情報処理学会論文誌 Vol.62 No.6 1373–1381 (June 2021)

咽喉マイクを用いた大語彙音声認識のための特徴マッピングによるデータ拡張と知識蒸留

鈴木 貴仁[1,a)] 緒方 淳[2] 綱川 隆司[1] 西田 昌史[1] 西村 雅史[1]

受付日 2020年3月16日, 採録日 2021年3月2日

概要：咽喉マイクは接話マイクのような一般的なマイクよりも外部雑音に頑健であるが，一般的なマイクとの音響ミスマッチが大きく，通常の音声認識システムでは認識精度が低下する．また，大量の音声データが利用可能という状況にもない．本研究では接話マイクと咽喉マイクで同時収録した小規模パラレルデータを活用した咽喉マイク音声認識のための学習手法を提案する．提案手法では，まず既存の大規模音声データベースから抽出した接話マイク特徴量を咽喉マイクの特徴量空間にマッピングし，咽喉マイク用音響モデル（DNN-HMM）の学習データを拡張する．このとき特徴マッピングはパラレルデータを用いてLSTM によって学習する．続いて，特徴マッピングによって得た特徴量で DNN-HMM を初期学習し，これを生徒モデルとする．そして，大量の接話マイク特徴量で学習した DNN-HMM を教師モデルとし，知識蒸留に基づき生徒モデルの再学習を行う．読み上げ音声を用いた評価の結果，提案法は咽喉マイク音声のみで学習した DNN-HMM と比べて約 36.5%の文字誤り率の削減を達成した．

キーワード：咽喉マイク，音声認識，データ拡張，知識蒸留

Feature Mapping-based Data Augmentation and Knowledge Distillation for Large Vocabulary Speech Recognition Using Throat Microphone

Takahito Suzuki[1,a)] Jun Ogata[2] Takashi Tsunakawa[1] Masafumi Nishida[1] Masafumi Nishimura[1]

Received: March 16, 2020, Accepted: March 2, 2021

Abstract: Throat microphones are more robust against external noise than conventional acoustic microphones such as close-talk. However, automatic speech recognition (ASR) performance is degraded when throat microphone speech signals are simply input to a general (clean) ASR system due to large acoustic mismatches. Moreover, the amount of throat microphone speech data is not enough to train accurate ASR systems. In this study, we propose a training approach for throat microphone ASR utilizing a small parallel corpus simultaneously recorded by close-talk and throat microphones. As a data-augmentation process, existing large-amount close-talk microphone features are transformed to a throat microphone feature space with the LSTM-based feature mapping which is trained from the parallel corpus. The DNN-HMM is then pre-trained with the mapped features, and fine-tuned by knowledge distillation from a DNN-HMM trained with a large amount of close-talk microphone speech data. Experimental results using read speech data showed that the proposed approach achieved 36.5% relative improvement of character error rate compared to the DNN-HMM trained only with throat microphone speech data.

Keywords: throat microphone, speech recognition, data augmentation, knowledge distillation

1. はじめに

近年，スマートスピーカやスマートフォンのような通信機能とマイクを備えたデバイスが普及し，日常生活のなかでも気軽に音声認識を利用できるようになった．今後も認

[1] 静岡大学
Shizuoka University, Hamamatsu, Shizuoka 432–8011, Japan
[2] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology, Koto, Tokyo 135–0064, Japan
[a)] suzuki.takahito.14@shizuoka.ac.jp

識性能の向上によって，より多くの場面で活用されることが期待される．一部の音声認識タスクでは，クラウドベースの潤沢な計算資源の活用と深層学習技術に基づくモデルの改良により，すでに人間と同等以上の認識精度を達成したと報告されている [1]．一方，高騒音などの状況下で認識精度を高く維持することはいまだに困難であり，引き続き多くの研究が行われている [2]．

外部雑音の影響を抑制した音声の収録方法の 1 つとして咽喉マイクを用いる方法がある．咽喉付近の皮膚振動をとらえる咽喉マイクは接話マイクのような一般的な気導マイクよりも外部雑音に頑健である．それゆえに高騒音環境下での音声認識 [3], [4], [5], [6], [7], [8], [9] や話者認識 [10], [11]，発話区間検出 [12], [13] において咽喉マイクの利用が検討されている．しかしながら，咽喉マイクは気導マイクとは特性が大きく異なる音が収録されるため，咽喉マイク音声をそのまま通常の音声認識システムに入力した場合，認識精度は著しく低下する．また，これまでに大規模な咽喉マイクの音声データベースは公開されておらず，利用可能な咽喉マイク音声が少量であるため，咽喉マイク音声のみで高精度な音響モデルを学習するのは困難な状況である．

本研究では特徴マッピングによるデータ拡張と知識蒸留を組み合わせた咽喉マイク用 Deep Neural Network-Hidden Markov Model（DNN-HMM）の学習手法を提案する．具体的にはまず，接話マイクの特徴量空間から咽喉マイクの特徴量空間へのマッピングを既存の大規模音声データベースに適用することで咽喉マイク用 DNN-HMM の学習データを拡張する．特徴マッピングは接話マイクと咽喉マイクで同時収録した小規模パラレルデータを用いて Long Short-Term Memory（LSTM）によって学習する．さらに，特徴マッピングによって得た大量の特徴量で初期学習した DNN-HMM をパラレルデータを用いた知識蒸留によって再学習することで DNN のパラメータを実際の咽喉マイク特徴量の入力に適応させる．

