社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

3次キュムラントのBispectrumとMFCCの統合による音声区間検出の検討

松田　博義†　　滝口　哲也†　　有木　康雄†

† 神戸大学自然科学研究科　〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: †matsuda@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

あらまし　雑音下において音声認識を行う際，音声非音声の判定により音声区間検出 (VAD: Voice Activity Detection) を行う必要がある．静かな状況ではゼロクロッシング法などにより区間検出を行うことが可能である．しかし雑音下，特に音声の大部分が雑音に埋もれてしまっているような状況においては，従来の手法では十分な結果を得ることができない．本稿では，雑音に対するロバストな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 Bispectrum を用いて，PCA による次元圧縮後，MFCC (Mel Frequency Cepstrum Coefficient) との初期統合を行う方法を提案する．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．

キーワード　キュムラント，音声区間検出，VAD，高次統計量

Voice Activity Detection with Integrated Bispectrum of 3rd Order Cumulant and MFCC

Hiroyoshi MATSUDA†, Tetsuya TAKIGUCHI†, and Yasuo ARIKI†

† Graduated School of Science and Technology, Kobe University　Rokkodaicho1-1, Nada-ku, Kobe, Hyogo 657-8501 Japan
E-mail: †matsuda@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract　The separation of speech and non-speech events is an important problem for speech recognition. In clean conditions, energy or zero-crossing features work well. However, a traditional voice activity detection (VAD) is not robust to noisy conditions, where speech signal is seriously contaminated by noise. A robust VAD algorithm based on the determination of the speech/non-speech bispectra of the third order auto-cumulants has been proposed. In this paper, we investigate the effectiveness of the integration between MFCC and the bispectra of the third order auto-cumulants. Experimental results show the proposed algorithm effective.

Key words　cumulant, voice activity detection, VAD, higher order statistics

1. まえがき

近年，音声認識技術は飛躍的な向上を遂げてきている．それに伴い，音声認識技術を実環境で生かすことが期待されている．しかし，イベント会場，車内，街中などの雑音が大きい環境では，非音声 (雑音) を音声として認識してしまい，それにより誤動作を起こしてしまうことも少なくない．そのため，音声と非音声を識別して，非音声区間では，認識をしないようにさせることが必要となってくる．このように音声非音声を識別し，音声区間を検出することを VAD (Voice Activity Detection) といい，さまざまな研究がなされている．中でも音声特徴抽出に MFCC (Mel Frequency Cepstrum Coefficient)，その識別に GMM (Gaussian Mixture Model) を用いたもの [1] などは優れた効果を挙げている．しかし雑音が強くなってくるにつれ，音声と非音声の識別が難しくなり，その結果，非音声区間を音声と誤認識してしまうことや，音声区間が途中で切れてしまい，その後の音声認識にかけられなくなってしまうことがある．

本稿では，音声に非常に強い雑音が重畳している様な環境においても，音声と非音声を精度良く分離できるような音声特徴の定式化を目的とする．

一般に音声は複数のフレーム間で，相関性のある波形をもつ．しかし従来手法である MFCC は，音声をフレームごとに独立して処理しているため複数のフレーム間 (約 100 ms) にわたる音声の特徴を得ることはできない．そこで，統計量であるキュムラントを用いることにより，複数のフレーム間での相関を計算し，それに対し 2 次元フーリエ変換を行なうことによっ

て MFCC では得られないフレーム間での相関を表した音声特徴を得ることを考える．こうして得られたフレームでの特徴を反映している音声特徴と MFCC を統合することにより，識別精度の向上を試みる．

2. 3 次キュムラントによる音声特徴

2.1 キュムラント (累積数) [2]

確率変数 x の k 次モーメント M_k 及びモーメント母関数 $G(\xi)$ は

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

$$G(\xi) = E[x^{x\xi}] = \int_{-\infty}^{\infty} x^{x\xi} p(x) dx \tag{2}$$

で定義される．$G(\xi)$ を $x\xi = 0$ においてマクローリン展開すると

$$e^{x\xi} = 1 + x\xi + \frac{1}{2!}(x\xi)^2 + \frac{1}{3!}(x\xi)^3 + \cdots \tag{3}$$

