多様な環境におけるレアリソースの音声認識

代表研究者　　王龍標　　長岡技術科学大学　電気電子情報工学専攻　准教授

1 研究調査の要旨

グローバル化が進むに従い語学力の重要性がますます高まっている。このような状勢で、非母国語話者を含む人間同士が、様々な場面で英語や日本語を使ってコミュニケーションをとる機会がますます増えている。しかし、非母国語を自由に操る能力を有する人は少ない。従って、情報処理技術を利用して人間同士のコミュニケーションを支援する技術が必要となる。流暢な外国語も必要なく、多様な環境（屋内・屋外環境、近接・遠隔発話）におけるいつ・誰・どこでも効率よくユビキタス環境における音声認識の研究が今までないが、グローバル化に欠けないものである。本研究では、これまでの世界先端な音声認識と話者認識技術を発展さて、多様な環境における実環境の音響信号処理、非母国語の音響モデルと言語モデルの自動適応、話者の出身国の自動推定と非母国語の音声認識（レアリソースの音声認識）の高精度化の研究を行う。自動で推定した残響成分を環境情報として DAE（denoising autoencoder）に追加する手法（dereverberation-aware DAE)、DNN（deep neural network）と GMM（Gaussian mixture model）の融合による母国語認識手法、ボトルネック特徴を用いる cross-accent SGMM（subspace GMM）に基づく非母国語話者の音声認識手法を提案し、残響環境下での音声認識の単語誤り率を従来法の 43.19%から提案法の 26.83%、母国語認識の認識精度を従来法の 90.7%から提案法の 97.5%、非母国語話者の音声認識（レアリソース音声認識）の単語誤り率を従来法の 31.13%から提案法の 25.26%へ大きく改善した。

2 研究背景

近年の国際化の発展により、英語を母国語としない話者による英語の利用が一般的となりつつある。母国語以外の言語を非母国語と呼ぶが、非母国語の利用が求められる状況において音声認識技術の応用を考えた場合、非母国語話者の音声認識の精度は不十分であるため、非母国語話者の音声認識の精度改善が求められる。本研究では、多様な環境におけるレアリソース音声認識の精度を改善するための残響除去、母国語認識と非母国語話者の音声認識を注目する。

2.1 残響除去

遠隔発話環境下で音声を収録する場合、雑音や残響の混入は避けられない問題である。雑音や残響によって引き起こされる音声の歪みは、学習時と認識時の収録環境にミスマッチを生じさせ、音声認識性能を著しく低下させてしまう。このような問題に対処するためのアプローチは多数提案されており、それらはフロントエンドベースとバックエンドベスの2 つに大きく分けられる。フロントエンドベースは観測された音声信号から雑音や残響の成分を推定しそれらを除去するアプローチであり、バックエンドベースは音響モデルやデコーダを雑音や残響の環境に適応させるアプローチである。

雑音や残響の成分を推定する方法として、シングルチャンネルもしくはマルチチャンネルマイクロフォンを利用したものが数多く提案されている。一般的に、マルチチャンネルマイクロフォンを利用した方が情報量も多く雑音や残響の成分を推定しやすい。しかし、マルチチャンネルマイクロフォンを用意するのは金銭的、時間的コストがかかる欠点がある。一方、シングルチャンネルアプローチは、一般的なスマートフォンに搭載されているようなマイクでも利用が可能であり、手軽さという点ではマルチチャンネルアプローチより優れている。そこで、本研究ではシングルチャンネルの音声信号を利用した方法に着目する。シングルチャンネルアプローチの代表例として、ケプストラムのフレーム平均を全フレームから減算することで特徴量の段階で補正を行うケプストラム平均正規化(CMN)が広く知られている。一般的に、短時間スペクトル分析窓長には25ms 程度の長さが用いられており、残響の長さが分析窓長よりも短い場合にCMN は有効な方法となる。しかし、一般的な室内の遠隔環境下では、残響の影響は100ms から1000ms 程度の範囲におよび、短時間スペクトル分析窓長より非常に長くなってしまう。そのため、CMNを用いることで、フレーム内の初期残響は除去できるが、フレーム長より長い後部残響は除去できないという問題がある。このような問題に対応するために、残響を加算性雑音とみなしスペクトルサブトラクションを利用して除去する方法が話者認識の分野で提案されている。また、シングルチャンネルとマルチチャンネルマイクロフォンの両方に対応したMulti Step

