情報処理学会論文誌　Vol. 51　No. 3　846–855 (Mar. 2010)

人名と番組名の言い換えに対応する音声認識インタフェース

大内一成[†1]　若木裕美[†1]　屋野武秀[†1]
住田一男[†1]　土井美和子[†1]

テレビでの番組検索やWeb検索など，キーボードが使えない機器での文字入力の機会が増加している．そのような場面に適した音声認識の課題は，受理可能な語彙数と音声認識率にトレードオフの関係がある点である．実用的な認識率を確保できる語彙数の制約の中で，発話語彙に対するカバー率を上げるため，検索対象のコンテンツごとに愛称などを加えた認識語彙を生成する手法を提案する．これにより発話語彙カバー率が9ポイント向上していることが確認できた．さらに音声認識とリモコン操作を併用したマルチモーダル操作手法により，テレビ番組検索実験では，スクリーンキーボードを用いた従来手法より，提案手法がタスク完了までの所要時間を約40%短縮できた．

Speech Recognition Interface Accepting Paraphrases of Celebrity Names and Program Titles

Kazushige Ouchi,[†1] Hiromi Wakaki,[†1]
Takehide Yano,[†1] Kazuo Sumita[†1] and Miwako Doi[†1]

The opportunities of character entry on keyboard less devices such as TV have increased. The problem of speech recognition which is suitable for the situation is a trade-off between the number of acceptable vocabulary and recognition accuracy. In order to improve the cover ratio in the limitation of the number to keep the practical accuracy, we propose vocabulary generation including paraphrases such as nicknames from the targeted contents. We confirmed that it improved the cover ratio of vocabulary by 9 points by adopting paraphrases. Moreover, in the experiment of TV program retrieval, the proposed multimodal interface by the combined use of speech recognition and remote control shortened the time required to finish all tasks by about 40% against the conventional method using a screen keyboard.

1. はじめに

様々な情報のデジタル化にともない，私たちの身の回りの情報は，その量だけでなく多様性も含めて日々増加している[1),2)]．この大量かつ多種多様な情報の中から，必要な情報を的確に検索する手法として，パソコンや携帯電話では，検索キーワードをキーボードあるいはボタンで入力し，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る[3)]．

さらに，パソコンや携帯電話以外のAV機器，車載機器などにおいても，扱う情報の増加にともなって情報検索の必要性が増してきている．たとえば，テレビの番組検索，カーナビゲーションシステム（カーナビ）の目的地検索などが例としてあげられる．これらの機器で検索キーワードの入力を行うためには，スクリーンキーボードを画面に表示して1文字ずつ入力する方法が主流であるが，ボタンの操作数が多く，目的遂行までにかなりの手間と時間を要するのが現状である．これは，特に高齢者など，機器操作を得意としないユーザにとっては，使いにくいものとなっている[4),5)]．しかも，今後はIPTV（Internet Protocol Television）や動画共有サービスなどの動画コンテンツのテレビでの視聴の増加などが見込まれ，既存の電子番組表からの番組選択とあわせて，文字入力によるコンテンツ検索の必要性は現在より増加すると思われる．

一方，音声認識入力は，特定業務での作業効率向上[6),7)]などで活用されてきている．上述のAV機器や車載機器などでの情報検索の使い勝手の悪さを改善するために，音声認識により検索キーワードを入力することが期待されている[8),9)]．

本論文では，まず，AV機器や車載機器など，パソコンや携帯電話以外の機器における情報検索に音声認識入力を使う際の課題を整理する．その解決方法として，発話語彙に対するカバー率を上げるために，検索対象コンテンツごとに愛称なども加えた認識語彙を生成する手法を用いる．また，自然な発話で検索キーワードを入力し，それ以外はリモコンで操作を行うマルチモーダルな入力方式を提案する．さらに，リモコンのボタンで文字入力を行うため現状非常に面倒なテレビの番組検索を題材とし，音声認識を活用することで使いやすさの向上を図った番組検索システムを構築した．最後に，提案手法とリモコンのボタンのみで検索キーワードの入力を行う従来手法との比較評価実験を行い，その有用性を示す．

†1 株式会社東芝研究開発センター
Corporate Research and Development Center, Toshiba Corporation

2. 音声認識システムの課題

音声認識（孤立単語認識，連続音声認識）を用いたシステムでは，音声認識により受理可能な語彙は事前に辞書登録されていなければならない．ユーザが辞書登録されていない語（未知語）を発話した場合，音声認識システムは，事前に登録された語彙の中から発話された音声に近い語彙を候補とする．そのため，未知語の存在は誤認識やそれに起因するシステムの誤作動の原因となる．

未知語発話による誤認識を減らすためには，発話が想定される語彙を事前に幅広く辞書登録しておく必要がある．しかし，一般に音声認識システムが受理可能な語彙数と音声認識率はトレードオフの関係にある．語彙数が多くなると類似候補が多くなるため，音声認識率は下がる（認識に要する時間も増加する）．逆に，語彙数が少ないと，類似候補が少ないので，大語彙の場合に比べて高い精度で音声認識入力を行うことができる（認識に要する時間も短い）．しかし発話語彙に対するカバー率（発話語彙カバー率）が低下するため，未知語による誤認識が多くなる．つまり，極力未知語を減らし，最適な語彙セットで辞書を構成できるかどうかが，実運用では非常に重要である．

