2.2.1 ニュース音声自動字幕化のための音声認識・音声信号処理

ニュース番組を対象としたアナウンサー音声からの自動字幕作成を目指して、我が国の大学や内外の研究機関との連携のもと、ニュース音声認識システムの研究開発を進めた。

研究を加速するため、「ニュース7」、「おはよう日本」など、NHKの代表的なニュース番組を対象としたニュース音声のデータベース化を進めており、1999年4月1日からは、毎日放送されている内容を継続してデータベース化している。また、「日曜討論」、「サンデースポーツ」などの番組音声の書き起こしを行い、言語データの蓄積に努めた。

音響モデル関連では、ニュース音声データベースを利用したモデルの作成方法を詳細に検討し、認識性能向上に努めた。また、不特定話者認識の特定話者化の第1段階として、アナウンサー音声をクラスタリングし、認識時には、対応するクラスター用の音響モデルを選択する方法を開発した。その結果、認識処理時間が1割から2割程度短縮されるという結果が得られた。

言語モデル関連では、放送直前に入稿される記者原稿の利用法の検討を進めた。また、直前原稿を利用しても登録できなかった新しい単語を手動で即座に登録し、ほかの単語との前後関係を自動推定する手法を開発した。そのほか、記者原稿中の単語出現位置を登録しておき、この出現順序に従った候補を重視して認識を行う方法を開発した。これらの言語モデル関係の改善により、4〜5%程度の認識率の向上を実現した。

デコーダー関連では、それまで文の終わりまでの情報を利用して認識結果を確定していたが、これを、逐次確定していく方式に変更した。この結果、認識結果確定に要する平均時間を、7.2秒から0.6秒に短縮した。

これらの成果を統合し、1998年9月30日放送の「ニュース7」、「正午のニュース」および「おはよう日本」の全国中継分の音声を認識したところ、スタジオアナウンサー部分（492文）に対して、平均値で認識率86%（このうち、メインアナウンサー部分の122文については認識率97%、スポーツ、天気などの213文については88%）を得た。認識時間は、メインアナウンサー部分について、認識遅れ時間3秒であった。特に、スタジオメインアナウンサーに対しては、目標性能である認識率95%以上を達成した。

一方、認識結果中の誤りを、人間が即座に発見し修正するシステムの開発を進めた。このシステムは、作業を、「誤りの発見」と「発見された誤りの修正」という2つに分けて修正する。修正作業の確実性を増すため、話速変換技術を応用し、誤り発見・修正者に、音を文字とほぼ同期させて呈示する方式も開発した。

図1　ニュース音声認識システム

スタジオメインアナウンサー部分に対する認識性能が、ほぼ実用化レベルに到達したため、2000年3月27日から夜7時のニュースについて、字幕放送を試行的に開始することとなった。この本番運用に備えて、音声認識システムの実用機を開発した。

