情報処理学会論文誌　Vol. 51　No. 2　324–333 (Feb. 2010)

センサ駆動ハンドヘルド型音声認識入力方法の提案 ——センサを用いた発話動作検出手法

大内一成[†1]　土井美和子[†1]

音声認識入力を日常的に使うには，リモコンにマイクを搭載するようなハンドヘルド型が望ましい．しかし，ハンドヘルド型マイクでは，口元とマイクの位置を一定距離に確保できないためヘッドセット型マイクに比べて認識精度が劣ること，音声認識の開始と終了をユーザがボタン操作で指示する必要があることが課題となっている．その解決策として，発話距離については距離センサ，発話動作検出については加速度センサの活用を検討した．実験の結果，距離センサの方が，音声認識精度，発話動作の検出ともに有効であることが分かった．しかし，距離センサ単独では，発話意図のない状況での誤検出と，消費電力が問題である．そこで，両センサを組み合わせたセンサ駆動ハンドヘルド型音声認識入力方法を提案する．その結果，発話動作の検出エラー率は 4.8%，音声認識精度は 82.4%と，プッシュトーク方式での問題である認識トリガボタン押し忘れ頻度，プレストーク方式での問題である音声認識精度と比べて，それぞれ 3.3 ポイント，6.9 ポイント改善することができた．提案方式は，特に高齢者など機器操作が不得意なユーザにとって効果が大きいことも確認できた．

Sensor-driven Speech Recognition Input Method by Handheld Device — Utterance Movement Detection Method Using Sensors

Kazushige Ouchi[†1] and Miwako Doi[†1]

It is preferable to use a handheld device equipped with a microphone for speech recognition input in our daily lives. However, when we use handheld devices for speech recognition, there are two problems to be solved. One is that the recognition accuracy by using handheld devices is less than that by using headsets because it is difficult to keep an appropriate distance between the microphone and the mouth. The other is that it is troublesome for users to indicate a section of speech recognition by button press action. We considered using a distance sensor for measuring a distance between them and an accelerometer for utterance detection respectively. The experiment showed that using the distance sensor was more effective for both speech recognition accuracy and utterance detection. But there still existed problems related to false detection and power consumption for daily use. Then, we proposed a sensor-driven speech recognition input method by using both sensors. As a result, we improved the error rate of utterance detection to 4.8% and the accuracy of speech recognition to 82.4%. They were superior to conventional methods such as push-talk and press-talk on the handheld devices. We also confirmed that the proposed method was especially effective for users such as elderly people who were not generally very good at using electric appliances.

1. はじめに

音声認識入力は，自動車運転中など両手が使えない状況での機器操作[1)]や，パソコンを用いた特定業務での作業効率向上[2),3)]などで活用されている．

一方，機能数の増加，ネットワーク化などにより，日常的に扱う一般的な家電機器の操作が複雑になってきている．特に高齢者など，機器操作を得意としないユーザ層にとってその影響は大きい[4),5)]．

また，情報家電はその役割が変わりつつある．たとえば，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を搭載したデジタルテレビの普及などにともない，テレビの役割は，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と考えられる．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，直接文字を入力する機会が増える．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 1 文字ずつ入力するという方式では，不十分であるし，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も，日常生活における音声認識による文字入力の重要性が増すと考えられる．

音声認識入力を日常的に使うとする場合，その入力方法としては，現状で次の 3 種類の方法があげられる．

A) ヘッドセット型マイクを装着
B) 操作対象機器にマイクを搭載（機器搭載型マイク）
C) リモコンにマイクを搭載（ハンドヘルド型マイク）

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

A) のヘッドセット型マイクは，現状では最も認識精度が高い．ヘッドセット本体に音声認識エンジンを内蔵し，直接外部機器を操作する装置なども開発されている[6]．ヘッドセットの利用は，ハンズフリーであるので，両手が使えない状況では有効である．また，パソコンを用いた特定業務などにおいても，ヘッドセットを介した音声認識入力とマウス/キーボードなどによる操作を組み合わせる UI（ユーザインタフェース）[7] により，作業効率の向上が期待される．しかし，日常生活で常時ヘッドセットを装着しつづけるのは，拘束性が強く，受け入れがたい．

B) の機器側にマイクを搭載する方法（機器搭載型マイク）は，ユーザは身体に何も装着する必要がないという利点がある．しかし，機器からの距離，つまりマイクからの距離が遠くなることにより環境ノイズの影響を受けやすくなり，認識精度が悪化するという問題がある．この問題を解決するため，マイクロフォンアレイを用いて話者の方向を推定し，ノイズの影響を軽減し音声認識精度を改善するといった試み[8),9)] もなされている．しかし，日常生活環境下で実際に使われるレベルの性能を確保するために，マイクの数を増やす必要があるなど現状ではまだ難しい．

