可視化情報 Vol.20 Suppl. No.1 (２０００年７月)

309 ウェーブレット画像処理法による音声認識

妹尾 勇○, 早野 誠治, 斎藤 兆古 (法政大学)

Voice Cognition by Wavelets Image Processing

I.SENOO,S.HAYANO and Y.SAITO

ABSTRACT

As is well known, voice cognition system of the personal computers becomes one of the popular command input methodologies. However, it is still remaining problem, which requires a hard training in order to build up a reliable voice database. Most of the voice cognition methods are based on a simple correlation analysis between the input and database voices. This leads to a relatively low cognition rate when using a poor trained database system. To overcome this difficulty, we are now developing a new voice cognition methodology, which cognizes an input voice as a solution of an inverse problem. Our inverse methodology requires each of the characteristic patterns representing essential feature of the single pronunciation. In the present paper, we describe that the characteristic patterns representing essential feature of the single pronunciation are derived by means of the simple Lissajous's diagram after the threshold operations. Using the database containing the characteristic patterns, an ill-posed linear system of equations for an each pronunciation input is established. Simple least squares leads to a good approximate solution of this system. Thus, we have succeeded in cognizing the single pronunciation. Further, we have tried to compress the characteristic patterns by means of the discrete wavelet transform. As a result, it is revealed that the Daubechie's 2^{nd} order base function makes it possible to reduce the database into 25% quantity.

Keywords: Voice Cognition, Eigen Pattern, Method of Least Squares

1. まえがき

パーソナルコンピュータの広汎な高性能･小型化・低価格化に伴い、パーソナルコンピュータは高機能文房具として急速な普及を遂げている。パーソナルコンピュータは、単純な電卓などのように単機能機器でなく、入力指令に基づいて複雑な作業も可能な機器である。複雑な作業を行える反面、パーソナルコンピュータのユーザは複雑な命令を使いこなす知識が必要である。しかし、ゲーム専用計算機で見られるように直感的操作が可能なGUI(Graphical User Interface)を用いることで、直接的な命令を生成する知識は不必要となりつつある。しかしながら、何らかの入力装置、例えばジョィスティツクなどの装置を操作する技術は必要である。

このような、機械と人間のインターフェースを改善する究極の方途は、人間の情報伝達方法として使われる、音声認識と考えられる。このような観点から、既に音声認識ソフトは商品として販売されている。音声認識の問題点は、人間が機械でなく生物であることから必然的に新陳代謝を伴うため、同一人物でも全く同じ音声を生成出来ない点にある。しかし、我々人間は音声を用いて風邪を患っている条件下でも意思伝達が可能である。計算機にこのような人間の音声認識が可能ならしめようとする技術が音声認識である。

本論文では、風邪を患っている条件下でも計算機へ命令伝達が可能である究極の音声認識技術の開発を目指し、音声の固有の特徴を抽出する固有パターン法を提唱する。音声の固有パターンとは,どのような条件下においても共通な音声の特徴を言う。人間は幼児期からの長い訓練で音声認識の技術、すなわち、固有パターンを抽出する技術を体得する。本論文では、音声の時間依存性をリサージュ図形で削減する固有パターン抽出の一方法を提案する。日本語50音中の45音の固有パターンを生成し、固有パターンから線形システムのシステム行列を生成する。この線形システム行列と入力音声の固有パターンを用いて、音声認識問題を線形システム方程式の解問題へ定式化する。当然ながら、得られる線形システムは不適切であり、厳密な解ベクトルを得ることが困難である。本論文では、最小自乗法を用いて近似解を導く。固有パターン生成に採用した音声は完全に最小自乗解から認識可能であることが判明した。また、実用化を勘案して、固有パターンへ離散値系直交ウェーブレット変換を適用してデータ量を削減する方法について検討した。その結果、ドビッシーの2次基底関数を用いたウェーブレット変換は、音声認識機能を低下することなく、固有パターンのデータ量を25%へ削減可能とすることを報告する。

2. 固有パターン法による音声認識

2.1 固有パターンの生成

音声の固有パターンとは、音声が持つ固有の不変量が生成するパターンである。言葉で固有パターンを定義するのは簡単である。しかしながら、計算機命令へ化すような具体的記述は不可能に近い。これを可能な限り具体化する技術が音声認識である。個々の音声が持つ固有の特徴は、以下のように、大まかに列記できるであろう。1）周波数に無関係である。これは同じ音声でも低音と高音があることに起因する。2）固有パターンは発声の長短に依存しない。3）音声は必ずしも音声の固有パターンのみで生成されない。これは、情報伝達の基本信号以外に個体差に起因する情報が音声に包含されるためである。すなわち、音声には個性があり、個性に起因する要素を削除した音声が固有パターンを生成する。

