E-012

カオス論的手法により音声信号に定義する発話者の覚醒度に相関する特徴量
Chaotic Characteristic Value Related to the Arousal Level of the Speaker Defined in Voice Signal

塩見 格一†
Kakuichi Shiomi

1. はじめに

人間の脳に関する研究は NIRS や fMRI 更には SPECT や PET 等の脳機能状態を視覚化する観測装置の普及により，嘗ての脳波程度しか観測できなかった状況に比較し大きく進展し，我々の脳に関する知見は急速に増大した。もっとも 21 世紀初頭からの約 20 年間がこのような状況ではあっても，脳については，未だに分からない事の方が遥かに多い状況であることもまた事実と思われる。今日，上記の脳機能状態観測装置がかなりの普及を見ているとは言え，微小な電磁波を観測するこれらの装置は比較的に高価であり，利用できる環境条件には多くの制限もあり，例えば，ビデオカメラで行動を撮影したり，IC レコーダで音声を収録したりするようには，誰もが簡単に利用できるものではない。

上記の脳機能観測装置により収録されたデータに対しては，今日既に，高度な機械学習の成果として実現された AI による（時に，人間の医師よりも正しい）診断が可能であるから，これらの観測装置が誰にでも日常的に利用できるようになれば，脳機能障害の予兆の発見等，個々人においてより高度な自己管理が可能となることが期待される。

しかしながら近い将来において上記の脳機能観測装置が家庭や職場に当たり前に存在する状況は想像し難く，そこで筆者は，安価で容易な脳機能状態の観測手法としての発話音声分析技術・装置を提案したい。脳の能動的な機能は身体を構成する筋肉の内の随意筋の制御だけであるから，発話音声信号は人間の脳が出力する最も信号雑音比に優れた生体信号と考えられる。声帯の基本周波数は男性の場合に 80 ~ 200 Hz 程度であり，女性では男性の倍程度である。日本語の場合には，言語的な意味を有している音声の周波数帯域は第２フォルマント周波数帯（~ 3 kHz）までであり，静かな部屋（騒音レベル 35 phon の部屋）で 70 ~ 80 phon の朗読音声を収録するとすれば，信号雑音比は 35dB 程度以上であり，他の生体信号に比較して遥かに高品位な生体信号としての音声信号を収録することが可能である。

2. カオス論的な特徴量

カオス論的な手法による時系列信号の分析は，特にリアプノフ指数の算出に関しては，1985 年の佐野・澤田のアルゴリズムの公表を契機として，広く行なわれるようになった[1]。時系列信号としての生体信号においては，脳波や指尖脈波の分析が行われるようになり，音声波形については，先ずは，数秒以上の継続時間を有する単母音の「あ～！」のような発話による音声信号波形の分析が行われた。

発話音声のカオス性に関する研究は，当初，多くは合成音声をより自然な発話音声に近付けるための知見等を得ることを目的として行われていたように思われるが，筆者は，人間の声とそれ以外の音（物がぶつかる時の音や，裂けたり，砕けたりする時の音）を区別する特徴量の発見を目的として開始した。1998 年当時，カオス論的な信号処理には従来の FFT 等に比較して数桁以上の演算処理を要したことから，数十分から 1 時間以上に及ぶ連続的な朗読音声の分析等は全く行われておらず，我々は，朗読の継続により，その朗読音声から切り出した音声データから算出される最大リアプノフ指数の時間的な平均値が，経時的に変化している事に気付く事になった[2]。

最大リアプノフ指数等のカオス論的な特徴量は，安定なダイナミクスが生成する時系列信号から再構成されるアトラクタにおいて定義可能なものであり，百数十ミリ秒程度しか継続しない一般発話中の母音波形による時系列信号によっては安定なアトラクタを再構成することは不可能であり，従って最大リアプノフ指数等を従来手法により計算することもできない。そこで，筆者等は「時間局所的な最大リアプノフ指数」と見做せる特徴量を定義し，発話者の覚醒度との相関が強くなるカオス論的な信号処理パラメータ（埋込次元，埋込遅延時間，他）の組み合せを探すことを目的として，仮説検証型の音声分析実験を行った[3]。

即ち，従来の音声分析技術の研究開発は音声データに人間の主観による属性（喜怒哀楽の表現，疲労の程度，等々）を付することから始められたが，筆者等は，強度の運動負荷や睡眠制限により強い眠気に襲われる状況を実験的に実現し，その都度の状況において発話音声を収録し，算出される特徴量の変化が，覚醒度指標としての臨界フリッカ周波数の変化と相関するように[4]，また昼食の前後や，眠気を誘う薬物服用の有無，更には等々の心身状態の差異が期待される複数の状況において収録された音声の相互識別の信頼性が高くなるように，信号処理パラメータの最適化を目指した。

