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

2003－ＳＬＰ－49　(28)
2003／12／18

音素弁別特徴間距離に基づくキーワード検出における音節単位サブワードモデルの検討

伊勢路 真吾　福田 隆　山田 博文　桂田 浩一　新田 恒雄

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: {iseji, fukuda, yamada}@vox.tutkie.tut.ac.jp, {katsurada, nitta}@tutkie.tut.ac.jp

あらまし　言語モデルをサブワードモデルに置換えた汎用 LVCSR エンジンを利用し，対話音声中のキーワードを高い精度で抽出する方法を研究している．これまで，汎用エンジンが持つ固有言語モデルの制約を緩めると共に，認識結果の音素列に音素弁別特徴 (DPF) ベクトルに基づくキーワードスポッティングを適用する方式を提案し，道案内タスク対話音声中のキーワード検出実験から，この方式が置換・脱落・付加誤りの少ないことを報告した．本報告では，日本語の全ての音節構造を考慮したサブワードモデルを利用することにより，少ない語彙数（20k 辞書の 1/40）で高い精度を維持することができることを示す．これによりメモリと計算負荷を大幅に低減できる．汎用エンジンの 20k 辞書 (0-gram)との比較実験では，最初に(a) 汎用エンジン中の音節長の短い単語のみを辞書に登録（辞書サイズ 1/2）することで同等の性能が得られることを示す．次に，(b) 辞書中の単語から短音節単位サブワードモデルを作成することを検討し，2 短音節の 3-gram で同程度の性能が得られることを示す．最後に，(c) 新聞記事から得られる日本語の全ての音節構造を検討し，短音節，長母音音節，撥音付き音節，促音付き音節，二重母音音節をサブワードとして採用することで，20k 辞書と比較し 1/40 の登録項目で同等の性能が得られることを示す．

キーワード　音声対話，キーワードスポッティング，音素弁別特徴，音節サブワードモデル

A Keyword-Spotter Based On Distinctive Phonetic Feature Vectors For Application-Independent ASR Engines

Shingo ISEJI, Takashi FUKUDA, Hirobumi YAMADA, Koichi KATSURADA, and Tsuneo NITTA

Graduate School of Engineering, Toyohashi University of Technology

1-1 Hibariga-oka, Tempaku, Toyohashi, 441-8580 JAPAN

E-mail: {iseji, fukuda, yamada}@vox.tutkie.tut.ac.jp, {katsurada, nitta}@tutkie.tut.ac.jp

Abstract　This paper describes a method for spotting key-words in spontaneous speech using a general-purpose LVCSR engine with sub-word model and distinctive phonetic feature (DPF) vector. The proposed method takes advantage of the potential abilities of (1) precise phoneme recognition in the LVCSR engine and (2) coping with the issues of substitution, deletion and insertion errors by combining a process of conversion from a phoneme into a DPF vector and a key-word spotting process. In this report, firstly we show (a) a language model (LM) of word sequences selected from an original LM of the LVCSR engine with 20k vocabulary words and composed of within three Japanese short syllables gives equivalent performance in comparison with the 20k-words LM. Next, we design a sub-word LM using the 20k vocabulary words and show (b) a trigram LM of sub-word sequences composed of within two Japanese short syllables can hold comparative performance. Finally, we investigate all the Japanese syllable structures using a news paper corpus and show (c) a bigram LM of sub-word sequences that consists of Japanese short syllables (CV) and long syllables with long vowels (CVV), independent nasal sound (CVN), glottal stop (CVQ), and diphthongs (CV1V2). The proposed sub-word LM that contains all the Japanese syllable structures achieves high performance with only 1/40 vocabulary size of the LM in comparison with the 20k-words LM of the LVCSR engine.

