2004－ＨＩ－109　(24)
2004－SLP－ 52　(24)
2004／7／17

高次統計量の分布モデルを用いた音声・環境音識別法の検討

吉村 隆[†]、浅野　太[†]、麻生　英樹[†]、北脇　信彦[‡]

産業技術総合研究所 情報処理研究部門[†]、
筑波大学大学院 システム情報工学研究科[‡]

yoshimur@ni.aist.go.jp , {f.asano , h.asoh}@aist.go.jp , kitawaki@is.tsukuba.ac.jp

URL - http://www.media-interaction.jp/

あらまし：　音響信号の種類毎にモデル化された高次統計量の分布を用いて、新たな入力の分布から音声と他の環境音を識別する手法について検討を行った。高次統計量として，独立成分分析などで用いられる尖度に着目した. 尖度を確率変数とした確率密度関数を信号の種類毎に推定し，入力信号との平均対数尤度が安定して最大となる信号の種類を識別結果としたところ，提案手法は実験結果より特に単発性環境音と音声を区別するのに有効である可能性が示された.

キーワード：　高次統計量，尖度，確率密度関数，ノンパラメトリック推定手法

Investigation of Voice/Sound Activity Classifier using Distribution Models of Fourth-Order Statistics

Takashi YOSHIMURA[†], Futoshi ASANO[†],
Hideki ASOH[†] and Nobuhiko KITAWAKI[‡]

AIST 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, JAPAN[†],
University of Tsukuba[‡]
yoshimur@ni.aist.go.jp , {f.asano , h.asoh}@aist.go.jp , kitawaki@is.tsukuba.ac.jp

URL - http://www.media-interaction.jp/

Abstract: We investigate an effective classifier of voice/sound activity. The classifier is constructed by distribution models of fourth-order statistics, especially kurtosis. The distribution models are estimated by kernel density of each voice/sound category. The mean logarithmic probabilities of each voice/sound category were calculated from kurtosis of input signals. Categories with stable and high probabilities were classification results of the input signals. Experimental results show that proposed classification methods are more effective in identification of voice and separated short sound.

Key words: Fourth-order statistics, Kurtosis, Kernel Density, Non-parametric estimation

1. はじめに

ロボットなど人間の生活を支援する機械が，実際に家庭の居間や職場の会議室などで，利用者とのやりとりを円滑に進めるためには，室内に存在するさまざまな物体や音の中から，機械に話しかけ指示を与えている利用者のいる方向を特定し，背景の雑音を取り除いて，利用者の発声する音響信号のみを抽出し，認識および理解することが不可欠となる．我々はこれまで，マイクロフォンアレイから得られた音響信号の情報と，カメラから得られた画像信号の情報を統合することにより，システムに向かって話しかけるユーザの方向を特定し，そのユーザの音声のみを分離抽出する研究を行ってきた[1][2].

これは確率ネットワークを用いて，音響情報から得られる音源方向の情報と，画像情報から得られる背景画像との差分情報を統合して，発話者の位置と発話区間を推定している．しかし，雑音源としてテレビが存在し，画面の場面が切り替わる場合などは，音源方向情報と背景差分画像情報の統合から推定される発話区間にテレビの音声が混在し，推定が不安定になる可能性がある．また，システムから見てユーザの後方から物音が聞こえてくる場合などは，推定した発話区間が，ユーザの話しかけていない時刻で物音のする区間となっていても，判別が困難である．したがって，画像情報との統合により発話区間推定を行う前処理として，音響情報から得られた音源方向ごとに分離されたそれぞれの音響信号の音声らしさを調べることにより，音声らしさの低い音響信号を本処理には用いず取り除き，発話区間推定の精度を高めることがのぞましい.

本研究では，この前処理にふさわしい，音声らしさを表す特徴量を検討し，推定を安定させるための補助的な情報源となりうる，処理時間の少ない，処理過程の複雑でない音声らしさを識別する手法について考察を行ったので，以下に述べる.