2. 関連研究

Deblin らの研究 [14] では音声強調モデル（Spectral Mapper）の学習に主眼を置いているのに対して，本研究では音響モデルの学習に主眼を置いている．Deblin らは Clean 音声と De-noised 音声，初期学習した Spectral Mapper に Noisy 音声を入力したときの出力の誤差だけでなく，Clean 音声と De-noised 音声をそれぞれ学習済みの音素分類モデルに入力したときの誤差を導入して音声強調の精度を改善している．音声認識時は学習した Spectral Mapper の出力を通常の音響モデルに入力する．一方，本研究ではマッピングモデルは拡張データを得るためのモデルであり，マッピングによって拡張したデータを使って初期学習した DNN-HMM を知識蒸留によって再学習するという音響モデルの学習方法である提案手法が Deblin らの研究との違いである．本研究でもマッピング手法を改善することでさらなる精度の改善が期待できるが，これらは今後の課題である．

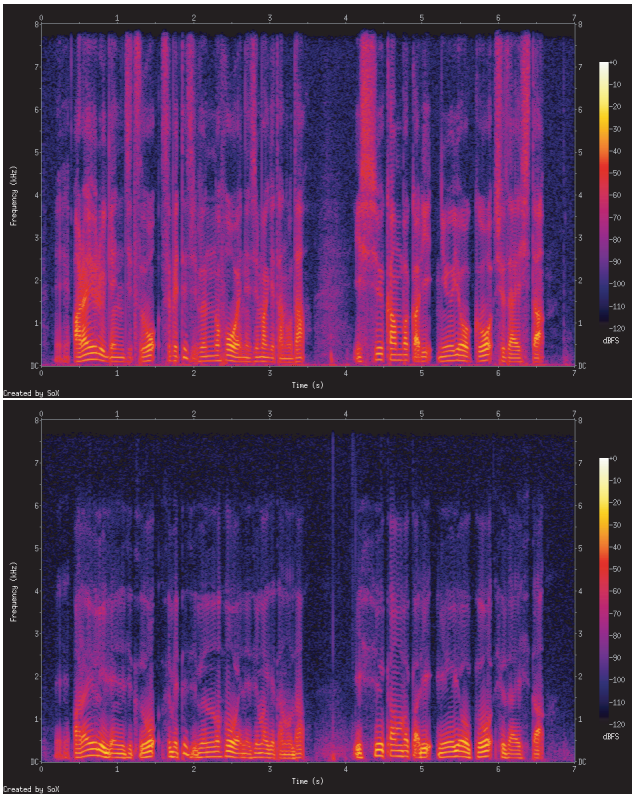

図 1 接話マイクで収録した音声のスペクトログラム（上）と咽喉マイクで収録した音声のスペクトログラム（下）

Fig. 1 Spectrogram of Close-talk microphone (upper plot) and spectrogram of throat microphone (lower plot).

2.1 咽喉マイクを用いた音声認識

皮膚から振動をとらえる咽喉マイクは音声スペクトルに +6 dB/oct 程度の傾斜を与えるとされる口唇からの放射特性が欠落する．また，咽喉マイクがとらえる振動を伝搬する脂肪や骨といった生体組織にはそれぞれ振動吸収に周波数特性があり，気導マイクで収録した音声とはフォルマントが異なる [15]．実際，**図 1** に示すように咽喉マイクで収録した音声は接話マイクよりも高域が減衰しており，不明瞭である．我々はこれまで咽喉マイクの特性について詳しく分析を行った [16]．その結果，接話マイクと咽喉マイクで音声を同時録音し，2 つの音声の LPC ケプストラム距離を算出したところ，7.14 であった．また，周波数特性については，接話マイクと咽喉マイクで収録した音声のスペクトルを比較すると，500 Hz 以下の低周波領域では約 5〜30 dB のずれ，500 Hz 以上の高周波領域では約 5〜20 dB のずれが見受けられた．さらに，接話マイクと咽喉マイクで収録した単語の音声を聴きとる実験を行ったところ，接話マイクでは誤り率 4.9%に対して，咽喉マイクで

は誤り率 27.5%という結果が得られた．以上の結果から，咽喉マイクが接話マイクと比べて特性が異なる点が明らかになった．それゆえに咽喉マイクだけではなく気導マイクと併用し，気導マイクの明瞭な音声とともに雑音環境下では咽喉マイクの音声情報を活用することで音声強調の性能や音響モデルの頑健性を向上しようとする研究が行われている [3], [4], [5], [6]．しかしながら，2 つのマイクそれぞれの音響モデルから算出した確率を統合する手法 [3], [4] では，咽喉マイク用音響モデルが高精度であることが求められる．そこで本研究では咽喉マイクの音響モデルの精度改善に着目する．

咽喉マイク音声をそのまま通常の音声認識システムに入力すると，気導マイクとの音響ミスマッチのために認識精度は著しく低下する．さらに，利用可能な咽喉マイク音声のデータ量が限られているため，咽喉マイク音声のみで高精度な音響モデルを学習するのも困難な状況である．そこで咽喉マイク音声の特徴量を気導マイク音声の特徴量へマッピングすることであらかじめミスマッチを抑制してから，通常の音声認識システムを用いて認識を行う手法が提案されている [7], [8]．マッピングの方法としては，Gaussian Mixture Model（GMM）で区分化した領域で線形変換を行う Stereo Piecewise Linear Compensation for Environment（SPLICE）[17] による手法，Feed-Forward Neural Network（FFNN）[7], [18] や LSTM [8] に基づく非線形変換による手法が提案されている．これらの特徴マッピングは咽喉マイクと接話マイクで同時収録したパラレルデータによって学習される．特徴マッピングによって音響ミスマッチが抑制され，認識精度が改善することが報告されている．しかしながら，骨伝導マイクの研究 [19] でも報告されているとおり，咽喉マイク音声においてもその高域減衰の度合いは人によって異なり，同一話者でも咽喉マイクの装着位置によって収録できる音の特徴が変化する [20] こともあり，気導マイクよりも情報が欠落している咽喉マイク特徴量から気導マイク特徴量へのマッピングを完全に行うことはきわめて難しく，クリーンな環境下では気導マイクと同等の認識精度を達成することはできていない．

