3R－5

エルゴディック隠れマルコフモデルを用いた単語境界の抽出

天野 真樹　船田 哲男　金寺 登
金沢大学 工学部　石川高専

1 はじめに

実環境での音声認識における問題の1つに音声区間の抽出がある。実環境 (S/N 比 10～20dB) においては 弱い摩擦音や音声の始端・終端にある振幅の小さい有声音などの検出が困難になる。これまで音声区間抽出法としては従来のパワーレベルと零交差数を用いた方法[1] を始め様々な方法が提案されているが、本研究では現在 言語識別などに応用されているエルゴディック HMM[2]を利用して、雑音と音声での音響的特徴の差異から雑音と音声 (単語) の境界の抽出を試み、従来のパワーレベルと零交差数を用いた方法との比較を行なった。

2 実験方法

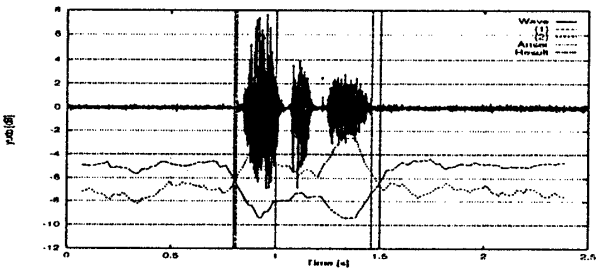

図 1: 雑音中の音声波形と尤度曲線

2.1 HMM の学習

雑音環境中の単語認識の前処理としての音声区間抽出に離散型のエルゴディック隠れマルコフモデル (EHMM) を利用した。まず雑音波形においてフレームピッチを 15[ms] とし、各フレーム (25.6[ms]) ごとに FFT ケプストラム分析を行ない低次の 15 次ケプストラム係数を 8 ビットでベクトル量子化し、シンボルに変換する。4 状態 EHMM の学習と尤度計算には 10 個のシンボル (フレーム) を 1 単位 (1 ブロック) とする。そして 1 ブロックごと重複なしでシフトしながら EHMM を学習する。ここで、雑音波形と音声区間の波形から作成したシンボル列で学習した EHMM をそれぞれ雑音 EHMM 及び音声 EHMM と呼ぶことにする。

2.2 区間抽出

尤度曲線は 1 ブロックを 1 フレームずつシフトしながら同様に尤度を計算して求める。この雑音 EHMM によって得られる尤度曲線を図 1 の (1) に示す。この曲線から音声区間での尤度の低下が見られ、音声の存在を確認することができる。同様に音声 EHMM で尤度曲線を求めると図 1 の (2) となる。この尤度曲線を利用した 2 つの音声区間抽出法を提案する。

[A] 雑音の尤度曲線に一定閾値を設けて閾値以下となる区間を音声区間として抽出。

[B] 音声 EHMM での尤度が雑音 EHMM での尤度曲線より大きくなる区間を音声区間として抽出する。

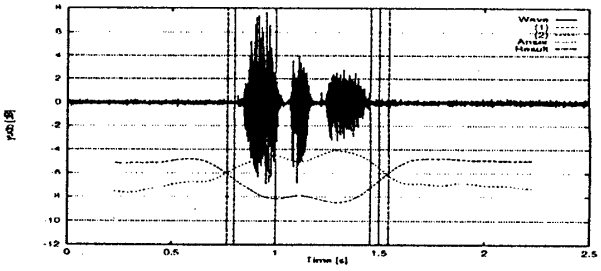

図 2: 尤度曲線の平滑化 (L=21)

また図 2 に示すように連続 L 個の尤度を平均し、尤度曲線を平滑化することにより、学習データ不足を補うことができる。[B] の方法においてこの平滑化を音声・雑音用 EHMM の両方に対して行なった。

3 実験結果と考察

今回実験に用いたデータは駅構内の騒音中での駅名読み上げ音声データで、話者 02.04(女性).08.10(男性) による駅名 01～15 の 15 種類をそれぞれ 2 または 3 回ずつ読み上げた計 164 発声を用いた。コードブック及び EHMM の学習には話者 02.08 によるデータを用い、04.10 によるデータは評価用とした。実験条件をまとめて表 1 に示す。単語境界の正解値は波形の視察で得た値に実際の聴取による補正をしたものを用いた。表中の誤差はフレームピッチ単位 (15[ms]) で表示

Extraction of Word Boundary Detection by Ergodic Hidden Markov Model
Masaki Amano
Faculty of Technology, Kanazawa University

しており、数値は誤差が E フレーム以内である発声数の割合 [%] を示している。

表 1: 実験条件

特徴ベクトル	15 次 FFT ケプストラム係数
ベクトル量子化	8 ビット
HMM	4 状態 EHMM
EHMM 学習データ (話者 02,08)	雑音用: 11300 フレーム 音声用: 3340 フレーム
評価データ (話者 04,10)	話者 04: 30 発声 話者 10: 43 発声