2. 音響信号の高次統計量

これまでに，独立成分分析などで用いられる高次統計量から音響信号のS/N比を求め[3]，雑音が重畳された信号の中から発話区間を検出する手法が報告されている．この発話区間を検出する手法では，発話区間信号と非発話区間信号のエネルギーの差を用いて閾値以上のエネルギーがある場合に発話区間と判別する手法，あらかじめ求めておいた雑音の平均スペクトルまたは平均ケプストラムと分析区間の信号との距離を求め閾値以上の距離がある場合には発話区間と判別する手法[4]などが代表的である．さらに発話区間検出の精度を向上させるため，細かく改良された分析結果を用いる手法が多数報告されているが，雑音の変化に応じて閾値を設定するために多くの計算量と経験を要してきた．対照的に文献[3]で述べられている S/N 比を求める手法は，事前の知識より閾値を設定する必要もなく計算量も実時間処理が可能なほど少なくて済む.

よって本研究では，分離された音響信号の性質を表す特徴量として，独立成分分析などで用いられる 4 次統計量について検討を行う.

観測された音響信号を $x(n)$ とすると，音響信号の分布形状を評価するために，平均値 m の周りの中心モーメントが用いられる.

$$M_k \equiv E\left[x^k(n)\right] = \frac{1}{N}\sum_{i=1}^{N}\left(x(i)-m\right)^k$$

$$m = \frac{1}{N}\sum_{i=1}^{N}x(i)$$

ある分析区間内の観測音響信号 N 点のｋ次中心モーメント M_k は上記のように表される．分散 σ^2 は 2 次中心モーメント M_2 である．ここで $E[\cdot]$ は期待値を表す．4 次中心モーメントは信号の分布の正規分布からのずれを表し，値の一部が他のものとかけ離れていると大きくなり，分布が平均値 m の周りにかたまっていると小さな値になる．それゆえに，分布の中央が尖っていて両側に長く広がっている分布では大きくなり，中央が平坦で台形状に両端が切れた分布では小さくなる，このことから，

$$t = \frac{M_4}{\sigma^4} - 3 \quad , \quad M_4 = \frac{1}{N}\sum_{i=1}^{N}\left(x(i) - m\right)^4$$

尖度（kurtosis）tは分布の‘裾’の長さを示す尺度となる．

これまでの研究で，音声や楽音などは尖度が正の値を示す super-Gaussian，定常雑音は尖度が負の値となる sub-Gaussian，さらに人工的なガウス雑音は尖度がほぼ 0 となり分類されることが報告されている[5]．これは，信号全体を一つの分析区間とした大まかな分類ではあるが，音響信号の種類により尖度が異なる範囲に現れ，信号の種類を識別する特徴量となりうる可能性を得ている．

我々は，さまざまな音響信号について，実時間上の尖度の変化を調べるため，分析区間を移動させ，信号の種類ごとに特徴が現れるかを考察したので，次節で詳細を述べる．

3. 4 次統計量の分析

各種音声・環境音・楽音データベースを用いて，尖度の時間変化に関する定量的な検証を行った. 分析に使用したデータベースは，音声に関しては日本音響学会 研究用連続音声 DB（音素バランス文を各話者が約 150 文読み上げ）[6]，楽音は RWC 研究用音楽 DB（クラシック，ジャズ，ポップス etc.）[7]，環境音はRWCP 実環境音声・音響DB(拍手，電話，ドア，ドライヤー etc.）[8]，さらに参考までに背景雑音として，AURORA-2J DB（レストラン, 展示場, 空港, 駅）[9]である．

分析条件をそろえるため，サンプリング周波数は，原データと周波数が異なる場合は変換して，すべて 16kHz とした．また分析フレーム長を 100msec，分析フレーム周期を 100msec として尖度の時間変化と分布を調べたところ，音響信号の種類ごとに以下の特徴が見られた. ただし, 尖度の分布を表す際，区間内がすべて無音である分析フレームより求めた尖度は除いている．

- 音声 : 話者ごとに尖度の分布を求めた．0 付近の正の値で分布が最大になる．+20 前後まで減衰しながら分布が広がる．負方向への分布の広がりはほとんどない．男性話者のほうが女性話者より分布の正方向の広がりが大きい．尖度の時間変化が大きい．（図 1）