これらに対し，C) のハンドヘルド型マイクは，ユーザの手元で音声を入力できるので，機器搭載型マイクよりは音声認識の精度が高く，ヘッドセット型より拘束感を少なくできる．また，テレビ視聴など，部分的に手による操作が可能な日常生活のシーンにおいては，面倒な文字入力などを音声認識で行い，ボタンによる操作と適切に組み合わせるのが現実的な使い方と考える．その面でも，1 つのデバイスで音声認識入力とボタン操作を行うことができるハンドヘルド型が望ましい．

本論文では，まずハンドヘルド型音声認識入力の課題を，実験を基に明らかにする．次に，それを解決するための方法としてセンサ駆動ハンドヘルド型音声認識入力方法を提案し，その有用性を評価する．

2. ハンドヘルド型音声認識入力の課題

ハンドヘルド型マイクはヘッドセット型マイクに比較し，拘束性が少ない半面，音声認識精度に問題がある．その問題を明らかにするために，まずヘッドセット型マイクと音声認識精度を比較する予備実験を行った．その結果を基に，ハンドヘルド型マイクでの口元距離と音声認識精度との関係を調べる予備実験を行った．また，音声認識精度向上に必要になると考えられる発話開始/終了指示のボタン押し忘れについても予備実験を行った．

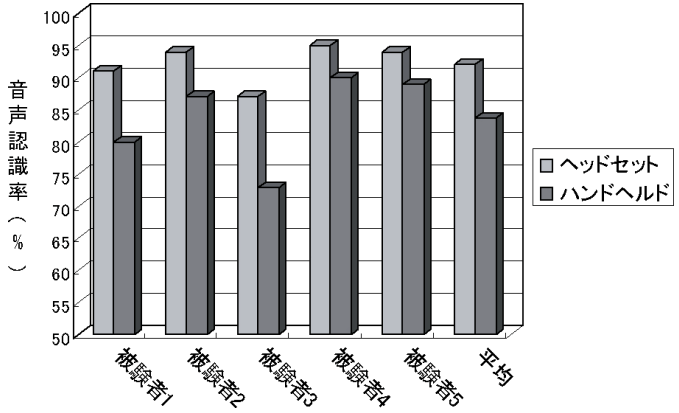

図 1　ヘッドセット型マイクとハンドヘルド型マイクの音声認識精度
Fig. 1　Speech recognition accuracy of headset and handheld microphone.

2.1 ヘッドセット型マイクとハンドヘルド型マイクの音声認識精度の比較

まず，ヘッドセット型マイクとハンドヘルド型マイクとの，音声認識精度の違いを評価した．本論文の全実験で使用した音声認識エンジンは，番組名，出演者名など 7,000 語彙が登録された孤立単語認識エンジンである．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．

被験者 5 名に，同一の 100 種類の単語（番組名および出演者名）をヘッドセット型マイク，ハンドヘルド型マイクのそれぞれを用いて連続で発話してもらい，録音した．全発話の録音完了後，手作業で発話区間を切り出して音声認識処理を施し，正しい音声認識結果が尤度上位 5 件に入るかどうかを評価した．なお，一般的な使用環境での特徴を把握するため，本論文の全実験は，家庭のリビングを模した実験室内で，被験者は椅子に座った状態で実施した．ドアの外は廊下となっているため，通行人の足音や話し声が聞こえる環境である．実際の家庭のリビングとは異なるが，雑音環境的にはほぼ同一レベルである．実験結果を**図 1** に示す．

5 名すべての被験者で，ハンドヘルド型マイクはヘッドセット型マイクよりも音声認識精度が低かった．全被験者の平均値は，ヘッドセット型マイクが 92.2%，ハンドヘルド型マイ

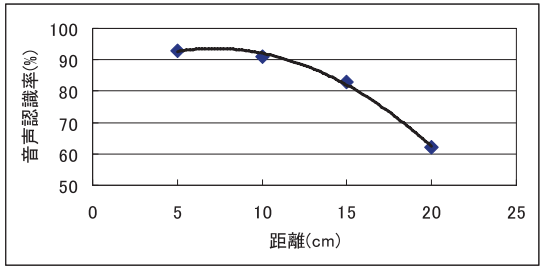

図 2　マイク–口元間の距離と音声認識精度
Fig. 2　Speech recognition accuracy with distance between the microphone and the mouth.

