音声言語情報処理　28－4
（1999．10．29）

Eurospeech99, IEEE MMSP99 会議報告

中村　哲[1]，大川茂樹[2]，伊藤彰則[3]，田本真詞[4]，
水野秀之[5]，鵜木祐史[6]，徳田恵一[7]，鏑木時彦[4]，畑岡信夫[8]

[1] 奈良先端大　情報科学研究科，[2] 千葉工業大学 情報ネットワーク学科
[3] 山形大学　工学部　電子情報工学科，[4] NTT コミュニケーション科学基礎研究所
，[5] NTT サイバースペース研究所，[6]ATR 人間情報通信研究所，
[7] 名古屋工業大学　知能情報システム学科，[8] 日立中央研究所

あらまし　本稿では，1999 年 9 月 5 日から 9 日にハンガリーのブタペストで開催された ESCA の Eurospeech99 および 9 月 13 日から 15 日にかけてデンマークのヘルシンガーで開催された IEEE　Multimedia Signal Processing Workshop の報告を行う．

キーワード　音声情報処理、マルチメディア信号処理

A Report on Eurospeech99 and IEEE Multimedia Signal Processing Workshop

Satoshi Nakamura[1]，Shigeki Okawa[2]，Akinori Itoh[3]，
Masafumi Tamoto[4]，Hideyuki Mizuno[5]，Masashi Unoki[6]，
Keiichi Tokuda[7]，Tokihiko Kaburagi[4]，Nobuo Hataoka[8]

[1] Nara Institute of Science & Technology，[2]Chiba Institute of Technology
[3] Yamagata University，[4] NTT Communication Science Labs.，
[5] NTT Cyber Space Labs.，[6] ATR Human Information Processing Labs.，
[7] Nagoya Institute of Technology，[8] Hitachi Central Research Labs.

Abstract　This paper summarizes the topics in ESCA Eurospeech99 held at Budapest, Hungary, from Sep.5 to Sep.9, 1999 and in IEEE Multimedia Signal Processing Workshop held at Helsinger, Denmark, from Sep.13 to Sep.15, 1999.

key words　Speech processing, Multimedia signal processing

1　はじめに

本稿では，1999 年 9 月 5 日から 9 日までの 5 日間にわたり，ハンガリーのブタペストで開催された Eurospeech99 および 9 月 13 日から 15 日の 3 日間デンマークのヘルシンガーで開催された IEEE Multimedia　Signal Processing Workshop についての概要報告を行う．

Eurospeech99 は今回で第 6 回目となる ESCA 主催の音声情報処理に関する会議であり，年々規模を拡大している．当初は、Eurospeech という名前通り参加者はほとんど欧州からで、ICASSP に比較して発表の質もいまひとつの感があったが，最近では米国，日本などのアジア諸国、オーストラリアからも多くの発表があり、それに伴って発表の質も向上しつつある．ESCA の Moore 氏の話では，欧州を超え国際的な組織にして，2001 年は Interspeech という名称で開催したいということであった．ちなみに，2001 年はデンマーク Aalborg にて 9 月 3 日から 8 日にわたって開催される予定である．

開催地であるハンガリーは，音声生成の研究のパイオニアである W.Kempelen や聴覚の研究でノーベル賞を受賞した G.Bekesy を輩出した地である．今回は，1000 件以上の論文申込みがあり、723 件の論文が採録された．セッションは，38 の口頭発表と 39 のポスターセッション，1 つのパネルディスカッション，5 つの KeynoteSpeech から構成されていた．

セッション数をおおまかに分類すると、音声認識関連が 25 セッション，話者識別 (照合) 関連が 5 セッション，音声合成が 6 セッション，音声符号化が 3 セッション，システム関連が 6 セッション，対話、韻律が 7 セッションと認識や合成の工学的な分野が半数近くのセッションを占めていた．数に限って見れば，ICSLP 等と比較すると音声科学分野の基礎的な研究論文はやや少ない傾向にあるといえる．

一方，IEEE Multimedia Signal Processing Workshop は，IEEE の SP Society の Multimedia Signal Processing Technical Committee 主催のワークショップで今回で 3 回目のワークショップであり，デンマークのヘルシンガーで開催された．このワークショップでは，画像，音声，マルチメディア { 信号処理，通信，データベース，システム } に関する 113 件の論文が発表された．最近では，画像，音声それぞれの技術が進展して，両方を統合的に利用することによる情報検索やそのためのインデキシング，符号化など新しい分野，アプリケーションの開拓が盛んに行われている．このワークショップは，マルチメディアの中の音声情報処理の今後の展開を議論する 1 つの場として位置付けられる．ただ，現状では発表，参加者とも画像分野の研究者が多く今後の音声研究者の参加が期待される．

