LSSN 0915-5961

AVIRG

会　報

Vol.33 No.3 (1999.11)
発　　行：視聴覚情報研究会(AVIRG)
代表幹事：伊　藤　崇　之
〒157-8510 世田谷区砧 1-10-11
日本放送協会放送技術研究所
TEL 03-5494-2361
FAX 03-5494-2371

Ⅰ．10 月例会報告

「MDL 基準を用いた音声認識単位の自動生成」

講演： 篠田 浩一 氏（ＮＥＣ Ｃ＆Ｃメディア研究所）
報告： 世木 寛之（ＮＨＫ）

《概要と感想》

現在の音声認識は，観測された音声波形と単語との音響的な類似度である音響スコアと，単語と単語のつながりやすさを表す言語スコアを計算していき，この２種類のスコアの合計が最終的に最大となる単語列を出力する．

そして，この音響スコアを計算するとき，HMM(Hidden Markov Model：隠れマルコフモデル)を用いる手法が一般的である．

本講演では，音声認識を専門としていない人でも理解できるようにHMMの初歩から説明があり，篠田氏ご自身が行ったクラスタリングに関する研究までの報告が行われた．

大語彙連続音声認識では，限られた学習データしか存在しないことと，HMMが大きくなってしまう（計算機のメモリに載りきらない）ことから，単語でHMMを作らずに，音素と呼ばれる単位でHMMを作ることが多い．例えば，入力である音声波形と「雨」という単語を比較するときには，入力された音声波形と/a/，/m/，/e/という各音素を比較することになる．

このとき，前後の音素環境を考慮したトライフォンを使用すると，より正確な音響スコアを求めることができる．例えば，同じ音素/m/でも，前述の「雨」の音素/m/と，「網(ami)」という単語の音素/m/では，後ろの音素が異なるので音響的な特徴が異なり，別々にモデルを作ったほうが精度が良くなるのである．

トライフォンを扱うときに問題になるのが，学習データが少なくなり統計的信頼性が低下することと，学習データに出現しなかったトライフォンのHMMが作れないということである．

そこで，トライフォンをいくつか集めてクラスターを作り，クラスター内のトライフォンを同じトライフォンとして扱うことにより，見かけ上の学習データを増やす手法がとられてきた．その際，パラメータを制御することによりクラスター数を制御するのだが，クラスター数を小

さくしてしまうとデータの音響的特徴を十分に表現できず精度が悪くなり，逆にクラスター数を大きくしてしまうと学習データ数が少なくなってやはり精度が悪くなる．つまり，適当なクラスター数を作るパラメータを求めるためには，いくつかのパラメータで実際にHMMを作成し，認識実験を行って認識率を求めなければならなかった．また，そのパラメータは学習データの量や質が変われば使えなくなってしまうので，新たに求めなおす必要があった．

篠田氏は，このクラスタリングする際のパラメータを情報量基準の一つであるMDL基準(Minimum Description Length Criterion：記述長最小基準)から決めることを提案している．

MDL基準は，与えられたデータに対し最適なモデルを選択する問題において有効であることが知られている．そこで，学習データに対するHMMの記述長を考え，この記述長を最小とするHMMを求めることにした．これを考慮すると，パラメータの推定式として，学習データの大きさに依存する式が得られる．このことを使うと，学習データに対し最適なモデルサイズを持つモデルを作成できるパラメータを得ることができるようになる．

実際に，従来法でいくつかのパラメータでHMMを作り認識実験を行った結果と，今回のMDL基準により推定したパラメータでHMMを作り認識実験を行った結果を比較すると，本手法の方が孤立発声単語認識実験で２％程度良い認識率が得られている．

ただ，MDL基準の最適性を調べるために，MDL基準により推定されたパラメータを数倍にしたいくつかのパラメータでHMMを作り認識実験を行ったところ，MDL基準により求めたパラメータの２倍の値が最高の認識率を示している．このことから，MDL基準によりパラメータを推定することは，最適ではないが，従来法に比べ非常に効率よくパラメータを選ぶことができることが分かる．また，あらかじめMDL基準によりパラメータを推定しておき，その値付近のパラメータをいくつか試せば，従来法でパラメータをいくつか試すよりも少ない数ですむと予想される．

