情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2011-SLP-89 No.25
2011/12/20

単語断片の候補選択が可能な音声入力インタフェースの実装と評価

張　用起[†1]　　甲斐　充彦[†1]　　王　龍標[†1]

大語彙の単語入力タスクを音声を用いて遂行する際、認識精度が低い話者や環境での誤認識による入力効率低下の問題を解決するため、複数の認識候補の出力と単語断片の利用を併用したマルチモーダルインタフェースを提案する。提案するシステムでは、入力音声に対する ASR(Automatic Speech Recognition) システムの出力結果から複数の単語候補を画面に表示し、ユーザへ GUI(Graphical User Interface) で選択させる。また、認識精度の低下によって正解の認識候補が候補リストの下位に位置する際、単語断片で認識結果を絞り込み、より効率良く単語入力を行わせるアプローチとその実現法について紹介する。本稿では、この音声入力インタフェースシステムに対するシミュレーション実験の評価結果と、実際の被験者を対象としたオンラインシステムでの被験者実験の結果を示す。語彙サイズ約 42,000 単語の施設名入力タスクに対して実験を行った結果、提案するインタフェースシステムは認識結果の N-best 仮説だけを提示するシステムと比べ、入力効率を改善できることが確認できた。また、シミュレーション実験の結果では認識精度や候補リストの長さの変化に対して頑健性を持つことを示しており、実際のユーザに対する実験では、シミュレーション実験より入力効率の改善は少ないものの、インタフェースシステムの設計段階における改良により更なる性能の向上が期待できる。

Evaluation of A Speech Interface System with a Funtion to Select Word Fragment Candidates

Yonggee Jang,[†1] Atsuhiko Kai[†1]
and Longbiao Wang[†1]

We propose a multimodal interface using multiple candidates of words and their fragments to solve a problem that a input efficiency declines by recognition errors for a large-vocabulary spoken-word input task. Our system displays multiple candidates from ASR (Automatic Speech Recognition) results for speech input and makes users select a correct candidate through GUI(Graphical User Interface). Also, we describe an approach that can improve a efficiency of word inputs by narrowing word candidates from recognition results using word fragments. In this paper, we show experimental results for the proposed interface system by simulation and on-line evaluations. According to the results, we can see that the proposed method improves an input efficiency compared with a traditional method which displays N-best hypotheses only, and has a robustness for the various environments where the recognition accuracy and the length of candidates list were changed. In the experiment by real users, even though the proposed system can show less improvement than that of the simulation, we expect that the input efficiency of the proposed method will be improve, by conducting a review of an multimodal interface design.

1. はじめに

コンピュータのハードウェア、ソフトウェア技術の発展と共に、音声認識技術を利用した様々なシステムが実用化されており、スマートフォンや無線インターネットの普及などによって音声認識技術に触れることが容易になっている状況である。しかし、音声認識における誤認識の問題は、依然として音声認識技術の普及を防ぐ主な原因となっている。孤立単語認識に対しても、雑音が強い環境や遠隔音声入力では誤認識が起こる可能性が非常に高い。

誤認識問題に対し、認識器から出力される複数の認識仮説を用いる方法は簡単かつ強力な手段であり、昔からその有効性が示されてきた [1,2,3,5]。複数の認識候補を利用するアプローチの実現法として、画面に複数の候補を出力し、ユーザが GUI を通して発話した単語を選択するマルチモーダルインタフェースが考えられる。しかし、複数候補を提示するインタフェースシステムを利用するとき、雑音によって発話した単語が認識結果の下の順位に位置すると、限られた画面の大きさで発話単語を探すには多くの単語候補の参照が必要となり、入力効率が低下してしまう。

この問題を解決するため、我々は単語断片の絞り込みによってより少ない操作回数で単語が入力できる候補リストの構成法を提案した [5]。単語断片は語彙辞書から自動的に抽出し、認識結果の信頼度スコアを利用して単語候補と単語断片で構成された候補リストをユーザに提示した。単語の一部を利用するアプローチは以前から存在し、認識結果と短い単位の基本単語を併用する研究が行われてきた。組織名の音声入力インタフェースにおいて、孤立単語認識と連続基本単語認識の併用を利用し、少ない数の語彙数で組織名のカバー率を

†1 静岡大学 工学部
Faculty of Engineering, Shizuoka University

Vol.2011-SLP-89 No.25
2011/12/20

向上させる研究から、語彙単語を分割し基本単語として併用する方法の有効性が示されている [2]。また、姓名入力インタフェースにおいて、単語認識結果と共に音節を基本単位とした連続音節認識を併用し、姓名のカバー率及びインタフェースによる入力可能割合を向上させる研究も行われた [3]。最近の研究では、部分単語を用いて略語や未知語に対応できるインタフェースシステムも開発され、より広い語彙をカバーできることを示した [4]。一方、我々の先行研究では単語断片と単語候補で構成された候補リストを利用するシステムが、認識結果からの単語候補のみを利用する時より良い入力効率を示し、提案するシステムが候補リストの長さや雑音のレベルに頑健性を持っていることも確認できた [5]。但し、この研究では入力効率の改善に関する明確な評価基準が無く、語彙サイズも約 12,000 単語と限定されていた。

