Vol.2018-SLP-123 No.8
2018/7/27

深層学習に基づく特徴量変換を用いた咽喉マイクの音声認識

鈴木貴仁[†1]　緒方淳[†2]　綱川隆司[†1]　西田昌史[†1]　西村雅史[†1]

概要：咽喉マイクはピエゾ素子によって皮膚の振動を受け取るマイクで一般的な空気の振動を受け取るマイクよりも外部騒音の影響を低減できる．しかしながら，咽喉マイクは一般的なマイク（気導マイク）と音響特性が大きく異なるため通常の音響モデルでは認識精度が低下する．気導マイクと咽喉マイクの音響ミスマッチを低減するため，本研究では咽喉マイク音声の MFCC から気導マイク音声のボトルネック特徴量への DNN に基づく特徴量変換を提案する．また，LSTM に基づく時系列を考慮した特徴量変換によって変換精度の更なる改善を図った．大語彙音声認識タスクによる評価を行った結果，咽喉マイク音声を気導マイク音声で学習した音響モデルで認識する方法と比べて，LSTM を用いた提案手法によって CER で約 46.8%(74.6% → 39.7%)の改善が得られた．

キーワード：咽喉マイク，音声認識，特徴量変換，深層学習

Throat Microphone Speech Recognition Using Deep Learning Based Feature Transformation

TAKAHITO SUZUKI[†1]　JUN OGATA[†2]　TAKASHI TSUNAKAWA[†1]
MASAFUMI NISHIDA[†1]　MASAFUMI NISHIMURA[†1]

Abstract：Throat microphones are more robust to environmental noises than usual acoustic microphones because they detect speech signals through skin vibrations rather than by air transmission. Throat microphones, however, cannot be used in conventional speech recognition systems because their acoustic characteristics are much different from those of the acoustic microphones. In this study, we propose a deep neural network (DNN)-based feature transformation method for throat microphone speech recognition. To utilize a large amount of training data recorded by acoustic microphones and effectively reduce the acoustic mismatch between the throat and acoustic microphones, we tried to use the bottleneck features to mediate between them. Moreover, we also propose a long short-time memory (LSTM)-based feature transformation method. Evaluation results for a large-vocabulary speech recognition task of Japanese free conversation revealed that the proposed method using LSTM had a 46.8% lower character error rate (74.6% → 39.7%) than the typical MFCC system trained from the acoustic microphone data.

Keywords：Throat microphone，Speech recognition，Feature transformation，Deep learning

1. はじめに

近年，高い識別性能を持つディープニューラルネットワーク（deep neural network; DNN）が音声認識に取り入れられるようになり，英語の電話会話音声認識（Switchboard）では人間と同等の認識精度に達したとの報告がされている[1]．しかしながら，非定常な高雑音環境下での認識精度の劣化を抑えることは未だに課題であり，さらなる改善が必要である．外部雑音の影響を抑える方法として咽喉付近の皮膚から直接振動を受け取る咽喉マイク（図 1）を用いて音声を収録する方法がある[2][3]. 咽喉マイクは空気の振動を受け取る一般的な気導マイクと比べて外部雑音に頑健である．しかし，咽喉マイクと気導マイクとの間には音響特性に大きな差があり，気導マイク音声で学習された一般的な音響モデルではモデルのミスマッチのために咽喉マイク音声の認識精度が低下する．さらに，咽喉マイク音声はデータ量が少なく，咽喉マイク音声だけで十分な精度を持つ音響モデルを学習させることが困難である．

図 1 咽喉マイク

咽喉マイクと気導マイクとの音響ミスマッチを解消するために様々な手法が提案されてきた[4]-[6]. Lin ら[6]は咽喉マイク音声の MFCC（mel-frequency cepstral coefficient）

†1 静岡大学
Shizuoka University
†2 産業総合技術研究所
National Institute of Advanced Industrial Science and Technology