であるから，モーメント関数は次のように展開できる．

$$G(\xi) = 1 + \sum_{k=1}^{\infty} \frac{1}{k!} M_k \xi^k \tag{4}$$

モーメント母関数を ξ で n 階微分すると，

$$\frac{d^n G(\xi)}{d\xi^n} = \frac{d^n}{d\xi^n} \int_{-\infty}^{\infty} e^{x\xi} p(x) dx$$

$$= \int_{-\infty}^{\infty} \frac{d^n}{d\xi^n} e^{x\xi} p(x) dx = \int_{-\infty}^{\infty} x^n e^{x\xi} p(x) dx \tag{5}$$

であるから，$\xi = 0$ において

$$\frac{d^n G(\xi)}{d\xi^n} \mid_{\xi=0} = \int_{-\infty}^{\infty} x^n p(x) dx = M_n \tag{6}$$

となる．すなわち，$\xi = 0$ におけるモーメント母関数の n 階微分係数は n 次モーメントに等しい．

いま，モーメント母関数の対数をとり，

$$c(\xi) = \log G(\xi) \tag{7}$$

とおくと，確率変数 x の n 次のキュムラント κ_n は

$$\kappa_n = \frac{d^n c(\xi)}{d\xi^n} \mid_{\xi=0} \tag{8}$$

で定義される．このため，$c(\xi)$ はキュムラント母関数と呼ばれる．たとえば 1 次キュムラントは

$$\begin{aligned} \kappa_1 &= \frac{dc(\xi)}{d\xi} \mid_{\xi=0} \\ &= \frac{d}{d\xi} \log G(\xi) \mid_{\xi=0} \\ &= \frac{1}{G(\xi)} \frac{dG(\xi)}{d\xi} \mid_{\xi=0} = M_1 \end{aligned} \tag{9}$$

となり，同様に 2 次キュムラントは

$$\begin{aligned} \kappa_2 &= \frac{d^2 c(\xi)}{d\xi^2} \mid_{\xi=0} \\ &= \frac{d}{d\xi} \frac{1}{G(\xi)} \frac{dG(\xi)}{d\xi} \mid_{\xi=0} \\ &= M_2 - M_1^2 \end{aligned} \tag{10}$$

と計算される．さらに高次のキュムラントについては

$$\kappa_3 = M_3 - 3M_2 M_1 + 2M_1^3 \tag{11}$$

$$\kappa_4 = M_4 - 4M_3 M_1 - 3M_2^2 + 12M_2 M_1^2 - 6M_1^4 \tag{12}$$

となる．なお確率変数 x の平均値がゼロである場合，すなわち $M_1 = 0$ である場合には，

$$\kappa_1 = 0 \tag{13}$$

$$\kappa_2 = M_2 \tag{14}$$

$$\kappa_3 = M_3 \tag{15}$$

$$\kappa_4 = M_4 - 3M_2^2 \tag{16}$$

となる．このとき，$\kappa_2 = M_2$ は分散であり，分布のばらつきを表す．$\sigma^2 = M_2$ とすると，$\kappa_3/\sigma^3 = M_3/\sigma^3$ は歪度と呼ばれ，分布の非対称性を表す．また $\kappa_4/\sigma^4 = M_4/\sigma^4 - 3$ は尖度と呼ばれる．4 次中心モーメントは値の一部に，ヒストグラムの形状の中央とかけ離れたものがあれば大きくなり，分布が固まっていれば小さくなる．したがって，その値はヒストグラムの形状が中央がとがっていれば大きくなり，中央が平らであれば小さくなる．正規分布については，3 次以上のキュムラントはすべてゼロとなるため，それらを正規分布からのずれを表す指標として用いることができる．

2.2 3 次キュムラントの bi-spectra [3] [4] [5]

以上より，正規分布から発生した乱数の 3 次以上のキュムラントは全てゼロである．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．そのため 3 次以上のキュムラントについては，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．

すなわち，3 次以上のキュムラントには音声と雑音を区別する能力が存在する．そこで計算コストも考慮に入れ，音声特徴抽出に 3 次のキュムラントを用いることを考える．$\{x(t)\}$ を音声信号とする．与えられた信号 $\{x(t)\}$ 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．

$$y_k(t) = x(t + k \cdot \tau + T) \quad (k = 0, \pm 1, \pm 2, \cdots \pm M) \tag{17}$$

k は信号の切り出し位置に対応しており，サンプル間での前後の遅延数を表している．τ はシフト幅である．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 $\{x(t)\}$ から，$y_k(t)$ として新しく $2 \cdot M + 1$ 個のベクトルセットを得る．