2.2 知識蒸留

知識蒸留では one-hot ベクトルで表現される正解ラベル（hard target）を教師信号とする通常の学習とは異なり，学習済みの高精度な DNN（教師モデル）の出力（soft target）を教師信号として DNN（生徒モデル）の学習が行われる．教師モデルは生徒モデルよりも大規模で表現力の高いネットワーク構造であることが多い．教師モデルを 1 つだけではなく複数用いる [21] 方法や，soft target だけでなく hard target も組み合わせて損失を計算する [22], [23] 方法もあり，具体的な学習方法にはいくつかのバリエーションが存在する．知識蒸留はモデル圧縮の手法として知られており，小規模な DNN を生徒モデルとして知識蒸留に基づき学習したところ，hard target で学習した場合よりも教師モデルに近い精度になることが報告されている [24], [25]．また，少量の学習データを用いて hard target による学習と soft target による学習を比較した結果，前者では過学習を引き起こしたのに対し，後者では学習が収束して前者よりも高い精度が得られたと Hinton らは報告しており [26]，知識蒸留は通常の学習よりも強い正則化効果が期待できる．

知識蒸留をドメイン適応に適用する手法も提案されている [27], [28], [29]．具体的には適応元データで学習した DNN を教師モデルとして適応元と適応先のパラレルデータをそれぞれ教師モデルと生徒モデルに入力して得た出力間の損失を計算し，生徒モデルを学習する．我々も大規模な接話マイクの音声データベースで学習した DNN を教師モデルとして，接話マイクと咽喉マイクのパラレルデータを用いて知識蒸留により咽喉マイク用の DNN（生徒モデル）を学習したところ，hard target によって学習した DNN よりも高い認識精度が得られることを確認している [9]．

3. 提案手法

提案する咽喉マイク用 DNN-HMM の学習方法の全体図を**図 2** に示す．提案手法は大きく 2 段階に分けることができ，本章ではまず，特徴マッピングに基づくデータ拡張手法により生成した擬似咽喉マイク特徴量を用いた初期学習に関して述べ，その次節でパラレルデータを用いた知識蒸留による再学習に関して述べる．

3.1 特徴マッピングによる咽喉マイクのデータ拡張

音声データの拡張手法として Vocal Tract Length Perturbation [30] や Speed Perturbation [31]，SpecAugment [32] などが提案されている．しかしながら，オリジナルデータを加工・変形してデータの拡張を行うため，オリジナルデータ量が少ない場合はデータ拡張によって得られるデータ量も限られる．そこで，既存のデータ拡張手法よりも多量でかつ多様性に富んだ拡張データを得るため，特徴マッピングによって咽喉マイク用 DNN-HMM の学習データを拡張する．具体的には，先行研究 [7], [8], [17], [18] で行われていた特徴マッピング（咽喉マイク → 接話マイク）とは逆方向の特徴マッピング（接話マイク → 咽喉マイク）を接話マイクで収録された既存の大規模音声データベースに適用し，大量の擬似咽喉マイク特徴量を生成する．特徴マッピングのモデルは接話マイクと咽喉マイクで同時収録したパラレルデータからそれぞれ特徴量を抽出し，接話マイク側の特徴量を入力信号，咽喉マイク側の特徴量を教師信号としてそれらの平均絶対誤差を最小化するように学習する．すなわち，咽喉マイク音声から抽出した音響特徴量を $x^{tm} = (x_1^{tm}, \ldots, x_n^{tm})$，対応する接話マイク音声から抽出した音響特徴量を $x^{cm} = (x_1^{cm}, \ldots, x_n^{cm})$ とすると，特徴

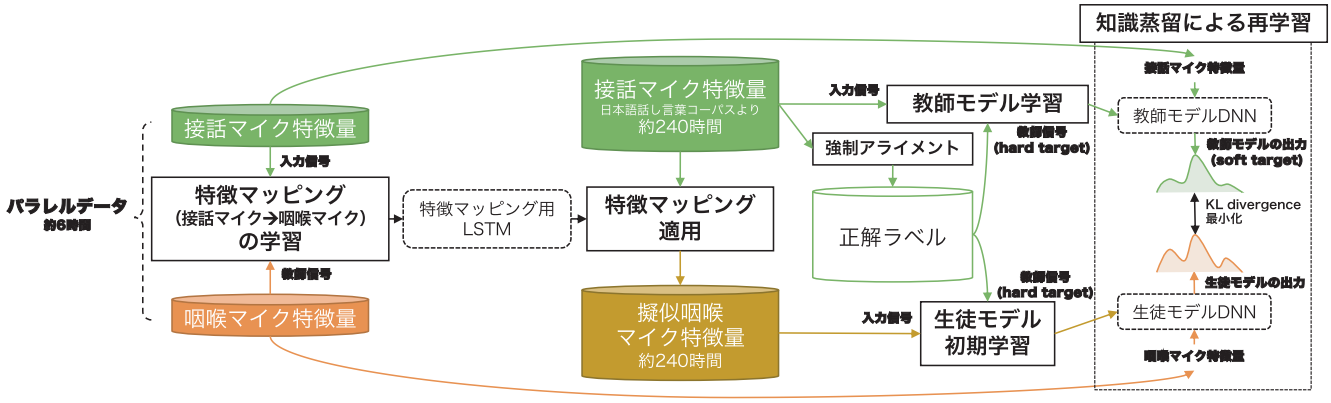

図 2 咽喉マイク用 DNN-HMM のための特徴マッピングによるデータ拡張および知識蒸留
Fig. 2 Feature mapping-based data augmentation and knowledge distillation for training DNN-HMM of throat microphone.

