LSSN 0915-5961

AVIRG 会報

Vol.34 No.3 (2000.11)
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

「東芝の不特定話者音声認識技術とその応用」
講演：正井 康之 氏（㈱東芝 研究開発センター）
報告：松井 淳 （NHK）

《概要と感想》

音声認識システムには，話者を特定して認識精度を高めるものと，話者を特定せずに個人差を許容するものとの２つのタイプがある．市販のＰＣ用音声認識ソフトウェアをお持ちの方であれば，前者はお馴染みかもしれない．初回の起動時に話者登録用のサンプル文を読み上げる作業は，意外と骨が折れるものである．このような話者適応ゆえの操作に煩雑さを感じている利用者は，案外と多いのではなかろうか．

本講演では，利用者の使い勝手の向上をテーマとして研究・開発された，(株)東芝の不特定話者音声認識の応用事例と，それらの要素技術が，デモを交えてわかりやすく紹介された．

まず，不特定話者音声認識技術の応用事例がスライドで紹介された．

●電話音声認識応答装置（1982年）
電話音声による銀行口座の残高照会システム．都市銀行で運用実績有り．

●行き先階音声登録エレベーター（1989年）
ハンズフリー行き先階登録システム．
センサーによるマイクスイッチ制御や，キーワードスポッティングが特徴．

●音声認識券売機（1989年）
世界初の音声認識券売機．難波の駅でフィールドテスト．

●マルチメディアＡＴＭ（1992年）
音声認識，手書き文字認識，タッチパネルを併用したバリアフリーＡＴＭ．

●音声地図検索（1996年）
音声で住所や目標物を入力することで，検索時間を大幅に短縮．(語彙サイズ：約11万)

●東芝音声システムver.2.0（1996年）
音声認識，文規則合成合成を搭載した対話形エージェント．

●マルチメディア教育端末（1998年）
音声認識，手書き文字認識，タッチパネルを利用したクイズ形式の教育端末．

●ディクテーション（1999年）
東芝音声システムver.5.0．声の登録が不要なので，すぐに使用可能．

●音声認識ミドルウェア（1999年）
RISCチップ向けミドルウェア．雑音環境下での高性能音声認識を実現（例：カーナビ）．

続いて，これら応用事例の実現にあたり開発された幾つかの技術のうち，とくに代表的な次

の２つの技術について，詳しい説明があった．

1．複合音響特徴平面（ＭＡＦＰ）
2．音声セグメント

複合音響特徴平面（ＭＡＦＰ）とは，入力音声の分析パラメータについて，その時間-周波数平面上での２次元的な変化を，複数方向の空間微分オペレーターを用いて抽出する手法である．分析パラメータの時間方向の変化を特徴量に加える手法は，音声認識の分野では一般的になりつつあるが，ＭＡＦＰでは更に周波数軸を含めた計４方向の変化を特徴量に加えることで，より高精度な認識を狙っている．講演ではＭＡＦＰによる認識性能の改善効果を裏付ける実験結果が示された．

また，音声セグメントとは，連続音声中に出現する様々な音声現象（音素環境）を記述する音声学的単位の一種である．講演では，話者による声の違いを吸収させる目的で，この音声セグメントを中間表現として音声認識システムに実装した例が紹介された．

最後に，ノートＰＣによるディクテーションと音声対話エージェントのデモが披露された．正井氏本人による新聞記事の読み上げでは，ほぼリアルタイムで良好な認識結果が得られていた．マシンの処理能力（Pentium II，266MHz）を考えれば，かなりの性能である．話者を交代しての実験では若干精度が劣化したようであったが，利用者の操作の習熟度にも因るので，と正井氏から説明があった．また，電子メールを想定した比較的自由な発話スタイルについては，現時点では技術的に難しい問題を抱えたタスクであるが，非常に魅力的な今後の研究課題としての印象を受けた．

質疑応答では，本講演のテーマともいえる使い勝手に関するものから，技術的・専門的な内容まで，多岐にわたる質問が寄せられ，音声認識技術に対する参加者の関心の高さが窺われた．

