音声ペン：音声認識結果を手書き文字入力で利用できる新たなペン入力インタフェース

Speech Pen: New Pen Input Interface Capable of Utilizing Speech Recognition for Digital Writing

栗原 一貴　　後藤 真孝　　緒方 淳　　五十嵐 健夫*

Summary. This paper introduces a multimodal input system, called "*speech pen*" that assists digital writing during lectures or presentations with background speech and handwriting recognition. The instructor basically freely speaks to the audience and writes on an electronic whiteboard as usual. The system recognizes those speech and handwriting in the background and provides the instructor with predictions for the further writing by using the recognition results. The instructor can accept a prediction and paste it in the board to save manual writing. If all predictions are wrong or useless, the instructor can simply ignore them. The speech-pen system also allows the sharing of context information for predictions among the instructor and the audience; the speech recognition result of the instructor is sent to the audience to support their own note taking. A preliminary study shows the effectiveness of this system and the implications for further improvements.

1　はじめに

音声や手書き文字は人間にとって古くから自然な表現手段であり，それを計算機への入力に活用することを目指した音声認識や手書き文字認識などの認識技術は，長年の研究により性能が大きく向上してきた．しかし入力した文字列には認識誤りが不可避であるため，入力後の訂正作業を必要としていた．訂正作業の効率を向上させる研究 [15] もなされてきたが，事前に用意した辞書に登録されていない未知語には対応できず，完全に訂正するためには依然として煩雑なインタラクションが必要であった．せっかく自然な表現による入力を目指しても，入力結果を正しい文字列とするための労力が大きいため，総合的に判断してキーボードより優れたインタフェースを実現することは難しかった．

そこで本研究では，ユーザが計算機へ文字列（活字）を一字一句間違わずに入力するために認識技術を用いるのではなく，ユーザが他の人達に読んでもらう手書き文字を入力する手助けを得るために認識技術を用いる新たなインタフェース「音声ペン」を提案する．音声ペンでは，音声認識と手書き文字認識を組み合わせて，活字ではない純粋な手書き文字入力[1]を効率化する．従来研究との大きな違いは,「認識誤りを全て訂正しなければならない」状況ではなく,「認識結果が誤っていれば使用せず，認識結果が正しいときに恩恵が受けられる」状況で音声認識および文字認識を用いている点である．また，ユーザが計算機へ向かって発話するのではなく，ユーザが他者に向かって発話した自然な音声を認識する．つまり，ユーザが音声認識の存在を特に意識せずに普通に発話すると，その人間へ向けた音声を計算機が勝手にモニタリングして認識し，ユーザを支援する点が新しい．

今回音声ペンでは講義，プレゼンテーション，およびミーティングのような，ユーザが講演者としてペンで板書しながら聴衆へ説明する状況を対象とする．こうした人前で説明する状況では，話している途中にキーボードを用いて文字入力することに社会的な違和感があるが，その一方ですべての文字をペンで板書する（例えばタブレット PC 上で文字を書く）のは労力がかかるため，文字入力の効率化に対する潜在的なニーズがある（[11] によれば講義時間の 18%が板書に費やされている）．上記の状況で特徴的なのは，ユーザは聴衆に読んでもらうために手書き文字で板書しているだけなので，計算機へ活字を一字一句間違わずに入力する必要はないことである．また，聴衆へ説明している音声は，板書内容と密接に関わるため，その音声を認識した結果がたまたま正しかったときには，それを板書する際に利用できる．例えば，ユーザが板書している文字列の続きを，システムが講演音声の認識結果中から見つけ出して提示することで，ユーザはそれを選択するだ

* Kazutaka Kurihara and Takeo Igarashi, 東京大学大学院 情報理工学系研究科 コンピュータ科学専攻, Masataka Goto and Jun Ogata, 産業技術総合研究所, Takeo Igarashi, 科学技術振興機構 さきがけ

[1] 本論文では手書きではなくフォントにより表示されるコンピュータ上の文字を「活字」と呼び区別する．

けで板書ができ，続きをすべて手書きする労力が削減される．

以上はユーザが講演者の立場で音声ペンを利用する場合について述べてきたが，聴衆が（例えばタブレット PC 上で）ノートを取る立場でも音声ペンは有用である．例えば講演者の音声認識結果や文字認識結果を，文字入力支援のためにコンテクスト情報としてネットワーク経由で共有し，聴衆がノートを取るときに候補として提示することが可能である．さらに過去の講演内容等も共有することで，現在の認識結果だけに限定されない総合的なコンテクストの共有も可能になる．