本研究では約 4 万単語の語彙を持つ施設名入力タスクに対して、様々な認識性能を持つ ASR システムで実際のユーザの操作を想定したシミュレーション実験を行い、提案するインタフェースシステムが典型的な複数候補を出力するシステムより高い入力効率を持つことを示す。そして、実際のユーザに対するオンライン実験を行い、単語断片を利用するアプローチに関する客観的・主観的評価を行う。

残りの部分は以下のように構成されている。2 章では単語候補と単語断片で構成される複数候補のリストを提示するインタフェースシステムについて説明する。3 章ではユーザの操作を想定したシミュレーション実験の結果を、4 章では実際のユーザの操作によるオンライン実験の結果を示す。5 章では実験からなる考察や今後の課題についてまとめる。

2. 提案するインタフェースの概要

本稿で提案するインタフェースシステムは、語彙辞書と認識結果を利用し複数の候補をリストとしてユーザに提示する。その候補リストは、認識結果からなる単語候補と、単語断片で認識結果の絞込みが可能な検索候補で構成される。

2.1 単語断片の抽出

大語彙を用いるタスクの語彙辞書では、複数の語彙単語の間で共通して含まれる部分（主に形態素単位）が多く存在する。我々はそのような単位を共通部分単語と呼び、本研究では語彙辞書から一定回数以上出現する形態素[⋆1]を利用する。図 1 に共通部分単語の抽出過程を示す。

⋆1 単独の形態素のみを使い、形態素列は利用しない。

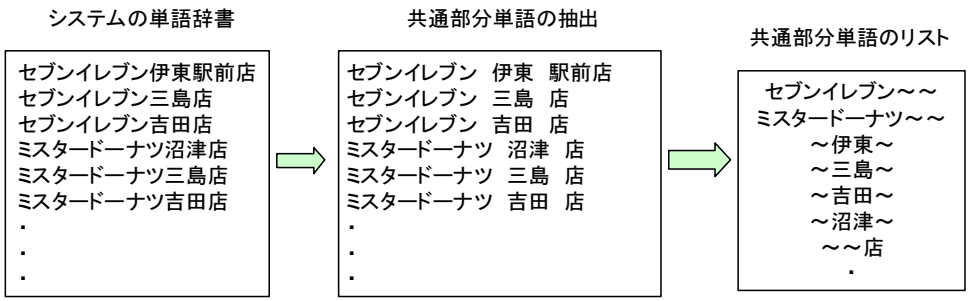


図 1 共通部分単語の抽出過程の流れ
Fig. 1 Flow of extracting common word parts

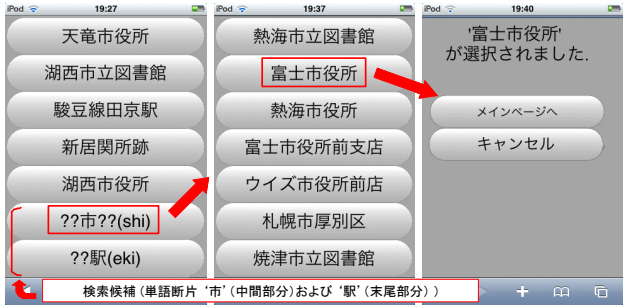


図 2 複数の候補を表示する GUI 併用の音声入力インタフェース
Fig. 2 Example of GUI-assisted speech interfaces displaying multiple candidates

本研究では、一度に画面へ表示される候補リストの長さを 10 に設定しており、共通部分単語の最低共起頻度も 10 に設定した。その結果、先頭・中間・末尾部分を合わせて 2,051 個の共通部分単語が自動的に抽出されている。なお、語彙単語の形態素解析には茶筌 [6] を利用している。

2.2 単語断片を含む候補リストの構成

複数候補を提示する典型的な方法は、認識尤度に従って出力された N-best 仮説を順位に従って出力することである [7]。提案手法における候補リストは、これらの単語候補に加え、信頼度スコアによって可変的な数の検索候補も一緒に提示する。提案するインタフェースシステムの表示例を図 2 に示す。

提案手法における候補リストの構成については様々な方法が考えられ、「どの単語候補・検索候補を挿入するか」と「単語候補と検索候補の割合をどのように決めるか」が最も重要