ここ数年で音声認識は大変な進歩をとげ，ディクテーションソフトやゲームなど実用的な製品にとりいれられ，また一方でニュース番組でのアナウンサーの音声をリアルタイムで字幕化するのに使用する試みも行われている．音声認識がいろいろな分野に普及していくためには，現在の音声認識システムに存在する様々なパラメータを，今回の研究発表のように合理的な観点から効率良く推定していくことが必要不可欠のように思われる．

「音声認識のための精密かつ頑健な音響モデル」

講演： 中村 篤 氏（ＡＴＲ音声翻訳通信研究所）

報告： 黒岩 眞吾（ＫＤＤ研究所）

《概要と感想》

限られた学習データを用い精密かつ頑健なモデルを構成することは，人間の学習を含め，学習の研究における永遠のテーマの一つである．音声認識においても，最近では数千から数万人の発声データを用いて音響モデルの学習が行われるようになってきたが，自然の産物である人間の音声を認識するには，それでもなお「限ら

れた」学習データに過ぎず，精密かつ頑健な音響モデルを構築するために，今日もなお多くの研究努力が続けられている．

中村氏の所属するATR音声翻訳通信研究所は，その前身であるATR自動翻訳電話研究所時代を含め日本における統計的手法を用いた音声認識研究の老舗の一つである．講演では現在ATRの音声翻訳システムの主力の音響モデルである隠れマルコフ網(HM-Net)について，精密性と頑健性のバランスをとりつつその構造を決定する手法について，改良の歴史を含め紹介された．

HM-Netはその構造の学習法であるSSS(Successive State Splitting)と共に，1991年，ATRの鷹見氏と嵯峨山氏によって提案された．SSSは従来，知識や経験等によって決定していたHMMの構造を，1状態のHMMを出発点としデータ駆動で最尤基準により逐次状態分割を進めていくことで，精密かつ頑健な音響モデルを得ようというものである[1]．

音声認識においては，一般に音素(例えば「か」は/k/という音素と/a/という音素から構成される)毎に3状態程度のHMMを作成し，それらを連結することで任意の単語や文に対するテンプレートを作成し入力音声とのマッチングが図られる．このようなモデルをさらに詳細化するためには，例えば前後の音素によってモデルを分ける(akaの/k/とakiの/k/では，後ろの/a/や/i/に引っ張られて舌の位置や口の形が異なり，音響的にも特徴量が違ってくるため)ことが考えられる(音素環境による分割と呼ぶ)．しかし，あまり分割しすぎると状態あたりの学習データが少なくなってしまうため頑健性が低下するという問題が生じる．

そこで，SSSでは，(1)すべての状態の中から分布の広がり(正確にはdivergence)が最大の状態を選択する，(2)その状態を，時間方向への２分割と音素環境による２分割の組み合わせのすべての方法で分割を試み最も尤度が上昇する２分割を行う，という２つの操作を繰り返すことで，１状態のHMMから出発し，最終的には数百から千程度の状態で構成されるHM-Netを構成する．本来ならば，(2)の分割の試みをすべての状態について行うべきであるが，計算量的に不可能であり，(1)の近似を用いたことが本手法を現実的なものとしている．このようにモデルを構成した結果，特定話者に関しては精密かつ頑健なモデルが実現された．