(中村　哲)

2 Eurospeech99

2.1 音声認識関連

音声認識関連のセッションは 25 で最もセッションの多かった分野である．大まかにセッション数を分類すると，音響モデル+信号処理 (7)，言語モデル (4)，適応 (4)，Search　+Conficence Measure +LVCSR(8) の他，Speaking Style(1)，Multilinguality(1) のセッションで構成されていた．Speaking Style，Multilinguality がそれぞれセッションを構成したことは，発話速度や多言語の音声認識が音声認識で大きな関心事になっていることを示している．また、Speech & Noise なる Robustness に関するセッションがあり，合計 4 つのうち 2 つのセッションで音声認識における騒音の問題が議論された．

(中村　哲)

2.1.1 音響モデル

音響モデルに関するセッションでは，口頭 15 件，ポスタ 33 件の発表が行われた．内容としては，HMM の状態や分布の共有に関するもの，音素決定木の分割基準に関するもの，尤度計算の基準に関するもの，ANN/HMM ハイブリッドシステムを用いたものなどが相変わらず多いが，新たな傾向として，Multi-stream ASR (Multi-band ASR を含む) に関するセッションが設けられていたことを報告しておく．以下，特に興味深かったものを口頭発表を中心にいくつか紹介する．

Albesano ら (CSELT) は，FFT および蝸牛フィルタを用いて周波数帯域上の gravity centers (重心周波数) を計算し，MFCC 特徴量に加えて用いるというシンプルな方法により，雑音下での単語誤り削減率 20.1 %を報告した [1].

Gales ら (IBM) は，HMM の出力確率を表現するのに，ガウス分布の代わりに Richter 分布および power exponential 分布を用いて，より忠実なデータ分布 (特に分布の裾部) の表現に挑戦した．認識性能向上は僅か [2].

Liu ら (OGI) は，decision tree の構成法として，ガウシアンを共有するノード (従来形) と，重みを共有するノードの 2 段を設定した two-level decision tree を提案し，単語誤り削減率 10 %を報告した．学習データに応じてパラメタ数を調整しやすいとしている [3].

Sanker ら (SRI) は，HMM の状態共有において音声学的な (音素カテゴリ毎の) クラスタリングを

行うと，性能を落とさずに 10 分の 1 のサイズのモデルを構成でき，かつ計算量を減らすことができると報告した [4].

Singh ら (CMU) は，異なるタスクのデータでモデルを設計する cross-domain modelling について言及し，別タスクで生成した decision tree を，自タスクの少量データで最適に融合する効果を示した [5].

(大川茂樹)

2.1.2 言語モデル、大語彙音声認識

言語モデル関連では，ポスターが 2 つ，口頭発表が 2 つの計 4 セッションが設けられた．まず，全体の傾向を見るために，関連論文を独断でおおまかに分野ごとに分けて（重複あり），2 件以上発表されたテーマを並べてみると，次のようになる．

> 文法などの利用 (6)　適応 (4)　最大エントロピー法 (4)　モデル評価法 (3)
> 複合語登録 (3)　クラス n-gram(2) 学習データクラスタリング (2)　単語クラスタリング (2)　識別学習 (2)　各種モデルの比較評価 (2) 話題依存モデル (2)

N-gram 自体を改良するというよりも，文法的な情報と n-gram を組み合わせるという方法が増えてきたことがわかる．Jelinek のグループによる structured language model[6] や，文法情報によるトリガ [7] などのほか，伝統的な文法をそのまま使ったシステム [8] や SCFG の利用 [9] など，n-gram の台頭でいったん下火になっていたテーマが復活した感がある．

言語モデル適応はもうひとつの中心テーマである．十分な学習データがない場合の対策という意味では，クラス n-gram も問題意識は同じといえるかもしれない．適応手法では，従来の n-gram カウント混合，n-gram 確率混合，最大エントロピー法などの他，MDI による適応手法が提案されていた [10].

