ブラインド音声抽出システムにおける音声認識率予測のための音声カートシス推定法*

☆平野佑佳，宮崎亮一，猿渡洋，中村哲 (奈良先端大)，高谷智哉（トヨタ自動車）

1 はじめに

近年，音声対話ロボットや音声検索などの音声認識に関するアプリケーションが研究されている．しかし，環境雑音によって音声認識性能が低下し，システム全体の性能が劣化するという問題がある．そのため，高精度な音声抽出手法が望まれる．先行研究において，我々は高精度な音声抽出手法として，ブラインド空間的サブトラクションアレー（blind spatial subtraction array: BSSA）[1] を提案している．BSSA では雑音推定器として独立成分分析（indipendent component analysis: ICA）[2] が用いられており，ICA で推定した雑音をパワースペクトル領域で減算 [3] することで音声を抽出する．BSSA は雑音推定に ICA を用いているため，非定常雑音を含む実環境においても，音声を高精度に抽出できる．

BSSA にはいくつかの内部パラメータが存在するが，音声認識性能が最大となるパラメータは手動で調節する必要がある．さらに，音声認識システムを運用する場所が変わると，最適な内部パラメータは変化するため，再度チューニングが必要となる．よって，環境ごとに最適な内部パラメータを自動チューニングする枠組みが望まれる．環境に合わせた内部パラメータを自動的にチューニングするには音声認識率の予測が必要となる．音声認識率を予測する要素として，雑音抑圧性能，音声歪み量，ミュージカルノイズ発生量が有効であると言われている [4]．音声歪み量の評価尺度として一般的にケプストラム歪みが用いられているが，ケプストラム歪みを求めるには雑音を含まない真の音声信号が必要となる．そのため，実環境下でケプストラム歪みを求めることはできない．そこで，教師無しで推定可能な音声歪み量の評価尺度として，音声カートシスが提案されている [4]．これまで，音声カートシスの推定法として，観測信号と推定雑音信号から直接求める手法（従来手法 1）[4] と，カートシステーブルを作成して，テーブルルックアップで求める手法（従来手法 2）[5] が提案されている．従来手法 1 は，推定雑音信号の信号長が短いと安定して音声カートシスを推定できない問題がある．また，従来手法 2 は入力 SNR 毎にカートシステーブルを作成するが，観測信号から入力 SNR を高精度に推定することは困難である．

本稿では，これらの問題点を解決するために，ICA で推定された雑音を用いて音声カートシスを推定する手法を提案する．また，従来手法と提案手法の比較実験より，提案手法における音声カートシスの推定精度は安定して高いことを確認した．

2 先行研究

2.1 音声カートシス直接推定法 [4]

目的音声信号と雑音信号が混合した観測信号は以下の式で表される．

$$x(f,\tau) = s(f,\tau) + n(f,\tau) \tag{1}$$

ここで，$s(f,\tau)$ は目的音声信号，$n(f,\tau)$ は雑音信号，$x(f,\tau)$ は目的音声信号と雑音信号が混合された観測信号を表し，f は周波数ビン，τ は時間フレーム番号を表す．実環境下で観測される信号において，音声信号は時間周波数領域上で常に雑音信号の影響を受けるため，観測信号から音声信号のカートシスを推定することは非常に難しい．そこで，雑音信号および観測信号における統計量を用いて，音声パワースペクトルのカートシスを逆算的に求める手法が提案されている．この手法はケプストラム歪みのような真の音声信号を必要としない教師無し手法である．

波形領域においては，音声信号と雑音信号は加法的な関係を持っているが，高次のモーメントにおいてはその加法性が保持されないため，雑音信号および音声と雑音の混合信号のモーメントから直接的に音声信号のカートシスを求めることは困難である．そこで，キュムラントと呼ばれる統計量を導入する．キュムラントにおいて加法的な信号に対する加法性は保持されるため，観測信号および雑音信号のキュムラントから音声信号のキュムラントを推定可能であると考えられる．一方で，波形信号からパワースペクトルを求める際には指数乗演算が行われる．指数乗演算において，キュムラントは直接的な関係を持っていないため，指数乗演算にはキュムラントよりもモーメントを利用することが望ましい．以上の性質より，先行研究 [4] においては，モーメント-キュムラント変換を利用した音声カートシス推定法が提案されている．