・ 学習内データ (話者 02,08)

	誤差 ≤E	[R]	[A]	[B] L=1	[B] L=3	[B] L=7	[B] L=15
始端	≤1	76.5	50.7	75.3	79.5	75.3	58.9
	≤3	95.3	68.5	94.5	95.9	97.3	94.5
	≤5	97.6	74.0	97.3	98.6	98.6	98.6
	>10	1.2	8.2	2.7	1.4	0.0	0.0
終端	≤1	61.2	64.4	71.2	79.5	79.5	68.5
	≤3	92.9	90.4	98.6	98.6	100	100
	≤5	95.3	93.2	98.6	98.6	100	100
	>10	1.2	0.0	1.4	1.4	0.0	0.0

・ 学習外データ (話者 04)

	誤差 ≤E	[R]	[A]	[B] L=1	[B] L=3	[B] L=7	[B] L=15
始端	≤1	66.7	10.0	40.0	36.7	33.3	26.7
	≤3	88.9	20.0	56.7	53.3	50.0	40.0
	≤5	97.2	20.0	76.7	73.3	70.0	43.3
	>10	0.0	6.7	3.3	0.0	0.0	0.0
終端	≤1	80.6	6.7	50.0	50.0	46.7	36.7
	≤3	97.2	30.0	70.0	66.7	66.7	56.7
	≤5	100	33.3	73.3	73.3	73.3	60.0
	>10	0.0	13.3	3.3	3.3	6.7	3.3

(話者 10)

	誤差 ≤E	[R]	[A]	[B] L=1	[B] L=3	[B] L=7	[B] L=15
始端	≤1	79.1	51.2	67.4	65.1	67.4	60.5
	≤3	97.7	65.1	81.4	81.4	83.7	83.7
	≤5	97.7	74.4	88.4	88.4	93.0	90.7
	>10	0.0	4.7	2.3	2.3	0.0	0.0
終端	≤1	81.4	48.8	72.1	53.5	48.8	55.8
	≤3	100	90.7	93.0	90.7	93.0	90.7
	≤5	100	93.0	95.3	95.3	93.0	95.3
	>10	0.0	0.0	0.0	0.0	0.0	0.0

Rabiner 法[1]のパラメータを一部変更したもの [R] と比較すると、学習内データにおいては EHMM を用いた [B] の方が明らかに良い抽出率を示している。学習外データに関しては話者 10 に対しては [R] と同等であるが、話者 04 においては音声 EHMM の尤度曲線にばらつきが見られ、抽出率は従来法に比べ低い結果となった。これは音声 EHMM の学習に話者 2 名しか用いておらず、学習データ (特に話者数) の不足が原因と考えられる。

また [B] の方法と比較して [A] の方はかなり抽出誤差が大きくなっている。これは、尤度曲線は話者・単語・雑音の種類によって始・終端での変化の割合や雑音部での平均レベルが異なるため、[A] の場合では閾値をそれぞれにおいて適応的に変化する必要があり、また学習不足による雑音部での尤度曲線の劣化が始・終端の誤抽出を招くと考えられる。従って雑音 EHMM のみを用いて音声区間を抽出するには複雑な抽出アルゴリズムと多くの学習データが必要となる。これに対して [B] の場合では、アルゴリズムもより簡単化され、雑音区間での学習不足による曲線の劣化も音声 EHMM による曲線のレベル以上の範囲のものであれば影響がないと考えられる。

尤度曲線の平滑化を施すことにより、10 フレーム以上離れた誤抽出が減少しているのがわかるが、必要以上の平滑化 (15 フレーム平均) では図 2 のように始端・終端付近での尤度が平滑化前の値からと大きく変化し、かえって正解値からの大きなずれを生じると考えられる。

4 まとめ

実環境下における音声認識の前処理としての音声区間抽出においてエルゴディック HMM を利用する方法を提案し、従来のパワーレベルと零交差数を用いる方法と比較した。学習内データでは従来法以上の結果を得ることができたが、学習外データに対しては音声用学習データ不足のため話者による抽出結果のばらつきが見られた。従って学習用データ (特に話者数) を増加して話者に依らない抽出率の改善と S/N 比に対する結果の検討を今後の課題としたい。

なお、本研究を進めるに当たりデータを提供して頂いたオムロン (株) システム総合研究所第 4 研究室に感謝致します。

参考文献

[1] L.R.Rabiner,M.R.Sambur: "An algorithm for Determining the Endpoints of Isolated Utterances", BSTJ, vol. 54, No. 2, pp. 297-315 (1975).

[2] 中川 聖一, 清野 崇, 上田 佳央: "エルゴディック HMM とその状態シーケンスを用いた音声による言語の識別", 電子情報通信学会論文誌 A, vol. J77-A, No. 2, pp. 182-189 (1994 年 2 月).

[3] 松本 弘: "実環境における音声認識の現状と課題", 日本音響学会講演論文集, 1-4-10 (1993 年 3 月).