社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MUS－74 (17)
2008－SLP－70 (17)
2008／2／9

動作を伴う発話の収集とその認識

三木まどか†, 宮島 千代美†, 西野 隆典‡, 北岡 教英†, 武田 一哉†

† 名古屋大学大学院情報科学研究科,

‡ 名古屋大学情報メディア教育センター

〒464-8603 名古屋市千種区不老町
{miki,miyajima,nishino,kitaoka,takeda}@sp.m.is.nagoya-u.ac.jp

あらまし 人間のコミュニケーション方法に倣ったより自然な人間とコンピュータのインタラクションの実現に向けて，発話と動作の統合理解を目指して，発話と指示動作の統合手法について検討する．タスクとして幾何の問題の解説を設定し，音声と指先の動作を収録した．指先の動作収録には磁気式位置センサを用いている．音声認識によって得られた解説において重要なキーワードと，指先の高さのヒストグラムの分布に基いて抽出した指示動作部分とを，発話と指示動作の開始タイミングのずれの確率分布に基いて対応付けを行った結果 88.7%の正解率を得た．また，音声情報を利用して動作認識の候補を絞ることで，動作の認識精度を 10.8 ポイント向上させることができた．

キーワード 音声認識, 動作認識, マルチモーダル

Multimodal data collection and understanding of speech accompanied with pointing gestures

Madoka Miki†, Chiyomi Miyajima†, Takanori Nishino‡,
Norihide Kitaoka†, and Kazuya Takeda†

† Graduate School of Information Science, Nagoya University,

‡Center for Information Media Studies, Nagoya University

Furo-cho, Chikusa-ku, Nagoya 464-8603, JAPAN
{miki,miyajima,nishino,kitaoka,takeda}@sp.m.is.nagoya-u.ac.jp

Abstract We investigated an integration method for the recognition of multimodal speech accompanied with pointing gestures. As an example of such multimodal speech, we selected a multimodal explanation for a mathematical problem of calculating an angle in a quadrilateral inscribed in a circle. Speech and fingertip movements were recorded while solving the problem with a close microphone and 3-D position sensor. Correspondence between keywords for the mathematical problem obtained in speech recognition and pointing gestures extracted using the histogram of fingertip height was found according to the probability distribution of the time gap between the starting points of an utterance and a pointing gesture. 88.7% of the keywords were of keywords were correctly matched to corresponding gestures. Gesture recognition performance was improved by 10.8 points using grammar networks constrained by speech recognition results.

Keywords speech recognition, gesture recognition, multimodal

1 はじめに

人間同士は日常，音声や身振り，表情といった様々な情報伝達様式を用いてコミュニケーションを行っている．そこで，そのコミュニケーション方法に倣ったより自然な人間とコンピュータのインタラクションの実現が期待されており，複数の情報伝達様式を利用し，情報統合を行うシステムに関心が高まっている．人間が用いる情報伝達様式の中でも，音声と身振りは互いに情報を補完し合う関係にあり，双方の情報があってはじめて曖昧性が解消される場面が多くある．その代表例が指示語により対象物を表現する場合である．人が，「ここ」，「この○○」などの音声を発しながら指先で対象物を示した時，指示表現の指す対象物は，音声だけでは特定することができないが，指示動作を考慮することではじめて対象物を特定できる．この観点から，音声と身振り，特に音声と指示動作は，コミュニケーションの中で意図を伝えるにあたり，最も重要な情報伝達様式であると言える．そこで本研究では，発話と動作の統合理解を目指し，幾何の問題の解説を対象とした発話と指示動作の統合認識手法について検討する．

本稿ではまず，幾何の問題を設定し，その解説の様子を，音声と指先の動作を同期して収録した．次に，音声と動作をそれぞれに検出した結果を対応付ける手法を検討した．さらに，音声認識の結果により動作認識に制約を加えることにより，動作認識の精度向上を試みた．

実験対象とする問題の設定と収録，音声，動作の認識手法とその時間情報を利用した統合手法を検討した．また，音声と動作の意味内容を考慮した統合に向けて，音声認識結果の動作認識への利用手法を検討した．