しかし，この手法を不特定話者のデータに適用した場合，話者による分布の広がりでが原因で状態が選択され，音素環境で状態を分割しても尤度がほとんど上昇しないという問題が生じた．これに対し1996年，ATRのSinger氏とOstendorf氏によって提案されたML-SSS(Maximum Likelihood SSS)[2]では，上記(1)の方法を排除する代わりに，音素環境による２分割をChouの最適分割アルゴリズム[3](K-meansクラスタリングやVQコードブック作成時に空間を２分割するのと類似の手法)により行うことで，尤度最大(正確には局大)条件で状態の選択および分割を同時に行うことが可能となった．その結果，不特定話者に関しても精密かつ頑健なモデルの実現に至った．しかしながら，その後，このモデルを自然発話の大語彙音声認識に適用した際，尤度の局所的な落ち込みが生じ探索の過程で正解候補が枝刈りされてしまう場合があることを，本講演者であるATRの中村氏が明らかにした[4]．

大語彙音声認識では，演算量および記憶容量の問題で探索処理の過程でどうしても枝刈りを行わなくてはならず，部分的にでも尤度が下がってしまうことは誤認識の増大につながる．この問題に対し中村氏は，局所的に尤度を落ち込ませてしまうデータに対し尤度が上昇するように状態間で分布を共有する（分布を再構成する）ことを提案した(HM-Netに限らず音声認識で用いるHMMの各状態は混合正規分布で表現するこ

とが一般的である．共有は混合分布の１要素を他の状態から借りてくることに相当し状態を共有するわけではない)．これは，ML-SSSでは２分割で状態を分割していくため，２つの分布の重なりでそれなりに尤度を保っていたデータが分割によって谷間になってしまうことがあるからと氏は説明された．(谷間以外の例として，例えば同じ「が(/g//a/)」の音もポーズの後では濁音，他の音に続く場合は鼻濁音とアナウンサー等，使い分けている人も多い．この音声でML-SSSを行いモデルを作成した場合，前の音素がポーズの/g/（濁音）と他の音素に続く/g/（鼻濁音）で状態が分割される可能性が高い．このモデルに対し，単語中の/g/も濁音で発声する人の音声では，分割前は尤度が高かったにも関わらず分割により尤度が下がるという現象が生じるものと予想される．)

以上の結果，ML-SSS＋分布の再構成法によるモデルを用いた現在のATR大語彙連続音声認識システム(ATRSPREC)では，不特定の人の自然な発話に対しても高い認識性能が達成されている（95年当時40%を超える誤り率があったものが，現在では27,000語彙の自然発話で単語誤り率13.5%，また，文献[5]によれば朗読発声では誤り率7.2%と報告されている)．

なお，文献[5]によれば現在用いているHM-Netの状態数は1,000から1,400となっている．この状態数は開発用テストセットにより経験的に決定しているとのことであるが，今後は，例えば前半の篠田氏の講演にあったMDL基準等により学習データ数やタスクが変わっても精密性と頑健性のバランスのとれた状態数を自動的に決定できる手法の導入が望まれる．

以上のようにATRでは国内外の優秀な研究者が集い，一つの手法に対しても様々な視点から研究開発が行われ目覚ましい成果を達成している．2001年にHAL9000程度の音声認識が実現するか否かは微妙なところであるが，海外のホテル予約が音声翻訳システムにより日本語だけでで行えることは，既にSFの中のお話ではないようである．

《参考文献》

[1] 鷹見淳一,嵯峨山茂樹,“音素コンテキストと時間に関する逐次状態分割による隠れマルコフ網の自動生成”，信学技法SP91-88，pp.57-64，1991.

[2] H.Singer，M.Ostendorf，〝Maximum Likelihood Successive State Splitting"，Proc．of ICASSP 96，pp．601-604，1996.

[3] Chou，"Optimal partitioning for classification and regression trees"，IEEE Trans．PAMI，13(4)，pp．340-354，1991.

[4] 中村篤，"ガウス混合分布の再構成による不特定話者音響モデルの改善",信学技法SP97-18，pp.9-16，1997.

[5] 内藤正樹，他，"旅行会話タスクにおけるATRSPRECの性能評価"，音響学会秋季講演論文集，pp.113-114，1999.

