外国語・擬音語と弁別素性を用いる音声カナ変換システムの評価*

宮城順一　高良富夫 (琉球大 工)

1　まえがき

現在、多くの音声認識システムが開発され、実用に供されている。これらのシステムでは、実用的な性能を達成するため認識用の単語辞書を用いている。これは、語彙を制限し、文法情報・文脈情報を利用するためである。

しかし、人間は外国語をカナ表記で書き表したり、人間の声以外の音を擬音語として表す能力を持っている。認識用に単語辞書を用いている場合、無意味単語である外国語や人間の声以外の音を認識することは不可能である。

そこで我々は、音声カナ変換システムを作成した。このシステムは、入力された音声を日本語のカナ表記へと変換することができる。

本研究では、このシステムを評価するために無意味単語として犬と猫の鳴き声とベトナム語を用いる。犬と猫の鳴き声の音声カナ変換の評価では、日本語と英語による鳴き声の擬音語との比較を行う。ベトナム語の音声カナ変換実験の評価では、人間の聴取実験結果との比較を行う。各実験結果の評価において単語間の比較を行う際に、類似性を、より定量的に評価するため、音素間の類似度を弁別素性によって表現する。

2　音声カナ変換システム

2.1　隠れマルコフモデル (HMM) による音声認識

音声カナ変換システムは、隠れマルコフモデル (以下 HMM とする) による音声認識システムが元になっている。本研究で使用する基本 Left-to-Right 構造の HMM の例を図 1 に示す。S_1, S_2, S_3 は HMM の各状態を表す。

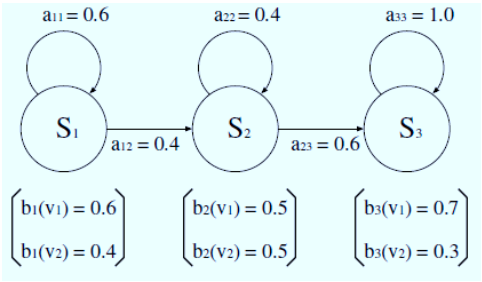

図 1　基本 Left-to-Right 構造の HMM

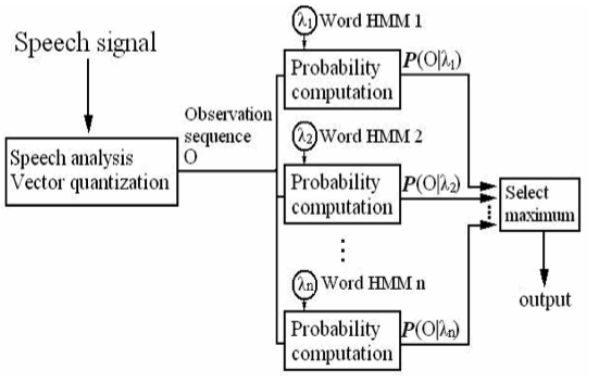

図 2　HMM による認識プロセス

a_{ij} は状態 S_i から S_j への状態遷移確率である。v_1, v_2 はシンボルであり、$b_j(v_t)$ は状態 S_j におけるシンボル v_t の出力確率である。

通常の単語音声認識は次のようなプロセスで行われる。まず、それぞれの単語の HMM モデル λ と観測系列 $O = (o_1, o_2, ..., o_T)$ からビタビアルゴリズムによって尤度 $P(O|\lambda)$ を計算する。そして、最も高い尤度を持つ HMM モデルが表す単語が選択され出力される。このプロセスを図 2 に示す。

2.2　音節 HMM による音声カナ変換

入力された無意味単語をカナ表記で表すためには、認識用の単語辞書を使わずに、音節 HMM を用いて考えうるすべての単語の単語

---

*Evaluation of speech Kana conversion system by using foreign words, onomatopoeic words and distinctive features, by Tomio TAKARA, Junichi MIYAGI(University of the Ryukyus)

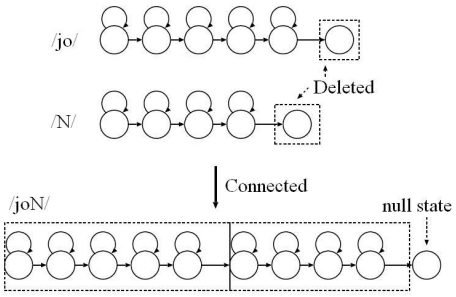

図 3　音節 HMM の連結学習

表 1　音節 HMM を利用した音声認識実験の結果

実験方法	認識率	正答数/単語数
Closed test	99.1%	6511/6570
Open test	91.8%	3016/3285

HMM を生成し尤度を求め比較する。本論文では計算時間の制約から 3 音節以下の単語に限って実験を行う。

音節 HMM の学習は、東北大-松下単語音声データベースから日本人男性 3 人分の音韻バランスを考慮した 3285 単語を用いて連結学習により行った。連結学習では各単語毎に、音節 HMM から単語 HMM を構成し、Baum-Welch アルゴリズムを用いて HMM パラメータの推定を行っている。音節 HMM の連結学習は図 3 のようにして行う。