最大エントロピー法は，複数の言語制約を組み合わせるスタンダードな手法として定着した感がある．識別学習によるモデルでも，最大エントロピー法と同じ GIS アルゴリズムを用いているので，GIS によるモデル推定は今後のトレンドになるかもしれない．

パープレキシティに替わる評価方法の発表が多かった [11, 12, 13] のも，今回の Eurospeech の特色である．パープレキシティの問題点については，前回の Eurospeech'97 で熱い議論があったので，その成果といえるかもしれない．しかし，決定版といえるようなものはなく，評価手法はこれからのテーマであろう．

N-gram に長距離制約を入れる手法の提案としては，上記の文法ベースのもののほか，複合語登録によるもの [14, 15]，distant bigram を組み合わせるもの [16]，LSA を用いるもの [17] などがあった．

その他に注目すべき発表として，Confidence のセッションの中に，単語の信頼度を言語モデルに統合するという発表があった [18].

LVCSR 関連では，LVCSR の口頭発表が 1 セッション，Broadcast News がポスターと口頭発表各 1 セッションであった．LVCSR のセッションはシステムの全体に関する発表（5 件）だったが，そのうち 3 件が中国語のシステム（台湾含む）なのが特徴的であった．AT&T のシステムについての発表 [19] では，細かい改良が積み重なって全体の性能が改善されていく様子が報告され，興味深かった．Broadcast News のセッションでは，主な話題は音響信号のセグメンテーションであった．話者の交代，音声・非音声，笑いなどの区間を自動的に推定し分割することが当面の課題であろう．また，自動キャプションを意識してか，デコーダの高速化に関する発表も目立った．

Search 関連では，Search and pronunciation modeling（ポスター）と，Search （口頭）の 2 つのセッションが設けられた．単語グラフに含まれる候補を比較して出力結果を改良するという発表が 2 件あり [20, 21]，原理的に同じ問題を違う方向から解いているのが面白かった．また，いったん音素系列を認識してからデコーディングを行うという発表があった [22, 23]．これは昔よく行われていた手法だが，現在では音素認識の精度が上がっているので，高速化のひとつとして試してみるのも面白いかもしれない．

(伊藤彰則)

2.1.3 適応化および雑音下音声認識

適応化に関しては 3 つのセッションがあった。ほとんどが MLLR に関する改良で，変換行列をクラスごとに求める，Sequential に求める，事前分布を推定し MAP 推定を行う，EigenVoice と組み合わせるものなどが発表された．

雑音下音声認識では，実用化が進んでいることを反映して，カーナビゲーション，携帯電話用の音声認識の発表が目立っていた．特に，欧州では組織的に自動車内の音声を収録しているようである [24]．手法的には，雑音減算法や Blind Deconvolution, モデ

ル適応化，Missing Data Theory による方法が用いられていた。

(中村　哲)

2.1.4　音声対話関連

音声対話の発表は、それぞれ2セッションずつ10件の講演と26件のポスター展示で行なわれた。話題ごとに主な研究発表の傾向を紹介する。

[特定ドメインの対話システム]

KPN,LIMSI-CNRS,CSELT の共同で3ヶ国語対応の列車情報対話システム"ARISE" を報告。対話状態ごとの言語モデル、対話管理、情報提示方式とサービス達成率による評価が報告された [25]。

[対話戦略]

Lucent では、タスク依存の知識から脱却した一般化 mixed-initiative 対話管理戦略を設計評価。対話マネージャに与えられた知識や対話履歴で解決できない問題が生じたときに必要な情報をユーザに要求するサブ対話を開始する [29]。同様の発表で認識率など話者の特徴に応じて対話戦略を変化させるシステムを設計し、PARADISE に基づく評価尺度で対話達成度の改善が報告された [27]。また、mixed-initiative な対話システムで、確認対話のストラテジとして選択肢を明示しない質問の場合は心的操作の順番に並べられた対話が対話を長引かせることなく達成できることが報告された [28]。

[対話管理や対話タスクの記述]

Lucent では、タスク記述とライブラリから対話の状態遷移図と各状態でのアクションを自動的に生成する Automatic Dialogue Generator(ADG) の概念を紹介、タスク記述はテンプレートにしたがってユーザが行う。ドメイン依存知識を格納するライブラリの拡充を目指している [32]。同様に、対話タスクを構成するサブタスク、サブタスクで要求される情報をフレームやスキームで記述する研究発表が散見される [33, 31, 36]。IBM の研究では、対話タスクを構成する個々のサブタスクをフレームで記述し、対話マネージャがユーザの意図に従って最適なフォームを選択することで必要な情報確認を行なう。ユーザが mixed-initiative に任意にサブタスクを遂行できる [34]。CMU からは、タスク遂行に必要な情報がスキーマで記述された旅行案内システムにおける音声認識、対話管理、ドメイン依存の知識処理を報告 [35]。認識方式の研究として MIT では、第一段にドメイン非依存、第二段にドメイン依存の音声認識を備えた対話システムの認識誤りの低減を評価した。第一段の認識結果は、音韻レベルの音響-音声的ネットワークを出力する [26]。