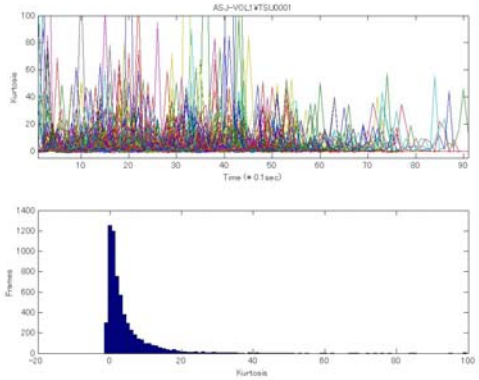

図 1 : 尖度の時間変化（上段）および分布（下段）
［ASJ 研究用連続音声 DB ― 男性話者 TSU0001］

- 楽音 : 音楽のジャンルごとに尖度の分布を求めた．各信号が非常に長いので，冒頭の 30 秒間のみ分析を行った．0 付近の負の値で分布が最大になる．分布は狭い．他種の音響信号と比べ，負の値の分布が多い．これは信号内で周期性（リズムおよび音程）が特に強調されていることが原因であると考えられる．尖度の時間変化は小さいが，音声を含むポップスはクラシックおよびジャズに比べて時間変化がやや大きい．（図 2）

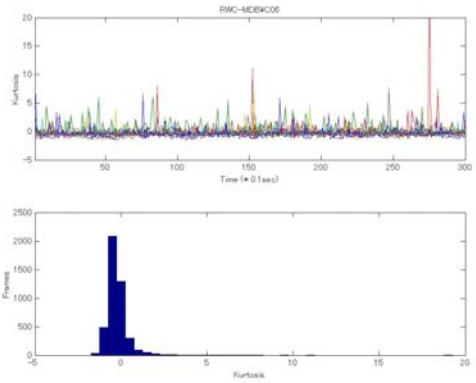

図 2 : 尖度の時間変化（上段）および分布（下段）
［RWC 研究用音楽 DB ― クラシック C06］

- 環境音：ドアや手をたたく音などの単発音では，短時間に+50～+150 または，それ以上の尖度を持ち，時間変化も非常に大きい．(図 3) 連続空調音や機械音などの定常的な音響信号の尖度分布は周期性のないガウス性雑音に近く，ほとんど 0 付近で変化が見られない．(図 4)

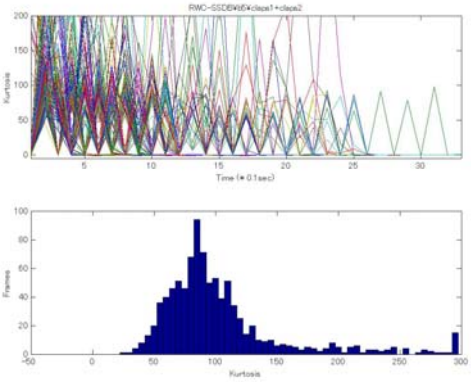

図 3：尖度の時間変化（上段）および分布（下段）
［RWCP 実環境音声・音響 DB － 拍手］

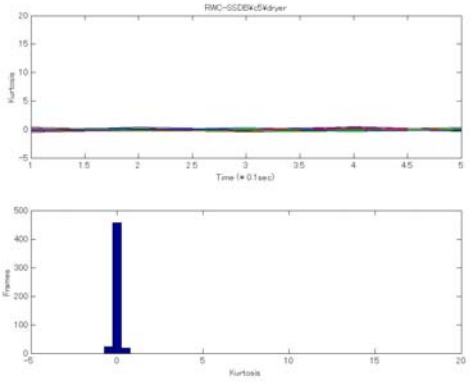

図 4：尖度の時間変化（上段）および分布（下段）
［RWCP 実環境音声・音響 DB－ ドライヤー］

さらに，単独の音響信号ではないが，複数の音響信号が混在する雑音についても，参考までに尖度の分析を行った．

- 背景雑音：定常的な環境音と同様に尖度の分布は 0 付近で最大となり，楽音よりさらに分布は狭い．時間変化も少なく．さまざまな音響信号が含まれる背景雑音では，周期性がほとんど見られないことが，この特徴の原因と考えられる．(図 5)

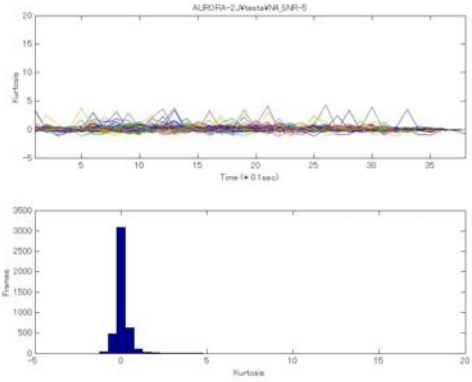

図 5：尖度の時間変化（上段）および分布（下段）
［AURORA-2J DB － テストセット A
Exhibition －5dB］

以上より，文献[5]における知見を満たしつつ，尖度の詳細な分析が行えたと考えられる．なお本研究では，あらかじめ分離された音響信号について識別を行うことを想定しているので，次節以降背景雑音を除く各種データベースを対象とした議論を行う．