2 問題の設定と収録

図1に示す問題の解説を収録した．図2のように，机の上に置かれた問題図を人差し指で指しながら解説し，その際の発話音声を接話マイクで，人差し指の指先の動きを位置センサで収録した．収録条件を表1に示す．被験者6名（男性4名，女性2名）を対象にして，各被験者に対して以下の2通りの条件で解説を行ってもらい，計16回の解説を収録した．

条件1 自由に説明を行う．(計7回)

条件2 「この角」などの指示表現を用いて表現を簡略化して説明を行う．(計9回)

(問1) 下の図で，∠c の大きさを求めよ．

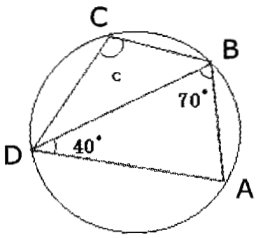

図 1: 図形問題

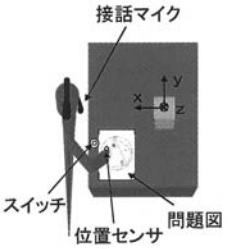

図 2: 収録環境

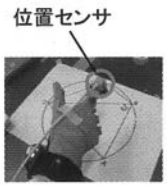

図 3: 位置センサ

表 1: 収録条件

音声の収録条件	
サンプリング周波数	48 kHz
情報量/サンプル	16 bit
位置の収録条件	
サンプリング周波数	100 Hz
情報量/サンプル	16 bit
収録情報	x, y, z 位置座標及び方位角
移動精度	2.5mm RMS
収録時間	計 417.8 [s], 26.1 [s/解説]

表 2: 音声認識の実験条件

サンプリング周波数	16 kHz
フレームサイズ	25 ms
フレームシフト	10 ms
特徴量ベクトル	MFCC12 次, Δ MFCC Δ 対数パワー
HMM 状態数	3

3 音声の認識と指示動作の認識

3.1 音声の認識

図1の問題の解説は，「∠ADB= ∠ACB」のような

等式1つに相当する要素項目の系列によってなされる．そこで，要素項目ごとに対応する発話のネットワーク文法を作成し，そのネットワーク文法を用いて，証明手順の多様性を考慮した解説全体のネットワーク文法を構成した．このネットワーク文法により，人による解説の手順の違いを考慮して音声の認識を行った．また，被験者が問題の証明手順や次に発する言葉や動作を考えながら解説を行うことから，「えー」などの言いよどみも多く現れるため，単語間の各所に言いよどみが入る可能性を考慮した文法を作成した．音響モデルには，話し言葉に適しているCSJ（日本語話し言葉コーパス）の性別非依存音韻モデルを使用した．実験条件は表2のとおりである．実験の結果，認識率 93.1%，認識精度 (挿入誤りを考慮) 71.8%という結果が得られた．

3.2 指示動作の認識

動作を伴う発話を機械が理解するためには，動作の認識を行い，動作が示している情報を得る必要がある．本研究で扱っている動作の全体は，問題の解説に直接関係する指示動作と解説に関係しないその他の動作から構成されており，指示動作は問題図形の部分図形要素 (角，線分，頂点など) を対象物として指し示している．そこで，本節では，問題解説中の一連の指先の動作から，指示動作が指し示している図形要素の系列を認識する手法を考える．指示動作部分の抽出と図形要素との対応付けの手順を以下に示す．

3.2.1 DP マッチングによる指示動作認識

解説中の一連の動作から指示動作部分を抽出し，その指先の軌跡 (位置情報の系列) から，DP マッチングを用いて指し示している図形要素の認識を行った．指示動作部分の抽出と図形要素との対応付けの手順を以下に示す．

- 指示動作の抽出
  まず，指示動作とその他の動作との切り分けを行った．問題図の面に垂直なベクトルの成分 ($-z$) について指先の動作軌跡の分布を見ると，その頻度グラフは問題図面付近 ($-z \approx -500$) に1つ，問題図面上部にもう1つ山のある双峰性のグラフとなる (図4参照)．問題図面付近で行われる指示動作 (紙面に触れている状態) の分布が谷より小さい部分の山に対応し，関係のないその他の動作の分布 (手が浮いてる状態) が谷より大きい部分の山に対応すると考えられるため，このグラフの谷を検出し，谷より小さい部分に分布する動作軌跡を指示動作の軌跡とする．