Vol.2011-SLP-89 No.25
2011/12/20

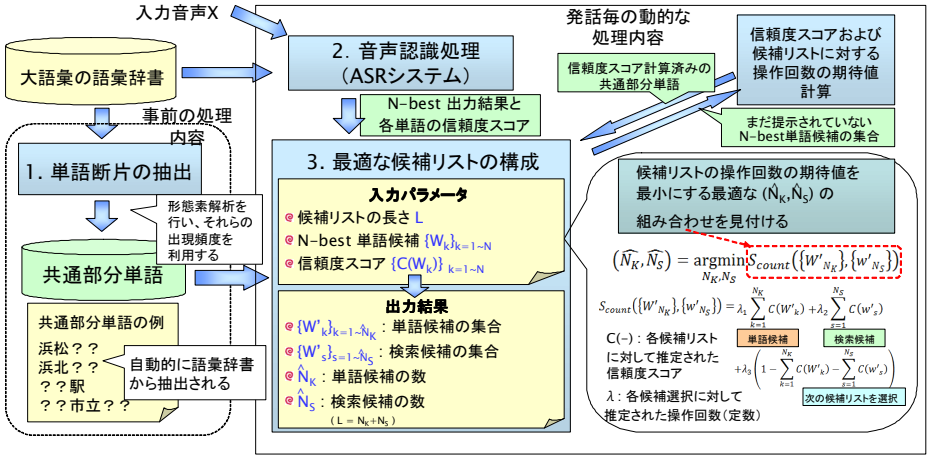


図 3 提案するシステムにおける候補リスト構成の仕組み
Fig. 3 A structure of making a candidates list in a proposed system

な点となる。まず、一度に表示する候補リストの長さを L とし、N_K 個の単語候補および N_S 個の検索候補を含むとする ($L = N_K + N_S$)。このとき、効果的に N_S や提示する候補を決定する方法として、ユーザが入力成功までに必要とする候補リストの表示回数（=ユーザの操作回数）の期待値を最小にするという考え方に基づき、それを反復的に実現するアルゴリズムを提案している [5]。また、ユーザの操作回数の期待値を求めるとき、候補リストのおける単語候補および共通部分単語の信頼度を利用し、最適な候補リストの中身を見付ける。図 3 に提案するインタフェースシステムにおける候補リストの構成法の概要を示す。

本研究では、ASR システムの認識結果から各単語候補 W_i に対する信頼度スコア $C(W_i)$ を計算し、信頼度スコアが高い順に候補リストに挿入する。検索候補に対しても同じように信頼度スコアが高いものを優先して表示する方法が考えられる。信頼度の推定方法には様々な方法が提案されているが [8]、我々は ASR システムより順位付けされた単語候補の集合 $\mathbf{W}_R = \{W_1, W_2, \cdots, W_n\}$ とそれらの尤度情報のみ利用し、予め抽出した共通部分単語 w_j の信頼度スコアを式（1）のように定義する。

$$C(w_j) = \frac{\sum_{i=1}^{n} P^{\alpha}(X|W_i)\delta(W_i, w_j)}{\sum_{i=1}^{n} P^{\alpha}(X|W_i)} \quad (1)$$

ここで、$P^{\alpha}(X|W_i)$ は入力音声 X に対する第 i 位の認識候補 W_i の重み付き尤度スコアである。$\delta(W_i, w_j)$ はデルタ関数であり、共通部分単語 w_j が単語 W_i の部分である場合は $\delta(W_i, w_j) = 1$，そうでなければ $\delta(W_i, w_j) = 0$ とする。

次に、候補リストに対する操作回数の期待値を計算する。例えば、0.5 の事後確率（信頼度スコア）を持つ単語候補は、0.5 の確率で正解としてユーザに選択されると推定でき、単語候補の選択は入力成功となるので操作回数は 1 になる。これに対し、0.5 の信頼度スコアを持つ検索候補は 0.5 の確率で絞り込みの操作を含め 2 回以上の表示回数を要する。この場合、N_K 個の単語候補 $W'_k(k=1,2,\cdots,N_K)$ と N_S 個の検索候補 $w'_s(s=1,2,\cdots,N_S)$ で構成された候補リストにおいて，ユーザが必要とする表示回数の期待値を次式のように推定できる。

$$S_{count}\{Candlist_t(N_K, N_S)\} = \lambda_1 \sum_{k=1}^{N_K} C(W'_k) + \lambda_2 \sum_{s=1}^{N_S} C(w'_s) + \lambda_3\{1 - \sum_{k=1}^{N_K} C(W'_k) - \sum_{s=1}^{N_S} C(w'_s)\} \quad (2)$$