具体的に上記の条件を満足する信号情報処理は簡単でない。本論文では、音声の周波数依存性1）や発声の長短依存性2）を削除するため、時間パラメータを削除する音声信号のリサージュ図形を生成する。時間領域信号のリサージュ図形は、x-y直交座標系で、x軸方向の座標値を原音声信号、y軸方向の座標値を原音声に対して時間位相が90度異なる信号でそれぞれ決めたx-y平面上の軌跡である。この意味で、1次元時間軸情報を2次元平面座標上へ可視化した映像情報である。次に個体差3）の情報をリサージュ図形の原点付近の情報と仮定し、閾値処理で削除する。閾値処理は必然的に経験的要素を必要とするため、好ましくない。しかし、本論文の目的は固有パターン法の着想を検証することを意図しているため、ある程度の経験的ファクターを勘案せざるを得ない。

リサージュ図形を生成するには、原音声信号以外に時間位相が90度異なる音声信号が必要である。時間位相が90度異なる信号は2種類生成可能である。一方は時間位相が90度進んだ信号であり、原信号を時間軸方向へ積分することで得られる。他方は時間位相が90度遅れた信号であり、原信号を時間軸方向へ微分することで得られる。一般に微分演算はノイズ若しくは個体差情報を拡大するため、本論文では原信号を時間積分して、原信号よりも90度位相の進んだ信号を生成して、リサージュ図形を生成する。

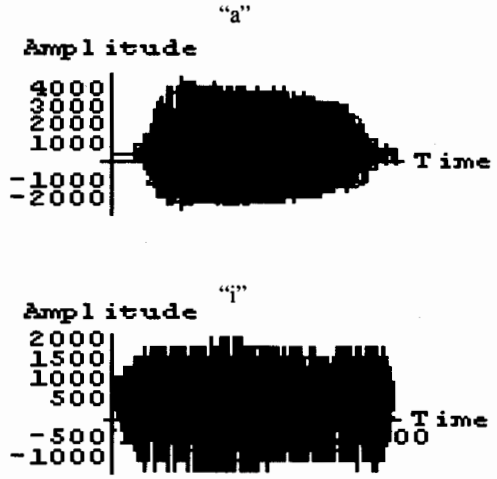

Fig. 1 Waveform of the sounds "a" (upper) and "i" (lower) signals

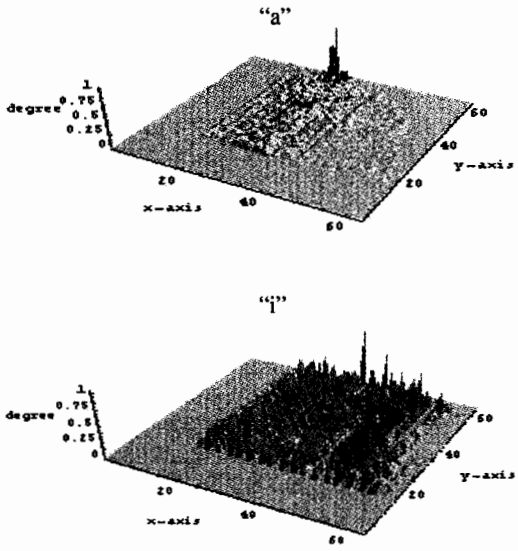

Fig. 2 Characteristic patterns of the voice signals "a" (upper) and "i" (lower)

Fig.1は母音「ア」と「イ」の原信号波形である。Fig.2は画素数を64x64とし、閾値を最大振幅の25％として作成した「ア」と「イ」に対する固有パターンである。尚、Fig.2で、固有パターンのz軸方向の高さは同一座標値の重複度である。明らかに、音声「ア」と「イ」では完全に異なる固有のパターンが生成されることがわかる。

2.2 システム方程式

Fig.2に示す固有パターンは、それぞれ64x64画素からなるため、1次元配列に並べ替えると、64x64次のベクトルとなる。このようにして得られるn個の固有パターンベクトル$\mathbf{c}_i$, i=1,2,...,nを使って、（1）式からn行64x64列の長方システム行列Cが構成できる。

$$C = [\mathbf{c}_1, \mathbf{c}_2, .., \mathbf{c}_n] \tag{1}$$

いま、任意の音声信号の固有パターンを1次元配列へ並べ替えて得られる入力ベクトルを$\mathbf{Y}$とすれば、解くべき線形システム方程式は（2）式で与えられる。

