音声言語情報処理 32−14
（2000．7．15）

ICASSP2000に見る世界の研究動向

武田一哉（名古屋大学）徳田恵一（名古屋工業大学）今井 亨（NHK）
中村 哲（ATR音声言語通信研究所）河原達也（京都大学）

あらまし： 2000年のICASSP (International Conference on Acoustic Speech and Signal Processing) が，6月5日から9日の間，イスタンブールヒルトンホテルにて行われた．3964件の投稿の中から採録された920の論文と56件の招待論文が発表された．音声に直接関連した247の論文が音声関連の21のセッションにおいて発表された．本稿では，会議の内容梗概を報告し研究動向をまとめる．

ICASSP 2000 Summary

Kazuya TAKEDA (Nagoya University), Keiichi TOKUDA (Nagoya Institute of Technology),
Toru IMAI (NHK), Satoshi NAKAMURA (ATR Spoken Language Telecommunication Laboratories)
and Tatsuya KAWAHARA (Kyoto University)

1. 音声強調・特徴抽出・音響モデル関連

雑音下，残響下における遠隔音声受音がいくつかの発表において問題として取り上げられた．アレイ信号処理を用いることで，複数受音から空間指向特性を構築し利用する方法が中心的手法である．（II-1037, II-1049）．同様の複数受音に基づく手法に，98, 99年のICASSPで独立したセッションとして取り上げられた独立成分分析（ICA）があるが，今回はこれを応用した手法が音声強調だけでなく音声認識のセッションでも発表された．（II-1041, II-1133, III-1747）いずれも，ベースラインのパフォーマンスを改善しているが，改善後の性能は未だ実用レベルには至っていない．

複数の特徴量を併用する音声認識手法であるサブバンドモデルに関して，特徴量間の時間的同期の観点からいくつかの報告がなされた．状態間の遷移を特徴量毎に非同期に行うモデル（II-1005）と，遷移の間の同期をモデル化する手法（II-1619）がともに報告されたことは興味深い．音声強調・特徴抽出全般には，複数信号（情報）の統合に基づく音声情報の頑健な処理・情報抽出が研究の基本方向である．

音響モデルに関しては，モデル化の基本単位，単語辞書の構成方法に関する検討が数多く発表された．（III-1663〜1691）対象とする発話の拡大において重要な研究の方向であり，今後の発展が期待される．その他，音響モデルの確率計算の高速（効率）化，識別的学習，セグメント特徴量の利用，など従来の研究の方向に沿った報告も数多く発表された．（武田）

2. 音声分析・合成・符号化関連

（符号化）ICASSPにおける音声符号化関連の発表は，同じ音声関連の会議であるEUROSPEECH, ICSLPに較べると実際的なものが多く，件数も多いのが通例である．但し，種々の標準化方式が定まってきたためか，一時のように，特定の標準化を目指して競合した方式が数多く発表されるという熱気溢れる雰囲気は若干薄らいできた．今回は，計40件の音声符号化関連の発表が行われた．セッションは，Low bit rate speech coding, Wideband speech coding, Topics in

speech coding-part1; part2 の４つであり，その他にオーディオ符号化関連のセッションがあった．

標準化関連では，ITU-T の 4kbps 電話帯域音声，16kbps 広帯域音声，ETSI GSM AMR-WB，MPEG4/CELP に関したものがあった．

低ビットレート音声符号化方式としては，従来の CELP (Code-Excited Linear Prediction)，WI (Waveform Interpolation)，MELP (Multi-Band Excitation Linear Predictive Coding)，正弦波符号化(Sinusoidal Coding)など，従来の方式をベースにしたものが多かった．また，これらを組み合わせるものもあった(CELP+MELP: Ⅲ-1379, WI+MELP: Ⅲ-1387)．WI あるいは正弦波符号化をベースに，各パラメータを ABS (Analysis-by-Synthesis) により量子化する方式 (CELP と同様，閉ループ系によりパラメータを量子化する) が目に付いた(正弦波符号化：Ⅲ-1371, WI: Ⅲ-1363)．インターネットや無線環境を意識した発表，また，情報源・通信路同時符号化に関する発表もあった．