モーメント-キュムラント変換において，m 次モーメントは m 次までのキュムラントを用いると以下のように表される [4]．

$$\begin{aligned}\mu_m(x) &= \left.\frac{\partial^{(m)} \exp(\log \phi_x(it))}{\partial it^{(m)}}\right|_{t=0} \\ &= \sum_{\pi(m)} \prod_{B\in\pi(m)} \kappa_{|B|}(x)\end{aligned} \tag{2}$$

ここで，$\mu_m(x)$ は確率信号 x の m 次モーメント，$\pi(m)$ は m の分割パターン，B はある分割パターンにおけるブロック，$|B|$ はブロックのサイズを表す．また，m

---

* “ Speech kurtosis estimation in blind speech extraction for speech recognition performance prediction, ” by Yuka Hirano, Ryoichi Miyazaki, Hiroshi Saruwatari, Satoshi Nakamura (Nara Institute of Science and Technology), and Tomoya Takatani (TOYOTA MOTOR CORPORATION.)

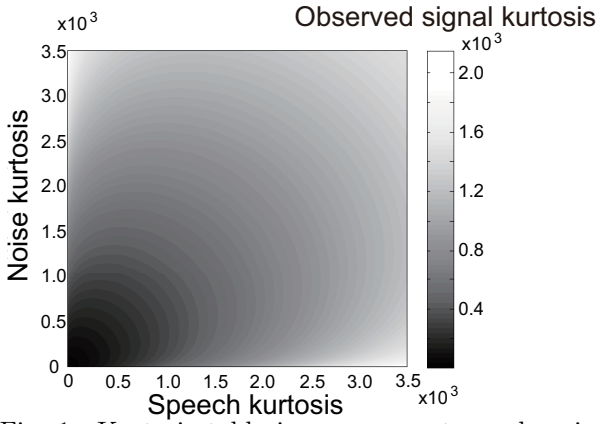

Fig. 1 Kurtosis table in power spectrum domain when input SNR is 0 dB.

次キュムラントは m 次までのモーメントを用いて以下のように表される．

$$\kappa_m(x) = \left.\frac{\partial^{(m)} \log \phi_x(it))}{\partial it^{(m)}}\right|_{t=0} = \sum_{\pi(m)} (-1)^{|\pi(m)|-1}(|\pi(m)|-1)! \prod_{B\in\pi(m)} \mu_{|B|}(x) \tag{3}$$

ここで，$\kappa_m(x)$ は確率信号 x の m 次キュムラントである．これらモーメント-キュムラント変換とキュムラントの加法性を利用することで，音声カートシスの推定を行う．観測信号，音声信号および雑音信号の複素スペクトルをそれぞれ $(x_\mathrm{R}+ix_\mathrm{I})$，$(s_\mathrm{R}+is_\mathrm{I})$ および $(n_\mathrm{R}+in_\mathrm{I})$ とするとき，推定されたパワースペクトル領域の音声カートシスは以下で与えられる．

$$\mathrm{kurt}_\mathrm{speech} = \frac{\mu_4(s_\mathrm{R}^2+s_\mathrm{I}^2)}{\mu_2^2(s_\mathrm{R}^2+s_\mathrm{I}^2)} = \frac{\mathcal{N}\left(\mu_m(x_\mathrm{R}),\mu_m(n_\mathrm{R})\right)}{\mathcal{D}\left(\mu_m(x_\mathrm{R}),\mu_m(n_\mathrm{R})\right)} \tag{4}$$