ここで、$C(W'_k)$ 及び $C(w'_s)$ は、各候補 W'_k や w'_s の信頼度スコアを表す。また、λ_1 は単語候補 W'_k の選択を想定したとき、その操作を含めて入力成功に必要とされる操作回数（=1)、λ_2 は検索候補 w'_s の選択、λ_3 は次のリストの選択を想定したときの操作回数（λ_2,$\lambda_3 \geqq 2$）である。式 (2) から t 回目の表示において、単語候補を N_K 個、検索候補を N_S 個含む候補リスト $Candlist_t(N_K, N_S)$ に対する操作回数の期待値 $S_{count}\{Candlist_t(N_K, N_S)\}$ が推定できる。本研究では、以前の予備的な検討 [5] に基づき、$S_{count}\{Candlist_t(N_K, N_S)\}$ を最小にする $\hat{N_K}$ と $\hat{N_S}$ の割合を近似的に決定するアルゴリズムとして、図 4 に示す方法を用いる。この方法では、λ_1〜λ_3 は定数として、1 回目の一覧表示内容から確定的に候補内容を決定するアルゴリズムであり、4.3 節で触れているように、効率良く候補リスト生成が可能である。

図 4 のアルゴリズムによって、システムが t 回目の候補リスト提示時点において判断した、操作回数の期待値を最小にする $(\hat{N_K}, \hat{N_S})$ の組み合わせが見つかる。

3. シミュレーション実験によるインタフェースの性能評価

この章では，2 章で述べた候補リスト生成に基づく音声インタフェースシステムとベースラインシステムの性能を比較するシミュレーション評価について述べる。シミュレーション実験は様々な条件下で行われたが、紙面の関係でオンライン実験と関連がある一部のみを示す。

Vol.2011-SLP-89 No.25
2011/12/20

1. $m \leftarrow 1(t=1$ のとき) または $m \leftarrow 0(t>1$ の時)。
2. $N_K \leftarrow m, N_S \leftarrow L-m$。
3. 認識結果に基づき、単語候補の集合を構成する。
4. $CandList(N_K, N_S) \leftarrow \emptyset$。
   (a) $CandList(N_K, N_S) \leftarrow CandList(N_K, N_S) \cup W_i$
   for $i = 1, 2, \cdots, N_K$: 単語候補を追加。
   (b) 選択された単語候補 W_i を単語候補の集合から外し、
   集合内の順位付けをやり直す
   (c) (c1) から (c3) を $N_S = (L-m)$ 回実行する。
   (c1) 共通部分単語 $w_j (j = 1 \sim CW_{max})$ に対して,
   信頼度 $C(w_j)$ を計算する。
   (c2) $CandList(N_K, N_S) \leftarrow CandList(N_K, N_S) \cup w_{\hat{j}}$,
   ここで $\hat{j} = \underset{j}{\mathrm{argmax}}(C(w_j))$。
   (c3) 検索候補 $w_{\hat{j}}$ を部分として持つ単語候補 $(\forall W_i \supset \hat{w_j})$ を
   単語候補の集合から外し、集合内の順位付けをやり直す
   (d) $S_{count}\{Candlist_t(N_K, N_S)\}$ を計算する。
5. $m \leftarrow m+1$。
6. $m \leqq L$ であれば 2 に戻る。
7. $CandList(\hat{N_K}, \hat{N_S})$ を候補リストとして出力。
   ここで $(\hat{N_K}, \hat{N_S}) = \underset{N_K, N_S}{\mathrm{argmin}} S_{count}\{Candlist_{t+1}(N_K, N_S)\}$

図 4 候補リスト構成のアルゴリズム
Fig. 4 An algorithm of making a candidates list in a proposed system

3.1 タスクおよびデータ

本研究で利用する施設名入力タスクは、東海地方の 5 つの県における様々な施設名を含んでおり、合計 42,693 単語で構成されている。また、前述したように共通部分単語を辞書から抽出し、2,051 個（先頭部分 568 個、中間部分 1341 個、末尾部分 367 個）の共通部分単語が得られた。

実験には 4 名の男性話者が発話した孤立単語 190 発話を利用しているが、5 県の中で静岡県の施設のみを発話している。これらの音声データは雑音が殆どないクリーンな環境で収録されており、シミュレーション実験には複数通りの SNR レベルで定常雑音を加えて利用した。なお、ASR システムは SPOJUS++[9]、音響モデルとしては 124 種類の音節カテゴリ単位の HMM を利用し、N-best 仮説数はオンライン実験の仕様に合わせた 200-best である。

3.2 実験方法と評価尺度

シミュレーション実験では予め収録された音声データを利用し、ユーザがインタフェースを操作することを想定している。候補リストの中に発話した単語が単語候補として含まれていればその単語候補を選択し、単語全体は見つからないが単語の部分が検索候補として含まれている場合には検索候補を選択する。もし、候補リストの中に単語の全体や部分も現れていない場合には次の候補リストを選択し、このような行動を許容する操作回数の上限に達するまで続ける。

