論文 / 著書情報
Article / Book Information

論題(和文)	音声認識の動向 ［I］ ―話し言葉音声認識―
Title(English)	
著者(和文)	古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	電子情報通信学会誌, Vol. 89, No. 8, pp. 746-751
Citation(English)	, Vol. 89, No. 8, pp. 746-751
発行日 / Pub. date	2006, 8
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2006 Institute of Electronics, Information and Communication Engineers.

解説

音声認識の動向［I］
――話し言葉音声認識――

Trends in Speech Recognition［I］: Spontaneous Speech Recognition

古井貞熙

Abstract

音声認識技術は近年大きな進歩を遂げているが，自然な話し言葉に対する性能は，まだ極めて不十分であり，現在の音声認識の最重要課題の一つになっている．音声認識タスクは，発声相手がコンピュータか人か，対話か独話かによって，四つのカテゴリーに分けられる．そのうちの二つのカテゴリー，すなわち人を相手とする独話と，コンピュータが相手の対話の研究が，当面最も重要となると考えられる．話し言葉の音声認識では，音声を単語列に変換するよりも，その内容や意味を抽出する音声理解のアプローチが重要となる．本稿では，「話し言葉工学」プロジェクトで得られた音声認識及び音声要約に関する成果を紹介し，今後の話し言葉音声認識の課題について述べる．

キーワード：話し言葉，音声認識，音声理解，話し言葉工学プロジェクト，音声要約

1. はじめに

音声認識技術は最近の10ないし20年間に大きな進歩を遂げている[1]．その進歩の概要は，認識対象語い数と，対象とする音声の発話スタイルからなる二次元平面を用いて，図1のようにまとめることができる[2]．この図で，左下が最も易しい領域，右上が最も難しい領域に対応し，技術は概して左下から右上へ向かって進歩している．3本の太い線は，それぞれ，1980年，1990年，及び2000年の技術レベルを示す．NTTが16単語を認識する孤立単語音声認識システムを用いたバンキングシステムを実用化したのは1981年で，当時としては，それが最高レベルの技術を用いた世界最大の音声認識システムであったことを顧みると，その後約20年間の音声認識技術の進歩がいかに大きいかが実感できる．

その進歩を実現した技術には，ケプストラムとその動的特徴の利用，HMM（隠れマルコフモデル）や統計的言語モデルに代表される統計的手法の確立，動的計画法をベースとする探索法の確立，大規模音声・テキストデータベース（コーパス）の利用，コンピュータやハードウェア技術の進歩，デコーダ（音声認識エンジン）などのソフトウェア技術の進歩などがある．しかもそれらは基本的に，英語，ドイツ語，フランス語，イタリア語，スペイン語，日本語など，どのような言語にも適用できる．

しかし，図1から分かるように，進歩を示す3本の線は，互いに並行ではない．すなわち，対象の発話スタイルによって，進歩の速度が大きく異なる．孤立単語，読み上げ音声などの認識は，数千語あるいは数万語の大語いを対象としても，既にほぼ実用レベルに達しているが，自然な話し言葉（spontaneous speech）に対しては，残念ながらまだ限られた性能しか得られていない．この課題が解決されないと，音声認識の用途が大きく拡大することは期待できない．

目　　次

古井貞熙　正員：フェロー　東京工業大学大学院情報理工学研究科計算工学専攻
E-mail furui@cs.titech.ac.jp
Sadaoki FURUI, fellow (Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo, 152-8552 Japan).
電子情報通信学会誌　Vol.89 No.8 pp.746-751 2006年8月

