# IEEE音響・音声・信号処理国際会議(ICASSP'99)報告

*新田 恒雄[1] 広瀬 啓吉[2] 嵯峨山 茂樹[3] 中川 聖一[1] 小林 哲則[4]*
*1豊橋技術科学大学 2東京大学 3北陸先端大学院大学 4早稲田大学*

1999年3月15－19日の期間，IEEE ICASSP'99が米国Phoenixにおいて開催された。音声関係は26のセッションで213の報告があり，全体の1/4を占めた。本報告は音声関連セッションの概要をまとめたものである。言語モデルが有効に働くディクテーションでは，現在，ニュース番組の自動筆記が主要な目標になっている。一方，車載応用，電話・ネットワーク応用では，騒音下あるいは異なる環境下におけるロバスト（頑健な）音声認識を目指して，新しい音響モデルや言語モデルの模索が続いている。また，未知語に対するリジェクト能力改善，キーワードスポッティング性能向上など，本格的音声対話を目指す様々な提案が行われている。

## A report on 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing （ICASSP '99）

*Tsuneo NITTA[1], Keikichi HIROSE[2], Shigeki SAGAYAMA[3]*
*, Sei'ichi NAKAGAWA[1], and Tetsunori KOBAYASHI[4]*
*1 Toyohashi Univ. of Tech. 2 Tokyo Univ. 3 JAIST 4 Waseda Univ.*

This paper summarizes the speech-related topics in IEEE ICASSP'99 held at Phoenix, U.S.A. In the conference, 213 papers were presented at 26 sessions. In the application of dictation in which a language model is most effective, their efforts are focusing on broadcast news transcription. On the other hand, various types of acoustic models and language models are proposed to achieve robust speech recognition under noisy environments and/or spontaneous speech recognition.

## 1. 項目と分担

本報告の主要項目と分担を以下に示す。

2．音声合成 (SP- 12 広瀬)
3．音声認識
3.1 特徴抽出 (SP- 13 嵯峨山)
3.2 音響モデル (SP- 4/ 11/ 18 嵯峨山)
3.3 言語モデル (SP- 17/ 23 中川)
3.4 大語彙連続音声認識 (SP- 2 中川)
3.5 音声理解システム (SP- 16/ 20 小林)
3.6 発話照合 (SP- 22 新田)
3.7 ロバスト音声認識 (SP- 9/ 14 新田)
3.8 音声認識応用 (SP- 5 新田)
3.9 話者認識と言語識別(SP- 10/ 26中川)

## 2. 音声合成

ICSLPの音声部門の発表は、例年、認識関係、コーディング関係が大半を占め、合成関係の比重が小さいが、本年は特にその傾向が顕著であった。合成のセッションは SP12（Speech Production and Synthesis）のポスターセッションのみであり、その8件の論文の内、音声合成を対象としたものは5件であった。以下、この5件を紹介するが、他のセッションにも合成関係の論文があり、特に、セッション SP7 の徳田らの韻律的特徴を包含するHMMの発表は注目される。これは、基本周波数を有声区間に対応する1次元空間と無声区間に対応する0次元空間の2つの空間から出力される観測事象として捉えた上でHMMで表現するものであり、HMM ベースの音声合成において、基本周波数パターンの統一的な枠組みでの処理を可能とする。音声認識への利用も有望と考えられる。

Donovanは、HMMによる波形切り出しとパラメータ重み学習に基づく波形選択合成で、従来から大きな成果をあげているが、ここでは、フレーズ単位で波形を切り出して接続することを併用する合成システムを、共著者とともに開発している。TD-PSOLAに基づく波形編集合成などの技術の開発により、高い品質の音声合成が可能となったが、韻律的特徴も含め、自然音声とは、なお、大きな隔たりがある。実際の音声応用システムでは、合成対象が限られることも多く、この様な場合、同様のフレーズが頻出することが予想される。ここではこの様なフレーズをそのまま合成波形に接続することによって、人間の発音に近い合成音声を得ている。ある意味で、匂坂らによる複合合成単位手法に包含されるものである。