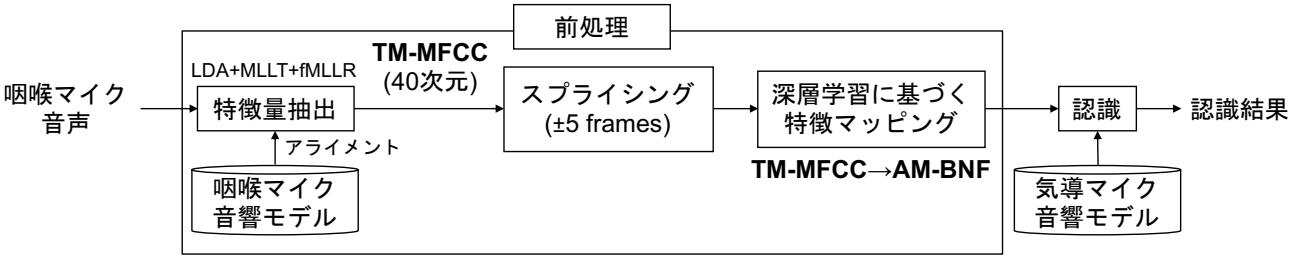

図 2 咽喉マイク音声認識システムの構成図（TM : 咽喉マイク, AM : 気導マイク, BNF : ボトルネック特徴量）

から気導マイク音声の MFCC への深層学習に基づく特徴量変換によって音響ミスマッチを抑える手法を提案し，気導マイク音声で学習された音響モデルでの咽喉マイク音声の認識精度を改善した．一方，遠隔発話音声認識タスクにおいて Hiwaman ら[7]は接話マイク(individual head microphone; IHM)と話者から離れた単一のマイク（single distant microphone; SDM）との音響ミスマッチを抑えるために SDM の MFCC 特徴量空間から IHM のボトルネック特徴量空間への特徴マッピングを DNN で学習する手法を提案している．ボトルネック特徴量(bottleneck feature; BNF)は入力特徴量の音素を識別するように学習した DNN のボトルネック層から抽出した特徴量で音素識別の本質的な情報が抽出できるとされており，従来の MFCC といった特徴量よりも音素識別に有効な特徴量の抽出が期待出来る．さらに Hiwaman ら[7]は特徴マッピングのための DNN の重みを IHM の BNF を抽出するための DNN の重みで初期化し，SDM の MFCC を入力信号，IHM の BNF を教師信号としてファインチューニングを行った．Das ら[8]は英語のアライメントを用いて学習した DNN を少量のトルコ語のアライメントを用いて再学習した DNN はランダムに初期化して学習した DNN よりもトルコ語の音素識別の性能が高くなったと報告しており，Hiwaman らが[7]で行ったように，DNN の初期値を工夫することで特徴マッピングの性能の改善が期待できる．

本研究では，遠隔発話音声認識において提案された BNF を介する特徴マッピング[7]に着目し，これを咽喉マイクの音声認識タスクに応用する．具体的には，本研究では咽喉マイクの MFCC から気導マイクの BNF への深層学習に基づく特徴マッピングを提案する．特徴マッピングは音声認識の前処理として行い，続く認識においては大量の気導マイクの BNF で学習した GMM-HMM 音響モデルを用いて認識を行う．さらに，特徴マッピングの DNN の重みを接話マイクの BNF を抽出する DNN の重みで初期化してファインチューニングを行うことで特徴マッピングの性能改善を図る．また，Lin ら[6]は LSTM（long short-term memory）に基づく時系列を考慮した特徴マッピングも提案しており，DNN よりもマッピング精度が高かったと報告している．そこで，LSTM による時系列を考慮した BNF を介する特徴マッピングに関しても検討を行ったので報告する．

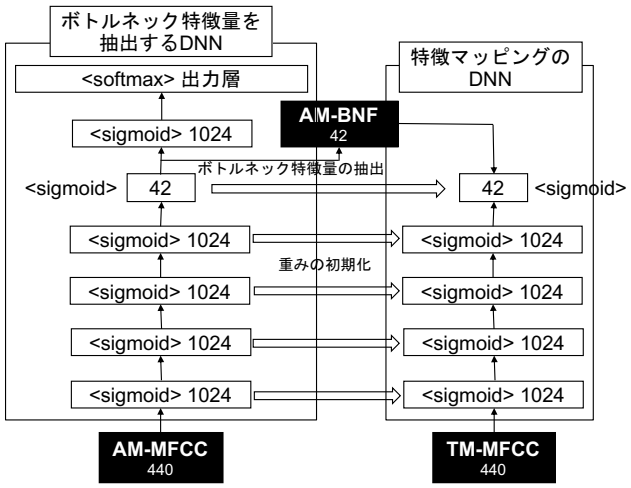

図 3 DNN の構造（全て全結合層．左 : ボトルネック特徴量を抽出するための DNN の構造，右 : 特徴マッピングのための DNN の構造）

本稿の構成を以下に示す．第 2 節では咽喉マイクの音声認識のための深層学習に基づく特徴マッピングの学習方法を述べる．第 3 節では大語彙音声認識タスクによる実験の方法と結果を述べる．最後に第 4 節では結論と今後の展望について述べる．