ここで，音声区間検出を行なうため，2 つの仮説をたてる．

$$H_0 = \begin{pmatrix} y_0 = n_0 \\ y_{\pm 1} = n_{\pm 1} \\ \vdots \\ y_{\pm M} = n_{\pm M} \end{pmatrix} \tag{18}$$

$$H_1 = \begin{pmatrix} y_0 = s_0 + n_0 \\ y_{\pm 1} = s_{\pm 1} + n_{\pm 1} \\ \vdots \\ y_{\pm M} = s_{\pm M} + n_{\pm M} \end{pmatrix} \tag{19}$$

s_k, n_k はそれぞれ，音声，非音声の信号である．すなわち H_0 は非音声，H_1 は非音声の上に重畳した音声である．すべての信号は定常で，平均 0 であると仮定しておく．

ここで 3 次キュムラントの式を，次のように定義する．

$$C_{y_k y_l} = E[y_0 y_k y_l] \tag{20}$$

$$= \frac{1}{N} \sum_{i=0}^{N-1} y_0(t_i) y_k(t_i) y_l(t_i) \tag{21}$$

式 (21) は，式 (15) を複数のフレーム間で計算するように拡張したものである．$k = l = 0$ すなわち $C_{y_0 y_0}$ は，式 (15) と同義である．式 (21) により，処理しているフレームの前後のフレームとの相関の度合いが計算される．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．

得られた 3 次キュムラント行列を，データ解析のため 2 次元離散フーリエ変換を行なうことを考える，$C_{y_k y_l}$ についての 2 次元離散フーリエ変換は，以下のように定義される．

$$\hat{C}(\omega_n, \omega_m)$$

$$= \sum_{k=-M}^{M} \sum_{l=-M}^{M} C_{y_k y_l} \cdot w(k,l) \cdot \exp(-j(\omega_n k + \omega_m l)) \tag{22}$$

ここで $\omega_{n,m} = \frac{2\pi}{M}(n,m)$．ただし $n, m = -M, \cdots, M$ は，離散周波数である．$w(k,l)$ は滑らかな値を得るための窓関数である．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 の値はできる限り小さくする．

スペクトルのままでは，非常にデータ量が多いので，得られた 2 次元フーリエ変換行列に対し，PCA(主成分分析) を行なうことによって，次元圧縮を行なう．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．

実際に適用した例を以下に示す．図 1，図 2，図 3 は，それぞれ音声，車内雑音 (H_0)，車内雑音重畳音声 (H_1) の 3 次キュムラントである．図 4，図 5，図 6 は，それぞれ音声，車内雑音 (H_0)，車内雑音重畳音声 (H_1) の 3 次キュムラントバイスペクトルである．雑音重畳音声は，音声データに車内雑音を足しこんだものを用いた．SN 比は 10 dB である．すべてのデータは 200 Hz のカットオフ周波数をもつハイパスフィルタを適用した．ここで，図 6 の雑音重畳音声のバイスペクトルに注目してみると，図 5 の雑音の特徴はほとんど反映されておらず，図 4 の音声の特徴を大きく反映していることがわかる．これより，3 次キュムラントバイスペクトルには，雑音の影響をあまり受けない性質があると考えられる．

