話者依存サブネットワークを用いた深層学習による多対一声質変換*

☆橋本哲弥, 柏木陽佑, 齋藤大輔, 広瀬啓吉, 峯松信明 (東大)

1 はじめに

近年，テキストから音声を合成する Text-to-Speech 変換技術が実用化され，カーナビゲーションシステムなど様々な用途で利用されている．しかし，それらの音声は得られる品質は高いものの，特定の話者が特定の発話様式 (一般には朗読調) で発声した多量の音声を前提とするもので，話者 (声質)，あるいは発話様式が異なる音声を合成するためには，そのような音声を新たに用意しなければならないという問題点がある．こうした問題を解決する技術として声質変換が存在する．

声質変換とは, 入力音声と出力音声の特徴の対応を記述したモデルによって，入力音声を言語情報を損なうこと無く他の声質や他の発話様式に変換して出力する技術である．基本的には入力音声と出力音声の特徴量空間上でのマッピングを構築するというタスクと考えられ, 統計的な変換モデルによる実装が多く試みられている [1]．変換モデルの構築の際の問題としては, 学習の際に入力話者 (以下では声質変換を念頭に置いて議論を進める) と出力話者の同一内容発声のパラレルコーパスが必要となるという点がある．そのため，新しい話者間の変換を扱う場合にはそれに対応した新しいパラレルコーパスを用意することになる．声質変換の実用的な側面を考えると, 変換元・変換先の話者が変わる度にそれに対応したコーパスを用意することはコストが大きいため, より少量のデータで話者に対して柔軟な変換を行える話者適応型のモデル構築が望ましい. このため, 現在は一対多変換・多対一変換・多対多変換を少量のパラレルデータ，あるいはパラレルデータなしに適応的に行う手法が研究されている. 声質変換では音響空間の統計的表現である Gaussian Mixture Model (GMM) が一般的に用いられており，大量の事前登録話者のデータを用いて話者空間をより少ないパラメータで表現する固有声変換 [2] などの手法が提案されている．

一方で，近年画像認識や音声認識などの分野で高い精度を実現している Deep Learning という手法がある [5]．Deep Learning は，特徴量抽出器を多層化したネットワークを用いた手法の総称であり，特徴量抽出 (pre-training) に用いる手法や教師あり学習 (fine-tuning) の有無によって様々な種類がある．声質変換においても Deep Learning を用いた手法が提案されており，GMM による変換よりも精度の高い変換を実現したという研究結果も存在する [3]．しかし，Deep Learning では，各層，各ノードがどのような処理を行っているかが不明瞭であるため，GMM のように柔軟なパラメータの適応が行いづらいという問題がある．

そこで，本研究では，多言語認識における松田らの手法 [4] を参考に，Deep Learning の一種である Deep Neural Network(DNN) に対して，変換先の話者毎のサブネットワークを導入した，多対一型の変換手法を提案する．松田らの手法では，Neural Network の層が深くなるにつれて入力情報が複雑に集約されていくという性質から，ネットワークの入力側に近い層 (浅い層) においては単純な処理を，出力側に近い層 (深い層) においては複雑な処理を行っているという仮定をおいている．本研究では，この仮定を基に，ネットワークを前半と後半に分割し，前半のネットワークはすべての話者に対して共通のものを用い，後半のネットワークを出力話者毎に用意して学習を行うことで，前半のサブネットワークには入力及び出力話者に依存しない特徴量抽出のような処理を，後半のサブネットワークには出力話者に依存する特徴量変換のような処理をそれぞれ集約することを目的としている．

実験では，初めに特徴抽出の段階で用いる話者数によって変換精度がにどのような影響が出るかを確認する．次に，提案手法による声質変換，従来の DNN による声質変換，GMM による声質変換の間で精度の比較を行い，提案手法の有効性を示す．最後に，学習時に用いていない未知の入力話者の特徴量を入力として各手法による変換を行い，変換精度を比較する．

2 Deep Neural Network

入力特徴量を出力特徴量に直接変換するモデルの 1 つとして，Artificial Neural Network がある．Artificial Neural Network は，人間のニューロン (神経細胞) を数式的に模した Artificial Neuron を用いてネットワーク構造を構築したモデルのことをいう． Artificial Neuron は数式化すると，以下の式 (1) で表される．

*Many-to-one voice conversion using speaker dependent subnetwork based on deep learning. by HASHIMOTO, Tetsuya. KASHIWAGI, Yousuke. SAITO, Daisuke. HIROSE, Keikichi. and MINEMATSU, Nobuaki. (The University of Tokyo)