広帯域音声符号化では，従来の電話帯域符号化を元に，スケーラブルの構成にする方式が目立った(Ⅱ-1141, Ⅱ-1145, Ⅱ-1149)．これは，広帯域音声符号化方式の各標準化において，スケーラブルであることが要求条件とされたためとのことである．4kHz 以上の周波数成分は，振幅変調した雑音によりモデル化できるとする発表が２件あった(Ⅱ-1153, Ⅱ-1157)．MELP を広帯域音声符号化に適用したものもあった(Ⅱ-1137)．

ベクトル量子化，励振源など，要素技術に関する発表もあったが，中では，GMM をスペクトルパラメータの予測ベクトル量子化に用いる手法が個人的に興味を引かれた(Ⅲ-1451)．予測ベクトル量化よりも若干よい客観・主観性能を得ているようであった．その他，背景雑音を考慮し，音声強調と符号化系を関連づける手法(Ⅲ-1479)，後処理を行う手法(Ⅱ-1165)などもあった．

（音声合成関連）同じ音声関連の会議である EUROSPEECH, ICSLP では，多様な音声合成方式の発表があるが，ICASSP では，現在主流の単位接続型音声合成(Concatenative Speech Synthesis) の発表が中心で，全体的な件数も少なめであるように思われる．音声合成に関連したセッションは，Speech synthesis および Topics in speech processing -part1: Speech synthesis & analysis の二つで，これらにおいて十数件の発表が行われた．

正弦波モデルによるピッチ変形やスペクトル変形の発表が目に付いた(スペクトル変形：Ⅱ-941，効率的な計算法：Ⅱ-957，時間周波数スケールの変形：Ⅲ-1295)．英語以外の言語（マルチリンガルを含む）に関連したシステムに関する発表もいくつかあった(Ⅱ-929, Ⅱ-933, Ⅲ-1281, Ⅲ-1285, Ⅲ-1291)．他に，単位接続型音声合成におけるセグメントの予備選択(Ⅱ-937)，ピッチに依存したスペクトルの変形法(Ⅱ-949)，ポーズ挿入法(Ⅲ-1289)，HMM 音声合成(Ⅲ-1315)などの発表があった．

（音声分析関連）音声分析に関しては，Topics in speech processing -part1: Speech synthesis & analysis, part2: speech analysis においてバラエティーに富んだ発表が行われた．

ピッチ抽出関連(Ⅲ-1307, Ⅲ-1339, Ⅲ-1343)，声道長の推定(Ⅲ-1319)，MFCC からの音声合成法(Ⅲ-1299)，ウェーブレット変換を用いたエポック検出(Ⅲ-1303)，ARX 分析(Ⅲ-1331)，メル周波数ウェーブレット変換の音声認識への応用(Ⅲ-1351)，話速推定(Ⅲ-1355)他があった．（徳田）

3. サーチ・言語モデル

サーチに関する発表は，主に Fast Decoding (SP-P6) と Search Techniques (SP-P8) と題された２つのポスターセッションにて行われ，高精度化・高速化を目的とした興味深い多数の報告があった．

探索の高精度化のための発表では，単語事後確率を用いたリスコアリングの報告(Ⅲ-1655, Ⅲ1587)の関心が高かった．これらは，最初の探索で得られ

た単語ラティス（あるいはグラフ）を同一時刻の同一単語でマージするなどして再構成しておき，forward-backwardアルゴリズムによって単語事後確率を求めて最適解を探索しようとするものである．単語事後確率は単語の信頼度としても使われており，今後の研究の発展が期待される．

サーチアルゴリズムの改良の研究では，スタックデコーダにおけるいくつかのスタック選択法の比較検討(III-1555)，単語終端のクロスワード用ノードを探索中に動的に生成する方法(III-1671)，状態継続時間長に基づく枝刈り(III-1583)，ビーム幅を動的に更新する方法(III-1535)，発話終了前に逐次的に第２パスを実行することによる単語の早期確定法(III-1559)などの探索高速化の報告や，話者交替を言語モデルで表現しつつ話者認識と単語認識を同時に行う探索法(III-1575)の報告があった．

音響尤度計算法の改良による探索高速化の研究では，ガウス分布をサブスペースでクラスタリングする方法(III-1519)，プロセッサのSIMD (Single Instruction Multiple Data) 命令を利用して音響尤度計算に必要なl_2ノルムを３倍高速に計算する方法(III-1531)などの報告があった．

