日常的使用を目指した音声認識検索システム

文字入力なしで簡単にテレビの番組を検索

文字入力に適したキーボードなどの入力手段を持たないテレビ（TV）や家電製品などにおいても，扱う情報量の増加とともに情報検索の必要性が高まっています。

東芝は，これらの機器でも手軽にコンテンツや情報を検索できるマルチモーダル音声認識検索システムを開発しました。加速度センサと距離センサを用いた自然な発話動作を認識する技術で音声認識精度を向上させ，愛称や略称などの言いかえ表現を的確に推定する自然言語処理技術によって日常使用する単語で音声を認識できます。

従来の番組検索方法に比べ所要時間を約40％短縮でき，高齢者など機器操作が不得意なユーザーでも容易に使えることを実験で検証しました。

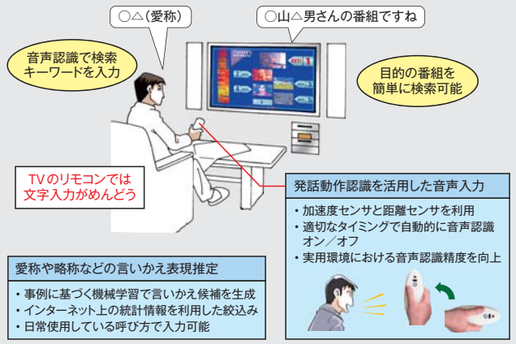

図1．音声認識を用いた番組検索システム ― TVのリモコンでは文字入力がめんどうですが，発話動作認識を活用した音声入力と，愛称や略称などの言いかえ表現の推定により，文字入力なしで簡単に番組を検索できます。

図2．試作したセンサ内蔵リモコン ― 加速度センサと距離センサを内蔵し，ユーザーの自然な発話動作を認識して，適切なタイミングで自動的に音声認識のオンとオフを切り替えます。

音声認識システムの現状と課題

パソコン（PC）や携帯電話では，キーワードをキーボードで入力し，関連するWebサイトなどを検索するテキスト中心の情報検索が広く利用されています。その一方，文字入力に適した手段を持たないTVや家電製品などでも，情報検索の必要性が高まっています。

従来，TVでは画面上のスクリーンキーボードから十字ボタンでカーソルを移動して文字を選択して入力したり，キー操作で1文字ずつ入力する方式が用いられますが，入力が遅く，使い勝手も良くありません。この問題を解決するうえで音声認識による文字入力が有効ですが，日常的に使うためには，音声認識エンジンの更なる性能向上に加えて，次に示す課題があります。

●入力方法

音声認識を日常的に使う場合，常時ヘッドセットを装着する入力方法は現実的ではありません。TV視聴のように手で操作できる日常生活シーンでは，めんどうな文字入力をリモコン搭載のマイクロホンからの音声入力で行い，ボタン操作と適切に組み合わせるマルチモーダルな使い方が効果的です。

また，入力する意図のある発話とそうでない発話を区別するために，ユーザーがボタン操作で音声入力の開始と終了を指示する方法が広く用いられています。しかし，ボタンの押し忘れや適切なタイミングで操作できずに正しく音声認識できないなどの問題があります。

●音声認識語彙

TV番組では，配信されている電子番組表（EPG）が検索対象です。EPGからは人や番組の正式名を取得でき，日々変化する番組に追従した音声認識語彙（ごい）を生成できます。しかし，人名の愛称や番組名の略称など言いかえ表現の情報は取得できません。このため，これらの語が発話されると音声認識辞書の未知語となり誤認識の原因となります。

東芝は，これらの課題の解決を図り，システム全体として使いやすさを大きく改善する技術を開発しました（**図1**）。

センサを利用した音声入力支援

手持ち型マイクロホンによる音声認識入力では，ヘッドセットを用いた場合より，マイクロホンと口元の距離が遠いことから音声認識率が低くなります。

また，音声認識区間を指示する方法としては，音声認識入力中に操作ボタンを押し続けるプレストーク方式や，音声認識の開始だけボタンで指示しシステム側が無音区間を検出して自動的に終了するプッシュトーク方式などが使われていますが，ボタン操作を忘れたまま発話してしまう例が多く見受けられます。

