5 日常的使用を目指した音声入力インタフェース

大内一成　若木裕美　（(株)東芝）

音声入力インタフェースの現状と課題

パソコンや携帯電話では，検索キーワードをキーボードで入力し，関連するWebサイト等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る．その一方で，文字入力に適した入力手段を持たない家電機器などにおいても，扱う情報の増加に伴って情報検索の必要性が増してきている．

たとえば，テレビは，地上波デジタル放送への全面移行，インターネット接続機能や録画機能の普及などに伴い，従来の映像コンテンツ表示装置としての役割だけでなく，インターネットの検索も行うリビングでの情報端末としての役割が増えていく．目的の情報／コンテンツをうまく探し出すための検索機能の重要性が増すほど，検索キーワードとしての文字を入力する機会が増える．しかし，従来のテレビ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する，あるいは携帯電話方式のキー操作で一文字ずつ入力するという方式では，入力速度が遅く，大変使い勝手が悪い．この使い勝手の悪さを改善するために，日常生活における音声認識による文字入力の重要性が今後増してくる．

しかし，日常生活で日常的に音声認識を使うためには，音声認識性能のさらなる向上はもちろん重要ではあるが，それ以外にも解決しなければならない課題がある．

◉入力方法

たとえばテレビで音声認識入力を日常的に使おうとする場合，その入力方法として，常時ヘッドセットを装着し続けるのは拘束性が強く，受け入れがたい．テレビ視聴のように，手による操作が可能な日常生活シーンにおいては，面倒な文字入力などはリモコンに搭載したマイクからの音声認識入力で行い，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ある．

また，現状は音声入力の意図を持った発話とそうでない発話を切り分けるために，ユーザが自ら発話時にボタン操作により音声入力のオン－オフを指示する方法が広く用いられている．しかし，この方法は，ボタンの押し忘れや適切なタイミングで押下できずに正しく音声認識できないなどの問題点がある．

◉音声認識語彙

音声入力インタフェースでは，音声認識入力として受け付け可能な語彙が多いほど，ユーザの多様な発話に対応できるが，その一方で語彙数の増加に伴って音声認識精度が低下する．そこで，日常的に音声認識入力を使うためには，認識率が低下しない程度の語彙数におさえる必要がある．しかし，語彙数が少ない場合，語彙として登録されていない単語(未知語)が発話されると認識できず誤認識が増える．さらに，音声入力では，テキスト入力に比べてより口語的な表現が使用されると想定され，さまざまな言い換え表現を含む多様な入力語でも認識できる必要が出てくる．このため，検索対象のコンテンツに基づいて適切な語彙を生成し，状況に応じて語彙を切り替えることが有効であると考えられる．

以上のように，現状の音声入力インタフェースでは，音声認識性能以外に，入力方式に関する課題と音声認識語彙に関する課題が存在する．我々はそれぞれの課題について，テレビの番組検索を題材に問題解決を図り，システム全体としての使いやすさを大きく改善できたので紹介する．

センサを利用した音声入力支援

テレビ視聴時に音声認識を利用する場合，上述の通り，マイクを搭載したリモコンによるマルチモーダルな操作が適している．そこで，手持ち型マイクによる音声認識入力の課題について整理する．

◉手持ち型マイクでの課題

・音声認識精度

一般家庭のリビングを模した実験室内において，ヘッドセットと手持ち型マイクで音声認識精度の比較を行ったところ，5名すべての被験者で，手持ち型マイクはヘ

ッドセットよりも音声認識精度が低かった．音声入力時のマイクと口元の距離がヘッドセットよりも手持ち型マイクの方が遠いことが原因と考えられる．

そのため，手持ち型マイクを被験者の口元から 5cm，10cm，15cm，20cm の距離に順に固定し，それぞれの距離で音声認識精度を比較した．その結果，10cm 以内の距離ではヘッドセットと同等レベルの音声認識精度が確認できた．しかし，10cm[☆1] を超えると距離に応じて精度が顕著に劣化していく．S/N 比の低下が原因である．このことから，一定レベルの音声認識精度を確保するためには，手持ち型マイクで音声入力を行う際に，ユーザがマイクに対して適切な距離で発話するための支援が必要であることが示唆される．

・音声認識開始／終了の指示操作

手持ち型マイクでの音声認識に限らず，従来の音声入力インタフェースでは，音声認識精度を高めるために，発話の開始と終了の切り出しが重要である．このため，ユーザが自らボタン操作などで音声認識の開始，あるいは終了を指示する方法が使われている．音声認識入力中に操作ボタンを押下し続ける方式（以下，プレストーク）や，音声認識の開始だけボタン押下で指示し，音声認識の終了はシステム側が無音区間検出により自動的に行う方式（以下，プッシュトーク）などがある．