Linear Prediction（MSLP）に基づく方法も提案されている。

2.2 母国語認識

母国語認識とは、発話者の発話から、その発話の言語には関係なく、その人の生まれ育った土地で学んだ言語を同定することを指し、これに対し言語識別は現在発話している言葉が何語かを同定する。近年では、このような母国語認識や、地方の方言を認識する方言認識といったアクセントを同定する技術が注目されており、アクセント認識率は向上しつつある。このような技術は、例えば国際会議において音声認識技術を用いる場合、録音装置に対し遠隔状態でも、このアクセント認識技術によってアクセント情報を同定し、情報に合ったシステムに切り替えることで、非ネイティブ話者のスピーチにおいても自動的な議事録の作成が可能となる。また、コールセンターでは、アクセントに合った話者に切り替えることで、円滑に対応できるようになる。しかしながら、遠隔環境下では残響が入ってしまうため、アクセント認識率が下がってしまう。そのため、本研究では、母国語認識に注目し、遠隔環境下に頑健な母国語認識を行う。

近年では、母国語認識と方言認識といったアクセント認識があるが、これは音声認識において認識率を改善する一つの鍵となっている。ここではそのアクセント認識のアクセント認識率を上げるための従来でのアプローチを紹介する。音声認識においてアクセントはパフォーマンス悪化の原因の一つであり、自動アクセント認識は、今の音声研究の話題となっている。アクセント認識である方言認識においては、従来のGMMベースとしたアクセント認識が存在する。しかし、一般に用いられるGMMはMLE（Maximum likelihood estimation）で、これは判別可能な学習方法ではなく、多数のGMMの間のEER(Equal Error Rate)を抑えられないため、アクセント認識率が下がってしまう。また、遠隔環境下においては性能が悪くなってしまう。この問題を解決する必要があるため、本研究では1つの工夫として判別可能な学習方法を持つDNN（Deep Neural Network）を用いる。DNNは、深い層の学習により段階的に識別可能となるように特徴量を最適化することで、識別最大化、無駄な情報の残響除去が行われ、結果的にＧＭＭより判別率の向上に繋げることができると考えられる。次に、GMMとDNNの性質の違いに着目した。DNNは、GMMよりも母国語認識率を向上させることが出来るが、学習特徴量に強く依存してしまう。そのため、学習特徴量から少しでも外れる特徴量が入力された場合、正しくマッピングできない。このような問題に対して、GMMはDNNより特徴量を大まかにマッピングする性質がある。これらの性質の違いにより、提案法として出力されるスコアを組み合わせることで、更に母国語認識率を向上させると考えられる。

2.3 非母国語話者の音声認識

非母国語の発話には、その発話者の母語による干渉が生じる。母語の干渉は、音素などに影響を与え、その特徴を変動させる。そのため、同一の言語であっても、発話者がもつ母語によって発話の音響的特徴に差が生じる。音声認識で使用される音響モデルの作成には、その言語を母語とする人が発話した音声を学習データとして利用するのが一般的である。ここで、ある言語について非母国語話者が音声認識を利用することを考えると、母語の干渉による音響的特徴の変動から、発話された音声の特徴と音響モデルがもつ特徴との間に差が生じる。この特徴の差が誤認識の原因となり、結果的に認識率の低下へと繋がる。このように発音の変動が音声認識の精度の悪化の原因となっている。このような認識率の低下を防ぐために、音響モデルの作成に使用する学習データをすべて同一の母語をもつ非母国語話者の音声とする方法が考えられるが、非母国語話者の音声は非常に少なく、音響モデルの学習データとして用いるには不十分であるといえる。このことから、非母国語話者の音声認識の精度を改善するには、少量の学習データでも非母国語話者に対応した音響モデルを作成する手法が必要となる。近年、音響モデルに深層ニューラルネットワーク（Deep Neural Network: DNN）を用いる音声認識の研究が数多くされている。DNNは任意の入出力の非線形関数を多数のネットワークによって表現することができる。DNNの音響モデルはこのような特徴から様々な音声認識のタスクで認識精度の向上を示した。しかし、非母国語話者の音声認識で利用するにはいくつかの問題点がある。DNNの学習は事前分布を仮定しないため、そのパラメータの調整に多量の学習データが必要となる。先も述べたように、非母国語音声は比較的学習データが少ない傾向にあるため、十分な学習を行うには学習データが不足することが考えられる。一方、低リソースの音声認識を対象とした手法として部分空間混合ガウスモデル（Subspace Gaussian Mixture Model ： SGMM）の研究が行われている。SGMMの基本の構造は混合ガウスモデル（Gaussian Mixture Model: GMM）と共通であるが、SGMMのパラメータは部分空間の概念を利用して設定される。これによりパラメータの低次元表現を実現し、低リソースの学習を可能とした。本研究では、非母国語話者の音声認識の精度改善を目的とし、非母国語話者に対応した音響モデル学習の手法と、DNNによる特徴量変換の手法を提案する。非母国語話者の音声認識は低リソースの条件であるため、音響モデルとしてSGMM