一方、実験における評価尺度としては、入力成功までの操作回数に注目し、どのぐらい入力効率が改善できたのかを表す「平均操作回数改善率」を用いる。ここで操作回数とは、最初の候補リストを含んで発話単語全体の入力に必要な候補リストの参照回数を意味する。図 2 の場合、2 回の操作回数で発話単語の全体の入力に成功している。平均操作回数改善率 IR は次式のように求めることができる。

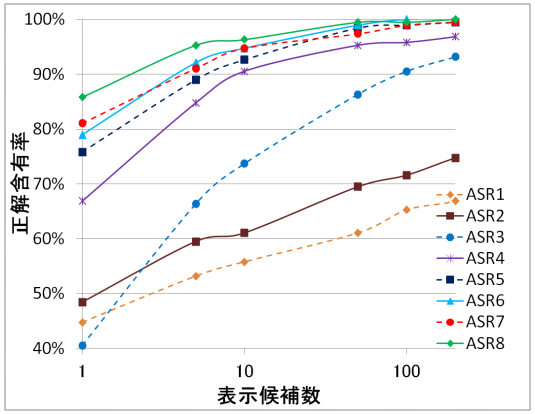
$$IR = \frac{1}{U} \times \sum_{i=1}^{U} \frac{Counts(Base) - Counts(Prop)}{max\{Counts(Base), Counts(Prop)\} - 1} \quad (3)$$

ここで、U は発話数、$Counts(Base)$ 及び $Counts(Prop)$ はそれぞれベースラインと提案法の操作回数である。但し、式 (3) において、ベースライン手法及び提案手法の操作回数が共に許容する操作回数の上限を超える場合、及びどちらも 1 回の操作で成功する場合（分母が 0 になるので）には 0 にする。例えば、入力成功までベースラインは 4 回、提案法は 2 回の操作回数が必要だったとすると、$(4-2)/(4-1) = 2/3$ となり、操作回数の改善が約 67%できたといえる。この改善値は提案法の操作回数がベースラインシステムよりも多い場合にはマイナスになり、必ず -1 から 1 の範囲の値を取る。

3.3 実験結果

実験で評価するインタフェースシステムは、複数候補を ASR システムからの N-best 認識結果に基づいてそのまま一定個数ずつ順に提示するベースラインシステムと、2 章で紹介したアルゴリズムで候補リストを構成する提案法のシステムの 2 種類である。また、式 (2) における λ_2, λ_3 は、0.1 刻みで行った予備実験から最適な組み合わせとして判断された $\lambda_2 = 2.8, \lambda_3 = 4.1$ に設定した。候補リストの長さ L は 10、最大に許容する操作回数は 10 回である。

Vol.2011-SLP-89 No.25
2011/12/20



(a) 各ASRシステムにおける表示候補数と正解含有率

ASRシステム	SNR	SSの使用	音響モデル
ASR1	15[dB]	未使用	クリーン条件
ASR2	10[dB]	使用	クリーン条件
ASR3	0[dB]	未使用	マッチ条件
ASR4	0[dB]	使用	マッチ条件
ASR5	5[dB]	未使用	マッチ条件
ASR6	5[dB]	使用	マッチ条件
ASR7	20[dB]	未使用	クリーン条件
ASR8	15[dB]	使用	クリーン条件

(b) 各ASRシステムの認識条件

図 5 シミュレーション実験における認識条件及び認識精度
Fig. 5 Recognition accuracy on simulation experiments and different ASR conditions

まず、シミュレーション実験に用いた認識条件と各条件に対する認識精度を図 5 に示す。認識条件は合わせて 8 種類となり、各システムは定常雑音のレベル、雑音抑圧 (Spectral Subtraction)、音響モデルの学習条件などが異なる。図 5 におけるマッチ条件とクリーン条件とは、それぞれ同じ雑音条件の音声で学習したものと雑音無しの音声で学習したものを意味する。また、図 5 のグラフは表示候補数 n に対して、第 n 位までの候補の中に正解の発話単語が含まれている割合を示している。認識条件によって正解含有率やその変化は異なっているが、全体的に見れば第 1 位の正解率が低下している状況では、表示候補数の増加に対して正解含有率の改善の割合が小さいことが分かる。

表 1 は、各 ASR システムに対するシミュレーション実験の結果である。この結果を見れば分かるように、どの条件においても提案するインタフェースシステムは入力効率を改善できることが分かり、特に認識精度が低いほど高い有効性を示している。例えば、ASR1 において平均操作回数改善率が 0.1386 であることは、発話単語を探すまでの全体的な操作回数が 13.86%減少したことを意味する。

一方、操作回数の改善ができる場合の平均操作回数改善率は、ベースラインシステムでは正解の単語が最初のリスト（上位 10 位）に含まれておらず、且つ N-best 認識結果（ここでは $N = 200$) の中には入っている場合に限って操作回数の改善率を計算した結果である。表 1 の結果を見れば、提案手法は改善できるケースに対しては入力効率を大きく改善して

