様式 C-19

科学研究費補助金研究成果報告書

平成２１年３月３１日現在

研究種目：若手研究（B）
研究期間：2006～2008
課題番号：18700174
研究課題名（和文）
ゆっくり喋ると認識されやすい音声認識システムの開発
研究課題名（英文）
Development of speech recognition system which can recognize slow speaking utterance
研究代表者
山本　一公（KAZUMASA YAMAMOTO）
豊橋技術科学大学・工学部・助教
研究者番号：40324230

研究成果の概要：音声認識システムを使用時に誤認識が発生すると、人間であるユーザは訂正発話をゆっくりとした発話速度で行う傾向があるが、現在の音声認識システムはゆっくりした発話速度の音声の認識精度が悪いため悪循環となっている。本研究では、発話速度を自動推定し、その結果から最適な音声認識システムパラメータ（挿入ペナルティ、言語重み）を自動的に調整することでゆっくりした発話速度の音声の認識精度を大幅に改善した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	1,700,000	0	1,700,000
２００７年度	800,000	0	800,000
２００８年度	800,000	240,000	1,040,000
年度			
年度			
総　計	3,300,000	240,000	3,540,000

研究分野：音声情報処理
科研費の分科・細目：情報学、知覚情報処理・知能ロボティクス
キーワード：音声認識、発話速度変動、挿入ペナルティ、言語重み、訂正発話、発話速度推定、対角共分散行列、全共分散行列

１．研究開始当初の背景

近年の音声認識技術の発達により、静かな環境で丁寧に読み上げられた音声に対しては、90%以上という高い単語認識率での音声認識が可能となってきた。しかし、自由に発話された音声（話し言葉音声）に対する認識性能は 70%～80%程度の単語正解精度であり、「自然で誰でも使えるユーザ・インタフェース」の実現には遠い状況である。話し言葉音声の認識精度向上は、自然なマン―マシーン・インタフェースの実用化に欠かすことができない重要な課題となっている。

現在、音声の特徴を表現するための音響モデルとして、隠れマルコフモデル（Hidden Markov Model; HMM）が重要な基幹技術として広く用いられている。しかし、HMM は統計的手法であることから、学習データ中の出現頻度が高い発話速度で最も認識率が高くなるモデルであり、それより速く発話しても、遅く発話しても認識精度が低下してしまうという問題がある。そのため、音声認識システムがうまくユーザの声を認識するためには、ユーザがある一定の速度で発声する必要があるが、これでは自然なマン―マシーン・インタフェースとは言えない。

一方で、現在種々の対話システムが研究・

開発されているが、音声認識率が 100%でないため、ユーザが音声認識誤りを訂正しなければならない場面が少なくない。訂正は言い直し等により行われるが、ユーザは言い直し発話をゆっくりと行う傾向がある。これは、「人間同士の会話ではゆっくりと発声することで、相手が聞き取りやすくなる」という経験から起こる現象であり、人間としては自然な行動である。しかし、前述の通り、現在の音声認識システムではゆっくり発声することで逆に認識精度が低下してしまうため、ユーザビリティ低下の一因となっている。

２．研究の目的

（１）本研究の目的は、ゆっくりした訂正発話を頑健に認識できるようにすることである。訂正発話の誤認識が減少することで、音声認識システムのユーザビリティが向上し、ユーザ・インタフェースとしての音声認識システムの可用性が高まることが期待される。

（２）もう一つの目的は、音声認識精度そのものを向上させることである。音声認識精度が向上すれば、音声認識システムを使用できる場面が増え、音声認識システムの実用レベルでの普及が見込まれる。

３．研究の方法

（１）まず、発話された音声がゆっくりと発話されたものかどうか判別するために、発話速度を推定する必要がある。本研究では、音節制約付きの連続音素認識結果から、母音部分だけを抜き出し、その継続フレーム数から発話速度を推定することとした。その際、最初に母音あたりの平均継続フレーム数を算出し、その平均値より 1.5 倍以上長い継続時間を持つ母音は 2 重母音や長母音である（2 母音分の継続長がある）として母音数を設定し直し、再度平均継続フレーム数を計算したものを母音の平均継続フレーム数とした。このようにして求めた母音の継続フレーム数の逆数と実際の発話速度の関係を図１に示す。図より、非常に高い相関を持っていることが分かる。このことから、母音部の平均継続フレーム数を求めることで、発話速度を推定することが可能であることが分かる。