Keyword　Spoken Dialogue , Keyword Spotting, Distinctive Phonetic Feature, Syllable-based sub-word Model

1. はじめに

VoiceXML[1]を利用した音声対話による Web サービスが始まっている．これらのサービスではトピックがページ毎に遷移するため，未知語への対応やトピックに適した言語モデルの構築が困難であるなどの問題を抱えている．また，リアルタイムにモデルを適応(オンライン適応)する場合には，十分なドメイン対応コーパスの収集と，モデル設計の計算時間が問題になる．一方，単語 N-gram に頼らず，ABNF (Augmented BNF) 形式などにより文法を認識エンジンに渡す手段も規定されているが (SRGS : Speech Recognition Grammar Specification [2])，Web サービスのような音声対話タスクにおいてユーザ発話を精度良く認識することは困難である．

これらの問題点を解決するために，我々はアプリケーションから独立した音声認識(ASR)エンジンと対話マネジャの間に SLP(spoken language processing)ユニットを設け，対話音声中のキーワードを抽出する方式を提案している[3]．この方式は，ASR の認識結果(1-best)と対話管理部からのキーワードの双方を DPF ベクトル系列に変換してワードスポッティングを行い，高精度にキーワードを抽出することを目指している．SLP ユニットは，連続的な特徴ベクトル系列ではなく音素セグメンテーション後の DPF ベクトル系列を対象としており，また DP マッチングに基づく距離計算もテーブル参照で行うため，処理量が少ない利点を持つ．また，ASR エンジンをアプリケーション独立にできることで，本提案方式は将来，分散音声認識（DSR）に適用することも可能である．

2. 音声対話システムの概要

2.1. 提案システム

図 1 に音声対話システムの全体構成を示す．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 (FEP)，音声言語処理部（SLP），および対話管理部 (DM) の三つの処理部で構成される．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．

ここで用いた DPF は，日本語の弁別特徴として提案されているもののうち[4]，“母音性／非母音性”と“子音性／非子音性”の二つの弁別特徴を除き，代わりに国際音声記号表を参考にして，“母音性／子音性”，“半母音性（/j,w,r/）／非半母音性”，および“摩擦性（/s,z,h/）／非摩擦性”を追加して 12 次元としたものである．予備実験の比較結果では，置き換え後の特徴が良好な結果を得た．また，今回使用した DPF は，母音グループと子音グループの分離が十分である反面，母音間は極く接近している．そこで，母音間距離を 2 倍に評価し補正することを

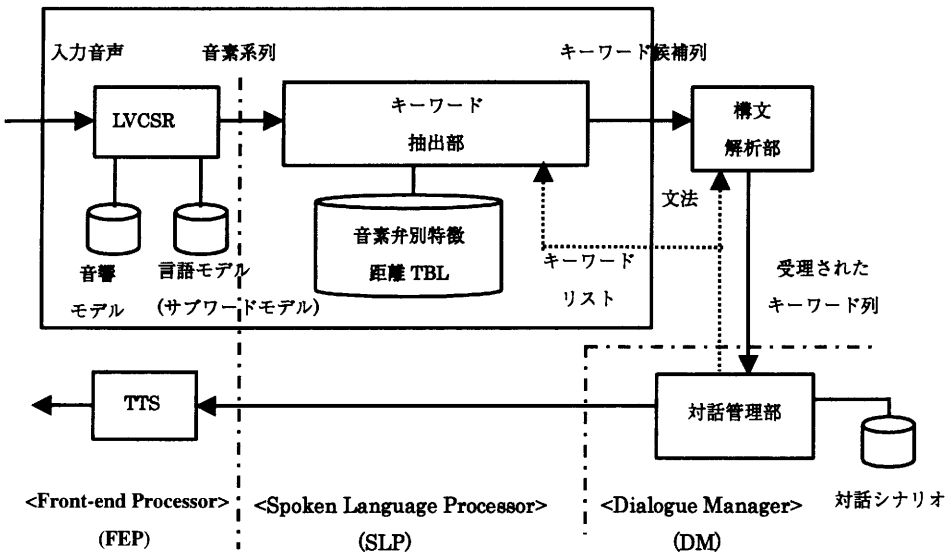

図 1. 音声対話システムの全体構成図

試みる[5].