一定の標本化周期で標本化された音声データは図１に示すような時系列信号 $f(t_n \mid n = 1, 2, \ldots)$ であって，ターケンスの定理により適正な埋込次元と埋込遅延時間 (τ_d) を設定すれば，図２に示すように位相空間上に

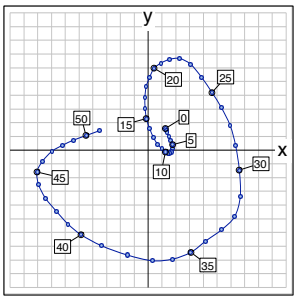

図２　ターケンス・プロット

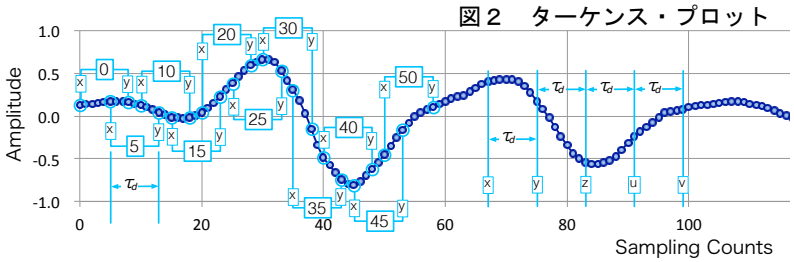

図１　音声信号波形（「お」音の一部）

† 福井医療大学　Fukui Health Science University

Copyright © 2019 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

ターケンス・プロットとしてストレンジ・アトラクタを再構成することが可能であって，そのストレンジ・アトラクタからは最大リアプノフ指数を計算することができる[5]。

図２は埋込次元を２次元として２次元平面に対して埋込点を $P_n(f(t_n), f(t_n - \tau_d))$ としてストレンジ・アトラクタの再構成手法を模式的に示したものであって，実際の音声信号に対しては４次元より高次の埋込空間に，例えば埋込次元を５次元とすれば埋込点を $P_n(f(t_n), f(t_n - \tau_d), f(t_n - 2\tau_d), f(t_n - 3\tau_d), f(t_n - 4\tau_d))$ としてストレンジ・アトラクタを再構成して，時間局所的な最大リアプノフ指数を計算する。この例は「お」音の信号波形から再構成したものであり，図２のターケンス・プロットにおいて埋込点を増やせば図３のようにストレンジ・アトラクタが再構成される。P_n, P_{n+1}, P_{n+2},…は時間的に標本化周期だけ離れた埋込点列であり，P_m は埋込点 P_n に対する近傍点であり，$\overline{P_nP_m}$, $\overline{P_{n+1}P_{m+1}}$, $\overline{P_{n+2}P_{m+2}}$, … と近傍点集合の外包円径が増大しており，これらの距離から時間局所的な最大リアプノフ指数を計算する。時間局所的な最大リアプノフ指数は全ての埋込点に対して計算し，これを統計処理してフレーズや一連の発話に対するマクロスコピックな「発話者の覚醒度に強く相関する特徴量」を算出する。

図４は，音声資源コンソーシアムの提供する AWA-LTR コーパスの内の A46 朗読音声の午前 10 時に収録した 52 レコードと午後 13 時に収録した 52 レコードの合計 104 レコードを処理した結果で，特徴量としてのカオス論的な指数値を計算し，その平均値（0.0）と標準偏差（1.0）を尺度として，その指数値の分布を示したものである。午前 10 時における 52 回の朗読音声から算出された指数値の平均値は 0.700，標準偏差は 0.865 であり，午後 13 時における 52 回の朗読音声から算出された指数値の平均値は-0.700，標準偏差は 0.522 であった。全 104 レコードから１つを任意に選んだ場合に，昼食の前後の識別が正しい確率は約 88%と期待される。

図４からも見られるが，カオス論的な指数値の特徴として，この指数値の分布は平均値を中心とした対称なものではなく，信号処理パラメータの設定に依らず，多くの場合にガンマ分布のように，大きな値の側の裾野が長くなる。