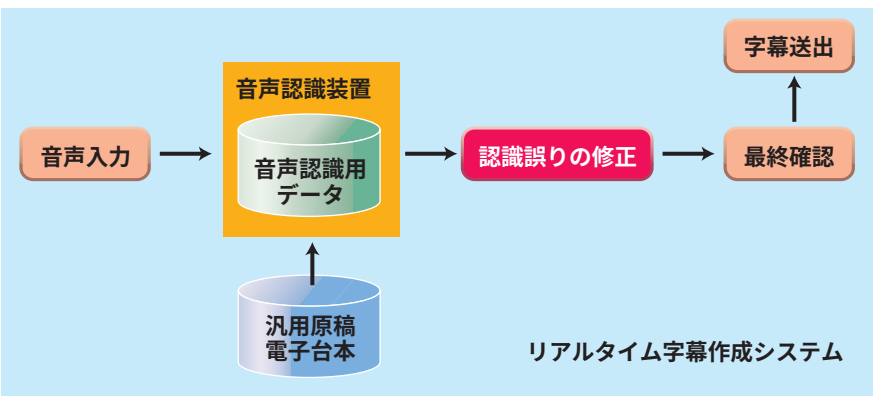


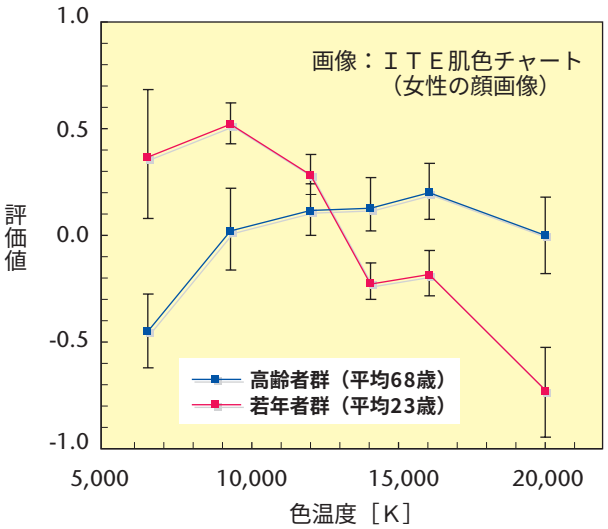
図2　ニュース字幕自動作成システム

2.2.2 人にやさしい情報呈示法

2眼式立体画像の観視時には、両眼の画像を融合するために生じる輻輳(ふくそう)眼球運動と焦点調節との間に食い違いが生じることが知られている。調節機能の衰えた高齢者での実態を明らかにするため、2眼式立体画像に対する調節応答と主観評価を行った。その結果、高齢者では、立体像の奥行き移動で誘導される調節応答量が若年者に比べて1/3以下で、立体像に反応しにくいことが明らかとなった。

加齢に伴う色覚特性の変化を考慮したカラーテレビの好ましい色再現条件を見いだすことを目的として、高齢者および若年者を対象とした好ましいカラーテレビの色温度条件を求めた。その結果、高齢者にとって好ましい色温度は若年者の9,300Kよりも高く、約16,000K付近であることがわかった（図）。また最も好ましい色温度条件での視覚心理効果を主成分分析法によって明らかにした。

また、ISDB見出しメニューなどGUI（Graphic User Interface）環境でのアクセシビリティー改善を目的に、点字とサイン音などの音声とを併用したマルチモーダルインターフェースの検討を進めた。さらにISDBメニュー画面などのGUI環境を、形状や表面の粗さなどの触覚と音声で呈示することを目的としたインターフェースの基礎的な検討を行った。



好ましい色温度の実験結果

2.2.3 次世代ヒューマンインターフェース

多様で数多くのサービスが想定されているBSデジタル放送を、誰もがやさしく利用できるデジタル受信機用リモコンの要件を探るため、各種リモコンの試作を行い、それに対応したテレビ受信機の模擬操作画面をコンピューターで生成して実際に操作したときの操作と表示の整合性を評価した。試作リモコンは、キーを極力減らしたボタン型、パソコンで広く用いられているポインティングデバイスの例としてトラックボール型、音声によって操作内容を指示する音声認識型である（図）。

試作したデジタルテレビ用リモコン

目の不自由な人や盲ろう者にデータ放送や電子データの文字情報を伝える受信端末の研究開発を進めている。文字放送やISDB情報へのアクセス性と操作性、および点字呈示方式について試用評価実験を実施し、結果を基にそれらの改善を進めた。また、将来の双方向サービスを視野に、盲ろう者を対象とした遠隔コミュニケーション機能を検討した。新たに開発した会話プロトコルにより遠隔での意思伝達が可能であることが実証された。また装着型の6指1文字点字端末を試作し、入出力の基本特性を確認した。

早口の聴き取りが苦手なお年寄りにも放送を楽しんでいただくために、音声を「ゆっくりした声」に変換する話速変換技術を開発してきた。1999年度は、話速変換技術をソフトウエア化し、パーソナルコンピューター上で音声データをリアルタイムに話速変換して出力できることを確認した。また放送現場での効率的な編集作業に寄与することを目的に、ノンリニア編集機の可変速再生において、映像に同期した音声を再生するシステムを試作し、5倍速程度までの音声の内容把握が可能な音声信号処理方法について検討した。

2.2.4 画像認識に基づく効率的映像検索

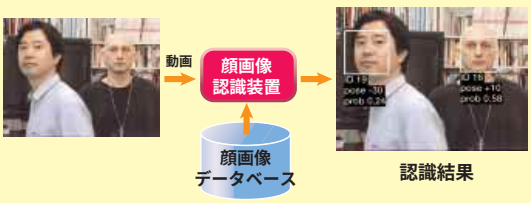
人物認識による映像への自動インデックス付けや編集支援などの応用を目指して、顔画像認識を中心に以下の研究を進めた。