2　音声ペンシステム

ここでは，まず音声ペンシステムのユーザインタフェースについて説明する．

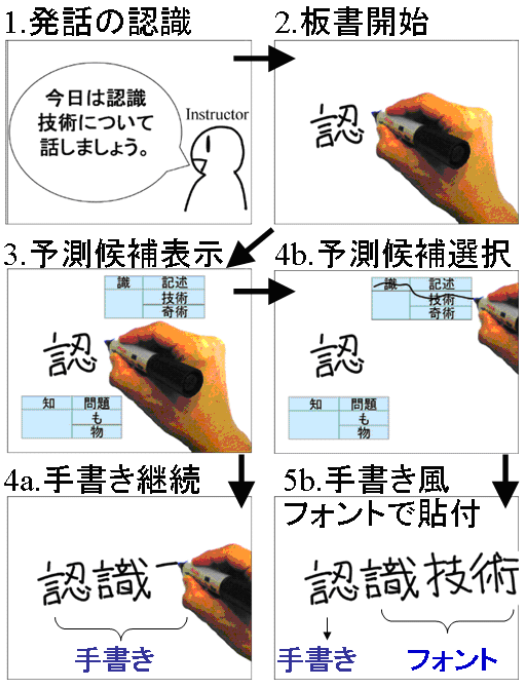

図 1. ユーザインタフェースの概要

2.1　ユーザインタフェースの概要

音声ペンシステムは音声認識と手書き文字認識を用いた予測入力により，講義・プレゼンテーション時の板書およびノート取り作業を支援するものである．図１はユーザの視点から音声ペンシステムがどのように動作するかを示した図である．講演者は自由に発話しながら電子白板に板書を行う（図１の１および２）．書くのを少し静止すると，システムは音声認識と手書き文字認識結果に基づき次に書く可能性が高い文字，語，文を提示する．これらの予測候補は書く作業の邪魔にならないように手の周りに表示される（図１の３）[2]．予測候補は過去の発言履歴（音声認識結果）や予め設定しておいた辞書から生成される．もし講演者が予測を利用したくない時や正しい予測候補が得られない場合はそのまま手書き作業を継続できる（図１の 4a）．入力したい候補を発見できた場合は候補をなぞるジェスチャーによりそれを白板上に挿入でき，(図１の 4b 5b）すべて手書きで入力する場合に比べて労力を軽減できる．この文字は，講演者の筆跡に似せて作られたフォントで表示される．

ここで重要である点は，各時点において既に書き込まれている手書き文字については活字への変換や訂正作業を行わないため（図１ 5b では “認” は手書き文字のままであり，“識技術” が手書き風フォントである），ユーザが予測候補を能動的に使いたいと思う時以外は音声ペンシステムの存在を無視できることである．つまり本システムは基本的にバックグラウンドで働くものであり，フロントエンドとして使用を強制するものではない．また，ユーザのスキルに合わせて段階的に作業効率を高められる学習スキームを持っており，初心者ユーザは初め普段のように手書きのみを行えばよく，次第にシステムのサポートを受けるように慣れてゆけばよい．

2.2　予測の表示

本システムでは複数の予測候補がユーザの最新の書き込み位置の周辺に表示される（図１の３）．これらの予測結果は主に過去の発話の音声認識結果に対応しており，直前の手書き文字認識結果に基づき音声認識結果データベースから検索された「最後に書かれた文字や語から始まるような過去の発言」である．得られた発言は音声認識結果の複数の可能性とともに尤度の高い順に並べられて表示される（図２）．この表示は，音声認識の誤り訂正用インタフェース「音声訂正」[15] の競合候補表示を応用したものである．

