論文 / 著書情報
Article / Book Information

論題(和文)	音系列のセグメンテーションとその評価法の検討
Title(English)	
著者(和文)	田熊竜太, 古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会 2000年春季講演論文集, Vol. , No. 1-Q-2, pp. 135-136
Citation(English)	, Vol. , No. 1-Q-2, pp. 135-136
発行日 / Pub. date	2000,

1—Q—2 音系列のセグメンテーションとその評価法の検討 *

◎田熊 竜太　古井 貞熙　(東工大)

1. はじめに

本稿では連続的に入力される音系列を, 話者交代や背景雑音の変化を自動的に検出して, セグメンテーションすることについて考える. 放送ニュースの完全自動ディクテーションシステムなどでは, このような技術が不可欠である.[1][2]

本稿で提案する手法は実時間から 1s の遅延で音声認識システムに結果の通知が可能な, オンライン向け高速自動セグメンテーションアルゴリズムである.

また本手法を正しく評価するための, 正解と検出結果の平均時間差に基づく評価方法を提案する.

2. 境界尺度

ある時刻 x(s) における境界尺度を次のように定義する. 区間 $[x-1,x][x,x+1][x-1,x+1]$ から得られる特徴量を用いて GMM(混合ガウス分布モデル) を生成する. 各 GMM 自身の対数尤度を L_h, L_t, L_a として時刻 x の境界尺度 $L(x)$ は以下のように表される :

$$L(x) = L_h + L_t - L_a \tag{1}$$

一般に $L(x) \geq 0$ であり前後区間の音響的特徴が異なれば $L(x)$ は増加する.

3. セグメンテーションの流れ

入力された音系列に対し 1s ごとに境界尺度 $L(x)$ を計算する. ここで $L(x)$ を計算するためには時刻 $x+1$ までの音響データが必要である. これは時刻 x の境界尺度を計算するためには前後それぞれ 1s の音響データを用いる必要があるからである.

次に計算された $L(x)$ を用いて境界を検出する. このとき以下の 3 つの方法を検討する:

$$L(x) > \alpha \text{ならば境界検出} \tag{2}$$

$$L(x) > L(x-1) + \alpha \text{ならば境界検出} \tag{3}$$

$$L(x) > \alpha L(x-1) \text{ ならば境界検出} \tag{4}$$

α を制御することによって境界数を制御することができる. どの式でも α を大きくすると境界数は減少する.

式 (2) は広く使われている方法で, 境界尺度が閾値である定数 α よりも大きければ時刻 x に境界があると判断する. 式 (3) 式 (4) は今回提案する方法で, 定数 α を用いて直前の境界尺度 $L(x-1)$ と比較している.

本アルゴリズムでは音響信号が入力されてから境界判定までの遅延は理論上 1s で, リアルタイム処理に特化した手法であるといえる.

4. 評価法

以下に二つの評価法を示す. ひとつは一般的に用いられる誤検出率 (FAR : False Alarm Ratio) と未検出率 (FRR : False Rejection Ratio) による評価法で, もうひとつは今回提案する正解と検出結果の境界の時間差に基づく評価法である.

4.1 評価法 1

手動で与えた正解は後述のように 100ms 単位で与えられているのに対し, 今回の実験では検出結果は 1s 単位で得られる. そこで誤検出および未検出を以下のように定義する.

誤検出 : 検出した境界の 0.5s 以内に正解の境界がない
未検出 : 正解の境界の 0.5s 以内に検出した境界がない

これらに基づいて誤検出率および未検出率を定義すると

$$\text{誤検出率 (FAR)} = \frac{\text{誤検出境界数}}{\text{検出境界数}} \times 100(\%) \tag{5}$$

$$\text{未検出率 (FRR)} = \frac{\text{未検出境界数}}{\text{正解境界数}} \times 100(\%) \tag{6}$$

となる.

