

論文 / 著書情報
Article / Book Information

論題(和文)	スペクトル時系列の大局的・局所的特徴の組合せによる単語音声認識
Title(English)	
著者(和文)	古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会1980年秋季講演論文集, , pp. 393-394
Citation(English)	, , , pp. 393-394
発行日 / Pub. date	1980, 10

1—1—18 スペクトル時系列の大局的・局所的特徴の 組合せによる単語音声認識*

古井 貞熙 (日本電信電話公社 武蔵野電気通信研究所)

I まえがき 単語音声認識システムを構成する際、対象とする語彙数(数十か数百か数千か)や語彙の種類がある程度明確な場合には、その対象にあつた方式を採用するのが有効かつ経済的であることは言うまでもない。我々はこのような観点から、いくつかの検討を並行して進めているが、ここでは比較的少数の、しかももう数か少ない語彙を対象とする簡易な認識方法について検討した結果を報告する。

II 認識方法 図1に示すように、音声を40ms毎に10次までの(LPC-)ケプストラムとパワーからなる特徴ベクトルに変換し、その時間パターンと、母音の変化のような比較的ゆっくりした大局的な特徴と、子音情報のような比較的局所的な情報にわけとり出して、両者を組合せて認識を行うことを考える。前者の特徴を表現するために、特徴ベクトルの各要素の音声区画全体のパターンをフーリエ級数展開し、その0次から6次までの展開係数計7成分のうち、単語の分離に有効な4成分(分散分析によりあらかじめ決めておく)を選択して用いる。未知入力とのマッチングには、各成分の変動性を考慮した重みつき距離を用いる。後者の特徴としては、ここでは簡単のために語頭のCV情報をとり出すこととし、語頭のフレームの特徴ベクトルの時系列をそのまま用いる。未知入力とのマッチングの尺度としては、標準パターンと未知入力の語頭部分を時間軸方向に互いに±3~5フレーム程度シフトして最もよく対応がつけられたときの、各要素の重みつき距離の総和を用いる。ここでは経験的に、 $l = \text{Min}\{L/3, 15\}$ (ただしLは全フレーム数、図2参照)とした。大局的・局所的特徴による距離の組合せ方法としては、経験的に定めた重みによる重みつき加算を行った。

III 認識実験 実験には、表1に示す2種類の音声サンプルセットを用いた。いずれの場合も各話者は各単語を11回ずつランダムな順序で発声している。計算機端末室の騒音レベルは50dB(A)、平均S/Nは20~25dBである。学習は、各話者の各単語毎に最初の1~3サンプルを用いて行い、2サンプル以上を用いる場合は、フーリエ展開係数は平均化し、語頭の時間パターンはそのまま複数個を標準パターンと

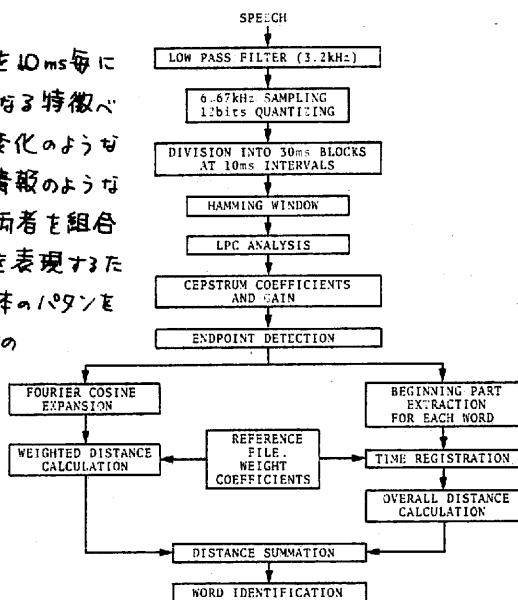


Fig.1. Block diagram indicating the principal operation of the word recognition system.

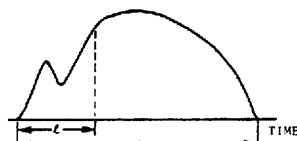


Fig.2. Power pattern of a word utterance.