表 1 シミュレーション評価実験の結果

認識条件	1-Best 認識率	平均操作回数改善率	操作回数の改善ができる場合の平均操作回数改善率（発話数）
ASR1	44.74%	0.1386	0.3497 (21/190)
ASR2	48.42%	0.0409	0.2989 (26/190)
ASR3	40.53%	0.0509	0.3155 (37/190)
ASR4	66.84%	0.0146	0.3138 (12/190)
ASR5	75.79%	0.0163	0.2376 (13/190)
ASR6	78.95%	0.0182	0.3452 (10/190)
ASR7	81.05%	0.0080	0.3914 (9/190)
ASR8	85.79%	0.0061	0.3095 (7/190)
平均	65.26%	0.0242	0.3202 (16.88/190)

いることが分かる。紙面の関係で載せられなかったが、候補リストの長さ L を 5，20 にした時も同じ程度の有効性が示された。

4. 実システムの被験者実験によるインタフェースの性能評価

本章では、実際に動作するシステムを用いて行った被験者実験の結果について述べる。

4.1 インタフェースシステムの実装

被験者に対する実験のため、本研究で比較しているベースラインシステムと提案法のシステムの実装を行った。インタフェースシステムはクライアント側とサーバ側に分かれ、ユーザはノートパソコンのディスプレイに表示されるウェブブラウザ基盤の GUI と、スタンドマイクを利用してインタフェースシステムを操作する。図 6 は同じ N-best 認識結果を用いたときのベースラインおよび提案手法によって作られた候補リストの画面表示例である（正解単語は「富士市役所」)。

図 7 が提案するインタフェースシステムの実装の仕組みであり、基本的なウェブブラウザだけで起動できるようになっている。サーバ側は入力音声に対する認識結果を出力し、2 章で述べた構成法に基づいて候補リストをウェブブラウザを通してユーザに提示する。

4.2 実験条件と実験方法

合計 8 人の被験者に対し、与えられた施設名のリストを入力させる実験を行った。8 人は全て大学生でありパソコンの操作に慣れているが、音声認識技術の利用経験に関しては 2 人を除いては利用経験が殆どなかった。各システムに入力する施設名リストは語彙辞書からランダムに抽出した 50 単語であり、事前に別の単語で練習した後システムごとに異なる単語セットを入力する。また、ベースラインシステムと提案法のシステムを利用する順番を入れ

Vol.2011-SLP-89 No.25
2011/12/20

(a) ベースラインシステムの候補リスト　(b) 提案するシステムの候補リスト

図 6　実装したインタフェースシステムの候補リスト提示例
Fig. 6　An example of candidates lists for each system using the same recognition result

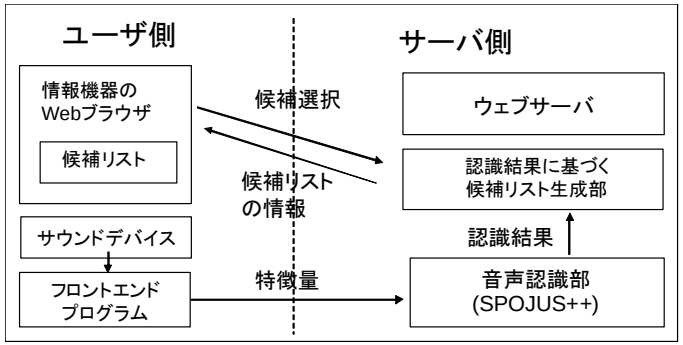
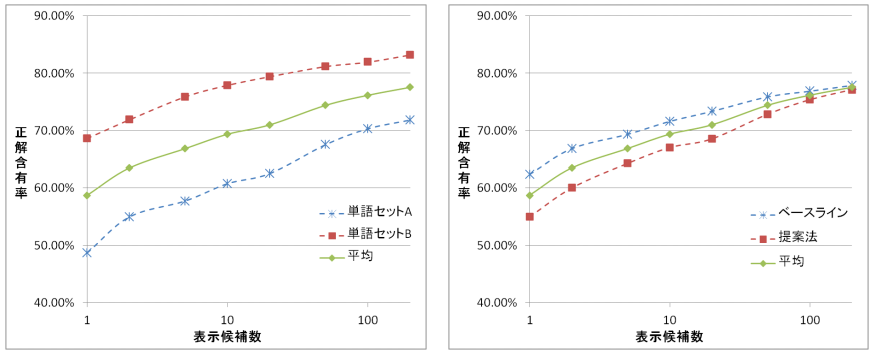