Stylianouは、波形選択合成において、合成用音声素片の音質の違いが合成音声の品質に影響することを考慮し、音質の違いを検出して修復する手法を提案している。波形選択合成で品質の高い音声を得るためには、多量の音声データが必要となるが、この収録に時間がかかるため、話者の発声態度や録音の条件が変化し、均一の音質の音声を得ることが困難である。この様な問題は従来から指摘されていたが、修復まで含めた対処手法を初めて提案している。まず、各録音セクションの始めの部分の音声を用いて音質に対するGaussian Mixture Model（GMM）を学習し、次に、セクションの各部分の音質の近さの度合いを、モデルに対する尤度として求める。尤度が閾値以下となった部分が音質の点から問題ありとし、平均パワースペクトル密度の違いに基づいて計算したcorrective filter（ARフィルタ）を用いて修復する。このフィルタの特性は、当然、音素の種類によって異なると考えられるが、とりあえず同一として実験している。音質の劣化なしに修復できたとして手法の有効性を主張しているが、実際に修復したのは、主として録音環境の違いに起因するものであり、今後の研究が期待される。

波形編集型の音声合成において、素片の特徴を変更する有効な手段として正弦波モデリングによる分析再合成が期待さている。分析再合成でのパラメータの変更によって各成分の位相関係が影響を受け、これに対処する手法が必要であるが、従来はpitch pulse onset timeを設定し、これに各成分を同期させることをしていたため、onset timeの抽出誤りの問題があった。これに対しO'Brienらは、各成分の大きさ、周波数、位相のスムースなつながりを、まず基本波成分について求め、高調波成分に拡張する手法を開発した。広い範囲の時間軸伸縮（すなわち基本周波数の変更）に対し、波形の変形が小さい高品質の音声が得られるとしているが、音声全体にわたって一様伸縮しただけであり、実際にテキスト音声合成等に適用した結果が待たれる。

連続音声中の音素長の定式化に有効な手法として、音素長に関与する種々の文脈の寄与を多重線形回帰式によって表現することが行われているが、その際、音素長を対数変換することで回帰式の表現能力が向上することが報告されている。Silvermanらは短い音素長の場合にも回帰式により予測困難な部分があることに着目し、シグモイド関数による変換を提案した。実験の結果、対数変換と比較して約半数の因子で同様の表現力が得られ、変換の有効性が示されたとしている。しかしながら、長い音素長と短い音素長を圧縮された範囲に変換すれば、回帰式による表現が容易になるのは当然とも言え、人間の知覚特性との関連を考慮した評価が不可欠である。

Childersらは有声音源を動的モデリングにより表現することを試みている。動的モデリングは、観測事象を生成する多次元動的システムを表現するものであり、そのモデル構築に音源波形を入力とするtime-delay neural networkを利用している。従来の線形予測分析に基づく逆フィルタ法によって音声波形から音源波形を求めており、その精度が気にかかる。有声音声のみからなる文の合成で、品質の高い音声が得られたとしているが、他の有声音源モデルとの比較がない。無声音源に対しても、同様の枠組みでの取り扱いが可能であり、今後、一般の文章を合成した結果が待たれる。

# 3 音声認識

## 3.1 特徴抽出 (SP13)

- SP 13.2: "LSP Weighting Functions Based on Spectral Sensitivity and Mel-Frequency Warping for Speech Recogntion in Digital Communication" by S.-H. Choi (KAIST), H.-K. Kim (AT&T Labs), H.-S. Lee (SK Telecom)
  ディジタル通信に使われる LSP パラメータをそのまま特徴量として、そのスペクトル感度を重みに用いて距離尺度を構成して音声認識 (VQ-HMM) に用いる。比較に mel 周波数伸縮を考慮。QCELP を入力に認識系を構成。ところで、最近、ヨーロッパで音声認識に適した音声符号化伝送の標準化を進めようという動きがある。

