2T-09

咽喉マイクの大語彙音声認識のための知識蒸留

鈴木貴仁† 緒方淳‡ 綱川隆司† 西田昌史† 西村雅史†
静岡大学大学院総合科学技術研究科情報学専攻† 産業技術総合研究所‡

1 はじめに

咽喉付近に密着させる咽喉マイクは一般的な接話マイクと比較して周囲の雑音の影響を抑制した音声の収録が可能であり，雑音環境下での音声認識への活用に向けた研究が行われている[1][2]．一方，咽喉マイクは接話マイクと音響特性が大きく異なるため，一般的な音声認識システムでは認識精度が低下する．また，咽喉マイクの音声データ量には限りがあり，咽喉マイク音声のみで十分な精度を持つ音響モデルを構築するのが困難である．そのため，[1]や[2]のように認識時に咽喉マイクと接話マイクを併用し，雑音レベルが高い時に咽喉マイク音声の情報を多く活用する手法が提案されているが，雑音レベルに応じて重みを動的に変化させるのは難しい．そこで本研究では認識時に接話マイクを併用しない咽喉マイク単体での音声認識の精度改善を目的とし，大量の接話マイク音声と少量のパラレルデータを活用した咽喉マイク音声のための知識蒸留に基づくハイブリッド方式の音響モデルの学習方法を提案し，日本語の大語彙音声認識による評価を行ったので報告する．

2 提案手法

HMMによる音声認識ではDNNを用いた音響モデルとしてハイブリッド方式とタンデム方式がよく利用される[4]．ハイブリッド方式ではDNNによってHMMの隠れ状態の出力確率を計算する．一方，タンデム方式ではDNNを特徴量抽出器として使用し，隠れ層の出力を特徴量としてGMMによってHMMの出力確率を計算する．出力を得る隠れ層はノード数を少なく設定することが多く，その層から抽出される特徴量はボトルネック特徴量(BNF)と呼ばれる．我々は先にパラレルデータを用いて咽喉マイク音声のMFCCから接話マイク音声のBNFへの特徴マッピングをLSTMで学習し，大量の接話マイク音声のBNFで学習したGMM-HMMを認識に活用することで認識精度を改善できることを示した[3]．GMM-HMMでは入力特徴量を正規分布に当てはめて出力確率の計算が行われるのに対して，ハイブリッド方式では出力確率の計算もDNNによって最適化される（DNN-HMM）．

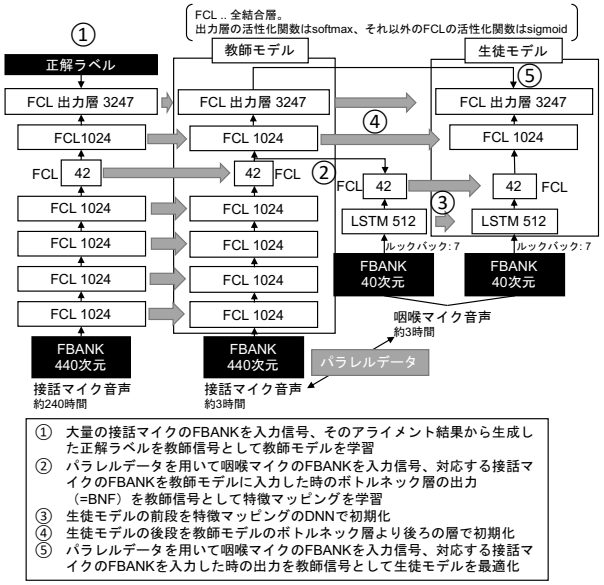

図1 提案手法の概要

本研究ではハイブリッド方式音響モデルによって咽喉マイク音声認識のさらなる性能向上を目指す．咽喉マイク音声（パラレルデータ）のデータ量が少量である場合，大規模データを前提とした通常のDNN-HMMを学習・適用することは性能面で問題があると考えられる．そこで本研究では，少量のパラレルデータに適したDNN-HMMの構成・学習法を提案する．具体的には，DNNの前段を咽喉マイク音声の音響特徴量から接話マイク音声のBNFへの特徴マッピング用LSTMのパラメータで初期化し，後段を接話マイク音声のBNFからHMMの隠れ状態を識別するネットワークのパラメータで初期化した後，パラレルデータを用いて全体を最適化する（図1）．全体の最適化では教師信号に正解ラベルを用いず，大量の接話マイクで学習したハイブリッド方式の音響モデルのDNN（教師モデル）にパラレルデータの接話マイクを入力した時の出

Knowledge Distillation for Throat Microphone Large Vocabulary Speech Recognition
Takahito Suzuki†, Jun Ogata‡, Takashi Tsunakawa†, Masafumi Nishida†, Masafumi Nishimura†
† Graduate School of Integrated Science and Technology, Shizuoka University
‡ National Institute of Advanced Industrial Science and Technology

