自 然 言 語 処 理 149－20
音声言語情報処理 41－11
（２００２．５．２４）

汎用 LVCSR を用いた対話音声の認識

新田 恒雄　浅見 弘道　伊勢路 真吾　福田 隆　桂田 浩一

豊橋技術科学大学　大学院工学研究科
〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1
E-mail : nitta@tutkie.tut.ac.jp

あらまし　本報告では汎用 LVCSR（ディクテーション用）ソフトウエアを利用して，対話文音声を高精度で認識する方式を提案する。提案方式は，LVCSR が出力する音素系列を弁別的な特徴ベクトル系列に変換した後，対話管理部から指示される対話記述（語彙と文法知識）を利用して，キーワードをスポッティングする。本方式の特長は以下の二点にある。(1) 言語モデルの制約を緩めて，LVCSR の持つ高い音素識別能力を最大限に利用している。(2) 音素系列出力を弁別的な特徴ベクトル系列に置き換えた後 DP マッチングを適用し，置換・脱落・付加誤りに対処している。本文では，道案内タスクの対話文音声データを用いて比較評価実験を行い，提案方式の有効性を示す。

キーワード　音声対話，LVCSR，VoiceXML，言語モデル，ABNF，弁別的特徴

Recognition of Spoken Dialogue by Using General Purpose LVCSR

Tsuneo Nitta, Hiromichi Asami, Shingo Iseji, Takashi FUKUDA, and Kouichi Katsurada

Graduate School of Engineering, Toyohashi University of Technology
1-1 Hibariga-oka, Tempaku, Toyohashi, 441-8580 JAPAN
E-mail : nitta@tutkie.tut.ac.jp

Abstract　This paper describes an attempt to recognize spontaneously spoken dialogue by using general purpose LVCSR (Large Vocabulary Continuous Speech Recognition) software. In the proposed method, a phoneme string outputted from LVCSR is converted into a sequence of vector represented with distinctive features, then keywords assigned by a dialogue manager are detected from the input vector sequence. The method takes advantages of the potential abilities of: (1) precise phoneme-discrimination achieved by relaxing the linguistic constraint in LVCSR, (2) coping with the issues of substitution, deletion and insertion errors by combining a conversion process from a phoneme into a distinctive feature vector and a key-word spotting process. The proposed method shows significant improvements in comparison with an LVCSR software in a experiment with a spoken dialogue corpus of a map guidance task.

Key words　Spoken Dialogue, LVCSR, VoiceXML, Language Model, ABNF, Distinctive Feature

1. はじめに

読み上げ文音声認識に続いて，対話文音声認識の実用化を目指す研究開発が始まっている。対話文音声認識へのアプローチとしては，読み上げ文音声で成功した方式，すなわち N-gram 言語モデルの強い拘束条件を適用して単語を限定し, HMM 音響モデルで表現した単語列を入力音声から決定する方式[1]が，多く試みられている。このアプローチでは，対話文に現れる語彙を，あらかじめ言語モデルに組み込む必要がある。このため，Web 上のサービスに見られるように，対話が不特定多数のトピックに遷移する場合には，言語モデルを頻繁に実時間で再構成する必要があり，このことは実用化を妨げる要因の一つになっている。一方，W3C VoiceBrowser WG では，音声ブラウジングの実現を目指して, VoiceXML 2.0 を策定中である[2]。この中には，N-gram 言語モデルと共に，語彙と文法を ABNF (Augmented BNF) 形式等により認識エンジンへ渡す手段が規定されており(SRGS : Speech Recognition Grammar Specification [3]), アプリケーションに関する語彙と文法知識を如何に巧く対話文音声認識へ利用するかが今後の課題となっている。

他方，対話音声では文中の息継ぎ・息漏れ，話し言葉特有の音響現象，様々な話し言葉表現，あるいは不要語や未知語（言語モデルがカバーしていない語彙，あるいは対話管理部から渡されるキーワードにない語彙）が出現するため，言語モデルの対応だけで高精度認識を実現することは困難である。こうした課題を解決する有力な方法の一つに，対話音声中のキーワードをスポッティングする方式がある[4], [5]。しかしこの方式は，正解単語区間外で多量の単語を湧き出すという新たな問題を伴うため，実用化は語彙数が極めて少ないか，もしくは構文規則から認識対象を小量の語彙に絞れる場合に限られる。さらに，入力音声（特徴ベクトル）に沿って，キーワード毎に端点フリーマッチングを行う必要があるため，演算量の多さも問題である。