マッピングの誤差 L_{MAE} は以下のように定義される．

$$L_{MAE} = \frac{1}{n}\sum_{k=1}^{n}|x_k^{tm} - \sigma(x_k^{cm})| \tag{1}$$

ここで σ は DNN に基づく特徴マッピングを表す．擬似咽喉マイク特徴量 x' は学習済みの DNN に接話マイク音声から抽出した音響特徴量 x^{cm} を入力して以下の式に従って計算される．

$$x' = \sigma(x^{cm}) \tag{2}$$

咽喉マイクから接話マイクへの特徴マッピングでは FFNN よりも LSTM が有効に働いたと報告されており [8]，本研究でも特徴マッピング用モデルとして LSTM を用いる．

続いて，生成した擬似咽喉マイク特徴量を用いて咽喉マイク用 DNN-HMM の初期学習を行う．特徴マッピングによって生成した擬似咽喉マイク特徴量を入力信号，対応する正解ラベルを教師信号としてクロスエントロピーを最小化するように DNN を学習する．正解ラベルは one-hot ベクトルになっており，入力特徴量に対応する HMM 状態を表すクラスは 1，それ以外のクラスは 0 となっている．HMM の状態数を N として，HMM の状態を $s = (s_1, \ldots, s_N)$，正解ラベルを $d = (d_1, \ldots, d_N)$ とすると，以下の式で表される誤差 L_{CE} を最小化するように学習する．

$$L_{CE} = -\sum_i (d_i \log(P(s_i|x'))) \tag{3}$$

ここで $P(s_i|x')$ は擬似咽喉マイク特徴量 x' を DNN に入力して算出した i 番目の HMM の状態 s_i の事後確率を表す．なお，正解ラベルは擬似咽喉マイク特徴量ではなくマッピング前の接話マイク特徴量に対して GMM-HMM で強制アライメントを行った結果に基づいて推定する．

本研究では，咽喉から気導へのマッピングは欠落した情報を生成・補完することに相当する困難な変換問題であるのに対して，気導から咽喉へのマッピングは情報を欠落・変質させる一種のフィルタリングであり，気導から咽喉へのマッピングのほうがより現実的な問題と考えて検討を行った．しかし，LSTM によるマッピングだけでは正確なマッピングが難しいと考えられるので，マッピングによる歪みのない咽喉マイク特徴量を用いた知識蒸留による再調整を行うことで特徴量の多様性の確保を試みた．

3.2 知識蒸留による再学習

特徴マッピングによって咽喉マイクの特徴量を完全に模倣した特徴量を生成することは困難であり，実際の咽喉マイク特徴量との間にはミスマッチ部分が残っていると考えられる．そこで初期学習済み DNN のパラメータをパラレルデータと知識蒸留を用いて咽喉マイクの特徴量に適応させる．生徒モデルの初期パラメータは擬似咽喉マイク特徴量で学習した DNN とし，教師モデルはマッピング前の大規模な接話マイク特徴量とそのアライメント結果に基づいて推定した正解ラベルで事前に学習しておく．

本研究では教師モデルが出力した事後確率分布と生徒モデルが出力した事後確率分布間の KL Divergence を最小化するように生徒モデルの学習を行う．すなわち，パラレルデータの接話マイク音声から抽出した特徴量 x_c を教師モデルに入力したときの HMM 状態 s_i の事後確率を $P(s_i|x_c)$，対応する咽喉マイク音声から抽出した特徴量 x_t を生徒モデルに入力したときの s_i の事後確率を $Q(s_i|x_t)$ としたとき，損失関数は以下のように定義される．

$$\begin{aligned} D_{KL}(P||Q) &= \sum_i P(s_i|x_c)\log\frac{P(s_i|x_c)}{Q(s_i|x_t)} \\ &= \sum_i P(s_i|x_c)\log P(s_i|x_c) \\ &\quad - \sum_i P(s_i|x_c)\log Q(s_i|x_t) \end{aligned} \tag{4}$$

なお，式 (4) の第一項は生徒モデルのパラメータの最適化に関係しないため無視できる．それゆえにパラメータの最適化では第二項のみを用いて損失を計算する．なお，第二項はクロスエントロピーの式と同等である．

認識時はまず咽喉マイク音声から抽出した特徴量 x_t を再学習済みの生徒モデルに入力し $Q(s_i|x_t)$ を得た後，ベイズの定理により事前確率 $Q(s_i)$ を用いて HMM の各状態の出力確率 $Q(x_t|s_i)$ を計算し，デコードを行う．ここで $Q(s_i)$ は教師モデルの学習時に使用したアライメント結果から各 HMM 状態 s_i の出現回数を数えてそれを総フレーム数で割ることであらかじめ計算しておく．デコードに使用する HMM はその強制アライメントに使用したものと同一のものを使用する．

なお，知識蒸留での逆温度パラメータは 1 とした．

4. 認識実験

4.1 データセット

本研究で使用した咽喉マイク（**図 3**）はネックバンドの先に小型のコンデンサマイクユニットが取り付けられており，装着すると咽喉付近の皮膚振動に由来する音をとらえることができる．この咽喉マイクは首元の血流によく反応して低周波信号が混入するため，事前にハイパスフィルタを適用しておく．咽喉マイクの装着位置に関して調査したところ，個人差はあるものの，咽頭寄りでやや上方にコンデンサマイクユニットを密着させることで良好な音声信号が得られる可能性が高いことが分かっており [20]，本研究で使用する咽喉マイク音声はすべてその位置で収録されている．

評価データとしては咽喉マイクと接話マイクで同時収録した男性話者 10 名による新聞記事読み上げ音声（約 40 分）を使用した．なお，評価の際はどちらか 1 ch のみを用いる．学習用の咽喉マイクと接話マイクのパラレルデータとしては男性話者 11 名から収集した音素バランス文読み上げ音声（約 6 時間）を用いた．いずれのデータも静かな環境で収録されている．接話マイクで収録された既存の大規模な音声データベースとしては日本語話し言葉コーパス（CSJ）から約 240 時間の音声を使用した．評価データには学習データの話者は含まれていない．なお，評価データ中に出現する単語のうち，約 2.8%が未知語であった．

接話マイクの学習データには，データ量の問題で JNAS ではなく CSJ を用いた．また，知識蒸留における教師データは高精度なモデルが必要であると考えたため，多種多様な音声を含む CSJ の学会講演を学習データに用いた．

本研究で用いた音声は，サンプリング周波数 16 kHz，フレーム長 25 ms，フレームシフト幅 10 ms で特徴量抽出を行った．

図 3 咽喉マイク
Fig. 3 Throat microphone.