発話語彙カバー率向上の 1 つの方法として考えられるのが，場面に応じて辞書を切り替える方法である．音声認識による操作が可能なカーナビでは，階層化したメニューを音声で選択し，メニューごとに認識語彙を切り替えるが，基本的には正式名称の固定語彙である．しかし実際には，「二子玉川（ふたこたまがわ）」は「にこたま」，「溝の口（みぞのくち）」は「のくち」のように，普段から愛称や略称で呼ばれる地名は多い．つまり，正式名称は人口に膾炙すると愛称や略称で呼ばれるようになり，固定語彙ではこのような語彙の変動に追随できない．愛称や略称のような語彙変動への追随は，特に AV 機器の利用において重要であり，語彙変動に追随できるように，検索対象コンテンツ変動に合わせて認識語彙を生成する必要がある．たとえば，テレビにおける情報検索は Web 検索と番組検索が主な対象となる．音声認識を用いた Web 検索については，検索対象テキストから言語モデルを作成することで認識率の改善を行う取り組みなどがある[10]．一方，番組検索については，検索対象は配信されている EPG（Electronic Program Guide）データである．EPG データに基づいて辞書を再構成すれば，未知語を削減できる．しかし，EPG データだけでは，正式名称以外の愛称や略称などの言い換え表現が含まれないため，これらが未知語として誤認識の原因となる．これまでに，登録語彙に対する略語の推定を行い，辞書へ追加する取り組み[11),12)]はあるが，番組検索においては，略語以外に人名の愛称が使われることも想定される．

また，従来音声認識では，操作も含めてすべて音声認識で行う方法がとられてきた．特にカーナビでは，運転中はすべての操作を音声で行う必要があり，各メニューで非常に限られた認識語彙数にして切り替えている．しかし，階層ごとに受入可能な語彙をユーザが意識しなければ，適切な語彙を入力できず，また，目的遂行まで複数回の入力操作が必要となり，煩雑になってくる．これに対しテレビでは，すべての操作を音声で行う必要はなく，手元のリモコンを併用したマルチモーダルな入力体系が適している[13)]．この点は車載用途との大きな違いである（もちろん，車載用途においても，停車中であればマルチモーダルな入力は利用可能ではある）．

以上をまとめると，発話語彙カバー率向上のために，検索対象コンテンツごとに愛称などを加えた認識語彙を生成することが必要である．さらに，音声認識インタフェースとして使い勝手を向上させるために，すべて音声認識で操作する方法ではなく，リモコン操作と音声認識を融合したマルチモーダルな操作が必要である．本論文では，この 2 つの提案手法の有用性を，テレビの番組検索機能を題材として実証する．なお，愛称などの言い換え表現の推定方法は参考文献 15) の方法を活用している．参考文献 15) は，推定方法の有用性を Web に出現する愛称数に対するカバー率（言語的語彙カバー率）により明らかにしている．これに対し，本論文は音声認識インタフェースを扱っているので，実際に発話された語彙に対するカバー率（発話語彙カバー率）により有用性を明らかにしている点が異なる．

3. 番組検索のための認識語彙生成

3.1 言い換え表現対応の必要性

EPG には，正式な出演者名，正式な番組名のみが記載されており，出演者の愛称や番組の略称の記載はない．ユーザが自然な発話で所望の番組を検索するためには，音声認識エンジンがこれらの言い換え表現にも対応することが必要と考える．そこで，まず言い換え表現対応の必要性を確認するため，

- テレビに出ている有名人を普段呼んでいる呼び方
- 番組名を普段呼んでいる呼び方

について，アンケート調査を実施した．被験者は 20 代から 50 代の計 32 名で，それぞれの設問に対し，思いついた人または番組について，自由に 10 個ずつ記述方式で回答してもらった．収集できたデータ数は，人名の呼び方が 326 個，番組名の呼び方が 317 個であった．人名の呼び方の内訳を**図 1** に，番組名の呼び方の内訳を**図 2** に示す．

人名の呼び方については，正式名称またはそれに準じる表現が 63.8%，愛称が 23.0%，そ

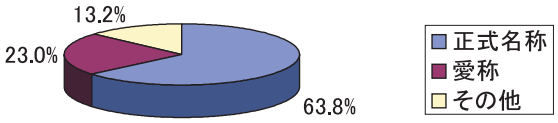

図 1　出演者名の呼び方の内訳
Fig. 1　The appellations of performers' names.

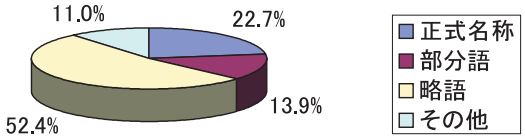

図 2　番組名の呼び方の内訳
Fig. 2　The appellations of program titles.

の他の呼び方が 13.2%であった．ここで，正式名称またはそれに準じる表現とは，本名（たとえば「森田一義」），芸名（たとえば「タモリ」），姓，名，あるいは姓＋「さん」（たとえば正式名称が「木村拓哉」の場合の「木村」「拓哉」「木村さん」）などを指す．愛称とは，特定の人物に対して定型的に使用される表現であり，一般に通用する呼び名であるとする（たとえば「木村拓哉」の愛称としての「キムタク」）．これら正式名称，愛称に分類できない表現をその他の呼び方とする．その他の呼び方としては，グループ名と組み合わせた表現（例：爆笑の太田），役職や職業名をつけた呼び名（例：三谷監督），説明的表現（例：麒麟の田村じゃない方）などがあった．つまり，EPG に記載されている正式名称およびそれから機械的に生成できるそれに準じた表現だけの対応では，ユーザが使う人物表現に対し 6 割程度の言語的語彙カバー率しか見込めないこととなる．

一方，番組名の呼び方については，正式名称（例：パネルクイズ　アタック 25）および正式名称を明示的な区切り文字（記号，空白，句読点など）で分割して生成できる部分語表現（例：アタック 25）を合わせた呼び方は，全体の約 3 分の 1 にあたる 36.6%であった．番組名の正式名称を省略して作られた 1 語（たとえば「ズームイン!! サタデー」を略した「ズムサタ」，「ミュージックステーション」を略した「M ステ」など）は，52.4%と全体の半数以上を占めることが分かった．