表 1 従来手法と提案手法の認識実験結果

モデル	CER
(1) TM-GMM-HMM	52.4 %
(2) TM-DNN-HMM	49.3 %
(3) LSTM-GMM-HMM	39.7 %
(4) TM-LSTM-DNN-HMM	45.4 %
(5) **LSTM-DNN-HMM-KD**	**39.5 %**

力分布を用いる．この学習方法は知識蒸留と呼ばれ，強い正則化効果があると報告されており[5]，過学習防止に効果が期待できる．

3 認識実験

3.1 実験条件

評価データとして咽喉マイクで収録した男性話者 10 名による自由発話の音声（約 20 分）を使用し，文字誤り率(CER)による評価を行った．教師モデルの学習に用いる接話マイク音声には日本語話し言葉コーパス（CSJ）より約 240 時間を使用した．咽喉マイクの音響モデルの学習に用いる接話マイクと咽喉マイクのパラレルデータには簡易防音室で同期録音した男性話者 8 名による約 3 時間の音素バランス文読み上げ音声を使用した．言語モデルには CSJ の書き起こし文から生成した 3-gram 言語モデルを用いた．

教師モデルの入力特徴量は fMLLR を適用した 40 次元の FBANK の前後 5 フレームを結合した 440 次元の特徴量である．教師モデルは Stacked Denoising Autoencoder による事前学習を行った後，ファインチューニングを行った．咽喉マイクの音響モデルの入力特徴量は fMLLR を適用した 40 次元の FBANK である．特徴マッピングの LSTM のルックバック数は 7 とした．DNN のノード数は図 1 に示す通りである．

3.2 実験結果

まず，従来手法と提案手法の実験結果を表 1 に示す．ここで表中のモデル(1)は約 3 時間の咽喉マイク音声の MFCC で学習した GMM-HMM を音響モデルとしたシステム，(2)は約 3 時間の咽喉マイク音声の FBANK とそのアライメントで学習したハイブリッド方式の音響モデルを有するシステム，(3)は先行研究[3]と同様に接話マイクの BNF で学習した GMM-HMM を音響モデルとし，咽喉マイク音声の FBANK から接話マイク音声の BNF へ LSTM によってマッピングした特徴量を入力とするシステムである．(4)は提案手法（図 1）と同じ DNN の構造を有するが，アライメントから生成した正解ラベルを用いて学習されている．(5)は提案手法のシステムである．(2)の DNN のノード数は[440: 1024: 1024: 1024: 1024: 42: 1024: 出力層]とした．提案手法(5)は従来の(1)と(2)のシステムと比較して CER の大幅な削減を達成でき，先行研究[3]の手法(3)と比べて性能が向上した．また，(5)は同じ DNN の構造を有する(4)のシステムよりも精度が高く，知識蒸留によって学習がうまく進んだと考えられる．

表 2 提案手法の初期化方法に関する認識実験結果

前段初期化	後段初期化	CER
Random	Random	41.0 %
FM-LSTM	Random	41.3 %
Random	CM-Hybrid	**39.3 %**
FM-LSTM	**CM-Hybrid**	39.5 %

次に，提案手法の初期化方法に関する実験を行った．実験結果を表 2 に示す．ここで表中の Random は DNN のパラメータをランダムに初期化すること，FM-LSTM は特徴マッピング用 LSTM のパラメータで初期化すること（図 1-③），CM-Hybrid は接話マイク音声で学習したハイブリッド方式の音響モデルの DNN の後段のパラメータで初期化すること（図 1-④）を表す．結果から DNN の後段を接話マイクのハイブリッド方式の音響モデルの DNN のパラメータで初期化する方法はランダムに初期化するよりも精度が高く，有効である可能性が示された．しかしながら，今回の実験では DNN の前段を特徴マッピング用 LSTM で初期化する方法の有効性は認められなかった．特徴マッピング用 LSTM の学習データが少なく，パラメータの学習が不十分で有効な初期値と成り得なかった可能性が考えられる．

4 おわりに

本研究では DNN の前段を咽喉マイクの FBANK から接話マイクの BNF へのマッピングを学習した LSTM，後段を接話マイクの BNF から音素を識別するよう学習した DNN のパラメータで初期化した後，パラレルデータを用いて知識蒸留に基づきネットワーク全体を最適化する手法を提案し，従来手法と比較して咽喉マイク音声の認識精度を改善することができた．しかしながら，今回の実験では DNN の前段を特徴マッピング用の LSTM のパラメータで初期化する方法の有効性は認められなかった．今後の課題としてモデルのハイパーパラメータの調整や学習データの追加などが挙げられる．

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] Dupont, S. *et al.*, *In Proc. of Robust 2004*, 2004
[2] Shengke, L. *et al*., *In Proc of NCSP,* pp. 363–366, 2018.
[3] 鈴木他, SLP, Vol.2018-SLP-123, pp.1-6, 2018
[4] 神田, 日本音響学会誌, 第 73 巻, 第 1 号, pp.31-38, 2017
[5] Hinton, G. *et al*, arXiv:1503.02531, 2015