4.2 実験方法

特徴量抽出やモデルの学習，評価には Kaldi ツールキットを用いた．教師モデルの学習と生徒モデルの初期学習用正解ラベルの推定には CSJ で学習した GMM-HMM を用いた．その HMM 状態数は約 9,300 である．GMM-HMM は 13 次元の Mel-Frequency Cepstral Coefficient（MFCC）を前後 4 フレームずつ結合して線形判別分析によって 40 次元に圧縮し，Maximum Likelihood Linear Transform（MLLT）および feature-space Maximum Likelihood Linear Regression (fMLLR）を適用したものを用いた．特徴マッピングや生徒モデルの入力特徴量には 40 次元の FBANK に対して MFCC と同様の処理を通して fMLLR を適用した 40 次元の特徴量を用いた．特徴マッピング用 LSTM のユニット数は 512 とし，LSTM の後に 40 ユニットの出力層を持つ構造とした．生徒モデルの DNN の入力特徴量に関しては fMLLR を適用した 40 次元の特徴量を前後 5 フレーム結合した 440 次元の特徴量を用いた．生徒モデルの隠れ層は 6 層の全結合層（1,024 ユニット）を持ち，出力層のユニット数は CSJ で学習した GMM-HMM の状態数に等しい．擬似咽喉マイク特徴量を用いた学習では Stacked Denoising Autoencoder [33] による教師なし事前学習を行い，その後に正解ラベルを用いた Fine-tuning を行った．

教師モデルにはユニット数が 1024 の Time Delay Neural Network（TDNN）[34] を重ねたネットワークを用いた．その構造の詳細を**表 1** に示す．なお，表 1 中の $[-n, m]$ は前 n フレームから後ろ m フレームまでの全フレームを結合すること，$\{-n, m\}$ は前 n フレームと後ろ m フレームのみを結合すること，$\{0\}$ は前後の結合を行わないことを表す．入力特徴量は 40 次元の high-resolution MFCC と 100 次元の i-vector を結合したものである．教師モデルの学習には CSJ の約 240 時間の音声に対して Speed Perturbation

表 1 教師モデルの構造
Table 1 Structure of teacher model.

Layer	活性化関数	結合フレーム数
第 1 層	ReLU	$[-2, 2]$
第 2 層	ReLU	$\{-1, 2\}$
第 3 層	ReLU	$\{-3, 3\}$
第 4 層	ReLU	$\{-3, 3\}$
第 5 層	ReLU	$\{-7, 2\}$
第 6 層	ReLU	$\{0\}$
出力層	Softmax	$\{0\}$

表 2 従来手法と提案手法の文字誤り率
Table 2 Character error rates (CER) of conventional and proposed approaches.

Model	Training data	CER
TM GMM-HMM	咽喉マイク音声（約 6 時間）	14.4%
TM DNN-HMM	咽喉マイク音声（約 6 時間）	9.6%
TM DNN-HMM + Speed Perturbation	咽喉マイク音声 + Speed Perturbation（約 18 時間）	9.4%
CM DNN-HMM + Feature Mapping	接話マイク音声（約 240 時間）	8.8%
Map-aug DNN-HMM	擬似咽喉マイク特徴量（約 240 時間）	9.0%
Map-aug DNN-HMM + KD	擬似咽喉マイク特徴量（約 240 時間）+ パラレルデータ（約 6 時間）	**6.1%**

（$\alpha = 0.9, 1.0, 1.1$）を適用して拡張したデータを用いた．

比較対象とする従来手法として咽喉マイク音声（約 6 時間）のみで学習した GMM-HMM と DNN-HMM を音響モデルとするシステムを用いた．この GMM-HMM は前述の学習方法と同様に fMLLR を適用した MFCC によって学習され，その HMM 状態数は約 4,000 である．DNN-HMM の入力次元数や隠れ層の構造は生徒モデルと同じであるが，咽喉マイクのみで学習した GMM-HMM による強制アライメント結果に基づいて推定した正解ラベルを利用して学習を行うため，出力層のユニット数は咽喉マイクのみで学習した GMM-HMM の状態数に等しい．

加えて，咽喉マイクから接話マイクへの特徴マッピングを適用する従来手法との比較も行った．音響モデルには CSJ で学習した DNN-HMM を用いた．この DNN は生徒モデルと同じ構造を持つ．また，この特徴マッピング（咽喉マイク → 接話マイク）は提案手法の特徴マッピング（接話マイク → 咽喉マイク）用 LSTM と同じ構造を持ち，同じパラレルデータを用いて学習した．

すべての認識実験において，3-gram 言語モデルを使用したデコードによって得た 100 の認識仮説に対して TDNN-LSTM 言語モデルによるリスコアリングを行い，最終的な認識結果を推定した．リスコアリング時には 3-gram 言語モデルの重みを 0.2，TDNN-LSTM 言語モデルの重みを 0.8 とした．TDNN-LSTM はユニット数が 2,048 の TDNN と LSTM-projection（LSTMP）を交互に 5 層重ねた構造とした．3-gram および TDNN-LSTM 言語モデルは CSJ の書き起こしを使用して学習した．