を利用することが有効であることが考えられる。さらにSGMMは異なる種類の音声を学習データとして複数用いた場合に、その差を考慮した学習が可能であるため、母国語話者の音声と非母国語話者の音声の両方を利用する学習方法を提案する。この手法で作成される音響モデルをcross-accent SGMMと定め、提案する。また、DNNは多数のネットワーク表現される非線形関数から任意の変換が可能であることから、DNNを特徴量変換器として利用する手法を提案する。これらの手法について非母国語話者の音声認識実験において評価を行う。

3 提案手法

3.1 Reverberation-aware Denoising autoencoder (DAE)

本研究では、図１に示すようにMSLPで推定した残響成分を環境情報としてDAE に追加する手法を提案する。本手法を略記するときには、Reverberation-aware DAEと呼ぶ。MSLP に基づく残響除去は，残響下の音声から残響成分を推定し、その成分を元の音声からスペクトル減算法（ＳＳ）を用いて取り除くことで残響除去を実現している。推定した残響成分は、音声を収録した部屋や話者とマイクロフォンの距離に依存するため、収録環境をよりよく表現した情報といえる。MSLPに基づく残響除去では、得られる残響成分をもとに残響成分のスペクトルを求め、それをSS法を用いて観測信号から減算するが、本手法ではMSLP で推定した残響成分のスペクトルをもとに残響成分のMFCC（Mel-frequency cepstral coefficient）を抽出し、その値を環境情報とする。本手法はMSLPに基づく残響除去法で残響除去を行っていたSS 法の部分をニューラルネットワークによるものに置き換えたといえる。SS法は線形的な処理であるが、ニューラルネットワークによる変換は非線形的であるため、SS法と比べてより表現力の高い処理で残響除去ができると考えられる。