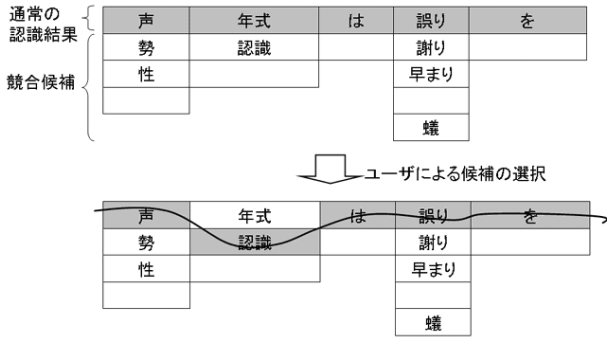

図 2. 予測候補の提示・選択方法

[2] 講演者の書き込み用画面と表示用画面を分けられる場合（例えば Tablet PC とプロジェクタ），これらの予測候補は書き込み用画面にのみ表示するとよい．

2.3 予測の選択と無視

入力予測候補が表示されたとき，ユーザはそれらを選択して挿入するか，無視して手書き作業を続行するかを任意に決定できる．予測の選択は一筆書きでリスト中の候補をなぞっていく crossing interface で行われ（図２），選択された文字列は手書き中の白板領域にユーザの筆跡を模したフォントを用いて挿入される．そうしたフォントは商用サービス[3] を利用して用意でき，フォントの表示サイズは直前の手書き文字を分析し自動的に決定される．一方，予測候補が役に立たないと思ったときは，再び手書きを始めるだけで予測候補は消えるので，他の余分な操作を必要とすることなくユーザは予測候補を無視することができる．また，最後に手書きを行ってから一定の時間が経過すると予測候補は同様に消え，ユーザに余計な混乱を与えることはない．

2.4 システム構成と ambient context の共有

図３に現在の音声ペンシステムのシステム構成を示す．講演者は音声認識用のマイクに向かって話し，プロジェクタに接続された Tablet PC もしくは電子白板にペンで書き込みを行うことで講義を進める．このとき聴衆もそれぞれ独立して各自のノートを Tablet PC で取る．講演者の音声は音声認識サーバで処理され，認識結果が講演者，聴衆を含むすべてのユーザにネットワーク経由で配信され，共有される．共有されている音声認識結果は各ユーザ（講

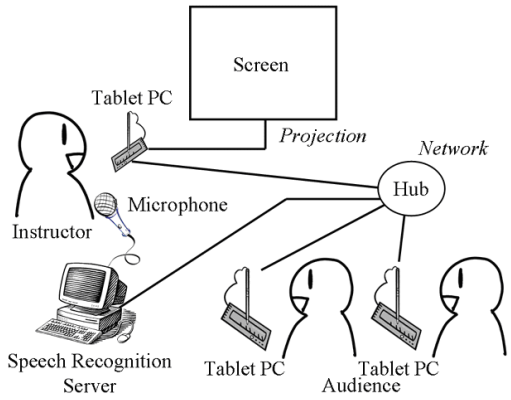

図 3. システム構成

演者，聴衆）がこれまでに述べたような予測つき手書き入力を行う際にデータベースとして用いられる．講演者の発言というある種のコンテクスト情報が，入力支援という最初からは目に見えない形で共有されるため，各ユーザの主体性を反映した資料作成が可能である．われわれはこれを ambient context の共有と呼んでいる．「○年△月□日の講義」,「～分野の専門用語」のように保存，読込が可能であるため，時間と空間を越えて文字入力用の辞書をカスタマイズすることも可能である．

現在のプロトタイプシステムでは共有する対象が音声認識結果や用語辞書に限定されているが，今後手書き認識情報も ambient context として共有したり，各ユーザがどのような認識候補を採用して挿入したか，といった情報も共有したりすることを検討している．

