図像情報を利用した講演調音声のディクテーション

7N－6

池田　徹志　荒木　雅弘　堂下　修司

京都大学大学院工学研究科情報工学専攻

1　はじめに

人間は自分の思考内容を他人に伝達するために，音声・図・ジェスチャ等複数のモードを適切に用いている．また，モードを組み合わせることにより，効率の良い伝達や，信頼性の高い伝達を行なっている．

近年，このようなモードの統合の現象に興味を持った研究がさかんに行なわれている．しかし，図像に表現されている深い意味を音声認識に積極的に利用する試みはない．

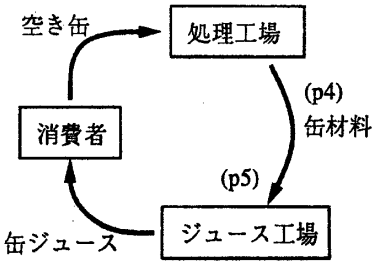
本研究では，図像を指示しながら行なわれる口頭発表の音声のディクテーションをタスクとして設定する．OHPに代表される図像の作成はコンピュータ上で行なわれるのが主流になってきており，提示される図像の情報として誤りのないものを利用できると考えられる．さらに本研究で利用する図像は，図像の意味する内容をも含む高次の表現がなされていると仮定する．これは作成者が図像の深い意味を指定しながら作図を行なうような，高度な作図システムを利用することで可能となる．我々はそのような作図システムの研究も並行して行なっている．

2　図像を用いたプレゼンテーション

本研究で利用する図像は，線・矢印といった部品の単なる集合ではなく，各図形シンボルの表している物の属性や図形シンボル間の関係の属性といった高次の表現がされている．図像中のテキストも図像要素の属性の１つである．図像と高次の表現の例を図１・図２に，それを指示しながら行なった発話を書きおこしたものを図３に示す．図１の(p4)は，図３の <p4>...</p4> の部分を発話している際にクリックした位置を示している．

ここで，図形要素の属性としてはEDR概念辞書などのシソーラス上の概念を与え，要素間の関係の属性としては各シンボルの属性および関係の属性を与える．要素間の関係の属性は，図の種類に依存して属性の種類が異なる．例のような要素間を結んだ図では，矢印やリンクが要素間の関係を表す．属性としては物の移動，処理の

Dictation of Spoken Language in Lecture Style using Diagramatic Information
IKEDA Tetsushi and ARAKI Masahiro and DOSHITA Shuji
Department of Information Science, Graduate School of Engineering, Kyoto University, 606-01, JAPAN
E-mail: tetsushi@kuis.kyoto-u.ac.jp



表題：ゴミのリサイクルについて

図 1: プレゼンテーションの際に用いる図像の例

```
(elem1
  (type 人間) (label "消費者"))
(elem2
  (type もの) (label "処理工場"))
(elem3
  (type 人間) (label "ジュース工場"))
(arc1
  (type 物の移動) (from elem1) (to elem2)
  (label "空き缶"))
(arc2
  (type 物の移動) (from elem2) (to elem3)
  (label "缶材料"))
(arc3
  (type 物の移動) (from elem3) (to elem1)
  (label "缶ジュース"))
```

図 2: 図１に付随する図像の高次表現

ごみのリサイクルについて，空き缶を例にとって説明します．消費 <p1> 者によって </p1> 消費されたジュースのごみは，<p2> 空き缶となって </p2><p3> 処理工場に運ばれ </p3> ます．処理工場では缶の材料の鉄やアルミなどが取り出され，<p4> 新たな缶の材料と </p4> してジュー <p5> ス工場に運ば </p5> れます．ジュース工場では再び <p6> ジュースが缶になってつめられて </p6>，<p7> 消費者のもとに </p7> 届けられます．

(... は指示が行なわれた区間を表す．)

図 3: 図１を用いたプレゼンテーションの発話

流れ，対応，適用，論理的含意などが挙げられる．

本研究ではこのような図に対して指示を行ないながら発話するという状況において，音声の認識を行なうことを目的とする．指示の情報を利用することで，図像と音声の対応をとることができる．対応がとれた音声区間に対しては，図像に表現されている意味表現を利用することで音声認識の精度を向上させることができると考えられる．