(田本真詞)

2.2　音声合成関連

音声合成関連の発表は、'Speech generation /synthesis' として6件、'Prosody' として45件の発表があり、これらのセッションで主に音声合成に関連する発表が行われた。その他に'Assessment'、'Speech Analysis architectures'、'Systems, architectures, interfaces'、'Speech Internet'、'Corpora'、'Speech analysis and segmentation' などのセッションにおいて音声合成に関連する発表も含まれていた。

音声合成の研究動向として、音声合成単位関係の研究について言えば従来の TTS（Tex t-to-Speech）システムでは Diphone をベースとしたシステムが主であったが、最近では non-uniform な単位の接続に基づく音声合成方式が主な研究テーマとなってきている。今回の EUROSPEECH においても non-uniform な単位での接続に基づく多くの発表があった。このような方式は大容量データベース中の音声素片を一部又は全て変形せず、接続・合成することで従来の Diphone や音素単位の合成システムと比較して極めて高品質な合成音声が生成できる可能性がある。しかし、データベースの構築、素片の選択・接続、韻律の再現等多くの課題があり、発表内容もそれらの課題に関するものが主であった。例えば、"Word and Syllable Concatenation in Text-to-Speech Synthesis" (Lewis, Bristol 大) や"Synthesis by Word Concatenation"(Stober, Bonn) の 2件の発表は、単語ベース単位接続による信号処理を用いない素片接続型の合成系の提案であり、タスクを限定すれば自然音声に近い品質の合成音声の生成が可能であることを示すものであった。素片選択に関しては、"Rapid Unit Selection from a Large Speech Corpus for Concatenative Speech Synthesis"(Beutnagel, AT&T) での事前学習により素片選択を高速化する提案や、最適な素片を選択するための選択基準についての発表として、"C hose the Best to Modify the Least:A New Generation Concatenative Synthesis System" (Balestri, CSELT)、"Selection of Waveform Units for Corpus-Based Mandarin Speech Synthesis Based on Dicision Trees and Prosodic Modification Costs" (Chou, Taiwan 大) などでは、ある程度

の韻律変形処理を許容した上で変形を最小とする音声素片選択方法についての提案がされていた。また、多言語対応も音声合成システム構成上のポイントとなりつつあり、既存システムの他言語への拡張、多言語合成システムについての発表もあった。

韻律生成の研究では、決定木やニューラルネットを利用した統計的な学習による韻律生成の方式に基づく複数の提案があった。高品質な韻律制御を実現する目的の他に、特定話者の韻律パターンを模擬したり、多言語対応のための自動学習、タスク限定で非常に高品質な韻律を実現する目的としたもの等、従来のヒューリスティックな規則での韻律制御では困難な様々な分野での音声合成の利用の拡大を図るものとなっている。

その他に、音声合成を応用したソフトウエアや音声合成を利用した対話システムの構築や評価について多くの発表があった。その中で子供向けにターゲットを絞ったシステムの構築等の報告があり、例えば"Child-Directed Speech Synthesis: Evaluation of Prosodic Variation for an Educational Computer Program"(House,KTH) では、大人と子供では同じ合成音声でも評価が異なることが報告されており、子供向けのシステムにおいては子供の評価が必要であることを示唆していた。

その他、EUROSPEECH'99 と同じ会場においてSigSyn の会合が 9/5、20 時からあった。内容としてはこれまで 1 年間の活動報告と今後の活動計画に関する討論であった。その中で昨年度の音声合成ワークショップの余剰金の使用方法についての討論があったがその場では決定されず、現在まだ意見を募集中である。また次回の音声合成ワークショップが 2001 年にスコットランドで行われることが決定された。

(水野秀之)

2.3 音声知覚、音声信号処理関連

知覚・信号処理に関連するセッションは、音声知覚 (24 件)、音声と雑音 (52 件)、音声信号処理 (5 件)、エンハンスメント、エコーキャンセレーション、音質評価のセッション (13 件) など、90 件弱の発表があった。