以上のように，出演者名，番組名の双方について，愛称，略称への対応の必要性が示唆される結果となった．本論文では，まず，人名の愛称への対応を行い，番組名の部分語対応と合わせて音声認識語彙を拡張し，その効果を確認する．

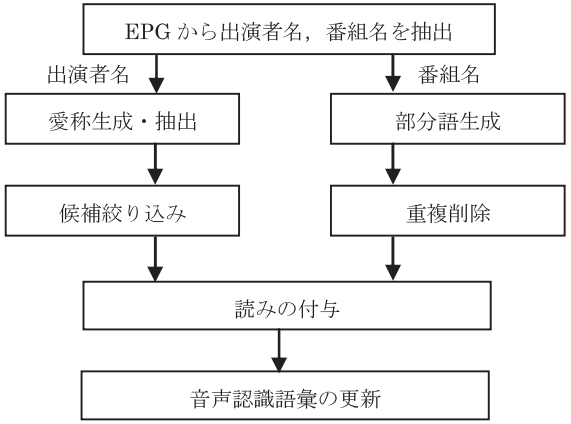

図 3　EPG からの音声認識語彙生成概要
Fig. 3　The overview of speech recognition vocabulary generation from EPG.

3.2　処 理 概 要

音声認識による番組検索を実現するために，テレビ放送で実際に配信されている EPG から，出演者名，番組名を抽出するとともに，前節で必要性が示唆された出演者名の愛称を追加した音声認識用の語彙セットを生成する．一連の処理の流れを**図 3** に示す．

まず，配信されている EPG（ADAMS-EPG[14] では 8 日分）に記述されている正式な番組名と正式な出演者名を抽出する．EPG はタグによって構造化されており，番組名として <TITLE> タグに囲まれた文字列を抽出する．出演者名は長い番組内容記述を表す <LONG_DESC> タグに囲まれた記述内容のうち，<ITEM> タグに「出演者」と記載された直後の <TEXT> タグに記述されており，それらを抽出する．なお，記述内容には，カンマ，括弧，「○○ ほか」など人名以外の記号や記述も多く存在するため，それらを除外して出演者名のみを抽出する．

このようにして抽出した出演者名に対する愛称推定，番組名に対する部分語生成は，それぞれ別プロセスで実行する．その処理内容については，次節で具体的に説明する．最後に，EPG から抽出した出演者名，番組名，および出演者名の愛称，番組名の部分語に含まれる漢字やアルファベット表記に対して読みの付与を行い，音声認識エンジンの認識語彙を更新する．読みの付与には，当社内で開発中の音声合成で用いられている読み生成ツールを活用した．

表 1　人名「滝沢秀明」の場合の記号への変換例（カタカナ表記）
Table 1　Converting characters of the name into symbols in the case of “Hideaki Takizawa”.

	文字種	姓名	文字位置	記号
タ	3	1	1	311
キ	3	1	2	312
:	:	:	:	:
ア	3	2	3	323
キ	3	2	4	324

3.3　言い換え表現による語彙拡張

3.3.1　出演者の愛称

EPG から抽出した出演者名について，以下の 2 通りの手順で愛称を推定する．

①　ルールに基づいた（人名由来の）愛称候補生成

人名に由来した愛称候補を生成するにあたり，まず，人名とその愛称の組を事前に学習データとして与え，人間が愛称をつけるルールを学習させる．ここで，正式名称（姓/名），その読みのひらがな，その読みのカタカナ，愛称の 4 つを学習データとして与える．各入力に対し，次に示すルールで番号を与え，1 文字ごとに 3 桁の数字へ置換する．まず，100 の位の値は文字種を表し，「1」が正式名，「2」がひらがな表記，「3」がカタカナ表記に対応する．10 の位の値はその文字が姓名のいずれに含まれるかを表し，姓ならば「1」，名ならば「2」とする．1 の位の値はその文字の各単語中での先頭からの位置を表す．**表 1** は，正式名称「滝沢秀明」のカタカナ表記に対して上記ルールで番号を付与した例である．「滝沢秀明」の愛称「タッキー」を学習する場合，「タッキー」は，表 1 の記号に置き換えると「311ッ 312 ー」となり，このルールを学習し保存する．

このようにして得られた愛称生成ルールをもとに，新たな人名（正式名称，ひらがな表記，カタカナ表記）に対応する候補を生成し，これらの各候補について，Web 上で正式名称を検索語とする検索結果上位の文書中で，正式名称（fullname）の前後に各候補の記述があるかどうかを調べ，該当するものを愛称候補（nick）とする．さらに，Web 上で「nick こと fullname」の連語を検索語として検索を行い，検索結果が一定数以下の nick は削除する．このようにして絞り込みを行い，人名由来の愛称候補を生成する．

②　Web 上の知識を用いた人名に由来しない愛称の抽出

愛称には，人名に由来せず上述のルールでは生成できないものも存在する（たとえば，「斎藤佑樹選手」の愛称である「ハンカチ王子」）．これらは単純なルールでは生成することが難しいため，Web 上の知識を活用した以下の手順で対応を試みる．

まず，「こと fullname」で Web 検索をし，検索結果上位のページ（Page）のうち，「こと fullname」の前の文字列を集める（この文字列集合を Str とする）．「こと」の前の 10 文字に対する全接尾辞 10 個に対し，Page と Str 中での直前の異なり数と先頭の異なり数を数える．各接尾辞について，直前の異なり数 > 1 かつ先頭の異なり数 $= 1$ のとき，その接尾辞を Data として登録し，また，直前の異なり数 $= 1$ かつ先頭の異なり数 > 1 のとき，その接尾辞の先頭の 1 文字を除いた残りの文字列を Data として登録する．Data のうち，Page 中で 1 回も出現しないものは削除する．残った Data について，末尾の共通する部分文字列は Str 中でのそれぞれの出現頻度を用いて絞り込みを行い，nick とする．最後に，① と同様に Web 上で「nick こと fullname」の連語を検索語として検索を行い，検索結果が一定数以下の nick は削除する．

