Internal Use Only（非公開）

TR－SLT－0031

発話速度の補正を用いた講演音声の認識

Speaking Rate Compensation in Lecture-Style Speech Recognition

新宮将久
Masahisa Shingu
新美康永
Yasuhisa Niimi
實廣貴敏
Takatoshi Jitsuhiro
中村哲
Satoshi Nakamura

2003年2月21日

概要

話し言葉の音声認識に対して、現状の音声認識技術では未だ十分な精度は得られていない。発話速度の変動が認識性能に影響を与えることが問題点の一つとして挙げられる。本研究では講演音声を対象とし、発話速度を補正して認識する手法を検討した。分析周期を5msecと細かくした上で、隣接するフレーム間でのMFCCのユークリッド距離を元に変動の小さいフレームを間引くことで発話変動を補正する。この手法とフレーム周期をいくつか変えたものとの比較を種々の条件で行った。更に音響モデルの学習にもこの手法を用い、認識実験を行った。また、発話速度、フレーム周期、認識精度の関係についても調べた。

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目2番地2 TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

1. はじめに

話し言葉の音声認識に対して、現状の音声認識技術では未だ十分な精度は得られていない。発話速度の変動が認識性能に影響を与えることが問題点の一つとして考えられる。本研究では、「日本語話し言葉コーパス(CSJ)」で提供されている講演音声を対象とし、発話速度を補正して認識する手法を検討した。

一般的な音声認識システムにおいて、音声信号は最初にフレーム単位に区分される。この際に一定の分析周期で分析する方法では、同一発話内の変動をとらえることができないと考えられている。速い音声に対しては脱落誤りや置換誤りが、遅い音声に対しては挿入誤りが多く発生することが報告されている。分析周期を固定するのではなく、発話速度に応じた適切な分析周期で各音声を分析することが望ましいが、認識時には発話速度が分からないので問題となっている。そこで、発話速度の正規化や補正が重要であると考えられる。

[1]では、複数の分析周期、窓長を用いて並列に認識した結果から尤度を比較し、発話毎に適した分析周期と窓長を選択するという手法が提案されている。この手法によって発話速度の変動の大きな講演の認識率を向上することができた。また、同様の手法を音響モデル学習時にも用いることで更に認識性能の改善を得ることが出来た。しかし、同一発話内での発話速度の変動には対応できないという課題があった

[2]では、予め分析周期を一般的に用いられる値より細かくとって分析し、隣接するフレーム間のMFCCのユークリッド距離を計算し、蓄積していき、その値が定められた閾値を超えた時点で、そのフレームを選び、累積値をクリアする。この処理を繰り返してゆき、認識用のパラメータを作成するという手法を提案している。これは、発話内での発話速度の変動を考慮したもので、特にSNRが低い場合に関して認識性能の向上が得られたと報告されている。

[2]の結果を受けて、[3]では、隣接するフレーム間のMFCCのユークリッド距離と発話速度との関係について考察し、更に発話速度の変動にロバストな音声認識を行うための手法を検討した。音声を予め細かく分析し、隣接するフレーム間のMFCCのユークリッド距離から定常区間を調べ、定常な区間を圧縮することにより、発話速度の補正を試みた。CSJの4話者による講演音声に対してこの手法を施し、新聞読み上げによって学習した音響モデルによる認識実験を行い、認識性能の僅かな改善が得られたと報告されている。しかし、認識精度はまだまだ十分ではなく、まだまだ検討が必要であると考えられる。

本研究では、まず、音響モデルを固定し、認識データに対して数種類の分析周期を用いて認識実験を行い、その結果を発話速度毎に集計することにより、発話速度と分析周期、認識精度との関係について検討した。更に、学習データの分析条件を変えた複数の音響モデルを用意し、音響モデルと認識データの分析条件を合わせた場合の認識精度について検討した。

また、[3]の手法を利用し、分析周期を従来用いる 10msec から 5msec と細かくした上で、隣接するフレーム間のMFCCのユークリッド距離と閾値を元に変動の小さな部分を間引くことで、変動の大きな部分を細かく、変動の小さな部分を粗く分析し、発話速度の補正を行った。更に同手法を音響モデルの学習時にも用い、更なる認識性能の向上を図った。

最後に、[1]の手法を利用し、発話文毎に最適な閾値と音響モデルを認識結果からの尤度を基準に自動的に選択するという方法について試みた。

2. VFR アルゴリズム

2.1 概要

本研究では発話速度の補正を行うための手法として VFR(Variable Frame Rate)アルゴリズムを検討した。この手法は、まず、従来の方法よりも細かく分析することで時間分解能を上げ、隣接するフレーム間の MFCC のユークリッド距離と閾値を元にして、変動の小さいフレームを間引くことを行うことによって、発話速度の補正を行う。実装したアルゴリズムについて図 1 に示す。各処理の詳細については後述する。