音声知覚では、単語知覚やセグメンテーション、知覚過程における同定や弁別に関係した発表が中心的であり、視覚情報による音声知覚への影響や声質・明瞭度に対する主観的／客観的評価法の提案、フォルマント遷移のマスキングのモデル化に関係した発表もあった。 特に興味深かったものは、重複した音節の知覚の議論や音源の大きさを正規化する聴覚的方略の研究であり、聴覚の情景解析やカクテルパーティ効果のモデル化の研究に役立つ良い研究成果であると感じた。

音声信号処理関係では、雑音で汚れた音声信号を如何にして分離／エンハンス／雑音抑圧するかということに主眼が置かれ、音質の主観的／客観的向上あるいは、ASR の前処理としての有効性を示す発表が大半を占めた。 特に、エンハンスメント、エコーキャンセレーション等のセッションでは、時間領域における分離／エンハンスメント／キャンセレーション法の研究が中心的に発表された。

音声と雑音のセッションでは、周波数領域におけるエンハンスメント／雑音抑圧法と雑音による影響を積極的に考慮した雑音抑圧法の研究が中心的に発表された。最近の手法の多くは、サブバンド信号処理、マイクロフォンアレイ、スペクトルサブトラクションに集中しているが、最近の聴覚生理・心理の知見を活かした新しい方法もいくつか発表され、注目を集めていた。 今回の発表で目に付いたことは、ASR をゴールとしたエンハンスメントの研究だけでなく、携帯電話におけるエコーキャンセラーや雑音抑圧の研究、雑音にロバストなラベリング／コーディングの研究発表が増えてきたことである。 また、研究のアプローチとして、音声に対する制約を考慮した方法よりも、雑音に対する制約あるいはマスキングといった雑音による影響を考慮した方法に関心が集まっているようであった。

(鵜木祐史)

2.4 音声符号化関連

通例，EUROSPEECH における音声符号化関連の発表件数は，ICASSP (IEEE International Conference on Acoustics, Speech and Signal Processing) ほどは多くない．これは，本会議の伝統的な性格の他に，毎回近い時期に IEEE Speech Coding Workshop が行われることと関係しているように思われる．今回は，30 件近くの音声符号化関連の発表が行われた．セッションは，Speech coding, Wideband and perceptually based coding, Joint source-channel coding の 3 つであり，その他に Enhancements, echo cancellation and quality measures および Speech and noise 1-3 の計 4 つのセッションに背景雑音および通信路誤りに関連した発表が何件か含

まれていた.

一時の ICASSP や Speech Coding Workshop 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ことはなく，その意味で活気や熱気に欠ける印象があるが，逆に，目先の標準化や製品化などの短期的な目標に捕らわれず，ある程度将来的な方向性をもった発表が行えるという見方もできる．標準化関連では，欧州 ETSI GSM-AMR に提案されたもの，ITU-T の 4kbps 電話帯域音声および 16kbps 広帯域音声の標準化を念頭においたものなどがあった.

音声符号化方式としては，CELP (Code-Excited Linear Prediction), WI (Waveform Interpolation), MBE (Multi-Band Excitation), 正弦波符号化 (Sinusoidal Coding), Harmonic Coding など，従来の方式をベースにしたものが多かった．WI あるいは Harmonic Coding 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) が興味を引いた．非線形予測器を CELP, ADPCM に導入する試みもあった．1〜2kbps 以下の極低ビットレートについては，まだ定番の手法が確立されておらず，いくつかの方式が提案された.

これまでと同様，ビットレートを可変とすることにより，平均的なビットレート下げる VBR (Variable Bitrate) 方式，通信路誤りの状況に応じて，適応的に，音声符号化と通信路符号化のビットレート比を変える方式などの発表も多かった．その他，ベクトル量子化，インデックス割当などの符号化系の要素技術，音声強調と符号化系を関連づける手法，通信路誤りを前提とした音声符号化系の構成法などについてもいくつかの発表があった.

(徳田恵一)

2.5 音声生成関連

音声生成関連では、"Articulatory measurements and modeling"と題したポスターセッションにおいて 16 件の発表があった。Eurospeech は ICSLP などの会議に比較して工学系の色彩が強いことを反映して、音声学に代表される分析的研究は比較的少なく、調音モデルや音響モデルに関連した数理よりの研究が多く発表された。研究動向としては、生成過程を反映した高品質で柔軟な音声合成や調音パラメータを用いた音声認識への適用を究極の目標とした、3 次元の調音モデルや声道音響モデル、動的調音モデル、音響・調音逆マッピングを挙げることができる。