以上に述べた対話音声認識構成上の課題に対処するため，我々は汎用 LVCSR (Large Vocabulary Continuous Speech Recognition)ソフトウエアが潜在的に持つ高い音素識別能力と，対話制御部から指示される対話記述（語彙（キーワード）と文法知識）の双方を活用して，対話文音声中の単語を高精度で認識する方式を提案する[6]。本報告では，まず 2 節で LVCSR の出力から音素系列を得た後，これを弁別的な特徴ベクトル系列に変換し，この中から対話記述に合致するキーワード列を抽出する方式を説明する。続いて 3 節では，実験条件と結果を示し考察を加える。

2. 音声対話システムの概要

2.1 提案方式の背景

現在の LVCSR ソフトウエアは，時折，発話内容と音韻的に近い間違いをおかす（利用者から見えるのは単語列であるが)。これは音響モデルの精度が，大語彙中の類似単語を確実に識別できる程度には至っていないことと，言語モデルの不整合という二つの理由に依る。一方，発話に未知語が含まれる場合には，おかしな（音韻的に近いとは言えない）単語列が出力される。これは LVCSR ソフトウエアの認識性能が，主に言語モデルの強い拘束に依存していることと関係している。

図 1 に,「ケンタッキーフライドチキンに行きたいんですけども」と発話した例を示す。言語モデルが trigram, bigram の場合を見ると，漢字かな混じり文については，意味的に通じる文である。しかし，( ) 内のかな文字文は，元の音声とかけ離れた結果になっている。

一方，言語モデルの拘束を和らげた unigram の結果を見ると，漢字混じり文は意味不明であるが，かな文字文からは内容を十分理解できる。このように，音響モデルの性能が一定のレベルに達し，かつ大語彙をカバーする LVCSR ソフトウエアは，未知語（ここではケンタッキーフライドチキン）を含む発話文が入力されても，言語モデルの強い制約を緩和することで，意味的に理解可能な音節列（音素列）を出力する能力を持つことが

trigram の場合：
現在 、 大豆 時期 荷 期待 する
（げんざいだいずじきにきたいする）
bigram の場合：
現在 、 ドイツ 人 時代 です
（げんざいどいつじんじだいです）
unigram の場合：
建 他 樹 歩 羅 伊豆 血 琴 遺棄 ― ん す
（けんたきふらいずちきんいきたいんす）

図 1　LVCSR ソフトウエア出力の例
入力音声：**「ケンタッキーフライドチキン**
に行きたいんですけども」

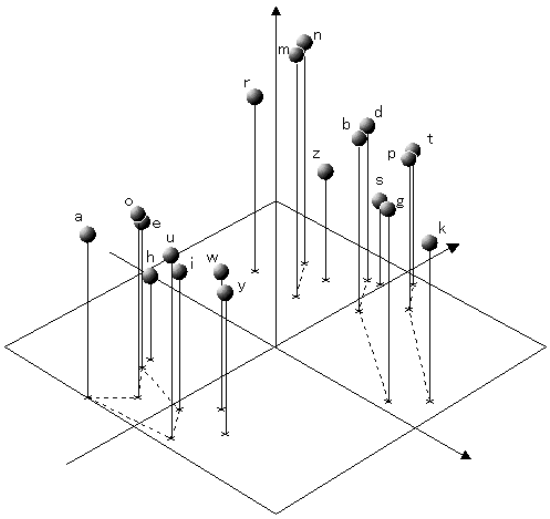

図 2　MDS による三次元弁別特徴空間

分かる。

これまでに，言語制約の強い汎用 LVCSR ソフトウエア（ディクテーションシステム）の出力から，音声データを検索する方式が提案されている[7], [8]。これらの方式では，LVCSR が出力する音素系列に含まれる置換・脱落・付加誤りに対処するため，置換誤りに対しては混同行列を，また脱落・付加に対しては DP マッチングを適用している。音素間の混同行列は，音声コーパスを用いて設計されるが，この場合，設計時と利用時とでは音響諸条件が異なることが問題になる。そこで本報告では，音響環境に依存しない方法として，弁別的な特徴を利用する方式を検討する。