上述の 2 つの方法で抽出した愛称が，実際に使われている有名人の呼称表現をどのくらいカバーできるかを調査したところ，本手法で抽出した愛称の言語的語彙カバー率は，81.5%であった．つまり，図 1 中の愛称 23.0%のうち，81.5%分に相当する 18.7%は提案手法によってカバーできるものと推定できる．すなわち，図 1 のうち 82.5%（$= 63.8 + 18.7$）がカバーできる計算となる．なお，出演者の愛称抽出については，参考文献 15) に詳しいので，そちらを参照されたい．

3.3.2　番組名の部分語

EPG から抽出した番組名については，特徴的な記号や空白を区切りとして以下の手順で部分語を生成する．

①　不要文字・不要区間の削除

[二]（二カ国語放送），[再]（再放送），［午前 5：00 まで放送を休止します］など，全角 “[]”，および半角 “[]” には番組名以外の放送に関する属性情報が記述されているため，それらは削除する．

また，“[字] あしたをつかめ〜平成若者仕事図鑑 ▽ 笑顔とセンスでお客様をつかめ〜デパート販売員” のように，番組名の中の “▽” 以降に番組内容に関する記述がある場合があるため，“▽” 以降の記述も削除する．

②　メインタイトルとサブタイトルの分割

次に，括弧（「」『』など）をサブタイトルと判断して，括弧前までをメインタイトル，括弧の中をサブタイトルとして分割する．

③　メインタイトルの分割

メインタイトル中に，空白，感嘆符 “！”，記号 “〜”，“—” などがあれば，それらの前後でタイトルを分割し，それぞれの組合せをメインタイトルのバリエーションとして登録する．

上記手順で生成した部分語のそれぞれについて読みを付与することで，たとえば下記のような部分語読みを生成する．

- EPG 記載番組名：NHK 高校講座　理科総合 A・B「ひんやり・ほかほかの物理」
- 生成した部分語の読み：
  えぬえいちけいこうこうこうざ
  りかそうごうえーびー
  ひんやりほかほかのぶつり
  えぬえいちけいこうこうこうざりかそうごうえーびー
  えぬえいちけいこうこうこうざりかそうごうえーびーひんやりほかほかのぶつり

なお，“NHK 高校講座　日本史「……」” などの他の教科の番組名からも “えぬえいちけいこうこうこざ” の部分語が生成されるため，同一の部分語が生成された場合は重複登録しないように削除する．また，部分語の読みに対応する表記は，その読みに対応した表記だけを登録しておく．よって，当該部分語を検索キーワードとした部分一致検索を実施することになる．

また，上記の例では，“えぬえいちけいこうこうこうざりかそうごうえーびーひんやりほかほかのぶつり” など，EPG 記載の番組名全部に対する読みも今回は登録したが，実際にこのように長い番組タイトルをそのまま全部発話することはあまりないと考えられる．一定文字数以上の読みは削除するなど，読みの長さを基準とした語彙の選別も今後検討する．

4.　マルチモーダル操作でのテレビ番組検索による実証実験

上述の手法により実際の EPG から出演者名の愛称，番組名の部分語による音声認識語彙の拡張を行い，音声認識を活用した番組検索の効果を検証するための実証実験システムを試作した．本システムにより，従来のリモコン操作のみによる番組検索に比べて，操作負担（操作時間）の軽減（短縮）と使い勝手の向上を目標としている．

検索対象とするのは，実際の地上波テレビ放送における EPG 8 日分のデータで，前章で説明した手法で，出演者名とそれらの愛称，番組名とそれらの部分語を抽出し，読みを付与して音声認識語彙を生成した．8 日分の EPG から抽出した正式な出演者名 3,573 語に，3.3.1 項の手法で抽出した 992 語の愛称を追加した計 4,565 語と，正式な番組名と 3.3.2 項の手法で生成した部分語を追加した 2,498 語，両者を合わせ音声認識語彙として登録した語彙数は，計 7,063 語とした．

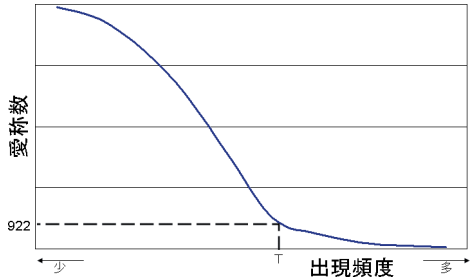

図 4　Web 上の出現頻度と愛称数の関係
Fig. 4　Relationship between the number of nicknames and the frequency of appearance on the Web.

追加した愛称 922 語の選定は次のようにした．**図 4** は愛称数と Web 上の出現頻度との関係を示す概略図である．ある出現頻度（今回は目視で決定した．その値を T とする．T の決定方法については今後の課題である）を境に，それ以下では愛称数が一気に増加する．これらにはめったに使われていないもの，一般に知られていないものが多数存在する．Web に存在する愛称全体に対する言語的語彙カバー率を高くするには，T はなるべく小さく選定すべきである．しかし，出現頻度が少ないものは，実際にユーザが知らない場合も多く，発話されない可能性も高い．つまり