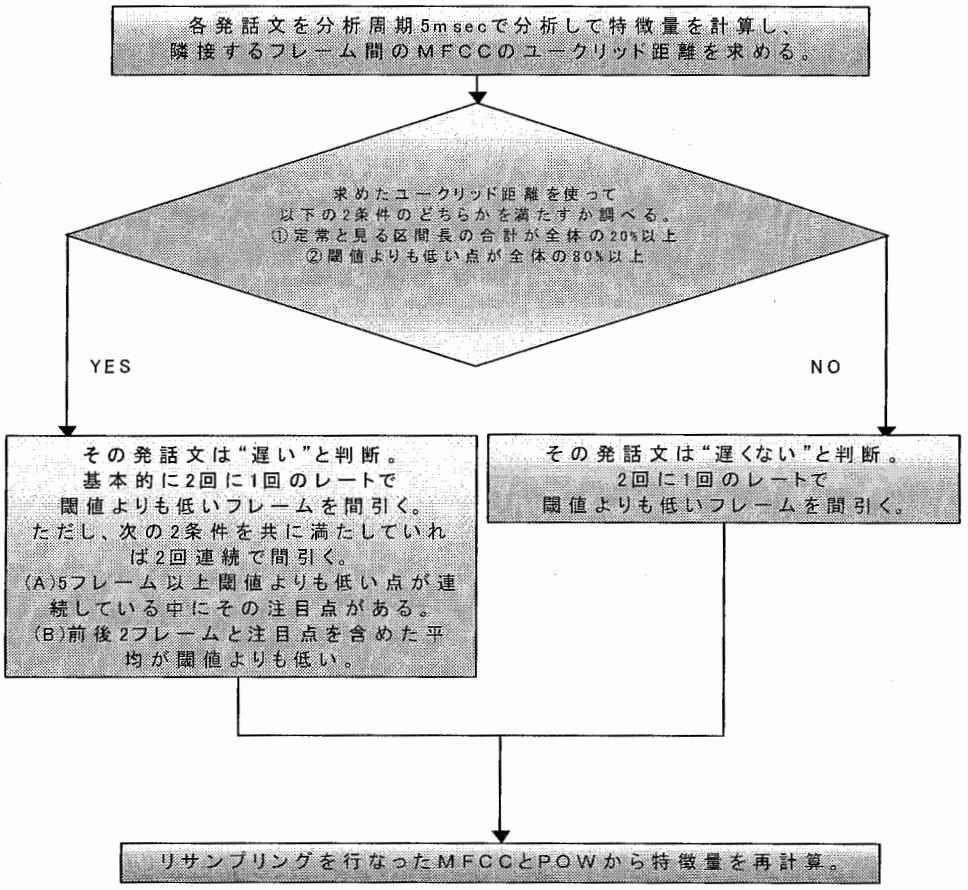


図 1：VFR アルゴリズム

2.2 分析

従来、分析周期は 10msec であったが、本手法では 5msec で分析を行い、MFCC とパワーを計算する。そして、それぞれのフレームに関して、式(1)のようにして、隣のフレームとの MFCC ユークリッド距離を求める。この式において、i はフレーム番号、j は MFCC の次元数(12 次元)をあらわしている。本手法において、窓長は 20msec に固定した。

$$dis = \sqrt{\sum_{j=1}^{12} \left(feature_{i,j} - feature_{i+1,j} \right)^2} \qquad (1)$$

2.3 発話速度の判定

VFR では同一発話内での発話速度の補正を行うのであるが、より精度を上げるために、各発話文に対して、定常な部分の多い『遅い』音声と、そうでない『遅くない』音声のどちらかに判定する。『遅い』音声に対しては後の間引きの処理でたくさん間引くように、『遅くない』音声に対してはそれほど間引かないようにという制約をつける。

発話速度の判定には 2.2 で求めた各フレームの隣のフレームとの MFCC のユークリッド距離を用いる。

5msec で分析した場合、50 フレーム(250msec)を定常区間として見ることが最適であるということが[3]により報告されており、発話速度を判定するために定めた閾値より低い点が 50 フレーム以上続いていれば、その区間は定常区間であると見なされる。その発話内における定常区間長を合計してゆき、その値が全体のフレーム数の 20%以上であれば、その発話文は定常区間の占める割合の多い『遅い』音声であると見なすことにする。また、閾値より低い点が全体の 80%以上を占めていれば、その音声は全体的に変動の小さい定常な音声、つまり『遅い』音声であると見なすことにする。これら 2 つの条件のどちらかを満たせば、『遅い』音声と判断され、そうでなければ『遅くない』音声であると判断する。

2.4 間引き処理

間引き処理は閾値より低いユークリッド距離を持つフレームが制約つきで間引くということを行っている。制約が『遅い』音声と『遅くない』音声とで異なる。

『遅い』音声は粗く分析することが望ましいので、たくさん間引くように、『遅くない』音声は細かく分析することが望ましいので、あまり間引かないように処理を考えた。

2.4.1 遅い音声に対する間引き処理

まず、『遅い』音声用の間引き処理について説明する。基本的に、2 回に 1 回の割合で閾値よりも低いユークリッド距離を持つフレームを間引く。但し、以下の 2 条件を共に満たしていれば、その注目点は局所的な定常区間に存在すると見なし、1 つ前のフレームを間引いていても間引くことにする。ただし、前 2 フレームが連続して間引かれている場合は、2 条件を満たしていても注目点を間引かないことにする。

(A)5 フレーム以上閾値より低い点が連続している中にその注目点がある。
　これは、注目しているフレームが局所的に定常な区間の中にあるかを見ている。
(B)前後 2 フレームと注目点を含めた 5 フレームの平均が閾値よりも低い。

この処理によって最大で 3 回に 2 回の割合(33%)で間引かれることになる。

2.4.2 遅くない音声に対する間引き処理

『遅くない』音声については、2 回に 1 回の割合で閾値よりも低いユークリッド距離を持つフレームを間引く。この処理によって最大で 50%まで間引かれることになる。

2.5 回帰係数の計算

間引き処理後、残された MFCC とパワーは不等間隔に並んでいる。それを詰めていくことで等間隔に並べる。そして、ΔMFCC とΔパワーを計算する。

3. ベースライン認識システムと評価データセット

3.1 ベースライン認識システム

本研究では以下に示す認識システムをベースラインとして用いた。

3.1.1 音響特徴パラメータ

音響特徴パラメータは、16kHz、16bit でサンプリング、分析周期 10msec、分析窓長 20msec の条件で抽出した 25 次元の特徴ベクトル(12 次メルケプストラム、12 次Δメルケプストラム、1 次Δ対数パワー)を用いている。これらの処理は ATR で開発された ATRSPREC[4]を使用した。

3.1.2 ベースライン音響モデル

ベースとなる音響モデルは、モニター版のうち 10 話者(A01M0035, A01M0007, A01M0074, A05M0031, A02M0117, A03M0100, A06M0134, KK99DEC005, YG99JUN001, YG99MAY005)を除く全ての学会講演、模擬講演男性話者データ 200 名(約 34 時間)のデータから作成されたものを用いた。モデルは状態共有化 HMM(HMNet)により構築された性別依存モデルであり、各音素モデルは 3 状態(状態の飛び越し遷移なし)、10 混合ガウス分布、総状態数 1400 で表現されている。なお、評価話者は全て男性であるため、モデルは男性モデルのみを用いている。