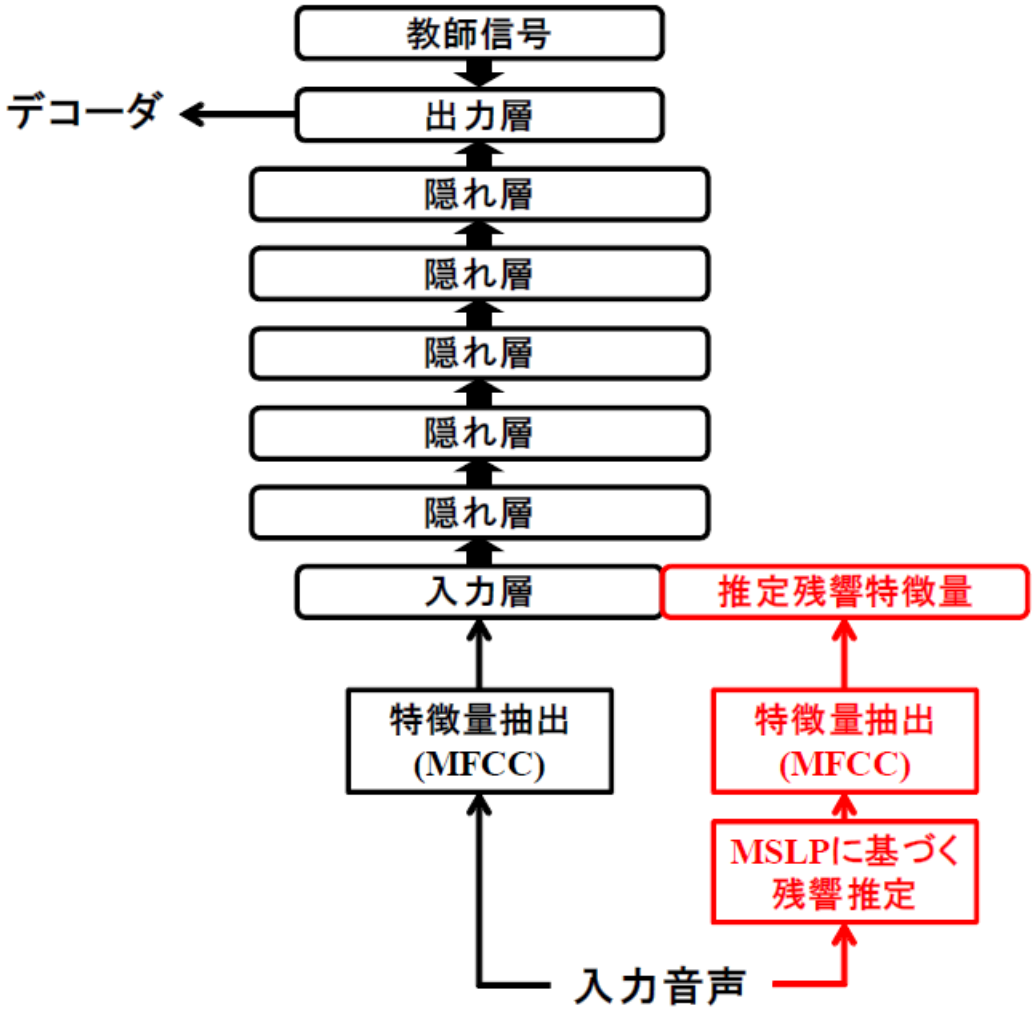

図1　Reverberation-aware DAE

3.2 DNNとGMMの融合による母国語認識

GMM と比べ、DNN にはいくつかの利点がある。それは、①識別学習　②多段階層化ネットワークによる表現力が大きいこと ③複数フレーム入力することで、データの前後相関が取ることが可能なことである。①識別学習は、ＤＮＮは、深い層の学習により段階的に識別可能となるように特徴量を最適化するため、識別が最大となる。そのため、母国語認識情報に必要のない残響情報を段階的に除去することが考えられ、遠隔発話においても効果を期待できる。また、特徴量の変換については画像認識分野でが有名である。②多段階層化ネットワークによる表現力は、いくつもの層を重ねることで、それぞれの層において任意の分布で特徴を表現することができる。これにより、GMM よりもより正確に特徴の分布を表現することが可能である。③複数フレーム入力することで、データの前後相関が取ることが可能というのは、DNN に特徴を入力すると、次の層のノードとのネットワークにより、複数フレームの相関が取れるということである。

ＤＮＮは GMM よりも母国語認識率を向上させることが出来る。しかし、ＤＮＮは表現能力が大きく、学習特徴量を任意の分布で忠実にマッピングするため、学習特徴量に大きく依存してしまい、少しでも学習特徴量に外れ値が存在すると、その外れ値もマッピングしてしまう。つまり、テスト時に微妙に外れた特徴量が入力されると正確に認識できないことになる。それに対し GMM は、ＤＮＮに比べ表現力が小さく混合ガウス分布によるマッピングを行うため、学習特徴量を大まかにマッピングする。これは、少々例外の特徴量が存在しても無視する。つまり、ＤＮＮで誤認識した特徴量をＧＭＭでは認識する可能性が出てくる。これらＤＮＮとＧＭＭの性質の違いをお互いにカバーするよう、お互いの出力スコアを組み合わせ母国語認識率の向上を図る。

3.3 ボトルネック特徴を用いるcross-accent SGMMに基づく非母国語話者の音声認識

音声認識に利用する音響モデルとして、cross-accent SGMM を提案する。学習に利用できる音声データが少ないことから、音響モデルには少量の学習データでも十分な学習を行うことのできる SGMM を利用する。さらに、UBM (universal background model)学習には SGMM 学習と同様の学習データを用いるのが通常の SGMM の学習方法であるが、UBM 学習に多くのデータが利用可能な母国語話者の音声を学習データとして利用することで認識率の改善を図る。この UBM 学習に母国語話者の音声、SGMM 学習に非母国語話者の音声を利用する SGMM を Cross-accent SGMM と定め、提案する。