プレストークやプッシュトークでは，ユーザが明示的に開始（プレストークでは終了も）を指定してくれるのが利点である．しかし，筆者らが過去に実施したプッシュトークによる音声認識を用いる実験では，始端でのボタン押下を忘れたまま発話してしまう事例が多く見受けられた．この傾向は，高齢の被験者など，機器の扱いに不慣れな被験者で特に顕著で，60 歳以上の被験者 6 名に対してその発生頻度を調査したところ 15.3% であった．これは操作の習熟により多少なりとも改善できる可能性はあるが，使い始めの段階から誰にでも使いやすいインタフェースとするためには，看過できない発生頻度であると考える．

◉センサによる発話動作検出

前節で述べた通り，手持ち型マイクにおける音声認識入力の精度向上，使い勝手向上には，口元とマイクの適正距離での発話支援，音声認識開始操作支援が必要である．手持ち型マイクでの音声入力は，ユーザはマイクを手に持って構え，口元に近づけて話すという動作を行う．そこで，マイクを構える動作を加速度センサで，口元への近接を距離センサでそれぞれ捉えることで，ロバストな発話動作検出と，口元とマイクの距離の適正化支援ができると考え，両センサを搭載した図 -1 のリモコンを試作し，それぞれの課題解決を試みた[1]．

図 -1　試作したセンサ内蔵リモコンデバイス

・発話動作検出方法

まず，加速度センサでリモコンがユーザに把持されたかどうかを判断する．ユーザに把持され，発話動作を検知した時点で距離センサを起動し，ユーザの口元とマイクの距離を計測する．加速度センサで発話動作を検知したにもかかわらず，マイクー口元間距離が 10cm よりも遠い場合は，上述の予備実験の結果から，音声認識にとっては適切な距離でないため，音声認識を開始とせず，口元をマイクに近づけるようにアプリケーション画面にメッセージ（例：口元をマイクに近づけてください）を表示する．加速度センサの出力が発話動作を検出した状態で，かつ距離センサによりマイクー口元間距離が 10cm 以内であることを検出した場合に，アプリケーション側へ音声認識開始コマンドを送信し音声認識を開始する．発話動作終了を検知した際には，音声認識終了コマンドを送信する．あるいは，発話動作終了前に一定時間以上の無音区間が続くと，音声認識エンジンが音声認識を終了する．この方法により，ユーザはボタン操作をすることなく，自然な発話動作を行うだけで適切に音声認識入力を行うことが可能となった．

・従来入力方式との比較評価

提案方式について，高齢者も含む被験者に対して，音声認識入力時にボタンを押下する従来方式とあわせて，比較評価を実施した．被験者は 21 名で，内訳は 20 ～ 30 代が 9 名（男性 4 名，女性 5 名），60 代が 12 名（男性 6 名，女性 6 名）である．図 -1 のリモコンに，従来入力方式（プレストーク，プッシュトーク）と提案方式（センサ駆動）による音声入力機能を実装し，家庭のリビングを模した実験室でそれぞれの方式について同一内容の発話（20 種類の人名）をしてもらった．プレストーク，プッシュトークについては，操作エラー率として全発話回数に対するボタン押し忘れ発話回数の割合を算出した．センサ駆動については，発話動作の検出漏れおよび誤検

☆1　今回使用したマイクの感度，実験室環境において 10cm であっただけのことであり，異なるマイク，環境ではその距離も異なる場合がある．

		プレストーク	プッシュトーク	センサ駆動
操作	全被験者	1.9%	8.1%	4.8%
エラー率	高齢者	2.9%	13.8%	5.0%
音声	全被験者	75.5%	81.9%	82.4%
認識率	高齢者	62.1%	71.3%	77.3%

表 -1　音声認識入力従来方式との比較結果

大分類	分類	例（括弧は正式名）
正式名称等 [63.8%]	正式名称	木村拓哉（木村拓哉）
	姓＋敬称等	小倉さん（小倉智昭）
愛称 [23.0%]	名前由来	キムタク（木村拓哉）
	名前非由来	ハンカチ王子（斎藤佑樹）
説明的表現 [13.2%]	グループ名	爆笑の太田（太田光）
	配役	やんくみ（仲間由紀恵）
	説明的	金麦の人（壇れい）

表 -2　人名の呼称表現の分類

大分類	分類	例（括弧は正式名）
正式名称等 [34.4%]	正式名称	NEWS23（NEWS23）
	部分語	アタック 25 （パネルクイズ アタック 25）
略称 [54.6%]	短縮語	スパモニ (スーパーモーニング)
	英語略称	WBS（ワールドビジネスサテライト）
その他 [11.0%]	通称	黒バラ (ブラックバラエティ)
	説明的	みののニュース （みのもんたの朝ズバッ！）