- SP 13.5: "Towards a Robust/Fast Continuous Speech Recognition System Using a Voiced-Unvoiced Decision" by D. O'Shaughnessy, H. Tolba (INRS-Telecom).
  有声／無声判別を連続音声認識に用いる。フロントエンドで有声／無声判別を行い、それに応じて Comb Filtering および Spectral Subtraction を用いて雑音を除去する。さらに、有声／無声判別によって音響モデルを切替える。Viterbi 探索空間が削減でき、高速化できた。

- SP 13.8: "Distinctive Feature Detection Using Support Vector Machines" by P. Niyogi, C. Burges, P. Ramesh (Bell Labs).
  学習理論で注目されている Support Vector Machine (SVM) の音声認識への応用。Structural Risk Minimization Principle に基づく SVM の理論についてかなり述べている。具体的な応用として、破裂音の特徴検出を扱っている。実験結果によると HMM より高い検出率・正解率が得られた。今後研究が必要な分野であろう。

## 3.2 音響モデル (SP-4, 11, 18)

- SP 4.4: "Automatic Clustering and Generation of Contextual Questions for Tied States in Hidden Markov Models" by R. Singh, B. Raj, R. M. Stern (Carnegie Mellon Univ).
  Tree-based のトップダウンの状態 tying(つまり異音クラスタリング) において使われる "linguistic questions" を音響的に自動的に生成して、認識性能が向上した。まずボトムアップにクラスタリングする。すべての ($2^{I-1}$ 通り) 組合せを試みて、尤度損失が最小になるように合体させる。こうしてルールを作成する。従来の手で作られたルール群より性能がやや高い。Kai-Fu Lee (1990) が (PEC 法の二分木の手法の影響を受けて) 音声学の知識を採り入れた二分木による異音クラスタリングを始めたが、結局 10 年かかって PEC や SSS 法のように音響的な基準を用いる方向に戻って来ているといえそう。

- SP 4.5: "Partly Hidden Markov Model and its Application to Speech Recognition T. Kobayashi, J. Furuyama, K. Masumitsu (Waseda Univ)"
  早稲田大学の小林先生の新しいアイディア。従来の HMM における一重マルコフモデルに代わって、二次のマルコフモデルを導入し、第一状態は隠れ、第二状態は観測できるとした。単語誤り削減率は 39%。

- SP 18.3: "Refining Tree-Based Clustering by Means of Formal Concept Analysis, Balanced Decision Trees and Automatically Generated Model-Sets" by D. Willett, C. Neukirchen, J. Rottland, G. Rigoll (Univ Duisburg).
  これも状態クラスタリングのための Dicision Tree の構成法である。10 通りの決定木の作成法を実験で比較している。性能差はあまり大きくない。新しい数学の理論 Formal Concept Analysis を用いる。学習データ量によってセット数を変えるような方法がよいとしている。

- SP 18.4: "Efficient Speech Recognition Using Subvector Quantization and Discrete-mixture HMMs" by S. Tsakalidis (Tech Univ of Crete), V. Digalakis (Tech Univ of Crete / SRI ), L. G. Neumeyer (SRI)
  離散混合 HMM の巧みな構成法。部分ベクトル量子化と組み合わせた。速度を上げるために、尤度計算における MFCC の次元の加算順序を調整し、途中で打ち切る。SRI の DECIPHER に組み込んだ。ATIS で評価。連続混合 HMM の精度を保ちながらかなり高速化。混合数が同じならば認識率はより高い。

- SP 18.5: "A Unified Approach of Incorporating General Features in Decision Tree Based Acoustic Modeling" by W. Reichl, W. Chou (Bell Labs).
  Dicision tree clustering に音素情報以外の要素も含めたもの。(このアイディア自体は PEC 以来論じられているのでは?)。実際には、単語中の音素の位置を含めている。単語誤り削減率 10%。

- SP 18.8: "On the use of Support Vector Machines for Phonetic Classification" by P. R. Clarkson, Pedro J. Moreno (Compaq).
  Support Vector Machine を用いた音素認識。まだ初歩的 (学生の夏実習)。母音パターンを固定次元にして一対多 (他) の識別を行った。64 混合正規分布より若干識別性能が高い。