4.2 評価法 2

セグメンテーションを行った結果の境界の集合を $X = \{x_1, x_2, \ldots, x_M\}$, 正解として与えられる境界の集合を $Y = \{y_1, y_2, \ldots, y_N\}$ とする. $M > N$ のときを過分割であるといい, $M < N$ のときを過少分割であるということにする. ここで以下の評価式を与える.

$$t_1 = \frac{1}{M} \sum_{x \in X} \min_{y \in Y} |x - y| \tag{7}$$

$$t_2 = \frac{1}{N} \sum_{y \in Y} \min_{x \in X} |x - y| \tag{8}$$

これは基準となる境界に最も近い対象の境界までの時間差の平均であり, $t_1 = 0$ かつ $t_2 = 0$ のとき $X = Y$ である.

4.3 評価法の比較

過分割の場合, t_1 および FAR の値は大きくなり t_2 および FRR の値は小さくなる. 過少分割の場合, t_2 および FRR の値は大きくなり t_1 および FAR の値は小さくなる. より良いセグメンテーションアルゴリズムではこれら 4 つの値ともに小さい値を示す.

$$T = t_1 + t_2 \tag{9}$$

によりアルゴリズムの総合的な評価が可能になる.

5. 実験

5.1 実験条件

実際に放送された 3 つの NHK のニュース番組を使いセグメンテーションの実験を行った:

P1	1996 年 6 月 1 日放送の「おはよう日本」	40 分
P2	1996 年 6 月 1 日放送の「ニュース 7」	30 分
P3	1996 年 6 月 2 日放送の「ニュース 7」	20 分

各番組とも背景に音楽の流れる部分を含んでいる. 番組 P1 は男女キャスターの発話中心である. 番組 P2 は男性キャスターの発話中心で女性の発話はない. 番組 P3 は男性キャスターの発話中心で女性の発話も含まれる.

これらの番組の音響データを 12kHz でサンプリング,16bit で量子化し, 窓幅 30ms, 分析周期 10ms のハミング窓を用いて 16 次元のケプストラムを生成し, 特徴量とした. 得られたケプストラムを用いて混合数 8 の GMM を作成した.

境界の正解は実験者一人が全放送を聞いて与えた. 境界は話者交代・背景音の変化・有音/無音の変化に対し 100ms 単位で与えた.

---

*A Study on Sound Sequence Segmentation and Evaluation Methods
By Ryuta Taguma and Sadaoki Furui (Tokyo Institute of Technology)

5.2 結果

式 (2)～(4) の α の値を変化させたときの各番組の実験結果を図 1,2,3 に示す．図 1,2,3 では右下に行くほど境界数は増加している．　式 (2)～(4) を評価するために，

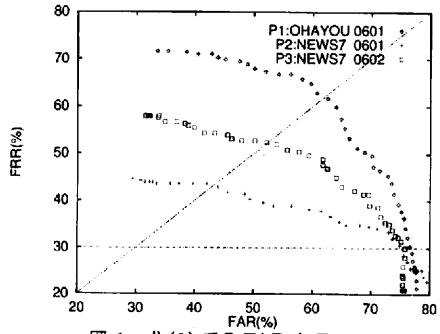

図 1. 式 (2) での FAR と FRR

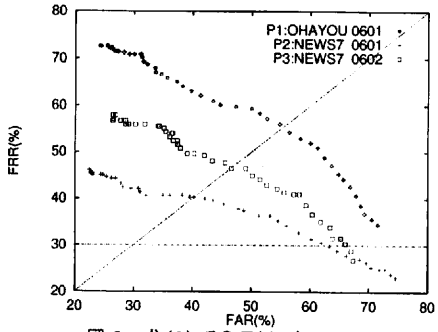

図 2. 式 (3) での FAR と FRR

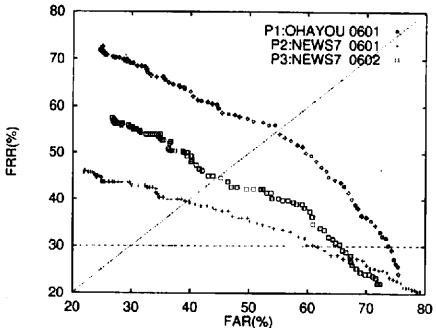

図 3. 式 (4) での FAR と FRR

FRR= 30% での番組 P2 についての詳しい結果を表 1 に示す．T の値を比較すると，各番組とも式 (4) の結果が最も良かった．FAR の値も (4) の結果が最も良くなっている．

