3)-type adaptations (10 weights) was compared to that of the speaker independent (SI) model.

	Real	Sim
SI	12.83	9.38
Eq. (2)-type adaptation	12.77	10.05
Eq. (3)-type adaptation	11.38	8.93

Table 4 Average WER [%] of the third CHiME challenge development set (dt) and evaluation set (et) for the SAT-DNN when all 440 utterances per adapted speaker were used for DNN adaptation. 'per speaker' shows the conventional SAT-DNN result.

	dt		et	
	Real	Sim	Real	Sim
SI	12.83	9.38	25.94	14.57
per speaker	**11.22**	**8.56**	**24.06**	12.15
84 weights	11.55	9.10	24.98	11.79
+offsets	11.55	9.04	25.04	12.46
10 weights	11.38	8.93	24.99	11.30
+offsets	11.28	8.98	25.37	**11.73**
2 weights+offsets	11.88	9.32	26.82	12.38
1 weight+offset	12.07	9.15	27.60	12.85

に比べて、式 (2) 型の適応はほとんど改善が見られないが、式 (3) 型の適応では改善が見られる。これより、話者ごとに 1 つの重みでは不十分なことがわかった。以降は式 (3) のタイプの適応を行うこととする。

表 4 に、開発セットと評価セットの各話者の全 440 発話を適応に使った場合の結果を示す。この場合、各話者に十分な適応データがあるため、従来の SAT-DNN('per speaker') が評価セットの Sim 以外では、最も効果が出ており、提案法は適応の効果が少ない。十分に適応データがある場合には、パラメータ数を多くしても問題がないことがわかる。ただし通常一人の話者に 440 発話もの適応データを用意できることは稀である。提案法の中では、重み行列数は 10 の時が最も性能が良かった。

5.4 適応発話が少ない場合

表 5 に、適応発話数を 10 発話に減らした場合の WER を比較している。この場合、従来法 ('per speaker') は開発セットにおいて 0.1%程度の改善幅にとどまっている。これに対して、提案法は 0.5%程度の改善がみられる。重み行列数は 10 の時が最も性能が良かった。オフセットの適応は効果がある場合と無い場合があるが、逆効果になっていることはなかった。

Table 5 Average WER [%] of the SAT-DNN when only 10 utterances per adapted speaker were used for DNN adaptation.

	dt		et	
	Real	Sim	Real	Sim
SI	12.83	9.38	**25.94**	14.57
per speaker	12.74	9.27	26.96	13.63
84 weights	12.86	9.51	27.74	14.29
+offsets	12.71	9.39	27.55	13.95
10 weights	12.39	**8.91**	27.88	13.30
+offsets	**12.35**	**8.91**	27.51	**13.10**
2 weights+offsets	12.65	8.93	28.74	13.46
1 weight+offset	12.41	8.93	28.39	13.31

た。どちらの適応法ともに、評価セットの Real では適応がうまく働いていない。これはもともとベースラインの性能が低い、かつ発話量の少ない教師データを使って適応をしたため、誤った適応が行われたためと考えられる。

6 まとめ

SAT-DNN の問題点である推定すべきパラメータ数が多いため、適応発話が少ない場合に適応の効果が低くなるという問題を解決するために、学習話者で求めた話者行列の重み付き和で、適応話者の話者行列を表現する方法を提案した。第 3 回 CHiME チャレンジで適応発話を減らした実験を行い、提案法の有効性を示した。今後の課題としては、SAT-DNN と識別学習の組み合わせが挙げられる。

参考文献

[1] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Computer Speech and Language, **12**, 75–98 (1998).

[2] M. Delcroix, K. Kinoshita, T. Hori, and T. Nakatani, "Context adaptive deep neural networks for fast acoustic model adaptation," Proceedings of ICASSP, pp.4535–4539 (2015).

[3] T. Ochiai, S. Matsuda, H. Watanabe, X. Lu, C. Hori, and S. Katagiri, "Speaker adaptive training for deep neural networks embedding linear transformation networks," Proceedings of ICASSP, pp.4605–4609, IEEE (2015).

[4] D. Povey and K. Yao, "A basis representation of constrained MLLR transforms for robust adaptation," Computer Speech and Language, **26**, 35–51 (2012).

[5] Y. Tachioka, T. Narita, S. Watanabe, and F. Weninger, "Dual system combination approach for various reverberant environments," Proceedings of REVERB challenge workshop, pp.1–8 (2014).

[6] 金川裕紀，太刀岡勇気，石井純，"事前情報を利用した基底fMLLR のための重み係数推定法," 日本音響学会研究発表会講演論文集 (春季)，pp.41–42 (2016).

[7] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," Proceedings of ASRU, pp.504–511 (2015).

[8] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," Prodeedings of ICASSP, **1**, pp.41–45 (2007).