3次キュムラントバイスペクトラム特徴と Real AdaBoost による音声区間検出*

○松田博義, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

雑音下において音声認識を行う際, 非音声の誤認識による認識精度の低下を防ぐため, 音声区間検出 (VAD: Voice Activity Detection) を行う必要がある. 本稿では音声区間検出の特徴抽出法として, 高次統計量として知られている 3 次キュムラント (3rd order cumulant) のバイスペクトラム [1] を PCA により次元圧縮した音声特徴の使用, 及び, その特徴と MFCC との統合手法 [2] について述べる. さらに, 識別手法に Real AdaBoost [3] を用い, GMM との比較を行なう. 評価には, CENSREC-1-C に収録されている, 実環境にて録音されたデータを用い, 提案手法の有効性を示す.

2 3次キュムラントバイスペクトラムによる音声特徴

確率変数 s の k 次モーメント M_k は

$$M_k = E[s^k] = \int_{-\infty}^{\infty} s^k p(s) ds \quad (1)$$

で定義される. このとき 3 次のキュムラントは,

$$\kappa_3 = M_3 - 3M_2M_1 + 2M_1^3 \quad (2)$$

と表される. なお確率変数 s の平均値がゼロである場合, すなわち 1 次のモーメント $M_1 = 0$ である場合には, $\kappa_3 = M_3$ となる.

ここで, 実際に 3 次キュムラントを用いて音声特徴を得る手法について述べる. $\{s(t)\}$ を音声信号とする. 与えられた信号 $\{s(t)\}$ を長さ N に切り分けることで以下のような信号系列 $y_k(t)$ を得る.

$$y_k(t) = s(t + k \cdot \tau + T) - M_{1k} \quad (3)$$

k は信号の切り出し位置に対応しており $-K \leq k \leq K$ である. τ はシフト幅である. データはできる限りフレーム間でオーバーラップするよう τ の値は小さくする. T は現在処理している音声信号の初期位置を示している. これにより $\{s(t)\}$ から, $y_k(t)$ として新しく $2 \cdot K + 1$ 個のベクトルセットを得る. ここで, すべての信号は平均 0 としておき, 定常であると仮定する.

通常, 平均が 0 の場合の 3 次キュムラントの計算式は $\kappa_3 = M_3 = E[y_0 y_0 y_0]$ であるが, 新たに k, l という変数を導入することにより, 現在処理している前後のフレームとの相関を表すよう拡張する.

$$\begin{aligned} C_{y_k y_l} &= E[y_0 y_k y_l] \\ &= \frac{1}{N} \sum_{i=0}^{N-1} y_0(t_i) y_k(t_i) y_l(t_i) \end{aligned} \quad (4)$$

ここで, $-K \leq k, l \leq K$ である.

得られた 3 次キュムラント行列に対し, データ解析のため 2 次元離散フーリエ変換を行なう. それにより得られたバイスペクトラムはそのままでは非常

にデータ量が多いので, PCA (主成分分析) を行ない主要な情報だけを取り出すことにより次元圧縮を行なう. こうして得られたベクトルをもって, 3 次キュムラントのバイスペクトラムによる音声特徴とする.

3 次キュムラントのバイスペクトラムによって得られる音声特徴はフレーム間での情報であり, MFCC によって得られる音声特徴はフレーム内での情報である. これらは相互に補完しあっていると考えられるので, 統合することを考える. ある時刻 t のフレームから得られた n 次元のキュムラント特徴と, m 次元の MFCC とを合わせ, 新たに $n + m$ 次元の音声特徴量とすることにより, キュムラントと MFCC による統合特徴量とした.

3 Real AdaBoost による音声非音声の識別

AdaBoost は, 単純な識別機を複数組み合わせることによって精度の高い識別器を構成する Boosting 法の中でも顕著な性能を示す手法である. 今回は AdaBoost を発展させた, Real AdaBoost [3][4] という手法を用いて音声非音声の識別を行った.