2. 音声認識タスクの分類

2.1 四つのカテゴリー

音声認識のタスクは，表1に示すように，認識対象に

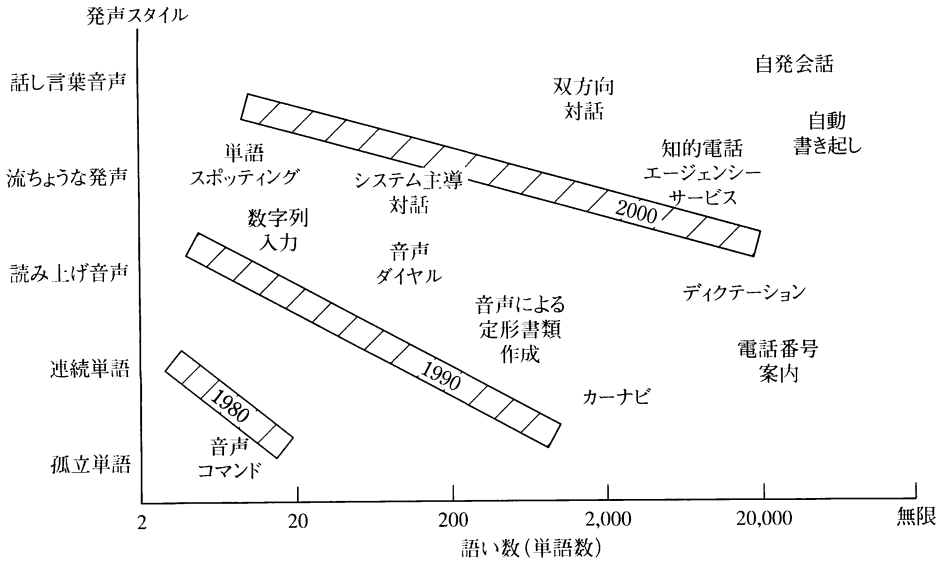

図1　音声認識技術の進歩

表1　音声認識タスクの分類

	対　話	独　話
人対人	(カテゴリーⅠ) 会議，討論， 電話対話 (Switchboard， Call Home (Hub 5))	(カテゴリーⅡ) 放送ニュース (Hub 4)， 講演，講義， ボイスメール
人対計算機	(カテゴリーⅢ) 情報検索，予約 (ATIS， Communicator)	(カテゴリーⅣ) ディクテーション

関する二つの規準，すなわち，

① 人がコンピュータに対して発声している音声か，人に対して発声している音声か
② 対話 (dialogue) か独話 (monologue) か

によって，四つのカテゴリーに分類することができる[3]．

カテゴリーⅠは，人と人との対話を対象とするもので，会議議事録の自動作成や，その検索などを目的とした研究が行われている．代表的な音声認識コーパス（データベース）に，電話の対話音声を対象としたDARPA (Defense Advanced Research Projects Agency) の Switchboard，Call Home (Hub 5) などがある．

カテゴリーⅡは，人が人に聞いてもらうことを前提に一方向的に話している独話を対象とするもので，ニュース音声の自動字幕化，講演録や講義録の自動作成，音声メールの文字化などの研究が行われている．これらは，至る所に存在する音声ドキュメントのアーカイブ化（コンテンツ化）の技術として，今後重要度が増してくると考えられる．DARPA のニュース音声を対象とした Hub4 タスクは，このカテゴリーの音声認識と，情報検索などの関連技術の発展に大いに貢献した．NHK のニュースの字幕作成システムは，世界で唯一の音声認識を用いたシステムであるが，アナウンサーが原稿を読んでいる音声や，きちんと発声している音声に対してのみ適用可能で，現場からの中継など，話し言葉に近い音声にはまだ適用できない．ニュースの字幕を含め，音声ドキュメントの認識では，音声を全部そのまま文字にするのではなく，むしろ要約や，メタデータ化することが必要となることが多い．

カテゴリーⅢは，人とコンピュータシステムとの対話を対象とするもので，情報検索，予約などを行う実用システムが，米国を中心に既に多数利用されている．研究としては，DARPA の ATIS (Air Travel Information System)，Communicator の両タスクが有名である．入力音声と望まれる動作との対応付け（意味理解）に関しては，種々の方法が研究されている．