ここで問題となるのが両者の対応の曖昧性である．一般には発話と指示は必ずしも同期をとって行なわれるわけではない．また，図像上の指示対象についても曖昧性が生じる．

3 時間・空間・意味スコアによる情報統合

個々の指示は本質的に音声中の1つの自立語と図像中の1つの要素を対応付ける物と考え，指示と自立語と図像要素の任意の組に対してコストづけを行なう手法を提案する．以下ではこの組を「対応」と記す．

3.1 音声・図像の対応へのコスト付与

指示と対応する音声区間は時間的に必ずしも一致しないという結果が報告されている[1]．しかし人間は同期するのを最も自然と感じ，指示と音声が時間的に離れるにしたがって両者が対応するのを不自然と感じる．したがって時間的ずれの程度に対しコストを持たせることが有効であると考えられる．

時間コスト = Ctime(指示時間，音声区間)

図3の発話の一部の時系列を図4に示す．音声認識の結果は文節単位で扱われ，1文節中に自立語は1つである．図4の1回めの指示は図1の(p1)印に対して行なわれたものである．この例では1回目の指示と各単語との5通りの組合せそれぞれに対しコストが与えられる．ここでは対応2が最も低いコストになる．しかし時間コストは悪いけれども後述する意味的整合性は良いという対応が存在することもあり，どの対応が良いかは情報を統合して判断を行なう．

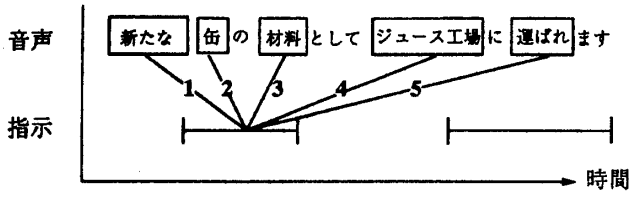


図 4: 音声・指示の入力の例

また，指示している位置と図像要素の間の曖昧性に対してもコストを導入する．

空間コスト = Cspace(指示位置，図像要素位置)

3.2 意味効果へのスコア付与

指示を介して対応付けられる自立語の音声区間と図像要素の意味的整合性に対してスコアを付与する．意味的整合性は，図像要素に付けられた属性(テキストも含む)と認識された単語との間の距離を，シソーラスを用いて与える．EDRのシソーラスは概念を表すノードの上位下位の関係にあるものの間の接続で表され，従来の単語が全てリーフノードに分類されるシソーラスと比べて距離の定義が困難であるが，単語間の距離の定義がいくつか提案されている([2]など)．

意味的整合性スコア
= Ssem(シソーラス上での両者の距離)

3.3 最良の対応の探索

以上のように求めたコスト関数Ctime, Cspace及びスコア関数Ssemより，対応全体の評価関数Fを定義する．

$$F = F(Ctime, Cspace, Ssem)$$

このFを最大にする対応を正しい対応と考える．

4 音声認識結果の再評価

N-bestの音声認識結果のそれぞれに対して，最高の評価値となる対応を求める．評価値はそれぞれの音声認識結果の図像・指示との整合の度合を表していると考えられる．そこで音響スコアとこの評価値の両方を考慮することによりN-bestの認識結果の再順序づけを行なう．このように両方のスコアを統合することにより，

- 対応が確かなときは図像情報を積極的に利用する
- 不確かなときには通常の音声認識を行なう

という処理が境目なく実現される．

5 むすび

本稿では図像情報と指示情報を音声認識の精度向上に積極的に用いる手法の枠組を提案した．これから実際にプレゼンテーションのデータに基づいて各種コスト関数の形状等を決定し，本手法の有効性を検証してゆく．

参考文献

[1] K.H. Loken-Kim, Suguru Mizunashi, Mutsuko Tomokiyo, Laurel Fais, and Tsuyoshi Morimoto. 翻訳通信環境におけるマルチモーダル入力の分析と統合. 95-SLP-7-12, 1995.

[2] 中山聡，峯恒憲，東優，谷口倫一郎，雨宮真人. EDRコーパスを利用した動詞の語義分類. 電子情報通信学会技術研究報告, NLC95-43, pp. 23–30, 1995.