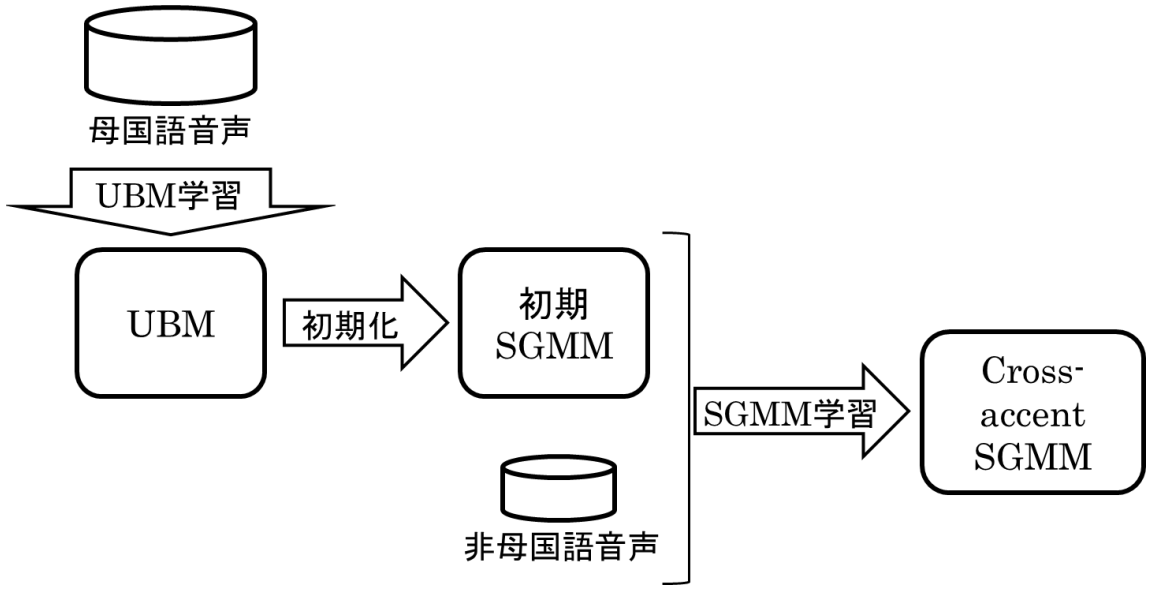

図 2　Cross-accent SGMM の学習手順

Cross-accent SGMM は、少量の学習データでも十分な学習を行うことのできる SGMM を母国語話者の音声と非母国語話者の音声の両方を使用して学習する音響モデルである。Cross-accent SGMM の学習の概略図を図 2 に示す。SGMM の学習は UBM 学習と SGMM 学習の 2 段階で行われる。通常の SGMM では、これらの学習は同一の学習データが使用される。それに対して、cross-accent SGMM では、UBM 学習では母国語話者の音声、SGMM 学習では非母国語話者の音声を使用して学習を行う。UBM 学習では言語内のすべての音素について一纏めに学習が行われるため、その言語の傾向を学習することとなる。そのため、母語の干渉の影響を受けた非母国語話者の音声を学習に利用する場合よりも、その言語を母語とする話者の音声を学習に利用することでより性能のよい UBM の作成が可能であると考えられる。また、それはデータの量の面から見ても同様であり、より豊富な学習データが存在する母国語話者の音声を利用したときによりよい UBM が作成可能である。

SGMM 学習は UBM から作成された初期 SGMM のパラメータの調整によって行われる。そのため、SGMM 学習は少量の学習データでも十分な学習を行うことができる。そこで、非母国語話者の音声を使用して学習することで母語の情報を考慮した非母国語話者に対応した SGMM を作成する。

本研究では DNN を非母国語話者の音声認識に有効な特徴量を得るための特徴量変換器として用いる手法も提案する。ボトルネック DNN とは、ユニット数が他の層と比べて極端に少ないボトルネック層を中間層に設けた DNN である。一般的に、ボトルネック DNN には複数フレームの特徴量を与え、中間に設けられたボトルネック層によって、入力特徴量を非線形圧縮する。この圧縮された特徴表現をボトルネック特徴量と呼ぶ。ボトルネック特徴量は高次元の入力特徴量を低次元表現に変形したものであるため、入力特徴量の冗長な成分を取り除き、より抽象的な特徴表現となることが期待される。

ボトルネック DNN の学習法は、一般的な DNN と同様であり、pre-training では各層ごとに RBM を事前学習し、DBN（deep belief network）を構築する。Fine-tuning では教師信号として音素クラスラベルを与え、BP（backpropagation）法によって学習する。このとき、ボトルネック DNN は音素の識別性を高めるようにパラメータを更新するため、ボトルネック特徴量は音素の識別性を間接的に反映する。ボトルネック特徴量の抽出の流れは図 3 に示すように、入力層に音声特徴量を入力したときのボトルネック層の値を特徴量として抽出することで行われる。非母国語話者の音声認識実験においてボトルネック特徴量を利用することで、母語の干渉の影響による変動した特徴量を補正して、認識性能を向上させることが期待される。