3.1.3 言語モデル

言語モデルは京都大学で作成され、モニター版と共に配布された講演音声用言語モデルの前向き単語バイグラムと後ろ向き単語トライグラムを用いた。辞書についても同様に配布されている 19k 単語のものを用いた。

3.1.4 デコーダ

デコーダとして Julius3.3[5]を使用した。

3.2 評価テストセット

近年、話し言葉の音声認識技術を高めることを目標とした開放的融合研究『話し言葉工学』プロジェクト[6]が開始された。

本研究では、上記プロジェクトで配布されている『日本語話し言葉コーパス(CSJ)』モニター版(2001)から、表 1 に示す男性話者 4 名の講演音声を用いて評価を行った。モニター版に含まれる書き起こしデータに記述されている時刻情報を元に、発話単位にファイルを分割している。

表 1：評価テストセット

話者	講演時間	単語数
A01M0035	28 分	6127
A01M0007	30 分	4302
A01M0074	12 分	2486
A05M0031	27 分	5305

4.発話速度と分析周期、認識精度との関係

発話速度と認識性能との関係を調べるために、ベースライン音響モデルを用いた認識実験を行った。まず、話者毎による集計を行い、更に、発話文毎に細かく集計を行った。
更に、学習データの分析周期を変えた数種類の音響モデルを用意した実験を行った。

4.1 話者毎に集計した発話速度と認識性能との関係

図2に認識データの分析周期を10msec、窓長を20msecに固定して分析し、ベースライン音響モデルを用いて認識を行った場合の各話者の単語誤り率と平均発話速度をまとめる。図中における各話者の平均発話速度は、文毎に算出した1秒あたりのモーラ数の平均を示している。1秒あたりのモーラ数は、ビタビアラインメントにより算出した音声区間の時間長で文中のモーラ数を割った値である。講演音声においてはフィラーの出現頻度が多いため、この値が必ずしも正確なものではないが、傾向を知ることはできると思われる。

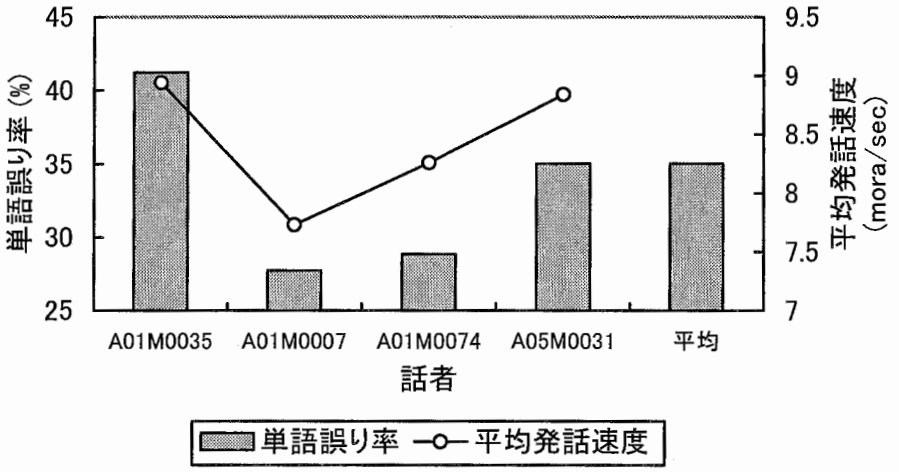


図2：単語誤り率と平均発話速度との関係

話者4名の平均単語誤り率は35.01%である。図2より発話速度の速い話者ほど単語誤り率が高いという傾向が観測される。このことより、特に発話速度の速い音声に対応できるように発話速度の補正をすることが必要であると考えられる。

4.2 話者毎に集計した分析周期と認識性能との関係

認識データの窓長を 20msec に固定し、分析周期を 5msec～12msec と変化させて分析し、ベースライン音響モデルで認識した場合の話者毎に最適な分析周期について調べ、表2にまとめた。

表 2：各話者の最適な分析周期とその認識率

話者	平均発話速度	分析周期	SUB	INS	DEL	WER	COR	ACC
A01M0035	8.94	10msec 固定	26.55	4.44	10.24	41.23	63.21	58.77
		最適 8msec	24.87	5.84	7.65	38.36	67.48	61.64
A01M0007	7.73	10msec 固定	17.13	3.95	6.62	27.71	76.24	72.29
		最適 9msec	16.69	4.56	5.88	27.13	77.43	72.87
A01M0074	8.26	10msec 固定	18.42	4.42	5.99	28.84	75.58	71.16
		最適 9msec	17.82	4.55	5.55	27.92	76.63	72.08
A05M0031	8.84	10msec 固定	24.67	2.39	10.61	37.68	64.71	62.32
		最適 9msec	24.41	2.88	8.97	36.27	66.62	63.73
平均		10msec 固定	22.47	3.69	8.85	35.01	68.68	64.99
		最適な結果の平均	21.69	4.43	7.31	33.43	71.00	66.57

どの話者もベースとして用いられている分析周期 10msec で固定して分析するという方法より細かく分析する方が認識精度が高いということが分かる。話者毎に最適な分析周期を選んだ場合と分析周期を 10msec で固定した場合とを比較すると単語正解精度に関しては 4 話者の平均で 1.58%の差があった。

話者数が少ないことなどもあり、表 2 の結果からは発話速度と最適な分析周期との関係について十分な考察を行うことはできなかった。

4.3 発話文毎に集計した発話速度と分析周期、認識性能との関係

認識に用いる 4 話者の講演音声から発話速度が求められた発話文だけを用い、認識率の集計を文毎に行い、発話速度毎の単語認識精度と分析周期との関係を調べた。実際にどの話者の発話を何文使ったかということについては、表 3 に示してある。図 3 にこの講演の各文を発話速度によって分類し、その頻度を示した。