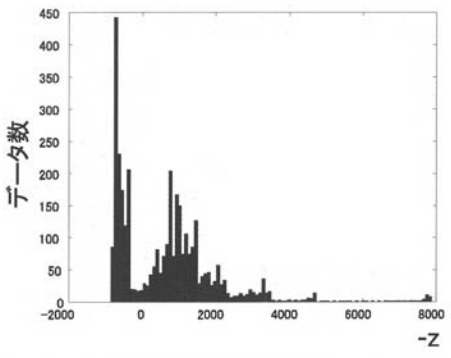

図 4: 解説における $-z$ のヒストグラムの例

- 指示対象 (図形要素) の認識
  次に，抽出した指示動作が指し示している図形要素を認識するため，指示動作の軌跡と問題図中の部分図形要素の対応付けを行った．「∠ABD を指し示している軌跡」のように，あらかじめ問題図中のどの図形要素を指している動作軌跡であるのかわかっている指示動作軌跡をテンプレートとし，評価用の指示動作軌跡と DP マッチングを行った結果，そのマッチングコストが最小となる指示動作軌跡同士に対応する図形要素を認識結果とする．

上記の手法を用い，条件2の8回分の収録のうち1回分を評価用データ，残りの7回分に含まれるすべての指示動作をテンプレートとし，評価用データを入れ替えて実験を行った．その結果，指示動作の要素図形認識正解率は 96.4%となった．

同じ図形要素を示す動作でも，人によってばらつきがあり，それが誤りの原因となった．例えば，同じ直線を指す動作においては，どの被験者の動作も同じ傾きで直線をなぞり示すという動作の本質は同じだが，描く長さが違うなどである．

3.2.2 移動ベクトルを用いた指示動作認識

3.2.1 から，人による指し方のばらつきを吸収し，様々な指し方に対応できるようにするためには，指示対象の図形要素に共通な性質を動作認識に用いる必要があると考えられる．そこで，指先の位置ではなく移動ベクトル (一定時間の指先の移動量と方向を表すベクトル) を用いて指示動作の示す対象図形要素を認識する手法を提案する．図形要素を指し示す動作は，指し示す位置や動きの大きさにはばらつきがあるが，指先の移動方向は共通しており，ばらつきが少ない．

表 3: HMM の学習条件

評価用データ	問 1 の 1 回分の収録データ	
学習用データ	問 1 の学習用データを除く 7 回分の収録データ	
状態数	3	
混合数	1	
特徴量	実験 1	$\Delta x, \Delta y, z, \Delta\Delta x$ $\Delta\Delta y, \Delta z$
	実験 2	$r\Delta\theta_i, r, z,$ $\Delta(r\Delta\theta_i), \Delta r, \Delta z$ (i=1,⋯,12)

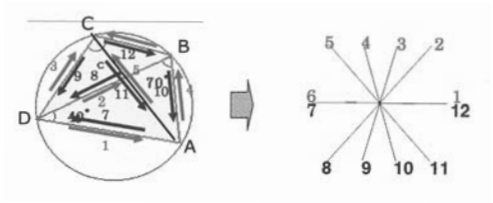

図 5: 問 1 の線分テンプレート

移動ベクトル $\boldsymbol{vec}[n]$ を以下のように定義する.

$$\boldsymbol{vec}[n] = \boldsymbol{q}_{xy}[n] - \boldsymbol{q}_{xy}[n-1] = \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \quad (1)$$

ここで n は時刻であり，実験では 0.1 秒単位である．q_{xy} は描画軌跡の xy 平面上の位置である．

HMM により表 3 の実験条件で図形の基本要素 (線分と頂点) の指示動作及び非指示動作部分のモデルを作成し，指示動作認識実験を行った．指示動作のモデルの種類は 18 種類である．線分を示す動作が 11 種類，頂点を示すが動作 4 種類，角を弧を描いて示す動作が 1 種類，その他非指示動作が 3 種類である．非指示動作のうち 1 種類は紙面に触れていない動作をモデル化したものであり，モデル名を「sil」とした．実験 1 では，移動ベクトルの x，y 成分 $(\Delta x, \Delta y)$ 及び垂直方向の位置 z とそれらの一次差分を特徴量とした．実験 2 では，指示対象となる問題図形の線分の傾き情報を移動ベクトルの表現方法に反映した．指先の動きの速度が早い時の偏角の方が遅い時よりも正確な指先の動きの方向を表していると考えれることから，$r\Delta\theta_i$ という特徴量を用いた．実験 2 の具体的な特徴量計算方法を以下に示す．