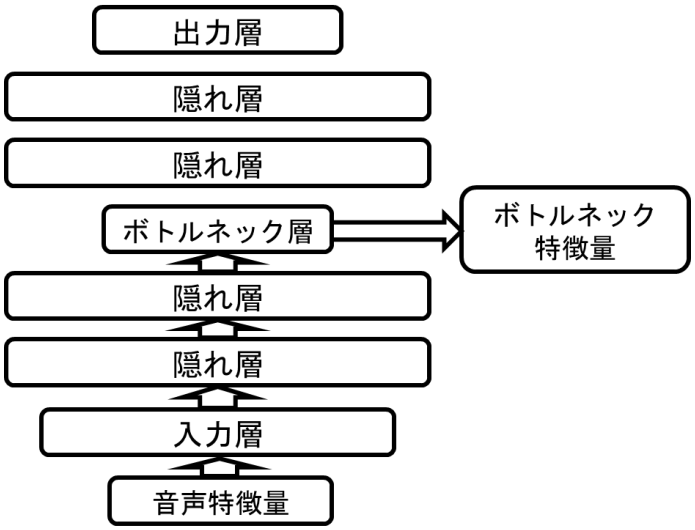

図 3　ボトルネック特徴量の抽出の流れ

このようなボトルネック特徴量を非母国語話者の音声認識に利用する場合、図 4 に示されるような流れになる。Cross-accent SGMM の学習にはすべてボトルネック DNN によって、ボトルネック特徴量に変換されたデータを使用する。そのため、非母国語音声データ、母国語音声データの両方に変換が適応される。ただし、それぞれの音声データは非母国語音声であれば非母国語音声で学習を行ったボトルネック DNN、母国語音声であれば母国語音声で学習を行ったボトルネック DNN で変換を行う。また、認識の対象となる音声についても変換が行われ、ボトルネック特徴量による認識を行う。これにより認識性能の向上を図る。

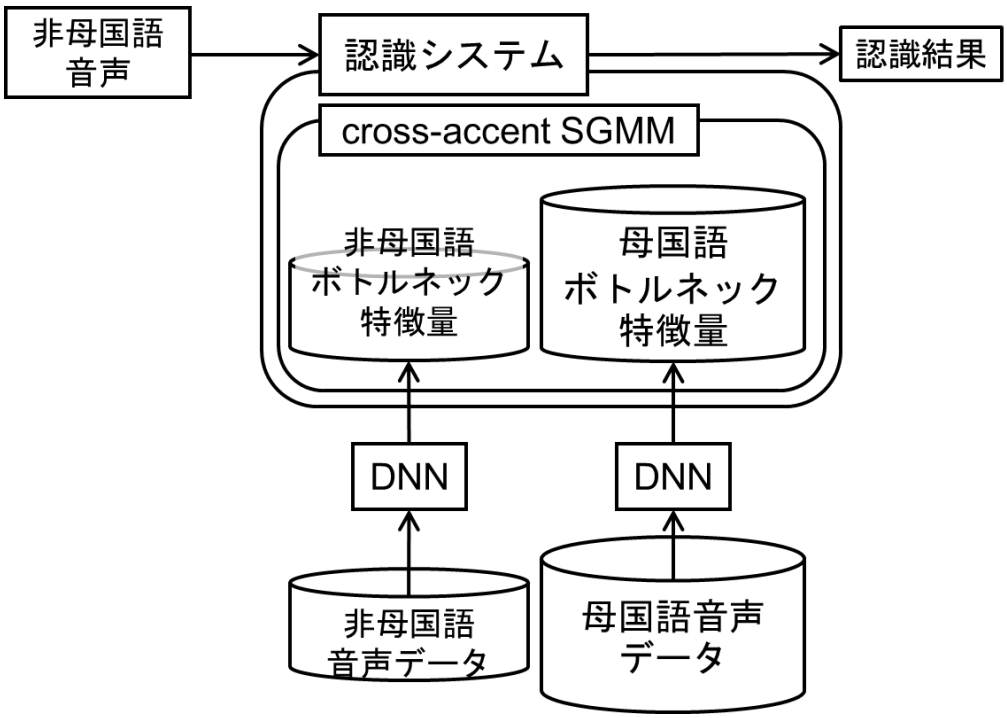

図1　ボトルネック特徴量を利用した非母国語話者の音声認識の概略

4 実験

4.1 残響除去実験

本研究では、“REVERB-challenge”(残響下音声の音声強調と音声認識ベンチマーク)が提供する学習データを使用する。残響除去のために用いるDAE へ与えるデータとして、音声の13 次元MFCC にそれぞれのΔ、ΔΔ特徴量も加えた1フレームあたり計39次元の特徴量を用いる。一般的に残響は音声の複数のフレームに渡って影響を及ぼすため、DAE の入力に複数フレームの特徴量を与えることで音声フレームの前後からの影響も同時に学習させることが有効であると考えられる。そのため本研究では、入力層および教師信号のそれぞれに9フレーム分の特徴量を与える。音声認識のために使用する音響モデルは、SGMMとDNN-HMM（hidden Markov model）の2種類を用いる。SGMMとDNN-HMMの2つの音響モデルから得られる音声認識結果をCNC（confusion network combination）で統合したものを最終的な認識結果とする。言語モデルは、5000単語トライグラムモデルを用いる。評価指標として単語誤り率(WER)を用いる。