提案方式の処理の流れは以下のようである．入力音声は，まず LVCSR エンジンで認識処理された後，出力音素系列が SLP に送られる．LVCSR エンジンには，日本語ディクテーションシステム Julius [6]を使用した．音響モデルは特徴パラメータとして“MFCC+Δt+ΔP（25 次元）”を，また HMM は 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）を使用している．今回は，言語モデルに 2.2 節で説明するサブワードモデルを利用することを検討する．

FEP と DM との間に設けた SLP は，FEP をアプリケーションから独立にすると共に，DM 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割を持つ．具体的には，LVCSR エンジンが出力した音素系列から，DM が指示するキーワードを抽出する．キーワード抽出は，以下に説明するワードスポッティングを用いる．まず，次の (1) 式から音素系列とキーワード間のハミング距離を計算する．

$$d_k(i,j)=\frac{1}{12}\sum_{m=1}^{12}\{x(m,i)\oplus r_k(m,j)\} \quad \cdots\cdots (1)$$

ここで，$d_k(i, j)$はキーワード k に対するハミング距離で，i は音素を単位とした入力フレームの番号，j はキーワード中の音素系列の番号，また m は DPF の次元数である．入力 DPF ベクトル系列 x(m,i), m=1,2,...,12, i=1,2,...,I と，キーワード $r_k(m, j)$, m=1,2,...,12, j=1,2,...,J とのハミング距離は，テーブルの形であらかじめ用意する．

次に，入力フレームに沿って全てのキーワードとの間で，端点フリーDP マッチングを行う．DP マッチングには以下に示す (2)~(4) の漸化式を利用した．

$$g_k(i,j)=min\begin{cases} g_k(i-1,j)+d_k(i,j) & (a)\\ g_k(i-1,j-1)+d_k(i,j) & (b)\\ g_k(i,j-1)+d_k(i,j) & (c)\end{cases} \quad \cdots\cdots (2)$$

$$c_k(i,j)=\begin{cases} c_k(i-1,j)+1 & if\,(a)\\ c_k(i-1,j-1)+2 & if\,(b)\\ c_k(i,j-1)+1 & if\,(c)\end{cases} \quad \cdots\cdots (3)$$

$$D_k(i)=g_k(i,J)/c_k(i,J)+\alpha/n_k \quad \cdots\cdots (4)$$

ここで，$g_k(i,j)$は累積距離，$c_k(i,j)$ は DP パスの重み，$D_k(i)$ はキーワード毎のスコアを示す．また，音素列長が長いキーワードは，短いキーワードに比べて累積距離が大きくなる傾向があるため，キーワードを構成する音素列長 n_k に反比例してペナルティが加算される項（α/n_k）を組み込んでいる．今回は α を 0.1 とした．DP マッチングの結果，$D_k(i)$ が一定の閾値以下のキーワードを抽出する．同時に，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は，最も距離が小さい候補のみを残した．

2.2. サブワードモデルの導入

現在の LVCSR(Large Vocabulary Continuous Speech Recognition) ソフトウェアは，発話内容に未知語を含む場合，入力音声と大きく異なる音素系列を出力する．これは LVCSR ソフトウェアの認識性能が，主に言語モデルの強い制約に依拠していることと関係している．図 2 に，異なる言語制約を適用した際の，LVCSR ソフトウェアの出力例を示した．3-gram は文法的には正しい文を出力するが，その音素列は入力音声と大きく異なることが分かる．他方で，言語制約を持たない 0-gram は意味不明な文を出力するが，音素列は正解に近いものを出力している．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 エンジンでは，言語モデルの強い制約を緩和することで，未知語（ここでは「ケンタッキーフライドチキン」）を含む発話文に対しても，意味的に理解可能な音素列を出力することができる．

以上の事実を自由発話音声の認識結果から確認したものを図 3 に示す[7]．この例は，3 節に述べる対話音声評価データに対して，上述した提案方式を用いてキーワード検出したもので

3 - gram: 現在 、 大豆 時期 に 期待 する
（げんざいだいずじきにきたいする）
2 - gram: 現在 、が いる 磁気 に 前 する
（げんざいがいるじきにぜんする）
1 - gram: 現在 期待 と 自身 、 再選
（げんざいきたいとじしんさいせん）
0 - gram: 軒 多 木 歩 ラ 伊豆 地 近 遺棄 帯 頭 ん
（けんたきふらいずちきんいきたいずん）
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.言語制約を変化させた場合の出力例