言語的語彙カバー率 $\neq$ 発話語彙カバー率

となっており，音声認識では発話語彙カバー率を優先すべきである．言語的語彙カバー率は Web に出現する全語彙数と選択した語彙数から単純に算出できるが，発話語彙カバー率は実際にユーザに発話してもらわないと分からない．よって，今回の実験では，この出現頻度 T より多い出現頻度の愛称 922 語を採用した．この場合の言語的語彙カバー率は 7.6%であるが，発話語彙カバー率は 4.3.4 項で述べるように 93.7%となっている．ただし，実用化に向けて実験を積み重ねることで，発話語彙カバー率と言語的語彙カバー率との関係についても明確にしていく必要がある．また，発話語彙カバー率を上げるために語彙数（特に愛称）を増やし続ければよいわけではなく，類似した発音が増え，音声認識率が悪化する可能性もある．発話語彙カバー率と音声認識率との関係も今後検討の必要がある．

ちなみに，本システムで使用する大語彙孤立単語認識エンジンにおいて，語彙数約 7,000 語，SNR が 20 dB の環境で，発話区間を手動で切り出した場合，音声認識率が 88.5%と実

用的な性能が確保できることを，システム構築前の調査で確認した．なお，音響信号のサンプリング周波数は 16 kHz で，フレーム幅 25 ms・シフト幅 8 ms で分析した．音響特徴量には 0–12 次の MFCC（Mel-Frequency Cepstral Coefficient）とその Δ および ΔΔ で構成される計 39 次元の特徴ベクトルを用いた．なお，前処理に雑音除去は適用していない．音響モデルには 3 状態 20 混合の left-to-right 型の HMM（Hidden Markov Model）を用いた．

4.1　マルチモーダル操作

実験用システムはノート PC 上に構築し，操作デバイスとして，市販のマイク内蔵型コントローラ（Philips 社製 SpeechMike Pro[16]）を用いることとした．出演者名，番組名およびその言い換えを含めて音声入力すると，音声認識候補リストが画面上に表示される．その中からの正解候補や目的の番組の選択などは，コントローラを用いたボタン操作で実施する構成とした．日常生活においてテレビの番組検索を行うシーンを想定すると，自動車の運転中などとは異なり，リモコンによるボタン操作が可能であるため，音声認識入力とボタン操作をそれぞれ組み合わせて使用するマルチモーダル操作が適切であると考え，上記のような操作体系とした．

なお，8 日分の EPG データは，XML データベースに格納しておき，キーワード検索が実行された際に XQuery を発行してデータベース内を検索する．

4.2　機 能 概 要

ここでは，音声認識による検索キーワード入力から，入力した検索キーワードによる EPG 検索および結果の表示までの機能概要について説明する．

まず，ユーザは目的の出演者名，あるいは番組名を正式名称，愛称，部分語など，好みの表現で音声入力を行う．音声入力の開始時に，マイク内蔵コントローラのボタン押下により音声認識開始をシステム側に通知する．音声認識の終了は，音声認識エンジンが無音区間を検出することにより自動的に行う．出演者名の愛称（この例では「そのまんま東」）を音声入力した際の一例，番組名の部分語（この例では「日本人テスト」）を音声入力した際の一例を**図 5** に示す．音声認識の結果は，尤度順に上位の 8 件がリスト表示される．それぞれ正式出演者名「東国原英夫」に対する愛称「そのまんま東」，正式番組名「全国一斉！日本人テスト [字][S]」の部分語「日本人テスト」が音声認識エンジンに受理されたことが分かる．なお，それぞれの候補がどんな読みに基づいて候補としてリストアップされたか分かるように，正式出演者名，番組名（部分語）に対応した読みを括弧内に記載しておくこととした．もし，正解がリスト中にない場合は，再度音声入力を行う．

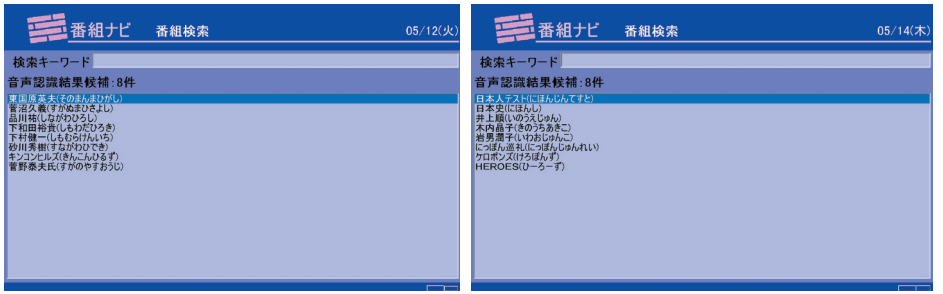

図 5　音声認識結果例
Fig. 5　Results of speech recognition.

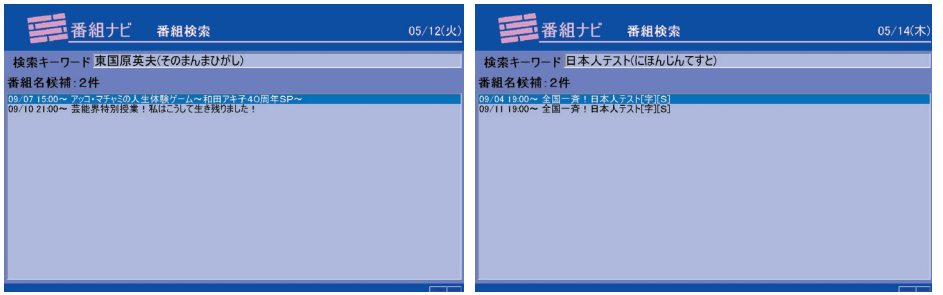

図 6　番組検索結果例
Fig. 6　Results of TV program retrieval.