「NHKニュース音声認識システム」
講演：今井 亨 氏（NHK）
報告：登内 洋次郎（㈱東芝 研究開発センター）

《概要と感想》

パターン認識の中でも音声認識は近年大幅な進歩を遂げており，実用化に向けた研究が益々盛んになってきている．中でも，大量の語彙を認識対象とするディクテーションプログラムが各社から商品化されている．ディクテーションプログラムを使えば，キーボードの代りにマイクに向かって音声で文章を入力することができる．音声認識は，人間にはまだ及ばないものの，ようやく実用化されつつある．

ＮＨＫでは毎晩７時のニュースの一部で，音声認識を利用した字幕放送を試行している．これは，ディクテーションプログラムと同じ音声認識技術を用いて実現されたものである．今回の講演で今井氏は，このニュース音声認識システムについて説明された．まず，ニュース音声システムの全体の概要を述べられた後に，それぞれの要素技術について具体的に解説された．

まず，ニュース音声システムの概要の説明があった．このシステムは,生放送のスタジオ・アナウンサーの音声をリアルタイムで音声認識して，誤りがあれば即座に人手で修正しつつ字幕を作成するものである．平成 12 年 3 月 27 日より，毎晩 7 時のニュースの一部で試行中であり，生放送としては日本初，音声認識を利用したものとしては世界初の字幕放送である． 字幕放送に音声認識を用いているのは，あらかじめ用意されたニュース原稿用に対してアナウンサーは修正を加えてから読み上げるため，ニュース原稿をそのまま字幕として表示することはできないからである．ちなみに，アメリカでは，速記用のキーボードが普及しているため，生放

送のニュースでは速記用キーボードを用いて字幕を作成しているということである．

次に，音声認識手法についての説明があった．ＮＨＫのニュース音声認識システムでは，ＨＭＭ（隠れマルコフモデル）という確率統計的な手法で認識を行う．音響パラメータとして，短時間スペクトル分析を用いてスペクトル包絡に関する 39 次元の特徴量を求める．音響モデルは，大量のアナウンサーの音声データから作成する．男性用，女性用それぞれの不特定話者のモデルを作成する．母音，子音の合計 42 音素ごとに連続HMMを用いてモデル化する．約２500 個のトライフォン（前後音素環境依存）を 3 状態の Left－To－Right 型の HMM で表す．状態共有化することで，約 1500 個の状態で表現する．HMM の出力には 8 混合のガウス分布を用いる．

言語モデルは，過去の NHK のニュース原稿 9 年分を用いて作成し，約 2 万単語のモデルを構築する．ただし，直前のニュース原稿から作成する言語モデルを重み付けして上記モデルに加えている．これは，日々変わっていくニュースの最新の話題に対応するために行うものである．言語モデルには，N 個の単語間の接続関係を確率的にモデル化した N-gram を用いる．最初に音響モデルとバイグラムで正解の候補をある程度絞り込み，最後にトライグラムで詳細な探索を行って認識結果を出力する．

通常，ディクテーションプログラムでは，1 文単位で音声認識を行うため，文末まで発声された後に結果が確定する．しかし，生放送で流れる音声に対してリアルタイムで字幕を送るのに，文末を待って音声認識の結果を確定するのでは遅くなる．そのため，逐次 2 パスデコーダという手法を用い，文の途中で探索をトレースバックして，安定した単語列があれば逐次確定していく．この処理で認識率は 0.2%程度悪くなるものの，単語確定遅れを 0.5 秒に押えることができる．

最後に、オペレータが音声認識の結果に対して確認を行い，間違いがあればその個所を修正する．放送中にリアルタイムで修正するために，2 クルー（各 2 名で，タッチパネルによる選択作業に 1 名，キーボード入力に 1 名）が一文ごとに交代で確認及び修正作業を行う．認識性能は，オペレータの修正前で 95%以上，修正後に 99.5 以上になる．認識時間はオペレータの作業を入れて 1〜2 秒の遅れになり，実際画面に字幕が放送されるのは，表示までのタイムラグがあるため，平均 10 秒の遅れとなる．

今回の講演では，実際に放送局で使われているシステムについて，ニュース原稿が出来上がってから，アナウンサーの音声を認識して，字幕を放送する所までを，実際の放送のビデオを交えて丁寧に説明していただいた．そのため，放送局の内部を全く知らない私でも，字幕放送ができあがるまでの過程がよくわかった。また，今井氏が音声認識方式の中身をわかりやすい解説してくださったおかげで，認識方式についてもよく理解することができた．今回の講演は，音声認識に携わる研究者はもちろん，それ以外の研究者にとっても，とても興味深い講演だったと思う．