クが 83.8%であった（$p < 0.01$）．この差は，実験者の観察では，ハンドヘルド型マイクは発話時のマイクと口元の距離がヘッドセット型マイクよりも遠く，かつ一定となっていない点が原因と考えられる．次に，その影響について調査した．

2.2　マイク–口元間の距離と音声認識精度の関係

前節の実験と同一の実験環境において，ハンドヘルド型マイクを被験者の口元から 5 cm，10 cm，15 cm，20 cm の距離に順に固定し，それぞれの距離で同一の 100 単語の発話を録音し，前節同様，後から手作業で発話区間を切り出して音声認識率を確認した．被験者は 1 名で実施した．結果を**図 2** に示す．

10 cm 以内の距離で 90%以上（前節の同一環境のヘッドセットによる性能とほぼ同一レベル）の精度が確保できている．しかし，10 cm を超えると距離に応じて精度が劣化していく．これは，距離が遠くなるとマイクに入力される音声の強度が弱くなり，SNR（Signal to Noise Ratio）が低下することが原因と考えられる．よって，一定レベルの音声認識率を確保するためには，ハンドヘルド型マイクで音声認識を行う際に，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．

2.3　操作面での課題

ハンドヘルド型マイクでの音声認識入力に限らず，従来の音声認識入力システムでは，認識精度を高めるために，発話開始と終了の始終端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，プレストークと呼ぶ）や，音声認識の開始だけボタン押下で指示し，音声認識の終了は認識エンジンによる無音区間検出により自動的に行う方式（以下，プッシュトークと呼ぶ）などがある．

プレストークやプッシュトークでは，ユーザが明示的に始端（プレストークでは終端も）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あった．

その発生頻度を調査するために，60 歳以上の高齢被験者 6 名（男性 3 名，女性 3 名）に対して次の実験を実施した．前節での連続して発話する方法は操作面での課題が明確に観測しにくいため，本節ではテレビの番組検索を音声認識およびボタン操作を組み合わせて実施するアプリケーションに変更した．そのアプリケーション使用時のボタンの押し忘れ頻度を観測した．ここで被験者に課したタスクは，課題（番組名あるいは出演者名）を提示し，それに基づいて任意の検索キーワードを音声認識で入力して番組検索を行い，ボタン操作で提示された番組あるいは提示された出演者が出演している任意の番組の詳細情報を得るというもので，各被験者につき合計 10 タスクずつ行った．その結果，全発話における押し忘れ発話の発生頻度は，15.3%であった．また，適切なタイミングでボタンを操作できず音声認識が正しく入力できない事例も観測できた．

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だれ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．

2.4　課題のまとめと対策案

以上のように，現状のハンドヘルド型マイクによる音声認識入力における課題として，以下の 2 点があげられる．

＜ハンドヘルド型マイクによる音声認識入力の課題＞

A)　マイク–口元間の距離がユーザ任せになってしまい，ヘッドセット型マイクに比べて音声認識精度が低い．
B)　音声入力時にボタン操作で音声認識開始/終了を指示させる方式（プレストーク，プッシュトーク）では，ユーザによっては押し忘れが発生するなど，適切なタイミングでの操作が難しい．

これらの課題の対策としては，以下の機能の必要性が示唆される．

＜課題を解決するための機能＞

A)　発話時にマイクと口元を音声認識にとって適切な位置にユーザを誘導する機能
B)　ユーザの発話動作を検知し，ボタン操作なしで適切なタイミングで発話の開始と終了を切り出し，音声認識する機能

これらの機能を実現するために，筆者らはそれぞれ距離センサ，加速度センサを活用することを検討した．その検討内容について次章以降で詳しく説明する．

3. センサの活用による課題解決の検討

3.1 距離センサによる発話動作検出

発話時のマイクと口元を音声認識にとって適切な位置にユーザを誘導する機能を実現するために，マイク近傍に距離センサを設置した実験用ハンドヘルド型マイクデバイスを試作した．音声認識にとって適切な距離は，図 2 の実験において 90%以上の認識精度が確保できる 10 cm 以内とした．そこで，10 cm を境に出力がデジタル的に変化する PSD（Position Sensitive Device）センサ GP2Y0D810Z0F（シャープ製）を距離センサとして採用した．

