音声言語情報処理 44－36
(2002. 12. 20)

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

伊勢路 真吾　福田 隆　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, katsurada}@vox.tutkie.tut.ac.jp, nitta@tutkie.tut.ac.jp

あらまし 本報告では汎用 LVCSR ソフトウェアを利用して，対話音声を高精度で認識する方法を提案する．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，対話管理部が指示する対話記述（語彙と文法）を利用してキーワードをスポッティングする．本方式の特徴は以下の二点にある．（1）LVCSR の言語制約を緩めることにより(0-gram, 挿入ペナルティ有)，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，キーワードスポッティングを行うことにより，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音声データを用いて，言語モデルにおける言語制約の違い，サブワードモデルとの比較，および混同行列を用いた整合方式との比較を行い，提案方式の有効性を示す．

キーワード 音声対話，LVCSR，キーワードスポッティング，言語モデル，サブワードモデル，音素弁別特徴，混同行列

Recognition of Spontaneous Speech by Using a General-Purpose LVCSR with 0-gram and Distinctive Phonetic Features

Shingo ISEJI, Takashi FUKUDA, Kouichi 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, katsurada}@vox.tutkie.tut.ac.jp, nitta@tutkie.tut.ac.jp

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

Keyword Spoken Dialogue, LVCSR, Keyword Spotting, Language Model, Sub-word Model, Distinctive Phonetic Feature, Confusion Matrix

1. はじめに

webへのアクセスは従来PCに限定されていたが，近年，携帯電話を初めとして，PDA，カーナビゲーション，及びデジタルTV等へと利用端末が拡大しつつある．また利用端末の拡大に伴い，webアクセス方法もキーボードとマウスから，音声入力，ペン入力およびそれらを組み合せたMulti-Modal Interaction（MMI）に対応することが要請されている．同時に，本格的なMMIに利用可能な高性能対話音声認識ソフトウェアの開発が期待されている．

音声認識の応用分野では，ディクテーション向けに高精度音声認識ソフトウェアが開発されているが，webアクセスのように頻繁にトピックが変化する用途に，こうした既存のソフトウエアをそのまま適用することは難しい．

対話音声認識では，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の出現など，多くの課題に直面する．これらの課題を解決する上で，ワードスポッティングは最も有効な手法である[1], [2]．しかし，この手法は正解単語区間外で多量の単語を沸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少量の語彙に絞れる場合に限られる．さらに，入力音声に沿って，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になる．

これらの問題を解決するため，我々は汎用LVCSR（Large Vocabulary Continuous Speech Recognition）ソフトウェアが本来持つ，高い音素識別能力を引き出して利用すると共に，対話制御部が指示するキーワードについて，これらを対話音声中から高精度に抽出する方法を提案している[3]．今回は，これまで提案した0-gram言語モデルとサブワードモデルを含む他の言語モデルとの比較，および音素弁別特徴ベクトルと混同行列との比較を中心に報告する．

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て，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．

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1. 言語制約を変化させた場合の出力例

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ソフトウェアは，時折，発話内容と音韻的に近い単語への識別誤りをおかす．これには，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まで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が考えられる．一方，発話に未知語が含まれる場合には，おかしな（音韻的に近いとはいえない）単語列が出力されることが多い．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1は，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ている．3-gramは文法的には正しい文を出力する一方，その音素列は入力音声と大きく異なる．他方，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ソフトウェアは，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素列を出力する能力を持つことが分かる．本報告では4種類のn-gram（n=0, 1, 2, 3）を比較すると共に，日本語のサブワードモデルについても検討を行う．近年，言語制約の強い汎用LVCSRソフトウェアの音素系列出力を利用して，音声データを検索する方式が幾つか提案されている[4], [5]．これらの方式では，音素系列中の置換・脱落・付加誤りを，混同行列（CM: confusion matrix）とDPマッチングを適用することで軽減している．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，設計時と利用時では音響諸条件が異なることが問題となる．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法を検討すると共に，両者の比較実験を行う．