目新しい研究では，言語処理で広く使われるようになってきた有限状態変換器(FST; Finite- State Transducer)を音声認識に適用した発表(III-1675)があった．これは，HMM，バイグラム，発音辞書をすべて重み付きFSTの合成で表現しておき，FST上でフレーム同期Viterbiサーチを実現するもので，HTKに比べて25倍高速にデコードできることが示された．また，FSTは特に発音辞書の自動生成にも適していることが示されている(III-1683)．

（言語モデル）言語モデルに関する発表も，主にFast Decoding (SP-P6)とLanguage Modeling (SP-P8)と題された２つのポスターセッションにて行われ，基本的にN-gramモデルをベースとした改良モデルの報告が多数あった．

III-1591：観測された単語頻度からポアソン分布にしたがってN-gramモデルを更新していくもので，キャッシュモデルに似た効果を実現している．放送ニュースのタスクにおいてパープレキシティの削減が報告されているが，スパースネスの問題のためにbigram以上では効果が小さくなる．

III1595：最大エントロピー法の枠組みで，意味的あるいは構文的な質問を単語トリガーモデルと併用したもの．ホテル予約タスクのリスコアリングにおいて，パープレキシティと認識率の改善が報告されている．

III-1639：ドメイン依存の文脈自由文法(CFG)とN-gramモデルを組み合わせた言語モデルを謳っているが，実際にはクラス・モデルと同等．スケジュール管理タスクにおいて大幅なパープレキシティの削減が報告されている．

III-1643：可変長N-gramに２単語フレーズを組み込んだもの．頻度の高い単語連鎖をフレーズとして自動的に検出するところが特徴．スイッチボード・コーパスに対して若干のパープレキシティと認識率の改善が報告されている．

III-1647：高次N(5)-gramモデルにキャッシュ・モデル，単語スキップ，クラス・モデルのすべてを組み合わせてスムージングしたもの．WSJタスクにおいてパープレキシティと認識率の改善が報告されている．

III-1695：学習コーパス中の記事を選択して言語モデルを構築することにより，特定の記事に対するパープレキシティと未知語の削減を図ったもの．モデルのサイズも1/3に削減される．

III-1699：N-best文から構文解析によって主辞(syntactic head)を検出し，最大エントロピー法によってN-gramモデルの改善を図ったもの．スイッチボード・コーパスに対して若干のパープレキシティと認識率の改善が報告されている．（今井）

4．適応・ロバスト処理

最近では，音響モデル，適応化，ロバスト処理の区別が非常に曖昧になりつつある．おおよそ，話者などの複雑な対象に対するモデル適応が適応化のセッションに，雑音に特化したモデル適応がロバスト処理のセッションに主に区分されている．適応化関係では，約 10 件の発表がなされた．内容としては，これまでの主流であるMLLR, MAPの精度をさらに高め

るための研究と(MLLR 5件, MAP 2件), 両者の組み合わせの同時最適化研究（3件）に2分される．すでに，両者を併用することが普通であるが，それぞれの手法の改良研究がそろそろ飽和し，2つの手法の統合へ研究課題が移りつつあることがわかる．まず，MLLR の改良については，複数の回帰行列の補間により精度を改善するもの（Boulis らII-989）に関する結果が報告された．また，回帰クラスの相関を利用するものも2件あり，Doh ら（III-1543）らは複数クラスからの相関を利用する方法，He ら（II-981）は近傍のクラスからの重み付き MAP-MLLR を，Wong ら（III-1551）は回帰クラスの設計法に関する報告を行ない，いずれも改善効果を確認している．学習データの不足を考慮した Discounted MLLR （II-985）もアイデアとしてはおもしろい．

MAP 関係では，Buhrke ら（II-993）により混合ガウス分布を Prior にする報告が，Wang ら（II-977）により階層的 MAP 適応において一般化ガウス分布を Prior 分布として利用する報告がなされた．また，デルタケプストラムの適応法の報告もあった（II-973）．MLLR と MAP の最適化に関しては，Siohan ら（II-965）により同時最適化の枠組みが示された．MLLR と MAP は利用される条件が異なるが，学習データ量が少量の時は MLLR が適用され，データ量に従って MAP 的に動く，非常に素性がよく合理的な方法である．

次に，ロバスト処理に関してであるが，音声認識をターゲットした発表を分類するとおおむね以下のようになる．但し，従来の近接発話の音声に対し雑音，歪みが加わった場合を対象とするものと，部屋の中での遠隔発話を対象としたものの2つに分類した．(括弧内は関連発表件数)