表 -3　番組名の呼称表現の分類

出を操作エラーとして扱った．なお，プレストーク，プッシュトーク，センサ駆動の実施順は被験者ごとに入れ替えを行い，実施順による影響を排除した．実験結果を表 -1 に示す．

それぞれの入力方式の特徴をまとめると次の通りとなる．まず，プレストークは，操作エラー率は最も低かったが，音声認識率は最も悪かった．これは，特に高齢被験者で，発話し始めてからボタンを押下したり，発話の途中でボタンを離してしまったり，ボタンを押し始めてからタスクを確認してしばらくしてから発話したりと，適切にボタンを操作できないため，誤認識となったケースが目立った．次に，プッシュトークは，プレストークに比べると操作に対する負荷は低く，うまく扱えないことに起因する誤認識は少なかった．一方で，操作に対する負荷が少ないことが，操作を忘れがちにさせる傾向があることが改めて確認できた．それらに対し，提案方式であるセンサ駆動は，高齢被験者にも特に使い方を丁寧に説明する必要なく音声認識入力を使ってもらうことができた．エラー率は 4.8% とプッシュトークのエラー率に比べて優位な性能を実現でき，音声認識率は最も良い結果が得られた☆2．

プレストークは機器操作が不得手なユーザにとっては適切に扱うのが難しく，プッシュトークによる音声認識入力の場合は，ボタンを押し忘れたまま発話してしまう頻度が高い．一方で，センサ駆動は，双方の欠点を補う特長があり，高齢者など機器操作が不得意なユーザでも，習熟なしで音声認識入力を扱うことができる方法として有用であることが分かった．

Web 情報を利用した愛称推定

テレビの番組検索では，配信されている EPG（Electronic Program Guide：電子番組表）が検索対象である．EPG からは人名や番組名の正式名称を取得することができ，日々変化する番組の変化に追従した音声認識語彙を生成できる．しかし，日常的に使われている人名の愛称や番組名の略称などの言い換え表現の情報は取得できない．このため，これらの語が音声入力されると音声認識辞書の未知語となり誤認識の原因になる．

◉人名・番組名の言い換え語の調査

言い換え表現対応の必要性を調査するため，『思い浮かんだ人名・番組名を，普段読んでいる呼び方で 10 個ずつ記載してもらう』というアンケート調査を 20 代から 50 代の男女計 32 名に対して実施した．その結果，人名 326 個，番組名 317 個の表現が得られた．表 -2, 3 にその例と種類を示す．また，正式名称やその簡単な派生表現（表中の正式名称等）だけでは，人名で 6 割程度，番組名で 4 割弱の表現にしか対応できないことが分かった．

◉人名愛称推定手法

愛称には，表 -2 のように，「名前由来の愛称」と「名前由来でない愛称」がある．そこで，各タイプの愛称を別々の方法で推定し，それらを最後に組み合わせる手法を開発した[2]．処理の概要を図 -2 に示す．

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，愛称生成ルールを自動的に作り出す．そして，新たに入力された人名（正式名）にそのルールを適用することにより，名前由来の愛称を推定する．

また，名前由来でない愛称については，「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する．たとえば，「ゴジラこと松井秀喜」という表現があった場合に，「こと」の前の「ゴジラ」を抽出する方法である．

さらに，上記 2 手法により別々に推定した愛称候補を Web 上の頻度を基に選定し，愛称推定結果とする．

☆2　センサ駆動の後にプレストーク，プッシュトークを実施した場合，センサ駆動での適正距離発話支援の影響を受け，その効果が確認しづらかった．しかし，センサ駆動の前に実施した場合の音声認識率は，プレストークが 74.3%，プッシュトークが 76.4% であり，適正距離発話支援の効果が読み取れる．