2.3　単語認識実験

本研究で用いる音声分析のパラメータは、13 次元の MFCC と ΔMFCC の計 26 次元、フレーム長 33ms、フレーム周期約 16.5ms である。音節 HMM は離散 HMM の基本 Left-to-Right 構造である。

この音節 HMM を使った通常の単語音声認識の性能を評価するため、学習に使用したものと同じ 3285 単語の認識用辞書を用いて単語認識実験を行った。この結果を、表 1 に示す。2 人分の音声を用いて学習し、同じ 2 人分の音声を認識したクローズドテストでは、99.1%、学習に使用した 2 人分の音声とは別の発声者が発声した音声を認識したオープンテストでは 91.8%の認識率となった。

表 2　弁別素性の表の一部

	a	i	m	t
母音性 (Vocalic)	+	+	-	-
子音性 (Consonantal)	-	-	+	+
高舌性 (High)	-	+	O	O
後方性 (back)	+	-	O	O
低舌性 (low)	+	-	O	O
前方性 (Anterior)	O	O	+	+
舌頂性 (Coronal)	O	O	-	+
円唇性 (Round)	-	-	O	O
緊張性 (Tense)	+	-	O	O
有声性 (Voice)	O	O	+	-
持続音性 (continuant)	O	O	-	-
鼻音性 (Nasal)	O	O	+	-
粗擦音性 (Strident)	O	O	-	-

3　弁別素性を用いる距離の計算

弁別素性 (distinctive feature) は、言語学的に音素の体系や調音結合などの音韻構造を表すものである。弁別素性を用いることにより、ある音声を他の音声から明示的に区別することができるとされている [1]。本研究では、Chomsky & Hall の弁別素性と日本語の弁別素性表 [2] を採用した。

各素性が示す特徴がある場合を +、当てはまらないものを −、該当しない素性は O としている。表 2 に弁別素性の表の一部を示す。

音素を弁別素性を成分とするベクトルとみなし、式 (1) のようにハミング距離を計算することにより、音素間の距離 D を表すことができる。

$$D = \sum_{i=1}^{l} |A_i - B_i| \qquad (1)$$

ここで、l は弁別素性の数,A_i, B_i はそれぞれ音素 A, B の i 番目の素性である。各素性と値の対応を表 3 に示す。”+”を 1、 ”-”を 0、”O”を ”don’t care” とする。”don’t care” では、比較相手の素性が ”+”, ”-” のいずれでもハミング距離は 0 とする。弁別素性では、対立する素性によって音素を表現しているので、対立する素性の無い”O”は無視できるものとする。

表 3　各素性と値の対応

素性	”+”	”-”	”O”
値	1	0	don’t care

表 4　日本語、英語の擬音語間の距離

(犬) wan と bau の距離	1.0
(猫) nyaa と myuu の距離	2.0

計算された音素間の距離をもとにDPマッチングによって、単語間の距離を求める。最後に音素数で距離を割り、1音素あたりの素性の違いを表すようにする。この距離を弁別素性距離と呼ぶことにする。

4　犬と猫の鳴き声のカナ変換実験

4.1　実験に用いる音声データ

この実験では、犬の鳴き声6回分、猫の鳴き声5回分の音声を使用する。音楽CDから44.1kHzで読み取ったデータを11.025kHzサンプリングに変換して使用している。

4.2　評価に用いる擬音語

この実験では、評価のために日本語と英語による犬と猫の鳴き声の擬音語を用い、音声カナ変換システムの認識結果のずれを両言語間の差異と比較する。日本語の犬の鳴き声として”wan”, 英語の鳴き声として ”bau”, 日本語の猫の鳴き声として”nyaa”, 英語の鳴き声として ”myuu” を用いた。

日本語と英語での擬音語間の弁別素性距離を計算した結果、それぞれ表 4 のようになった。

4.3　実験結果

音声カナ変換システムによって犬と猫の鳴き声を認識しカナ表記を求め擬音語との距離を求め比較した。この結果は表 5 のようになった。これらの関係を図 4 と図 5 に示す。カナ変換の結果は、両言語間の違いと同程度であり、犬猫どちらの音も英語に近い結果になった。

表 5　擬音語と音声カナ変換結果の距離

擬音語 (犬)	wan (日)	bau (英)
変換結果との距離	2.0	1.5
擬音語 (猫)	nyaa (日)	myuu (英)
変換結果との距離	2.06	1.69