ここで，$\mathcal{N}(\cdot)$ および $\mathcal{D}(\cdot)$ は以下のとおりである．

$$\begin{aligned}
&\mathcal{N}\left(\mu_m(x_\mathrm{R}),\mu_m(n_\mathrm{R})\right)\\
&= \mu_8(x_\mathrm{R}) - \mu_8(n_\mathrm{R})\\
&\quad + [4\mu_2(x_\mathrm{R}) - 32\mu_2(n_\mathrm{R})]\,\mu_6(x_\mathrm{R})\\
&\quad + [-32\mu_2(x_\mathrm{R}) + 60\mu_2(n_\mathrm{R})]\,\mu_6(n_\mathrm{R})\\
&\quad + [-76\mu_4(n_\mathrm{R}) - 96\mu_2(x_\mathrm{R})\mu_2(n_\mathrm{R})\\
&\qquad + 516\mu_2^2(n_\mathrm{R})]\mu_4(x_\mathrm{R})\\
&\quad + [-60\mu_2^2(x_\mathrm{R}) + 1056\mu_2(x_\mathrm{R})\mu_2(n_\mathrm{R})\\
&\qquad - 1416\mu_2^2(n_\mathrm{R})]\mu_4(n_\mathrm{R})\\
&\quad + 3\mu_4^2(x_\mathrm{R}) + 73\mu_4^2(n_\mathrm{R})\\
&\quad + 468\mu_2^2(x_\mathrm{R})\mu_2^2(n_\mathrm{R}) - 3456\mu_2(x_\mathrm{R})\mu_2^3(n_\mathrm{R})\\
&\quad + 2988\mu_2^4(n_\mathrm{R})
\end{aligned} \tag{5}$$

$$\begin{aligned}
&\mathcal{D}\left(\mu_m(x_\mathrm{R}),\mu_m(n_\mathrm{R})\right)\\
&= 2(\mu_4(x_\mathrm{R}) - \mu_4(n_\mathrm{R}) + \mu_2^2(x_\mathrm{R})\\
&\qquad - 8\mu_2(x_\mathrm{R})\mu_2(n_\mathrm{R}) + 7\mu_2^2(n_\mathrm{R}))^2
\end{aligned} \tag{6}$$

2.2 カートシステーブル法 [5]

従来手法 1 を用いることで，真の音声信号を利用することなく音声パワースペクトルのカートシスを求めることが可能となる．しかし，従来手法 1 による音声カートシスの推定精度は非常に不安定であることが知られている [5]．これは，従来手法 1 において音声カートシスを推定するには 6 次や 8 次といった非常に高次の統計量を求める必要があり，これらの高次の統計量を実際に観測可能な有限サンプルから安定的に求めることが困難なためである．そこで，波形信号より高次の統計量を計算せずに，パワースペクトル領域において直接的に音声カートシスを推定する手法が提案された [5]．この手法では，パワースペクトル領域における音声信号，雑音信号および混合信号のカートシスの直接的な関係を表すルックアップテーブル（カートシステーブル）を作成し，観測された雑音信号および混合信号より得られるカートシスの値から，音声パワースペクトルのカートシスの推定を行う．カートシステーブルの作成には，パワースペクトル領域における信号間でのカートシスの関係を求める必要がある．しかし，同じパワースペクトルカートシスを有する信号には無限のパターンが存在するため，一意に信号を同定することは難しい．そこで，本手法では音声および雑音の波形信号に対して一般化ガウス分布に従うという統計的仮説を与えることにより，カートシステーブルを作成する．

一般化ガウス分布の確率密度関数（p.d.f.）は以下の式で定義される．

$$p(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)}\exp(-(|x|/\alpha)^\beta) \tag{7}$$

ここで，α は尺度母数，β は形状母数，$\Gamma(\cdot)$ はガンマ関数を表す．さらに，一般化ガウス分布における m 次モーメントは以下の式によって与えられる．

$$\mu_m(x) = \int_{-\infty}^{\infty} x^m p(x)dx = \alpha^m \frac{\Gamma((m+1)/\beta)}{\Gamma(1/\beta)} \tag{8}$$