試作したハンドヘルド型マイクを使って自由に音声入力を行い，その際の距離センサの出力（口元が 10 cm 以内にあるかどうかの検出/非検出）を測定した．**図 3** は 4 回連続して発話したときの一例である．発話の際にハンドヘルド型マイクを口元に近づけることで，おおむね発話を検出できている．しかし閾値である 10 cm 前後の距離で発話した図 3 の 1 番目と 4 番目の発話では，出力が不安定になってしまっている（手のぶれなどにより 10 cm 前後を行き来するため）．これを改善するため，口元がマイクから 10 cm 以内であることを検出した後は，少なくとも一定時間（たとえば 1 秒間）音声認識入力を継続するようにした．つまり 10 cm を閾値として，10 cm 以内になったときに音声認識入力を開始，一定時間音声認識入力を継続した後，距離が 10 cm 以上になったときに認識終了とした．また 10 cm 以上になる以前に，無音区間が続くと（ハンドヘルド型マイクを持ったまま沈黙した状態になった場合），音声認識エンジン自身が音声認識を終了する．ここで，無音区間の検出は，音声認識開始から実際の発話開始までの雑音レベルの平均値を測定し，発話終了後，音声の入力レベルが発話開始前の雑音レベルに戻ることを基準として検出を行っている．よって，発話開始時が静かで，発話中あるいは発話終了後に雑音が増加した場合などは，この無音区間検出は正しく動作しない場合がある．この点は今後の課題である．

3.2 加速度センサによる発話動作検出

ボタン操作なしで適切なタイミングで音声認識の開始と終了を切り替える機能を実現するために，3 軸加速度センサによりユーザの発話動作を検出する実験用ハンドヘルド型マイクデバイスを試作した．加速度センサの各軸の方向は**図 4** のとおりとした．図 3 で 4 回連続して発話したときに同時に測定した加速度センサの出力を**図 5** に示す．

発話動作は鉛直方向の動作であるため，その Y 軸および Z 軸に特徴的な加速度の変化が

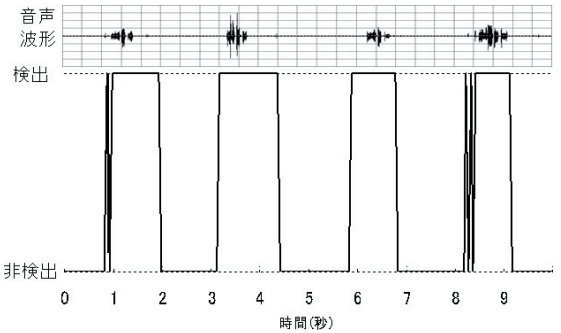

図 3　発話動作と距離センサの出力
Fig. 3　Output of the PSD sensor corresponding to utterances.

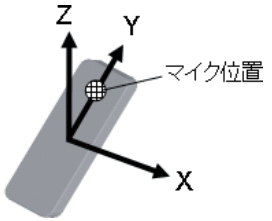

図 4　内蔵 3 軸加速度センサの軸方向
Fig. 4　Directions of built-in 3-axis accelerometer.

発生することが分かる．これは，重力加速度に対するデバイスの傾きが変化することに起因する．

発話動作の波形全体の特徴量により，発話動作を検出する認識手法も考えられる．ただし，音声認識の開始をユーザの発話動作の開始に応じて遅滞なく切り替えるためには，発話動作の波形全体を考慮する手法では発話動作開始の検出に遅れが発生することが想定される．

そこで，立ち上がり/立ち下がりが明瞭である Z 軸に着目して，−0.7 G を発話開始と発話終了の閾値とする単純な検出方法を採用した．つまり，放置状態では重力加速度が Z 軸のマイナス方向に加わることで −1.0 G を出力しているが，ユーザの把持・発話動作による傾き変化で −0.7 G を上回ると発話開始とし，−0.7 G を下回ると発話終了とした．ただし，距離センサの場合と同様に，発話開始から発話終了まで最低 1 秒間は音声認識を継続する．また，−0.7 G を下回る（発話動作を終了する）前に無音区間が続くと（ハンドヘルド型マ

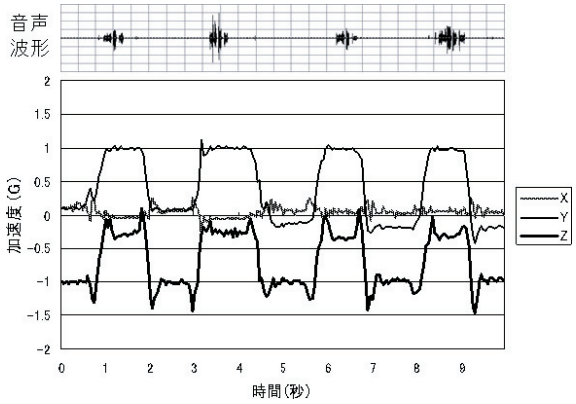

図 5　発話動作と加速度センサの出力
Fig. 5　Output of the accelerometer corresponding to utterances.