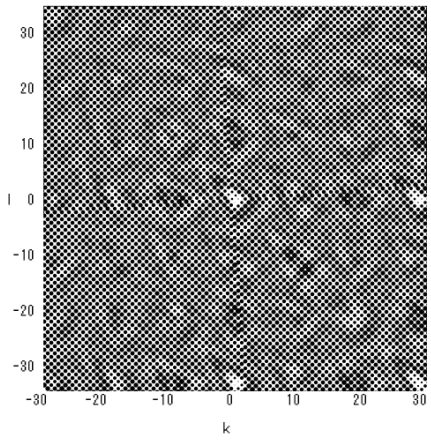

図 1　Clean speech
3rd order cumulant

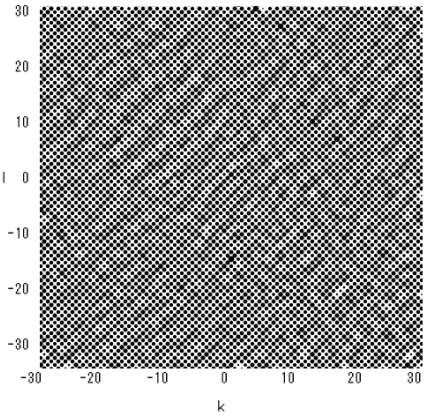

図 2　Car noise
3rd order cumulant

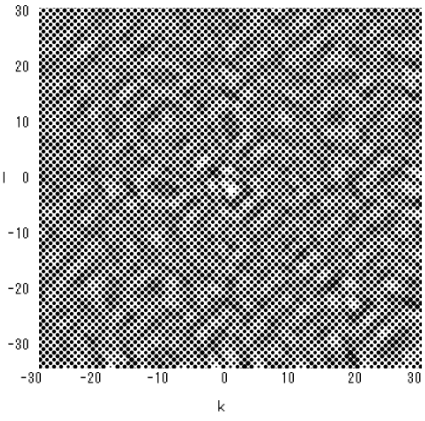

図 3　Noisy speech
3rd order cumulant

2.3　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

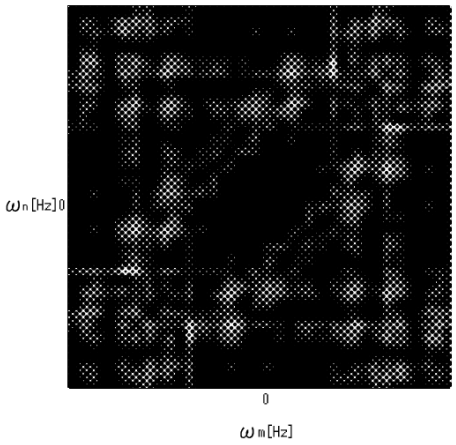

図 4　Clean speech
Bispectrum magnitude

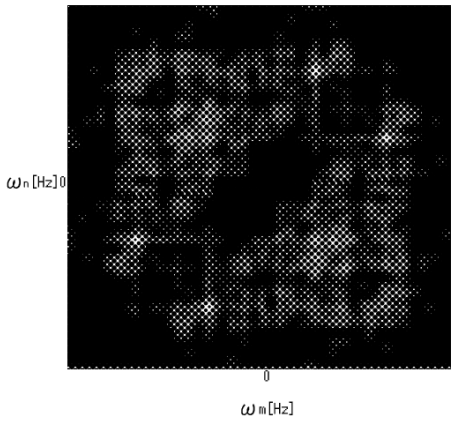

図 5　Car noise
Bispectrum magnitude

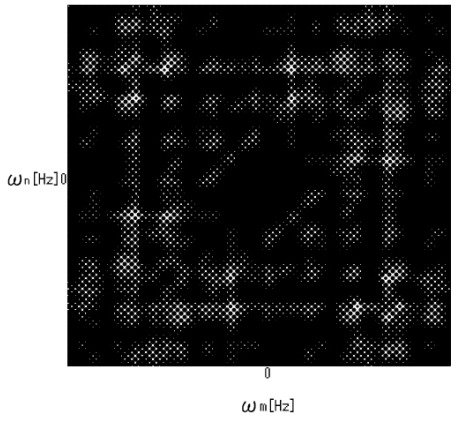

図 6　Noisy speech
Bispectrum magnitude

徴はフレーム間での相関であり，MFCC によって得られる情報は，各フレーム内での音声情報である．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．その方法として，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 次元の音声特徴とする．それをもってキュムラント特徴と MFCC による初期統合音声特徴とする．これを用い，H_0, H_1 についてそれぞれ GMM (Gaussian Mixture Model) を作成する．GMM 学習，及び検定の際，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る．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現される．

$$b(x_t) = \lambda_m b_m(x_{mt}) + \lambda_c b_c(x_{ct}) \tag{23}$$

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，キュムラント特徴ストリーム重みである．ここでは実験により最適なストリーム重みを決定した．

図 (7) は特徴抽出及びその統合，検定までの流れ図である．

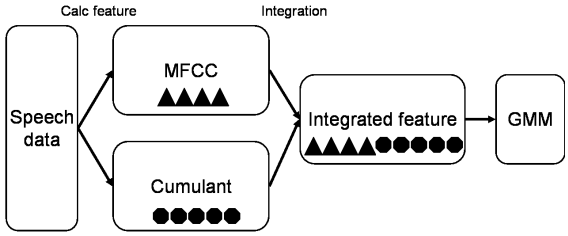

図 7　Integration flow chart

2.4　音声区間検出

得られた GMM を用い，区間検出を行なうデータの各フレームごとに H_0 及び H_1 の尤度を計算する．それらを用いて，式 (24) により対数尤度比 $L(x_t)$ を得る．

$$L(x_t) = \log \frac{P(x_t \mid H_1)}{P(x_t \mid H_0)} \tag{24}$$

ここで，$P(x_t \mid H_1) = \exp(b_w(x_t \mid H_1))$ は音声の尤度，$P(x_t \mid H_0) = \exp(b_w(x_t \mid H_0))$ は非音声の尤度である．

