様式 C-19

科学研究費補助金研究成果報告書

平成２３年　４月　１日現在

機関番号：13901
研究種目：若手研究(B)
研究期間：2009～2010
課題番号：21700577
研究課題名（和文）　遠隔パソコン要約筆記における筆記者支援に関する研究

研究課題名（英文）　A Study on Supporting Transcribers for Remote PC Transcription

研究代表者：竹内　義則（TAKEUCHI YOSHINORI）
名古屋大学・情報連携統括本部情報戦略室・准教授
研究者番号：60324464

研究成果の概要（和文）：
本研究では，遠隔パソコン要約筆記における筆記者支援について研究を行う．特に発話数式に対応するスライドの数式画像を自動で抽出する手法を研究する．この手法は数式発話時に講師が発話数式に対応するスライドの数式を指示しているという特徴を利用している．音声認識により，数式発話抽出を行う．また，映像から講師の行っている指示動作を抽出する．実際に収録したデータに対して，数式画像抽出処理を行った．その結果，約 70%の再現率，約 91%の適合率を得た．

研究成果の概要（英文）：
In this research, we propose a method solving the problem of captioning when an instructor utters a mathematical formula with different interpretations, using remote PC transcription system. We adopt an approach to extract an image corresponding uttered mathematical formula from lecture slides. This method uses a fact that a instructor usually points a mathematical formula on the slide when he/she utters it. We use speech recognition software to detect pronunciations of mathematical formula. Then, we extract pointing gestures from the images of the lecture. We have conducted experiments by using recorded audio-visual signals from a real lecture. Experimental result shows that recall ratio is about 70% and relevance ratio 91%.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,000,000	300,000	1,300,000
2010 年度	1,000,000	300,000	1,300,000
総　計	2,000,000	600,000	2,600,000

研究分野：福祉情報工学
科研費の分科・細目：人間医工学・リハビリテーション科学・福祉工学
キーワード：情報システム，ユーザインタフェース，福祉情報工学，情報保障

１．研究開始当初の背景

本研究では，聴覚障害者が大学での高等教育を受ける際の情報保障について取り扱う．手話で講義が出来る教育者は限られており，専門知識を持った外部の非常勤講師に講義を依頼する場合，その講師は，手話ができないことが多い．この場合，パソコン要約筆記によって情報保障が行われてきた．これは，健聴者のボランティアをその講義室に配置し，その場で講師の声をパソコンでタイプし，前のスクリーンにタイプした文字を投影する手法である．

また，講師の声をマイクロホンから入力し，黒板，スライドなどの映像をカメラで撮影し，それらを自宅などの遠隔地に送り，遠隔地側で文字を入力して送り返す「遠隔パソコン要約筆記」が行われている．遠隔パソコン要約筆記では，講師の声だけでなく，スライドや

黒板などの映像も重要である．講師の声を入力しただけでは，難解な専門用語を誤って入力するかもしれないし，指示語に対応することができなくなる．しかし，遠隔では，カメラの画角，解像度などの問題で，見たい部分がよく見えない問題が新たに生じている．

このような背景の下，これまでに遠隔パソコン要約筆記で問題となる指示語を検出し，指示対象物を抽出する研究を行ってきた．要約筆記では，要約文中の指示語の指示対象物が要約文だけではわからないため，入力者が指示語を指示対象物に置き換えて入力している．遠隔パソコン要約筆記では，カメラの画角，解像度の問題で，講師が指示語を発生したときに指示対象物を見落とすことがある．そこで，指示対象物を抽出し，要約筆記者へ提示することにより，補うことができる．

２．研究の目的

本研究では，遠隔パソコン要約筆記において，カメラから得られる視覚情報から要約筆記に必要な情報を抽出し，入力者に提示することを目的とする．これにより遠隔の入力者は，必要な視覚情報をもれなく参照することができ，正確な要約文を入力することが可能となる．

字幕作成者の対応が困難である対象の一つとして数式が挙げられる.発話される数式には多様性が存在する.例えば,「エックスマイナス二分の一」と講師が発話したとする.このとき,数式の文字情報もなく音声情報のみで,その数式を解釈しようとした場合,x-1/2 と(x-1)/2 というような二通りの解釈が可能である.しかし,キーワード入力はテキスト入力であるため,上記のようなあいまいさが解消する表記ができない.これによって字幕入力が途切れるという問題が起こりうる.この問題に対して,数式を自動的に字幕作成者に提示するシステムが求められる.