カテゴリーⅣは，人がコンピュータに独話で話している音声を対象とするもので，ディクテーションがこれにあたる．この場合も，音声認識誤りを避けることはできないので，その修正を含むシステムは，通常，キーボード入力を含む対話形システムの構成をとることになる．

2.2　カテゴリー間の比較

カテゴリーⅢでは，システムへの入力手段として音声が用いられるのに対し，他のカテゴリーⅠ，Ⅱ，Ⅳでは，音声そのものがドキュメント，言い換えると情報コンテンツとして扱われる．人がコンピュータを意識して発声しているカテゴリーⅢ及びⅣの音声や，原稿を読み上げた音声は，人が人と話しているカテゴリーⅠ及びⅡの音声と，音響的にも言語的にも大きく異なることが分かっている．音声の自発性 (spontaneity) は，カテゴリーⅠが最も高く，カテゴリーⅣが最も低いと考えられる．カ

テゴリーⅠは極めて困難で，実用化までの道のりはかなり遠い．カテゴリーⅣのニーズは，当初考えられたほど大きくない．これからは，カテゴリーⅡとⅢを主たるターゲットとした研究が進められていくであろう．

3. 話し言葉の音声認識

3.1 話し言葉の音声認識はなぜ難しいか

読み上げ音声に比べて，話し言葉の音声認識性能が大きく低下するのは，話し言葉音声が音響的にも言語的にも読み上げ音声と大きく異なるためである[4]．話し言葉の音声認識を実現するためには，話し言葉の大規模コーパスを構築し，それを用いて音素モデルや言語モデルを作成することが必須であるが，話し言葉の大規模コーパスは極めて少ない．話し言葉を録音して，人手で書き起し，種々の情報を付加するのに，多大な労力と資金を要するためである．

話し言葉の変動には，言い直し，言いよどみ，繰返し，間投詞，不正確な発音など，種々の現象が含まれ，これらをモデル化するのは，極めて難しい．話し言葉の音声認識では，音声をすべてそのまま文字化するというよりも，不要な情報はむしろ除去して，読みやすい形に変換するのが望ましい．

3.2 話し言葉工学プロジェクト

1999年より約10億円の科学技術振興調整費を投入して，5年計画のプロジェクト「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」が推進された[5]．本プロジェクトでは，次の三つのサブテーマが進められた．

① 大規模話し言葉コーパスの構築
② 話し言葉を音声認識・理解・要約するためのツール及び基本技術の構築
③ 話し言葉の音声要約プロトタイプシステムの構築

3.3 話し言葉コーパス

本プロジェクトでは，2.1で述べたカテゴリーⅡにおいて，タスクを限定しない独話，特に講演（学会講演と模擬講演）を対象とした．延べ約650時間，単語（形態素）にして約700万語規模の話し言葉コーパス（CSJ：Corpus of Spontaneous Japanese）を作成した[6]．独話，特に講演を対象としたのは，

① 対話音声に関しては別の幾つかのプロジェクト（例えば文献（7））があったこと
② 聴衆に向けて一方向的に語る発話は比較的モデル化しやすく，認識率向上にもつながりやすい上，そのモデルは対話音声のモデル化の基礎になり得ること
③ 自動書き起しによるコンテンツや字幕作成のニーズが多いこと

などによる．

ただし，研究の幅を広げるため，コーパスの一部に，インタビューなども含めた．コーパス全体について，セグメンテーション，書き起し（正書法による基本形と発音形），形態素解析などを行ったが，全体の約8%（50万語）の「コア」については，人手による形態素解析結果，構文構造，要約のほか，韻律情報などパラ言語情報まで含めたコーパスとした．コアの形態素解析結果を用いて，コーパス全体を自動的に形態素解析するためのツール（解析プログラム）の学習を行った．

3.4 話し言葉音声認識の試み

多数の話者による講演音声について，音声認識実験を行った[8]．音響モデルや言語モデルの学習には，テストセット以外のCSJの音声と，その書き起しを用いた．その結果，次のような知見が得られた．

