3次キュムラントのバイスペクトラムとPCA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手法として，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][2][3] 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 (Mel Frequency Cepstrum Coefficient) との初期統合手法について述べる．さらに，実データを用いた実験により，提案手法の有効性を検証する．

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

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

で定義される．このとき 3 次のキュムラントは，

$$\kappa_3 = M_3 - 3M_2M_1 + 2M_1^3 \tag{2}$$

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．$\sigma^2 = M_2$(分散) とすると，κ_3/σ^3 は歪度と呼ばれ，分布の非対称性を表す．例えば，正規分布は，3 次以上のキュムラントはすべてゼロとなっている．一般的に，雑音は音声に比べ正規分布から発生した乱数に近いという性質を持っている．そのため，3 次以上のキュムラントは，音声であれば大きな値となり，雑音であれば小さな値になると考えられる．

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．

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

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 $\kappa_3 = M_3 = E[y_0y_0y_0]$ 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間での相関を表すよう拡張する．

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

ここで，$-K \leq k, l \leq K$ である．今処理しているフレームと k 及び l フレーム離れたフレームとの 3 次キュムラントを計算することにより，雑音であれば値が小さな、音声であれば大きな値が得られる．式 (4) より，式 (2) だけでは得られなかった，前後のフレームにどのような音声が分布しているかという情報を得ることが出来るようになる．$k = l = 0$ の場合，すなわち $C_{y_0y_0} = E[y_0y_0y_0]$ は，式 (2) と同義である．

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，本稿では，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．こうして得られた数次元のベクトルをもって，3 次キュムラントのバイスペクトルによる音声特徴とする．

3 MFCC との初期統合

3 次キュムラントのバイスペクトル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は，各フレーム内での詳細な音声情報である．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．ここでは，各フレームから得られた n 次元キュムラント特徴，x_{ct} と，m 次元 MFCC，x_{mt} とをあわせ，あらたに n+m 次元の音声特徴とした．これを用い，音声非音声についてそれぞれ GMM (Gaussian Mixture Model) を作成する．

MFCC，及びキュムラント特徴はそれぞれをストリームに分け適切な重みを与えた．マルチストリーム GMM では，音響特徴 x_t の観測確率は，対数尤度 $b(x_t)$ を用いて以下のように表される．

$$b(x_t) = \lambda_m b_m(x_{mt}) + \lambda_c b_c(x_{ct}) \tag{5}$$

ただし，t は時刻，$b_m(x_{mt})$，$b_c(x_{ct})$ はそれぞれ音響特徴量 x_{mt}，x_{ct} に対する対数尤度，λ_m，λ_c は GMM における MFCC，キュムラント特徴に対するストリーム重みである．

4 評価実験

4.1 データ概要

音声の学習には，ASJ 男性話者 8 名，計 1200 文，および ASJ 女性話者 8 名，計 1200 文にそれぞれ非音声の学習に用いた車内雑音を重畳させたものを用いた．非音声の学習には，空調が弱，及び中に入った状態で車内にて収録された走行音計 4 分弱を用いた．

評価に用いたデータは，アイドリング時，高速道路走行時，高速道路走行時に音楽をかけている状況を想定した 3 つの発話データセットである．アイドリング時，高速道路走行時の発話データは実環境にて録音されたもの，音楽環境を想定した発話データは高速道路走行時の発話データに音楽を重畳させることにより作成した．それぞれ男性 4 名，女性 4 名，各話者 100 発話で計 800 発話からなる．発話内容は日本各地の地名である．SN 比はアイドリング時で 10〜25 dB，平均約 17 dB，高速道路走行時で 0〜8 dB，平均約 5.5 dB，音楽環境で-1〜6 dB，平均約 3.5 dB である．

すべてのデータは 12,000 Hz で，低域に集中する車内雑音を取り除くため，カットオフ周波数 200 Hz をもつハイパスフィルタを適用している．

---

*Voice activity detection by 3rd order cumulant bispectrum and PCA. by Hiroyoshi MATSUDA, Tetsuya TAKIGUCHI, Yasuo ARIKI(Graduated School of Science and Technology)

4.2 比較対象

比較は，音声特徴として，

1. MFCC, 16 次元
2. MFCC + Δ, 32 次元
3. Cumulant, 32 次元
4. Cumulant + MFCC, 48 次元
5. Cumulant + MFCC + Δ, 64 次元

の 5 通りを用いて行なった．キュムラントは式 (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 は 5 フレームから計算した．

実験では，それぞれの特徴量から作成された音声の GMM と非音声の GMM から計算された尤度を用いて対数尤度比を取り，閾値判定することにより区間検出を行なった．評価の方法は，検出された音声区間の始端終端のなかに，あらかじめ人手によってラベル付けされた始端終端が両方とも含まれていれば正解とした．検出された区間がラベル付けされた始端終端の内側にある場合や，片側しか含まれていなければ不正解とした．検出された区間のうち，ラベルと関係の無い区間であれば，それを湧き出しとした．