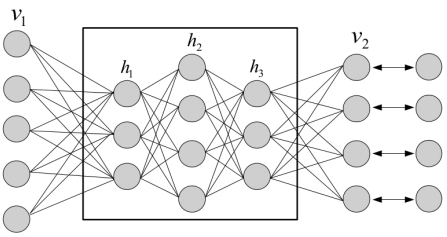

Fig. 1　Multi-layer perceptron

$$y = f(\sum_{i=1}^{N}(w_i x_i + b_i)) \tag{1}$$

ここで，x_i は入力信号を表し，w_i は各入力信号にかけられる重み，y は出力信号をそれぞれ表す．$f(\cdot)$ は閾値関数であり，シグモイド関数などが用いられる．また，b はバイアス項，N は入力の次元数をそれぞれ表す．これは他のニューロンからの入力信号がシナプスを通して現在のニューロンに伝達し，それが閾値を越えたとき現在のニューロンが発火し，結合している他のニューロンへ信号を伝達するという仕組みをモデル化している．この Artificial Neuron を複数接続し，ネットワーク構造を成したものを Artificial Neural Network (以下 Neural Network) と呼ぶ．Neural Network の 1 つであり Deep Neural Network と同じ構造を持つ多層パーセプトロンの例を Fig. 1 に示す．多層パーセプトロンは，複数の Artificial Neuron を層状に配置し，隣接する層との間で結合したものである．h_i は隠れ層，v_1 と v_2 は可視層と呼ばれ，v_1 と v_2 がそれぞれ入力層と出力層となる．このモデルに対して，入力特徴量と正解データ (とする特徴量) の組からなるパラレルコーパスを用いて，入力に対する出力と正解データの間の誤差を計算し，その誤差の値によって出力層側から順に各層の重み w_{ij}，バイアス b_j を更新することで学習を行う．この学習を誤差逆伝搬法という．DNN においては，この誤差逆伝搬法によるパラメータの学習を行うステップを fine-tuning と呼ぶ．

Neural Network を用いた声質変換としては [6][7] などがあり，声質変換において一般的に用いられている Gaussian Mixture Model (GMM) による変換精度を上回る結果が出ていると報告されている．

Neural Network の問題点としては，各層のノード数，または層の数を増やすことで，任意の関数を再現することができる一方で，層を増やす程に，誤差逆伝搬法によるパラメータの更新が入力側に近い層まで伝達しづらくなるという点がある．

DNN ではこの問題を改善するために，誤差逆伝搬法を行う前に各パラメータの初期値を計算する処理を行う．この処理を pre-training といい，本研究では Restricted Boltzmann Machine (RBM) を用いる．

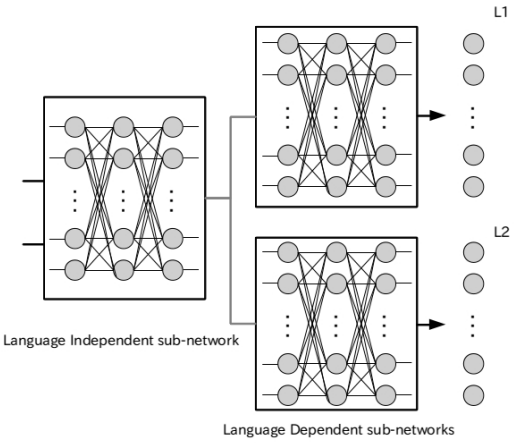

Fig. 2　Connections of language-independent sub-network and language-dependent sub-networks

3　言語非依存サブネットワークによる多言語音素認識

松田らは DNN によって多言語音素認識を行う際に，言語に依存しない処理をサブネットワークに集約する手法を提案している [4]．手法の概要を Fig. 2 に示す．

この手法では，Neural Network の層が深くなるにつれて入力情報が複雑な処理を経た結果，より集約的かつコンパクトな表現になるという性質から，ネットワークの浅い層においては周波数などの基本的な特徴量抽出処理を，深い層においては言語に依存した複雑な特徴量抽出処理を行っているという仮定をおいている．この仮定を基にして学習の際に Fig. 2 のように，言語に関わらず共通のものを用いる言語非依存サブネットワークと，出力側の言語によって異なる言語依存サブネットワークを用いる．これにより，言語に依存しない処理が入力側の言語非依存サブネットワークに集約され，言語に依存した処理が出力側の各言語のサブネットワークに集約される．言語依存・非依存サブネットワークを用いた音素認識の精度を，ベースラインとして通常の DNN の精度と比較したところ，英語・中国語ではベースラインを上回り，日本語においてもほぼ同程度の結果が得られている．また，この言語非依存サブネットワークのパラメータを固定し，新たな言語の音声に対して言語非依存サブネットワークからの出力によって pre-training と fine-tuning を行った結果，ベースラインよりもわずかに高い精度が得られている．松田らはこれらの結