音声区間が分断されることを避けるため，式 (25) により，$L(x_t)$ に対し，隣接する n フレーム間でスムージングを行なう．

$$L'(x_t) = \frac{1}{n} \sum_{j=t-\frac{n}{2}}^{t+\frac{n}{2}} L(x_j) \tag{25}$$

得られた $L'(x_t)$ が，閾値 θ 以上であれば音声，以下であれば非音声とし，暫定的な音声区間を得る．

こうして得られた音声非音声の区間から，連続時間が短いものを取り除くことにより，最終的な音声区間を得る．

3.　評 価 実 験

3.1　データ概要

学習に用いたデータは，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．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

内雑音を加算したものを用いた．

評価に用いたデータは，アイドリング時及び高速道路走行時にて録音された発話データである．どちらも男性 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およそ 10〜25 dB，平均約 17dB，高速道路走行時でおよそ 0〜7 dB，平均約 4 dB である．アイドリング時，高速道路走行時ともに背景雑音として排気音，走行音等が含まれるのみで，音楽，クラクション，ウィンカー音などは含まれていない．

なお，SN 比は以下の式により計算した．

$$P_{noise} = \sum_{\text{非音声のみの区間}} Amplitude^2 \tag{26}$$

$$p_{speech} = \sum_{\text{音声の含まれる区間}} Amplitude^2 - P_{noise} \tag{27}$$

$$SNR = 10 \log \frac{P_{speech}}{P_{noise}} \tag{28}$$

すべてのデータは 12,000 Hz にリサンプリングし，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た．

3.2 比較対象

比較は，音声特徴として，

（1） MFCC
（2） MFCC + Δ
（3） Cumulant
（4） Cumulant + MFCC
（5） Cumulant + MFCC + Δ

の 5 通りを用いて行なった．

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する．片側だけしか検出できていない，若しくはまったく区間を検出できなければ不正解とする．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，それを湧き出しとする．

評価には，以下のものを用いた．

再現率 (*recall*)

$$= \frac{\text{発話区間であると正しく検出された区間の数}}{\text{発話区間の総数}} \tag{29}$$

適合率 (*precision*)

$$= \frac{\text{発話区間であると正しく検出された区間の数}}{\text{検出された区間の総数}} \tag{30}$$

3.3 パラメータ

MFCC は，窓幅 32 ms，シフト幅 8 ms である．CMS (Cepstrum Mean Subtraction) は行っていない．Cumulant は，窓幅 32 ms，シフト幅 1 ms，最大で前後 30 遅延までの計算を行なっている．尚，データは 8 ms ごとに書き出している．

GMM は，H_1(雑音重畳音声) は 64 混合，H_0(非音声) は 32 混合で実験を行なった．尚，H_1 については，男声，女声のモデルをそれぞれ別途に作成した．実験を行なう際，男声尤度，女声尤度の 2 通りを計算し，値が高いほうを H_1 の尤度として採用した [6]．

3.4 ストリーム重み

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．図 8 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．

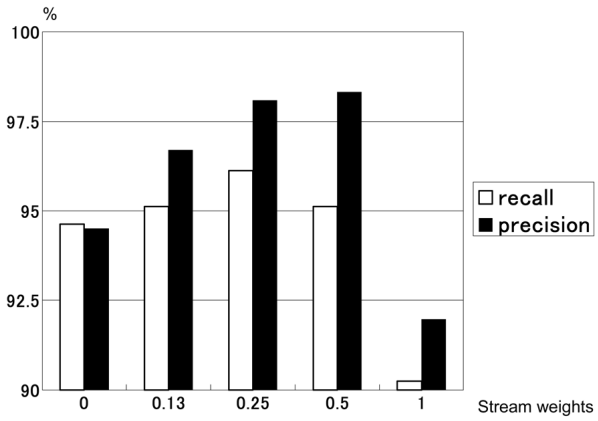

図 8　ストリーム重みを変化させた時の検出率

図より，(5) Cumulant + MFCC + Δ ではキュムラントに対するストリーム重みは，0.25 が最適である事が分かる．

3.5 実験結果

図 9，図 10 は，それぞれアイドリング時，及び高速道路走行時における実験結果である．図 9 より，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．MFCC+Δ とキュムラント特徴とを統合したものは，MFCC+Δ の Recall，Precision が，98.38 %，98.50 %であるのに対し，99.38 %，100.0 %と上回った．