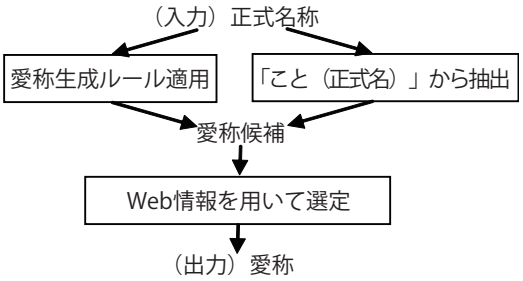

図 -2　人名愛称推定の流れ

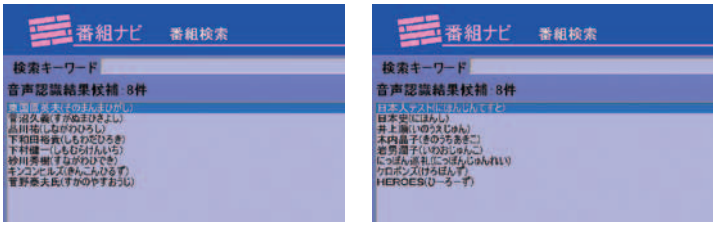

図 -3　音声認識結果例

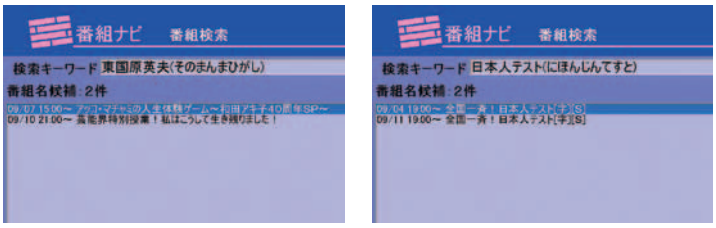

図 -4　番組検索結果例

本手法によって推定した愛称を，先にアンケートで収集した人名に適用して評価した．その結果，カバー率（アンケート中の出現頻度を考慮した再現率）は 81.5% で，正式名称等のみを用いた場合（6 割強）から 2 割近く改善されていることを確認した．なお，番組名略称推定については，文献 3）を参照されたい．

テレビ番組検索による実証実験

センサを用いた発話動作検知，適正距離発話支援と，人名の正式名称からの愛称推定手法を盛り込んだテレビ番組検索システムを試作した．音声認識語彙には，8 日分の実際の EPG から出演者，番組の正式名称を抽出し，人名の愛称推定に加え，単純なルールで番組名の部分語を生成し，正式名称と合わせて約 7,000 語を登録した．

提案手法の有用性を評価するため，本システムを用いた番組検索の際の検索キーワードの入力を，下記の 3 種類の入力方法で行い，それぞれの特徴を比較評価する実験を実施した．

① キーボード入力＋マウス操作（PC 環境を想定）
② スクリーンキーボード入力＋ボタン操作（既存のテレビでの入力を想定）
③ 音声認識入力＋ボタン操作（提案手法）

それぞれの入力方法について，20 代から 70 代までの 34 名の被験者に対して，同一の番組検索タスクを実施し，タスク完了までの所要時間を計測した．なお，タスクは，具体的な番組名や出演者名を提示し，番組名であればその番組を，出演者名であればその人が出演している任意の番組を検索するというものである．③における音声認識結果例，番組検索結果例を図 -3, 4 に示す．

結果は，発話に対する音声認識語彙のカバー率は 93.7% であった．このうち愛称表現対応を行っていないとした場合のカバー率は 85.1% で，愛称表現対応によりカバー率を約 9 ポイント向上できていることが確認できた．

また，全被験者の平均所要時間で比較すると従来手法の②に対し，提案手法の③は，所要時間を約 40% 短縮できたことが分かった．特に高齢者では，提案手法により所要時間を従来手法の②からおよそ半減でき，PC 環境の①に対しても約 30% 短縮できた．さらに，うち 4 名の被験者は①と②はうまく操作できずタスクを途中で断念したが，③では全員がすべてのタスクを完了することができた．主観アンケートでも，実際に使ってみたいかどうかという質問については，③が最も良好な結果となった．

このように，音声入力インタフェースを日常生活で使用するためには，音声認識エンジンそのものの性能向上はもちろん重要であるが，それ以外に入力方式や音声認識語彙についてもターゲットに応じた検討が必要である．実用化に向けて，各機能のさらなる性能向上および実環境での評価を引き続き行っていく．

参考文献

1）大内他：複数のセンサと自然言語処理技術による使いやすい音声入力インタフェース，マルチメディア，分散，協調とモバイル（DICOMO 2009）シンポジウム論文集，pp.1804-1814 (July 2009).
2）若木他：Web 情報を用いた人物の愛称抽出，日本データベース学会論文誌，Vol.7, No.1, pp.169-174 (June 2008).
3）Wakaki, et al.：Abbreviation Generation for Japanese Multi-Word Expressions, Proceedings of ACL-IJCNLP 2009 Workshop on Multiword Expressions, pp.63-70 (Aug. 2009).

（平成 21 年 10 月 30 日受付）

大内一成（正会員）
kazushige.ouchi@toshiba.co.jp
1998 年早稲田大学大学院理工学研究科物理学および応用物理学専攻修了，同年（株）東芝入社．現在，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にて，状況認識技術を応用したヒューマンインタフェースの研究開発に従事．ヒューマンインタフェース学会会員．

若木裕美（正会員）
hiromi.wakaki@toshiba.co.jp
2007 年東京大学大学院情報理工学系研究科電子情報学専攻博士課程修了，同年（株）東芝入社，博士（情報理工学）．現在，研究開発センター 知識メディアラボラトリーにて，自然言語処理，情報検索，音声対話等の研究開発に従事．