2. 提案手法

2.1 音声認識システムの全体像

咽喉マイクの音声認識システムの構成図を図 2 に示す．まず，入力した咽喉マイク(throat microphone; TM)音声から CMN（cepstrum mean normalization）を適用した 13 次元の MFCC を抽出してその前後 4 フレームずつ結合した後，それを LDA（linear discriminant analysis）で 40 次元に圧縮して MLLT（maximum likelihood linear transformation）を適用した特徴量を抽出する．次に抽出した特徴量に fMLLR（feature-space maximum likelihood linear regression）を適用する．本研究ではこの咽喉マイク音声の 40 次元の特徴量を TM-MFCC と呼ぶこととする．咽喉マイク音声のアライメントは咽喉マイクのみで学習した GMM-HMM（gaussian mixture model – hidden markov model）音響モデルを用いて推定する．そして，TM-MFCC を前後 5 フレームずつ結合した 440 次元の特徴量を DNN に入力して気導マイク

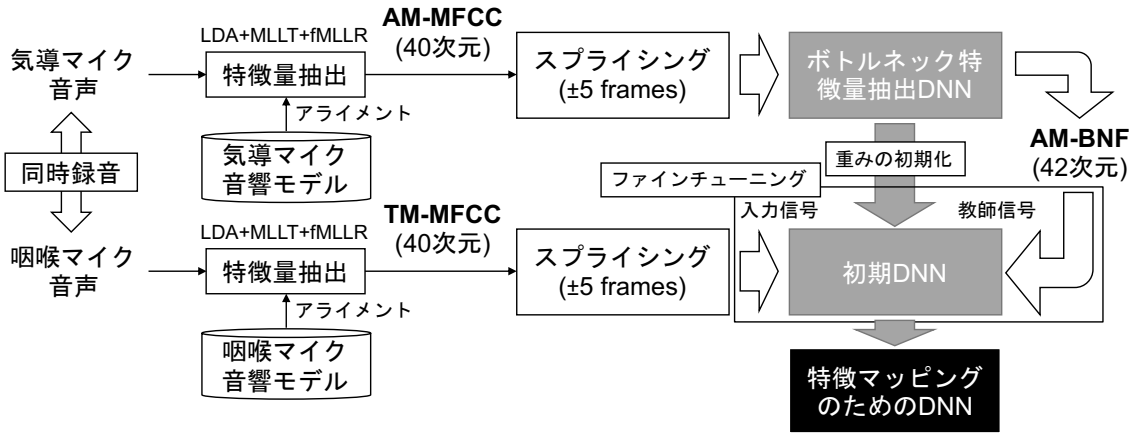

図 4 ボトルネック特徴量抽出の DNN の重みを活用した DNN に基づく特徴マッピングの学習方法

(acoustic microphone; AM)の BNF 空間に向けた特徴マッピングを行う．最後に，推定した特徴量を気導マイクの BNF で学習した GMM-HMM 音響モデルを持つ音声認識システムに入力して認識結果を得る．

2.2 ボトルネック特徴量抽出器の学習

BNF を抽出する DNN の構造を図 3-左に示す. TM-MFCC と同様にして気導マイク音声から fMLLR が適用された 40 次元の MFCC（AM-MFCC）を抽出する．AM-MFCC を前後 5 フレームずつ結合した 440 次元を入力信号，対応するアライメントを教師信号として入力特徴量の音素を識別するように学習を行う．気導マイクのアライメントは気導マイクのみで学習した GMM-HMM 音響モデルを用いて推定する．学習済みの DNN に気導マイクの特徴量を入力してボトルネック層から 42 次元の BNF（AM-BNF）を抽出する．