3. おわりに

機械学習に係る情報処理技術の発展は驚くべきものであり，シンギュラリティを待つまでもなく，囲碁や将棋等のボードゲームにおいて人間がコンピュータに勝利することは粗不可能になり，医療用の AI は医者よりも信頼性の高い診断を下す状況になっている。発話音声の分析においても，発話者の鬱状態の程度の評価等を目的とした特徴量の探索等が試みられており，成果が期待されている。

しかしながら如何に機械学習による特徴量の検出が進んだとしても，音声分析に関しては，ソノグラム上のパターンとして現れる特徴量のみを対象としていたのでは，その成果は些かの感情の分析等を可能にする程度の限定的なものに留まってしまうと危惧され，また人間の主観に頼ってラベル付けされた音声データを基にする限り，原理的にも人間の感覚や能力を超越することは不可能ではなかろうか。

筆者の定義したカオス論的特徴量によれば，発話者の覚醒度の評価において，ソノグラム上のパターンとしての特徴量等では実現し得なかった高い信頼性を実現している。

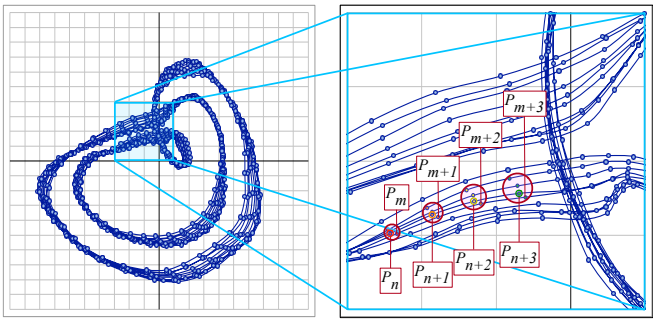

図３　埋込点と近傍点

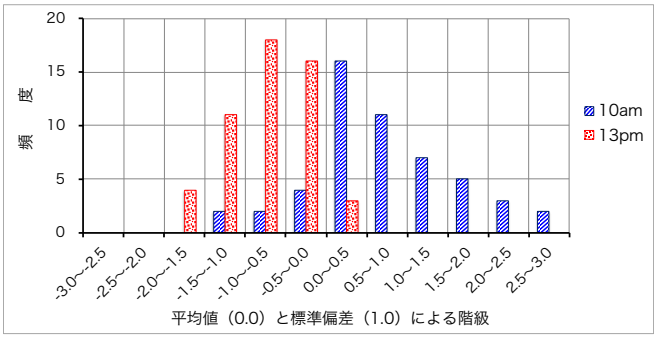

図４　昼食の前後における覚醒度の差異を観測した例

機械学習により発話音声信号から何等かの特徴量の発見を目指すのであれば，その特徴量探索の範囲はカオス論的な手法を利用して定義可能な特徴量の存在する空間まで拡張することが，より合理的であると考えられる。現時点において，筆者の定義したカオス論的特徴量の差異は，ソノグラム上に視覚化すること等が未だ不可能であり，これを人間の耳で識別することも，また何等かの練習により特徴量を演出することもできない。従って，原理的にも発話者の演技や演出が排除されており，結果的に客観的信頼性の高い特徴量となることが期待される。

実験的に音声信号のフラクタル次元は４より大きいと考えられており，このことは発話行為が脳の２つの機能の連携，あるいは相互作用によることを予言しており，現時点における脳の発話メカニズム（複数の脳機能野の連携）との整合性が期待される。これから機械的な手法により音声信号のビッグデータから，何等かの属性を定量的に評価する特徴量の発見を目指す方々には，その特徴量の探索空間をカオス論的な手法による特徴量が含まれる領域まで拡張されることを強く勧めたい。

参考文献

[1] M. Sano, Y. Sawada, “Measurement of the Lyapunov Spectrum from a Chaotic Time Series,” Physical Review Letters, Vol.55, No.10 (1985).

[2] K. Shiomi, S. Hirose, “Fatigue and Drowsiness Predictor for Pilots and Air Traffic Controllers,” Proc. of 45th ATCA Conference, (2000).

[3] K. Shiomi, “Voice Processing Technique for Human Cerebral Activity Measurement,” Proc. SMC2008, P0600 (2008).

[4] K. Shiomi et al., “Experimental Results of Measuring Human Fatigue by Utilizing Uttered Voice Processing,” Proc. SMC2008, P0557 (2008).

[5] F. Takens, “Detecting strange attractors in turbulence”, in Lecture Notes in Mathematics, No. 898 (Springer-Verlag, 1981).

Copyright © 2019 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.