映像中の人物の顔を検出、追跡、認識するプロトタイプシステム（図）を改良し、数十名規模の認識人数で実用可能なレベルの認識精度に向上させた。すなわち、顔の特徴量の求め方を改善することにより、顔の3次元的な動きによって画像中の顔が拡大・縮小、あるいは回転しても、より柔軟に対応できるシステムを実現した。また、データベースの自動登録を可能とするために、入力画像の顔の向き推定法を検討した。

人物認識の映像編集支援への利用法を検討するため、動画像クリップを人物認識した結果をデータベース化し、そのデータベースサーバーからウェブブラウザーを使って、人物や人物の構図をキーにして画像の検索が行える動画像検索システムを試作した。

人間の感覚に適合した柔軟な映像検索手法を実現するために、色彩情報、構図、背景情報などに着目した特徴抽出手法について検討した。色彩情報に関しては、（株）ATR人間情報通信研究所と共同で、柔軟な検索手法に対して最適な色彩統計量の抽出法の検討を行った。また構図および背景の複雑性を手がかりとする画像検索手法を提案し、種々の静止画像を用いた検索実験を行った。

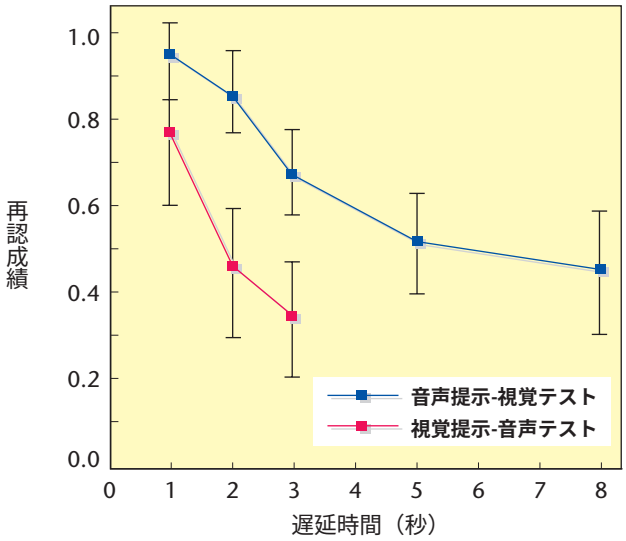
アニメなどの映像表現手法が視聴者の健康に影響を及ぼす問題に関して、引き続き、東京女子医大および岐阜大医学部などの外部研究機関への研究協力を行うとともに、研究動向に関する調査を行った。



人物認識システムの概要

2.2.5 わかりやすい映像表現手法

わかりやすく、使いやすいEPG（電子番組ガイド）の基本的な構成要素を明らかにするために、番組カテゴリーがどのような階層構造で構成されているかを検討した。検討に際しては、強制分類実験による階層的区分けをトップダウン的な手法として課し、また、提示した番組名の類似度の評定実験をボトムアップ的な手法として行った。なお、被験者は20才台24名、番組数はNHKでの295番組名である。番組名の強制分類から構成された番組カテゴリー化に伴うジャンル数は平均値で7.8である。また、被験者の80%がジャンル数を10とした。一方、番組名の類似度の評定結果をクラスター分析により解析した。これらの結果から、EPGの最上位ジャンルは10前後が適しているとの結論を得た。



文字・音声情報の提示時間差が再認に及ぼす影響

また、テレビでの情報提示手段である映像、文字、音声情報のうち、特に文字、音声情報の2つの情報を用いた場合に相互の提示時間差が表示内容の理解度にどのような影響を与えるかを、短期記憶のふるまいを指標として検討を行った。実験では単語を、音声、文字あるいは、文字、音声の順序で表示し、前者の表示に対して後者の表示の際に、単語が同じか異なるかの識別をさせた。結果は図のように音声・文字の順序での表示の方が正しく識別される率が著しく高く、かつ、時間差が2秒まででは80%を超える識別が可能であることが明らかとなった。これらの結果から「文字」映像、「音声」情報を利用した情報提示、インターフェースなどでの提示時間差に関してのわかりやすい表示方法を得た。