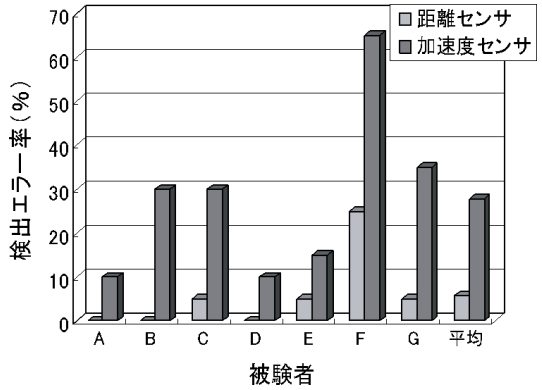

図 6　発話動作開始の検出エラー率
Fig. 6　Error rate of utterance detection.

イクを持ったまま沈黙した状態になった場合），3.1 節と同様に音声認識エンジン自身が音声認識を終了する．突発的に大きな加速度変化があった場合に誤検出しないようにするため，一定時間以上連続で閾値を超えるかどうかを判定するなどのノイズ対策を施したうえで，この閾値によって発話開始と終了を検知し，自動的に音声認識区間を検出できるようにした．

3.3　実験と考察

距離センサを活用しユーザの口元がマイクに近づいたときに音声認識を開始する機能と，加速度センサを活用しユーザがハンドヘルド型マイクを構えて発話する動作を検知して音声認識を開始終了する機能について，それぞれの特徴を明らかにするための実験を行った．

被験者は，20〜30 代の男女 7 名で，2 章の実験と同一の，番組名，出演者名など 7,000 語彙が登録された孤立単語認識エンジンを使用した．それぞれ 20 名分の人名を提示して，順にハンドヘルド型マイクに発話してもらった．センサが発話動作を検知し音声認識が開始されると，パソコンの画面に音声認識開始と表示され，被験者はその表示を確認して発話を行う．なお，それぞれの検出方式の違いについては，被験者には説明していない．

まず，距離センサと加速度センサのそれぞれについて被験者の発話開始動作検出時に，音声認識開始が正しく認識されなかったエラーについて検討する．ここでは，実験者が観測した被験者の発話動作回数に対して，発話した際に音声認識開始とならない検出漏れと，発話していないのに音声認識開始となる誤検出の両方をエラーとしてカウントし，式 (1) で検出エラー率を算出した．その結果を**図 6** に示す．

$$\text{検出エラー率（\%）} = \frac{\text{検出漏れ数} + \text{誤検出率}}{\text{発話動作回数}} \tag{1}$$

いずれの被験者も，加速度センサで発話動作を検出するよりも距離センサで口元とマイクの接近を検出して音声認識開始とした方が，エラー率が低いことが分かった．全被験者のエラー率の平均値は，距離センサの場合が 5.7%，加速度センサの場合が 27.9%であった．これは，加速度センサを用いた発話動作検出では重力加速度に対する傾きに閾値を設定して検出を行ったが，閾値として設定した角度に近い状態で把持し続けることが多い被験者でエラー率が高かった．

一方，それぞれの音声認識精度を**図 7** に示す．なお，ここでは発話動作の検出エラーにより音声認識が正しく動作しなかった場合は除外している．すべての被験者で距離センサを活用した場合の方が，加速度センサを活用した場合に比べて同等かそれ以上の音声認識率を示した．平均値では，距離センサの場合が 81.4%，加速度センサの場合が 73.6%であった．これはマイクと口元の距離を 10 cm 以内に確保して音声認識入力させる距離センサ活用方式が音声認識精度に良い影響を及ぼしていると考えられる．

以上により，発話動作検出および音声認識精度において，距離センサの使用が適していることが分かった．距離センサによる発話動作検出により，距離センサがない場合よりも平均値で 7.8 ポイント精度の高い音声認識を実現できた．ただし，距離センサによる発話動作検出は，ハンドヘルド型マイクが使用されずに机上などに置かれている場合でも，距離センサ

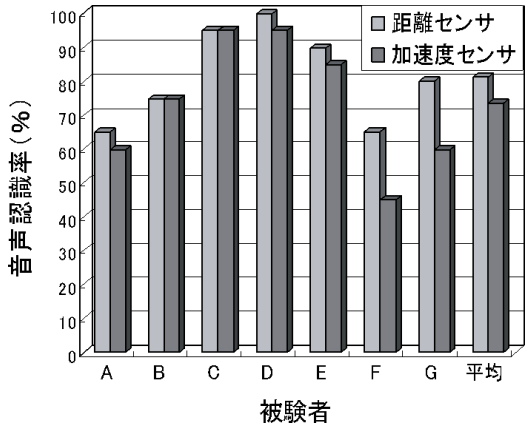

図 7　音声認識精度（距離センサ–加速度センサ）
Fig. 7　Speech recognition accuracy (PSD sensor vs. accelerometer).