図2に今回使用した音素の弁別的な特徴を示す．ここで用いた特徴は，日本語の弁別特徴として提案されているもののうち[6]，“母音性／非母音性”と“子音性／非子音性”という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．予備実験での比較結果では，置き換えた特徴が良い結果を得ている．

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[7],[8]．音素を単位とする認識方式の最大の利点はセグメンテーションである．提案方式は，LVCSRが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．

	a	i	u	e	o	N	w	y	r	m	n	p	t	ts	ch	k	b	d	g	z	j	s	sh	h	f
母音性/子音性	+	+	+	+	+	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
半母音性/非半母音性	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
摩擦性/非摩擦性	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	+	+	+	+	+	+
高舌性/非高舌性	-	+	+	-	-	-	+	+	-	-	-	-	-	+	+	+	-	-	+	-	+	-	+	-	+
後方性/非後方性	+	-	+	-	+	-	+	-	-	-	-	-	-	+	-	+	-	-	+	-	-	-	-	-	+
低舌性/非低舌性	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-
前方性/非前方性	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	-	+	+	-	+	+	+	+	-	+
舌端性/非舌端性	-	-	-	-	-	-	-	-	+	-	+	-	+	+	+	-	-	+	-	+	+	+	+	-	-
破裂性/破裂性	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	-	-	-	-
有声性/非有声性	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	+	+	+	+	-	-	-	-
連続性/中断性	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
鼻音性/口音性	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-

図 2. 今回使用した弁別的な特徴

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，音声言語処理（SLP）部，および対話管理部の3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．

入力音声は，まず言語モデルの制約を0-gramとしたLVCSRソフトウェアで認識処理され，出力される音素系列が音声言語処理部に送られる．LVCSRソフトウェアには日本語ディクテーションシステムJulius [9]を使用する．Juliusは2パス探索を行い，1stパスに2-gram，2ndパスに3-gram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）