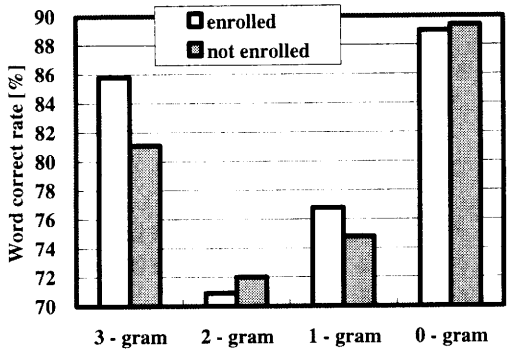

図3. 単語言語モデルの言語制約による性能比較

ある．言語制約を無くした 0-gram が最も高い性能を示しており，言語制約を弱めたことの効果は，未知語に対して顕著であることがわかる．また，未知語を辞書に登録する前後の性能を比較すると，0-gram について登録前後でほぼ同等の性能である．すなわち，0-gram を採用することで LVCSR エンジンに未知語を登録する必要が無くなり，認識エンジンはアプリケーション独立になるという利点が得られる．

一方，0-gram を採用した LVCSR エンジンの出力を観察すると，音節長の短い単語の組み合わせで構成されており，その多くは 1~2 音節程度である．これは登録する単語もしくはサブワードは，音節長の短いもののみで十分なことを示唆している．また，サブワードのみで辞書を作成することにより LVCSR の計算負荷を減らすことができる．そこで，以下では音節を単位とするサブワードモデルの構築と評価を行う．

2.2.1. m 音節以下の語彙から成る辞書

まず，音節長の短い単語のみで辞書を構成することを検討する．IPA 辞書 (語彙数 20k) から，短音節単位(全ての CV (C は子音，V は母音)．ただし撥音/N/と促音/Q/もここでは 1 音節に含めた)で m 音節 (m=1,2,...,6) 以下の単語を抽出して辞書を構成した．辞書中の単語数と異なり語数を表 1 に示す．異なり語数は辞書中から同音異義語を削除した際の登録単語数である．これ以降，m 音節以下の単語で構成した辞書を m 音節辞書と呼ぶ．

2.2.2. m 音節単位サブワード言語モデル

次に，音節長の短い単語から成る言語モデル（サブワード言語モデル）を作成する．ここでは，IPA 辞書 (語彙数 20k) にエントリされている単語を，短音節単位で m 音節(m=1,2)単位に分割し，その結果得られるサブワード系列から N-gram (N=1,2,3) 言語モデルを作成した．サブワードの数は 1 音節モデルで 121，2 音節モデルでは 3477 であった．

表 1. 20k 辞書中単語の(短)音節数と累積語数

(短)音節の数	語 彙	異なり語数
1	482	82
2	4,016	1,436
3	10,683	6,574
4	18,512	13,309
5	19,967	14,701
6	20,730	15,452

2.2.3. 日本語音節単位サブワードモデル

最後に，日本語の全ての音節をサブワード単位の対象とする言語モデルを作成する．日本語の音節は大きく短音節と長音節に分けられ，長音節は長母音音節(CVV)，撥音付き音節(CVN)，促音付き音節(CVQ)，二重母音音節(CV_1V_2)に分類される．これらの音節をサブワードの単位としたサブワードモデルと，このうち二重母音音節を除いたサブワードモデルの二種類を新聞記事コーパスから作成した．サブワードの数は二重母音を含んだモデルで 1018，除いたモデルで 499 であった．

3. 評価実験

3.1. 言語・音声試料

以下の二つのデータセットを評価に用いた．

D1. 言語モデル学習データセット
毎日新聞記事テキストコーパス 1991 年度分（ヨミ系列に変換し，日本語音節単位に区切った結果得られるサブワード系列から N-gram(N=1,2,3)言語モデルを作成）

D2. 評価データセット
電総研（ETL）の道案内対話コーパス（以下 ETL コーパスと呼ぶ）[8]のうち話者 14 名からなる 100 発話(全発話時間 305[sec])
男性話者 23 名からなる 100 文（全発話時間 581.9[sec]）

3.2. 実験概要

DM から送られるキーワードは，ETL コーパスから 109 単語（異なり語）を選んだ．このう

ち LVCSR に登録されている単語は 66 語，未知語は 43 語である．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．

評価は，キーワード抽出部の出力で行った．評価基準としては FA / WH（1 キーワード当たりの単位時間湧き出し数）を 45 前後の値に揃えると共に，単語正解率（ $(N-S-D)\times 100/N$: N, S, D は各々全キーワード数，置換誤り数，および脱落数）によりキーワード検出性能を調べた．