図 5 の音声認識結果のリストから目的の人名あるいは番組名を選択すると，その正式名称で EPG が格納された XML データベースを検索する．検索結果の例を**図 6** に示す．出演者名の場合はその出演者が出演している番組が，番組名の場合はその検索キーワードの文字列が含まれる番組が検索されてリスト表示される．あとは，これらのリストの中から所望の番組を選択すると，EPG の詳細情報を確認することができ，簡単に録画予約や視聴ができる．

なお，検索した出演者の当該期間での出演番組数が非常に多い場合，あるいは，検索キーワードの部分語文字列にヒットする番組数が非常に多い場合は，現状では番組候補リストを上下にスクロールして多数の候補の中から目的の番組を探し出す必要がある．**図 7** に当該出演者の出演番組数が多い場合，部分語文字列にヒットする番組数が多い場合の例をそれぞ

図 7　番組検索結果例
Fig. 7　Results of TV program retrieval.

れ示す．このように多数の候補の中から，リモコンのボタンだけで目的の番組を探し出すことは負荷が高い．また，別のキーワードによる再検索を行うにしても，どのようなキーワードで効果的な検索が可能かをユーザが判断することは難しい．よって，検索結果に応じて絞り込み検索用の認識語彙を生成し効率良く音声で絞り込みを行うことができる機能，さらには，絞り込みに適したキーワードを提示する機能などの必要性が示唆される．これらについては，今後の課題とする．

4.3 実　　験

4.3.1 実験概要

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（PC 環境を想定）
② スクリーンキーボード入力（既存のテレビ/レコーダでの入力を想定）
③ 音声認識入力（提案手法であるマルチモーダル操作）

① では，検索キーワード入力をノート PC のキーボードで行い，その他の番組選択などの操作はマウスを用いた．② では，本システム上に既存のテレビ/レコーダで使用されているものを模して作成したスクリーンキーボードを用意し，すべてマイク内蔵型コントローラで操作を行った．なお，テレビ/レコーダでは専用のリモコンによる十字キーあるいは携帯電話式の数字キー複数回押下により文字入力を行うのに対し，マイク内蔵型コントローラではトラックボールとボタンによる操作となる．事前に実際のレコーダ＋専用リモコンでの文字入力と，本システム＋マイク内蔵型コントローラでの文字入力を同一タスクで実施し，

(1)　風林火山(40)「三国同盟」
(2)　「みのもんた」が出ている番組
(3)　爆笑問題の検索ちゃん（10 月 13 日放送）
(4)　「森繁久彌」(もりしげ　ひさや)が出ている番組
(5)　日曜洋画劇場「アルマゲドン」
(6)　「木村拓哉」(きむら　たくや)が出ている番組
(7)　タモリ倶楽部(10 月 13 日放送)
(8)　「坂下千里子」(さかした　ちりこ)が出ている番組

図 8　実験タスク
Fig. 8　Experimental task.

後者が前者より有意差（$p < 0.01$）を持って短時間でタスク完了できたことが確認できたため，既存手法相当の入力として本システム＋マイク内蔵型コントローラを使用することとした．これに対して提案手法が有利であることが確認できれば，提案手法は既存手法に対しても有利であるとすることができる．③ では，マイク内蔵型コントローラのマイクに向かって音声認識入力を行い，音声認識候補や番組の選択はマイク内蔵型コントローラのトラックボールとボタンにより行う．

なお，①，② でも出演者名の愛称，番組名の部分語による番組検索を実施できるよう，入力されたテキストについて，音声認識語彙を参照することで正式名称を検索できるようにした．特に，入力した任意の文字列に対して部分一致検索を行えるようにしたため，③ よりも ①，② の方が入力の自由度は高く，有利な条件で実験を実施した．

4.3.2 タ　ス　ク

実験でのタスクは，被験者に録画予約する番組名あるいは出演者名を検索タスクとして画面に提示し，被験者は提示された検索タスクに基づいて検索キーワードを自由に入力して目的の番組を検索する．なお，被験者には本システムが愛称や部分語に対応していることは事前に説明しておいた．

検索タスクが番組名の場合は，その番組を録画予約できてタスク完了とする．レギュラー番組など同一名の番組が複数ある場合は日付も指定して提示することとした．出演者名が提示された場合は，その出演者が出演している任意の番組が録画予約されればタスク完了とする．1 つのタスクが完了すると，続いて次の課題が提示され，次々にタスクを実施する．実験で使用したタスクは，**図 8** のとおりである．なお，これらの番組名，出演者名に関連する音声認識語彙の読みがすべて正しく付与されていることを確認したうえで実験を実施した．

表 2　入力方法別全タスク完了までの平均所要時間（秒）
Table 2　Average time required to finish all tasks (second).

キーワード入力方法	非研究職一般者			IT 系研究者	全被験者
	若者	高齢者			
①キーボード	119.6	323.8	224.7	101.1	171.4
②スクリーンキーボード	217.2	426.6	323.8	243.3	277.0
③音声認識	123.8	228.9	180.8	139.9	165.8

4.3.3 実験条件

家庭のリビングを模した実験室において，合計 34 名の被験者に対して，図 8 のタスクを ① キーボード入力，② スクリーンキーボード入力，③ 音声認識入力のそれぞれについて実施した．なお，各入力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．被験者 34 名の内訳は以下のとおりである．

- IT 系研究者（20〜50 代）：12 名（男性 7 名，女性 5 名）
- 非研究職一般者：22 名
  - 若者（20 代）：10 名（男性 5 名，女性 5 名）
  - 高齢者（60〜70 代）：12 名（男性 6 名，女性 6 名）

4.3.4 実験結果

各入力方法別に図 8 のタスクがすべて完了するまでの平均所要時間を**表 2** に示す．

ここで，高齢者 4 名は，① キーボードと ② スクリーンキーボードでのタスクを途中で断念（タスクごとの制限時間 10 分以内に完了しなかったか，被験者がギブアップした）したため，表 2 の集計には含んでいない．ちなみに，タスクを途中で断念した被験者は，①，② とも同一の 4 名であった．