＜実験 2 の特徴量計算手順＞

(a) 問題図から各線分 (計 12 本) の相対角度を求め，その問題図形の線分テンプレート (図 5) を作成 (但し，線分 DA の角度を 0，線分 AD の角度を π とする)

(b) **3.2.1** 節の指示動作抽出手法によって指示動作を抽出し，式 (1) に基いて解説の一連の描画軌跡から 0.1 秒ごとの移動ベクトルを算出

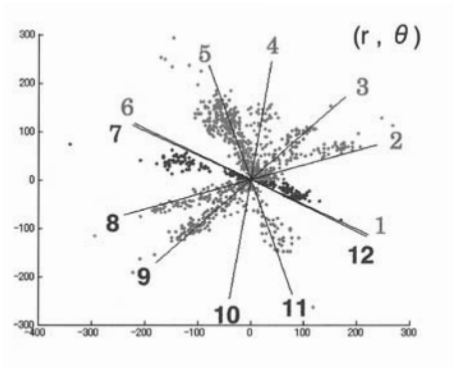

図 6: 回転後の線分テンプレートと vec の分布

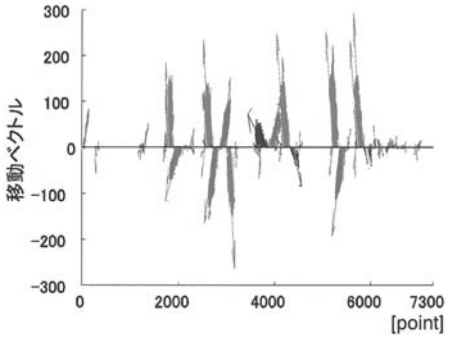

図 7: 時系列に並べた分類結果

(c) (b) の $\boldsymbol{vec}$ の分布と線分テンプレートの距離 D が最小となるよう線分テンプレートを回転 $\boldsymbol{vec}[n]$ の大きさを $r[n]$，偏角を $\theta[n]$ とし，$\boldsymbol{vec}[n]$ に最も近い線分テンプレートの偏角を $\theta_k[n]$ とすると，距離 D は次式で表すことができる．

$$D = \sum_n r[n]\,|\sin(\theta[n] - \theta_k[n])| \quad (2)$$

(d) $r\Delta\theta_i (i = 1, \cdots, 12)$ を計算 $\Delta\theta_i (i = 1, \cdots, 12)$ は線分テンプレートの線分 i との反時計回りの偏角差を示す．

D が最小となるように回転した線分テンプレートと $\boldsymbol{vec}$ の分布を図 6 に示す．各線分に対応する移動ベクトルが色分けされている．また，これを時系列に並べ替えて可視化した図を図 7 に示す．各移動ベクトルが，その時刻に対応する軸上の点を起点に描かれている．こうして算出された特徴量を基に，表 3 に示す特徴量を用いて認識実験を行った．実験 1 と実験 2 の実験結果を図 8 に示す．Corr は指示動作認識率，