本手法においては，式 (8) およびモーメント-キュムラント変換を用いて，入力 SNR 毎のカートシステーブルを作成する．詳細は文献 [5] を参照されたい．作成されたカートシステーブルを利用して，雑音信号および観測信号におけるパワースペクトルのカートシスから，テーブルルックアップにより音声パワースペクトルのカートシスを求める事が可能となる．Figure 1 に入力 SNR が 0 dB の場合のカートシステーブルの例を示す．

3 ICA 推定雑音を用いた音声カートシスの推定

3.1 概要

2.2 節で述べたように，安定して音声カートシスを推定するためには従来手法 2 を用いることが好ましい．しかし，従来手法 2 では正確な入力 SNR が必要であるが，実環境において正確な入力 SNR を求める

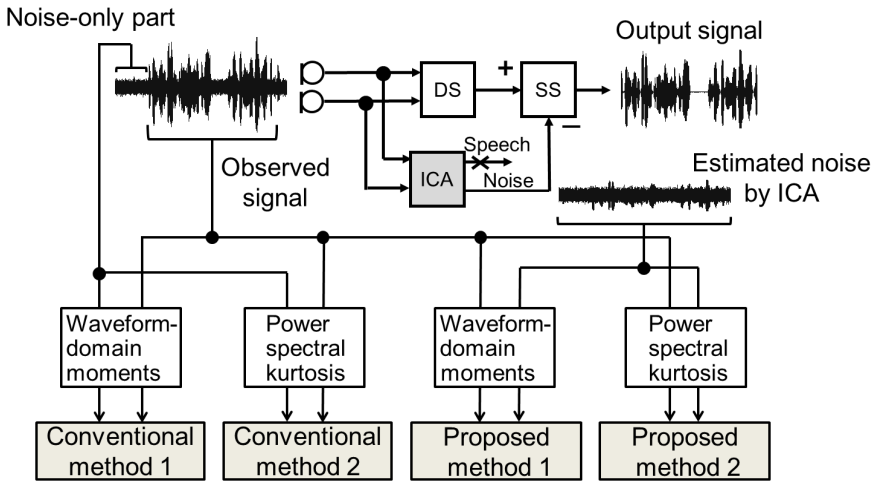

Fig. 2　Block diagram of BSSA, ICA-based noise estimation, conventional methods, and proposed methods.

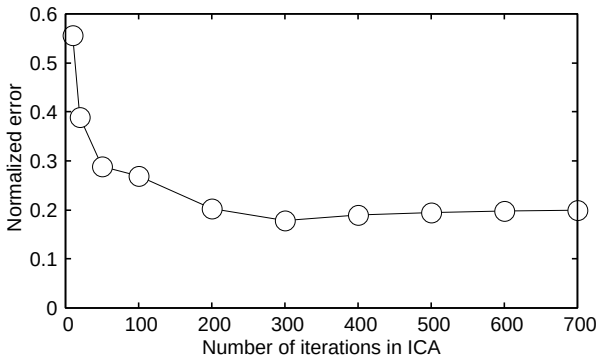

Fig. 3　Experimental results of speech kurtosis estimation (proposed method 1).

ことは困難である．仮に異なる入力 SNR のカートシステーブルを参照した場合，推定精度は大幅に低下する．そのため，充分な長さの信号を用意できるのであれば，従来手法 1 による音声カートシス推定精度が従来手法 2 よりも良い可能性がある．実環境において，観測信号は充分な長さを用意できるが，雑音信号は混合信号中の非音声区間より得られるため，充分な長さの信号を用意できない場合が多い．そこで，ICA による推定雑音信号を用いた音声カートシス推定法を提案する．