### 3.3 言語モデル

Warnkeら（Erlangen-Nurenburg大）は語彙サイズ4,500の Verbmobil データに対して，１８発話行為を分類するための n-gram言語モデルの線形補間と合理的( rational ) 補間（出現頻度の大きいほど重みを大きくする）に対して，補間係数を ML, MMI, MCEで求めて比較した。両補間共に MMIがよいが，合理的補間で比較すると差は少ない。

Wesselら（Univ. of Tech. Aachen）は，列車時刻案内タスクの対話の状態を２７個に分類し，全体の言語モデルと状態別の言語モデルの内挿によって，パープレキシティを，12.1から9.5に減少させている。

Samuelssonら（Bell Lab.）は，品詞の複合（例えば，{noun,aux verb}）を新しいクラスとするクラスn-gramの言語モデルを発表した。28品詞から305クラスを生成し，20,000語彙のWSJタスクに適用した。これで単語 trigram を補間することによって，パープレキシティを187から180に減少させた（主に，出現頻度の少なく unigram を使う単語に有効)。

Mahajanら（Microsoft）は，トピックに依存して動的に言語モデルを構成し，20,000語彙のWSJに対して，パープレキシティを168からキャッシュ使用により143，トピック依存モデルとの併用により113まで減少させている。適応化を類似な50記事，100記事，400記事を用いて，それぞれ言語モデルを構築し，混合して用いると，パープレキシティは109に減少した。

Khudanpurら（Johns Hopkins Univ.）も同様なトピック依存モデルの適応化により，スイッチコーパスに対し，約10%のパープレキシティの減少を得ている。

Martinら（Univ. of Tech. Aachen）は，通常のtrigramと距離２のtrigram（p(w|t・uv), p(w|tu・v) など)を用いて，スイッチコーパスに対して，約10%のパープレキシティの減少を得ている。

Bellegarda（Apple）は，ICSLP-98でも発表していたが，LSAモデルを用いて，ドキュメントの統計的性質によるn-gramの適応化を行い，WSJに対して，パープレキシティで24.7%，WERで17%の減少を得ている。

Kalaiら（CMU）は，混合言語モデルで混合重みをオンラインで適応化する方法を提案し，重みを最適に固定した場合よりも良い結果を得ている。

Gorinら（ATT Lab.）は，電話応対対話システムの言語モデルに人―人の対話データにより作成した可変長n-gram確率オートマトン(VNSA)（６状態：挨拶，確認など）を人―機械の対話データで適応化する方法を発表した。

Aielloらは音声データベースを収集する場合に，テキストでシナリオを提示する場合と画像で提示する場合を比較し，収集されたデータは，前者の方がパープレキシティが小さい，未知語は少ない，キーワードが多いと発表している。

Harper（Purdue Univ.）は，制約依存文法を用いて，RM文を意味理解する実験結果を発表した。 bigram 言語モデルよりも，スロットの抽出精度を向上させている。

### 3. 4 大語彙連続音声認識システム

ニュース音声データやスイッチコーパスなどのディクテーションに関する発表が，Dragon, IBM, BBN, LIMISI,ケンブリッジ大学からあった。

Dragonでは，特徴パラメータの線形変換によるクラス内分散行列の対角化後，クラス間の分散行列も対角化するIMELDA を適用したが，分散行列を結びにした方法よりも精度が悪い。その他，性別モデルと周波数軸の非線形伸縮の効果も比較した。

IBMは，次の改善を発表した。(1) 発音辞書の改善（効果小），(2) ガウス分布の選択基準にBICの導入（効果有），(3) 各単語辞書の語尾の短いポーズの追加（効果小），(4) ガウス分布で表現できない分布のピークネスをモデル化するためにガウス分布の変形である HAM ( homogeneous alpha mixtures )分布の提案（特別の場合として，フイリップスで用いられているらプラス分布を含み，効果有），(5) 判別分析による全分散行列の対角化 ( FACILIT ) の使用（効果小），(6) 発音ネットワークの導入（効果小）。最後に，それぞれの改善法を並行に走らせ，投票形式で最終決定する ROVER 法を適用し，改善効果を得ている ( WER : 21.5% → 20.2% )。