3 次元調音モデル (Engwall [42]) や音響モデル (Miki [44]; Matsuda [45]) は、MRI などの調音観測技術の進展や、声道音響系の等価電気回路網表現を背景としている。とくに 3 次元調音モデルでは、10 個という非常に少数のパラメータで声道形状を制御可能であり、音声合成や発声トレーニングなどへの応用が期待できる。今後の課題としては、音素固有の声道形状から連続性の意味で最適な調音運動軌道を計算する動的調音モデル (Kaburagi[43]) の適用により、連続音声における調音結合の生起を含んだパラメータ制御技術を確立することが重要である。

音声認識に調音パラメータを用いる動機は、調音パラメータの連続性やパラメータ空間における音素特徴の局存性・不変性を利用可能な点にある。音響・調音マッピングは、この目的のために調音パラメータを推定するものであり、調音・音響パラメータの値の組を多数保持するデータベースの検索に基づいた手法が多い。今回は、音響・調音マッピングにおけるデータベース構築法に関連して、順方向マッピングの感度を考慮する方法 (Ouni[46]) や調音系の幾何構造を制約として用いる方法 (Silvae[47]) が提案された。

(鏑木時彦)

3 MMSP99 報告

3.1 概要

標記国際会議が、デンマークのコペンハーゲンから列車で約一時間ほど北へ行ったヘルジンガーで 9 月 13 日〜15 日の 3 日間開催された。映像・画像、音声・オーディオ等の信号処理及び応用が主で、CG 等を使ったインタフェースに関する発表もあった。参加者は約 150 名で、田舎のリゾートホテルに缶詰で、3 日間発表と議論に集中した。発表の題材と質に関しては、「玉石混合」で、分野にまたがるテーマをサマリーだけで査読審査するのは難しいことを示していた。また、発表と予稿集の掲載順序に対応がなく、聴講の際に発表論文を探すのに苦労した。

3.2 招待講演とパネル討論

会期中の朝一番に各 1 件、計 3 件の招待講演があった。第一日目は、MIT メディアラボ Pentland 教授が "Smart Rooms, Smart Clothes:Personalized Multimodal Inter-faces for Everyday Living"と題して、メディアラボでの Wearable Computing、ジェスチャ

認識、オーディオ・画像処理の研究紹介があった。オーディオとビデオ画像情報を信号レベルで融合して認識する方式。第二日目は、プリンストン大学のKung教授が,"State of the Art of Application and Technology of Multimedia Signal Processing"と題して、インタネット産業は経済的でかつ、有効な情報を供給する必要があるとして、3C(Computing,Content, Consumers) が地理的に離れていることを考慮した研究開発が大事とした。実際は、MPEG 標準やニューラルネット利用の映像トラッキングや検索の研究紹介。第三日目は、Philips 研究所の Huisken 博士が"Components for Hand-held Multimedia Devices"と題して、ハードとソフトを伴うメディア処理の実現に関して、CPU,DSP, Memory 一体化の構成が主となると力説。

パネル討論のテーマは"Human-Machine Interfaces: Bridging Algorithms and Ex-pectations"、モデレータはAT&Tラボの Cox 博士、パネリストはIBM Basu 博士、南デンマーク大 Bernsen 教授、他4名。討論は多岐に渡り、(a) 心理学出身の研究者が必要、(b) 進歩を制限しているのは何か、(c) 異分野の研究者間の研究はなぜ難しい、(d) 研究方向は間違っていないか、等。QoP (Quality of Perception) の導入、オーディオ・映像処理での統合 HMM の導入、CHI デモに no science 等々のコメントあり。また、IBM が ViaVoice デモを行い、今後は雑音環境下での認識が課題とした。

3.3 一般発表論文

13セッションがシリーズに構成され、発表総数は109件。分野別発表の件数は、①マルチメディア通信 (インタネットアクセス・伝送、モバイル通信等) 14件、②オーディオ・ビデオ処理4件、③ヒューマンインタフェース9件、④音声・オーディオ・音楽9件、⑤ビデオ・画像圧縮19件。⑥映像シーン解析4件、⑦マルチメディアデータベース14件、⑧マルチメディアシステム・応用20件、⑨デザイン／装置化9件、⑩画像・音声品質管理7件であった。注目発表としては、RTP(Real Time Pro-tocol)Gatewayでの MPEG Audio on Demand の品質管理に関して (Yao, 米国 Fujitsu)、ソフトウェアだけの実現とスケーラビリティがキーワード。IP Packet ロスが MPEG1 品質にどう影響するか (Hayashi, TAO)、Emotional Recognition(Nakatsu, ATR)、TV ニュースのキャプション対象にオーディオと映像 (lip 形状等) を統合して音声認識の性能向上を図る (Basu, IBM) 等の発表があった。その他に日本からは、東大、KDD、会津大、北大、NTT、ATR、電総研、工学院大、日立からの発表。来年は ICME2000(Int. Conf. on Multi-media and Expo.) となり、7月末 N.Y. 開催。