2.5 関連研究

音声認識のインタフェースとしての新たな可能性を論じた関連研究として，後藤ら [13] の非言語情報を活用した「音声補完シリーズ」があげられるが，本論文もそのような試みのひとつであると位置づけられる．また，複数のモダリティの認識技術を相補的に組み合わせることで全体の認識率を向上させる Oviatt[9] 提唱の mutual disambiguation を，本論文では音声認識と手書き文字認識を組み合わせた文字入力に適用している．また Oviatt[10] は，人間の発言と書き込み動作の順やタイムラグには個人差が存在することを報告している．音声ペンでは，発言の前に書き始めてしまうと原理的に対応できない．しかしその場合でも書き続けることでシステムが破綻することは無く，また重要な語句は繰り返し用いられる場合が多いので後に活用される可能性は大きい．Kaiser[4] も音声と手書きのマルチモーダル入力手法を提案しているが，誤認識を意識させないインタフェースを追求した本論文の主張とは異なる．中川ら [16] は手書き，音声の統合的な認識エンジンを開発しテキスト入力インタフェースの一例を示したのに対し，本論文では音声認識器用ではない日常の自然な発話の活用を取り扱う．このテーマについては Hindus ら [3]，Lyons ら [7] が取り組んでいるが，文字入力へと活用する我々の目的とは異なる．

音声ペンシステムは [8][17] などの手書き認識技術を用いた従来の予測型テキスト入力システムとは異なり，ユーザが既に入力したものをシステム側の都合で訂正，再入力する作業が必要ない．これは雑音下で誤りを起こしやすい音声認識技術および文字認識技術を有効に活用できる，特記すべき特長の一つである．

講義，プレゼンテーションにおけるコンテキスト共有手法で一般的なのは，[5][1] のように講師のプレゼンテーションスライド資料を聴衆の PC に配信するものである．これはいわば陽に共有するコンテキストであり，最初から目に見える形で与えられた情報に対しアノテーションなどの操作を行っていく．一方 ambient context の共有はいわば陰に共有するコンテキストであり，最初から目には見えず個々人が各自必要と思う情報を記録しようとするとき初めて具現化する．これらの共存は可能であり，双方とも重要なものであると我々は考えている．[2] は

[3] http://www.techno-advance.co.jp/product/myfont/

キーボード入力やペン入力を複数人で共有するインタフェースを提案したが，ペン入力を単に画像で保持するなど，マルチモーダルに拡張した際に再利用性が乏しい点が問題である．一方 ambient context はマルチモーダルなテキスト情報を図２のような形で管理するため，検索や修正の作業が可能であり，再利用性が確保される．

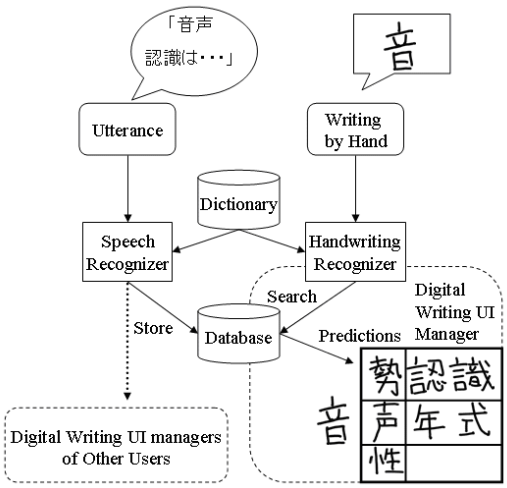

図 4. システムのアーキテクチャ

3 実装

ここでは音声ペンの実装方法について述べる．

3.1 アークテクチャ

「音声ペン」を実現するシステムは．図４のように，主に音声認識部と予測つき手書き文字入力管理部で構成される．音声認識部は，ユーザの発話を常時認識しており，手書き入力予測候補の元となる confusion network（ambient context の実体となるデータ，次節参照）を生成してデータベースに蓄える．それと平行して，手書き文字入力管理部では，ユーザの手書き文字を認識し，その先の予測候補を画面表示する．通常の使用では，講演者のみが音声認識部，手書き文字入力管理部を両方用い，聴衆は後者のみを用いる．これらの構成要素は別々のプロセスとして実装され，ネットワーク (LAN) 上の複数の計算機で負荷分散して実行することが可能である (前者をワークステーション (Xeon 3.06 GHz CPU, Linux 2.4)．後者をタブレット PC(Pentium M 1.4GHz CPU, Windows XP Tablet Edition) 上で実行した)．プロセス間の通信には，音声言語情報をネットワーク上で効率よく共有することを可能にするネットワークプロトコル RVCP (Remote Voice Control Protocol)[14] を用いた．