表 3：集計に使った文の数

話者	A01M0035	A01M0007	A01M0074	A05M0031	全体
使った文の数	554 (-23)	783 (-4)	253 (0)	449 (-34)	2039 (-61)
全発話文の数	577	787	253	483	2100

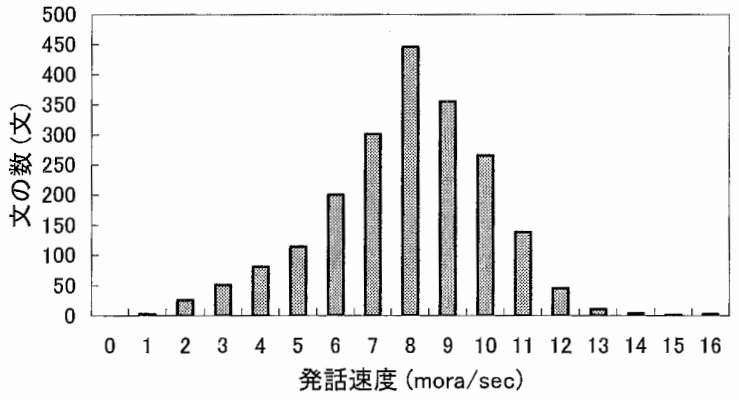


図 3：集計に用いた文の発話速度による分類

図 4 に発話速度毎に各分析周期での単語正解精度を計算しグラフにしたものを載せておく。この図には、5msec で分析し、閾値 1.0 で VFR による間引きを行った場合の結果(5 節で詳しく説明する)についても載せておく(太い線)。更に、単語正解精度で色分けして図 5 に、それを順位付けして図 6 に示す。

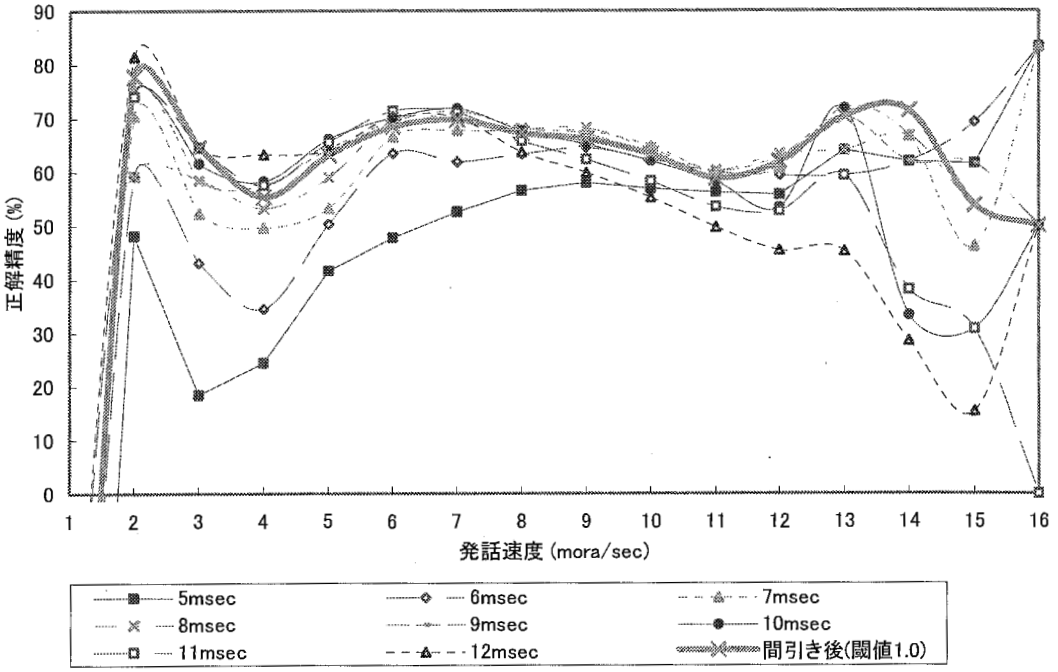
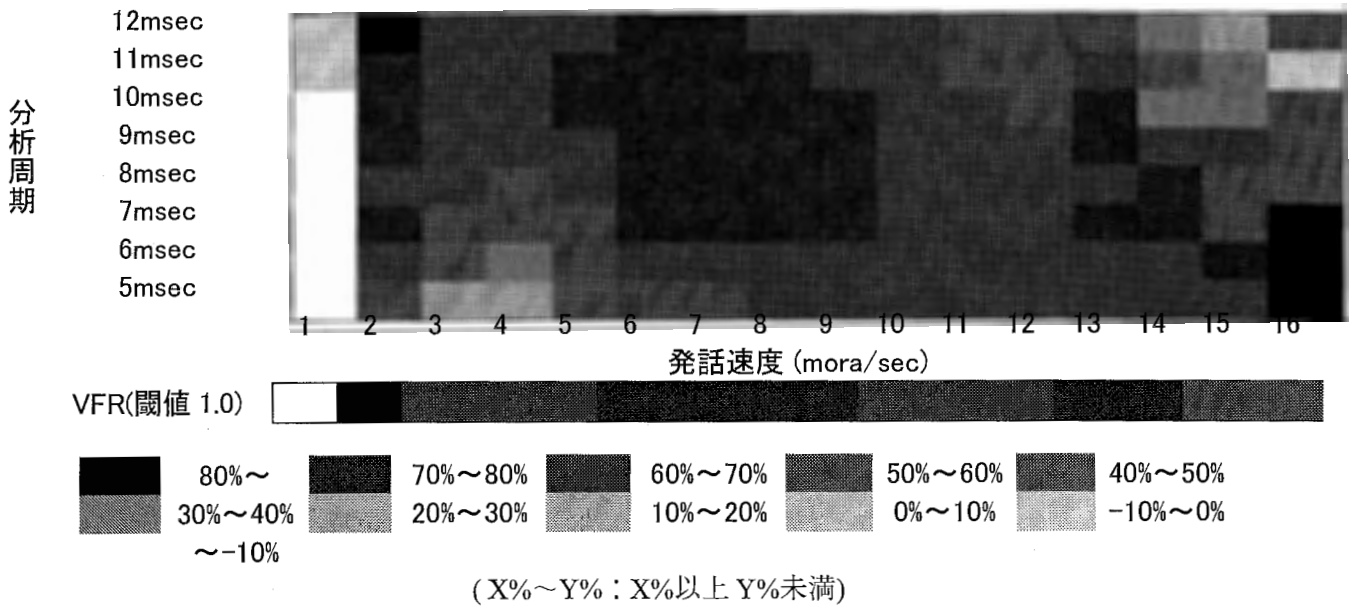