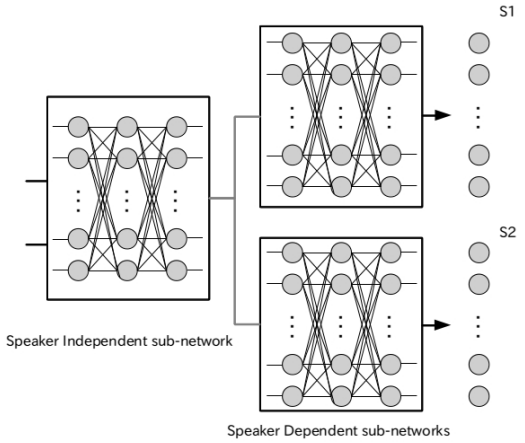

Fig. 3 Configuration of the proposed method: Deep neural network with speaker independent and dependent sub-networks.

果から，言語非依存サブネットワークが学習に用いていない新たな言語にも適用可能な汎用ネットワークになっているとしている．

4 提案手法

本研究では，DNN に対して変換先の話者 (目標話者) 毎に話者依存サブネットワークを導入することで，多対一声質変換を行う．提案手法の概要を Fig. 3 に示す．提案手法では，通常の DNN と同様に pre-training を行った後，ネットワークを 2 つに分割する．分割した前半のネットワークを話者非依存層 (Speaker Independent Layer: SI 層) とし，入力話者・出力話者に依らず同一のものを使用する．対して，分割した後半のネットワークを学習データ中の出力話者の数だけ複製したサブネットワークを話者依存層 (Speaker Dependent Layer: SD 層) とする．fine-tuning の際には，話者 (S1, S2, ...) に対するパラレルデータが存在しているとすると，話者 S1 への変換を学習する時は SI 層に話者 S1 に対応した SD 層を結合したネットワークを用い，話者 S2 への変換を学習する際には同様に SD 層を切り替えていく．このように，目標話者に依らず SI 層は共通のネットワークを使用し，SD 層を目標話者に応じて切り替えていくことで，SI 層に話者性の正規化または除去のような処理が集約されることが期待できる．

5 実験

実験として，提案手法と従来 2 手法 (GMM, DNN) の間で声質変換の精度を比較した．また，予備検討として，一対一声質変換を行う DNN の学習において，pre-training に用いる話者の数が変換精度にどのような影響を与えるかの確認を行った．

Table 1 Mel-CD for each pre-training condition

	Mel-CD[dB]
setA	4.538
setB	4.449
setC	4.256

初めに，予備検討について述べる．pre-training に用いる学習データを setA(M1)，setB(M1, M2)，setC(M1, M2, M3) とし，それぞれを基に話者 M1 から話者 M3 への fine-tuning を行い，テストデータに対する変換精度を式 (2) で表されるメルケプストラム歪み (Mel cepstral distortion:Mel-CD) によって評価する．

$$\text{MelCD[dB]} = \frac{10}{\ln 10}\sqrt{2\Sigma_{d=1}^{24}(mc_d - \bar{mc}_d)^2} \qquad (2)$$

ここで mc_d, $\bar{mc}_d$ はそれぞれ正解データ，変換データの d 次元目のメルケプストラムを表す．学習データとしては ATR 音声データベース [8] から選択した男性話者 3 人 (M1, M2, M3) の各 50 文を使用し，fine-tuning の際には入力話者と出力話者のデータの間に DP-matching によってアライメントを取り，パラレルコーパスとする．評価では，DP-matching によってアライメントを取った学習データとは別の 50 文をテストデータとして用い，話者 M1 から話者 M3 への変換を行う．特徴量としては音声データから STRAIGHT 分析によって得られたメルケプストラムのパワーを除いた 24 次元を使用する．DNN は 6 層とし，各層のノード数は入力層を 24 次元，隠れ層を各 1024 次元，出力層を 24 次元とした．

予備検討の結果を Table 1 に示す．Table 1 から，pre-training に用いる話者を増やすと，一対一の変換の精度が向上していることが分かる．これは，pre-training が特徴量抽出の処理であるため，変換対象の話者以外のデータを用いることによる入力話者・出力話者へのミスマッチの増加よりも入力への汎化性能の向上の効果が大きいためであると考えられる．

次に，提案手法と従来手法 (GMM, DNN) の間で話者 M1 から話者 M3 への声質変換の精度を比較する．GMM は全共分散型を使用し，学習時に入力話者 M1 と出力話者 M3 のパラレルコーパスのみを用いたものを GMM1，入力話者 M1 と出力話者 M3，入力話者 M2 と出力話者 M3 の 2 つのパラレルコーパスを用いたものを GMM2 とした．また，GMM は入力データによって最適混合数が異なるため，64 混合と 32 混合の両方で学習を行い，精度の高かった方の値を示す．DNN に対しても同様に 1 対 1 の学習を行ったものと，2 対 1 の学習を行ったもの (DNN1, DNN2) を