Ⅱ． 11月例会予定

AVIRG 11月例会は,

日時： 11月25日（木）14時～17時

場所： 東京大学工学部6号館

3Fセミナー室A・B

で開催いたします．テーマは,『マルチメディア処理』です．講演者およびタイトルは以下の2件を予定しております．奮ってご参加ください．

「情報空間の知覚化」

講演者： 広池 敦 氏

(日立製作所中央研究所)

我々が開発中の類似画像検索システムについて紹介する．我々のシステムの特色は，ユーザインタフェイス上での検索結果の表現にあり，基本コンセプトは「なるべく多くの画像をユーザに見せる」ことである．実際の表現上では，2000件規模の検索結果が，３次元空間中を群れを形成しながら運動する．また，視覚表現のみではなく，ユーザインタフェイス上の音響効果についても論じたい．

《参考文献》

[1] 広池，武者，杉本，“VR空間を用いた画像特徴量空間の可視化 ― 画像データベースの検索・ブラウジングのためのユーザインタフェイス”，信学技報，PRMU98-86，pp.17-24，1998.

[2] 武者，広池，“類似画像検索における検索結果の可視化インターフェース - 可視化軸として意味軸を用いる方法 -”，信学技報，PRMU99-57，pp.59-64，1999.

[3] Hiroike A., Musha, Y., Sugimoto, A. and Mori Y., "Visualization of information spaces to retrieve and browse image data", Third International Conference on Visual Information Systems, Springer-Verlag, pp.155-162, 1999.

「撮像面上に処理機能を統合したイメージセンサ」

講演者： 浜本 隆之 氏

(東京理科大学工学部電気工学科)

イメージセンサと処理回路を１つのチップに統合し，センサ上で直接画像処理を行う，高機能イメージセンサについて発表する．このような処理機能一体型イメージセンサは，スマートセンサ，コンピュテーショナルセンサ，ビジョンチップ等と呼ばれ，近年研究が盛んになりつつある．高機能イメージセンサの大きな利点は，蓄積時間等の撮像パラメータを直接制御できるということと，画像情報の2次元性を直接利用することで，高速並列処理を行なえる点にある．画像システムの一部あるいは全部の処理を撮像面上で行うことで，後段で必要とされる情報のみを出力し，データ転送量を大幅に削減できる．このような技術は，画像処理システムの小型化，高速化，低消費電力化に貢献するものと期待されている．

本発表では，高機能イメージセンサの研究動向について簡単に述べるとともに，我々のグループが検討している動画像圧縮センサ，適応蓄積時間イメージセンサ等について説明する．

《参考文献》

[1] http://www.eleceng.adelaide.edu.au/Groups/GAAS/Bugeye/visionchips/index.html, “Vision Chips or Seeing Silicon”, by A.

Moini, third revision, 1997.

[2] 相澤, 大野, 江木, 浜本, 羽鳥, 丸山, 山崎, 大竹, 小林, 大久保, 阿部, “動画像圧縮イメージセンサ”,テレビ誌, Vol.50, No.2, pp.257-265 (1996)

[3] 浜本, 大塚, 相澤, 羽鳥, “動画像圧縮センサ--列並列処理構成による設計と試作–”, 映情学誌, Vol.51, No.12, pp.2141-2148 (1997)

[4] 浜本, 相澤, 羽鳥,“動き適応イメージセンサの試作と機能評価”,映情学誌, Vol.51, No.12, pp.2149-2157 (1997)

[5] 大塚, 浜本, 相澤, 羽鳥,“空間可変サンプリングを撮像面上で行う新しいイメージセンサの設計・試作”,映情学誌, Vol.53, No.2, pp.261-268 (1999)

～会員登録情報の変更のお願い～

AVIRG会員の御所属, 会報送付先など登録情報に変更がありましたら, お手数ですが以下のいずれかにご連絡ください.

◎ (財)日本学会事務センター 会員業務係

◎ 電子メール（1999年度中） avirg-member@vision.STRL.nhk.or.jp (AVIRG幹事宛)

(注) 会員の確認のために, 御氏名とともに, 必ず会員番号を明記して下さい.
会員番号および学会事務センターの連絡先は会報郵送時の封筒に印刷されています.