図 4：発話速度毎の各分析周期における単語正解精度



(X%～Y%：X%以上 Y%未満)

図 5：発話速度と各分析周期における認識精度

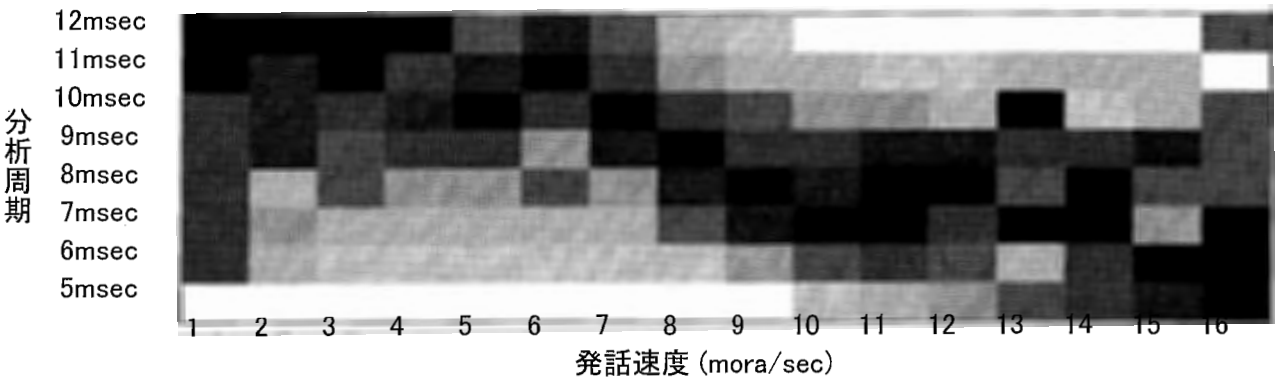


図 6：各分析周期の単語正解精度を基準とした順位の発話速度による分類
(黒に近い色が順位が高い)

(見方) 縦一列を見ると、分析周期 8 条件があり、
それらを単語正解精度の高い順に濃い黒->白と色分けしている。

結果があまりきれいではなかったことや、データが十分ではなかったことなどがあるが、この結果から速い発話文は分析周期を小さくして、遅い発話文は分析周期を大きくすると精度が良くなる傾向にあるということが言える。

4.4 音響モデルと評価データとの分析条件を同じにした場合の認識

音響モデルの学習に用いるパラメータの分析周期を変化させ、数種類の音響モデルを作成することを試みた。ベースライン音響モデルの条件において分析周期を変えた 4 条件 (6msec, 8msec, 9msec, 10msec)で分析したデータを学習に使った 4 つのモデルを作成した。
各モデルと同じ条件で評価データを分析し、認識を行った結果を表 4 に示す。

表 4：モデルと評価データとの分析条件を同じにした場合の最適な分析周期と認識率

話者	分析周期	SUB	INS	DEL	WER	COR	ACC
A01M0035	10msec 固定	26.55	4.44	10.24	41.23	63.21	58.77
	最適 8msec	24.93	5.78	7.79	38.49	67.29	61.51
A01M0007	10msec 固定	17.13	3.95	6.62	27.71	76.24	72.29
	最適 9msec	16.60	4.58	5.88	27.06	77.52	72.94
A01M0074	10msec 固定	18.42	4.42	5.99	28.84	75.58	71.16
	最適 8msec	17.70	5.39	4.91	28.00	77.39	72.00
A05M0031	10msec 固定	24.67	2.39	10.61	37.68	64.71	62.32
	最適 9msec	24.41	2.90	8.88	36.19	66.71	63.81
平均	10msec 固定	22.47	3.69	8.85	35.01	68.68	64.99
	最適な結果の平均	21.66	4.54	7.24	33.44	71.10	66.56

話者間において、最適な分析周期が多少異なるものの、8msec もしくは 9msec が最適であることが分かる。また、表 2 の最適な結果の 4 話者の平均単語認識精度は 66.67%であった。表 4 の結果はこれに比べ 0.11%劣っている。モデルと評価データの分析条件を同じにするよりも、モデルに合うような評価データの分析周期を選んだ方が効果があると言える。

5. VFR アルゴリズムを用いた認識実験

5.1 認識データへの VFR の適用

VFR を用いて認識データを処理し、ベースライン音響モデルを用いて認識実験を行った。発話速度の判定および、間引きに用いる閾値はあらかじめ実験を行い、0.4～1.2 の範囲の値を使うことに決めた。結果を表 5～表 9 に示す。また、4 話者の平均を図 7 にまとめた。

表 5：VFR を適用した評価データの認識率(A01M0035)

閾値	間引き後何%になったか	SUB	INS	DEL	WER	COR	ACC
0.4	93.95	27.93	10.87	5.95	44.75	66.12	55.25
0.5	88.82	26.42	9.95	6.70	43.06	66.88	56.94
0.6	83.22	25.52	8.34	6.98	40.85	67.50	59.15
0.7	77.31	25.39	7.46	7.33	40.18	67.29	59.82
0.8	71.63	25.33	7.16	7.38	39.87	67.29	60.13
0.9	66.62	25.56	6.08	7.86	39.51	66.58	60.49
1.0	62.60	25.39	5.51	8.36	39.26	66.25	60.74
1.1	58.97	25.60	4.99	8.72	39.32	65.68	60.68
1.2	54.24	26.55	4.74	9.81	41.11	63.63	58.89