4. 音声・環境音識別および考察

尖度 t を確率変数ととらえ，確率密度関数 $p(t)$を以下のような代表的ノンパラメトリック手法で推定する[10]．$t(i)$を確率変数 t の分析区間 i における観測値とすると，t における密度 $p(t)$の推定のためのカーネル密度推定量 $p_h(t)$は，以下の式で定義される．

$$\hat{p}_h(t) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t(i)-t}{h}\right)$$

ここで $K(u)$はカーネル関数を表し，h はバンド幅を表す．カーネル関数にはガウス関数

$$K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$$

を用いて，推定値が

$$\int \hat{p}_h(t)dt = 1$$

を満たすように正規化を行う．例えば図 3 下

段における尖度分布より，図 6 のように正規化された確率密度関数が推定される．ただし，ここでは計算量低減のため，尖度 t の各観測点を 0.1 ごとに区切られた階級の中心点に置き換えるという近似を行っている．

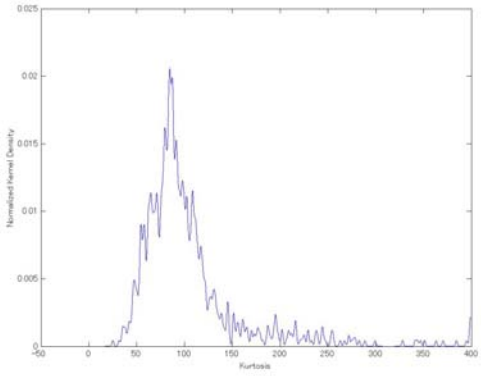

図 6：正規化推定確率密度関数
［RWCP 実環境音声・音響 DB － 拍手］

音響信号の種類 C_j ごとに推定された尖度分布の密度関数を用いて，未知入力信号の平均対数尤度を求める．

$$P_{C_j} = \frac{1}{N}\sum_{i=1}^{N} \log \hat{p}_h\left(t(i)\middle|C_j\right)$$

ここでは，未知入力信号冒頭の分析区間数 N の尖度 $t(i)$［$i=1,\ldots,N$］のみを用いて尤度が求められるので，未知入力信号がすべて入力されることを待たず，実時間逐次処理への拡張を検討することも可能である有用な導出法であると考えられる．分析区間数が増えるごとに尤度を更新し，安定して高い尤度を得る音響信号の種類 C_j が現れたところで，それを識別結果とする．

平均対数尤度が更新される例を以下に示す．男性話者（1 名）音声，女性話者（1 名）音声，クラシック音楽，ポップス音楽，拍手環境音，ドライヤー環境音，ホイッスル環境音の 7 種類の音響信号について，尖度の確率密度関数を推定し正規化を行った．図 7 は，これらを用いて，他の男性話者の 1 発話が未知入力信号となった場合の平均対数尤度が更新された様子である．横軸は分析フレーム数の増加を表し，縦軸はそれに伴う各確率密度関数から求められた尤度の変化を表している．なお，分析フレーム長および分析フレーム周期は分析時と同じ，各 100msec である．未知信号入力開始から十数フレーム経過した時点から，2 種類の尤度が大きくなり安定する．これは男性話者および女性話者の確率密度関数から得られた尤度を表し，1～2 秒程度で，音声と識別できていることを示している．もう 1 種類，比較的尤度が大きいまま更新を続けているものはホイッスル環境音から求められた尤度であり，突発性と周期性をあわせ持つ，この環境音の特徴が他の信号に比べて音声に類似していることが原因と考えられる．