BBNは以下の改善を行った。声道長の正規化，話者ごとのケプストラムバイアス除去，１音素当たり２フレーム以上の対応付け（以前は３フレーム以上 ），quinphon の採用， SAT と MLLR による話者適応化，類似データベースを用いた trigram の追加学習， POSによる trigram のスムージング等である。３種類のフレームレートによる quinphon 使用のシステムと通常のシステムとの ROVER 法により，効果を得ている ( WER : 47.9% → 45.9% )。

ケンブリッジ大学は，電話による会話音声ディクテーションシステムを発表した。用いた技術は，非線形伸縮を考慮した声道長の正規化（効果有）や MLLR による話者適応，quinphone の利用，単語 trigram とクラス ４-gram（３５０クラス）の併用などで，ROVER 法により，大きな改善を得ている ( WER : 49.3%→ 39.3% )。

MITは天気予報を対象とする電話案内対話システム JUPITER を発表した。実際のユーザから1,100呼, 60,000発話の書き起こしデータを作成している。語彙は1,893語，音響モデルは diphone，セグメント統計量を用いる SUMMIT を使用，言語モデルは２００クラスによる trigram で，性能はドメイン内の発話で WER 13%程度である（エキスパートの発話では2.3% ）。

この他では，LIMSI がニュース音声認識の評価 ( 語彙サイズ vs 未知語率，低頻出語の認識率など )，ケンブリッジ大学が音声ドキュメント検索の評価の発表を行った。

### 3.5 音声理解システム

音声理解システムに関しては，口頭発表８件，ポスタ８件，合計１６件の発表が行われた。内容的には，対話制御戦略に関するものが２件，検索応用１件，スピーチアクト／意図などの推定３件，翻訳・話題抽出における統計モデルの応用２件，語彙単位に関するもの１件，認識システム関連４件，評価関連２件，テキストから音の合成１件と多種・多用な話題が提供された。以下口頭発表を中心に講演の概要を紹介する。

Bouwman らは，彼らが提案する音声認識の確信度を，ARISE 対話タスクにおけるユーザ発話の確認戦略に利用する方法を提案した。対話において，暗示的な確認発話（発話の理解は正しいと仮定して話題を展開したうえで，発話にシステムが理解している内容を添える方法）は，発話総数を減らす上から有効と考えがちであるが，理解内容が実際とかけ離れている場合には，ユーザはシステムの発話を理解できずかえって戸惑うことがある。筆者らは，理解内容の確信度に応じて明示的な確認と暗示的な確認とを切り替える対話戦略を導入することでスムーズな対話を実現することに成功したとしている。

Ries は，CallHome タスクにおけるスピーチアクトの推定に，HMM とニューラルネットワークを用いることを試みた。スピーチアクトを隠れ状態におき，対話の流れを状態の遷移に，発話の内容を各状態での出力において，単語列からスピーチアクトの列を推定する枠組みにおいて，ニューラルネットを出力確率の表現に用いることの有効性を主張している。

Lamel らは，ARISE タスクにおける対話制を，対話の原則（ユーザを迷わせないこと，ユーザの質問には直接的に答えること，いつでもユーザに修正の機会を与えること，誤りを減らすこと）に基づいて行うことを提案した。この中で，特にシステムの理解の状態に関する正確なフィードバックをユーザに対し行うことの重要性を説いている。２段階の対話モードを持って，通常混合主導のモードで，暗示的な確認を行いながら対話を進め，理解の状態とユーザの発話とが矛盾を生じるなどの問題を生じた段階で，システム主導のモードに移り，明示的な確認を行いながら対話を進める方式を採用している。

Siegler らは，音声文書(Spoken Document)に対する情報検索において，音声認識結果の信頼性を加味することを提案している。Lattice Occupation Density と呼ぶ単語ラティスの複雑さを表現する尺度を導入し，これを用いて尤度を補正することで，単純に第一位の音声認識結果を用いて情報検索する場合に比べ，検索性能を向上できるとしている。