学習データ数を n , 繰り返し回数を M とする. これらの値はあらかじめ決定しておく必要がある. 学習データを $x_i (i = 1, \dots, n)$, 各データの重みを $w_i (i = 1, \dots, n)$ とする. 各データ x_i にはあらかじめラベルとして $y \in \{-1, +1\}$ を与えておく. すなわちデータ x_i が音声であれば $y = +1$, データ x_i が非音声であれば $y = -1$ とする.

- 各データに対する重みを $w_{1i} = \frac{1}{n}$ で初期化する.
- $m = 1, \dots, M$ まで, 以下を繰り返す.
 - w_{mi} を重みとして, x_i から重複を許して n 個, リサンプリングしたものを x'_i とする.
 - x'_i に対して弱識別器 $f_m(x)$ を構成する. ここでは, 弱識別器 $f_m(x)$ に CART を用いた.
 - 得られた弱識別器 $f_m(x)$ を用いて $c_m(x_i)$ を得る.

$$c_m(x_i) = \frac{1}{2} \log \left(\frac{f_m(x_i)}{1 - f_m(x_i)} \right) \quad (5)$$

- 各データ x_i の重みを $w_{(m+1)i}$ に更新する.

$$w_{(m+1)i} = \frac{w_{mi} e^{-y_i c_m(x_i)}}{\sum_{r=1}^n w_{mr} e^{-y_r c_m(x_r)}} \quad (6)$$

- 弱識別器の線形結合により, 強識別器 $F(x)$ を得る.

$$F(x) = \sum_{m=1}^M c_m(x) \quad (7)$$

AdaBoost の重要な性質として, 誤ったデータに対するサンプリング重みを増し, 次回以降の学習でそ

*Voice Activity Detection with 3rd Order Cumulant and Real AdaBoost. by Hiroyoshi MATSUDA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Graduated School of Science and Technology)

Table 1 VAD results on “restaurant, high”

	Baseline	GMM			AdaBoost		
		MFCC	Cumulant	MFCC+Cumulant	MFCC	Cumulant	MFCC+Cumulant
FAR	3.7%	11.8%	24.3%	12.0%	18.1%	11.9%	8.9%
FRR	32.1%	10.3%	6.7%	34.0%	5.9%	27.2%	9.0%

Table 2 VAD results on “restaurant, low”

	Baseline	GMM			AdaBoost		
		MFCC	Cumulant	MFCC+Cumulant	MFCC	Cumulant	MFCC+Cumulant
FAR	14.0%	36.7%	44.5%	34.8%	26.6%	45.8%	31.3%
FRR	48.1%	12.3%	12.9%	13.9%	17.4%	7.5%	10.6%

Table 3 VAD results on “street, high”

	Baseline	GMM			AdaBoost		
		MFCC	Cumulant	MFCC+Cumulant	MFCC	Cumulant	MFCC+Cumulant
FAR	54.3%	3.5%	0.5%	3.5%	5.3%	0.7%	0.6%
FRR	14.0%	12.6%	28.2%	12.6%	8.9%	23.6%	12.4%

Table 4 VAD results on “street, low”

	Baseline	GMM			AdaBoost		
		MFCC	Cumulant	MFCC+Cumulant	MFCC	Cumulant	MFCC+Cumulant
FAR	40.7%	3.0%	7.6%	1.3%	4.7%	15.1%	4.1%
FRR	32.9%	24.2%	23.1%	19.9%	21.2%	12.5%	14.5%

Table 5 VAD results on average

	Baseline	GMM			AdaBoost		
		MFCC	Cumulant	MFCC+Cumulant	MFCC	Cumulant	MFCC+Cumulant
FAR	28.2%	13.7%	18.6%	13.2%	11.7%	17.4%	11.9%
FRR	31.8%	13.2%	18.9%	12.7%	13.7%	17.6%	11.0%