　発話速度の推定が可能であれば、発話速度に合わせて最適な認識用パラメータ（挿入ペナルティ、言語重み）を推定すれば良い。推定を行うために、開発データセットに対して、挿入ペナルティ、言語重みを様々に変化させて認識実験を行い、認識精度が最も良かったパラメータと発話速度の関係を回帰分析し、発話速度から最適なパラメータを推定するようにした。挿入ペナルティの適用方法としては、従来用いられている、「単語数に比例した挿入ペナルティ」と、発話速度により良く対応していると考えられる「音節数に比例した挿入ペナルティ」の両方を用いた。

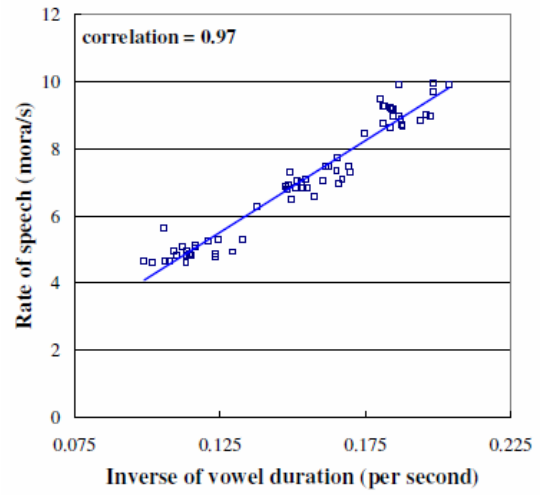

図１　実際の発話速度と
母音の継続フレーム数の関係

（２）現在の音声認識システムでは、音響モデルとして隠れマルコフモデル（HMM）が広く用いられている。HMM の状態は、ガウス混合モデル（GMM）を持ち、それによって特徴ベクトルの出力確率を与えるが、学習データが十分でない場合が多いため、頑健性のためにガウス分布には無相関分布（対角共分散行列（図２））が使われるのが一般的である。しかし、対角共分散行列はパラメータ間の相関を無視しているため、分布の表現能力は十分ではない。従来は、混合数を増やすことによってこれに対処してきたが、最近は使えるデータが増えたことによって、全共分散行列やそれに近いブロック共分散行列（図３）が使われるようになってきた。

　本研究では、より効率的な共分散行列のパラメータの使用方法を検討した。現在の音声認識システムで一般的に用いられるのは、メル周波数ケプストラム係数（MFCC）と、その時間方向の１次微分（Δパラメータ）、２次微分（ΔΔパラメータ）である。実際に全共分散行列で HMM のガウス分布を学習してみたところ、同じ次元の MFCC とΔパラメータ、MFCC とΔΔパラメータの間に相関が見られ（図４）、ブロック共分散では無視されている静的パラメータと動的パラメータ間の相関を考慮することが望ましいことが分かった。ただし、闇雲にパラメータを削減すると共分散行列の正定値性が損なわれるため、正定値性を確保した上でパラメータを削減しなければならない。そこで、図５～図８のようなパターンを提案し、これにより認識実験を行った。