全被験者の平均所要時間では，従来手法としての ② スクリーンキーボードの場合は 277.0 秒，提案手法の ③ 音声認識によりキーワード入力をした場合は 165.8 秒と，提案手法により所要時間を約 40%短縮することができた（$p \ll 0.01$）．

高齢者に対してはさらに顕著で，提案手法により所要時間を約半減でき，PC 環境を想定した ① キーボードに対しても約 30%短縮することができた．しかも，上述のとおりうち 4 名の被験者は ① と ② のタスクを途中で断念しており，提案手法の有用性を明らかに確認することができた．

また，非研究職一般者の若者と高齢者について，入力方法別の所要時間を**図 9** に示す．この集計にも ① キーボードと ② スクリーンキーボードにおけるタスク途中断念 4 名のデータは除外している．高齢者は若者に比べて ① キーボードと ② スクリーンキーボードの分散が大きく，個人差が大きいことが分かる．その一方で，③ 音声認識の分散は小さく，機器操作の得意/不得意にかかわらず，音声認識による検索キーワード入力は個人差が少ないことが読み取れる．

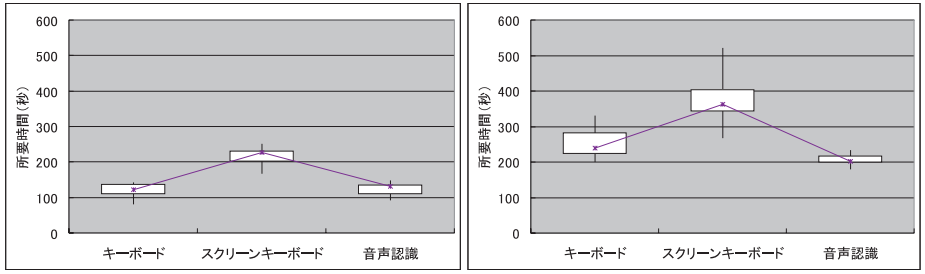

図 9　入力方法別所要時間（左：若者，右：高齢者）
Fig. 9　Time required to finish all tasks (left: young subjects, right: elderly subjects).

一方，発話に対する音声認識語彙の発話語彙カバー率は 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発話語彙カバー率は 85.1%で，愛称表現対応により発話語彙カバー率を約 9 ポイント向上させることができていることも確認できた．

音声認識語彙に登録されていない未知語発話のうち最も多かったものは，検索タスク「爆笑問題の検索ちゃん（10 月 13 日放送）」に対する「けんさくちゃん」という発話で，全被験者で 8 回の発話が観測された．形態素解析を用いた部分語生成，さらには，番組名の略称対応の必要性が示唆される結果となった．ただし，「けんさくちゃん」が受理されないことは少なくとも 2 度目の認識失敗で理解され，「ばくしょうもんだい」「ばくしょうもんだいのけんさくちゃん」など，別の受理可能な表現で再入力され，目的の番組を探し出すことができた．

ちなみに，1 タスクあたりの平均発話回数は 1.38 回で，これは誤認識発生の場合，あるいは未知語を発話した場合でも，たいていの場合は多くとも 3 回目の発話までに（未知語発話の場合は受理可能な表現を言い直して）正しく入力できるレベルである．なお，音声認識語彙に対応した発話における音声認識率は平均で 77.7%であった．事前に確認した認識率より低い結果となったが，ボタン押下による音声認識開始指示の後，目的のキーワード発話の前に「えーと」など，入力を意図しない発話に起因する誤認識が特に高齢者で多く発生したためであると考えられる．この点については今後の課題である．しかし，逆の見方をす

表 3　主観アンケートの結果
Table 3　The results of subjective test.

キーワード入力方法		非研究職一般者			IT 系研究者	全被験者
		若者	高齢者			
①	使いやすい	3.6	4.7	4.1	4.8	4.4
	面白い	3.3	3.1	3.2	2.8	3.0
	使ってみたい	3.1	4.3	3.6	3.7	3.7
②	使いやすい	1.8	2.1	1.9	1.4	1.7
	面白い	3.5	3.9	3.7	2.2	3.2
	使ってみたい	2.4	2.6	2.5	1.7	2.2
③	使いやすい	4.4	4.3	4.4	3.9	4.2
	面白い	4.6	4.3	4.5	4.4	4.4
	使ってみたい	4.4	3.9	4.2	4.2	4.2

れば，音声認識率がこの程度であっても，従来手法よりも所要時間を大幅に短縮できたことは，本手法の大きな特長であるともいえる．

各入力方法について，主観アンケートも実施した．①〜③ のそれぞれについて，使いやすいかどうか，面白いかどうか，実際に使ってみたいかどうかを，1（Negative）〜5（Positive）の 5 段階で評価してもらった．結果を平均値で**表 3** に示す．IT 系研究者や若者は，普段からキーボードに慣れているため，使いやすさに関しては ① キーボードが最も高い結果となったが，実際に使ってみたいかどうかについては ③ 音声認識が最も高い結果となった．日常生活におけるテレビ視聴時には，わざわざキーボードで文字入力を行うよりも，手元のリモコンで手軽に入力できることが求められていると考えられる．

また，高齢被験者に対しては，テレビ番組の録画予約を自分でできるかどうかも質問した．自分でできると回答した被験者は 12 名中 3 名のみで，その他の 9 名は同居している若い人にお願いして録画してもらっているとのことであった．これらの被験者からも，提案手法の音声認識を用いた番組検索が実際に使えるのであれば，番組予約を自分でやってみたいとのコメントが数多く得られた．

