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次キュムラント音声特徴を用いた音声区間検出

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，音声特徴に高次統計量として知られているキュムラント (Cumulant) を用いること，および，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 3rd Order Cumulant

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

† Faculty of Engineering, Kobe University　Rokkoudaicho1-1, 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) での相関を反映した特徴を得ることができない．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，それをフーリエ変換することによって 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提案する．さらに提案した音声特徴と，MFCC を統合することにより，それらの特徴がお互いにどのように影響を及ぼすかを確認する．

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ほか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．

こうして得られた 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．

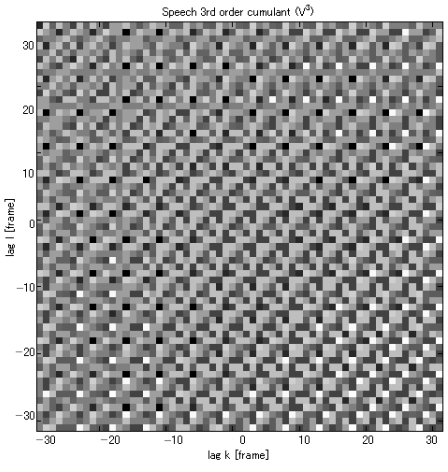

図 1　Speech 3rd order cumulant

図 (1)〜図 (4) に実際に音声から得られたキュムラント，及び 2 次元フーリエ変換のパワースペクトルを載せている．図 (1) に H_1 の 3 次キュムラント，図 (2) に H_1 の 3 次キュムラントをフーリエ変換しパワースペクトルを表示したもの，図 (3) に H_0 の 3 次キュムラント図，図 (4) に H_0 の 3 次キュムラント

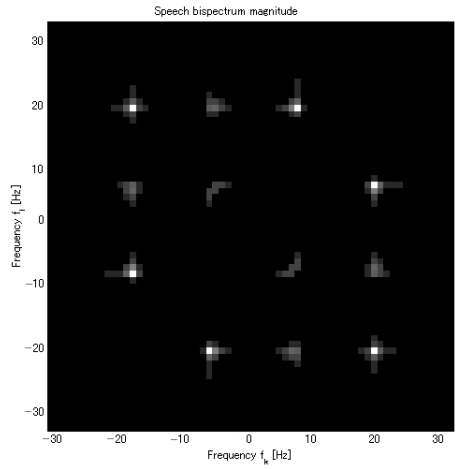

図 2　Speech bispectrum magnitude

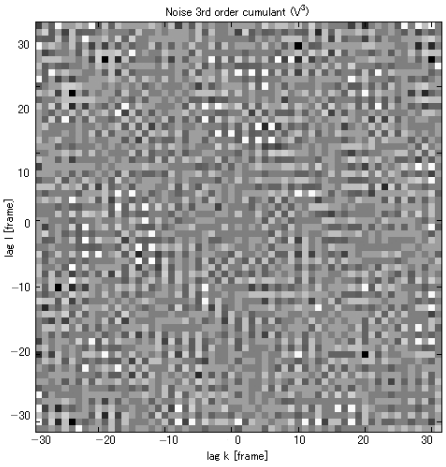

図 3　Non-speech 3rd order cumulant

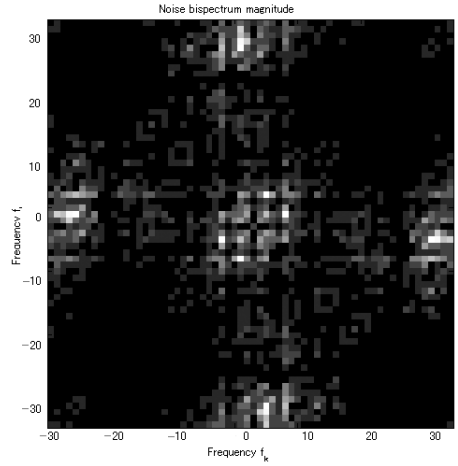

図 4　Non-speech bispectrum magnitude

をフーリエ変換し，パワースペクトルを表示したものを示す．3 次キュムラント図の中心は現在処理しているフレーム，その周辺は現在処理しているフレームと k, l フレームはなれたフレームとの，3 次キュムラントの値である．

スペクトル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から数点を抽出し，そのパワーからなるベクトルで，3 次キュムラント音声特徴とする．

2.3 MFCC との初期統合

キュムラントによって得られる音声特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，これらを補完する方法を考える．その方法として，各フレームから得られたキュムラント特徴 n 次元と MFCC 特徴 m 次元とをあわせ，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と MFCC による初期統合音声特徴量とし検定を行う．図 (5) は特徴抽出及びその統合，検定までの流れ図である．