Figure 2 に従来手法および提案手法で用いる信号を明示したブロック図を示す．音声抽出手法である BSSA では，Fig. 2 のように ICA を用いて雑音を推定する．実環境において，ICA は高い雑音推定精度を有することが知られている．ICA により推定された雑音信号の長さは，入力された信号の長さで出力される．従来手法では，雑音信号には観測信号中の非音声区間を用いるため，充分な信号長が得られないことが問題となったが，提案手法では雑音カートシスの算出に ICA による推定雑音を用いるため，充分な信号長が確保でき，安定的に音声カートシスを推定できることが期待される．以降，この手法を提案手法 1 と呼ぶ．また同様に，従来手法 2 に ICA 推定雑音を適用する．以降，この手法を提案手法 2 とする．

3.2　ICA の反復回数による音声カートシス推定精度の評価

3.2.1　実験条件

予備実験として，提案法における ICA の反復回数と音声カートシス推定精度の評価実験を行った．実験には二素子のマイクロホンアレーを用い，素子間隔は 2.1 cm，目的音声の方位は正面に設定した．実験に用いた部屋は 4.2 × 3.5 × 3.0 m^3 で，残響時間は約 200 ms であった．実験に用いた信号のサンプリング周波数は 16 kHz で，量子化ビット数は 16 bit であった．観測信号は 1 話者（女性）の目的音声信号からなり，環境雑音として実収録した駅雑音を用いた．入力 SNR は 0 dB，FFT 長は 1024，シフト長は 256 であった．音声カートシス推定法には提案手法 1 を用いた．また，本実験では周波数サブバンド毎にカートシスを算出し，その平均値を評価の対象とした．ただし，7 kHz から 8 kHz に音声成分はほとんど含まれていないため，本実験では評価対象から除外した．また，サブバンドは 16 とし，各周波数帯域幅は 500 Hz となる．

音声カートシス推定精度の評価には正規化誤差を用いた．正規化誤差は $e_{\mathrm{norm}} = |\mathrm{kurt}_{\mathrm{oracle}} - \mathrm{kurt}_{\mathrm{speech}}|/\mathrm{kurt}_{\mathrm{oracle}}$ で表される．ここで $\mathrm{kurt}_{\mathrm{oracle}}$ は真の音声カートシス，$\mathrm{kurt}_{\mathrm{speech}}$ は推定した音声カートシスである．

3.2.2　実験結果

実験結果を Fig. 3 に示す．Figure 3 より，ICA の学習回数の増加とともに音声カートシス推定の精度は向上することが確認された．また，ICA の学習回数が 400 回を超えたあたりで正規化誤差は収束することが確認できる．よって，以降の実験では ICA の学習回数は 400 回に設定する．

4　客観評価実験

4.1　実験条件

次に，提案手法および従来手法の比較実験を行った．実験では提案手法 1，2 および従来手法 1，2 の

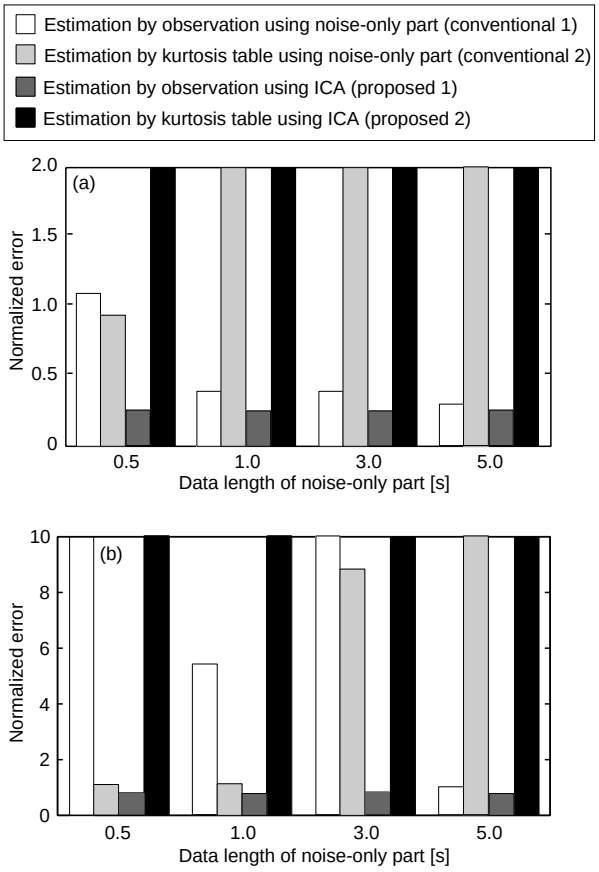

Fig. 4 Experimental results of speech kurtosis estimation: (a) number of subband is 16 and (b) number of subband is 32.

