音声言語情報処理 27－9
（1999．7．23）

Workshop on Robust Methods for Speech Recognition in Adverse Conditions報告

赤木　正人（北陸先端科学技術大学院大学）
武田　一哉（名古屋大学）

概要：
本文は、1999年5月25日から26日にかけて、フィンランドのタンペレで開催されたWorkshop on Robust Methods for Speech Recognition in Adverse Conditionsの会議報告である。

Report on "Workshop on Robust Methods for Speech Recognition in Adverse Conditions"

Masato AKAGI (Japan Adv. Inst. Sci. & Tech.)
Kazuya TAKEDA (Nagoya Univ.)

ABSTRACT:
This paper reports workshop on Robust Methods for Speech Recognition in Adverse Conditions, held at Tampere, Finland, on May 25-26, 1999.

まえがき

5月25日から26日にかけて、フィンランドのタンペレで、表記のワークショップがNOKIA, COST249の協賛、IEEEの共催で開催された。

本ワークショップでは、音声認識技術を実用化するために必要不可欠な要素である、通信路に含まれる様々な歪みに対処し頑健性を向上するための方法について、様々な角度から議論された。議論された内容は、音声強調、頑健な特徴抽出、伝送路正規化、頑健な認識モデルおよび適応法、実環境音声データベースなどである。

プログラムは、Robust Feature Extraction、Adaptation and Compensation/Databases 、Robust Modeling and Feature Extraction 、Techniques for Real-World ASR の4つのセッションにより構成され、52件（うち9件は招待講演）の口頭発表、ポスター発表が行なわれた。発表形式は、各セッションとも共通で、まず2～3件の招待講演があり、その後でポスターセッション、最後に参加者全員によるディスカッションにより構成された。参加者はのべ約300名であったが、日本からの参加者は5名であった。以下、各セッションごとにその内容について報告する。

各セクションの内容

1. Robust Feature Extraction

第一セッションのキーワードは音声強調、すなわち得られた信号波形の中の有益な音声特徴をどのようにして強調するかであり、これについて様々な発表があった。

招待講演では、AT&TのGhitzaが、CVC音節を時間—周波数方向双方にタイル状に切り分け、それぞれのタイルを変形させて再合成したCVC音節の聴取実験から得られた知見をもとに、各CVC音節に有効な音声強調（特にCV, VC変化の強調）とこれを考慮

した距離尺度について講演した。また、STLのJunquaが、応用場面を考慮した音声認識システムの構築法について講演した。

ポスターでは計13件の発表があり、その傾向は大きく二つに分かれた。一方は音声特徴の時間方向の変化をどのように強調するかについての発表であり、他方はスペクトルサブトラクションを用いた雑音除去による音声強調についての発表であった。

前者はさらに、AM変調音に対する聴覚特性、マスキング特性などを考慮して音声特徴の変化を強調する方法、時間方向フィルタ（RASTA, MFCC変形など）を用いたランニングスペクトル変形、さらに、時間—周波数二次元フィルタを用いた変形などの種々の方法、に細かく分けられる。

後者は、雑音を推定する場合に、適応的に雑音スペクトルを推定する方法について主に議論している。

ディスカッションでは、人間の聴覚特性をどのように特徴抽出アルゴリズムに取り入れるかに話が集中した。主な意見は以下の通りである。

現在音声強調に用いられている聴覚特性はほとんどが心理学的知見であり、生理学的知見はほとんど取り入れられていない。このため、今後はいかに生理学・心理学の知見を取り入れながら頑健な特徴抽出を行うかが重要である。また、このようなbottom-up処理による音声強調のみではなく、top-down処理での候補の絞り込みも含めた総合的なシステムを構築する必要もある。

2. Adaptation and Compensation/Databases

第二セッションのキーワードは，モデルやパラメータの適応であり、新たなアルゴリズムの提案よりも、既存の適応アルゴリズムの実環境下での有効性に関する議論が中心であった。

招待講演は、C.H.Lee（Bell研）、嵯峨山（北陸先端大）により行われた。

C.H.Leeは、学習時と認識時のシステムのミスマッチを、分布パラメータの揺らぎという観点から定式化し，MAP決定則を基本とする2つのロバストな決定則、minimax classification rule 、 Bayesian Predictive Classificationを解説した。これらの決定則は、分布パラメータの揺らぎが、(1) 学習データ中のパターンクラスの出現頻度、(2) 先見的にあたえられた「分布パラメータの分布」として、それぞれ考慮されることから、本質的にロバストな決定が可能であることが、説明された。

嵯峨山の講演は、音響特徴空間の操作の観点から、適応アルゴリズムを説明した。講演では音響特徴空間の局所的な歪を特徴づけるヤコビ行列式により、加法的な雑音やチャネル歪の正規化をケプストラム領域で効率的に行えること、特徴パラメータの局所的な変形を補間平滑化したベクトル場を構成することにより、話者適応を高精度に行うことができることが解説された。

本セッションでの発表13件の内容はおおよそ、以下の表1にまとめられる。認識対象とする環境は、電話、車内、ノンネイティブ音声など多様であり、狭い範囲の適応技術だけでなく、音声区間の検出を含む様々な問題に関する検討結果が発表された。