これらのデータを重点的に学習する，ということがあげられる。それにより，前段の弱識別器が誤識別を起こしてしまったデータに対しても，後段の弱識別器が正しく識別するため，最終的に正しい識別結果を得ることができる。

4 CENSREC-1-C による性能評価

4.1 CENSREC-1-C

提案手法の評価は，VAD の評価用に設計されたデータベース CENSREC-1-C[5] を用いて行なった。CESREC-1-C には人工的に作成したシミュレーションデータと，実環境にて収録されたデータの 2 種類が含まれている。ここでは，実際の環境での VAD 性能を評価するため，実環境にて録音された実データを用いて実験を行なった。実データは高速道路付近 (STREET) と学生食堂 (RESTAURANT) の 2 種類の環境で，それぞれ高 SNR，低 SNR の状況で収録されたデータセットがある。

4.2 実験条件

GMM, AdaBoost の学習には，非音声として，CENSREC-1-C に含まれている実データの内，評価対象とされていない男声のデータの非音声部を切り出したものを用いた。また音声として，AURORA-2J に含まれているクリーン音声の学習用データに，切り出した非音声データを重畳させたものを用いた。

キムラントは式 (3) において， N (窓幅)32 ms, τ (シフト幅)1 ms, K は 30 とした。MFCC は窓幅 32 ms, シフト幅 8 ms, Δ MFCC は ± 5 フレームから計算した。

AdaBoost は， $M = 1000$ ，男声・女声を音声，道路付近・食堂の雑音をあわせて非音声として学習を行なった。GMM は，音声は 64 混合，非音声は 32 混合で実験を行なった。尚，ここでは音声の GMM として，男声モデルと女声モデルの二通りを，非音声の GMM として，道路付近の雑音モデル，食堂の雑音モデルの二通りをそれぞれ作成した。検定の際は，各モデルで尤度を計算し，それぞれ値が高いものを音声クラス，非音声クラスの尤度として採用し，対数尤度比により音声・非音声の識別を行なった。

実験ではフレーム単位での評価を行なった。評価尺度は次式の FRR (False Rejection Rate) と FAR (False Acceptance Rate) である。

$$FRR = N_{FR}/N_s \times 100[\%] \quad (8)$$

$$FAR = N_{FA}/N_{ns} \times 100[\%] \quad (9)$$

N_s は音声フレーム数， N_{FR} は音声を非音声と検出したフレーム数， N_{ns} は非音声フレーム数， N_{FA} は非音声を音声と検出したフレーム数である。

4.3 実験結果

表 1~4 にそれぞれの環境での実験結果を，表 5 に各環境での結果を平均したものを示す。ここで，ベースラインとは CENSREC-1-C に付属の実験結果で，音響信号のパワーを特徴量とした手法である。実験の際の閾値は，各データセット間で共通である。

表 5 より，GMM, AdaBoost による手法が，ベースラインを上回っていることが確認できる。特に，AdaBoost と，MFCC, キムラントの統合特徴を組み合わせた結果が $FAR = 11.9\%$, $FRR = 11.0\%$ と最も良かった。AdaBoost の実験結果が GMM を上回った理由は，GMM は学習においてすべてのデータを均等に用いるのに比べ，AdaBoost は識別が難しいデータに対して重点的な学習を行なうためである。さらに，GMM と比べ，AdaBoost は音声・非音声の分離に特化した識別器を構成する，すなわち，音声・非音声の分離に必要な情報だけを見て，より正確な規則を構築しているためである。

参考文献

- [1] J. C. Segura *et al.*, Lecture Notes in Artificial Intelligence, pp. 174-185, No. 817, 2005.
- [2] 松田他，信学技報，SP2006-85-114, pp. 89-94, 2006.
- [3] J. Friedman, *et al.*, The Annals of Statistics, pp. 337-407, Vol. 28, No. 2, 2000.
- [4] <http://research.graphicon.ru/general-projects/about-us.html>.
- [5] 北岡他，音講論，1-P-27, pp. 103-104, Sept. 2006.