Golden らは，電話での情報サービスシステムにおけるメニュー選択をタスクとして，発話に現れる単語からその意図（その発話に対してとるべき行動。メニュー階層の移動や実際のサービスの提供など。）を推定する手法を，単語列と意図の多項関係モデルに基づいて実現した。

Gotoh らは，大語彙連続音声認識の言語モデルにおける，名前を表す名詞の扱いについて検討した。名前の属性を表すタグレベルの統計と，共通タグを有する単語グループ内における単語レベルの統計の２階層構造によって言語モデルを構成することで，名前の検出率を上げ，OOV の影響を減じることに成功している。

Ney は，音声認識から翻訳までをベイズ決定則に基づいて統合的に扱う翻訳システムを提案した。局所的な平均近似操作や monotone alignment などを導入することで，テキスト入力ベースのベイズ的翻訳手法を音声入力に拡張している。

Walls らは，ニュース放送における，話題検出・追跡のための確率モデルについて検討した。ストーリーと話題の関係や，質問とストーリーの適格性の関係など複数の確率モデルを用意し，これら統合することで，話題の検出・追跡の性能を向上できるとしている。

その他ポスターでは，既存単語に囚われずに語彙単位を構成しようとした発表があり興味深かった。（文責　小林）

### 3.6 発話照合 (Uterance Verification)

単語認識におけるreject機能やキーワード認識では，confidence measureを観測して性能を改善する必要がある。この場合，1-passで照合と同期して計測できれば効率的だが（新しいLLRの提案 ；Leungら(HKUST)），HMM decoderと別に独立した韻律情報(Chenら(NCKU))や他の方式(MLP等；Kirchhoff ら(Bielefeld大))を使用する2-passが，より高次の情報を組込めるため性能的には有利である。また，複数の方式で得たmeasures（知識）を組合せることで性能を改善する報告も複数あった(NNによる統合；Wendemuthら(Philips)ほか)。

### 3.7 ロバスト音声認識

分析手法の改良により耐ノイズ性能を向上する方向では，サブバンド毎にTeager Energyを計算した後ケプストラムを計算する(Jobloun ら(Bilkent大))，2本のマイクからcoherenceを計算してケプストラムを補正する(Peters(BMW))，対数スペクトル上でKalman Filterにより補償する(Kim(Seoul大)，PMCに基づくノイズ補償法を改良する(Hongら(NCTU))などの発表があった。

また，EMアルゴリズムの中で伝達関数やノイズを補正する発表も相変わらず多い(Gongら(TI), Wongら(HKUST), Fischerら(Philips), Giulianiら(ITC-IRST))。

この他では，GSM-encoder出力を(decodeせずに)使うことで，エラーに強い安定した性能を得たという発表があった(Gallardo-Antolinら(CarlosIII大)。

### 3.8 音声認識応用

音声認識の応用では車載向けが非常に多かった。貴重な発表としては，高齢者(68-98歳（平均79歳）297名；うち40名がtest set)による音声入力評価の発表があった（ Andesonら (Dragon))。記事の検索タスクを用いた客観的評価結果は，タイプと比較し，速度，ヘルプの必要回数共変わらなかったが，音声の方が圧倒的に速いと感じて好んだとのこと。

### 3.9 話者認識・言語識別

J.Heらは，発話の長さと話者同定率・照合率の関係を理論的に考察し，実験結果とよく合致することを示した。これは，尤度の分散はサンプル数が増大すると小さくなるということを利用したモデルである。この他，話者認識に有効な音声区間に重みを付ける方法が２件発表された（正規化尤度のJensen距離の利用，音韻カテゴリの利用)。

話者適応技術による詐称音声の合成によって，話者照合率が極端に悪くなることを示したB.L.Pellomらの発表は注目に値する。また，話者認識用評価DBの発表もあった。応用では，会話音声からの話者分割とドキュメントからの特定話者抽出の発表があったが，結果はいずれもそれほど良くはない。

言語識別の発表では，母音区間の抽出に基づく混合ガウスモデルを用いた方法の１件だけで低調だった。 ■