4.3 PCA の圧縮次元数の決定

3 次キュムラントバイスペクトラムは一辺が 2·30+1 (実際には FFT を行なっているので，ここでは 64) となっており，全体では 64^2 で，4096 次元ある．バイスペクトラムの対象性を考え，全体の 1/4，すなわち 1024 次元を使って PCA により次元圧縮する際，何次元までが有意な情報を持っているかを実験により調べた．表 1 より，PCA により次元圧縮された 3 次キュムラントバイスペクトルは，32 次元までが有意な情報を持っていると考えられる．

Table 1　PCA の次元数による検出率の変化

次元数	Recall	Precision
8	58.63 %	59.14 %
16	62.00 %	62.23 %
32	66.00 %	70.59 %
64	66.13 %	69.06 %

4.4 ストリーム重みの決定

3 次キュムラントバイスペクトル音声特徴を MFCC と統合する際，それぞれをストリームに分け，実験により最適な重みを与えた．表 2 に，(5) Cumulant + MFCC + Δ，高速道路走行時の実験データにおけるキュムラントに対するストリーム重みを変更した際の実験結果を示す．表 2 よりストリーム重みは 0.25 が最適である事が分かる．

Table 2　ストリーム重みによる検出率の変化

ストリーム重み	Recall	Precision
1.0	90.63 %	91.31 %
0.5	95.25 %	96.95 %
0.25	96.00 %	98.08 %
0.125	95.25 %	97.94 %
0	94.63 %	94.51 %

4.5 実験結果

表 3，表 4，及び表 5 にそれぞれアイドリング時，高速道路走行時，音楽環境下における実験結果を示す．アイドリング時のような SN 比が比較的よい環境下では MFCC，キュムラント特徴ともに識別率に大きな差は見られない．高速道路走行時において，MFCC+Δ に比べ，キュムラント特徴単体では識別率が大きく落ちている．これは 3 次キュムラントバイスペクトルは，例えば車が何かを踏んだ音など，突発的な雑音を音声として誤検出することが多かったためである．しかし，Cumulant+MFCC+Δ とすることにより，MFCC+Δ を上回った．音楽の環境下では，MFCC，キュムラントは共に大きく検出率を落としたが，Cumulant+MFCC+Δ において，最も良い検出率を得ている．これらの結果は，フォルマントなど MFCC によって得られるフレーム内での音声の特徴，そして 3 次キュムラントバイスペクトルによって得られるフレーム間での音声の特徴が，互いに補完しあったためと考えられる．

Table 3　アイドリング時の実験結果

	Recall	Precision
MFCC	96.88 %	97.12 %
MFCC+Δ	98.38 %	98.50 %
Cumulant	97.13 %	99.87 %
Cumulant+MFCC	98.00 %	99.56 %
Cumulant+MFCC+Δ	98.25 %	100 %

Table 4　高速道路走行時の実験結果

	Recall	Precision
MFCC	92.25 %	92.95 %
MFCC+Δ	94.63 %	94.51 %
Cumulant	66.00 %	70.59 %
Cumulant+MFCC	93.25 %	94.91 %
Cumulant+MFCC+Δ	96.00 %	98.08 %

Table 5　音楽環境下での実験結果

	Recall	Precision
MFCC	48.25 %	50.79 %
MFCC+Δ	54.38 %	60.67 %
Cumulant	48.88 %	48.09 %
Cumulant+MFCC	62.88 %	65.84 %
Cumulant+MFCC+Δ	63.50 %	67.46 %

5　まとめ

本研究では，車室内での音声と非音声の識別による音声区間検出に関して，3 次キュムラントバイスペクトルを用いた手法，及び従来手法である MFCC との統合手法を提案した．キュムラント特徴単体では，MFCC を超える識別結果を得ることはできなかった．特に、高速道路走行時等の比較的 SN 比の悪い環境下においては，識別結果は MFCC を大きく下回った．しかし，MFCC とキュムラント特徴を統合することにより，MFCC がもつフレーム内での特徴，キュムラントがもつフレーム間での特徴が相互に補完しあい，MFCC を上回る識別結果を得ることができた．

謝辞　今回の実験に当たり，学習に用いた車内雑音，及び実験に用いた発話データは，富士通テンにより収録されたデータを使用させていただきました．

参考文献

[1] J. M. Gorriz, C. G. Puntonet, J. Ramirez, and J. C. Segura: “Bispectrum Estimators for Voice Activity Detection and Speech Recognition,” Lecture Notes in Artificial Inteligence, pp. 174-185, No. 817, 2005.

[2] J. M. Gorriz, J. Ramirez, J. C. Segura and S. Hornillo: “Voice Activity Detection Using Higher Order Statics,” Lecture Notes in Computer Science, pp. 837 - 844, Vol.3512/2005.

[3] J. M. Gorriz, J. Ramirez, J. C. Segura, and C. G. Puntonet: “Improved MO-LRT VAD based on bispectra Gaussian model,” IEE Electronic Letters, Volume 41, Issue 15, pp. 877-879, July, 2005.

[4] Norbert Binder, Konstantin Markov, Rainer Gruhn, Satoshi Nakamura: “SPEECH NON-SPEECH SEPARATION WITH GMMS,” 日本音響学会講演論文集, pp. 141-142, 2001 年 10 月.