図 7　提案するインタフェースシステムの構成
Fig. 7　Construction of the proposed interface system

替え、システムに入力する単語セットも入れ替えを行い単語セットや実施順による影響を最小化した。

なお、ユーザに候補リストをどこまで参照するか、何回まで音声入力の段階からやり直すかについては自由に任せ、施設名の入力を諦めることも可能にした。音声入力からやり直して施設名を入力した場合、成功した結果をその発話に対する最終的な結果とする。

システムに関する実験条件はシミュレーション実験と一緒であるが、被験者実験ではVAD(Voice Activity Detection) 技術を用いて自動的に音声区間を推定する。また、実環境で認識精度が低下した状況を近似的に与えるため、入力した音声は指定された SNR レベル (=20[dB]) の雑音が重畳され、サーバ側に送られるようにした。候補リストの長さ L は 10、認識候補数 n は 200 であり、被験者実験に用いるタスクもシミュレーション実験と同様である。



(a) 単語セット別のASR出力時点の認識精度　(b) 操作システム別のASR出力時点の認識精度

図 8　被験者実験における認識精度
Fig. 8　Recognition accuracy on subjective experiments

4.3 実験結果

まず、2.2 節で述べた提案するシステムのアルゴリズム及びシミュレーションの実験条件は、システムの実装を踏まえた仕組みとなっており、被験者実験においてほぼ時間の遅延が発生されないように設計された。実験の結果、音声認識の結果が出力されてから候補リストの作成が終わるまでベースラインシステムでは約 0.02 秒、提案法のシステムでは普通の提示は約 0.52 秒、絞込みの候補リスト作成には約 0.01 秒がかかり、ほぼリアルタイムで候補リストを構成することができる。

次に、シミュレーション実験の結果と同じく、表示候補数と正解含有率の関係を図 8 に示す。図 8 の (a) は被験者毎に 2 つのシステムで個別に与えた 2 種類の単語セットに対する認識精度であり、(b) は比較する 2 種類のシステムそれぞれで実際に与えた単語セットに対する認識精度である。この結果を見れば分かるように、単語セットの規模が小さいため、音声認識システム (ASR) の N-best 出力時点において、ランダムに抽出した単語セット間に大きな認識精度の差が表れている。また、少ない候補数では、ベースラインシステムと提案法のシステムを操作するときも認識精度の差が現れた。この結果は、後で紹介するアンケートによる主観的評価にも影響を与え、提案法のシステムに対して不利な状況となっている。ユーザ別の認識精度は更に変動が激しく、1-best で 90%以上認識できた人もいれば、1-best で 30%ほどしか認識できなかった人も存在した。

次に、提案法の有効性の評価に関する集計結果を表 2 に示す。理想的には、音声入力後

表 2 被験者実験における提案法の有効性

被験者番号	1	2	3	4	5	6	7	8	合計	全発話に対する割合
有効性を示せる発話数	2	4	10	7	1	7	0	9	40	10.05%
有効性を示した発話数	1	3	3	4	0	1	0	1	13	3.27%
ユーザが誤った操作を行った数	0	0	5	3	0	5	0	8	21	5.28%
効果が無かった発話数	1	1	2	0	0	1	0	3	8	2.01%

表 3 被験者実験における主観的評価

被験者番号	1	2	3	4	5	6	7	8	平均
ベースラインの全体的な性能	3	4	4	5	2	2	4	3	3.38
提案法の全体的な性能	4	3	2	2	1	4	4	2	2.75
入力効率の差を感じた程度	4	5	4	5	-	4	-	4	4.33
検索候補が役に立った程度	4	4	2	4	-	4	-	4	3.67
検索候補のデザイン	4	4	3	2	-	2	-	4	3.17

から提示される最初の候補リストに正解の単語が現れなかったが、N-best 認識結果の下位には正解が含まれているケースが提案法で有効性を示す可能性を持つ。表 2 における「有効性を示せる発話数」がそのケースであり、「有効性を示した発話数」は実際に入力効率が改善したケース数である。しかし、シミュレーション実験と異なり、ユーザは正しくインタフェースを操作する保証がないため、様々な誤操作を行う可能性がある。本研究におけるインタフェースシステムはデザインやシステムの機能に対しては検証が不十分であったため、高い誤操作の可能性を持っており実験でもそのような結果が現れた。有効性を示せなかったケースの原因としては「正解の候補を見落として諦める」、「候補リストの参照を途中で諦める」、「候補リストの参照や候補選択から前に戻り、操作回数を増やしてしまう」、「正しい検索候補を見落として諦める」などがあった。

結果的に、被験者実験における平均操作回数改善率は 0.014、有効性を示せるケースに限っても 0.138 になり、有効性が示せる発話数はシミュレーション実験と同様にあったものの、実際のユーザが操作を間違って有効性を示せなかったケースが多かった。

