Vol.2020-SLP-131 No.4
2020/2/13

咽喉マイクの大語彙音声認識のためのデータ拡張と知識蒸留

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：咽喉付近の皮膚振動を捉える咽喉マイクは外部雑音に頑健であるが，接話マイクとは大きく異なった特性を持つ音が収録されるため，通常の音声認識システムでは認識精度が低下する．また，通常のマイクのように大量の音声データが利用可能という状況にもなく，咽喉マイク収録音声のみで高精度なモデルを学習するのは困難な状況である．本研究では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の特徴量を咽喉マイクの特徴量空間にマッピングし，咽喉マイク用の音響モデルの学習に活用する手法を提案する．特徴マッピングは接話マイクと咽喉マイクで同時に収録された小規模なパラレルデータを用いて LSTM によって学習する．特徴マッピングによって得た擬似的な咽喉マイク特徴量で DNN-HMM ハイブリッド音響モデルを初期学習し，これを生徒モデルとする．一方，大量の接話マイク音声で学習したハイブリッド音響モデルを教師モデルとし，パラレルデータと知識蒸留法による生徒モデルの再学習を行う．新聞記事読み上げ音声を用いた評価の結果，大幅な性能改善が得られたので報告する．

キーワード：咽喉マイク，音声認識，データ拡張，知識蒸留，DNN-HMM，特徴マッピング

1. はじめに

近年，スマートフォンやスマートスピーカの普及に伴い，音声認識技術が様々な場面で活用されるようになった．クラウドベースの潤沢な計算資源の活用に加え，深層学習を用いたモデルの改良によってすでに一部のタスクでは人間と同等以上の認識性能を達成した [1]．一方，高騒音など様々な状況下で高い認識性能を維持することは未だ困難であり，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する方法の一つとして咽喉マイクで音声を収録する方法がある．咽喉マイクは咽喉付近の皮膚振動を捉えるマイクで外部雑音の影響を受けづらく，高雑音環境下での音声認識 [3–9] や発話区間検出 [10, 11]，話者認識 [12, 13] に咽喉マイク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，咽喉マイクは接話マイクとは大きく異なった特性を持つ音が収録されるため，通常の音声認識システムでは認識精度が著しく低下する．また，これまでに咽喉マイクで収録された大規模な音声データベースは存在せず，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モデルを学習するのは困難な状況である．

本研究では咽喉マイク音声の認識精度を改善するための音響モデルの学習手法を提案する．まず，接話マイクと咽喉マイクで同時収録した小規模なパラレルデータを用いて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 Long Short-Term Memory（LSTM）によって学習する．次に，学習済の特徴マッピング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することで大量の擬似咽喉マイク特徴量を生成し，咽喉マイクのデータを拡張する．そしてこの擬似咽喉マイク特徴量を用いて Deep Neural Network-Hidden Markov Model（DNN-HMM）ハイブリッド音響モデルの初期学習を行う．最後に，擬似咽喉マイク特徴量で初期学習済の DNN-HMM を生徒モデル，接話マイク音声で学習された DNN-HMM を教師モデルとして，パラレルデータと知識蒸留法による生徒モデルの再学習を行い，そのパラメータを実際の咽喉マイク特徴量に適応させる．

2. 関連研究

2.1 咽喉マイクを用いた音声認識

咽喉マイクで収録した音声は高周波成分が大きく欠落しており，一般的な気導マイクで収録した音声よりも不明瞭である．それゆえに咽喉マイクのみではなく気導マイクと併用し，雑音環境下で咽喉マイク音声の情報を利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 [3–6]．しかしながら，気導マイク

[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

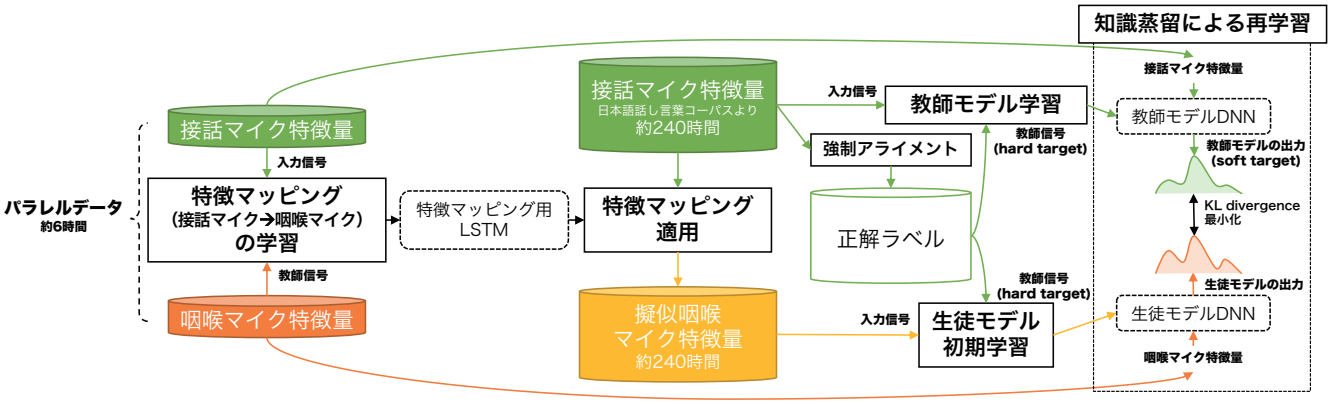

図 1　咽喉マイク用 DNN-HMM のための特徴マッピングによるデータ拡張及び知識蒸留法
Fig. 1　Feature mapping-based data augmentation and knowledge distillation for training DNN-HMM of throat microphone

