多様な環境における会議音声の音声認識と発言者識別に関する研究

Robust speech recognition and speaker diarization under various environments

2031005

研究代表者　長岡技術科学大学　准教授　王　龍標

［研究の目的］

音声は訓練なしで誰もが使えること，情報伝達が高速であること，使用に際して特に道具を必要としないことなどの利点を持つ，自然で重要な人間のコミュニケーション手段である。人間と機械のコミュニケーションでは，音声情報処理が注目されて期待は大きい。最近，スマートフォンで音声認識や音声検索，音声翻訳など様々な技術が実用化され，大きく注目されている。

多人数の会議では，便利性・コストを考慮しマイクを意識せずマイクから離れた遠隔環境下の自然発話が多い。自動会議録作成支援システムは人手で作成する会議録に比べてコストが低く，アーカイブ化により会議内容の検索が容易になるなどの利点があり，学術的・商業的にも非常に意義がある。しかし，雑音・残響を含む多様な実環境下で，外乱の影響により遠隔発話音声認識は非常に難しくなる。さらに，会議録の作成では，発話内容の書き起し以外に，いつ誰が発話したか（speaker diarization：以降，発言者識別と定義）も発話者毎の内容の整理・検索に重要な情報である。これに対して本研究は，申請者の世界先端の音声情報処理技術を発展させ，人間と機械の調和を促進するため，①音響信号処理と人間の音声認知の観点から遠隔環境下の自然発話に頑健な音声認識・話者認識・発言者識別手法および，②話者変動・チャネル変動の補正技術を用いる話者識別に関する研究を行う。

［研究の内容，成果］

・MCLMS に基づく残響環境下の話者認識

音声に含まれる残響を除去することは非常に困難である。残響（伝達特性）を除去する方法として，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（Cepstral Mean Normalization : CMN）が一般的に用いられている。一般的に，短時間スペクトル分析窓長には 25 ms 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，残響の長さは 100 ms から 1000 ms 程度の範囲をとり，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。そこで，申請者の先行研究において，マルチチャンネル最小平均二乗（Multi-channel Least Mean Square : MCLMS）アルゴリズムに基づくスペクトル減算法を用いた残響除去法という音声認識ための残響除去技術が提案された。これは，スペクトル領域における後部残響が付加雑音とみなし，MCLMS アルゴリズムに基づいて推

定したインパルス応答から得られる後部残響を雑音除去技術であるスペクトル減算法を用いて除去する方法である。本研究は MCLMS 手法を遠隔環境下の話者認識へ発展させた。しかし，MCLMS に基づく後部残響除去手法では，残響を除去するが，話者の特性も除去してしまう。この問題に対して，話者モデルを作るときに，人工的に雑音・残響を音声に付加して，スペクトル減算による雑音・残響を除去する。270 名の話者で様々な残響環境で，話者認識実験を行った。提案法は，学習段階と評価段階の話者のひずみを軽減できるため，性能を従来法の 88.4% から提案法の 91.7% に改善した。さらに，残響除去の補正パラメータの自動選択と高速化の方法（提案法の改善法）も提案し，認識結果がさらに良くなって，93.6% になった。以上の結果は EURASIP の海外論文誌に採択し，掲載された「成果 4」。

・DAE に基づく雑音・残響環境下の音声認識

MCLMS は線形変換に基づく手法で，残響による非線形の伝達関数の悪影響を軽減するのは限界がある。そこで，本研究は Deep Neural Network（DNN）の非線形変換の特徴を活かして，残響除去に発展させた。Denoising Autoencoder（DAE）はニューラルネットワークの 1 種で，入力にノイズを付与したデータを，教師信号に入力データに対応するクリーンなデータを与えることで，ノイズの乗ったデータをクリーンなデータに変換させるモデルであり，画像処理の分野でのノイズ除去技術として利用された。本研究では，入力に残響を畳み込んだ音声特徴量を，教師信号に対となるクリーン音声特徴量を与えることで残響音声をクリーン音声に変換するような非線形変換関数を学習する。一般的に，残響は観測フレームよりも過去のフレームに依存しているため，本研究では観測フレームに加えて，前方数フレームを入力特徴量として与え，教師信号として観測フレームのクリーン信号を与えるようなモデル構造を使用する。DAE で使用するニューラルネットワークのモデル構造を図 1 で示す。DAE によって出力される特徴量は以下の式で計算できる。