表 6：VFR を適用した評価データの認識率(A01M0007)

閾値	間引き後何%になったか	SUB	INS	DEL	WER	COR	ACC
0.4	93.45	21.62	10.48	4.28	36.38	74.11	63.62
0.5	89.43	21.15	9.25	4.18	34.59	74.66	65.41
0.6	84.78	20.13	7.67	4.51	32.31	75.36	67.69
0.7	79.56	19.41	7.07	4.97	31.45	75.62	68.55
0.8	74.08	19.15	6.44	4.93	30.52	75.92	69.48
0.9	68.89	18.46	6.23	5.07	29.75	76.48	70.25
1.0	64.51	17.81	5.81	5.44	29.06	76.75	70.94
1.1	60.81	17.95	5.23	5.79	28.96	76.27	71.04
1.2	57.07	17.55	4.67	6.07	28.29	76.38	71.71

表 7：VFR を適用した評価データの認識率(A01M0074)

閾値	間引き後 何%になったか	SUB	INS	DEL	WER	COR	ACC
0.4	90.95	20.51	8.57	3.94	33.02	75.54	66.98
0.5	85.63	19.07	7.24	4.22	30.53	76.71	69.47
0.6	80.62	18.79	6.76	4.59	30.13	76.63	69.87
0.7	75.57	18.79	6.6	4.59	29.97	76.63	70.03
0.8	70.48	19.03	5.99	4.63	29.65	76.35	70.35
0.9	65.87	18.62	6.32	5.11	30.05	76.27	69.95
1.0	62.03	18.42	5.79	4.87	29.08	76.71	70.92
1.1	58.71	18.87	5.35	5.31	29.53	75.82	70.47
1.2	54.47	19.43	4.83	6.03	30.29	74.54	69.71

表 8：VFR を適用した評価データの認識率(A05M0031)

閾値	間引き後 何%になったか	SUB	INS	DEL	WER	COR	ACC
0.4	92.52	29.61	8.24	5.45	43.30	64.94	56.70
0.5	86.89	28.86	6.33	6.41	41.60	64.73	58.40
0.6	81.25	27.43	6.03	6.94	40.40	65.64	59.60
0.7	75.54	26.30	4.58	7.03	37.91	66.67	62.09
0.8	70.01	25.69	4.24	7.75	37.68	66.56	62.32
0.9	65.06	25.07	3.41	8.29	36.78	66.64	63.22
1.0	61.02	25.67	2.85	8.82	37.34	65.50	62.66
1.1	57.06	25.82	2.79	9.05	37.66	65.13	62.34
1.2	49.80	27.05	2.26	12.08	41.39	60.87	58.61

表 9：VFR を適用した評価データの認識率(4 話者の平均)

閾値	間引き後 何%になったか	SUB	INS	DEL	WER	COR	ACC
0.4	92.97	25.81	9.64	5.09	40.54	69.09	59.46
0.5	88.01	24.80	8.28	5.63	38.71	69.57	61.29
0.6	82.77	23.80	7.24	6.01	37.05	70.19	62.95
0.7	77.25	23.23	6.36	6.26	35.85	70.51	64.15
0.8	71.75	23.00	5.92	6.49	35.41	70.51	64.59
0.9	66.76	22.65	5.33	6.91	34.89	70.44	65.11
1.0	62.65	22.59	4.81	7.27	34.68	70.13	65.32
1.1	58.95	22.80	4.43	7.60	34.83	69.60	65.17
1.2	53.88	23.45	3.98	9.04	36.46	67.52	63.54

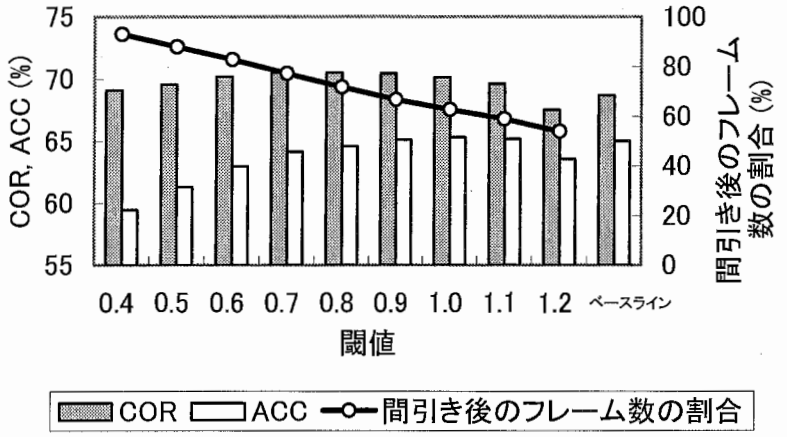


図 7：閾値と認識性能および、間引きの割合の関係(4 話者の平均)

結果より、話者毎に最適な閾値が異なり、閾値が少し変わっただけで、認識精度も影響を受けてしまうことがわかる。

閾値を固定しようとすると、1.0 のものが最も安定していると考えられる。このとき、ベースラインの方法よりも 4 話者の単語正解精度に関して 0.33%の改善が得られた。また、発話速度毎に単語正解精度を集計した結果を図 4 に示す。どの発話速度に対しても、ある程度の精度が得られている。分析周期を固定する方法に比べ、速い音声や遅い音声に対してロバストな認識ができるようになったと考えられるが、認識精度を大幅に改善したわけではない。

5.2 VFR アルゴリズムの高精度化

アルゴリズムの高精度化をねらい、最初に分析する際の分析周期をより小さくすることを考えた。改良点は、最初の分析周期をより小さくしたことと、それによってフレーム数が増えるので、より多くのフレームを間引くようにしたことである。現在のアルゴリズムには連続間引きに関して制約があるので、閾値をある程度大きくしないと思ったように(たくさん)間引かれないので、閾値を大きくするようにするか、間引きの制約を変更するなどを考える必要がある。