最後に、実験後に被験者に対して行ったアンケートから提案するインタフェースシステムに対する主観的評価を行った。アンケートの質問としては、「ベースラインと提案法の全体的な性能はどうだったか」、「施設名入力の際、入力効率の差を感じたか」、「提案法の検索候補が役に立ったか」、「検索候補のデザインは分かりやすかったか」をユーザに与え、1(Negative)〜5(Positive) の 5 段階で評価をしてもらった。その結果を表 3 に示す。

表 3 は、各ユーザによる主観的評価の結果であり、5 に近いほど肯定的な結果であるといえる。但し、一部のユーザは検索候補を使う機会が無かったため、評価できなかった部分もあった。アンケートの結果から注目するところは、ベースラインと提案法システム操作時の認識精度の差 (1-best で約 10%程度) をユーザも感じ、システムの性能に対する評価につながったことである。一方、検索候補による変化はユーザも感じており、誤操作が多かったものの役に立ったと評価した人が多かった。しかし、検索候補のデザインの評価は少し低くなっており、実験結果を反映している。

被験者実験の結果をまとめると、以下のように整理できる。

- シミュレーション実験と同じく、提案するインタフェースシステムが有効性を発揮する機会はあった。
- インタフェースシステムのデザインなどの影響で、ユーザの間違った操作が多く発生し、実際の有効性が縮小されてしまった。
- インタフェースのデザイン面での改善で、提案するシステムの有効性は増加でき、ユーザビリティも向上するはずである。

5. ま と め

本稿では、大語彙の単語音声入力タスクに対し、複数候補の提示及び選択を併用する音声入力インタフェースにおける新しい候補リストの構成法について提案し、シミュレーション実験及び簡単なオンライン実験を通して雑音環境下における入力効率の改善の有効性を示した。複数候補をリストとして提示する際、画面サイズによって限られた数の候補数を参照すると、正解の単語が候補リストの出力結果において低い順位に位置する場合、多くのリストを参照するか発話を再度行う方法でしか対処できない。しかし、認識精度が低い状況では発話を再度行ってもうまく認識できない可能性が高く、被験者実験でもそのような傾向が現れ、結局入力を諦めるケースが多かった。

提案するシステムは単語断片で認識結果の絞込みを行い、より少ない操作で単語を入力できるようにする。単語断片の抽出や候補リストの構成は自動的に行われ、実験によって様々な認識条件の変化にも頑健であることが示された。シミュレーション実験では全体的に約 2.4%の平均操作回数の改善が行われ、N-best の上位に候補が含まれない場合では約 32%まで改善できた。一方、インタフェースシステムを実装し、被験者に対して行ったオンライン実験からは、有効性は低かったものの、インタフェースのデザイン面での改良で入力効率の改善が期待される。

今後の課題としては、提案するアプローチを発展させて他の音声応用システムの一部とするなどより実用的な面で使えるようにすることや、デザイン面での改善によって、入力効率

Vol.2011-SLP-89 No.25
2011/12/20

の改善が更にできるかを検証することなどが挙げられる。

参 考 文 献

1) 趙　國，宮山章子，山下洋一，“ N-best 音声認識における認識スコアを利用した候補提示数の決定 ”, IEICE Trans., Vol. J88-D-II, pp.1003-1011, (2005)
2) 北岡教英，押川洋徳，中川聖一，“ 孤立単語認識と連続基本単語認識の併用に基づく組織名の音声入力インタフェース”, Proc. of INTERSPEECH 2005, pp.1201-1204 (2005)
3) 押川洋徳, 北岡教英, 中川聖一：“ 音節 N-gram と単語辞書併用による姓名入力インタフェース”, 電子情報通信学会技術研究報告, Vol.2003-SP-103, No.519, pp.175-180 (2003)
4) 大内一成, 若木裕美, 屋野武秀, 住田一男, 土井美和子：“ 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”, 情報処理学会論文誌 Vol. 51, No. 3, pp.846-855（2010)
5) 張　用起, 甲斐充彦, 王　龍標：“ 単語断片を含む複数候補の動的構成によるマルチモーダル単語入力インタフェース”, 日本音響学会秋季研究発表会講演論文集, 1-Q-24 (2010)
6) 形態素解析ツール「茶筅」, http://chasen-legacy.sourceforge.jp/
7) 藤原敬記, 伊藤敏彦, 荒木健治, 甲斐充彦, 小西達裕, 伊東幸宏：“ 認識信頼度と対話履歴を用いた音声言語理解手法”, 電子情報通信学会論文誌, Vol. J89-D, No. 7, pp. 1493-1503 (2006)
8) H. Jiang, “ Confidence measures for speech recognition : A survey”, Speech Communication, Vol. 45, pp.455-470 (2005)
9) 藤井康寿, 山本一公, 中川聖一：“ 大語彙連続音声認識システムの改善：SPOJUS++ ”, 第 4 回音声ドキュメント処理ワークショップ講演論文集, pp. 1-10 (2010)