2.3 特徴マッピングの学習

本研究では特徴マッピングを学習するネットワークの構造として DNN と LSTM を検討する．まず，DNN による特徴マッピングの学習方法を図 4 に示す. 気導マイクと咽喉マイクで同時録音したパラレルデータを用意し，それぞれから AM-BNF と TM-MFCC を抽出する．特徴マッピングの DNN は図 3-左に示した気導マイクのボトルネック特徴量を抽出する DNN のボトルネック層より後段の層を外した構造（図 3-右）であり，ボトルネック特徴量を抽出する DNN の重みで初期化する．TM-MFCC を前後 5 フレームずつ結合した 440 次元の特徴量を入力信号，対応する 42 次元の CMN を適用した AM-BNF を教師信号として DNN をファインチューニングする

LSTM による特徴マッピングも DNN と同様にパラレルデータから抽出した AM-BNF と TM-MFCC を用いて学習を行う．ただし，ボトルネック特徴量を抽出する DNN と構造が異なるので, LSTM の重みはランダムに初期化する．LSTM の構造を図 5 に示す．入力信号は DNN の場合とは異なり, 40 次元の TM-MFCC に手前 6 フレームを時系列データとして結合したものを入力する．教師信号は DNN の場合と同様に 42 次元の CMN を適用した AM-BNF である．

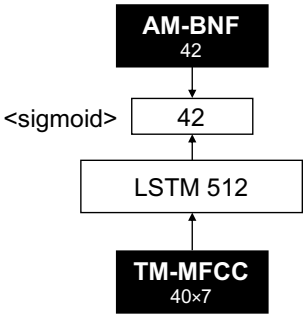

図 5 特徴マッピングの LSTM の構造.
出力層は全結合層である

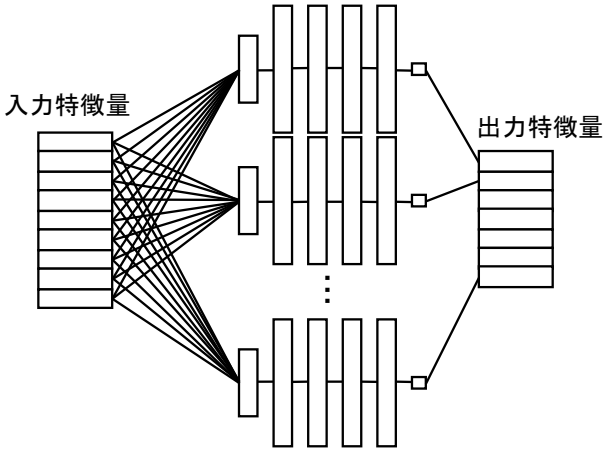

図 6 出力特徴量の次元ごとに特徴マッピングを学習する DNN

2.4 出力特徴量の次元ごとに DNN を学習する特徴マッピング

Lin ら[6]は出力特徴量の次元ごとに特徴マッピングの

DNN を学習することで一つの DNN で出力特徴量の全ての次元を推定する方法よりも特徴マッピングの精度が向上したと報告している．そこで，本研究でも特徴マッピングの学習方法として出力特徴量の次元ごとに特徴マッピングの DNN を学習する方法の検討も行う. 出力特徴量の次元ごとに特徴マッピングを学習する DNN の構造を図 6 に示す. 出力特徴量の次元数分 DNN があり, 各 DNN の入力層と隠れ層の構造は図 3-右に示した構造と同じであるが, 出力層は対応する次元の特徴量を出力するのでそのユニット数は 1 である．各 DNN は個別に学習が行われ，入力信号には 440 次元の咽喉マイクの MFCC を与え，教師信号には入力信号に対応する AM-BNF の 42 次元の うちのある 1 次元の特徴量を与える.

LSTM の場合も各次元の LSTM の出力層以外の構造は図 5 と同じで，出力層のユニット数は 1 である．入力信号は 40 次元の TM-MFCC に手前 6 フレームを時系列データとして結合したもので, 教師信号は DNN の場合と同様に AM-BNF の 42 次元のうちのある 1 次元である.

3. 実験

3.1 実験方法

特徴マッピングの学習に用いる気導マイクと接話マイクのパラレルデータはマルチトラックレコーダー(ZOOM R24)を使用して咽喉マイク(NANZU　SH-12iK)とピンマイク（Sony EXM -CS3）で同時に録音して収集した．音声データのサンプリング周波数は 16,000Hz である．男性話者 8 名から簡易防音室にて ATR503 音素バランス文の読み上げ音声（約 3 時間）を収録した．このパラレルデータの咽喉マイク音声は咽喉マイクの音響モデルの学習にも使用した.