① 読み上げ音声から作成した音素モデルと，Web上の講演録から作成した言語モデルを用いた場合に比べて，CSJから作成したモデルを用いることにより，音声認識誤りをほぼ半分にできた．
② 発声速度，言い直し頻度，未知語率，及びパープレキシティが，話者による認識率変動の主要な要因となっている．すなわち，早口の音声，言い直しの多い音声，及び音声認識用辞書に登録されていない語いを多く発声している音声の誤認識が大きい．
③ 単語の連鎖確率に基づく言語モデルを用いているため，1%の言い直し頻度と未知語率の増加は，それぞれ，3.3%と2.3%の単語正解精度の低下をもたらす．
④ フィラー（「えー」「あー」などの言葉）の回数は個人差が大きいが，フィラーを含む言語モデルを用いれば，フィラー自体の認識は特に難しくはない．
⑤ 話し言葉では，文の区切りが必ずしも明確でないため，文単位の音声を自動的に切り出して認識するのは極めて難しい．適切なポーズで区切れば大きな性能低下はないが，文単位に区切らずに連続してデコーディング（音声認識処理）するのが有効で，認識結果から，文の区切りが推定できる．
⑥ 話し言葉の音響的・言語的変動は極めて多岐にわたるため，CSJのほとんどすべてのコーパスをモデルの学習に用いることによって，初めてモデルの精度が飽和する．
⑦ 入力音声の話題（分野）が学習に用いられたコーパスでカバーされていないときには，その分野の教

科書などを用いて，言語モデルを適応化するのが有効である．

⑧ 声の個人差，話し方の癖，話題の変化などに対処するため，音素モデルと言語モデルの教師なし話者適応を行うのが有効である．

⑨ CSJを学習に用い，更に話者適応を行うことにより，講演音声に対して，ほぼ80％の単語正解精度を得ることができる．

4. 話し言葉の音声要約

4.1 音声要約の目的

発声内容をキーワード，要約文，その合成音声などで提示する自動要約技術について，研究を進めた[5]．音声要約技術には，音声ドキュメントの検索や理解の容易化，字幕の高度化など種々の応用があるが，要約結果から話題を特定し，その情報を音声認識部にフィードバックすることによって，音声認識理解の精度を向上させることもできると期待される．音声認識結果にはどうしてもある程度の認識誤りや不確実性が避けられないので，音声要約には，書き言葉の要約とは異なる技術の構築が必要となる．

4.2 音声要約の方法

図2に示すように，重要文抽出と文圧縮の組合せによって音声要約を行う方法の検討を進めた[9],[10]．まず，音声認識結果を用いて，自動的に文の区切りを推定する．情報量の多い重要な単語（形態素）を多く含み，かつ音声認識誤りの少ない文を選ぶため，各文について，下記のスコアの重み付き線形和を計算し，その値が大きい文（重要文）を抽出する：

- 重要度スコア：tf-idf尺度を変形した情報量尺度（各単語が持つ情報量）
- 言語スコア：単語トライグラムによる言語ゆう度（言語的正しさ）
- 信頼度スコア：各単語の事後確率（音声認識結果に対する音響的・言語的信頼度）

選択された重要文について，更に部分単語列を選択し，最終的な要約文を作成する．部分単語列の選択は，各単語についての上記のスコア，及び

- 単語間遷移スコア：係り受けSDCFG（Stochastic Dependency Context Free Grammar）によって表される各単語の係り受け関係（構造）に基づく，単語間の接続可能性

の重み付き線形和が最大となるように行う．単語間遷移スコアは，部分単語列を，文法的かつ意味的に正しい連鎖になるように選ぶことを目的として導入した．

与えられた要約率（圧縮率）に対して，重要文抽出によって圧縮する割合と，文圧縮によって圧縮する割合を決め，それに従ってスコアを最大化する単語セットを，動的計画法を用いて選択し，要約を作成する．