と咽喉マイクを組み合わせる手法では雑音レベルに応じてどちらのマイクの情報を重要視するかの重みを動的に調整することが困難である．また，二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である．そこで咽喉マイクと気導マイクで同時収録したパラレルデータを利用して，咽喉マイク特徴量空間から気導マイク特徴量空間へのマッピングを Gaussian Mixture Model（GMM）や DNN を用いて学習しておき，認識時にこれを適用することで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 [7–9]．特徴マッピングによって音響ミスマッチを抑制することができ，認識精度も改善することが報告されているが，気導マイクと比べて音素識別に必要な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極めて難しく，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，学習済みの大規模で高精度な 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．教師モデルを一つだけではなく複数用いる [14] 方法や，soft target だけでなく hard target の損失も組み合わせて損失を計算する [15, 16] 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 [17, 18]．また，Hinton らは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報告しており [19]，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 [20–22]．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．我々も接話マイクで収録された大規模な音声データベースを用いて学習した DNN を教師モデルとして，接話マイクと咽喉マイクで同時収録したパラレルデータを用いて同様の手法で咽喉マイク用の DNN を学習した結果，hard target を用いて学習したモデルよりも高い認識精度が得られることを確認している [23]．

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 1 に示す．提案手法は大きく二段階に分けることができ，本章ではまず，特徴マッピングによる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と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（VTLP）[24] や Speed Perturbation [25], SpecAugment [26] などが提案されている．しかしながら，オリジ

図 2 咽喉マイク
Fig. 2 Throat microphone

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る咽喉マイク音声データの拡張手法を提案する．具体的には，先行研究 [7–9] 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．咽喉マイクから接話マイクへの特徴マッピングでは Feed-Forward Neural Network（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．

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．なお，正解ラベルは擬似咽喉マイク特徴量ではなく接話マイク特徴量で学習した GMM-HMM で接話マイク特徴量に対して強制アライメントを行った結果に基づいて推定する．

3.2 知識蒸留法による再学習

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．そこで擬似咽喉マイク特徴量で学習した DNN のパラメータをパラレルデータと知識蒸留法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音声から抽出した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時の 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時の s_i の事後確率を $Q(s_i|x_t)$ とした時，損失関数は以下のように定義される．

$$\begin{aligned} D_{KL}(P||Q) &= \sum_i P(s_i|x_c) \log \frac{P(s_i|x_c)}{Q(s_i|x_t)} \\ &= \sum_i P(s_i|x_c) \log P(s_i|x_c) \\ &\quad - \sum_i P(s_i|x_c) \log Q(s_i|x_t) \end{aligned} \tag{1}$$

なお，式 1 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，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計算しておき，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．

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 2**）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捉えることができる．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わかり [27]，本研究で使用する咽喉マイク音声はいずれも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 6 名による新聞記事読み上げ音声（約 30 分）を使用した．なお，テストの際はどちらか 1ch のみを用いる．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．

4.2 実験方法

特徴量抽出や音響モデルの学習，デコードには Kaldi ツールキット [28] を，特徴マッピングの学習や知識蒸留の実装には Tensorflow [29] を用いた．教師モデルの学習と生徒モデルの初期学習用の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 9300 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）及び 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 の後に全結合層（40 ユニット）を持つ構造とした．なお，LSTM の入力は過去 7 フレームを結合した時系列データとした．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（1024 ユニット）を持ち，出力層のユニット数は CSJ によって学習した 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 [30] 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 Kaldi ツールキットの CSJ レシピに則り，ユニット数が 1024 の Time Delay Neural Network (TDNN) [31] を 6 層重ねたネットワークを用いた．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．学習データには CSJ の約 240 時間の音声に対して 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 4000 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．

表 1　従来手法と提案手法の文字誤り率
Table 1　Character error rate (CER) of conventional and proposed approaches

Model	CER
TM GMM-HMM	15.1 %
TM DNN-HMM	10.8 %
TM DNN-HMM + Speed Perturbation	10.0 %
CM DNN-HMM + Feature Mapping	9.1 %
Map-aug DNN-HMM	8.6 %
Map-aug DNN-HMM + KD	**6.6 %**

いずれの認識実験においても 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 言語モデルとしては Kaldi ツールキットの CHiME-4 レシピに則り，ユニット数が 2048 の TDNN と LSTM-projection（LSTMP）を 5 層重ねたネットワークを用いた．また，3-gram 及び 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 1** に示す．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を適用して学習した 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，今回の 3 種類の係数とした場合よりも高い認