図 2 は，日本語音素の弁別特徴[9]を多次元尺度構成法(MDS) により，元の 11 次元空間から 3 次元空間に圧縮して, 音素の布置をみたものである。この図を見ると，標準的に利用される弁別特徴は母音グループが隔離され過ぎていることが分かる。理由は“ 母音性／非母音性 ”と“ 子音性／非子音性 ”という二つの弁別素性が採用されているためである。分離だけに着目するなら，母音と子音は残りの弁別素性からも十分可能である。そこで国際音声記号表を参考にして，この二つの代わりに“ 半母音性(/ j, w, r /)／非半母音性 ”と“ 摩擦性(/s, z, h /)／非摩擦性 ”を入れたものを，ここでは改めて「弁別的な特徴」と呼び変えて使用する。

弁別特徴は古くから音声認識システムに組み込まれ利用されてきた[10]。音素を単位とする認識方式の最大の課題はセグメンテーションである。提案方式は，LVCSR が音素セグメンテーションと音素識別に対して持つ，潜在的な能力を利用することを狙っている。

2.2 提案システム

図 3 に音声対話システムの全体構成を示した。このうち対話文音声認識サブシステムは，大きくフロントエンド部（汎用 LVCSR ソフトウエアを使用），音声言語処理(SLP)部，および対話管理部から構成される。システムの構成要素中，アプリケーションに依存する部分は, 対話管理部の対話シナリオのみである。

入力音声は，まず言語モデルの制約を unigram に緩和した LVCSR で認識処理され，結果の 1-best の音素系列が音声言語処理部に送られる。LVCSR と対話管理部との間に設けられた音声言語処理(SLP)部は，LVCSR をアプリケーションから独立にすると共に，対話管理部で解釈可能なキーワードだけを入力音素系列から抽出して渡す役目を持つ。具体的には，LVCSR が出力する音素系列を弁別的な特徴(11 次元)ベクトル x(m, i), m=1, 2,…11, i=1,2,…, I に変換した後，この中からキーワードを抽出する。対話管理部から指示されるキーワード k は，同様に音素系列に置換えた後, 弁別的な特徴ベクトル系列 r_k(m, j),

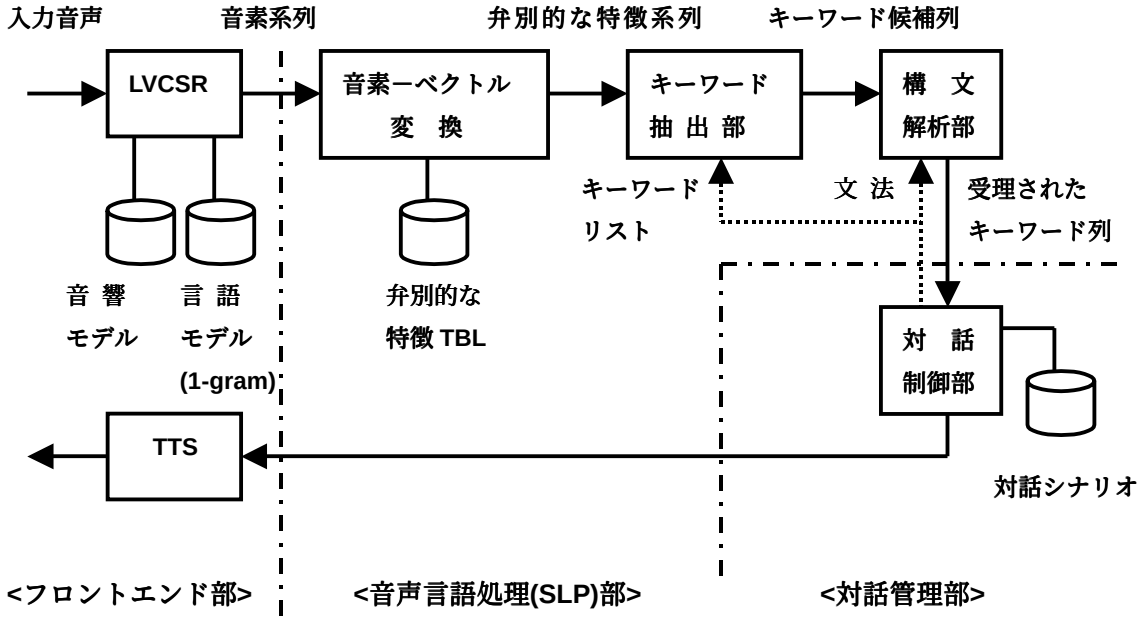

図 3　音声対話システムの全体構成

m=1,2,…11, j=1,2,…J の形に展開しておく。入力 x(i, j) からキーワード $r_k(i, j)$ を求める際には，以下の音素間距離（ハミング距離）と漸化式を用いた。