付近に手などをかざすだけで音声認識開始としてしまう誤検出が実験準備中などに発生した．また，距離センサは消費電力が大きいため，つねにアクティブにしておくのは実運用上望ましくない．そこで，距離センサに比べると低消費電力で動作する加速度センサと組み合わせ，低消費電力化するとともに，ロバストな音声認識開始/終了検出を行うセンサ駆動ハンドヘルド型音声認識入力方法を提案する．

4.　センサ駆動ハンドヘルド型音声認識入力方法の試作

4.1　アルゴリズム検討

距離センサと加速度センサを活用してユーザの自然な発話動作を検出し，適切なタイミングで音声認識の開始，終了を自動的に切り替えるセンサ駆動ハンドヘルド型音声認識入力方法を開発した．その処理の概要を**図 8** のフローチャートに示す．

まず，消費電力が小さい加速度センサ（typ. 0.36 mA at 3 V）でデバイスがユーザに把持されるかどうかを重力加速度に対するデバイスの傾きをもとに判断する．ユーザに把持され，発話動作を検知した時点で初めて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–口元間距離が閾値（10 cm）よりも遠い場合は，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

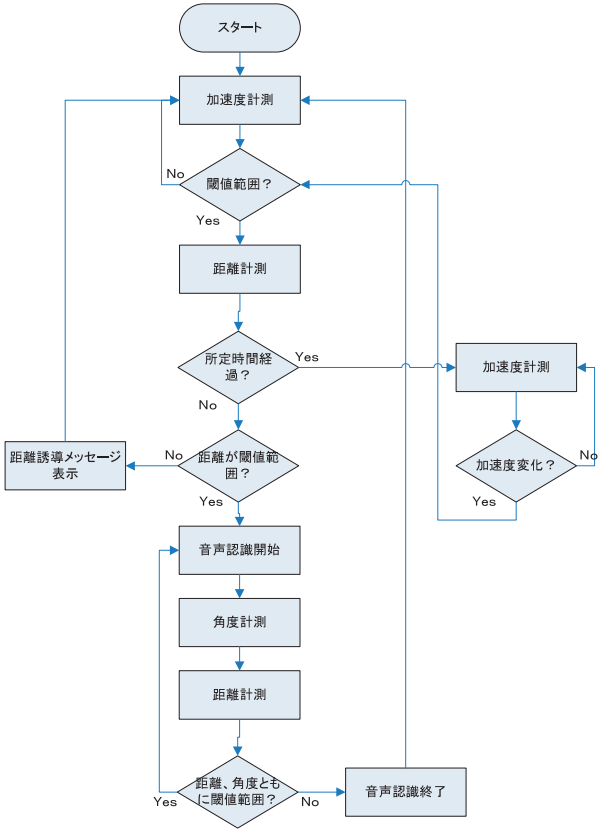

図 8　処理動作概要
Fig. 8　Outline of the algorithm.

をマイクに近づけてください）を表示する．メッセージが表示できる画面を搭載していない機器の場合は，ハンドヘルド型マイクに LED を搭載してその光で通知する方法や，音声合成の活用も考えられる．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–口元間距離が 10 cm 以内であることを検出した場合に，音声認識開始と判断し，アプリケーションへ音声認識開始コマンドを送信する．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．また，発話動作終了前に無音区間が続くと（ハンド

図 9　センサ駆動ハンドヘルド型マイク
Fig. 9　Developed handheld microphone equipped with sensors.

ヘルド型マイクを持ったまま沈黙した状態になった場合），音声認識エンジン自身が音声認識を終了する．

なお，発話動作に似た角度でデバイスが置かれた状態になった場合は，つねに距離センサがアクティブになってしまうため，加速度変化がない状態が所定時間を超えた場合はいったん距離計測を終了し，再びユーザがデバイスを把持して加速度が変化するまで距離計測を行わないこととした．

以上の処理により，ハンドヘルド型マイクが使われていないときに距離センサ付近にユーザの手などが近づいた場合の誤検出をなくすことができ，かつ消費電力の大きい距離センサを必要最小限に起動することで，消費電力を抑えることが可能となる．また，ユーザに対して適切な距離での音声認識入力を支援し，ユーザの自然な発話動作に応じて適切に音声認識開始/終了検出を行うため，ユーザに負担をかけずに高い音声認識精度を実現できるものと期待できる．

4.2　試作デバイス仕様

上述の機能を実現するために，距離センサ，加速度センサ，およびセンサの出力から発話動作，マイク–口元間距離の検知を行うマイクロコントローラ（MCU）などを搭載したセンサ駆動ハンドヘルド型マイクを試作した．外観を**図 9** に，概略仕様を**表 1** に示す．

MCU は，センサの出力から発話動作，マイク–口元間距離を検知し，それらに基づいて音声認識アプリケーションの処理動作を変更するコマンドをそれぞれ発行し，RS-232C 経由で外部機器へ送信する．今回の試作システムでは，パソコン上の音声認識アプリケーションに各コマンドを送信して音声認識エンジンの動作を切り替える．

表 1　試作デバイス仕様
Table 1　Specifications of the prototype.