また、分析周期を変えたことで、隣接するフレーム間のユークリッド距離も変わる。そのため、5msec で最適であった閾値も他の分析周期では最適であるとは必ずしも言えない。より細かく分析しているので、フレーム間の特徴量の変動は小さくなっていると考えられる。5msec で分析した場合よりも閾値を小さくとる必要があると考えられる。更に、閾値を変化させて認識性能を調べるときもより細かく見ていかなければならないと思われる。

連続して間引くという制約だけでどのくらい間引くかを決めているので、閾値の決め方が間引きに大きな影響を与えると考えられる。

また、細かく分析しておいて、間引き処理を繰り返すという方法も試してみた。

分析周期をいくつか変えて実験を行った(表 11)。各分析周期におけるリサンプリングレートは以下のとおりである。

表 10：リサンプリングレート

最初の分析周期(閾値)	リサンプリングレート (何回中何回間引くか)	
	対象が遅い場合	対象が速い場合
5msec	2/3	1/2
2.5msec	9/10	4/5
1msec	19/20	9/10
1msec(繰り返し間引き)	2/3	1/2

表 11：各分析周期における 4 話者の認識率の平均

最初の分析周期(閾値)	SUB	INS	DEL	WER	COR	ACC
5msec (1.0)	22.59	4.81	7.27	34.68	70.13	65.32
2.5msec (0.6)	28.49	6.92	7.53	42.94	63.98	57.06
1msec (0.45)	38.42	4.76	19.22	62.40	42.36	37.60
1msec 繰り返し間引き(1.0)	24.25	6.66	6.11	37.02	69.64	62.98
デフォルト(分析周期 10msec 固定)	22.47	3.69	8.85	35.01	68.68	64.99
最適フレームシフトの結果の平均	21.69	4.43	7.31	33.43	71.00	66.57

今まで考えてきたアルゴリズムはただ単にフレーム数を増やして間引いているだけであるので、これだけでは認識率を改善することに関して作用があるとは考えにくい。隣のフレームとの MFCC のユークリッド距離から間引きを行っているが、実際にどう間引くかについては、閾値を上下させて、連続で間引く制約をつけている。最終的に何割くらい間引くかということで見ているので、この間引き処理が認識率を上げるように働いているとは必ずしも言えない。良い効果もあれば悪い効果もあると思われる。より認識率を上げるためには、間引き以外の処理が必要となってくる。アルゴリズム自体を考え直さないと高精度化は得られないと考えられる。

アルゴリズム自体をより認識率を上げる方向に改善できれば、5msec で分析するよりも 1msec で分析した特徴量を使う方が望ましいと思われる。1msec で分析し、単純に 1/10 まで間引くという処理(「単純間引き 1/10」：××××××××××○；○残されるフレーム、×間引かれるフレーム)について検討した。この方法による実験結果は 10msec で分析した場合と近いものになった。

表 12：単純間引きによる認識結果(4 話者の平均)

	SUB	INS	DEL	WER	COR	ACC
単純間引き 1/10	22.72	3.81	8.80	35.33	68.48	64.67
10msec で分析	22.47	3.69	8.85	35.01	68.68	64.99

6. VFR を用いた認識における尤度基準による閾値と音響モデルの自動選択手法

各分析周期・窓長で認識した結果から文毎の尤度(音響+言語)が分かる。認識結果から尤度が高くなるような条件を選択して最終的な認識結果にするという方法が[1]では提案されている。この手法を用いることにより、発話文毎に最適な分析周期と窓長を選択することで、発話速度の補正を行うことができると考えられる。

本研究でも同手法を利用して次の4項目について検討した。

(1)『ベースライン音響モデル』を使い、間引きを行った認識データ9通り(閾値0.4～1.2)に対し、文毎に尤度が最良になるような閾値を選択する。

このことにより、『ベースライン音響モデル』に合うように認識データが間引かれるように閾値を選ぶことができる。

(2)『閾値1.0で間引きを行ったデータで学習した音響モデル』を使い、間引きを行った認識データ9通り(閾値0.4～1.2)に対し、文毎に尤度が最良になるような閾値を選択する。

リサンプリングによって、モデルがよりよいものになったという仮定のもと、そのよくなったモデルに合うように認識データが間引かれるように閾値を選ぶ。

(3)『閾値0.8～1.2で間引きを行ったデータで学習した音響モデル』それぞれを使い、間引きを行った認識データに対し、文毎に尤度が最良になるような閾値を選択する。

(2)において学習データと認識データの閾値を合わせることによって更に認識精度を上げることができるのではないかと考えられる。

(4)『閾値0.8～1.2で間引きを行ったデータで学習した音響モデル』および『6, 8, 9, 10msecで分析したデータで学習した音響モデル』と間引きを行った認識データを用いて文毎に尤度が最良になるような閾値を選択する。

リサンプリングすれば学習データは認識しやすいものになる筈であると考えられる。リサンプリングされたデータに合うモデルを本研究で作成した全てのモデルから選ぶことを行う。

異なる分析条件における音響尤度をそのまま比較することはできない。音響尤度をフレーム数により正規化する。発話毎の対数音響尤度をAM、フレーム数により正規化した対数音響尤度をAM'とすると、次のようにしてAM'を算出することにした。

AM' = AM * (10msecで分析した場合フレーム数) / (VFRによって間引き後のフレーム数)

これまでの結果を表13と表14にまとめる。表13において『尤度選択なし』には、話者毎に最適な条件による結果の平均を載せている。表14において『尤度選択なし』には、4話者の平均が最も良かった結果が載せてある。

表 13：分析周期を変化させた場合

学習データ	評価データ	尤度選択	SUB	INS	DEL	WER	COR	ACC
ベースライン	10msec 固定	なし	22.47	3.69	8.85	35.01	68.68	64.99
	変化	なし	21.69	4.43	7.31	33.43	71.00	66.57
		あり	25.43	9.35	5.01	39.78	69.57	60.22
変化	学習データと同じ	なし	21.66	4.54	7.24	33.44	71.10	66.56
		あり	23.02	7.82	5.00	35.84	71.98	64.16