5. ま と め

本論文では，音声認識システムの課題である発話語彙カバー率を向上させることを目的とし，検索対象コンテンツごとに愛称などを含めた認識語彙を生成する手法を用いたシステム構築を行った．出演者名の愛称に対応することにより，実用的な音声認識精度を確保できる語彙数の範囲において，正式名称だけの場合より，発話語彙カバー率が約 9 ポイント向上していることを確認した．また，音声認識インタフェースの使い勝手を向上させるために，すべての操作を音声認識で行うのではなく，リモコン操作と音声認識を融合したマルチモーダル操作を提案した．テレビ番組検索による実証実験では，スクリーンキーボードを用いた従来手法による番組検索に比べて，提案手法の方がタスク完了までの所要時間を約 40%短縮することができた．また，提案手法は，高齢者などで機器操作が得意でないユーザにとって特に効果的であることも分かった．

今回の実証実験では，出演者名の愛称と番組名の部分語に対応した音声認識を活用することの有用性を示す結果を得ることができた．今後は番組名の略称への対応も検討し，さらなる発話語彙カバー率の向上を検討する．また，今後はより実用的なシーンを想定した検索タスクでの評価を行うとともに，発話語彙カバー率と言語的語彙カバー率の関係，発話語彙カバー率と音声認識率との関係についても明確にし，実用化に向けさらなる検討を進めていく．

参 考 文 献

1) 喜連川優，松岡　聡，松山隆司，須藤　修，安達　淳：情報爆発時代に向けた新しい IT 基盤技術の研究，人工知能学会誌，Vol.22, No.2, pp.209–214 (2007).
2) 小川克彦：情報大航海時代の到来—リアルとネットを結ぶ知的情報アクセス基盤，電子情報通信学会誌，Vol.91, No.8, pp.727–731 (2008).
3) 藤田澄男：フィールドを広げる自然言語処理：自然言語処理を利用した情報の検索・分類へのアプローチ，情報処理学会誌，Vol.40, No.4, pp.352–257 (1999).
4) 財団法人共用品推進機構：高齢者の家庭内での不便さ調査報告書 (1999).
5) 原　紀代，志田武彦，中　俊弥，南部美砂子，原田悦子：家電操作における高齢者の認知特性の研究，*Matsushita Technical Journal*, Vol.51, No.4, pp.29–33 (2005).
6) Pausch, R. and Leatherby, J.H.: An empirical study: Adding voice input to a graphical editor, *Journal of the American Voice Input/Output Society*, Vol.9, pp.55–66 (1991).
7) Karl, L.R., Pettey, M. and Shneiderman, B.: Speech versus mouse commands for word processing: An empirical evaluation, *International Journal of Man-Machine Studies*, Vol.39, Issue 4, pp.667–687 (1993).
8) 西崎博光，中川聖一：音声キーワードによるニュース音声データベース検索手法，情報処理学会論文誌，Vol.42, No.12, pp.3173–3184 (2001).
9) 北岡教英，角谷直子，中川聖一：音声対話システムの誤認識に対するユーザの繰返し訂

正発話の検出と認識，電子情報通信学会論文誌 D-II，Vol.J87-D-II, No.7, pp.1441–1450 (2004).
10) 伊藤克亘，藤井　敦：NTCIR-3 ワークショップにおける音声入力型ウェブ検索タスク，情報処理学会研究報告—音声言語情報処理（2002-SLP-043），pp.25–32 (2002).
11) 榎　将功，皇甫美華，大田健紘，柳田益造：日本語における略語自動生成法の検討とその音声インタフェースへの応用，情報処理学会研究報告—音声言語情報処理（2007-SLP-069），pp.313–318 (2007).
12) 勝丸真樹，駒谷和範，尾形哲也，奥乃　博：音声対話システムにおける簡略表現認識のための誤認識増加を抑制する自動語彙拡張，情報処理学会研究報告—音声言語情報処理（2008-SLP-071），pp.71–76 (2008).
13) 新田恒雄：GUI からマルチモーダル UI（MUI）に向けて，情報処理，Vol.36, No.11, pp.1039–1046 (1995).
14) ADAMS-EPG. http://www.tadv.jp/service/adams.html
15) 若木裕美，藤井寛子，福井美佳，住田一男：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 (2008).
16) SpeechMike Pro. http://www.dictation.philips.com/index.php?id=1425

(平成 21 年 5 月 25 日受付)
(平成 21 年 12 月 17 日採録)

大内　一成（正会員）

1998 年早稲田大学大学院理工学研究科物理学及応用物理学専攻修了．同年（株）東芝入社．状況認識技術を活用したヒューマンインタフェースの研究開発に従事．現在，(株）東芝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．本会ユビキタスコンピューティングシステム研究会幹事．ヒューマンインタフェース学会会員．

若木　裕美（正会員）

2002 年東京大学工学部電子情報工学科卒業．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．2007 年（株）東芝入社．現在，(株）東芝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．博士（情報理工学）．

屋野　武秀

1996 年神戸大学自然科学研究科博士前期課程修了．同年（株）東芝に入社．音声を利用したインタフェース，音声対話管理の研究開発に従事．現在，(株）東芝研究開発センター知識メディアラボラトリー研究主務．電子情報通信学会会員．

住田　一男（正会員）

1982 年東京工業大学大学院総合理工学研究科修士課程修了．同年（株）東芝入社．以来，自然言語処理，ナレッジマネジメント，情報検索，音声インタラクション，音声翻訳等の研究開発に従事．現在，(株）東芝研究開発センター知識メディアラボラトリー研究主幹，人工知能学会副会長，言語資源協会理事．博士（工学）．

土井美和子（フェロー）

1979 年東京大学大学院工学系修士課程修了．同年（株）東芝入社．「ヒューマンインタフェース」を専門分野とし，日本語ワープロ，機械翻訳，道案内サービス，ネットワークロボットの研究開発に従事．現在，(株）東芝研究開発センター首席技監．日本学術会議連携会員，東工大経営協議会委員，国立情報学研究所運営会議委員，科学技術振興機構運営会議委員，ヒューマンインタフェース学会会長等を務める．本会フェロー．博士（工学）．