気導マイクの GMM-HMM 音響モデルの学習データとして日本語話し言葉コーパス（corpus of spontaneous Japanese; CSJ) より約 240 時間, ボトルネック特徴量を抽出する DNN の学習に同じく CSJ より約 114 時間を使用した.

テストデータとして男性話者 10 名による合計約 19 分の自由発話の音声を使用した．自由発話の内容はグループディスカッションで 3 人グループを作って各メンバーがそれぞれ異なる資料を予め読んで勉強しておき，勉強したことを他のメンバーに説明するという内容である．収録は 1 グループごとに簡易防音室で行われたため外部雑音の影響は少ない.

特徴量の抽出，音響モデルの学習と認識実験には Kaldi [9]を使用した．ボトルネック特徴量を抽出する DNN の学習は Kaldi+PDNN [10]を使用した．DNN のボトルネック層までの 4 層は SdA（stacked denoising auto-encorders）による事前学習を行った．特徴マッピングのための DNN の学習には Keras を使用した．特徴マッピングのための DNN, LSTM の学習ではミニバッチサイズは 4096，エポック数は 100 とし，最適化手法には Adam（学習率 : 0.001）を用いた．言語モデルは 3-gram で CSJ の書き起こし文より作成した.

表 1　特徴マッピングを用いない従来手法と提案手法の認識実験結果（CER : 文字誤り率）

音響モデル	入力特徴量	CER
(1) AM GMM-HMM	MFCC	74.6%
(2) AM Tandem	BNF	96.6%
(3) TM GMM-HMM	MFCC	52.4%
(4) TM Tandem	BNF	48.9%
(5) AM Tandem	特徴マッピングした特徴量（提案手法-DNN）	42.2%
(6) AM Tandem	特徴マッピングした特徴量（提案手法-LSTM）	39.7%

3.2 従来のシステムとの比較

従来の特徴マッピングを用いない音声認識システムとの比較を行うため，以下に示す 6 つの認識実験を行った.

(1) 咽喉マイクの MFCC を入力特徴量として気導マイクの MFCC で学習した GMM-HMM（AM GMM-HMM）を音響モデルとしたシステム
(2) 咽喉マイクの BNF を入力特徴量として気導マイクの BNF で学習した GMM-HMM（AM Tandem）を音響モデルとしたシステム
(3) 咽喉マイクの MFCC を入力特徴量として咽喉マイクの MFCC で学習した GMM-HMM を音響モデル（TM-GMM-HMM）としたシステム
(4) 咽喉マイクの BNF を入力特徴量として咽喉マイクの BNF で学習した GMM-HMM（TM Tandem）を音響モデルとしたシステム
(5) DNN に基づく特徴マッピングによって推定した特徴量を入力特徴量として気導マイクの BNF で学習した GMM-HMM（AM Tandem）を音響モデルとしたシステム（提案手法-DNN）
(6) LSTM に基づく特徴マッピングによって推定した特徴量を入力特徴量として気導マイクの BNF で学習した GMM-HMM（AM Tandem）を音響モデルとしたシステム（提案手法-LSTM）

それぞれのシステムで咽喉マイク音声認識の評価を行った．なお，(2)の咽喉マイクの BNF は気導マイク音声で学習した DNN から抽出しており，(4)の咽喉マイクの BNF は咽喉マイク音声で学習した DNN から抽出している．この実験において特徴マッピングの DNN はランダムに初期

表 2　咽喉マイクの MFCC または BNF から気導マイクの MFCC または BNF への特徴マッピングを用いた認識実験結果（CER : 文字誤り率）

マッピング方法	CER（DNN）	CER（LSTM）
(1) TM-MFCC→AM-MFCC	48.0%	46.5%
(2) TM-BNF→AM-MFCC	59.1%	51.1%
(3) TM-BNF→AM-BNF	51.1%	45.2%
(4) TM-MFCC→AM-BNF（提案手法）	42.2%	39.7%

化されている．

実験結果を表 1 に示す．(1)を見ると CER（文字誤り率）が 74.6%と高く，気導マイクと咽喉マイクの音響ミスマッチが大きいことが確認できた．さらに，(2)を見るとほとんど認識できておらず, AM-MFCC で学習した DNN では TM-MFCC を AM-BNF の特徴量空間へ正しく変換できないことがわかった．一方で提案手法の(5)，(6)では CER が 40%付近まで改善しており，特徴マッピングによって AM-BNF とのミスマッチを抑えた特徴量を推定できていることを確認できた．大量の気導マイクで学習した音響モデルは 3 時間程度の咽喉マイク音声で学習した音響モデルよりも音素の識別性能が高く，結果として提案手法の(5)，(6)は咽喉マイクのみで学習した(3)や(4)のモデルよりも認識精度が高くなったといえる．(5)と(6)を比較すると LSTM を用いた(6)の方が DNN を用いた(5)よりも CER が低かった．