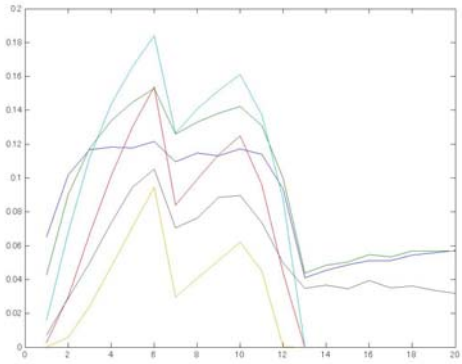

図 7：平均対数尤度の更新のようす
［未知入力信号：男性話者 NEC0001 より 1 発話］

次に，同じ 7 種類の音響信号から得られた確率密度関数を用いて，拍手環境音の 1 信号が未知入力信号として与えられた場合の平均対数尤度の更新の様子を調べた．図 8 が，その結果を表している．入力開始直後より，拍手環境音の確率密度関数から得られた尤度が他種類を大きく引き離し優位となって，単発の環境音と識別できていることを示している．

以上の試行実験の結果より，本研究で提案した手法を用いて，音声・環境音識別が比較的短い分析時間で，かつ少計算量で実現できる可能性を得たと考えられる．

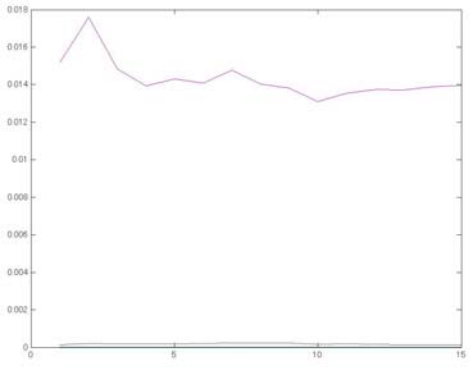

図 8：平均対数尤度の更新のようす
［未知入力信号：拍手環境音より 1 信号］

5. まとめ

本研究では，4 次統計量である尖度の分布を確率密度関数という形で推定したモデルを用いることにより，音声と環境音を識別する手法を提案した．提案手法を用いて試行実験を行った結果，未知信号入力開始から短時間で各モデルとの平均対数尤度に安定した差が現れ，音声・環境音識別器として有効である可能性を得た．今後の課題として，さらに多くの入力信号を用いて提案手法の定量的な評価をすすめる予定である．

試行実験の結果より，提案手法では単発的な環境音を他種類の音響信号と区別することが特に容易であると考えられるので，初期段階として単発性環境音のみを識別し，次の段階において音声と定常的な環境音（楽音など）を識別する段階的な処理の有効性について検討をすすめる．その際，識別時の分析フレーム長および分析フレーム周期を調整し，無音判別を伴わない，より簡便な手法を検討する必要があると考えている．また，本報告における識別手法では，尖度の時間変化情報を用いなかったが, 定常音どうしの識別では, その特徴の類似度を加味した尤度となることが必要である．

謝辞

日頃よりご支援とご討論をいただく，産業技術総合研究所 情報処理研究部門 メディアインタラクショングループ，ならびに筑波大学 電子･情報工学系 マルチメディア研究室の皆様に感謝いたします．

参考文献

[1]F. Asano et al., “Detection and Separation of Speech Segment Using Audio and Video Information Fusion”, *Proc. Eurospeech 2003*, pp. 2257-2260, Sep. 2003

[2]T. Yoshimura et al., “Detection of Speech Events in Real Environments through Fusion of Audio and Video Information Using Bayesian Networks”, *Proc. IWAENC 2003*, pp. 319-322, Sep. 2003

[3]S.E. Bou-Ghazale et al., “A Robust Endpoint Detection of Speech for Noisy Environments with Application to Automatic Speech Recognition”, *Proc. ICASSP 2002*, vol.4, pp. 3808-3811, May 2002

[4]E. Nemer et al., “SNR Estimation of Speech Signals Using Subbands and Fourth-Order Statistics”, *IEEE SP Letters*, vol.6, no.7, pp. 171-174, Jul. 1999

[5]Te-Won Lee, *Independent Component Analysis -Theory and Applications-*, Kluwer Academic Publishers, Boston, 1998

[6]http://www.milab.is.tsukuba.ac.jp/corpus/asj_move.html

[7]http://staff.aist.go.jp/m.goto/RWC-MDB/

[8]http://tosa.mri.co.jp/sounddb/

[9]山本 ほか，“AURORA-2J/AURORA-3J データベースとその評価ベースライン”, 情報処理学会研究報告，SLP-47-19，Jul. 2003

[10]Richard O. Duda et al., *Pattern Classification (2nd Edition)*, John Wiley & Sons, Inc., NJ, 2000