表 2　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 2　Character error rate (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	18.3 %	98.2 %
CM DNN	10.5 %	8.5 %
Map-aug DNN	8.3 %	**6.6 %**

識精度は得られなかった．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更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．

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．生徒モデルの初期化方法として Glorot の一様分布による初期化（Random）[32]，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）の三種類を比較した．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法による方法（soft target）を比較した．なお，正解ラベルは CSJ で学習した GMM-HMM でパラレルデータの接話マイク音声に対して強制アライメントした結果に基づいて推定した．各手法の文字誤り率を**表 2** に示す．乱数によって初期化した場合，知識蒸留法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高域が減衰するなど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一つだと考えられる．

4.3.3　クリーン環境下での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．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生徒モデルと同じ構造を持つ DNN-HMM（CM DNN-HMM）と教師モデル（CM TDNN-HMM）を用いた．接話マイク用音響モデルはいずれも CSJ で学習した．各モデルの文字誤り率を**表 3** 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同規模の音響モデル（CM DNN-HMM）を用いた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．ただし，より大規模で高性能な音響モデル（CM TDNN-HMM）を用いた接話マイク音声の認識精度とは未だに差は大きく，今後も更なる改善が必要である．

表 3　クリーンな接話マイク音声と咽喉マイク音声の文字誤り率
Table 3　Character error rate (CER) of close-talk and throat microphones clean speech

Model	Input	CER
Map-aug DNN-HMM + KD	**Throat mic**	**6.6 %**
CM DNN-HMM	Close-talk mic	5.6 %
CM TDNN-HMM	Close-talk mic	4.7 %

5.　おわりに

本研究では既存の大規模な接話マイク音声データと特徴マッピングを用いたデータ拡張手法に加えて咽喉マイクと接話マイクの小規模なパラレルデータと知識蒸留法による咽喉マイク用 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 40%（10.8%→6.6%）の文字誤り率の削減を確認した．今後は複数の教師モデルからの知識蒸留や雑音環境下での評価等を行う予定である．

謝辞　本研究の一部は科研費（16H01817, 18H03260）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 Transactions on 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, *Proceedings of Robust 2004 (Workshop (ITRW) on Robustness Issues in Conversational Interaction)* (2004).
[4] Panikos, H., Jani, E., Ishi, C. T., Takahiro Miyashita 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] Nigade, A. S. and Chitode, J. S.: Throat Microphone Signals for Isolated Word Recognition Using LPC, *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2, pp. 401–407 (2012).
[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.: Bottleneck feature-mediated DNN-based feature mapping for throat microphone speech recognition, *APSIPA ASC 2018*, pp. 1738–1741 (2018).
[10] 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).
[11] 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*, No. January 2010, pp. 1978–1982 (2010).
[12] Yegnanarayana, B., Shahina, A. and Kesheorey, M. R.: Throat microphone signal for speaker recognition, *Interspeech* (2004).
[13] 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*, Vol. 08-12-Sept, pp. 1720–1724 (2016).
[14] Chebotar, Y. and Waters, A.: Distilling Knowledge from Ensembles of Neural Networks for Speech Recognition, *Interspeech*, pp. 3439–3443 (2016).
[15] Yang, Z., Zhang, C., Zhang, W., Jin, J. and Chen, D.: Essence Knowledge Distillation for Speech Recognition, *ArXiv*, Vol. abs/1906.10834 (2019).
[16] Tachioka, Y.: Knowledge Distillation Using Soft and Hard Labels and Annealing for Acoustic Model Training, *GCCE*, No. 2, pp. 715–716 (2019).
[17] 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).
[18] Chan, W., Ke, N. R. and Lane, I.: Transferring knowledge from a RNN to a DNN, *Interspeech*, pp. 3264–3268 (2015).
[19] Hinton, G., Vinyals, O. and Dean, J.: Distilling the Knowledge in a Neural Network, *NIPS Deep Learning and Representation Learning Workshop*, (online), available from ⟨http://arxiv.org/abs/1503.02531⟩ (2015).
[20] 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).
[21] 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).
[22] 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).
[23] Suzuki, T., Ogata, J., Tsunakawa, T., Nishida, M. and Nishimura, M.: Knowledge Distillation for Throat Microphone Speech Recognition, *Interspeech*, pp. 461–465 (2019).
[24] 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).
[25] Ko, T., Peddinti, V., Povey, D. and Khudanpur, S.: Audio Augmentation for Speech Recognition, *Interspeech* (2015).
[26] 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).
[27] Suzuki, T., Ogata, J., Tsunakawa, T., Nishida, M. and Nishimura, M.: Effects of Mounting Position on Throat Microphone Speech Recognition, *GCCE*, pp. 897–898 (2019).
[28] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. and Vesely, K.: The Kaldi Speech Recognition Toolkit, *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society (2011).
[29] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems (2015).
[30] 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).
[31] 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).
[32] Glorot, X. and Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Teh, Y. W. and Titterington, M., eds.), Proceedings of Machine Learning Research, Vol. 9, Chia Laguna Resort, Sardinia, Italy, PMLR, pp. 249–256 (2010).