以降の結果は式 (4) を用いた場合のみを示す．

等誤り率 (EE)，つまり FAR=FRR になる α での結果を表 2 に示す．このとき正解と結果の境界数は非常に近い値になっている．

FAR が 30% になる α での結果を表 3 に示す．このとき結果の境界数が正解の境界数よりもかなり大きくなっている．正しい境界を誤って落とさないように過分割の状態になっているからである．後処理として，さらに詳細な分析を行ったり，クラスタリングを行うことでこの過分割状態は解消することが可能である．

6. 考察

境界検出に用いた式 (2)～(4) のうち，固定閾値を用いた式 (2) は高い誤り率を示した．これに対し直前の境界尺度との比較による式 (3)，式 (4) は比較的低い誤り率を示していた．

式 (4) において結果として得られる境界数を，正解の境界数に近づけるためには $\alpha \simeq 1.30$ にすれば良いことがわかった．これは異なる番組間で共通であった．また，未検出率を 30% に保つためには $\alpha \simeq 1.05$ とすれば良いことがわかった．この値も番組には依存していなかった．

t_1, t_2 および T による式 (2)～(4) の評価結果や，番組の違いによる手法の比較結果は，誤検出率/未検出率による評価結果とほぼ等しいことがわかった．

7. まとめ

本稿ではリアルタイム処理に向いた音系列の高速セグメンテーション法を提案し，その評価を行った．

従来使われてきた誤検出率/未検出率 (FAR / FRR) による評価と本稿で提案する時間差に基づく評価法により評価した．閾値 α の値を制御することにより誤り率と境界数を制御でき，タスクや後処理に応じて選択することができる．

本稿の結果は総じて誤り率が依然として高いので，GMM による音の状態に基づくクラスタリング等の後処理を導入することで，誤り率の削減をはかる必要がある．また今回の実験では 1s 単位でセグメンテーションを行ったが，実用化のためには 100ms 程度の単位で，セグメンテーションを誤り率を増加させることなく実現することが望ましく，今後さらに改良する必要がある．

表 1. FRR=30%での P2:ニュース 7 06/01 の結果

式	境界数		評価法 1	評価法 2(s)		
	正解	結果	FAR(%)	t_1	t_2	T
(2)	278	929	75.5	2.83	0.78	3.61
(3)	254	609	62.6	2.32	0.68	3.00
(4)	278	601	**60.9**	**2.24**	**0.65**	**2.89**

表 2. 式 (4) で FAR=FRR になるときの結果

番組	α	境界数		評価法 1	評価法 2(s)		
		正解	結果	EE(%)	t_1	t_2	T
P1	1.21	514	585	54.7	2.09	1.48	3.57
P2	1.39	278	321	39.6	1.31	1.28	2.59
P3	1.31	254	289	44.5	1.33	1.43	2.76

表 3. FRR=30(%) になるときの結果

番組	α	境界数		評価法 1	評価法 2(s)		
		正解	結果	FAR(%)	t_1	t_2	T
P1	0.96	514	1507	73.7	3.71	0.47	4.18
P2	1.16	278	601	60.9	2.32	0.53	2.85
P3	1.04	254	617	65.5	2.24	0.65	2.89

参考文献

[1] Daben Liu and Francis Kubara, "Fast Speaker Change Detection for Broadcast News Transcription And Indexing", Proc.Eurospeech.1031-1034, 1999.

[2] Alain Tritschler and Ramesh Gopinath, "Improved Speaker Segmentation and Segments Clustering Using the Baysian Information Criterion", ProcEurospeech.679-682, 1999.