4.3 実験結果

4.3.1 従来法との比較

まず，従来手法と提案手法との認識精度を評価し，比較を行った．各モデルの文字誤り率（CER）を**表 2** に示す．表 2 の各手法間での認識精度に対して有意水準 5%にて t 検定を行った結果，各手法間での認識精度に有意差があることが明らかになった．表中の TM GMM-HMM と TM DNN-HMM はそれぞれ咽喉マイク音声のみで学習した音響モデルを用いたシステム，TM DNN-HMM + Speed Perturbation は咽喉マイク音声に対して Speed Perturbation（$\alpha = 0.9, 1.0, 1.1$）を適用して拡張したデータで学習した音響モデルを用いたシステム，CM DNN-HMM + Feature Mapping は，咽喉マイク特徴量を接話マイク特徴量に変換し，接話マイクで学習した音響モデルに入力するシステム，Map-aug DNN-HMM は擬似咽喉マイク特徴量で学習した DNN-HMM を音響モデルとしたシステム，Map-aug DNN-HMM + KD は Map-aug DNN-HMM の DNN のパラメータを知識蒸留によって咽喉マイク特徴量に適応させたシステムである．咽喉マイク音声のみで学習した DNN-HMM は GMM-HMM よりも高精度であり，加えて Speed Perturbation を適用することで認識精度が改善したが，提案法の特徴マッピングによるデータ拡張手法で生成した擬似咽喉マイク特徴量で学習した DNN-HMM はさらに高い認識精度を示した．咽喉マイク音声に対して変形・加工を行う拡張手法に比べて提案した拡張手法はより多様な特徴量を生成することができ，それが認識精度の改善に寄与したと考えられる．なお，Speed Perturbation の係数を 5 種類（$\alpha = 0.9, 0.95, 1.0, 1.05, 1.1$）とした場合の認識実験も行ったが，その文字誤り率は 10.2%と係数を 3 種類とした場合よりも認識精度が低かった．さらに，知識蒸留による再学習によって実際の咽喉マイク特徴量に適応したことでさらなる認識精度の改善が得られ，従来の咽喉マイクから接話マイクへの特徴マッピングを用いる手法よりも高い性能が得られた．

約 6 時間のみの音声を使用した TM DNN-HMM や +Speed Perturbation と，変換データとはいえ CSJ の約 240 時間を使用している Map-aug DNN-HMM との間の認識精度の差が小さくなっている．これは，Map-aug は DNN-HMM の学習データ量が多いものの不完全なマッピングのため，咽喉マイク特徴量との間にミスマッチが残っているため，大きな改善が得られていないと考えられる．それに対して，不完全なマッピングでできた疑似咽喉マイク特徴量で初期学習したモデルを知識蒸留で再調整することが有効であると考えられる．

4.3.2 生徒モデルの初期化方法と再学習方法の比較

次に，再学習時の生徒モデルのパラメータの初期化方法と学習方法ごとの性能を評価した．活性化関数が sigmoid である生徒モデルの初期化方法として Glorot の一様分布

表 3 生徒モデルの初期化方法と再学習方法ごとの文字誤り率
Table 3 Character error rates (CER) of student model fine-tuned by hard target or soft target in each initialization approach.

Initialization approach	Fine-tuning approach	
	hard target	soft target
Random	17.4%	98.1%
CM DNN	8.7%	7.2%
Map-aug DNN	8.1%	**6.1%**

による初期化（Random）[35]，CSJ で学習した DNN を初期パラメータとする方法（CM DNN），そして提案手法の擬似咽喉マイク特徴量で学習した DNN を初期パラメータとする方法（Map-aug DNN）の 3 種類を比較した．一方，再学習方法として正解ラベルを教師信号とする方法（hard target）と提案手法の知識蒸留による方法（soft target）を比較した．なお，hard target による学習ではパラレルデータの接話マイク音声に対して CSJ で学習した GMM-HMM で強制アライメントした結果に基づいて推定した正解ラベルを教師信号として用いた．各手法の文字誤り率を**表 3** に示す．乱数によって初期化した場合，知識蒸留ではうまく学習が進まなかったが，学習済みの DNN のパラメータで初期化されている場合，hard target で学習するよりも高い精度が得られ，知識蒸留の有効性を確認した．また，初期パラメータを接話マイクで学習した DNN とするよりも提案法である擬似咽喉マイク特徴量で学習した DNN とする方法が高い認識精度を示した．咽喉マイク音声は接話マイク音声と比べて音素識別に必要な情報が欠落しており，擬似咽喉マイク特徴量もこのような咽喉マイクの特徴をある程度再現できていると考えられる．したがって，接話マイク音声で学習した DNN よりも擬似咽喉マイク特徴量で学習した DNN は情報が欠落している特徴量から音素を識別する点で優れ，その知識を再学習で活用できたことがより高い識別性能を獲得できた要因の 1 つだと考えられる．

4.3.3 接話マイクの認識精度との比較

最後に，評価データの咽喉マイク音声と接話マイク音声それぞれの認識精度の比較を行った．評価にはクリーンな環境で収録した音声だけでなく，それに雑音を人工的に重畳して特定の騒音レベルの環境をシミュレートした音声も用いた．騒音環境をシミュレートするため，騒音レベルが約 60 dB と約 80 dB のときの接話マイクと咽喉マイクの Signal-to-Noise Ratio（SNR）を調査した．SNR の調査結果を**表 4** に示す．この調査結果に従って 60 dB と 80 dB 程度の騒音環境をシミュレートするように重畳する雑音のレベルを調整した．重畳雑音には接話マイクと咽喉マイクを装着して収録したレストラン内の騒音を使用した．なお，暗騒音のレベルは約 35 dB であった．

咽喉マイク用音響モデルとしては提案手法モデル（Map-aug DNN-HMM + KD）を用いた．一方，接話マイク用音響モデルとしては提案手法の再学習時に利用した教師モデルを用いた．各騒音レベルでの接話マイクと咽喉マイクそれぞれの文字誤り率を**表 5** に示す．咽喉マイク音声の認識精度は提案手法モデルを用いることでクリーンな接話マイク音声の認識精度に迫る結果が得られた．また，60 dB 程度の騒音環境をシミュレートした評価データでは咽喉マイク音声の認識精度が接話マイク音声に劣る結果となったが，80 dB 程度の評価データでは接話マイク音声よりも高い認識精度を示した．

表 4 各騒音レベルでの接話マイクと咽喉マイクの SNR
Table 4 SNR of close-talk and throat microphones per noise level.