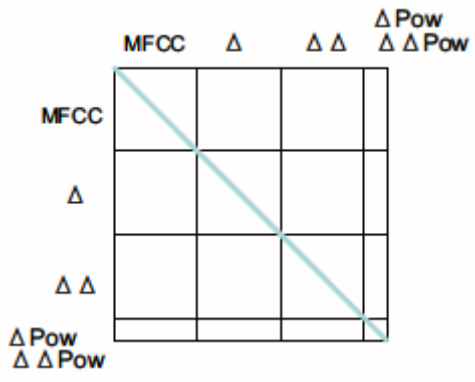

図２　対角共分散行列

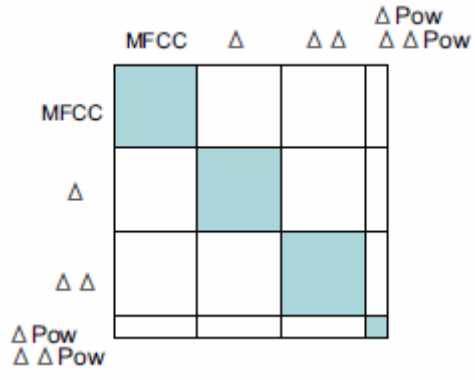

図３　ブロック共分散行列

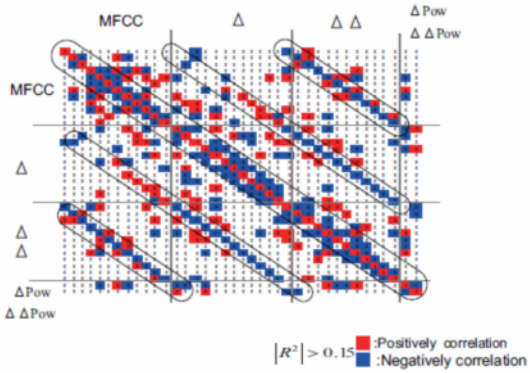

図４　実際に学習した全共分散行列

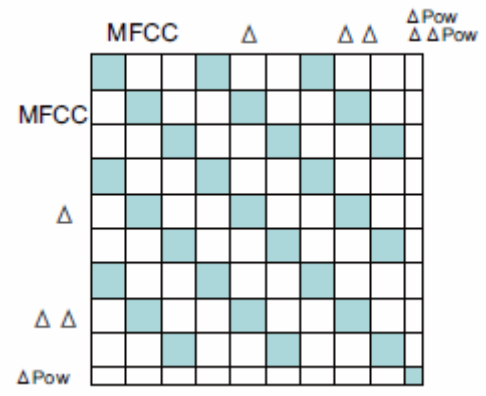

図５　提案手法：パターン A

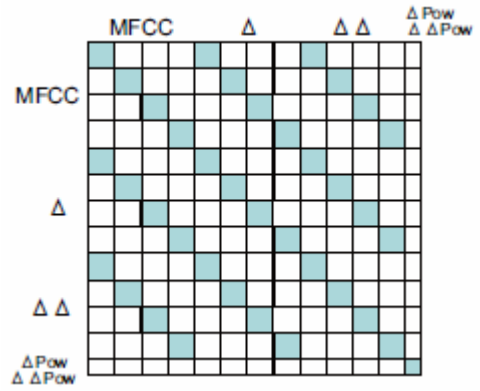

図６　提案手法：パターン B

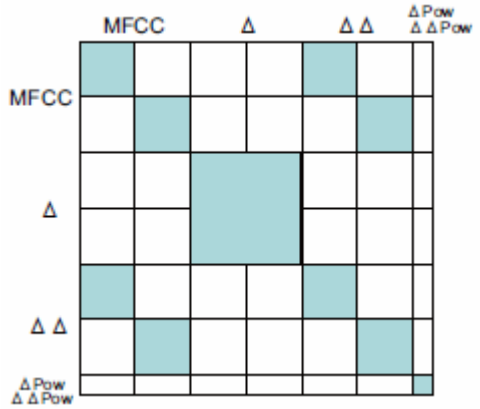

図７　提案手法：パターン C

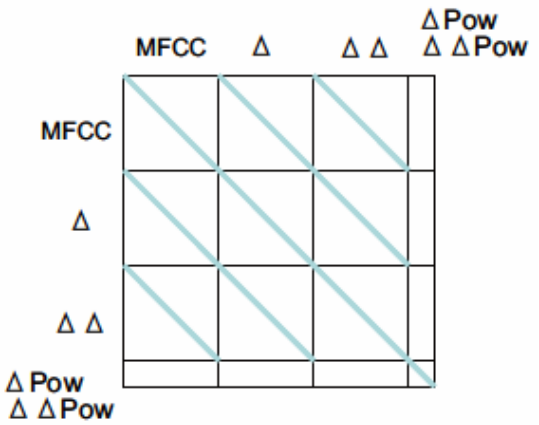

図８　提案手法：パターン D

４．研究成果

（１）発話速度をコントロールして読み上げた文章に対する大語彙連続音声認識により評価した結果を表１に示す。表中、”baseline”が、普通話速（normal）で認識率が最適になるように手動でパラメータを調整し、そのパラメータを速い話速(fast)と遅い話速(slow)にそのまま適用した場合の結果である。挿入ペナルティは単語数に比例した挿入ペナルティである。”wordmax”は従来法である単語単位の挿入ペナルティと言語重みを、それぞ