音声認識部の音響モデル，言語モデルには，CSRC ソフトウェア 2000 年度版 [6] から，PTM triphone モデル，新聞記事テキストより学習された 20000 語の bigram をそれぞれ用いた．手書き文字入力管理部は，Microsoft Tablet PC Platform SDK を用いて実装した．また音声ペンシステムを動作させる講義，プレゼンテーション環境として，手書き電子プレゼンテーションツール「ことだま」[12] を用いた．

3.2 音声認識と入力予測候補の生成方法

提案する音声ペンシステムを実現するためには，逐次入力される講演者の発話に対して認識を行い，図２に示されるようなシンプルな入力予測候補をリアルタイムで生成する必要がある．本システムでは，大規模な単語グラフを効率よく圧縮した形式である confusion network を，ユーザ側に提示する入力予測候補として利用する [15]．confusion network を利用することにより，図２で示されるように，各単語候補間の競合関係が明確化し，ユーザは効率よくペン等による候補の選択が可能になる．

音声ペンでは，ユーザは誤りを含めた全ての音声認識結果を利用することは想定しておらず，認識誤りを避けながらユーザの欲しい結果だけを積極的に利用するインタフェースとなっている．したがって，ディクテーション目的の音声認識システムのように，言語モデルや語彙の不足による認識誤りが，システム全体に大きく影響することはないと考えられる．実際に本システムでは，講演者の音声を認識するための言語モデルとしては，より多くの話題をカバーし，比較的学習テキストも利用しやすい新聞記事から学習された N-gram を用いている．認識結果は複

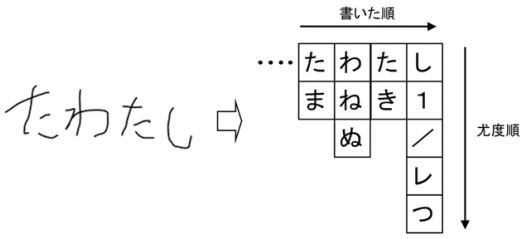

図 5. 手書き文字（左）と手書き文字認識結果（右）

数の区画からなり，その一つ一つが既定値では最大５個までの認識結果からなる

3.3 手書き文字認識方法

音声ペンシステムでは，ユーザは電子白板上の任意の場所に任意の大きさで手書きを行うことができる．即ち [17] などの多くの従来の手書き文字認識によるテキスト入力システムとは異なり，文字入力用のセル（長方形領域）への書き込みを強制されない．その反面，文字認識に先立ちシステムはまずストロークのセグメンテーション（ストロークを文字単位にグループ化する作業）を行う必要がある．図

５にセグメンテーションと手書き文字認識の結果を示す．手書き文字認識の結果は N-best リストの系列として次の処理段階に送られる．現在の実装では，Microsoft Tablet PC Platform SDK の文字認識エンジンを用いており，セグメンテーション結果に複数の可能性が考えられる場合については考慮していない．

3.4 入力予測候補の決定方法

システムは手書き文字認識の結果をクエリとして confusion network のデータベースを検索する．その際，まず一番最近書かれた文字（もしくは語）を取り出してクエリとし，対応するデータベース上の confusion network を検索する．もしもたくさんの候補がマッチした場合は，クエリに最近書かれた文字の一つ前の文字を加える．つまり confusion network の中から２文字の文字列と同じものを探す．この作業により，一般的にマッチした候補の数は１文字クエリの場合よりも減少する．この作業を繰り返し，マッチする候補がなくなるまでクエリの文字数を多くしていく．図５右の例では,「し」,「たし」,「わたし」,「たわたし」の順で検索を行っていく．最終的にシステムはもっとも長いクエリにマッチした候補を出力とする．

この方法でははじめから検索候補が見つからない場合がある．つまり confusion network データベースの中に最近書いた文字が存在しない場合である．このような場合，システムは代わりに次に尤度の高い文字認識結果を用いる．そしてマッチする検索結果が多かった場合は，先述のように最近の文字からさかのぼってマッチしなくなるまで検索を進めていく．図５右の例では，もしも「し」が見つからなかった場合,「１」,「た１」のように進める．

システムは今まで述べてきたような作業を，予め設定してある数（現在の実装では３つ）のマッチする検索結果が得られるまで行う．得られた検索結果は尤度の高い順にソートされ，ユーザに提示される．予備実験ではこの単純なアルゴリズムでも比較的うまく機能していたが，今後の改善の余地は多い．例えば現在は尤度を評価関数としているが，利用頻度，データの新しさなども評価に加えれば性能向上が期待できる．