近接発話認識：

エンドポイント検出（2），特徴抽出（3），モデル適応・モデル合成（5），マルチバンド（2），ミッシングデータ理論（1），セルラーネットワーク（1），IP ネットワーク（1）

遠隔発話認識：

モデル適応（2），マイクロホンアレー（7），ブラインド分離（2）

これを見てもわかるように，特徴抽出，モデル適応関連の研究が中心であり，また，遠隔発話認識の報告がかなり多いこともわかる．自動車を始めとしたハンズフリーインタフェースのニーズに応じたものと思われる．

近接，遠隔発話のいずれも雑音，歪みへの対処をいかに行うかということが問題になるが，最近では定常的な雑音，歪みへの対処法から非定常の雑音，歪みへの対処法に問題が移行しつつあると思う．特に，実環境の雑音は非定常なので，これらへの対処は不可欠である．発表の中からいくつかを紹介する．

特徴抽出では，線形判別分析(LDA)を用いてフィルタ設計する方法やPLP を用いる方法が報告された(II-1105)．モデル適応では，非定常雑音に対処する方法としてカルマンフィルタを用いる方法が Yao ら，Fujimoto らにより報告された（II-1125, III-1727）．Yao らは非定常雑音に対し，非定常項をカルマンフィルタにより推定する手法が PMC より優れていることを報告した．（ただし PMC の雑音の状態数は1状態）Zho（II-1109）は，伝達特性とノイズを同時に推定するアルゴリズムを提案している．音声を混合ガウス分布に従う信号，雑音をガウス分布に従う信号としてモデルパラメータを EM アルゴリズムで推定する．音声のモデルとして HMM を使わず GMM を使う点が特徴である．遠隔発話音声認識では，主にマイクロホンアレーを利用した方法，多チャンネル受音後，ブラインド分離の原理に基づいて目的信号を抽出する方法，モデル分解合成法により認識する方法に分けられる．マイクロホンアレーを用いた研究では，猿渡らによる相補的マイクロホンアレーにより少数マイクロホン素子で SNR を大幅に改善し認識率を改善する報告（II-1049），その他，マイクロホンアレー処理と後段の補正処理を併用することによる性能改善に関する報告が多数なされた（III-1723, III-1407, III-1411）．また，ブラインド分離に関する多数の報

告がなされたが，複数の移動話者の音声認識への適用結果が興味を引いた（Ⅱ-1137）．また，多数マイクロホンを尤度で選択しながら認識を行う手法（Ⅲ-1747），単一マイクロホンでHMM分解合成法に基づき伝達特性を推定し移動話者の音声認識をする手法（Ⅲ-1747）が提案された．（中村）

5．音声対話・信頼度尺度・話し言葉認識

今回のICASSPにおける音声認識関係の発表では，
(1)言語モデルのセッションがなかった
(2)信頼度尺度，音声対話のセッションがそれぞれ単独に設けられた，という点が目新しかった．
信頼度尺度(Confidence Measure)については，(a)競合モデル(anti-model)，(b)音素識別器の結果，(c)N-best候補などとの尤度比を用いる，のが基本的な考え方であるが，これらを組み合わせる他，言語モデルや音素識別の混同表などを考慮するなどの工夫がみられた．これとは別に，大語彙連続音声認識のセッションにおいても，Cambridge(Ⅲ-1655)とAachen(Ⅲ-1587)から独立に，単語の事後確率を基準としてデコーディングする方法が発表されたが，文全体の認識よりも単語認識精度を優先し，信頼度尺度計算を指向したものである．

音声対話においては，シミュレーションによって評価を行う発表が2件あった．

言語モデルやサーチアルゴリズムの発表はあまりなかった半面，ロバストネスや話し言葉を指向した音響モデル・発音モデルの論文が目立った．発話速度や方言を考慮したり，韻律情報を利用したりする試みもいくつか発表されていた．音響モデルでは，話速別にモデルを用意して長さ0のモデルを許したり(Ⅲ-1779)，言い直し音声(hyper-articulated speech)専用のモデルを用意する(Ⅲ-1779)などの発表があった．発音モデルでは，発音の生起確率を考慮したり(Ⅲ-1659)，baseformとsurface-formの対でモデルを用意する(Ⅲ-1679)などの発表が興味深かった．ただし，いずれの発表も認識率の向上は大きいとは言い難く，問題の難しさを改めて感じさせるものであった（河原）