図 10 より，高速道路走行時の SN 比が悪い環境において，MFCC の Recall，Precision が，92.25 %，92.95 %，及び MFCC+Δ が，94.63 %，94.51 %であることに比べ，キュムラント特徴単体では，61.75 %，62.14 %と識別率が大きく落ちていることがわかる．これは 3 次キュムラントバイスペクトルは波形を直接処理しているため，図 6 程度の SN 比であれば，雑音の影響を低減することが出来るが，高速道路走行時程度にまで SN 比が悪くなると，ピッチなどの音声情報が雑音の中に埋もれてしまい，取り出せなくなってしまうためである．しかし，キュムラント特徴と MFCC を統合することにより，93.25 %，93.72 %と MFCC+Δ と同程度の結果が得られている．さらに，Cumulant+MFCC+Δ とすることにより，結果は 96.13 %，98.09 %と，MFCC+Δ を上回った．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，ピッチなど，3 次キュムラントバイスペクトルによっ

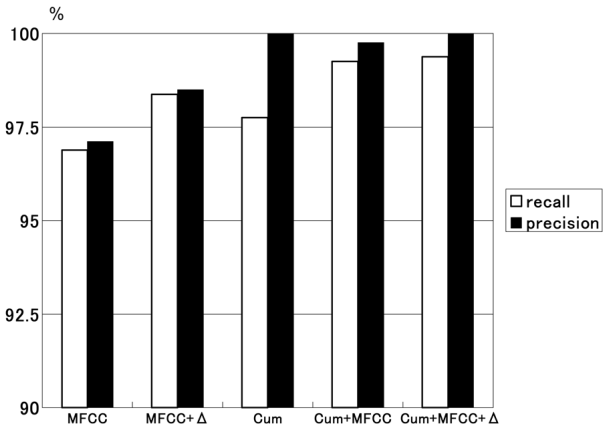

図 9　アイドリング時における実験結果

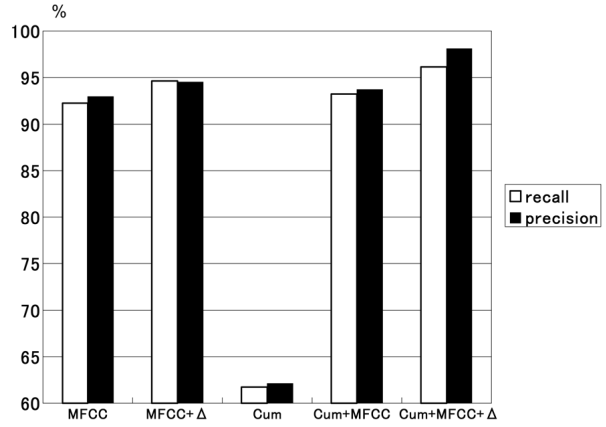

図 10　高速道路走行時における実験結果

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．

4. 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．

従来手法である音声特徴 MFCC 単体では，高速道路走行時の発話など SN 比の悪い環境では音声非音声の分離ができず，区間検出を行えないことがあった．

音声は一般に複数のフレーム間に渡って分布していることから，複数のフレーム間での相関を表すことができれば，雑音に対して強い特徴量を得ることができるのではないかと考え，統計量であるキュムラントにより音声特徴を得る手法を用いた．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．

今後の予定として，現在，波形から算出している 3 次キュムラントバイスペクトル特徴を MFCC から算出，SN 比変えての実験，雑音に音楽などを加えた状況下での実験，パーティ会場など車内環境以外での実験，音声認識への適用などがあげられる．

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．

文　　献

[1] 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.

[2] http://www-pse.cheme.kyoto-u.ac.jp/ kano/document/text-ICA.pdf.

[3] 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.

[4] 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.

[5] 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.

[6] 中村啓介，西村竜一，李晃伸，猿渡洋，鹿野清宏，“実環境音声認識システムのための GMM を用いた環境雑音及び不要発話の自動識別”，日本音響学会講演論文集，pp. 47-48，2004 年 3 月.

[7] Sohn, J., Kim, N.S., and Sung, W.: “A statistical model-based voice activity detection”, IEEE Signal Process. Lett., pp.1-3, 1999, 16, (1).