4 ユーザスタディ

提案システムの有効性を確認するとともにさらなる改善へ向けての知見を得るため，簡単なユーザスタディを行った．８人のテストユーザがボランティアで参加した．

4.1 手順

用意したタスクは，講演者と聴衆に扮して模擬的な講義を行い，板書およびノート取り作業を行うというものである．それぞれのテストユーザは，講演者と聴衆どちらか一方を一度だけ演じる．前もって数分間簡単な操作トレーニングを行い，その後にタスクを実行する．本実験では模擬的な講義のテーマとしてノート１ページ，５分程度の分量の「たこ焼きの作り方」を選んだ．音声認識エンジンの語彙や言語モデルには特に変更は加えていない．これは本システムが，認識誤りを起こしやすい環境にあっても有用であることを示すためである．音響モデルも，インフォーマルな会話用のものではなく，話者適応もしていない．

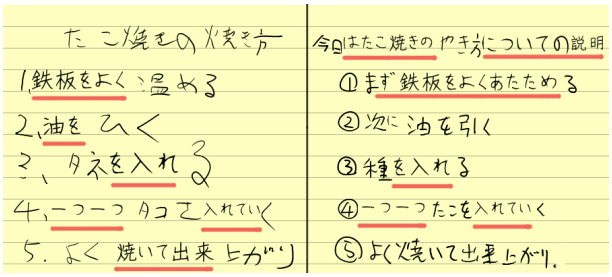

図 6. ユーザスタディで得られた板書・ノートの例．(左) 講演者の板書，(右) 聴衆のノート．

4.2 結果

図６に得られた板書・ノートの例を示す．予測により挿入されたテキストが区別できるように下線を図の作成時に引いた．[5][1] などの従来のコンテクスト共有システムでは見られない，同じコンテクストの共有から個性豊かな表現が得られる特徴が観察された．これは本システムの自由度の高さを示すものである．

4.3 サポート率

音声ペンシステムがどの程度ユーザを支援できるかを分析するため，以下の「サポート率」という評価尺度を提案する：

$$S = \frac{N_{sup}}{N_{all}} \tag{1}$$

ここで N_{sup} は予測候補の中からユーザが選択して挿入された文字のストローク数[4]，N_{all} は全ストローク数である．サポート率はすべてのストロークが手書きによって書かれた場合に最小値０となり，すべてのストロークがシステムによって生成された場合，理論値として最大値１を取る．本システムではユーザは基本的に手書きを行っており，入力予測は必要なときのみ用いるという方針のため，サポート率を最大値１に近づけることが目標ではない．また，箇

[4] 厳密に言えば挿入される文字は活字であり，ストローク情報は無い．ここでは活字をもし手で書くとしたときに必要な手書きストローク数を数えている．

条書き記号やアノテーション記号などの非文字列ストロークもサポート率を低下させる要因となる．

図 7 に全テストユーザのサポート率を示す．訓練時間はわずかであったが，テストユーザはシステムのサポートを得ることができた（0.22 から 0.70）．この結果を分析すると，後援者・被験者間のサポート率の差はそれほど顕著ではないようである．また「予測入力は気が向いたときに使ってください」という教示を行ったにもかかわらず novelty effect がバイアスとしてテストユーザの行動に影響を与え，積極的にシステムのサポートを得る傾向が現れた可能性がある．別のインフォーマルな評価実験では，講演者のサポート率は 0 だが聴衆のサポート率は 0.41 というケースも存在した．これはそれぞれのユーザが各自のスキルや状況に合わせてシステムを活用し，タスクを完遂させた例である．今後より長期的で実際の講義に近い評価実験を行うことにより，このような音声ペンの特徴を示す堅牢なデータが得られることだろう．