本研究では講師音声から発話された数式要素を抽出し,講師映像からの指示動作抽出結果を統合することによって,発話された数式要素に対応するスライド上の数式画像を抽出することを目的とする.

３．研究の方法

入力は講師の発話した音声（以下,講師音声）と講師とスライドを撮影した映像（以下,講義映像）である.講師音声からは音声認識を用いて,x や n，といった数式を構成する要素（以下,数式要素）を抽出する.講義映像からは講師の行っている指示動作を取得する.講師音声から抽出した数式要素の発話開始時刻において,抽出した数式発話と指示動作の統合を行う.これについては 3．4 節で述べる.そして,講義映像から講師が指示している対象を画像として抽出すること

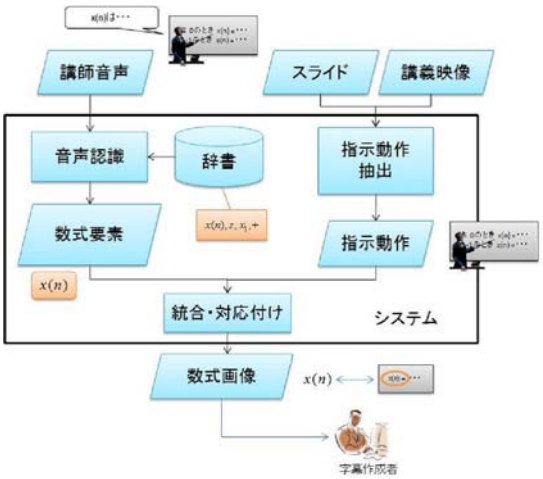

図１　システムの概要

で,発話数式に対応する数式画像が取得できる.システムの概要図を図 1 示す.

数式要素抽出には音声認識を用いる．本システムの音声認識には Julius を用いる．音響モデルと言語モデルは日本語話し言葉コーパスの性別非依存音韻トライフォンモデルを使用している.言語モデルは単語 N-gram モデルを用いている．また，単語辞書に関数や添字付き文字，アルファベット，符号，数字といった数式を構成する要素(数式要素)を辞書登録する．登録する数式要素は事前に配布される講義資料，講義スライドに記述されているものである．なお，辞書登録した数式要素は未知語として扱われる．これは単語辞書に辞書登録した数式要素が言語モデルには含まれていないためである．その代わりに登録した数式要素は未知語として扱われることで，認識が可能となる．

さらに，数式要素の認識率を上げるために,言語モデルの改良を行う．日本語話し言葉コーパスは，学会講演・模擬講演のデータから構成されており，そのデータを用いて音響モデル・言語モデルが作成されている．そのため数式を多用する講義と比較して，数式要素の出現する頻度が低い，もしくは出現していない可能性がある．本研究では，言語モデルの未知語の生起確率を高く設定することで,数式要素の認識率向上を試みている.

講義においてパソコンで提示されるスライドは，テキストで表現された個々の項目あるいは文や図を単位とする領域の組み合わせにより構成されている．これらを図字領域と呼ぶ.

Lilian らの研究では,講義における指示動作を人により客観的に分類・解釈している.それを計算機で利用することを念頭にまとめると，指示動作時の指示棒の使い方には,講師により個人差が存在するが，スライド中の図字領域上で止める・なぞる・囲むという