音声認識を含めパターン認識の研究では，認識性能を上げることが重要である．しかし，現状では 100%の認識率を得るのは困難であり，実際の問題で利用するためには，用いる対象に即した形に認識システムをうまく当てはめることが非常に重要である．今井氏に発表して頂いたニュース音声認識システムは，その意味においても非常に価値があると思う．

Ⅱ．11月例会予定

11月の例会は，

日時：11月30日（木）14時～17時

場所：東京大学工学部 6号館 2F 63号講義室で開催します．

テーマは，『マルチメディア情報処理』です．講演は以下の2件を予定しております．奮ってご参加ください．

●「超高速映像検索と国際標準化：MPEG-7」

講演者：山田　昭雄 氏

（NEC）

映像データの高速高精度検索に関する技術について，関連する国際標準化動向の解説を交えて報告する．ISO/IEC-15938通称MPEG-7は，マルチメディアデータのコンテンツ検索を目標として1998年に標準化活動が開始され，現在標準化終結を直前にむかえている．

本報告ではNECから提案を行ない採用された高速映像検索＆ブラウジング技術を中心に，最新の映像検索技術についてそれぞれの技術内容，特徴,用途などについて述べる.また高速映像検索の応用例として，インターネット映像ポータルサイトでのアーカイブシステム，インターネット連動型高機能TVinPC，CM放送等のビデオコンテンツをリアルタイム識別するシステム等をとりあげ紹介する．

《参考文献》

[1] イメージからの超高速映像検索方式
画像ラボ2000年6月号 GA0004-05

[2] Visual Program Navigation System based on Spatial Distribution of Color Proc. on ICCE2000, May 2000.

[3] Multimedia-Content Filtering, Browsing, and Matching using MPEG-7 Compact Color Descriptors Proc. on Int. Conf. on Visual Information Systems (Visual2000), Nov 2000.

[4] NECテレビコマーシャル調査システム
http://www.labs.nec.co.jp/cmsearch/

●「分身生成のためのマルチモーダル表情合成」

講演者：森島　繁生 氏

（成蹊大学工学部）

音声，画像，センサー情報等を駆使して，人物の特に表情のコピーをリアルに実現する技術の最近の成果について，デモ映像を豊富に交えて報告する．

顔モデルのカスタマイズ方法の提案．口の動きに関しては，音声駆動の方法，テキスト駆動の方法について述べ，前者の応用としてのネット対話システム，インタラクティブ映画の実例，さらに後者の応用例としてビデオ翻訳システムについて紹介する．表情に関しては，表情筋モデルの提案と，これを利用して1台のカメラからの情報で３次元の表情をコピーする方法，ＥＭＧを利用する方法を紹介．さらに表情以外の髪型の表現方法とその運動制御方法についても述べる．

さらに基礎技術として，音声への感情付加，顔特徴点のトラッキング技術等についても述べ，最後に最近のプロジェクトについてその現状報告を行う．

《参考文献》

[1] 顔の認識・合成と新メディアの可能性
第6回画像センシングシンポジウム論文集，特別講演S-2，2000年6月．

[2] 顔の認識・合成のための標準ツール
システム/制御/情報,Vol.44,No.3,pp.119-126,2000年3月．

[3] 感情表現のリアリティを追求
日経CG，2000年特集，pp.120-121，2000年1月．

ドメイン名取得のお知らせ

6月の総会において承認されました2000年度活動計画（AVIRG会報Vol.34，No.1）に基づいて，会報の電子化等を目的とするAVIRGのホームページ開設に向けて準備を進めております．
すでにドメイン名「avirg.org」を取得し，10月末よりInterNICにも登録されております．近々のうちにホームページを開設する予定です.ご期待下さい．
なお，ホームページのアドレスは

http://www.avirg.org

です．また，会報発行や例会開催，その他の重要なお知らせを会員の皆様にお届けしたり，議論の場として使って頂けるメーリングリストも追って開設の予定です．その節は会報等でお知らせ致します．
また，AVIRG幹事へのご意見・お問い合わせは以下のアドレスにお寄せ下さい．

幹事連絡先：**kanji@avirg.org**

～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　**kanji@avirg.org** (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．