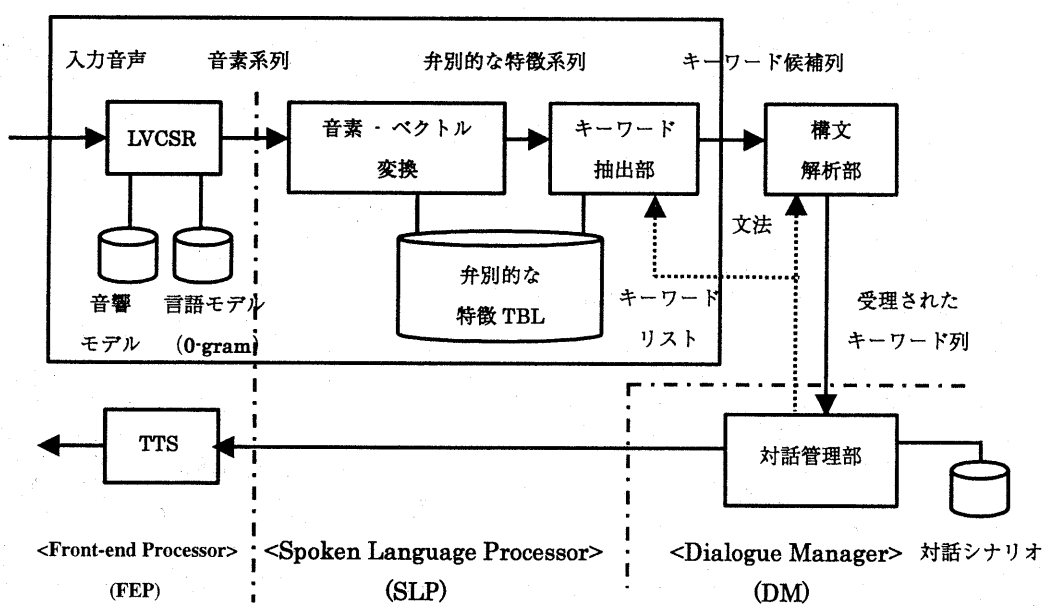

図 3. 音声対話システムの全体構成図

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）．

フロントエンド部と対話管理部との間に設けられた SLP 部はフロントエンド部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割を持つ．具体的には，最初に LVCSR ソフトウェアが出力する音素系列を 12 次元の弁別的な特徴（母音／子音性，半母音性，摩擦性，高舌性，後方性，低舌性，前方性，舌端性，遮音性，有声性，連続性，鼻音性）ベクトルに変換する．次に，I フレームの弁別的な特徴ベクトル系列 x(m,i)，m=1,2,...,12, i=1,2,...,I. を用いて，対話管理部が指示するキーワードを抽出する．キーワードは，音素系列に置き換えた後，キーワードを構成する J 個の音素を弁別的な特徴ベクトル系列 $r_k(m, j)$, k=1,2,...,29, j=1,2,...,J. に展開しておく．入力 x(i,j)からキーワード $r_k(m, j)$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．

$$d_k(i,j)=\sum_{m=1}^{12}\{x(m,i)\oplus r_k(m,j)\} \;/\; 12 \qquad (1)$$

$$g(i,j)=min\begin{cases} g(i-1,j)+d(i,j) & (a)\\ g(i-1,j-1)+d(i,j) & (b)\\ g(i,j-1)+d(i,j) & (c)\end{cases} \qquad (2)$$

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

$$D(i)=g(i,J)/c(i,J) \qquad (4)$$

$d_k(i, j)$はハミング距離，g(i,j)は累積距離，c(i,j)は DP パスの重み，D(i)はそのキーワードの持つ距離である．ここで，音素弁別特徴は元々音素間の対立を考慮して作成されたものである．すなわち音素間距離としてのバランスを考慮したものではないため，これを音素列間の累積距離として評価するとき，識別結果に誤りを生じることがある．例えば，音素弁別特徴では図 2 から知られるように母音グループ内の距離が小さいため，そのまま使用すると母音同士の距離が低く評価される．このため，母音間の距離は 2 倍にしている．

キーワードは端点フリーDP マッチングによりワードスポッティングを行った後，距離 D(i)が一定の閾値以下のものを抽出する．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を残した．

抽出されたキーワード系列は構文解析部に渡され，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．次節以降ではキーワード道案内タスクにタスクにおけるキーワード抽出実験結果を示し，考察を加える．

3. 評価実験

3.1. 音声資料

以下に示す 2 つの評価データセットを使用した．

D1. 電総研（ETL）の道案内対話音声コーパス[10]のうち話者（男女）14 名の 100 発話（全発話時間 305[sec]）

D2. 新聞記事読み上げコーパス（ASJ-JNAS）のうち男性話者 23 名からなる 100 文

3.2. 実験概要

フロントエンド部（Julius）において，様々な言語モデル重みと単語挿入ペナルティを設定し，キーワード抽出性能を評価した．設定した言語重みと単語挿入ペナルティを表 1 に示す．

LVCSR の語彙数は 20k である．対話管理部から与えるキーワードは 109 単語（異なり語）を評価コーパスから選んだ．そのうち 66 語は LVCSR ソフトウェアの辞書に登録されており，残りの 43 語は未知語となる．全てのキーワードを登録するとキーワードに関して未知語は無くなるが，発話の最初や最後にしばしば出現する不要語や，話し言葉特有の表現に含まれる未知語が存在する．

3.3. 実験結果

[A] 言語制約の比較

最初に，表 1 に示した各言語制約間の性能差を検証すると共に，全てのキーワードを Julius の辞書に登録した場合についても比較を行った．

表1. Juliusにおける言語制約

	1st Pass	2nd Pass	Weight (LM)	Insertion Penalty
3-gram:	2-gram	3-gram	8	-2
2-gram:	2-gram	2-gram	8	-2
1-gram:	1-gram	1-gram	8	-2
0-gram:	0-gram	0-gram	0	-5