れの話速に対して最適になるように手動で設定した結果である。”Auto”が提案法で、発話速度を自動推定し、その結果に応じて、音節数に比例した挿入ペナルティ・言語重みを自動的に設定した場合の結果である。”manual”は発話文毎に認識率が最適になるように音節数に比例した挿入ペナルティ、言語重みを手動で調整した場合の結果であり、提案法の上限となる。表中の数値は、”Cor”が単語正解率、”Acc”が単語正解精度、”Del”が単語脱落誤り率”Sub”が単語置換誤り率、”Ins”が単語挿入誤り率を示す。表から分かるように、従来法である単語数に比例した挿入ペナルティを調整した場合よりも、提案法の方が発話速度が遅い場合に大幅に認識精度を改善できる。また、普通話速、速い話速の場合の認識精度劣化も最低限に抑えられていることも分かる。

本研究では、音節数（モーラ数）を発話速度の指標として用いているため、音節言語以外の言語（例えば英語）に直接適用することは難しいと考えられるが、発話速度が推定できさえすれば、同じ枠組みで認識率を改善することができると考えられる。比較的シンプルな手法だが、効果は大きく、音声認識システムのユーザビリティ向上に貢献すると考える。本研究では、発話速度をコントロールした読み上げ音声での評価に留まったが、今後は発話中に話速が変化する自然発話音声にこの手法を適用することを考える。

表１　大語彙連続音声認識結果

	method	Cor	Acc	Del	Sub	Ins
fast	baseline	74.8	72.5	4.9	18.0	3.0
	wordmax	74.8	72.5	4.9	18.0	3.0
	Auto	74.7	71.9	4.6	17.9	2.8
	manual	85.0	84.3	2.8	11.4	0.8
normal	baseline	89.5	87.6	1.6	7.0	1.9
	wordmax	89.5	87.6	1.6	7.0	1.9
	Auto	88.2	87.3	2.9	8.1	0.9
	manual	94.3	93.8	1.0	4.2	0.5
slow	baseline	59.3	45.6	1.9	25.1	13.7
	wordmax	62.5	50.8	2.7	23.2	11.6
	Auto	78.0	76.3	6.7	13.6	1.8
	manual	88.5	87.0	2.3	7.8	1.4

（２）それぞれのパターンにおける共分散行列のパラメータ数を表２に示す。”Mix.”はガウス混合分布の混合数を示す。それぞれのパターンにおける連続音節認識実験の結果を表３に示す。提案パターンにおいて、静的パラメータと動的パラメータ間の相関を考慮することで、同程度のパラメータ数においてもより高い認識精度が得られていることが分かる。この手法により、従来の音声認識システムの枠組みを変えることなく、音声認識精度の向上を図ることができた。

表２　各パターンにおけるパラメータ数

Mix.	diag.	block	pattern			
			A	B	C	D
1	76	276	276	222	276	112
2	152	552	552	444	552	224
4	304	1104	1104	888	1104	448
8	608	2208	2208	1776	2208	896
16	1216	-	-	-	-	1792
32	2432	-	-	-	-	-

表３　各パターンにおける
連続音節認識実験結果

Mix.	diagonal		block		paternA		paternB		paternC		paternD	
	Cor.	Acc.	Cor.	Acc.	Cor.	Acc.	Cor.	Acc.	Cor.	Acc.	Cor.	Acc.
1	64.8	41.6	70.1	50.6	71.3	54.3	70.2	53.1	71.0	53.1	65.7	46.1
2	68.8	48.8	74.0	57.8	74.5	60.6	73.4	58.8	74.2	60.2	70.5	54.6
4	72.0	54.5	76.9	62.1	77.3	65.8	76.3	64.3	77.3	65.0	73.4	59.1
8	75.2	60.3	79.1	65.2	79.3	68.8	78.4	67.5	79.7	69.0	76.0	63.5
16	76.9	62.8	-	-	-	-	-	-	-	-	77.8	66.4
32	77.9	64.5	-	-	-	-	-	-	-	-	-	-

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計　２　件）

①荻山将成、山本一公、藤井康寿、中川聖一、”挿入ペナルティの自動推定による遅い発話に対する音声認識性能の改善”、日本音響学会 2009 年春季研究発表会、2009 年 3 月 19 日、東京工業大学.

②末吉英一、山本一公、中川聖一、”音声認識における多次元ガウス分布の全共分散行列の要素制限手法”、日本音響学会 2009 年春季研究発表会、2009 年 3 月 17 日、東京工業大学.

６．研究組織
(1)研究代表者

山本　一公（KAZUMASA YAMAMOTO）
豊橋技術科学大学・工学部・助教
研究者番号：40324230