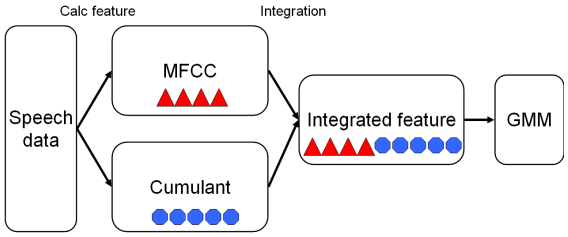

図 5　Integration flow chart

2.4 音声区間検出

これらの音声特徴を用い，H_0, H_1 についてそれぞれ GMM (Gaussian Mixture Model) を作成する．こうして得られた GMM を用い，各フレームごとに H_0 及び h_1 の尤度を計算する．各フレームの尤度に対し，式 (23) により対数尤度比 $L(x)$ を計算する．

$$L(x) = \log \frac{P(x \mid H_1)}{P(x \mid H_0)} \tag{23}$$

ここで，$P(x \mid H_1)$ は音声の尤度，$P(x \mid H_0)$ 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 (24) により，$L(x)$ に対し，隣接する n フレーム間でスムージングを行なう．

$$L'(x_i) = \frac{1}{n} \sum_{j=i-\frac{n}{2}}^{i+\frac{n}{2}} L(x_j) \tag{24}$$

得られた $L'(x_i)$ 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

実験により，音声区間検出の評価を行う．

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 0〜5 dB，平均約 2 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．

$$P_{noise} = \sum_{\text{非音声のみの区間}} Amplitude^2 \tag{25}$$

$$p_{speech} = \sum_{\text{音声の含まれる区間}} Amplitude^2 - P_{noise} \tag{26}$$

$$SNR = 10 \log \frac{P_{speech}}{P_{noise}} \tag{27}$$

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 のみ
（2） MFCC + Δ
（3） Cumulant
（4） Cumulant + MFCC

の 4 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，以下の式を用いた．

再現率 (*recall*)

$$= \frac{\text{発話区間であると正しく検出された区間の数}}{\text{発話区間の総数}} \tag{28}$$

適合率 (*precision*)

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

表 1　VAD result：idling

Data type	Recall	Precision
MFCC	98.50 %	99.24 %
MFCC+Δ	98.37 %	98.50 %
Cumulant	93.13 %	97.51 %
Cumulant+MFCC	99.25 %	98.76 %

表 2　VAD result：highway

Data type	Recall	Precision
MFCC	84.75 %	95.09 %
MFCC+Δ	93.50 %	95.74 %
Cumulant	60.60 %	66.56 %
Cumulant+MFCC	94.63 %	95.89 %

3.3　パラメータ

MFCC は，窓幅 32 ms，シフト幅 8 ms，CMS (Cepstrum Mean Subtraction) を行なっている．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(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　実 験 結 果

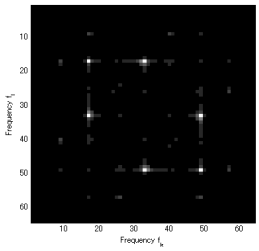

図 6　Noisy speech bispectrum magnitude in idling

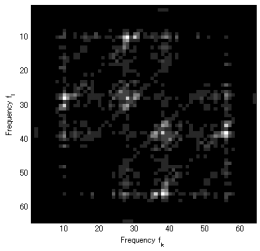

図 7　Noise bispectrum magnitude in idling

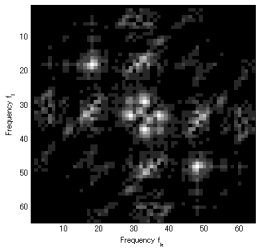

図 8　Noisy speech bispectrum magnitude in highway

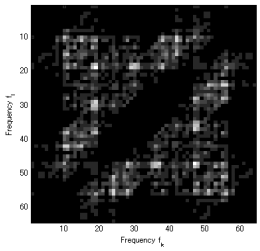

図 9　Noise bispectrum magnitude in highway

図 (6)〜図 (9) に一例として，アイドリング時，及び高速道路走行時のデータより得られた H_1 及び H_0 の 3 次キュムラント 2 次元フーリエ変換図を示す.

図 (6) はアイドリング時に車内にて発話された音声を変換した図である．最大値は約 2.8×10^{10} である．図 (7) はアイドリング時ににおける車内雑音を変換した図である．最大値は約 3.5×10^{6} である．図 (8) は高速道路走行時に車内にて発話された音声を変換した図である．最大値は約 3.8×10^{10} である．図 (9) は高速道路走行時における車内雑音を変換した図である．最大値は約 9.9×10^{8} である.

これらの図より，アイドリング時のような SN 比が良い環境下では，音声にはピッチによる周期性が見られ非音声と模様が異なる上，パワーの値も大きく異なるので，音声非音声の判定は容易に行うことができそうである．しかし，高速道路走行時など非音声が大きくなってくると，音声が，非音声に埋もれてしまい，ピッチによる周期性も見えにくくなっている．パワーの値もアイドリング時ほどの差は見られなくなり，音声非音声の判定は容易には行なえないことがわかる.

アイドリング時における発話区間検出の結果を表 1 に，高速道路走行時における発話区間検出の結果を表 2 に，それぞれ示す.

表 1 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，キュムラントともに識別率に大きな差は見られない．MFCC とキュムラントを統合したものは，Recall , Precision ともにベースラインを上回った.

表 2 より，SN 比が悪くなると，キュムラント単体では MFCC に比べ，識別能が大きく落ちていることがわかる．これはキュムラントは MFCC に比べ分散が非常に小さく，学習データにオーバーフィッティングしているためと考えられる．しかしキュムラントと MFCC を統合したものは，ベースラインより，良い結果を得られている.

4.　ま　と　め

本研究では，者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，キュムラントを用いて音声特徴を得る手法を提案した．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，これらの特徴量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，2 次元フーリエ変換したものから数点選ぶ際に PCA を用いる，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, pp141-142，2001 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，pp47-48，2004 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).