Table 1. Utterance sets

	Speakers	Recording Cond.
10 digits	15 male	Quiet room
CV syllables	10 male	Terminal room

* An Isolated Word Recognition System Based on the Combination of Global and Local Spectral Transitions. By Sadaoki FURUI (Musashino ECL, NTT)

して用いた。

10数字音声の場合の平均認識誤り率と学習サンプル数の関係を図3に、大局的・局所的特徴を組合せるときの重み係数と誤り率の関係を図4に示す。^{注)} 図3には、2種の特徴をそれぞれを単独に用いる場合と、最適な重み係数によって組合せる場合の結果が示してある。全音声サンプルのうち、学習サンプルを除いた残りを未知入力音声として用いた。図3には、1単語学習の場合について、単語全区間のDPマッチングを行ったときの結果をあわせて示してあるが、誤り率は2特徴の組合せによるものが1/2程度に小さい。DPによる場合の誤認識は無声化が関連しているものが多い。3サンプル学習を行い、2特徴を組合せれば平均99.8%の認識率を得ることができる。

C/a/ および C/i/ の形の単音節の認識結果を図5に示した。Cとして用いた音韻と認識誤りは表2に示す通りで、/a/ と /i/ の場合でCの音韻が共通していないが、誤りの傾向はよく似ている。2特徴の組合せ効果は10数字の場合ほど大きくないが、誤り率はDPによる場合よりも小さくなっており、3サンプル学習の場合はC/a/ で93%、C/i/ で85%の認識率を得られている。

注) 図4で重み係数 $r=0$ は大局的特徴のみ、 $r=\infty$ は局所的特徴(語頭スペクトル)のみを用いることに相当する。

Ⅳ むすび スペクトル時系列の大局的・局所的特徴の組合せによる本方法は、DPマッチングを用いないため処理が簡単で、10数字認識の場合、単語全区間のDPマッチングを行う場合の1/10程度の計算量で済むという利点があり、有効性も高い。単音節認識に関しては、提案されている他の方法(富士通・東芝・NEC等)との比較等を行う必要がある。
[謝辞] 日頃御指導頂く、当所野田基礎研究部長、小池孝四研究室長、ならびに熱心に討論頂く研究員の皆様に感謝する。

Table 2. Confusion matrices for CV syllable recognition

out in	a	ka	sa	ta	na	ha	ma	ya	ra	wa
a	76					4				
ka		77				3				
sa			69	11						
ta			1	15	63		1			
na					80					
ha		6	7			67				
ma							80			
ya								79	1	
ra			1		1	4			74	
wa										80

out in	i	ki	ji	ti	ni	hi	mi	ri	bi	pi
i	77			1		2				
ki		55		1		23				1
ji			63	17						
ti			1	21	55		2			1
ni					75		4	1		
hi		4	16		1	59				
mi						1	79			
ri		1					1	75	1	2
bi									1	76
pi			1	5		1	4			69

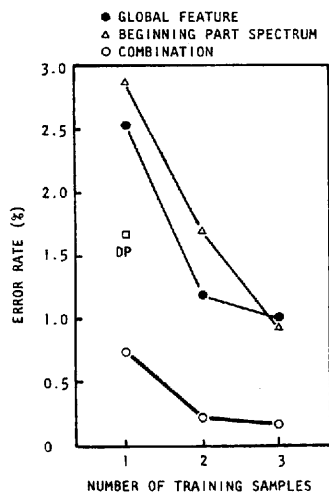


Fig.3. Results for 10 digits.

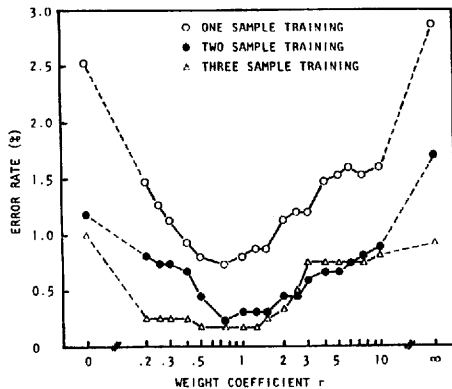


Fig.4. Error rate vs. weight coefficient in 10 digits recognition.

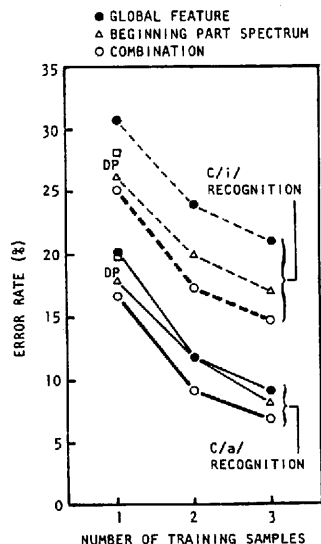


Fig.5. Results for CV syllables.