これまでに，この要約手法を日本語及び英語の，ニュース音声と講演音声に適用し，10～70％に要約する条件で，評価を行ってきた．評価法として，多数の被験者が単語抽出によって作成した要約結果をまとめた単語ネットワークを用いる方法を提案した．このネットワークは，被験者による正解要約文の変動を網羅していると考えられる．このネットワークから，評価対象の要約文に最も近い単語列を抽出し，一致度を，音声認識における正解

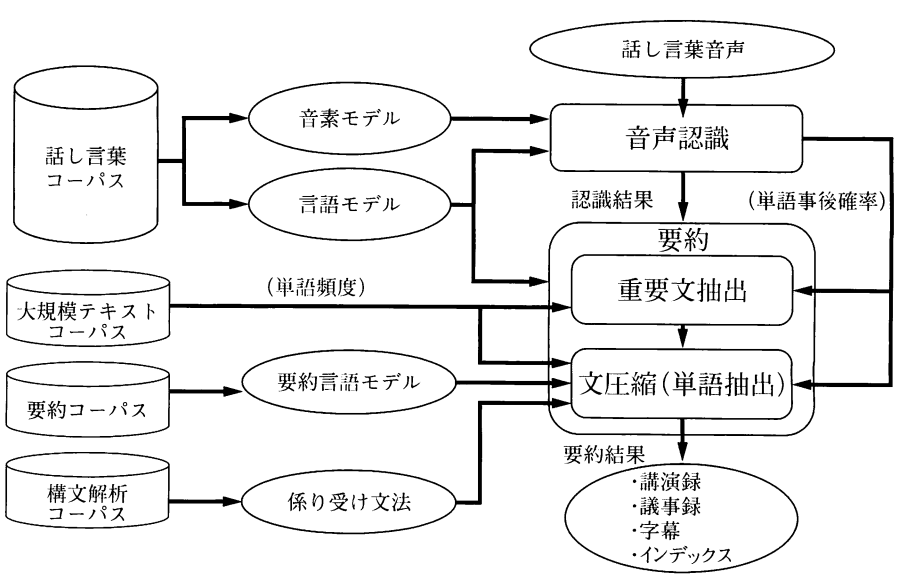

図2 重要文抽出と文圧縮の組合せによる音声要約

精度を測る場合と同じ方法で測る．その結果，上で述べた要約手法が有効であることが確認された．

4.3 音声から音声への要約

音声の要約結果をテキストで提示する場合は，音声認識誤りが避けられないため，読みづらくなるだけでなく，誤った情報をユーザに伝える可能性がある．また，韻律によって伝えられる話者の意図や感情情報を伝えることが難しい．元の音声をそのまま用いて提示すれば，それらの問題を回避することができる．

そこで，上で述べた音声要約手法を用いて，要約結果を音声で提示する「速聞きシステム」について研究を行った[5]．要約音声は，テキストで出力された音声要約結果に対応する音声を，元の音声から切り出し接続して作成する．切り出した音声同士をそのまま接続したのでは，音声の振幅や波形の差から，雑音が生ずる可能性があるので，音声の両端の振幅を徐々に減衰させ，前後の言語的関係に依存した長さを有するポーズ（無音区間）を挿入してから接続する．

要約音声を作成するのにふさわしい切り出し単位を調べるため，単語単位，文単位，文を更にフィラーで区切った単位，の三つの単位の選択により作成した要約音声について，「聞きやすさ」と「要約音声としてのふさわしさ」の2項目で，聴取実験による評価を行った．講演音声を用いた実験の結果，文単位選択による要約が最も有効であることが確認された．

5. 今後の課題

話し言葉音声認識・理解の精度を上げるためには，基本的な音響モデル，言語モデルに関して，まだ研究の余地が極めて大きい．認識精度を低下させている種々の要因について，統計的かつ系統的に解析し，問題点を明らかにするとともに，新たな特徴量の開発について研究する必要がある．重要な課題として，イントネーション，アクセントなどの韻律情報の利用法，形態素解析ツール，間投詞・言い直しなどのモデル化，音声要約のための構文及び意味表現法などがある．未知語への対処法も重要な課題である．これらの解決のためには，コーパスに基づいた地道な分析とともに，新しいアプローチの開拓が必須である．