(畑岡信夫)

参考文献

[1] D. Albesano, R. DeMori, R. Gemello and F. Mana: "A study on the effect of adding new dimentions to trajectories in the acoustic space", pp.1503-1506, Eurospeech99

[2] M. J. F. Gales and P. A. Olsen: "Tail distribution modelling using the Richter and power exponential distributions", pp.1507-1510, Eurospeech99

[3] C. Liu, X. Wu and Y. Yan: "High accuracy acoustic modeling using two-level decision-tree based state-tying", pp.1703-1706, Eurospeech99

[4] A. Sankar and V. R. R. Gadde: "Parameter tying and Gaussian clustering for faster, better, and smaller speech recognition", pp.1711-1714, Eurospeech99

[5] R. Singh, B. Raj and R. M. Stern: "Domain adduced state tying for cross-domain acoustic modelling", pp.1707-1710, Eurospeech99

[6] Ciprian Chelba, Frederick Jelinek: "Recognition Performance of a Structured Language Model", pp.1567-1570, Eurospeech99

[7] Ruiqiang Zhang, Ezra Black, Andrew Finch: "Using Detailed Linguistic Structure in Language Modelling", pp.1815-1818, Eurospeech99

[8] Arnaud Gaudinat, Jean-Philippe Goldman, Eric Wehrli: "Syntax-Based Speech Recognition: How a Syntactic Parser Can Help a Recognition System", pp.1587-1590, Eurospeech99

[9] Joan-Andreu Sanchez, Jose-Miguel Benedi: "Learning of Stochastic Context-Free Grammars by Means of Estimation Algorithms", pp.1799-1802, Eurospeech99

[10] Wolfgang Reichl: "Language Model Adaptation Using Minimum Discrimination Information", pp.1791-1794, Eurospeech99

[11] Akinori Ito, Masaki Kohda, Mari Ostendorf: "A New Metric for Stochastic Language Model Evaluation", pp.1591-1594 , Eurospeech99

[12] Don McAllaster, Larry Gillick: "Studies in Acoustic Training and Language Modeling Using Simulated Speech Data", pp.1787-1790, Eurospeech99

[13] Philip Clarkson, Tony Robinson, "Towards Improved Language Model Evaluation Measures", pp.1927-1930, Eurospeech99

[14] Christel BEAUJARD, Michele JARDINO: "Language Modeling Based on Automatic Word Concatenations", pp.1563-1566 , Eurospeech99

[15] Hong-Kwang Jeff Kuo, Wolfgang Reichl: "Phrase-Based Language Models for Speech Recognition", pp.1595-1598 , Eurospeech99

[16] D. Langlois, K. Smadli: "A New Based Distance Language Model for a Dictation Machine: Application to MAUD", pp.1779-1782, Eurospeech99

[17] Jerome R. Bellegarda: "Context Scope Selection in Multi-Span Statistical Language Modeling", pp.2163-2166, Eurospeech99

[18] Richard C. Rose, Giuseppe Riccardi: "Automatic Speech Recognition Using Acoustic Confidence Conditioned Language Models", pp.303-306, Eurospeech99

[19] Andrej Ljolje, Michael D. Riley, Donald M. Hindle: "The AT&T Large Vocabulary Conversational Speech Recognition System" pp.807-810, Eurospeech99

[20] Lidia Mangu, Eric Brill, Andreas Stolcke: "Finding Consensus Among Words: Lattice-Based Word Error Minimization", pp.495-498 , Eurospeech99

[21] Vaibhava Goel, William Byrne: "Task Dependent Loss Functions in Speech Recognition: A* Search over Recognition Lattices", pp.1243-1246, Eurospeech99

[22] Yoshiharu Abe, Hiroyasu Itsui, Yuzo Maruta, Kunio Nakajima: "A Two-Stage Speech Recognition Method with an Error Correction Model", pp.443-446 , Eurospeech99

[23] Paolo Coletti, Marcello Federico: "A Two-Stage Speech Recognition Method for Information Retrieval Applications", pp.459-462 , Eurospeech99

[24] H.Heuvel,et al, "The Speechdat-Car Multilingual Speech Databases for In-Car Applications: Some First Validation Results", pp. 2279-2282, Eurospeech99