音声認識の精度と使い勝手を向上するには，口元とマイクロホンの距離が適正に保持されたときに音声認識を開始する支援が必要です。これらの支援を実現するため，マイクロホンを構える動作をとらえる加速度センサと，口元への近接を検出する距離センサを内蔵したリモコンを開発しました（**図2**）。

リモコンがユーザーの手に持たれたことを加速度センサで検知し，かつ距離センサでマイクロホンと口元が設定距離（約10 cm）以内であることを検出したときに音声認識を開始します。この方式について，高齢者も含む21名の被験者に同一内容の発話をしてもらい，ボタン操作が必要な従来方式と操作エラー率(注1)及び音声認識率の比較評価を実施しました（**表1**）。プレストーク方式は音声認識率が悪く，プッシュトーク方式は操作エラーが多いですが，センサ駆動方式は，プッシュトーク方式より操作エラー率が低く，かつ音声認識率がもっとも良好な結果となりました。特に，高齢者など機器操作が不得意なユーザーでも，習熟なしで音声認識入力を扱えることがわかりました。

表1．音声認識入力方式の比較評価

項　目		音声認識入力方式		
		プレストーク	プッシュトーク	センサ駆動
操作エラー率（%）	全被験者	1.9	8.1	4.8
	高齢者	2.9	13.8	5.0
音声認識率（%）	全被験者	75.5	81.9	82.4
	高齢者	62.1	71.3	77.3

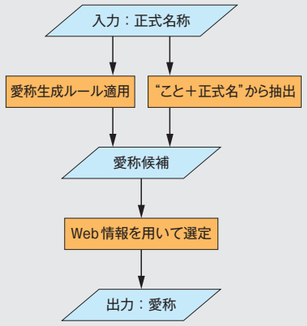

図3．人名の愛称推定の流れ ― 名前由来の愛称と名前由来でない愛称を別の方法で推定した愛称候補を，Web上の頻度を基に推定し，愛称推定結果とします。

Web情報を利用した愛称推定

人名には，“名前由来の愛称”と“名前由来でない愛称”があります。そこで，**図3**に示すように，各タイプの愛称を別の方法で推定し，それらを最後に組み合わせる手法を開発しました。

名前由来の愛称では，まず既知の愛称リスト（正式名と愛称のペア）から愛称生成のルールを自動的に作り出します。そして，新たに入力された正式名にそのルールを適用することで，名前由来の愛称を推定します。

また，名前由来でない愛称では，“こと+正式名”という表現パターンを利用してWeb上の表現から愛称を抽出します。更に，別々に推定した愛称候補をWeb上の頻度を基に選定し，愛称推定結果とします。

この手法による愛称推定のカバー率，つまり出現頻度を考慮した再現率は81.5％で，正式名称だけを用いた場合から20％近く改善されていることが確認できました。

（注1）　**プレストーク方式及びプッシュトーク方式ではボタン押し忘れ頻度，センサ駆動方式では発話動作検出漏れと誤検出の頻度。**

テレビ番組検索による実証実験

センサによる音声入力支援と，Web情報を利用した愛称推定によりTV番組を検索するマルチモーダル音声認識検索システムを開発しました。音声認識の語彙数は，正式名と愛称を合わせ約7,000語です。

このシステムを用いて，番組検索のキーワードを入力する実証実験を行いました。番組検索タスクに対する20代～70代までの34名の被験者の平均所要時間は，従来のスクリーンキーボードとボタン操作による方法に対し，約40％短縮できることがわかりました。特に高齢者では，所要時間は約50％で，キーボードとマウス操作によるPC環境を用いた方法に比べても約30％短縮できました。

また，発話に対する音声認識語彙のカバー率は93.7％でした。愛称推定を行っていない場合のカバー率は85.1％であり，愛称推定によりカバー率を約9ポイント向上できていることも確認できました。

今後の展望

音声認識システムを日常生活で使うためには，音声認識エンジンそのものの性能向上に加え，入力方式や音声認識語彙についても目標仕様に対応した検討が必要です。

実用化に向けて，各機能の更なる性能向上及び実環境での評価を継続していきます。

大内　一成
研究開発センター
ヒューマンセントリックラボラトリー
研究主務