Table 1 Mean and variance of (i) estimation by observations using noise-only part (conventional 1), (ii) estimation by kurtosis table using noise-only part (conventional 2), (iii) estimation by observation using ICA (proposed 1), and (iv) estimation by kurtosis table using ICA (proposed 2).

Subband		(i)	(ii)	(iii)	(iv)
16	Mean	0.55	241	**0.26**	842
	Variance	0.10	153620	–	–
32	Mean	8.23	36.5	**0.79**	8493
	Variance	28.3	3238	–	–

計 4 手法を用いて，音声カートシスの推定精度について評価した．サブバンド数は 16 および 32 に設定し，非音声区間は 0.5，1.0，3.0 および 5.0 秒に設定した．その他の実験条件は 3.2.1 項と同様である．

4.2 実験結果

4 手法における音声カートシス推定の正規化誤差を Fig. 4 に示す．なお，Fig. 4 は話者 5 名の正規化誤差を平均した値である．また，Table 1 に手法毎の正規化誤差の平均値と分散値を示す．Figure 4 および Table 1 より，サブバンド数に依らず，提案手法 1 が最も安定して音声カートシスを推定できていることが確認された．また，サブバンド数を 32 に設定して推定を行った場合（Fig. 4(b)），非音声区間を用いて推定した音声カートシスは充分な推定精度を得ることはできなかった．これはサブバンド数が増えることでサブバンドあたりのサンプル数が減少したためである．また，提案手法 2 は音声カートシスを精度よく推定できなかった．これは，ICA で推定した雑音信号にはわずかに残留音声成分が含まれており，それが入力 SNR 推定精度の低下を招いたためである．よって，正しい入力 SNR に対するカートシステーブルを参照することができず，結果として音声カートシス推定精度の劣化を招いたと考えられる．以上より，提案手法 2 よりも提案手法 1 の方が優位であるということが確認された．

5 まとめ

本稿では，頑健な音声カートシス推定を目的として，ICA による推定雑音を用いた音声カートシス推定法を提案した．また，評価実験の結果より，提案法の優位性を示した．

謝辞 本研究の一部は科学技術振興事業団・戦略的基礎研究推進事業（CREST）の支援を受けた．

参考文献

[1] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, K. Shikano, “Blind spatial subtraction array for speech enhancement in noisy environment,” *IEEE Transactions on Audio, Speech, and Language Proc.*, vol.17, no.4, pp.650–664, 2009.

[2] P. Comon, “Independent component analysis, a new concept?,” *Signal Proc.*, vol.36, pp.287–314, 1994.

[3] S. F. Boll, “Suppression of acustic noise in speech using spectral subtraction,” *IEEE Trans. Acoustics, Speech, Signal Processing,* vol.ASSP-27, no.2, pp.113–120, 1979.

[4] R. Miyazaki, H. Saruwatari, R. Wakisaka, K. Shikano, T. Takatani, “Theoritical analysis of parametric blind spatial subtraction array and its application to speech recognition performance prediction,” *Proc. Hands-free Speech Communication and Microphone Arrays (HSCMA2011)*, pp.19–24, 2011.

[5] R. Wakisaka, H. Saruwatari, K. Shikano, T. Takatani, “Speech kurtosis estimation from observed noisy signal based on generalized gaussian distribution prior and additivity of cumulants,” *Proc. International Conference on Acoustics, Speech, and Signal Processing 2012 (ICASPP2012)*, pp.4049–4052, 2012.