表2. 言語制約を変化させた際のキーワード抽出結果

	substitution	deletion	word error rate [%]			FA/WH
				detail		
				enrolled	unknown	
3 - gram	29	19	18.9	16.3	26.6	46.9
2 - gram	51	19	28.0	27.9	28.1	45.4
1 - gram	46	18	25.2	21.6	35.9	43.8
0 - gram	18	9	10.6	12.1	6.2	46.7
3-enrolled	25	11	14.2	—	—	45.7
2-enrolled	52	21	29.1	—	—	44.5
1-enrolled	43	16	23.2	—	—	43.9
0-enrolled	19	9	11.0	—	—	46.9

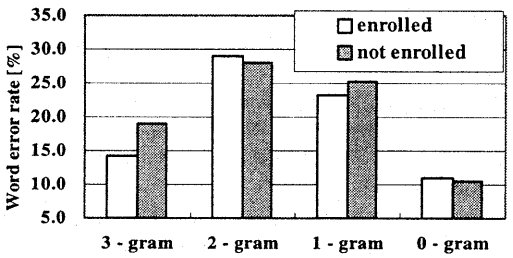

図4. 未知語登録前と後の性能比較

表2にキーワード抽出結果を示す. FA/WH (False Acceptance per Word and Hour) を同程度にし, キーワード抽出性能を調べた. 表中 "n-enrolled" は n-gram において 43 個の未知語を Julius の辞書に登録した場合の結果である. 結果から分かるように, 言語制約を無くした 0-gram が最も高い性能を示している. 特に, 内訳欄に示した通り, 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った. 3-gram, 2-gram, 1-gram において, 性能が低下した理由は, 言語モデルの強い制約により, 音響モデルで出力された音素が別の音素に置き換えられたためである. なお, Julius 単独での単語誤り率 (WER) は, 全てのキーワードを登録した場合で 44.9%であった.

図4に未知語を辞書に登録する前後の性能比較結果を示す. 0-gram について, 未知語の登録前後の結果を比べると, ほぼ同等の性能であった. この結果から, 0-gram を利用する場合は LVCSR ソフトウェアに未知語を登録する必要は無く, アプリケーションを独立にできることが分かる.

[B] 20k 言語モデルとサブワードモデルの比較

次に, サブワードモデル[11]を使用した場合と, 語彙数 20k 単語の言語モデルを使用した場合の性能を調べた.

表3. サブワードモデルを使用した際のキーワード抽出結果

	substitution	deletion	WER [%]	FA/WH
20k-LM	18	9	10.6	46.7
1 syllable	20	15	13.8	45.2
2 syllables	27	14	16.5	44.4
2 syllables (selected)	22	13	13.8	45.3

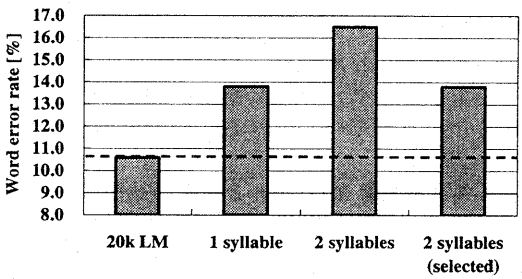

図5. 20k言語モデルとサブワードモデルの比較

日本語の1音節と2音節のサブワードモデルを用いて評価実験を行った. 表3に 0-gram 使用時において, 比較したキーワード抽出結果を, 図5に WER を示す. 表中, "2-syllables (selected)" は 20k 辞書の中に実際に存在する2音節を抜き出し, 1音節と併せてサブワードモデルとしたものである. サブワードモデルを利用した場合, 20k 言語モデルと比較して低い性能にとどまった.

図6にサブワードモデルを使用した時の音素系列出力を示す. これから分かるように, 20k 言語モデルは未知語"ロッテリア"をサブワード"ロッテ"と"リア"の組み合わせで表現

20k:
蛇あっ打っロッテリア矢土えっ茶李んんでしょ下?
(じゃあっだっろってりあやどえっちゃりんんでしょか)
1 syllable:
じゃあずどうてりあええああどいじゃいんでしょうか
2 syllables:
にだあとぶえげやえらあどいたいんでしょうか
2 syllables (selected):
じゃあどってりあやあどいっざいんでしょうか
入力音声:
じゃあー, ロッテリアへはどう行ったらいいんでしょうか