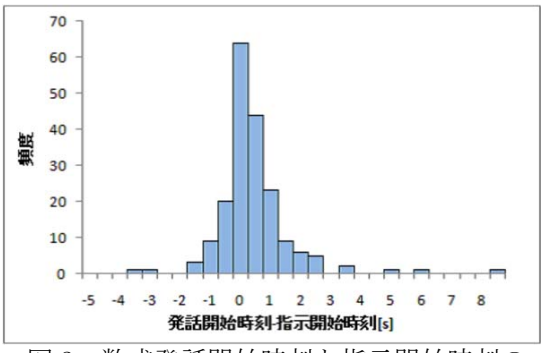

図 2　数式発話開始時刻と指示開始時刻の差のヒストグラム

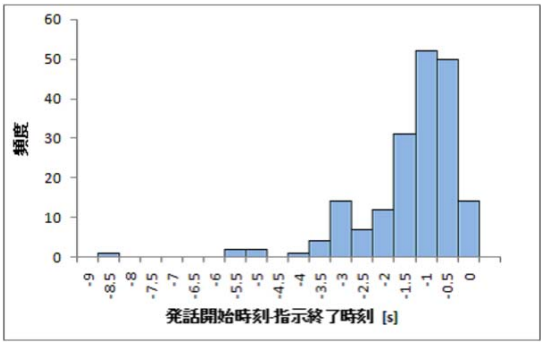

図 3　数式発話開始時刻と指示終了時刻の差のヒストグラム

方法がある．これにより，点・直線・楕円の 3 種類の特徴的な軌跡が現れる．本研究では，このように指示点の軌跡が点・直線・楕円となるように指示棒を使う行為を指示動作とした．

講師の指示動作を撮影し，指示棒先端を抽出する．その動きを追跡し，一定時間の間止まっている場合は点，狭い長方形内で移動している場合は直線，軌跡が交差する場合は楕円と分類することによって指示動作を検出する．

音声処理と映像処理で得られた結果を統合するために，数式発話と指示動作の時間関係の分析を行った．分析には，信号処理という実際の講義を撮影した音声と映像を用いたその講義では数式発話の数が多かった．講義内で行われる，数式発話 190 個のデータに対して分析を行った．まず，各指示動作において指示動作開始時刻と指示動作終了時刻を手動で記録した．次に，各数式発話において数式発話開始時刻と数式発話終了時刻を手動で記録した．この二つのデータを用いて，数式発話と指示動作の時間関係を調べた．

まず，指示動作開始時刻と数式発話開始時刻の分析を行う．数式発話開始時刻と指示開始時刻の差をヒストグラムで表したものが図 2 である．横軸は指示動作開始時刻から数式発話開始時刻を差し引いた時間である．指示動作開始時刻と数式発話開始時刻の差が −1.5 秒よりも正の軸方向にある数式は 185 個で，全体の 97.4%であった．また，−2.0 秒の場合では，188 個の数式が存在し，全体の 98.9%であった．

表 1 発話数式に対応する数式画像抽出結果

講義データ	1	2
発話数式の総数	205	68
音声認識の数式要素抽出総数	173	50
指示対象抽出された数式	170	48
発話数式に対応する対象が抽出された数式	150	45
非対応の対象が抽出された数式	20	3 個
未抽出の数式	3	2 個
再現率	73.1%	66.2%
適合率	88.2%	93.8%

次に，指示動作終了時刻と数式発話開始時刻の時間関係の分析を行う．数式発話開始時刻と指示終了時刻の差をヒストグラムで表したものが図 3 である．横軸は指示動作終了時刻から数式発話開始時刻を差し引いた時間で，秒単位で表している．これによると，全ての数式において指示動作終了時刻の前に数式発話が開始されていることが分かる．

この 2 つの分析により式(1)が成立する．

$$t1 - 2.0 < s < t2 \qquad (1)$$

t1，t2，s はそれぞれ，指示動作開始時刻，指示動作終了時刻，数式発話開始時刻を表す．本研究では，式(1)を満たす数式発話と指示動作を統合し，対応する数式を提示する．なお，一つの発話数式に対して複数の指示動作が対応する場合には，式(1)の時間範囲内にある全ての指示動作に対する指示対象を抽出している．

４．研究成果

講師音声から得られた数式要素の抽出結果と，講義映像から得られる指示動作抽出結果を式(1)により統合させることで，指示対象抽出を行った．使用する数式要素データは数式要素抽出結果(講義データ 1 では 177 個，講義データ 2 では 50 個)を用いた．

なお，式(1)の時間範囲にある指示対象を全て抽出するため，一つの数式要素に対して，複数の指示対象が抽出される場合もある．その場合，その複数の指示対象は一つの指示対象の集合とみなし，その中で一つでも対応する数式画像がある場合を正解とする．そして，その他の抽出された対象は抽出対象から除外する．

発話数式に対応する数式画像抽出結果を表 1 に示す．再現率と適合率は，

再現率 =（発話数式に対応する対象が抽出された数式)/(発話数式の総数)

適合率 =（発話数式に対応する対象が抽出された数式)/(指示対象抽出された数式)