距離センサ	PSD センサ GP2Y0D810Z0F（シャープ製）
加速度センサ	H34C（日立金属製）
MCU	PIC18LF1320-ML（Microchip Technology Inc.製）
マイク	モノラルマイク AT9642（オーディオテクニカ製）
外部入出力	マイク出力，RS-232C
操作／表示	操作ボタン：1，状態表示用 LED：2，電源スイッチ

表 2　性能評価実験結果
Table 2　Result of the performance evaluation experiment.

	距離センサのみ	加速度センサのみ	距離センサと加速度センサ
検出エラー率	5.7%	27.9%	0.7%
音声認識率	81.4%	73.6%	84.3%

4.3　性能評価実験

試作したセンサ駆動ハンドヘルド型音声認識入力方法の性能を評価するため，3 章の実験と同一の被験者 7 名（20〜30 代男女）に，3 章の実験と同様に，それぞれ 20 名分の人名を提示して，順にセンサ駆動ハンドヘルド型マイクに発話してもらった．結果（全被験者平均）を，距離センサのみ，加速度センサのみで実施した際の結果とあわせて**表 2** に示す．距離センサ，加速度センサの両方を活用することで，特に発話動作の検出エラー率を大きく改善できることが確認できた．

考察

音声認識率については，距離センサにより音声認識に適したマイクと口元の位置が良好な状態に確保できている．距離センサと加速度センサの両方を活用した場合は，距離センサのみの場合と同様に，加速度センサのみの場合よりも精度が良いことが確認できた．

一方，発話動作検出エラー率については，2 つのセンサの特徴を適切に活用した提案手法により，性能改善が確認できた．ただし，今回の被験者は比較的機器操作に慣れている 20〜30 代の被験者であったため，高齢者も含む一般的な被験者に対して，提案手法の性能を改めて検証する必要がある．

5. 従来手法との比較評価

開発したセンサ駆動ハンドヘルド型音声認識入力方法について，高齢者も含む一般的な（技術系の職種でない）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手法とあわせて，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．実験条件はこれまでと同一で，リビングを模した実験室内で，番組名，出演者名など 7,000 語彙が登録された孤立単語認識エンジンを使用した．音声入力には，試作した図 9 のセンサ駆動ハンドヘルド型マイクを用い，センサ駆動ハンドヘルド型音声認識入力方法と，従来手法であるプレストーク，プッシュトークによる音声認識入力が同一デバイスで実施できるように機能を実装し，同一タスク（人名 20 名分発話）に対する音声認識入力を行った．センサ駆動，プレストーク，プッシュトーク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．

被験者には，各入力方法によるタスク実施直前に使用方法を記載した説明資料のみを提示し，練習なしで直接タスクを実施してもらった．プレストーク，プッシュトークについては，エラー率として実験者が観測した被験者の発話動作回数に対する，ボタン押し忘れ数の割合を算出した．センサ駆動については 3 章の実験と同様に式 (1) によって算出した．実験結果として，全被験者平均と高齢被験者の平均（括弧内）を**表 3** に示す．

プッシュトークのエラー率（押し忘れ率）が 2.3 節で計測した 15.3%より低く出ている．これは，高齢者以外に 20〜30 代の被験者もあわせた平均となっているためで，高齢被験者だけの平均は 13.8%と，やはりかなり高い頻度で押し忘れが発生していることが確認できた．

音声認識率については，センサ駆動が全被験者の平均値としては最も良い数値となった．ただ，センサ駆動でのタスクを実施した後に，プレストーク，あるいはプッシュトークを実施した場合は，センサ駆動の後もセンサ駆動で指示された位置（マイクから 10 cm 以内）での発話となる場合が多く，距離センサによる適切な口元位置支援機能の効果を確認しにくかった．しかし，全タスクの最初の音声認識率だけピックアップすると，プレストークが 74.3%，プッシュトークが 76.4%，センサ駆動が 80.0%である．つまり距離センサによる適切な口元位置支援機能の効果が読み取れる．さらに，高齢被験者だけの音声認識率を比較すると，プレストークが 62.1%，プッシュトークが 71.3%，センサ駆動が 77.3%と，センサ駆動による音声認識率改善の効果が大きいことが分かる．

以上をふまえ，プレストーク，プッシュトーク，センサ駆動のそれぞれの音声認識入力方法について，本実験の結果から分かった特徴を以下に示す．

表 3　従来手法との比較実験結果
Table 3　Experimental result for comparing conventional technique and proposed technique.