$$O_i = f_L(\cdots f_l(\cdots f_2(f_1(X_i, X_{i-1}, \cdots, X_{i-N}))))$$

ここで，f_lは l 層における非線形変換関数，N は入力特徴量として使用する前のフレーム数，X_iと O_iはそれぞれ i フレーム番目の残響を畳み込んだ音声特徴量と変換した特徴量を表す。学習は制限付きボルツマンマシン（Restricted Boltzmann Machine : RBM）による事前学習を行った後に，特徴量と教師信号間のクロスエントロピー誤差を最小にするために誤差逆伝播法を用いてパラメータを更新する。事前学習後，誤差逆伝播法によって各パラメータの調整を行う。誤差逆伝播法とは，1 対の信号（入力信号と教師信号）が与えられたときに，出力値と教師信号の誤差を減らすようにネットワークの重みを調整するアルゴリズムである。本研究では，この誤差をクロスエントロピーで定義する。誤差を最小にするためのアルゴリズムは conjugate gradients を用いる。

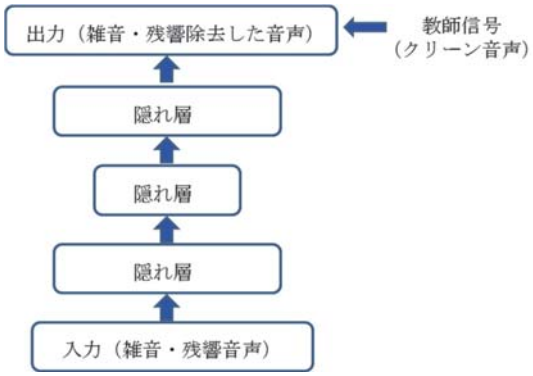

図 1　DAE の流れ図

本研究では，“REVERB-challenge”（残響音声の強調と認識ベンチマーク）が提供する学習データセットを使用する。このデータセットは，WSJCAM0 学習セットとマルチコンディション（MC）学習セットで構成されている。MC 学習セットは，WSJCAM0 学習セットのク

リーン音声に室内インパルス応答と背景雑音を重畳することで作成される。ここでの室内インパルス応答の残響時間は，0.1 秒から 0.8 秒の範囲のものを使用する。この学習セットは，DAE と音響モデルの学習の両方に使用する。評価データも学習データと同様に“REVERB-challenge”が提供するデータセットを用いる。このデータセットは，人工的に雑音と残響を重畳した音声（SimData）と実環境で収録した音声（RealData）から構成されている。SimData は WSJCAM0 コーパスの音声から生成され，RealData は MC-WSJ-AV コーパスの音声を用いている。ここでの室内インパルス応答や雑音は，学習データの環境とは異なるものを使用している。認識結果を表 1 に示す。DAE による雑音・残響除去を用いることで，単語の正解精度が改善していることがわかる。“SimData”においては 17.3% から 11.4% までの削減，“RealData”においては 54.0% から 37.0% までの削減を達成した「成果 1」。

表 1　雑音・残響音声の単語正解精度（%）

	Sim Data	Real Data	平均
CMN（従来法）	82.7	46.0	64.4
DAE（提案法）	88.6	63.0	75.8

また，ニューラルネットワークとマイクロフォンアレイの併用により，会議中の同時発話音声を目的話者のクリーン音声に変換した方法も提案した。認識率を従来法の 72.0% から提案法の 88.0% に改善した「成果 2」。