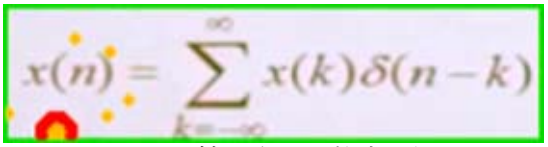

図 4　抽出された数式の例

として求められる．

抽出された数式画像の一例を図 4 に示す．これは講師が「x(n)はインパルスを使って」と発話している際に抽出された画像である．

講義データ 1 においては再現率が 73.1%，適合率が 87.2%であった．また，講義データ 2 においては再現率が 66.2%，適合率が 93.8%であった．対応しない対象が抽出された原因として，数式発話から遅れて指示開始されたため，式(1)を満たさず，別の数式を抽出した場合(1 個)と，指示棒先端が図字領域から遠い場所にあり，指示棒先端により近い別の数式を抽出した場合(1 個)，講師の体がスライドと重なり，講師の体の一部が指示棒先端認識されたことによって，別領域を抽出した場合(1 個)がある．また，音声認識の非発話数式が数式発話と誤認識したものに対して指示対象を行った場合(20 個)もある．

音声認識で正しく認識された数式要素に限定すると，講義データ 1 において 155 個のうち 150 個が対応した画像を抽出しており、その割合が 96.8%である．講義データ 2 において 46 個のうち 45 個が対応した画像を抽出しており，その割合が 95.7%である．このことから，高い割合で発話数式に対応する数式画像が抽出されていると言える．全体結果においては，音声認識で数式要素が抽出されなかった数式に対しては指示対象抽出処理を行っていないため，対応する数式抽出の個数を減少させている原因である．すなわち，音声認識の数式要素の抽出率が低いことによって，全体の数式画像抽出率が低くなったと考えられる．今回の結果においても，音声認識の認識率向上が重要であると考えられる．

本論文では発話された数式に対応するスライドの数式を抽出する手法の提案を行った．まず，音声認識で認識される数式要素の抽出においては，約 71%の再現率，約 90%の適合率を得た．また，数式発話と指示動作の時間関係を調査した．その結果，数式発話は指示動作開始の 2.0 秒前から指示動作終了までの間に開始されていることが分かった．そしてこの時間法則を用いて，実際の講義データで発話数式に対応する数式画像の抽出処理を行ったところ，約 70%の再現率，約 91%の適合率を得た．音声認識で認識されなかった数式に対しては指示対象抽出を行っていないこと，音声認識で認識した数式発話に対して，約 96%が対応する数式画像を取得していることから，音声認識における数式要素の認識率が低いことが原因と考えられる．

今後の課題として，音響モデルの検討や数式に対応した言語モデルの利用，数式要素の誤認識しやすい単語の生起確率を下げるなどの手法を用いた，音声認識の数式要素の認識率向上が挙げられる．また，本研究では数式発話時に対応するスライドの数式を指示しているものに限定しているため，指示動作を伴わなかった発話数式に対する処理の検討も必要と考えている．そして，実際にリアルタイム字幕作成者に抽出数式画像を提示する事によって，システムの有効性を検証する必要がある．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　１件）

① Yoshinori Takeuchi, Hajime Ohta, Noboru Ohnishi, Daisuke Wakatsuki, Hiroki Minagawa, Extraction of Displayed Objects Corresponding to Demonstrative Words for use in Remote Transcription, Proc. of 12th Int. Conf. on Computers Helping People with Special Needs, 査読有, Vol.2, 2010, 152-159

〔学会発表〕（計　３件）

① 太田　創，竹内義則，松本哲也，工藤博章，大西　昇，遠隔パソコン要約筆記における指示語に対応する指示対象抽出，2009 年映像情報メディア学会年次大会，査読なし，2009 年 8 月 26 日
② 太田　創，竹内義則，松本哲也，工藤博章，大西　昇，指示語に対応する指示対象抽出による遠隔パソコン要約筆記者支援の提案，電子情報通信学会福祉情報工学研究会，査読なし，2010 年 3 月 12 日
③ 川口弘哲，竹内義則，松本哲也，工藤博章，大西　昇，リアルタイム字幕作成支援のための数式抽出，電子情報通信学会福祉情報工学研究会，査読なし，2011 年 2 月 18 日

６．研究組織

(1)研究代表者

竹内　義則　（TAKEUCHI YOSHINORI）
名古屋大学・情報連携統括本部情報戦略室・准教授
研究者番号：60324464

(2)研究分担者　なし

(3)連携研究者　なし