ディスカッションでは、「真のリアルワールドの複雑さは発表された状況よりも、より深刻であり」「全く事前知識を利用しない適応を実現することが、必要である。」なる主張と「事前知識の問題は学習データ量の問題に帰着可能である。」との主張の対立や、「モデルの適応に比して音響空間の適応はより非線形性が高いのではいか。」といった主張がなされた。人間の聴覚システムの持つ適応機能に議論が及ぶに至って、議論は迷走の様相を呈し、聴覚機能に関する客観データの蓄積が未だ絶望的に不充分であることが、改めて浮き彫りにされた。

3. Robust Modeling and Feature Extraction

第三セッションのキーワードは特徴抽出および記述、すなわち外乱を受けた信号波形からどのように音声特徴を捉えるか、また、特徴を統合するためのルールをどのように構築するかを、音信号処理の観点から

議論するセッションであった。

招待講演では、スペクトルをマルチバンド（サブバンド）に分割した後、それぞれの周波数バンドで特徴抽出・強調を行う二つの方法が紹介された。

一つ目はBourlandによる講演であった。ある時刻において少なくとも一つのバンドは正しい認識結果を出力するという仮定の下で、各周波数バンドを一つのストリームと考え、ストリームごとの確率の計算法とその統合法を提案している。二つ目はHermanskyによる講演であり、各周波数バンドごとの振幅変化パターンの抽出法とその応用に関する発表であった。

ポスターでは計13件の発表があり、表2に示すような様々な信号処理法が提案された。

マイクロホンアレイでは、Delay-and-Sum法による雑音抑圧、Delay-and-Sum法とHMMを結合し音源方向を含めた3次元HMMによって音声認識を行う方法、適応フィルタ、スペクトルサブトラクションのための適応的雑音推定についての発表があった。また、マルチバンド処理については各サブバンドでの処理結果をどのように重みをつけて統合するかが最大の課題であり、これに関する発表がほとんどであった。さらに、ニューラルネットワークによる非線形な特徴抽出法、新しい発想で設計されたマイクロホンの耐雑音性についても発表があった。

ディスカッションでは、マルチバンド処理において各サブバンドでの結果をどのように統合するかについて、人間の処理機構との対比、音響特徴の解析と分類どちらに重点を置くかなどの点から議論が行われた。主な意見は以下の通りである。

マルチバンド処理は、聴覚末梢系における蝸牛までの処理のように周波数方向に分割して処理するだけか。有効に用いようとするならば、もっと上位のレベルとの関連を考える必要がある。また、各サブバンドでの結果の統合法について、音響信号の解析であれば、含まれる音信号に忠実に統合する必要があるが、音響信号の分類であるならば、タスクに依存した統合ルールが必要である。

4. Techniques for Real-World ASR

実環境での連続音声認識、モバイル端末での音声入力、自動車内での音声認識など、現在および今後問題として扱われる項目について、3名の講演者（Rahim, Haavisto, Hunt）による招待講演と4件の一般講演が行われた。演題に上った項目は

・自然に話された音声の認識
・モバイル環境での音声認識の現状と今後の問題
・自動車内での連続音声認識での課題

であり、本ワークショップを締めくくるにあたりまとめと今後の課題を明らかにする目的を担っていた。問題点を明らかにしたことは重要であるが、これを克服するための手法については、まだ手探りの状態である。

まとめ（感想に代えて）

1. 特徴抽出

特徴抽出における本ワークショップでの最大の論点は、スペクトルあるいはこれを変形した値の中の、どのような特徴をどのように強調するかであった。スペクトル系列のどの部分をどのように強調するか？これにより、人間の聴覚特性を模擬しようとしたもの、スペクトルの時間—周波数方向の変化を強調しようとしたもの、周波数をサブバンドに分けて独立に処理しようとしたもの、など、様々な方式が提案された。

しかし、これらはすべてまだまだ中途半端であるように思える。心理学的な知見だけではなく生理学的知見をも含んだ人間の聴覚特性の模擬による強調方式の充実、Top-Down処理によるどの部分をどのように強調するかの決定支援、Top-Down処理を有効に働かせるためのタスクの定義法など、総合的なシステムの構築が望まれる。

2. 雑音除去

雑音除去では、スペクトルサブトラクシ

ョン（SS）が多く使われていた。しかし、ここで問題となるのは、引き去るための雑音スペクトルをいかに精度よく推定するかである。バックグランドの雑音は一般には定常ではないので、時々刻々更新を繰り返さなければならない。この手法を確立することが、SSを実環境で使えるかどうかを左右する。本ワークショップにおいても、まだ完全な推定方法は提案されていない。

3. 適応・正規化

現在のところ、モデルのパラメータを適応化させることは、認識システムの頑健性を向上させる上で、最も有効なアプローチと考えられる。しかし、適応には、実に様々な実現方法があり、実際のアプリケーションにおける有効性を、比較実験なしに予測することは、困難である。個々の適応アルゴリズムの有効範囲や、性能の上限・下限を明確に整理することが望まれる。特に学習データ量と学習可能なパラメータ数の関係に対する理解を、深めていくことが望まれる。

表1　第二セッションにおけるポスター発表の内訳

対話／手法	PMC	MAP	LR	その他
電話（GMS）	1		2	2（端検出＋フレームエラー対策、混合重み係数の適応化）
車内		1		3（データベース設計、EMによる再推定、ニューラルネット）
話者適応（ノンネイティブ話者）		1		2（ツリークラスタリング、スムージング）
遠隔音声		1		
その他				1（SNRの推定）

表2　第三セッションにおけるポスター発表の内訳

手法	発表件数
マイクロホンアレイ	4
マルチバンド処理	3
ニューラルネット	3
missing feature theory	2
マイクロホン作成	1