Noise level	Close-talk mic	Throat mic
60 dB	26.7 dB	39.2 dB
80 dB	13.9 dB	30.3 dB

表 5 各騒音レベルでの接話マイク音声と咽喉マイク音声の文字誤り率
Table 5 Character error rates (CER) of close-talk and throat microphones per noise level.

Noise level	Close-talk mic	Throat mic
clean	**4.8%**	6.1%
60 dB	**6.4%**	7.6%
80 dB	18.0%	**13.4%**

5. おわりに

本研究では既存の大規模接話マイク音声データと特徴マッピングによるデータ拡張手法と咽喉マイクと接話マイクの小規模パラレルデータを用いた知識蒸留を組み合わせた咽喉マイク用 DNN-HMM の学習方法を提案した．新聞記事読み上げ音声を用いた認識実験の結果，DNN-HMM を咽喉マイク音声のみで学習する従来手法と比較して約 36.5%の文字誤り率の削減を達成した．

今後は特徴マッピングの高精度化や複数の教師モデルからの知識蒸留，既存の雑音抑圧技術との比較実験，8 kHz サンプリングでモデルを構成する方法などについて検討を行う予定である．

謝辞 本研究の一部は JSPS 科研費（16H01817, 18H03260），国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務（JPNP20006）の助成を受けた．

参考文献

[1] Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M.L., Stolcke, A., Yu, D. and Zweig, G.: Toward Human Parity in Conversational Speech Recognition, *IEEE/ACM Trans. Audio, Speech, and Language Processing*, Vol.25, No.12, pp.2410–2423 (2017).
[2] Zhang, Z., Geiger, J.T., Pohjalainen, J., Mousa, A.E.-D. and Schuller, B.W.: Deep Learning for Environmentally

Robust Speech Recognition: An Overview of Recent Developments, *ACM TIST*, Vol.9, pp.49:1–49:28 (2017).
[3] Dupont, S., Ris, C. and Bachelart, D.: Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise, *Proc. Robust 2004 (Workshop (ITRW) on Robustness Issues in Conversational Interaction)* (2004).
[4] Panikos, H., Jani, E., Ishi, C.T., Miyashita, T. and Hagita, N.: Fusion of Standard and Alternative Acoustic Sensors for Robust Automatic Speech Recognition, *ICASPP2012*, pp.4837–4840 (2012).
[5] Graciarena, M., Cesari, F., Franco, H., Myers, G.K., Cowan, C. and Abrash, V.: Combination of standard and throat microphones for robust speech recognition in highly noisy environments, *Interspeech* (2004).
[6] Lin, S., Tsunakawa, T., Nishida, M. and Nishimura, M.: Conversational Speech Recognition Using Multiple Wearable Microphones, *NCSP*, pp.363–366 (2018).
[7] Suzuki, T., Ogata, J., Tsunakawa, T., Nishida, M. and Nishimura, M.: Bottleneck feature-mediated DNN-based feature mapping for throat microphone speech recognition, *APSIPA ASC 2018*, pp.1738–1741 (2018).
[8] Lin, S., Tsunakawa, T., Nishida, M. and Nishimura, M.: DNN-based Feature Transformation for Speech Recognition Using Throat Microphone, *APSIPA ASC 2017*, pp.596–599 (2017).
[9] Suzuki, T., Ogata, J., Tsunakawa, T., Nishida, M. and Nishimura, M.: Knowledge Distillation for Throat Microphone Speech Recognition, *Interspeech*, pp.461–465 (2019).
[10] Yegnanarayana, B., Shahina, A. and Kesheorey, M.R.: Throat microphone signal for speaker recognition, *Interspeech* (2004).
[11] Sahidullah, M., Hautamäki, R.G., Thomsen, D.A.L., Kinnunen, T., Tan, Z.H., Hautamäki, V., Parts, R. and Pitkänen, M.: Robust speaker recognition with combined use of acoustic and throat microphone speech, *Interspeech*, pp.1720–1724 (2016).
[12] Dekens, T., Verhelst, W., Capman, F. and Beaugendre, F.: Improved speech recognition in noisy environments by using a throat microphone for accurate voicing detection, *European Signal Processing Conference*, pp.1978–1982 (2010).
[13] Otaka, Y., Tshunakawa, T., Nishida, M. and Nishimura, M.: Voice Activity Detection Using Throat and Lavalier Microphones for Multi-Party Conversations, *NCSP*, pp.369–372 (2017).
[14] Bagchi, D., Plantinga, P., Stiff, A. and Fosler-Lussier, E.: Spectral Feature Mapping with Mimic Loss for Robust Speech Recognition, *ICASSP* (2018).
[15] 林　伸行：骨伝導マイクイヤホン，テレビジョン学会誌，Vol.50, No.3, pp.351–357 (1996).
[16] 鈴木貴仁，緒方　淳，綱川隆司，西田昌史，西村雅史：咽喉音による音声認識のための収録デバイスに関する検討，日本音響学会講演論文集，1-R-17 (2018).
[17] Zheng, Y., Liu, Z., Zhang, Z., Sinclair, M., Droppo, J., Deng, L., Acero, A. and Huang, X.: Air- and bone-conductive integrated microphones for robust speech detection and enhancement, *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pp.249–254 (2003).
[18] Shahina, A. and Yegnanarayana, B.: Mapping Speech Spectra from Throat Microphone to Close-Speaking Microphone: A Neural Network Approach, *EURASIP Journal on Advances in Signal Processing*, Vol.2007, No.1, p.087219 (online), DOI: 10.1155/2007/87219 (2007).
[19] Kondo, K., Fujita, T. and Nakagawa, K.: On Equalization of Bone Conducted Speech for Improved Speech Quality, *2006 IEEE International Symposium on Signal Processing and Information Technology*, pp.426–431 (2006).
[20] Suzuki, T., Ogata, J., Tsunakawa, T., Nishida, M. and Nishimura, M.: Effects of Mounting Position on Throat Microphone Speech Recognition, *IEEE GCCE*, pp.897–898 (2019).
[21] Chebotar, Y. and Waters, A.: Distilling Knowledge from Ensembles of Neural Networks for Speech Recognition, *Interspeech*, pp.3439–3443 (2016).
[22] Yang, Z., Zhang, C., Zhang, W., Jin, J. and Chen, D.: Essence Knowledge Distillation for Speech Recognition, *ArXiv*, Vol.abs/1906.10834 (2019).
[23] Tachioka, Y.: Knowledge Distillation Using Soft and Hard Labels and Annealing for Acoustic Model Training, *IEEE GCCE*, No.2, pp.715–716 (2019).
[24] Ba, L. and Caruana, R.: Do deep nets really need to be deep?, *Advances in Neural Information Processing Systems*, Vol.3, pp.2654–2662 (2014).
[25] Chan, W., Ke, N.R. and Lane, I.: Transferring knowledge from a RNN to a DNN, *Interspeech*, pp.3264–3268 (2015).
[26] Hinton, G., Vinyals, O. and Dean, J.: Distilling the Knowledge in a Neural Network, *NIPS Deep Learning and Representation Learning Workshop* (2015) (online), available from ⟨http://arxiv.org/abs/1503.02531⟩.
[27] Li, J., Zhao, R., Chen, Z., Liu, C., Xiao, X., Ye, G. and Gong, Y.: Developing Far-Field Speaker System Via Teacher-Student Learning, *ICASSP2018*, No.1, pp.5699–5703 (2018).
[28] Li, J., Seltzer, M.L., Wang, X., Zhao, R. and Gong, Y.: Large-Scale Domain Adaptation via Teacher-Student Learning, *ArXiv*, Vol.abs/1708.05466 (2017).
[29] Yi, J., Tao, J., Wen, Z. and Liu, B.: Distilling Knowledge Using Parallel Data for Far-field Speech Recognition, *ArXiv*, Vol.abs/1802.06941 (2018).
[30] Jaitly, N. and Hinton, E.S.: Vocal Tract Length Perturbation (VTLP) improves speech recognition, *Iin International Conference on Machine Learning (ICML) Workshop on Deep Learning for Audio, Speech, and Language Processing* (2013).
[31] Ko, T., Peddinti, V., Povey, D. and Khudanpur, S.: Audio Augmentation for Speech Recognition, *Interspeech* (2015).
[32] Park, D.S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E.D. and Le, Q.V.: SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition, *Interspeech*, pp.2613–2617 (2019).
[33] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. and Manzagol, P.-A.: Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion, *The Journal of Machine Learning Research*, Vol.11, pp.3371–3408 (2010).
[34] Peddinti, V., Povey, D. and Khudanpur, S.: A time delay neural network architecture for efficient modeling of long temporal contexts, *Interspeech*, Vol.2015-Janua, pp.3214–3218 (2015).
[35] Glorot, X. and Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks, *Proc. 13th International Conference on Artificial Intelligence and Statistics*, Teh, Y.W. and Titterington, M. (Eds.), *Proc.*