表 14：VFR における閾値を変化させた場合

学習データ	評価データ(閾値)	尤度選択	SUB	INS	DEL	WER	COR	ACC
ベースライン	変化	なし	22.59	4.81	7.27	34.68	70.13	65.32
		あり	25.43	9.35	5.01	39.78	69.57	60.22
閾値：1.0 固定	1.0 固定	なし	24.02	4.79	8.04	36.85	67.93	63.15
		あり	23.02	7.82	5.00	35.84	71.98	64.16
閾値：変化	変化	なし	22.92	7.03	5.40	35.36	71.67	64.64
		あり	22.27	7.04	5.20	34.50	72.53	65.50
分析周期：変化 閾値：変化	変化	なし	22.59	4.81	7.27	34.68	70.13	65.32
		あり	22.22	6.91	5.28	34.41	72.51	65.59

[1]では、分析周期を 8msec, 9msec, 10msec の中から選んでいたが、本実験では、5msec～12msec と分析周期の幅を広げた。その結果、フレーム数の多い条件での対数音響尤度が有利となった。フレーム数で正規化を試みたが、対数音響尤度の値は負であるので、[1]の手法では更にフレーム数の多い条件が選ばれやすくなった。このようなことになった原因はわからないが、音響尤度を適当に正規化する方法を検討する必要がある。

表 13 において、評価データを分析周期 5msec～12msec で分析し、それぞれをベースライン音響モデルで認識し、尤度による選択を行った場合について、それぞれの分析周期の選ばれた頻度を表 15 に示す。

表 14 において、評価データに VFR(閾値 0.4～1.2)を適用し、それぞれをベースライン音響モデルで認識し、尤度による選択を行った場合について、それぞれの閾値の選ばれた頻度を表 16 に示す。

表 15：ベースラインモデルで認識した場合、尤度基準により各分析周期が選択された頻度

分析周期[msec]	5	6	7	8	9	10	11	12
選ばれた文の数	1999	68	14	12	2	4	0	1

表 16：ベースラインモデルで認識した場合、尤度基準により各閾値が選択された頻度

閾値	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2
選ばれた文の数	1969	87	28	5	2	2	7	3	2

7. まとめ

音声をあらかじめ細かく分析し、隣接するフレーム間のMFCCのユークリッド距離を元に変動の小さなフレームを間引くということにより、発話内における発話速度の補正を行うVFRアルゴリズムを検討した。従来法に比べ、発話速度にロバストであるということが実験により確認できたが、その効果は僅かなものであった。その要因として考えられるのが、音響モデルとして、講演音声で学習したモデルを使ったということである。ある程度多くの講演音声で学習することにより音響モデルがある程度発話速度の変動に対してロバストになっていると考えられる。モデルと合っていない認識データを認識しやすいように変形させるというVFR法の効果は、講演音声と合っているモデルでは殆ど効果が見られないことが分かった。それだけではなく、VFR法は閾値に依存するアルゴリズムであるにも関わらず、現在の段階では認識時に発話文毎に最適な閾値を決めることができない。あらかじめ実験を行って安定した値を選んでいるが、それではあまり効果が得られないと思われる。

更に音響モデルの学習時にもVFRを適用して、更なる認識性能の向上を図ったが、効果は殆どなかった。間引き処理というものが安定しているものではなく、MFCCのユークリッド距離の情報だけでは本当に間引かれるべきフレームが間引かれているのかということが確かではない。閾値に依存するアルゴリズムである以上、それぞれの学習データに合った閾値を選ぶ必要があるが、本研究ではそれができなかった。また、HMMの構造(状態数など)をベースライン音響モデルのままのものを使ったが、VFRに合ったような構造を考える必要があるのではないかと思われる。

また、VFRの高精度化を狙って分析周期を更に細かくして間引きを行ったが、現在の単純なアルゴリズムでは精度を下げてしまう結果になった。より複雑な処理を行う必要があると思われる。また、窓長も可変にすることが望ましいと考えられる。

複数の条件によって認識した結果から尤度を基準として最適な閾値と音響モデルとの組み合わせを自動的に選択する手法についても検討したが、対数音響尤度の正規化がうまく行えず、良い結果が得られなかった。この手法についても改善が必要であると考えられる。

発話速度に対してロバストな音声認識を目指す上で、隣接したフレームのMFCCのユークリッド距離に目をつけることは意味のあることであったが、精度を上げるためには、この情報だけでは不十分であると考えられる。発話速度と関連の強いパラメータについて検討する必要があり、アルゴリズムを作り直すことがより精度を高めるためには不可欠であると言える。音響モデルを高精度にモデリングすることも重要であると思われる。

また、自然発話の認識が困難である要因は発話速度の変動によるものだけではないので、今後は他の要因についても検討していきたいと思う。

参考文献・URL

[1] 奥田浩三, 河原達也, 中村哲, “尤度基準による分析周期・窓長の自動選択手法を用いた発話速度の補正と音響モデルの構築”, 電子情報通信学会論文誌(D-II), J86-D-II No.2,pp204-211, 2003-2.

[2] Q.Zhu and A.Alwan, “On the use of variable frame rate analysis in speech recognition”, Proc.ICASP, Vol.3, pp1783-1786, 2000.

[3] 新宮将久, “講演音声を対象とした自然発話の認識精度の改善に関する検討”, 豊橋技術科学大学特別実験報告書, 2002-12

[4] 籠宮隆之, 菊地英明, 小磯花絵, 前川喜久雄, “大規模話し言葉コーパスにおける発話スタイルの諸相――書き起こしテキストの分析から――”, 日本音響学会研究発表会講演論文集, 2-Q-9, 秋季, 2000.

[5] ATRSPREC(Ver06r06)ユーザマニュアル, http://lab.slt.atr.co.jp/dept1/SPREC/SPREC/doc/manual/manual/manual.html

[6] Julius rev.3.3 ユーザマニュアル, http://julius.sourceforge.jp/3.3/doc/