表１　各手法による単語誤り率（%）

手法	結果
CMN	43.19
MSLP	35.81
DAE	30.56
Reverberation-aware DAE（提案法）	26.83

各手法による単語誤り率の比較により提案手法の評価を行う。各手法による音声認識結果を表１に示す。CMN による結果と比較して、他の手法は大きく改善していることがわかる。このことから、CMNだけでは初期残響の影響は抑圧できても、後部残響の影響を抑圧するには十分ではないことがわかる。MSLP による推定残響とDAEを組み合わせたReverberation-aware DAE が従来法と比較して大きな改善が見られた。特に基盤技術

である従来のDAEとWERで比較すると、12.4%の誤り削減率を達成した。また、同様の推定残響を用いているMSLPと比較しても、WER は大きく改善している。このことから、従来のように未知環境の音声からMSLP で残響を推定しSS 法を用いて線形的に残響を除去するのではなく、推定残響を明示的にDNN に与えることで非線形的に除去を行うことが有効であることがわかる。

4.2 母国語認識実験

本節では、人工残響環境下母国語認識タスクにおける評価実験を行い、提案手法の有効性を確認する。本研究では、音響特徴量として MFCC を用い、話者モデルとして 128 混合の GMM を用いる。評価実験を行うために、事前に DNN 最適なモデルパラメータを決定する必要がある。本研究では、DNN のパラメータを変化させ、母国語認識率が最良となったパラメータを最適なモデルパラメータとして経験的に決定する。評価実験では、決定したパラメータを使用し、従来法との性能を比較することで、提案手法の有効性を確認する。

音声認識技術を利用する実際の環境を考えると、音声には残響がついてしまうので、母国語の認識の妨げになる。よって、本研究では収録環境を想定し、残響を付加したものも使用し、実験を行う。その前準備として Clean 音声に対しインパルス応答を畳み込む。学習音声には CENSREC-4 のインパルス応答を畳み込み、テスト音声には RWCP のインパルス応答を畳み込んだ。

提案法は、手法 1 である GMM と手法 2 の DNN を組み合わせた方法で、GMM の尤度と DNN の出力を 0～1 の値に正規化し、正規化した 2 種類の手法の出力に対し(1)式のように 1 発話毎に線形結合をし、その和の最大値を判別することによって母国語認識をする。提案法での DNN は、手法 2 で最も結果が良くなったパラメータを使用する。

$$L = \alpha L_{GMM} + (1 - \alpha) L_{DNN} \qquad (1)$$

ここでのL_{GMM}は手法 1 である GMM の尤度を正規化した物を、L_{DNN}は手法 2 である DNN の出力を正規化した物を示す。また、α は結合の重みであり、0～1.0 まで 0.1 刻みで計算する。この時、α は組み合わせ手法の母国語認識率が最良となるように、経験的に決定する。

遠隔環境下における従来法と案法の実験結果を表２に示す。表２から、提案法は従来法より母国語認識率が良くなったことがわかる。73.1%の相対誤り率の削減を達成した。

表２　各手法による遠隔環境下における母国語認識率(%)

	従来法	提案法
認識率	90.7	97.5

4.3 非母国語音声認識実験

この節では、非母国語話者の音声認識タスクにおける音響モデルの評価実験を行う。本研究では、音響特徴量として、特徴量として MFCC を用いる。各音響モデルの学習データには、非母国語話者音声として日本人英語音声の ERJ (English speech database Read by Japanese students)、母国語話者音声としてネイティブ英語音声の TIMIT (the Texas Instruments / Massachu-setts Institute of Technology) を用いる。表 3 にそれぞれの学習データの詳細を示す。評価データも同様に、非母国語話者音声として ERJ、母国語音声として TIMIT を用いる。表 4 に評価データの詳細を示す。また、言語モデルには WSJ (Wall Street Journal) コーパスによって学習を行った trigram の言語モデルを使用する。

評価実験では、非母国語話者の音声認識において従来の音響モデルとの性能を比較することで、提案手法である Cross-accent SGMM の有効性を確認する。

表 3　学習データの詳細

	ERJ	TIMIT
発話言語	英語	
母国語	日本語	英語
話者数	100 人	462 人
発話数	1000 発話	3696 発話
発話時間	0.8 時間	3 時間