図6. サブワードモデルを使用した際の音素系列の例

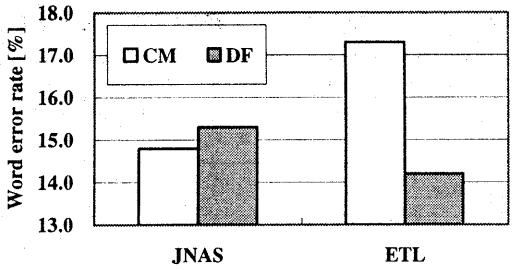

図 7. 弁別的な特徴と CM の比較

している．このように，20k 言語モデルはサブワードモデルと比較して，辞書中に正解単語が存在する場合や，単語を複合した形で表現できる場合があるため，高い性能を得ることができる．

[C] 弁別的な特徴ベクトルと CM の比較

キーワード抽出過程における DP マッチングの距離算出に，弁別的な特徴ベクトル（DF）と CM を使用した場合を比較した実験結果を図 7 に示す．ここでは評価実験[A]，[B]で行っていた母音同士のマッチングにおいて距離を 2 倍にする処理は行っていない．CM は D2 データセットから作成しているため，D2 音声コーパスに対しては，DF と同程度の性能を得ている．一方，音響環境が異なる D1 音声コーパスにおいては，DF が性能を維持したのに比べ，CM は性能が大きく劣化した．CM を利用する場合，その性能は音響環境に大きく依存すると言える．

4. まとめ

本報告では，対話音声の中からキーワードを抽出する新しい手法を提案した．本手法は以下の特徴を持つ．

(a) LVCSR の言語制約を緩めることにより(0-gram，挿入ペナルティ有り)，LVCSR の持つ高い音素識別能力を最大限に利用している

(b) 音素系列出力を弁別的な特徴ベクトル系列に置き換え，キーワードスポッティングを行うことにより，置換・脱落・付加誤りに対処している

対話音声コーパスを用いた評価実験において，汎用 LVCSR の 20k 言語モデルは日本語サブワードモデルと比較して，高いキーワード抽出性能を示した．さらに，弁別的な特徴ベクトルを利用することにより，音響環境に頑健なキーワードスポッティングを達成した．

今後は，MMI システム[12]への導入を行うと共に，実際の利用環境における性能を調査したい．

文 献

[1] J.R.Rohlicek, W.Russel, S.Roucus, and H.Gish, "Continuous HMM for Speaker Independent Word Spotthing."Proc.ICASSP, pp.627-630, May.1994.

[2] H.Matsu'ura, Y.Masai, J.Iwasaki, S.Tanaka, H.Kamio, and T.Nitta, "A Multi-modal, Keyword-based Spoken Dialogue System – MultiksDial," Proc.ICASSP, pp.II-33-36, 1994.4.

[3] 伊勢路，福田，桂田，新田，"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"，音学講論 2-9-11，pp.83-84，2002.9.

[4] 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.

[5] 西崎，中川，"未知語を考慮したニュース音声記事の検索"，情処研報，SLP-39-29，pp.171-176，2001.11.

[6] 比企静雄 著，"音声情報処理"，東京大学出版会，1973

[7] T. B. Martin, "Practical Application of Voice Input to Machine," Proc. IEEE, 64-4, 1976.

[8] 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, pp.210-214, 2001

[9] A.Lee, T.Kawahara, and K.Shikano, "Julius – an Open Source Real-Time Large Vocabulary Recognition Engine," EuroSpeech2001, pp.1691-1694, 2001

[10] K.Ito, T.Akiba, S.Hayamizu, and K.Tanaka, "A Spontaneous Speech Dialogue Corpus Collected Using WOZ System," Proc. Acoustic Society of Japan – Autumn Meeting, 1-1-19, pp.37-38, 1998.9.

[11] K.Ng, "Toward Robust Methods for Spoken Document Retrieval," Proc. ICSLP, pp.939-942, 1998.11.

[12] 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.