3.3 従来のマッピング手法との比較

提案した TM-MFCC から AM-BNF への特徴マッピングの有効性を検証するため，以下の 4 種類のマッピング方法で実験した．

(1) TM-MFCC から AM-MFCC への特徴マッピング
(2) TM-BNF から AM-MFCC への特徴マッピング
(3) TM-BNF から AM-BNF への特徴マッピング
(4) TM-MFCC から AM-BNF への特徴マッピング（提案手法）

これらの特徴マッピングは DNN を用いる方法と LSTM を用いる方法の 2 種類の検討を行った．いずれも DNN の隠れ層の構造は図 3 に示したものと同じで，LSTM の構造は図 5 に示したものと同じであるが，入力層と出力層のユニット数については対象とする特徴量が MFCC か BNF かによって違いがある．具体的には DNN に入力する特徴量が MFCC である場合は TM-MFCC を前後 5 フレームずつ結合した特徴量を入力するため入力層のユニット数は 440 であり，BNF である場合は TM-BNF を前後 5 フレームずつ結合した特徴量を入力するため入力層のユニット数は 462 である．また，LSTM に入力する特徴量が MFCC である場合は入力層のユニット数は 40, BNF である場合は入力層のユニット数は 42 である．一方，DNN を用いる方法でも LSTM を用いる方法でも出力する特徴量が MFCC の場合は気導マイクの CMN が適用された 13 次元の MFCC を推定するように学習するので出力層のユニット数は 13，BNF の場合は CMN が適用された 42 次元の AM-BNF に変換するように学習するので出力層のユニット数は 42 である．いずれも特徴マッピングの DNN，LSTM の重みの初期値はランダムとして学習を行った．認識では TM-MFCC と同様にマッピングした特徴量に CMN，LDA，fMLLR を適用した特徴量を用いる．音響モデルはいずれも気導マイクで学習したモデルである．

表 3　異なる方法で初期化された DNN によって推定された特徴量を用いた認識実験結果（CER : 文字誤り率）

初期化方法	CER
(1) ランダムに初期化	42.2%
(2) 咽喉マイクの BNF 抽出器の重みで初期化	41.8%
(3) 気導マイクの BNF 抽出器の重みで初期化	40.9%

実験結果を表 2 に示す．DNN でも LSTM でも[6]で行われていた従来の TM-MFCC から AM-MFCC への特徴マッピングよりも提案した TM-MFCC から AM-BNF への特徴マッピングの方が CER は低くなった．一方で(2), (3)のような TM-BNF を入力に用いた場合，それぞれ(1), (4)のような TM-MFCC を入力に用いた手法よりも CER が高くなっており，提案した TM-MFCC から AM-BNF へのマッピング方法が最も良い結果となった．

3.4 DNN の重みの初期化方法

特徴マッピングの DNN の重みの初期値の違いによって認識結果にどのような差があるのか確認するため，以下の 3 種類の重みの初期化方法で実験した．

(1) ランダムに初期化する方法 (Random)
(2) 咽喉マイクの BNF を抽出する DNN の重みで初期化する方法（TM-DNN）
(3) 気導マイクの BNF を抽出する DNN の重みで初期化する方法（AM-DNN）（提案手法）

それぞれの DNN が出力した特徴量を用いて認識実験を行った. 音響モデルには気導マイクの BNF で学習したものを用いる．

実験結果を表 3 に示す．ランダムに初期化するよりも BNF を抽出する DNN の重みで初期化する方法を用いた方

表 4　出力特徴量の全次元を一つのネットワークで推定する方法と出力特徴量の次元ごとに異なるネットワークを推定する方法を用いた認識実験結果（CER：文字誤り率）

ネットワークの構成方法	CER（DNN）	CER（LSTM）
(1) 単一ネットワーク	42.2%	39.7%
(2) 次元別ネットワーク	40.4%	40.1%