Table 2　Results of objective evaluation

	M1-M3		M1-M3
GMM1	4.313	subnet1:5	4.214
GMM2	4.571	subnet2:4	4.221
DNN1	4.449	subnet3:3	4.209
DNN2	4.252	subnet4:2	4.209
		subnet5:1	4.211

Table 3　Results of objective evaluation for an unknown speaker

	M4-M3		M4-M3
GMM1	4.767	subnet1:5	4.433
DNN2	4.393	subnet2:4	4.448
		subnet3:3	4.462
		subnet4:2	4.453
		subnet5:1	4.449

用いる．この際，各種条件は予備検討において pre-training に setB(M1, M2) を用いたものと同様とする．提案手法 (subnetDNN) では，話者 (M1, M2, M3) が入力と出力が同一話者にならない組み合わせをとり，その全てに対して学習を行う．pre-training に対しても setC(M1, M2, M3) を使用し，fine-tuning の際には 6 層のネットワークに対して SI 層と SD 層の割合をそれぞれ 1:5 から 5:1 まで変動し，結果を確認する (subnet1:5 から subnet5:1)．DNN 部分のその他の条件は予備検討と同様とする．

実験の結果を Table 2 に示す．Table 2 から，提案手法による変換精度が，既存手法である GMM1/2, DNN1/2 を上回っていることが分かる．これは，提案手法は，学習に用いる話者のデータを入力・出力を固定せず，様々な組み合わせで用いることができるため，実質的により多くのデータを用いて学習した場合と同等の効果が得られているためと考えられる．

また，学習に用いていない未知話者 M4 に対する話者 M3 への変換精度を，Table 2 での GMM，DNN において最も精度が高かったもの (GMM1, DNN2) と，提案手法のそれぞれによって比較を行った．

Table 3 から，未知話者に対する変換では，DNN を用いた手法が GMM1 を上回る一方で，提案手法が通常の DNN よりも低い精度となっていることが分かる．また, 提案手法の中では subnet1:5 が最も精度が高くなっている．このことから，未知話者に対する変換においては出力話者への変換である SD 層の効果が大きく，また，SI 層が未知話者に対して有効に働いていないとも考えられる．この原因として，3 話者 (M1, M2, M3) それぞれを入力話者，出力話者とするという他 2 手法と比べて多くの組み合わせを学習している一方で，学習に用いる話者数が 3 話者では十分でなく話者非依存層が過学習を起こし，未知話者に対する汎化性能が下がっているということが考えられる．

6　おわりに

本研究では，より柔軟な多対一声質変換の実現を目的としたサブネットワーク型 Deep Neural Network の声質変換を提案した．実験の結果，pre-training に複数話者を用いることによる変換精度の向上と，提案手法による精度の向上の両方が確認された．しかし，学習に用いていない未知話者に対する変換においては，通常の DNN による変換よりも精度が低くなっていた．原因としては，学習に用いた話者の数が十分で無く，且つ既存手法に比べてより多くの組み合わせによる学習を行っているため，学習に用いた話者に対して過学習を起こしているということが考えられる．現在は，このような問題の検証のため，pre-training と fine-tuning の両処理においてより多くの話者を用いた場合に提案手法の精度がどのようになるかを確認することを検討している．

参考文献

[1] 齋藤大輔, 他. “話者空間のテンソル表現に基づく任意話者声質変換” Technical report of IEICE 2011, pp.7-12, 2011.

[2] Y. Ohtani *et al.* “Adaptive training for voice conversion based on eigenvoices”, IEICE TRANS. INF. &SYST., VOL.E93-D, NO.6, pp.1589-1598, 2010.

[3] 中鹿亘, 他. “Deep Belief Nets による低次元空間表現を用いた声質変換の検討”, 日本音響学会春季研究発表会講演論文集, 3-P-46b, pp.517-520, 2013-03.

[4] S. Matsuda, X. Lu, and H. Kashioka. “Automatic localization of a language-independent sub-network on deep neural networks trained by multi-lingual speech” ICASSP 2013, pp.7359-7362, 2013.

[5] G. Hinton *et al.* “Deep neural networks for acoustic modeling in speech recognition,” IEEE Signal Processing Magazine, pp.82-97, 2012.

[6] S. Desai *et al.* “Voice conversion using artificial neural networks”, ICASSP 04, pp.3893-3896, 2004.

[7] Z. Chen, and L. H. Zhang, “A ANN based high quality method for voice conversion”, WiCOM, pp.1-4, 2010.

[8] A. Kurematsu *et al.* “ATR Japanese speech-database as a tool of speech recognition and synthesis”, Speech Communication, vol.9, pp.357-363, 1990.