$$d_k(i, j) = \sum_{m=1}^{11} \{ x(m, i) - r_k(m, j) \}^2 \qquad (1)$$

$$g(i, j) = \min \begin{cases} g(i-1, j) + d_k(i, j) \\ g(i-1, j-1) + 2d_k(i, j) \\ g(i, j-1) + d_k(i, j) \end{cases} \qquad (2)$$

$$D(i) = g(i, J)/(i+J) \qquad (3)$$

端点フリーDP マッチングの結果，距離 D(i) が一定の閾値以下のキーワードを抽出する。また，過剰な湧き出しを抑えるため，抽出区間に一定の重なりがある場合，最も距離が小さい候補のみを残した。なお，漸化式は他の形式も試したが，今回は上記のものが最も良い結果を与えた。抽出されたキーワード列は，構文解析部に送られ，対話管理部が提供する文法に適合するか否かの判断が行われる。文法は ABNF を用いた。最後に，対話管理部は受理した入力に対して，タスク記述に従い応答を返す。

3. 評価実験

3.1　音声試料

評価データセットは，電総研の道案内対話音声コーパス[11]のうち話者（男女）14 名の 100 発話（全発話時間 305[sec]）を使用した。

3.2　実験概要

LVCSR には日本語ディクテーションシステム Julius (2001 年度版)[12]を使用する。Julius は 2 パス探索を行い，1st パスに bigram，2nd パスに trigram を用いている。音響モデルは，特徴パラメータに“ MFCC+Δt＋ΔP （25 次元）”を, また HMM に 2000 状態の tri-phone モデル（対角化共分散，性別非依存モデル，混合数 16）を使用した。言語モデルは，毎日新聞の記事データ 75 ヶ月分（1991.1～1994.9, 1995.1～1997.6，約 118M 単語）を用いて設計したものを使用した（語彙数 20ｋ）。

対話管理部から与えるキーワードは，109 単語（異なり語）を用意した。今回の実験では，タスク達成に必要な単語という基準で，キーワードを直接評価データの書き起こしテキストから選んでいる。このうち LVCSR に登録されて

いる単語は 66 語で，43 語が未知語である。

音素－ベクトル変換では, LVCSR(Julius)が出力する ATR の音素表記を使用した。このため，弁別的な特徴テーブルもこれに合わせて作成した。

構文解析で用いた ABNF は，場所を尋ねる質問，店の内容を尋ねる質問，付近の店を尋ねる質問，および時間や距離を尋ねる質問に対応する規則を作成した。

［例］店 A が n 階にあるか尋ねる

$ask_place_floor = $place $floor $teach_where

3.3 評価実験

実験は，Julius の言語制約を表 1 のように変化させて行った。評価は，キーワード抽出部の出力，構文チェック後の出力，および意味的な正解判定から行った。評価基準は，単語正解率（ (N－S－D)×100 /N : N,　S，D は各々全キーワード数，置換，および脱落数），FA / WH（1 キーワード当たりの単位時間湧き出し数）の二つを用いた。なお，キーワード抽出部の出力については，キーワードが LVCSR で登録されている場合と，未知語の場合を分けたものについても評価した。

表 1. 実験条件（言語制約）

実験	言語制約（Julius における設定）
A	2nd パス（trigram）を使用 1st: 言語モデル重み 8, 単語挿入ペナルティ-2 2nd: 言語モデル重み 8, 単語挿入ペナルティ 2 ※デフォルト設定（言語制約の変更なし）
B	1st パス（bigram）のみ使用 1st: 言語モデル重み 8, 単語挿入ペナルティ-2
C	1st パス（unigram）のみ使用 1st: 言語モデル重み 0, 単語挿入ペナルティ-5

3.4 実験結果

キーワード抽出部の出力に対する実験結果を表 2 に示す。言語制約を unigram のみとした C は，LVCSR にキーワードを全て登録した時の結果（表の A-登録）と同等の，高い単語正解率を得ることが出来た。特に，表 2 の内訳にある LVCSR での辞書登録／未知語の違いをみると，言語制約を弱めたことの効果は，未知語の部分に対して非常に大きい。A，B で未知語に対する単語正解率が悪くなった理由としては，言語モルの強い制約により，音響モデルで出力した音素が別の音素へ置き換えられたためと考えら

表 2. キーワード検出結果

実験	置換数	脱落数	単語正解率 [%]	内訳 登録語	内訳 未知語	FA/WH
A	36	6	83.5	86.3	75.0	48.0
B	47	8	78.3	79.5	75.0	49.8
C	29	4	87.0	85.8	90.6	48.5
CM1	49	6	78.3	75.6	85.9	47.1
CM2	47	9	77.9	75.2	85.9	45.3
A-登録	25	5	88.2	—	—	49.3