$$\mathbf{Y}=C\mathbf{X} \tag{2}$$

（2）式で解ベクトル$\mathbf{X}$の要素を

$$\mathbf{X}= [X_1, X_2, \ldots, X_n] \tag{3}$$

とすれば、最大値を取る要素が識別された音声となる1,2)。

（1）式の固有パターンベクトル、それぞれに対応する原音声信号を、$\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n$とすれば、（3）式の解ベクトル$\mathbf{X}$から生成される音声信号$\mathbf{G}$は（4）式で与えられる。

$$\mathbf{G}=\sum_{i=1}^{n} X_i \mathbf{V}_i \qquad (4)$$

2.3 最小自乗解

（2）式のシステム方程式はｎ個の未知数に対し、64X64個の式の数であり、64X64>ｎとすれば、全ての式を同時に満足する解は特別な場合を除いて存在しない。このため、誤差ベクトルのノルム

$$\varepsilon=\left|\mathbf{Y}-C\mathbf{X}\right| \qquad (5)$$

を最小にする解ベクトル、すなわち、最小自乗法による解ベクトルを（6）式から計算する。

$$\mathbf{X}=\left(C^T C\right)^{-1} C^T \mathbf{Y} \qquad (6)$$

五十音中の「ア」から「ロ」の45音に対する固有パターンを生成し、（6）式の解ベクトルを計算した。その結果、システム行列生成に採用した原音声を入力とした場合、全てを完全に認識できた。Fig.3に音声「ア」と「イ」に対する解ベクトルの例を示す。Fig3の例では、認識された要素の値は1であり、他はすべてゼロである。明らかにこれらの解は（2）式を厳密に満足する特別な例である。換言すれば、システム行列生成に採用した原音声を入力とした場合、（2）式を完全に満足する解がえられる。

2.4 ウエーブレット変換による固有パターンの圧縮

Fig.2に示す音声の固有パターンを2次元映像情報として、離散値系直交ウェーブレット変換の領域法を用いて25％のデータ量へ圧縮する。2nd orderのドビッシーの基底関数を用いたいウェーブレット変換で25％のデータ量へ圧縮されたデータから再現された”ア”と”イ”に対する固有パターンをFig.4に示す。但し、採用した基底関数はドビッシ-の２次である。圧縮したデー

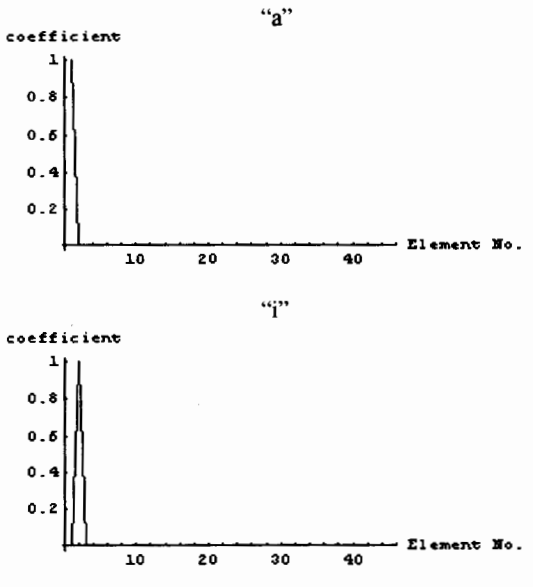

Fig. 3 Solution vectors for the input voices

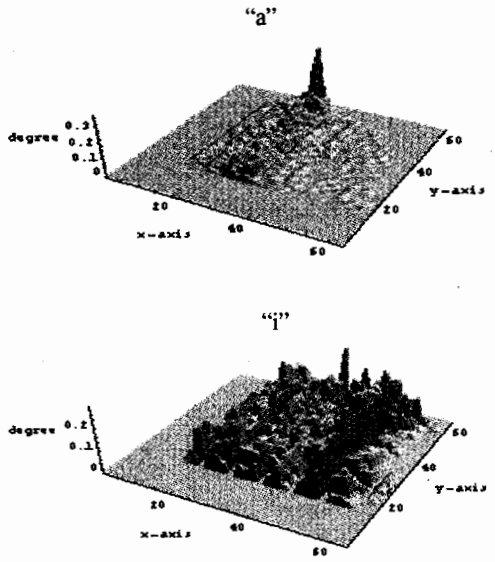

Fig. 4 Recovered characteristic patterns of the voice signals "a" (upper)and "i" (lower) from the compressed 25% data by means of the Daubechie's 2 nd order base function.