Machine Learning Research, Vol.9, pp.249–256 (2010).

鈴木 貴仁 （学生会員）

2018 年静岡大学情報学部情報科学科卒業．2020 年同大学大学院総合科学技術研究科情報学専攻修士課程修了．在籍中，音声認識に関する研究に従事．日本音響学会会員．

緒方 淳 （正会員）

2003 年龍谷大学理工学研究科博士後期課程修了．同年産業技術総合研究所に入所し，現在，人工知能研究センター主任研究員．音声認識・理解に関する研究に従事し，音声情報検索サービス PodCastle 等，音声認識技術を活用したシステム，アプリケーションの研究開発を推進．2000 年日本音響学会粟屋潔学術奨励賞，2001 年電子情報通信学会学術奨励賞，2006 年情報処理学会山下記念研究賞，2012 年度情報処理学会論文賞等受賞．日本音響学会会員．博士（工学）．

綱川 隆司 （正会員）

2005 年東京大学大学院情報理工学系研究科コンピュータ科学専攻修士課程修了．2008 年同博士後期課程単位取得退学．東京大学特任研究員，静岡大学学術研究員，同情報学部助教を経て，2019 年より静岡大学情報学部講師．自然言語処理に関する研究に従事．言語処理学会会員．博士（情報理工学）．

西田 昌史 （正会員）

1999 年龍谷大学大学院理工学研究科電子情報学専攻修士課程修了．2002 年同博士後期課程修了．千葉大学助手，同助教，同志社大学准教授，名古屋大学特任准教授を経て，2015 年より静岡大学情報学部准教授．音声情報処理，行動信号処理，福祉情報工学に関する研究に従事．2011 年度情報処理学会山下記念研究賞．電子情報通信学会，日本音響学会，人工知能学会，ヒューマンインタフェース学会各会員．博士（工学）．

西村 雅史 （正会員）

1983 年大阪大学大学院基礎工学研究科博士前期課程修了．同年日本アイ・ビー・エム（株）東京基礎研究所入社．2014 年より静岡大学大学院総合科学技術研究科教授．音声言語情報処理，特に音声認識や生体音認識等に関する研究に従事．2018～2019 年度情報処理学会音声言語情報処理研究会主査，2019～2020 年度日本音響学会東海支部支部長．1998 年情報処理学会山下記念研究賞，1999 年日本音響学会技術開発賞等受賞．IEEE，電子情報通信学会，日本音響学会，人工知能学会，信号処理学会，ISCA，日本咀嚼学会各会員．博士（工学）．