[25] Os,Els Den. and Boves,Lou. and Lamel,Lori. and Baggia,Paolo., "Overview of the ARISE Project", pp.1527-1530, Eurospeech99

[26] Chung,Grace and Seneff,Stephanie. and Hetherington, Lee., "Towards Multi-Domain Speech Understanding Using a Two-Stage Recognizer", pp.2655-2658, Eurospeech99

[27] Jose,Reano Gil. et.al, "Flexible Mixed-Initiative Dialogue for Telephone Services", pp.1179-1182, Eurospeech99

[28] Lavelle,C.Alexia. and Calmes,Martine De. and Perennou,Guy, "Confirmation Strategies to Improve Correction Rates in a Telephonic Inquiry Dialogue System", pp.1399-1402

[29] Chu-Carroll,Jeniffer., "Form-Based Resoning for Mixed-Initiative Dialogue Management in Inforamtion-Quiry Sysytems", pp.1519-1522, Eurospeech99

[30] B.Vromans et.al., "Extending the SUSI Sysytem with Negative Knowledge", pp.2667-2670, Eurospeech99

[31] Grisvard, Olivier. and Gaiffe, Bertkland., "An Event-Based Dialogue Model and its Implementation in MultiDial2", pp.1155-1158, Eurospeech99

[32] Pargellis,Andrew. and Kuo,Jeff. and Lee, Chin-Hui., "Automatic Dialogue Generator Creates User Defined Applications", pp.1175-1178, Eurospeech99

[33] Ute Ehrlich, "Task Hierarchies Representing Sub-Dialogs in Speech Dialog Systems", pp.1375–1378, Eurospeech99

[34] Papineni,K.A. and Roukos,S. and Ward,R.T, "Free-Flow Dialog Management Using Forms", pp.1411-1414, Eurospeech99

[35] Rudnicky,A.I. et.al, "Creating Natural Dialogs in the Carnegie Mellon Communicator System", pp.1531-1534, Eurospeech99

[36] Rosset,Sophie. and Bennacef,Samir. and Lamel,Lori., "Design Strategies for Spoken Lnaguage Dialog Systems", pp.1535-1538, Eurospeech99

[37] E.Lewis, M.Tatham, "Word and Syllable Concatenation in Text-to-Speech Synthesis", pp.615-618, Eurospeech99

[38] K.Stober, T.Portele, P.Wagner, W. Hess, "Synthesis by Word Concatenation", pp.619-622, Eurospeech99
M.Beutnagel, M.Mohri, M.Riley, "Rapid Unit Selection from a Large Speech Corpus for Concatenative Speech Synthesis", pp.607-610, Eurospeech99

[39] M.Balestri, A.Pacchiotti, S.Quazza, P.L.Salza, S.Sandri, "Chose the Best to Modify the Least:A New Generation Concatenative Synthesis System", pp.2291-2294, Eurospeech99

[40] F.Chou, L.Lee, "Selection of Waveform Units for Corpus-Based Mandarin Speech Synthesis Based on Dicision Trees and Prosodic Modification Costs", pp.2295-2298, Eurospeech99

[41] D.House, L.Bell, K.Gustafson, L.Johansson, "Child-Directed Speech Synthesis: Evaluation of Prosodic Variation for an Educational Computer Program", pp.1843-1846, Eurospeech99

[42] O.Engwall, "Modeling of the Vocal Tract in Three Dimensions", pp.113-116, Eurospeech99

[43] T.Kaburagi, M.Honda,T.Okadome, "A Trajectory Formation Model of Articulatory Movements Using a Multidimensional Phonemic Task" pp.121-124, Eurospeech99

[44] N.Miki, T.Yokoyama, T.Ohtani, S.Masaki, I.Shimada, I.Fujimoto, Y.Nakamura, "A Vocal Tract Model Using Multi-line Equivalent Circuits" pp.129-132, Eurospeech99

[45] M.Matsuda, H.Kasuya, "Acoustic Nature of the Whisper", pp.133-136, Eurospeech99

[46] S.Ouni, Y.Laprie, "Design of Hypercube Codebooks for the Acoustic-to-Articulatory Inversion Respecting the Non-linearities of the Articulatory-to-Acoustic Mapping" pp.141-144, Eurospeech99

[47] C.Silva, S.Chennoukh, I.Trancoso, "On Improving the Decision Algorithm for Articulatory Codebook Search", pp.153-156, Eurospeech99