3.3. 実験結果

以下では 2.2 節に述べた三つの言語モデルを比較評価した結果を述べる．

[A] m 音節辞書の評価

2.2.1 節で述べた m 音節辞書を使用した場合のキーワード抽出性能を図 4 に示す．同音異義語を削除した場合 (merged) と，その後に同音異義語の頻度情報 (1-gram) を利用した場合についても図中に併せて示す．

抽出する単語の音節数が増えるに従い，性能が改善することが分かる．語彙数 20k の辞書を利用した場合，単語正解率は 89.4%であったが，図から 3 音節辞書で同等の 89%の単語正解率を得られることが分かる．3 音節辞書の単語数は 10,683（同音異義語含む）であり，20k に比べて約半分で十分な性能が得られた．

また，同音異義語を削除した場合，性能が低下することが分かる．これは同音異義語の存在が，ビーム探索の枝狩りに抗して生き残ることに繋がったためである．異なり語のみから構成される辞書では，同音異義語の数を頻度情報 (1-gram) として組み込むことにより，性能を顕著に改善することができる．これから，少数のサブワードに加えて，頻度情報や共起を利用することが，枝狩りから生き残る際に有効であることが分かる．

[B] m 音節単位サブワードモデルの評価

次に，2.2.2 節で述べた短音節単位のサブワード言語モデルを使用し，サブワード単位と言語制約に対する比較性能実験を行った結果を図 5 示す．

全ての N-gram において 2 音節単位とした場合に高い性能を示した．言語制約に関しては，1 音節，2 音節共に制約を強めるほど性能が向上している．また，1 音節と 2 音節の性能差は，言語制約が大きいほど開く傾向が見られた．なお，両者の誤り傾向を検討すると，2 音節単位の場合，長母音（音響モデルに含んでいる）を最初から含む構造になっていることが性能向上に貢献していることが観察された．

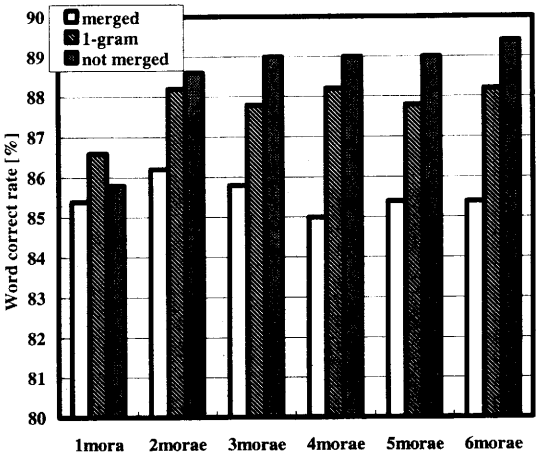

図 4. m 音節辞書の性能比較

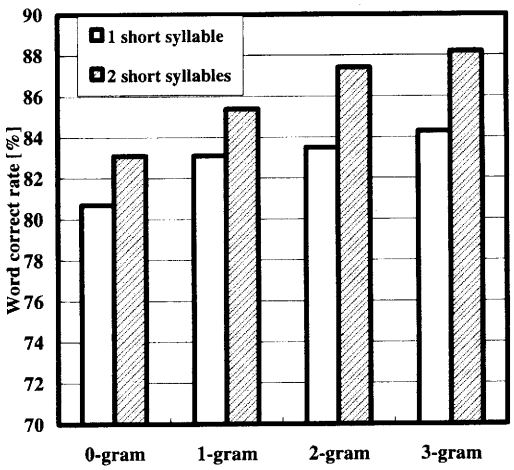

図 5. m 音節単位サブワードモデルの性能比較

[C] 日本語音節単位サブワードモデルの評価

上述した，長母音付き音節を言語モデルに含むことの効果を考慮して，2.2.3 に示した日本語音節単位の言語モデルを作成した．音節単位サブワードモデルを利用し，キーワード抽出性能を比較した結果を図 6 に示す．図中“1 short syllable”は語彙を短音節のみに限定し，短音節