が特徴マッピングの精度が高くなり，特に気導マイクの BNF を抽出する DNN で初期化する方法が最も良かった．

3.5 特徴マッピングのネットワークの構成方法

出力特徴量の各次元を異なるネットワークで推定する方法（次元別ネットワーク）を検討するため，ランダムに初期化した DNN と LSTM の 2 種類のネットワークで次元別ネットワークを構成して学習し，マッピングした特徴量を用いて認識実験を行なった．認識実験では気導マイクの BNF で学習された音響モデルを用いた．それぞれ出力特徴量の全次元を一つのネットワークで推定する方法（単一ネットワーク）と認識実験結果を比較した．

実験結果を表 4 に示す. DNN を用いた場合では次元別ネットワークにすることで単一ネットワークと比較して CER が改善したが LSTM を用いた場合では改善は見られなかった．

3.6 気導マイクの認識結果との比較

以上より従来法と比べて提案手法によって CER を削減することができ，咽喉マイク音声の認識精度の改善を達成できた．一方，今回の認識実験のテストデータと同時録音した気導マイク音声を用いて AM-MFCC で学習した GMM-HMM を音響モデルとする音声認識システムで認識実験を行ったところ，CER が 29.0%だった．外部雑音の影響が少ない環境下ではまだ咽喉マイク音声と気導マイク音声の認識性能に差があり、改善の余地がある．

4. おわりに

本研究では咽喉マイクの MFCC から気導マイクの BNF への深層学習に基づく特徴マッピングを提案した．特徴マッピングによって咽喉マイクと気導マイクの音響ミスマッチを抑えて大量の接話マイクの BNF で学習した高い識別性能をもつ音響モデルを用いて認識を行うことで特徴マッピングを用いない従来法と比べて認識精度が改善した．また，提案法は咽喉マイクの MFCC から接話マイクの MFCC への DNN に基づく変換手法よりも高い認識精度を確認できた．さらに，特徴マッピングの DNN の重みの初期値を気導マイクの BNF を抽出する DNN の重みとすることでランダムにした場合よりも特徴マッピングの精度が高くなった．今後は特徴マッピングの精度改善や外部雑音の多い環境下で収録した音声での評価を行う予定である．

謝辞　本研究の一部は科研費(16H01817, 16K01543)の助成を受けた．

参考文献

[1]. G. Saon et al., “English Conversational Telephone Speech Recognition by Humans and Machines,” *Proc. Interspeech 2017*, pp. 132–136, 2017.
[2]T. Dekens, W. Verhelst, F. Capman, F. Beaugendre, “Improved Speech Recognition in Noisy Environments by Using a Throat Microphone for Accurate Voicing Detection,” *Signal Processing Conference*, pp. 1978–1982, 2010.
[3]W. Amano, K. Noguchi, R. Takeda, K. Honma, “Automatic Speech Recognition Using Throat Microphone Under Highly-Noisy Environments,” *journal of EICA*, pp. 182–186, 2014, in Japanese
[4]A. Shahina, B. Yegnanarayana, “Mapping Speech Spectra from Throat Microphone to Close-Speaking Microphone: A Neural Network Approach,” *EURASIP Journal on Advances in Signal Processing,* vol. 2007, no. 2, pp. 1–10, 2007.
[5]K. Vijayan K. Sri Rama Murty, “Comparative Study of Spectral Mapping Techniques for Enhancement of Throat Microphone Speech,” *Twentieth National Conference on Communications,* pp. 1–5, 2014.
[6]S. Lin, T. Tsunakawa, M. Nishida, M. Nishimura, “DNN-based Feature Transformation for Speech Recognition Using Throat Microphone,” *APSIPA ASC 2017*, pp. 596–599, 2017.
[7]I. Himawan et al., "Learning Feature Mapping Using Deep Neural Network Bottleneck Features for Distant Large Vocabulary Speech Recognition," *ICASSP*, pp. 4540–4544, 2015.
[8]A. Das, M. Hasegawa-Johnson, “Cross-lingual Transfer Learning during Supervised Training in Low Resource Scenarios,” *INTERSPEECH*, pp. 3531–3535, 2015.
[9]D. Povey et al., “The Kaldi Speech Recognition Toolkit,” IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, 2011
[10]Y. Miao, "Kaldi+PDNN: Building DNN-based ASR Systems with Kaldi and PDNN," arXiv:1401.6984, 2014.