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

- プレストーク

エラー率（押し忘れ率）は最も低かった．これは，発話している間だけボタンを押し続けるため，操作に対する負荷が他の 2 つより高く，被験者にはボタン操作に対する意識が強く働き，押し忘れが少ないものと推測できる．実際に実験後の主観アンケートでもそのような感想を持つ被験者が多かった．

一方，音声認識率は最も悪い．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．実際，高齢被験者（12 名）のみのプレストークでの音声認識率は 62.1%であった．逆に機器操作に抵抗のない被験者にとっては，音声認識の開始と終了を自分の意思ではっきり切り替えられるので操作は少々面倒でも使いやすいといった感想も少数であるがあった．

- プッシュトーク

プッシュトークでは，音声認識の開始だけ被験者がボタンを押下し，音声認識終了は認識エンジンが無音区間を検出することで自動的に行うこととした．そのため，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り，エラー率（押し忘れ率）は全体で 8.1%，高齢被験者だけだと 13.8%と，プレストークに比べて高い結果となった．

- センサ駆動

センサ駆動は，自然な発話動作をセンサが検出することで，ボタン操作をすることなく適切なタイミングで自動的に音声認識の開始/終了を切り替えることができる．そのため，プ

レストーク，プッシュトークによる入力を先に実施した被験者にとっては本当に操作なしで入力できるのか少々不安に感じる様子もあったが，いったん慣れてしまうと非常に使いやすいとの感想が多数寄せられた．

発話動作の検出は，距離センサと加速度センサを併用することで頑健性が増し，高齢被験者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．発話動作の検出漏れと誤検出をあわせた検出エラー率は 4.8%（高齢被験者だけで集計しても 5.0%）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た．

これらの特徴をまとめると，次のように結論づけることができる．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．本論文で提案するセンサ駆動は，双方の欠点を補う特長があり，習熟なしで音声認識入力を適切に扱うことができる方法として有用であると考えられる．

6. ま と め

本論文では，複数のセンサを用いることで簡単かつ高精度な音声認識入力を実現する，センサ駆動ハンドヘルド型音声認識入力方法を提案した．まず，実験により，従来のハンドヘルド型マイクによる音声認識入力の課題は口元とマイクとの距離，発話開始と終了の検出にあることを明らかにした．その解決策として発話距離については距離センサ，発話動作検出については加速度センサの活用を検討した．実験の結果，距離センサの方が，音声認識精度，発話動作の検出に有効であることが分かった．しかし，距離センサ単独では，発話意図のない状況での誤検出が想定されるとともに，つねに動作させ続けるには消費電力が大きいという問題がある．そこで，両センサを組み合わせたセンサ駆動ハンドヘルド型音声認識により，発話動作検出のロバスト化と低消費電力動作を実現した．実験により，センサ駆動ハンドヘルド型音声認識は，プッシュトークでのエラー率（8.1%）より低いエラー率（4.8%）で発話開始を検出できること，プレストークより高い音声認識精度（プレストーク：75.5%，センサ駆動：82.4%）を得られることが明らかになった．これは，ハンドヘルド型マイクを使った音声認識入力において，ユーザへの負荷を少なくし，かつ高精度の音声認識を実現する手法として期待できるものである．

今回の実験では，検索課題を提示した限定的なタスクであったが，今後は被験者の好きな番組などを自由に検索するタスクを，実際の家庭のリビングで実施するなど，より実用的なタスクでの性能評価を実施する．また，発話動作認識精度，音声認識精度のさらなる向上を図るとともに，コスト面も含め実用化に向けた検討を進めていく．

参 考 文 献

1) 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).
2) 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).
3) 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).
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) 金澤博史，友田一郎，高畠由彰，竹林洋一：ユビキタス社会に向けた Bluetooth ヘッドセットの開発，日本音響学会 2002 年春季研究発表会，pp.219–220 (2002).
7) 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).
8) Kawamoto, M., Asano, F., Asoh, H. and Yamamoto, K.: Particle Filtering Algorithm for Tracking Sound Sources Using Microphone Arrays, *Proc. ICASSP 2007*, Vol.I, pp.129–132 (2007).
9) Herbordt, W., Horiuchi, T., Fujimoto, M., Jitsuhiro, T. and Nakamura, S.: Noise-robust hands-free speech recognition and communication on PDAs using microphone array technology, *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp.307–310 (2005).

(平成 21 年 4 月 20 日受付)
(平成 21 年 11 月 6 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．

土井美和子（フェロー）

1979 年東京大学大学院工学系研究科修士課程修了．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，東京工業大経営協議会委員，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副会長等を務める．本会フェロー．博士（工学）．