"There's no data like more data" といわれるように，データ収集には限りがない．幾ら大量のコーパスを構築しても，将来の音声認識理解の対象となるすべてのタスク（応用場面，話題）をカバーすることはできない．したがって，話し言葉に含まれるタスクに独立な現象と，タスクに依存する現象を分離したモデルを構築し，タスクに依存する部分について，新しいタスクに自動的に適応する方法を作り上げることが不可欠である．音声の音響的・言語的個人差，マイクロホンの種類や距離，周囲雑音，電話などによるひずみなどに対してどう対処するかも，極めて重要な課題である．認識対象音声自体を用いて，音素モデルや特徴パラメータを教師なしオンライン適応する方法の研究に力を入れる必要がある．

人間が，どのような手掛かりを用いて，どのように音声認識しているのか，各情報はどの程度有用なのかなどを明らかにすることによって，音声認識アルゴリズムの指針を明らかにすることも，一見回り道のようではあるが，着実な進歩を実現するために必要であろう[1]．

話し言葉音声認識の研究は，単に音声を単語列に変換するだけでなく，発声者の意図の理解，話題語あるいはキーフレーズの抽出，内容の自動要約へと展開するであろう．これらの技術の進展によって，だれが何をいったかといった情報（メタデータ）を含め，いろいろな知識を即座に流通させ，活用することができるようになると期待される．更にそれらの知識を，Web などの多様な知識と関連付け，体系化することができるかどうかが，21 世紀の重要な研究課題となると予想される[11]．そのような体系化された知識が利用できるかどうかが，逆に，音声認識・理解技術の発展の鍵となるであろう．音声のみでなくマルチモーダル環境でのコンテンツ作成も，今後の重要な研究課題である．

コンピュータの小型化，高性能化，低価格化により，ユビキタス（遍在）計算とウェアラブル計算環境の時代がくるといわれている．このような環境下では，キーボード入力はなじまないので，音声認識がヒューマンインタフェースの基本的手段の一つとして広く使われるようになるであろう．その際，音声認識の多くは，個人が身に付け，個人に特化したマイクロホン及びコンピュータで行われるようになり，そのシステム構成は，これまでとは大きく異なった形になると予想される．雑音などの環境情報は常時モニタでき，タスクに依存した知識は個人のコンピュータへ即座にダウンロードできるので，タスクへの適応性とロバスト性の向上が期待できる．ただし，それらの実現のためには，解決しなければならない課題も多く，今後の多角的な研究の展開が必要である．

文　献

（1） S. Furui, "50 years of progress in speech and speaker recognition," Proc. Int. Conf. Speech and Computer (SPECOM 2005), pp.1-9, 2005.

（2） 古井貞熙，"話し言葉の音声認識・理解を目指して，" 信学技報，SP2002-47, pp.31-36, 2002.

（3） S. Furui, "Recent progress in corpus-based spontaneous speech recognition," IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.366-375, March 2005.

（4） S. Furui, "Why is the recognition of spontaneous speech so hard?" Proc. 8th Int. Conf. Text, Speech and Dialogue (TSD 2005), pp. 9-22, 2005.

（5） 古井貞熙，"話し言葉の音声認識と自動要約，" 第3回話し言葉の科学と工学ワークショップ講演予稿集，pp.1-6, 2004.