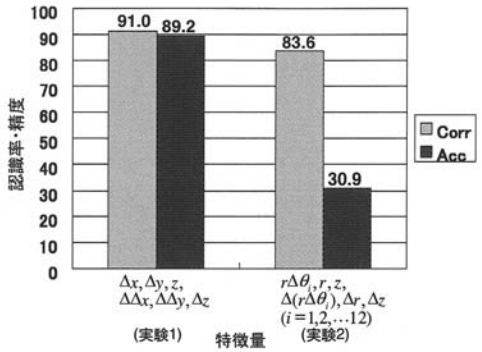

図 8: 移動ベクトルを用いた動作認識結果

Acc は認識精度を表している．実験 1 の場合の方が認識率，認識精度ともによい結果が得られた．実験 1 の場合，「∠DCA」などの指示動作軌跡が線分から成る図形要素を指し示す指示動作の認識は 100%正解が得られ，**3.2.1** 節の位置情報を用いた場合よりもよい結果が得られた．しかし，移動ベクトルでは，位置情報が失われるため，頂点を表す動作の誤りがあった．実験 2 では，$r\Delta\theta$ という特徴量を用いたため，線分を示す動作の始めや終わりの速度が遅い部分を点と認識してしまい，認識精度が悪い．しかし，線分の相対角度を用いることには，学習データがない場合に簡単にモデルを作ることができる，問題図が xy 平面上で並進移動や回転移動をした場合にも正規化など特別な処理をする必要がないという利点がある．問題図に固有な線分の相対角度の情報を効果的に用いるために，指示動作の速度が遅い部分も対応できる特徴量を検討する必要がある．

4 音声と指示動作の統合認識

音声と動作は相互に補完しあう関係にあり，音声のみあるいは動作のみからでは表現したい内容がわからない場合があることから，表現内容を正しく理解するためには音声と動作の対応関係を把握した上で，統合的に発話動作を理解する必要がある．そこで，同じ図形要素を指し示している音声区間と動作区間を対応付け統合的に認識する手法について検討する．

4.1 時間情報を利用した統合

一般に動作と発話の開始 (終了) の時刻は一致しない．図 9 の例では，対応する発話区間と指示動作区間では，発話区間の方が指示動作区間よりも遅れて始まる傾向があることがわかる．

$$\tau = (\text{発話の開始時刻}) - (\text{指示動作の開始時刻}) \quad (3)$$

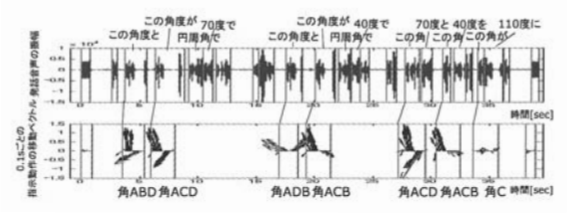

図 9: 発話と指示動作の対応関係

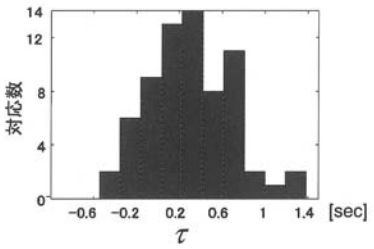

図 10: 発話と指示動作の開始時刻差のヒストグラム

の分布を図 10 に示す．図 9，図 10 から，(発話区間の開始時刻) と (指示動作区間の開始時刻) の差 τ には，確率的な偏りがあると考えられる．そこで，τ がガウス分布に従うという仮定のもと，ガウス分布を発話区間と指示動作区間の対応付けに利用した．具体的には，**3.1** 節の音声認識の結果得られた認識結果文からキーワードを含む区間を抽出してその区間の発話開始時刻を参照し，**3.2.1** 節の指示動作抽出手法により抽出した指示動作部分の開始時刻を参照して対応付けを行った．ここで，キーワードとは，「ここ」「この角」「40 度」などの，発話する際に指示動作を伴い図上で説明する可能性のある語のことである．

各指示動作区間に対して τ の分布を近似したガウス分布にしたがって確率的にキーワードを含む発話区間と対応付け，1 つの指示動作区間に対し複数の対応候補となる発話区間が存在する場合には，ガウス分布に基づいて対応する確率の高い発話区間を選ぶ．どの指示動作区間とも対応付けられなかった発話区間は対応する指示動作区間がないと判定した．対応付けの「正解」を以下のように定義し，発話中の全てのキーワードに関して正解の対応付けをすることができた場合を正解率 100%として対応付けの性能を評価した．

- 対応する指示動作のあるキーワード：
  正しく発話区間と指示動作区間の対応付けがとれ，尚且つ指示動作から正しく指示対象の図形要素を認識できた場合に正解

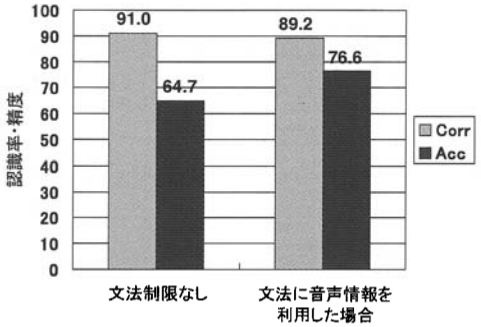

図 11: 音声認識結果による文法制約を加えた場合の動作認識結果 (移動ベクトルを利用)