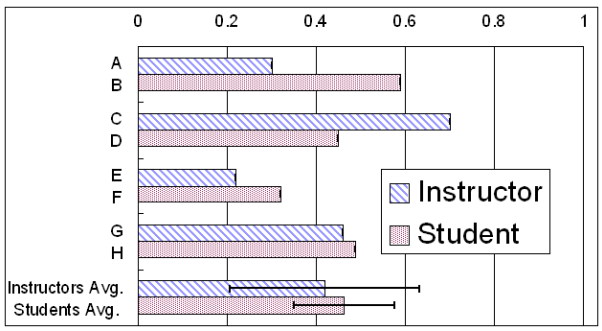

図 7. 8 人のテストユーザ $A-H$ のサポート率．

4.4 テストユーザからのフィードバック

タスク終了後にテストユーザに対しインタビューを行った．まず，本システムに対する一般的な印象を尋ねたところ，8 人全員がポジティブな印象を持っており，特に「使用を強制されるのではなく，活用したいときだけ存在を意識すればよい」という点が魅力的だと指摘されていた．次に，本システムの更なる改善に向けてのコメントおよび提案を尋ねたところ，(1) 入力予測候補の表示場所には改善の余地がある，(2) 各予測候補の表示が小さすぎる，(3) すべての文字を手で書く時間とそう変わらないようであれば，予測候補を選ぶメリットは小さいだろう，などが得られた．

5 まとめ

本研究では，音声認識と手書き文字認識を用いつつユーザが手書き文字を書く作業を支援する「音声ペン」システムを開発した．簡単な実験により本システムの有効性が示され，更なる改善に向けてのユーザからの意見が得られた．今後は得られた知見を元にシステムを改良し，より実際の講義に近い条件で評価実験を行う予定である．

謝辞

本研究は，文部科学省 21 世紀 COE プログラム (研究拠点形成費補助金)，および日本学術振興会科学研究費補助金 (若手研究 (B)) の支援を受けた．ここに記して謝意を表す．

参考文献

[1] Anderson et al.. A Study of Digital Ink in Lecture Presentation. *CHI'04*, pp.567–574, 2004.

[2] Denoue et al.. Shared Freeform Input for Note Taking across Devices. *CHI'03*, pp.170–171, 2003.

[3] Hindus et al.. Ubiquitous Audio: Capturing Spontane-ous Collaboration. *CSCW'92*, pp.210–217, 1992.

[4] Kaiser. Multimodal New Vocabulary Recognition through Speech and Handwriting in a Whiteboard Scheduling Application. *IUI'05*, pp.51–58, 2005.

[5] Kam et al.. A System for Cooperative and Augmented Note-Taking in Lectures. *CHI'05*, pp.531–540, 2005.

[6] Kawahara et al.. Recent Progress of Open-source LVCSR Engine Julius and Japanese Model Repository. *ICSLP*, pp.3069–3072, 2004.

[7] Lyons et al.. Augmenting Conversations Using Dual-Purpose Speech. *UIST'04*, pp-237–246, 2004.

[8] Masui. An Efficient Text Input Method for Pen-based Computers. *CHI'98*, pp.328–335, 1998.

[9] Oviatt. Mutual Disambiguation of Recognition Errors in a Multimodal Architecture. *CHI'99*, pp.576–583, 1999.

[10] Oviatt et al.. Individual differences in multimodal integration patterns: what are they and why do they exist?. *CHI'05*, pp.241–249, 2005.

[11] 岩田 他. 対話型電子白板を用いた電子化授業への遠隔受講者参加方式の試作. **情処研報** *2002-CE-67*, pp.33–40, 2002.

[12] 栗原 他. ことだま：ペンベース電子プレゼンテーションの提案. *WISS'04*, pp.77–82, 2004.

[13] 後藤. 非言語情報を活用した音声インタフェース. **情処研報** *2004-SLP-52-7*, pp.41–46, 2004.

[14] 後藤 他. 音声補完: 音声入力インタフェースへの新しいモダリティの導入. **コンピュータソフトウェア**, Vol.19, No.4, pp.10–21, 2002.

[15] 緒方, 後藤. 音声訂正: 認識誤りを選択操作だけで訂正ができる新たな音声入力インタフェース. *WISS'04*, pp.47–52, 2004.

[16] 中川 他. 音声と手書き文字の同時入力インタフェース. **情処研報** *2005-SLP-56*, pp.29–34, 2005.

[17] 福島, 山田. 予測ペン入力インタフェースとその手書き操作削減効果. **情処学論**, Vol. 37, No. 1, pp. 23–30, 1996.