・DAE による特徴量変換手法を用いた発言者識別

一般的な発言者識別では，ある話者の発話区間セグメントの音声特徴を統計モデルで表し，類似する特徴のセグメントを統合する bottom-up 型クラスタリングの手法で発言者識別を実施する方法が一般的である。このような方法では，用いる特徴量や統計モデルなどの与え方が重要な要素となる。また，各発話区間の比較を行う音声特徴量にメル周波数ケプストラム係数（MFCC）が用いられることが多いが，単純な MFCC だけでは雑音や残響の影響を低減したクラスタリングを行うことは困難である。そこで，本研究では DNN を用いた特徴量変換手法を発言者識別システムに利用する。残響に頑健な特徴量変換手法として Denoising Auto Encoder（DAE）を提案し，会議や討論のような複数の話者が存在する長時間収録音声の発言者識別実験において評価を行う。これにより，雑音や残響により頑健なシステム構築が可能となることを示す。本研究の評価実験には NIST RT-05S データベースを利用した。このデータベースには収録環境が異なる 5 箇所，各 2 回の計 10 回の会議音声が収録されており，会議毎に遠隔発話音声と近接発話音声が複数チャンネルで収録されている。DAE について，コンテキストサイズは入力として与えるフレーム数を表し，現在フレームだけでなく，前後のフレームを含めることにより，残響の影響も考慮した特徴変換を学習することができる。表 2 は評価実験の結果を示す。この結果より，平均 DER（Diarization Error Rate）において，提案法（DAE-frontend-based System）は従来法を上回ることが示された。各 Set 毎に結果を比較してみても，Set1 において若干の悪化が見られるものの，他の Set においては改善が見られている。

表 2　従来法及び提案法における DER 比較（%）

手法	NIST RT-05 S データベース					平均
	Set 1	Set 2	Set 3	Set 4	Set 5	
従来法	19.9	20.3	26.1	20.4	21.8	21.7
提案法	20.4	11.2	19.7	15.3	15.7	16.4

・チャンネル変動に対して頑健な話者認識

登録発話とテスト発話間でチャネル変動（収録設備，収録環境の変動など）があると，従来の混合ガウスモデル（GMM）の話者認識性能が低下してしまう。会議音声では，会議毎にチャンネルや発話様式が変動するため，二つ発

話セグメント間の距離が大きく変動し，この二つのセグメントが同じ話者が発言するかどうかを判定する閾値の決定は難しくなる。最先端の話者認識では，音声の特徴空間を話者独立の特徴空間とその以外のチャンネル・話者変動による全変動空間に分解し，話者の情報を全変動空間から次元圧縮して数百次元の固定長ベクトルで表現する。しかし，全変動空間から次元圧縮して，一部の話者の情報がなくなる。そこで，全変動空間から次元を圧縮しないベクトル抽出の方法を提案した。さらに，大きい行列の高速な計算方法を提案し，確率線形判別分析を使って，どの環境のどの発話様式でも同じ閾値を使って頑健な話者認識を行った。標準な NIST SRE 08 データベースを用いて，評価実験を行った。話者照合の等誤り率を従来法の 6.6% から 4.4% に軽減した「成果 3」。

［今後の研究の方向，課題］

最近，音声認識の様々な技術が実用化され，大きく注目されている。しかし，残響・雑音による環境変動や発話スタイル，発話者によるアクセント（方言や日本人英語発音など）などによる発声変動などに対しては技術がまだ不十分である。大量の音声を用いて機械学習技術に基づいて近接発話で高い性能が得られているが，実環境による音声認識の性能がまだ非常に低い。

一方，グローバル化が進むに従い実環境で多様な発話様式による音声コミュニケーションの重要性がますます高まっている。例えば，2020 年の東京オリンピックで，実環境の会場で多様な発話スタイル（読み上げ音声や自由発話など）や多様なアクセント（native 話者の英語，中国人英語，日本人英語など）の音声コミュニケーションをとる機会が予想され，音声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。そこで，実環境の環境変動と発声変動を同時に効率よく除去する手法の確立が非常に重要であり，今後の研究の課題である。

［成果の発表，論文等］

1. 上田雄磨，王龍標，甲斐充彦，“Cepstral domain denoising autoencoder および DNN-HMM による雑音･残響下音声認識”，日本音響学会 2015 年春季研究発表会講演論文集，2-1-2，Mar. 2015.
2. W. Li, L. Wang*, Y. Zhou, J. Dines, H. Bourlard and Q. Liao, “Feature Mapping of Multiple Beamformed Sources for Robust Overlapping Speech Recognition Using a Microphone Array”, IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 21, No. 12, pp. 2244-2255, Dec. 2014.
3. Y. Jiang, K. Lee, L. Wang*. “PLDA for Speaker Verification in the I-supervector Spaces”, Eurasip Journal on Audio, Music and Speech Processing, 2014 : 29, 2014.
4. Z. Zhang, L. Wang* and A. Kai, “Distant-talking speaker identification by generalized spectral subtraction-based dereverberation and its efficient computation”, Eurasip Journal on Audio, Music and Speech Processing, 2014 : 15, 2014.