表 3. 構文規則で受理／棄却した後のキーワード検出結果

実験	置換数	脱落数	単語正解率[%]	FA/WH
A	46	26	71.7	1.73
B	47	31	69.3	1.08
C	33	29	75.6	1.08
CM1	45	42	65.8	1.08
CM2	44	42	66.1	1.41
A-登録	30	29	76.7	1.08

れる。因みに，LVCSR の出力（1 位）中に正解キーワードが含まれる率は，全てのキーワードを登録した場合で，55.1%であった。なお，表中 CM1, CM2 は，混同行列（各々JNAS と評価データから作成）を使用した場合の結果を示しているが，今回はデータ数が少なかったこともあり，低い正解率しか得られなかった（追試験を予定)。

構文解析部の出力に対する実験結果を表 3 に示す。表 2 の結果と同様，C が最も高い単語正解率を示した。構文規則を適用することで，FA / WH は大幅に減らすことができるが，一方，正解キーワードの脱落も増えたことで，表 2 と比べるとキーワード検出性能が低下している。

最後に，対話文が意味的に正しく受理された割合を表 4 に示す。例えば「パルコは何処ですか？」という入力に対して，正解スロットは「パルコ－何処」であるが，「パルコ－行き方－教えて」に誤って受理された場合も正解としている。誤って受理される例の多くは，店名・場所の誤り（キーワードの置換誤り）であった。

表 4. 意味的に正しく受理された文の割合

実験	正解率[%]	誤り率[%]	棄却率[%]
A	62	32	6
B	65	29	6
C	75	19	6
CM1	64	28	8
CM2	65	26	9
A−登録	83	12	5

4. まとめ

アプリケーションから独立した汎用 LVCSR が出力する音素系列に対して，対話管理部が提供するキーワード および文法を用いて，対話音声を高精度で認識する方式を提案した。提案方式は次の二つの特長を持つ。

(1) LVCSR が潜在的に持つ高い音素識別能力を，言語モデルの制約を緩めることで引き出した。

(2) 音素系列を弁別的な特徴ベクトルに置換え，端点フリーDP マッチングを適用したことで，音響環境への依存の少ない方式が得られた。

今後は，言語モデルの単語間制約からサブワード間の制約への置き換え[13]，混同行列との比較を行うと共に，より実用的な音声対話システムを目指して，キーワード抽出部と構文解析部の改良を行い，マルチモーダル対話システム[14]への実装を目指したい。

参 考 文 献

[1] 鹿野ほか編著：音声認識システム，オーム社(2001).

[2] http://www.voicexml.org/

[3] http://www.w3.org/TR/speech-grammar/

[4] J. R. Rohlicek, W. Russel, S. Roucus, and H. Gish, “Continuous HMM for speaker independent word spotting”, Proc. ICASSP, pp. 627-630 (1994.5).

[5] 神尾，松浦，正井，新田：マルチモーダル対話システム MultiksDial，信学論 J77-D-II，8, pp.1429-1437 (1994. 8).

[6] 浅見，福田，桂田，新田：汎用 LVCSR を用いた対話音声の認識について，音学講論 1-5-25, pp.49-50 (2002.3).

[7] 前田，島津：音素認識に基づく音声全文検索，人工知能学会研究会資料，SIG-SLUD-A102-1 (2001.11)

[8] 西崎，中川：未知語を考慮したニュース音声記事の検索，情報処理学会研究報告，SLP-39-29, pp.171-176 (2001.12).

[9] 早田輝洋，“日本語音形論”，東北大学電気通信研究所 第 8 回シンポジウム論文集，音声情報処理，I-2, pp.I-2-1 – I-2-26 (1971.2).

[10] S. Makino, S. Homma, and K. Kido, “Speaker independent word recognition system based on phoneme recognition for a large size (212 words) vocabulary, J. Acoust. Soc. Jpn., (E) 6, 3, pp171-180 (1985).

[11] 伊藤 他，音学講論 1-1-19, pp.37-38 (1998).

[12] 河原 他，音響学会誌，Vol.57，No.3，pp.210-214 (2001).

[13] K. Ng, ”Toward Robust Methods for Spoken Document Retrieval”, Proc. ICSLP, pp.939-942 (1998.11).

[14] 桂田他，情処研究報告 2002-SLP-40，pp.51-56 (2002-02).