表 4　評価データの詳細

	ERJ	TIMIT
発話言語	英語	
母国語	日本語	英語
話者数	102 人	24 人
発話数	1020 発話	192 発話
発話時間	0.8 時間	0.2 時間

以下に、評価実験に用いる提案手法と従来手法を示す。

提案手法

Cross-accent SGMM: 非母国語音声と母国語音声で学習を行った SGMM

従来手法

Non-native SGMM: 非母国語音声で学習を行った SGMM

表1　音響モデルによる単語誤り率(%)の比較

評価データ	Non-native SGMM	Cross-accent SGMM
結果	31.13	29.18

SGMM について、従来手法と cross-accent SGMM で同じパラメータを利用する。SGMM は混合数 400 に設定したものを利用する。非母国語話者の音声認識タスクによって cross-accent SGMM の評価実験を行う。表 5 に非母国語話者の音声認識の認識結果を示す。Cross-accent SGMM は従来の Non-native SGMM を上回る結果となった。理由としては、cross-accent SGMM の場合、UBM 学習では母国語音声、SGMM 学習では非母国語音声を使用する。それぞれの音声を異なる学習の段階で利用することにより、母語が異なっていても学習への影響を回避することができたと考えられる。さらに、母国語音声で UBM 学習を行うことで母語に依存しない情報をより多くのデータで学習することができ､その後の非母国語音声による SGMM 学習で母語に依存する情報を学習することで非母国語話者に対応した音響モデルが作成できた。これにより、cross-accent SGMM の非母国語話者の音声認識への有効性が確認できた。

Cross-accent SGMM では非母国語音声と母国語音声の両方のデータを使用するため，ボトルネック特徴量を抽出するためにはそれぞれ別のボトルネック DNN を必要とする。そのため、非母国語音声で学習を行ったボトルネック DNN と母国語音声で学習を行ったボトルネック DNN の 2 つを作成した。以下よりボトルネック DNN の作成手順を述べる。はじめに、それぞれの音声データによる事前学習で各層の事前学習を行う。次に、同様の音声データで教師信号を音素ラベルとした fine-tuning を行うことでボトルネック DNN を作成する。学習では特徴量はすべて MFCC を利用する。音響モデルの学習の際には、学習音声データから得られる MFCC をボトルネック DNN に入力し、それによって抽出されるボトルネック特徴量を学習データして cross-accent SGMM の学習を行う。認識の際も同様にして抽出されたボトルネック特徴量を認識に利用する。ボトルネック

特徴量（BF-feature）は識別性の高い特徴量である。それにより認識の精度の改善が期待できる。結果を表６に示す。BN-feature が従来の MFCC を上回る結果となった。その相対改善率は 13.43%である。これにより非母国語話者の音声認識におけるボトルネック特徴量の利用の有効性が確認できた。

表 6　特徴量変換による単語誤り率(%)の比較

特徴	MFCC	BN-feature
結果	29.18	25.26

5 まとめ

本研究では、多様な環境における実環境の音響信号処理、非母国語の音響モデルと言語モデルの自動適応、話者の出身国の自動推定と非母国語の音声認識（レアリソースの音声認識）の高精度化の研究を行う。自動で推定した残響成分を環境情報として DAE に追加する手法（dereverberation-aware DAE)、DNN と GMM の融合による母国語認識手法、ボトルネック特徴を用いる cross-accent SGMM に基づく非母国語話者の音声認識手法を提案し、残響環境下での音声認識・母国語認識と非母国語話者の音声認識の性能を大きく改善した。

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
機械学習を用いた母国語判別	音声研究会（SP） 音学シンポジウム 2014	2014.5
GMM と DNN を組み合わせた遠隔環境下での母国語認識	音声研究会（SP） 第 6 回集合知シンポジウム	2014.12
SGMM と DNN を併用した非母国語話者の音声認識	日本音響学会 2015 年秋季研究発表会 講演論文集, pp. 69-72	2015.9
A Spectrum Smoothing Method for Speaker Verification	Proc. of APSIPA ASC 2015, pp. 1291-1295	2015.12
Environment-dependent denoising autoencoder for distant-talking speech recognition	Eurasip Journal on Advances in Signal Processing, 2015:92, pp. 1-11	2015.12
Distant-talking accent recognition by combining GMM and DNN	Multimedia Tools and Applications, Vol. 75, No. 9, pp. 5109-5124	2016.1