表 4: 単語から動作への変換ルール

発話中の単語	例	変換する動作仮説
指示語	「ここ」	全ての角，線分，頂点
角度を示す単語	「この角」	全ての角
線分を示す単語	「この線分」	全ての線分
点を示す単語	「この点」	全ての頂点
頂点名を用いた単語	「角 ADB」	特定の図形要素を示す動作
図形と関係しない単語	助詞，動詞	sil

- 対応する指示動作のないキーワード：
  各指示動作区間に対して対応する発話区間を選んだ結果，対応する指示動作区間がないという結果が得られた場合に正解

2 節に述べた**条件 2** の 9 回分の収録のうち 1 回分を評価用データ，残りの 8 回分を学習用データとし，評価用データを入れ替えて実験を行った．その結果 88.7%の対応付け正解率を得た．

4.2 音声情報を利用した指示動作認識

音声と指示動作は互いに情報を補完し合っていることから，動作の認識時に音声の認識結果を利用し，動作の指示対象候補を絞ることで，より正確な動作認識をすることができると考えられる．そこで，表 4 に示す単語から動作への変換ルール (単語が示しうる動作候補を記述したもの) を作成し，音声の認識結果文から，その変換ルールに従って動作認識用のネットワーク文法を生成する．そのネットワーク文法によって受理される指示動作系列のみが認識されるように制約を加えた．また，認識結果からのネットワーク文法の生成例を図 12 に示す．

3.2.2 節の実験 1 の条件において，音声認識結果による文法制限を加えて実験した結果を図 11 に示す．文法制限なしの場合に対し，認識率は 1.8 ポイント低下したが，認識精度が 10.8 ポイント向上した．

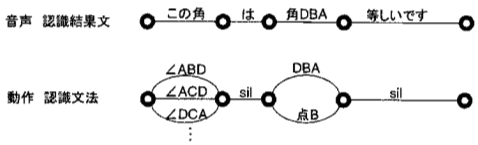

図 12: 音声認識結果からの動作認識文法作成例

意図せず問題図の紙面上に触れている際に図形要素に関連する単語を発していた場合を含む解説に関しては認識率が低下したが，その他の解説は認識率も向上した．

5 まとめと今後の課題

図形問題の解説を対象として発話と指示動作の統合手法を検討した．発話と指示動作の開始時刻の差の確率的な偏りを利用して対応付けを行い，88.7%の正解率を得た．また，音声認識結果から，起こりうる動作のネットワーク文法を作成し，動作認識に利用することで動作の認識精度が向上した．この結果に基づいて，発話，動作，各々が指し示す意味と時間的な対応関係を利用して発話と指示動作の統合認識を行うことが今後の課題である．

参考文献

[1] R.A. Bolt, "Put-that-there: Voice and gesture at the grahics interface," ACM Computer Graphics, Vol.14, No.3, pp.262–270, 1980.

[2] N. Krahnstoever, S. Kettebekov, M. Yeasin, and R. Sharma, "A real-time framework for natural multimodal interaction with large screen displays," Proc. ICMI 2002, Oct. 2002.

[3] P. Hui, and Helen M. Meng,"Joint interpretation of input speech and pen gestures for multimodal human computer interaction," INTERSPEECH-2006, pp.1197-1200, Sep, 2006.

[4] L. Wu, S.L. Oviatt, and P.R. Cohen,"Multimodal integration-a statistical view," IEEE Trans. Multimedia, Vol.1, No.4, pp.334–341, 1991.

[5] S. Kettebekov, M. Yeasin, and R. Sharma, "Prosody based co-analysis for continuous recognition of co-verbal gestures," Proc. ICME, 2002.

[6] 中井 満, 嵯峨山 茂樹, 秋良 直人, 小場 久雄, 下平 博, "ストローク HMM によるオンライン手書き文字認識の性能評価," 信学技報, PRMU2000-36, pp.9–16, June 2000.

[7] 鹿野, 伊藤, 河原, 武田, 山本,"音声認識システム," オーム社, 2001.

[8] 大語彙連続音声認識エンジン Julius, http://julius.sourceforge.jp/