タから最構成された固有パターンを用いて（6）式の最小自乗法による解を得る。Fig.5はドビッシーの2次基底関数を用いて圧縮された固有パターンから得られた音声「ア」から「オ」に対する解ベクトルの例である。それぞれの最大値を取る要素が認識された音声に対応するから、全ての音声が正確に認識できた。しかし、Fig.3に示す解ベクトルと異なり、（2）式のシステム方程式は完全に満足できない。ドビッシーの高次基底やCoifmanやBaylkinなどの基底関数を用いて同様の計算を行ったが、ドビッシーの2次基底関数以外は全てを正確に認識できなかった。この理由は、Fig.2の固有パターンを観察すれば、固有パターンの特徴は比較的ピーキィな値の分布に拠るため、ウェーブレット変換の領域法はこのピーキィな特徴を削減し認識率の低下を喚起していると考えられる。ドビッシーの2次基底関数の波形は矩形波状であるため、比較的均一にピーキィな高周波情報を保持する。このため、他の基底に比較して最も良い認識率を与えたと考えられる。

3.まとめ

本論文では、音声情報から音声固有の情報を抽出した固有パターンの概念を提案し、固有パターンの具体的抽出法の一方法として、音声情報のリサージュ映像情報を生成した。抽出した固有パターンを使って線形システム方程式を導き、音声認識問題を線形システムの解析問題へ化した。線形システム方程式の近似解ベクトルを得る一方法として最小自乗法を採用した。その結果、音声「ア」から「ン」45音声信号の完全な識別が可能であることが判明した。また、固有パターンを離散値系ウェーブレット変換の領域法で圧縮し、小規模のデータベースから音声認識の可能性を検討した。

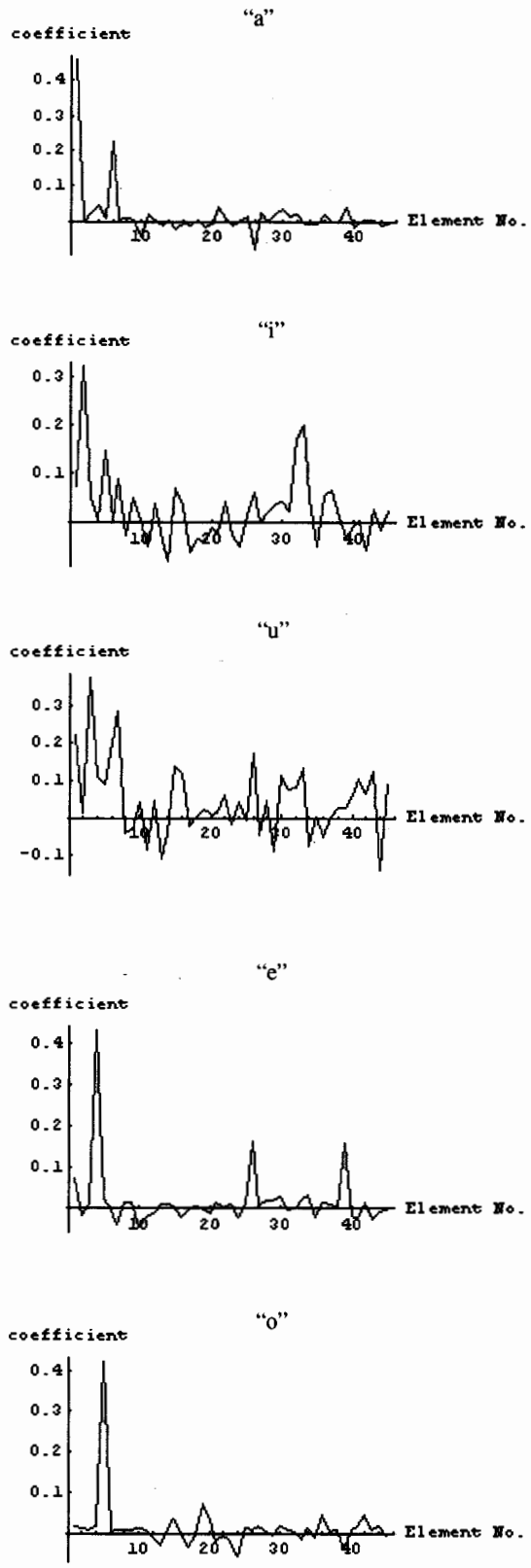

Fig. 5 Solution vectors for the input voices "a","i","u","e" and "o" where the characteristic patterns have compressed into 25% data quantity by means of the Daubechie's 2nd order base function.

その結果、ドビッシーの2次基底関数は25％のデータから100％の認識率を確保できることが判明した。

本論文の主要な目的は、音声固有のパターンを提唱し、音声認識の可能性を検討する点にある。この意味で、所期の目的は達成できたと考える。

参考文献

1) 斎藤兆古　著、ウェーブレット変換の基礎と応用（朝倉書店、1998年4月）

2) H. Takahashi, S. Hayano, Y. Saito, "Visualization Of The Currents On The Printed Circuit Boards", *IEEE* Visualization 1999, Late Breaking Hot Topics, pp. 37-40, Oct. 1999