(6) 前川喜久雄, "『日本語話し言葉コーパス』公開版の仕様," 第3回話し言葉の科学と工学ワークショップ講演予稿集, pp.7-14, 2004.
(7) 日本学術振興会未来開拓学術研究推進事業研究プロジェクト「音声言語による人間－機械対話システムの研究」成果報告会資料集, 2001.
(8) 篠崎隆宏, 古井貞熙, "日本語話し言葉コーパスを用いた講演音声認識," 情処学論, vol.43, no.7, pp.2098-2107, 2002.
(9) 堀 智織, 古井貞熙, "単語抽出による音声要約文生成法とその評価," 信学論(D-II), vol. J85-D-II, no.2, pp.200-209, Feb. 2002.
(10) C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel, "Automatic summarization of English broadcast news speech," Proc. Human Language Technology 2002 Workshop, San Diego, pp.228-233, 2002.
(11) 古井貞熙, "21世紀COEプログラム「大規模知識資源の体系化と活用基盤構築」," 信学技報, SP2003-162, pp.47-52, 2003.

(平成17年11月10日受付 平成18年2月6日最終受付)

古井 貞熙(ふるい さだおき)(正員：フェロー)
昭43東大・工・計数卒. 昭45同大学院修士課程了, 日本電信電話公社(現NTT)研究所入社. ベル研究所客員研究員, NTT基礎研究所第四研究室長, ヒューマンインタフェース研究所音声情報研究部長, 古井特別研究室長を経て, 現在, 東工大大学院情報理工学研究科計算工学専攻教授. 工博.

Asia and South Pacific Design Automation Conference 2006 (ASP-DAC 2006, アジア・南太平洋設計自動化会議2006)

主 催：IEEE Circuits and Systems Society, ACM/SIGDA, IEICE ESS, and IPSJ SIGSLDM
日 時：2006年1月24～27日(4日間)
会 場：Pacifico Yokohama(横浜)
参加者：約760名
主要参加国：日本, アメリカ, 韓国, 台湾, 中国, インド, フランス, スペイン, イギリス, イラン, ほか
セッション数及び論文数：36セッション, 135件(一般論文), 19件(デザインコンテスト), 14件(招待論文), 8件(デザイナーズフォーラム)
Proceedings発行所：IEEE
主たるトピックス：

ASP-DACは, VLSIと電子システムの設計技術に関するアジア・南太平洋地区最大の国際会議で, 1995年の第1回開催以来, 今回で11回目を迎えた. 米国のDAC (Design Automation Conference), ICCAD (International Conference on Computer-Aided Design), 欧州のDATE (Design Automation and Test in Europe) の姉妹会議として, 国際的にも認知度が高い.

全日, 半日合わせて6件の有料チュートリアルが, 初日に行われた. キーノートスピーチでは, 次の3件があった.

1. Automotive Electronics: Steady Growth for Years to Come !
Alberto Sangiovanni-Vincentelli (The Edgar L. and Harold H. Buttner Chair of Electrical Engineering and Computer Science, University of California, Berkeley and Chief Technology Advisor, Member of the Board and Co-founder, Cadence Design Systems)
2. Challenging Device Innovation
Satoru Ito (President & CEO, RENESAS Technology Corp.)
3. Effective Platform-based Development for Large-Scale Systems Design
Yukichi Niwa (Senior Advisory Director, Group Executive of Platform Technology Development Headquarters, CANON INC.)

セッションは, 一般講演／特別講演セッションが31, ユニバーシティデザインコンテストが1, デザイナーズフォーラムセッションが4で, 合計36セッションで構成され, 上流設計から物理設計, 設計事例など幅広い分野にわたる講演が行われた. これらのセッションには世界各国から750名を超える研究者が出席し, 活発な議論と討論が行われた. 特に, 今回新たに「デザイナーズフォーラム」と呼ばれる四つのセッションが開催され, 産業界の視点から招待講演として, Low Power Design (セッション5D), "Cell" Processor (セッション8D), 並びに, パネル討論としてFunctional Verification——now and future—— (セッション6D), Top 10 Design Issues by LSI Designers versus EDA Developers (セッション9D)があり, 非常に多くの聴講者を集めた. 次回は, 2007年1月に横浜にて開催される予定である.

(執筆者 戸川 望 正員 早稲田大学理工学部 コンピュータ・ネットワーク工学科)