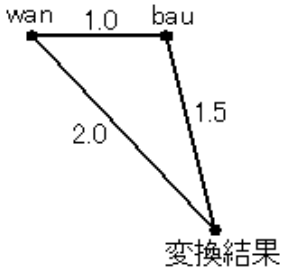

図 4　犬の擬音語と変換結果

5　ベトナム語のカナ変換実験

5.1　実験に用いる音声データ

実験に用いる音声データには、聴取実験の被験者の日本人男性 4 人が聞いたことのないベトナム語の基礎語彙 200 単語 [3] を使用した。ベトナム語を母国語とする男性1名が防音室内で発声したものを、 16bit モノラル 48kHz サンプリング で録音し、それを 11.025 kHz サンプリング に変換し利用する。

5.2　人間による聴取実験

カナ表記の書き取りの結果を得るため、4人の成人日本人による聴取実験を行った。4人の被験者には、防音室内でベトナム語の各単語毎に2回音声を書き取らせた。書き取り時間は5秒である。使用できる表記はカタカナの五十音表により指定した。

図 5　猫の擬音語と変換結果

表 6　カナ変換結果と平均距離との差

実験方法	聴取実験	システム
平均距離との差	0.14	0.80

5.3　評価方法

聴取実験によるカナ表記の書き取り結果とカナ変換システムのカナ変換結果を比べることにより音声カナ変換システムを評価する。まず、人間の聴取実験から任意の３人の結果を選び、それぞれのカナ表記間の弁別素性距離を平均する。これを平均距離と呼ぶことにする。次に、他の１人のカナ表記の結果と、この3人のカナ表記との間の弁別素性距離を計算し平均する。この平均値と上記の平均距離との差を求める。これをすべての単語について行う。４人分の聴取実験を行ったため、４通りの組み合わせで差を計算し、それらを平均したものを結果の比較に用いる。音声カナ変換システムの結果についても同様に、３人のカナ表記との間の弁別素性距離を計算し平均した値と平均距離との差を求める。人間の聴取実験結果と音声カナ変換システムの結果、それぞれから計算した平均距離との差を比較する。

5.4　実験結果

実験結果は表 6 のようになった。表 6 は平均距離との差を表している。この結果から人間の聴取実験の場合は、人間の平均的な聴取実験結果より７音素あたり１つ程度の素性が異なるカナ表記を表すことがわかる。音声カナ変換システムでは、人間の平均的な聴取実験結果より１音素あたり１つ程度の素性が異なるカナ表記に変換することがわかる。なお、この実験での平均距離は0.53であった。これらの距離の関係を図 6 に示す。ここで、聴取結果 a1,a2,a3 の点は平均距離を算出するために用いた聴取実験結果の例示である。聴取結果ｂの点は、他の１名の結果である。

6　むすび

無意味単語の音声をカナ表記で書き表すために音声カナ変換システムを作成した。このシステムの新しい評価法として、弁別素性を

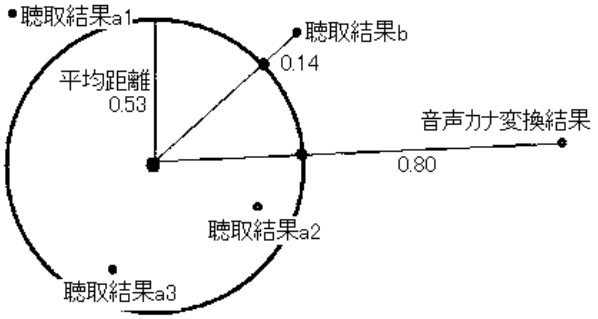

図 6　各距離の関係

用いてDP マッチングにより単語間距離を計算する方法を提案した。擬音語と音声カナ変換システムの結果を比較した結果、音声カナ変換システムの結果は日英間の擬音語間の距離と同程度であることが分かった。ベトナム語による聴取実験結果と音声カナ変換システムの結果を比較した実験では、システムの性能が聴取実験の結果よりもかなり低く、認識精度を大きく向上させる余地があることが分かった。

一般に、人間の言語音声の認識においては、辞書など言語情報を大いに利用していると思われる。従って、言語情報の利用できない無意味単語や動物の鳴き声などの認識においては、音声認識システムの方が性能が優れているようにできると予想される。もし既存の方法でできないのであれば、人間の音声認識の音声分析に未知のメカニズムがあることになる。

今後の課題として、音声カナ変換システムをこの評価法で評価し、人間の聴取能力以上の性能を達成するよう、まずは、フレーム周期・フレーム長・特徴パラメータなど、システムのパラメータを調整することが挙げられる。

参考文献

[1] 柴谷方良, 影山太郎, 田守育啓 :”言語の構造 音声・音韻篇”, くろしお出版, pp.74-75, (1987-04).

[2] 板橋秀一, 赤羽誠, 石川泰, 大河内正明, 粕谷英樹, 桑原尚夫, 田中和世, 新田恒雄, 矢頭隆, 渡辺隆夫 : ”音声工学”, 森北出版, pp.8-9, (2005-02)

[3] 安本美典, 本多正久 : ”日本語の誕生”, 大修館書店, pp.309-297, (1978-11)