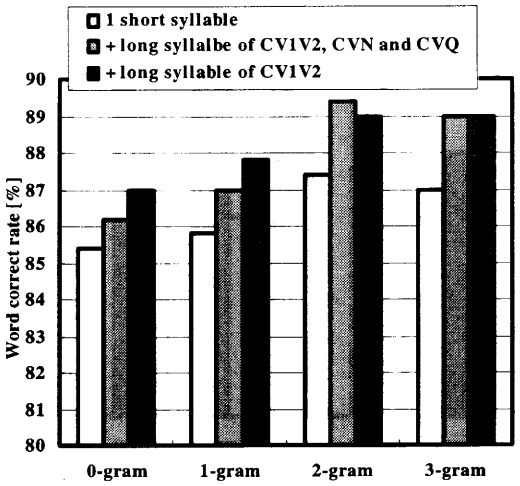

図 6. 日本語音節サブワードモデルの性能比較

単位に区切った新聞記事テキストコーパスから言語モデルを作成した場合の結果である.

サブワード単位に関わらず,2-gram までは言語制約を強めることで性能が向上した. また, 全ての言語制約において, 長音節を含むサブワードモデルが短音節単位サブワードモデルよりも高い性能を得た. 日本語発話の最小単位として, 長母音音節, 撥音付き音節, および促音付き音節を組み込んだことが性能向上に繋がったと考えられる.また,二重母音音節の有無では, 2-gram, 3-gram の強い言語制約下では性能差は見られないが, 0-gram, 1-gram の比較的弱い言語制約下では性能が向上した. これは, 二重母音音節を含まない場合には 2-gram, 3-gram モデルが短音節の組み合わせで二重母音音節の情報を保持し, 表現できるためと考えられる.

日本語音節単位サブワード言語モデルは, 2-gram で 89.4%と非常に高い性能を得ており, 語彙数 20k の 0-gram を利用した場合と同等の性能を得た[5]. 他方で, 語彙数は 20k 辞書の 40 分の 1 の約 500 と大幅にメモリおよび計算量を減らすことができた.

4. まとめ

汎用の LVCSR エンジンにサブワード言語モデルを組み込み, さらに DPF ベクトルを利用することで, 対話音声からキーワードを高精度に抽出する方法を提案した. 音節サブワード言語モデルを利用することで, LVCSR のメモリと計算負荷を大幅に減らしながら, 高いキーワード抽出性能が得られることが分かった. 同時に, 評価実験を通じて以下の点が明らかになった.

(a) LVCSR エンジンの登録辞書は, 20k 辞書中の 3 モーラまでの単語 (同音異義語含む) で, 20k 辞書 (0-gram) と同等の性能が得られる.

(b) 同音異義語の存在は, 頻度情報として認識時に有効に働き, 3 音節単語の頻度情報を 3 音節のサブワード言語モデル (1-gram) として採用することにより, 言語モデルの性能が大きく向上する.

(c) 短音節モデルに加え, 日本語音節に含まれる長母音音節, 撥音付き音節, 促音付き音節を単位としてサブワードモデルを構築することで, 語彙を減らしながら高い性能が得られる.

(d) 語彙に二重母音音節を含まないサブワードモデルにおいては, 言語制約を強めることで, 短音節を組み合わせて二重母音を表現できると考えられる.

今後は, 二重母音音節を選択的に使用した日本語音節サブワードモデルや, 新聞記事中に高い頻度で出現するサブワードに注目したサブワードモデルについて検討したい. 加えて, マルチモーダル対話システム[9]への導入を通して, 実際の利用環境における性能を調査したい.

参 考 文 献

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

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

[3] 伊勢路ほか, 音学講論, 1-6-27, pp.53-54 (2003-9).

[4] 比企静雄 著, "音声情報処理", 東京大学出版会 (1973).

[5] 伊勢路ほか, 音学講論, 2-4-12, pp.81-82 (2003-3).

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

[7] 伊勢路ほか, "0-gram 汎用 LVCSR と音素弁別特徴ベクトルを利用した対話音声認識の検討", 信学技報, SP2002-156, pp.49-54 (2002-12).

[8] 伊藤ほか, 音学講論 1-1-19, pp.37-38 (1998).

[9] K. Katsurada, Y. Ootani, Y. Nakamura, S. Kobayashi, H.Yamada, and T.Nitta, "A Modality-Independent MMI System Architecture, " ICSLP, pp.2549-2552, 2002.9.