K-011

音響構造モデルの提案とスペクトル推定成分との対応

Primary Modeling of Acoustic Structure

吉田 秀樹†‡　Xie Wei‡　藤原 祥隆†‡
Hideki Yoshida†‡　Xie Wei‡　Yoshitaka Fujiwara†‡

1. まえがき

音波は疎密波によるエネルギーの伝播であり、複数の音源から放射（反射）されたエネルギーは、位相が偶然重なり合って強め合うこともあれば、逆に弱め合い消失することもある。音源が生体起源であれば話は非常に複雑で、先ず信号源自体が一様には働かない。次に調音器官には個人差が著しく、調音器官を時々刻々と形状変化させることで、音響エネルギーに様々な情報を暗号化し添付させている。音響エネルギーの伝播路は非線形であり、エネルギーが伝わり易くもなれば、伝わり難くもなる。更に難解であるのが、聴性認識機構、即ち暗号解読のメカニズムである。意識して聴き耳を立てている状態、浅い睡眠の中で無意識の内に聞いている状態、意識はあるのに人の話を聞いていない状態と、音響エネルギーを受信する側すら、精神状態によって知覚の度合いを時々刻々と変化させている。以上の困難にも関わらず、人にとって最も簡易で満足のいくコミュニケーション手段は、音声手段に頼ったものである。極めて不安定かつ不確定な状況の中で、生物起源の音響と非生物起源の音響の識別、言語的な情報と非言語的な情報のコミュニケーションを成立させている事実は驚異である。

音響コミュニケーションの解明と応用へ向けた取り組みには枚挙に暇が無い。局所的に見れば線形システムであり、かつ伝達される信号も定常波に近似できることを前提にして、音響波動を、無限長の三角関数を無限に重ね合わせて理解する数理モデルが基礎とされた[1-3]。時間・周波数分解能も格段に向上し[4-5]、音声認識技術を始めとする最先端の IT 技術が花開いた。音響波動による全てのコミュニケーションは、顕著なスペクトルの集合で表わせる統計的な現象として理解され、正確無比であることに疑いの余地は無く、工学への貢献は今もめざましい。

我々の研究目的は、生きた人間である聞き手の立場に立ち、かつ、電算機処理にも適した音響モデルを提案することにある。如何なる複雑な形状をした音響波形であっても、分解することで、音響構造が観察できるようになる。音響を理解し易いものとするために、最短時間だけ持続する音響エネルギーを仮定した。仮定した音響エネルギーは、組み立て式のブロックの様な操作ができることを意識した。第２の仮定は、ヒトの聴覚能が、関心のある音響エネルギーと、無視している音響エネルギーの自動選別を実現していると云った立場をとることにした。音響構造を理解した後で、統計処理を経て初めて、音響波動に巧みに暗号化された全情報を紐解くことができると考えた。最後に、提案した音響モデルと従来のスペクトル推定値との対応を考察する為に、小規模ながらデータの収集を試みた。

†北見工業大学情報システム工学科
‡北見工業大学サテライト・ベンチャー・ビジネス・ラボラトリー

2. アルゴリズム

入力信号を 1 オクターブ帯域の帯域通過フィルターに通す。帯域通過フィルターは、観察する周波数帯域を網羅できる様に複数使用した、いわゆるフィルターバンク構成とする。フィルター処理した波形について、極大値から極小値までの時間，および極小値から極大値までの時間 t_e を計測する。各 t_e 時間の波動断片を、正弦波の一部として近似する。（近似可能な区間もあれば、実際には、誤差の無視できない区間も混在している。）時間 t_e を２倍して逆数をとった値を、近似のために使用した正弦波の周波数（瞬時値）とみなす。本計算により、如何なる複雑な波動も、ベクトルの時系列として表現でき、ベクトルの成分数は、フィルターバンクに使用したフィルター数（チャンネル数）と等しくなる。

3. 音響モデル

図１(a)は音声波形の一部であり、母音[i]を発音した区間である。図１(b)は、図１(a)の音声波形を帯域通過フィルターに通した後の波形であり、分解され情報は失われて、歪んだ正弦波状をしている。ここで通過帯域幅は 512 Hz から 1024 Hz までの 1 オクターブとした。波形を時間軸方向に細かく分割していき、隣り合う極値から極値までを最小単位とする考えに至った。するとフィルター処理した波形は、正弦波を振幅変調、周波数変調して造り出した波形で近似できるのではないかと考えた。元来、周波数は正弦波に対して定義されているので、歪んだ波形の極値から極値までの微少区間について定義できる性質のものではない。近似に使用した正弦波からの誤差が、高調波のエネルギーを生み出すことになる。この相違がヒトの聴覚心理に影響を与える様であれば、提案したモデルは意味をなさなくなる。仮定が成立する場合、保存すべき情報はそれぞれの極値となるであろう。この様子を図１(b)の波形に黒丸で示した。極値から次の極値までの時間変化は、再構成音の聞き心地に影響を与えることの無い様に、後から適切に補間される技術が求められる。図１(c)に、提案する音響モデルを示した。空間に、実際には音響エネルギーが存在しない時間があっても、ヒトの聴覚識別閾値以下の音響エネルギーが存在すると仮定した。するとフィルター処理後の波形は、極値から極値までの微少区間を最小単位とする音響エネルギーが、無限に数珠つなぎに連なるモデルとして表現される。個々の微少区間のエネルギー量は、振幅（瞬時値）と密接な関係がある。微少区間の変化やゆらぎ具合が、ヒトには音色として知覚されることになる。

4. スペクトル推定との比較実験

詳細は文献[6]に譲る。男性被験者 7 名について、椅子に座って安静にしてもらい、「良い香りだ」と 3 回繰返して発話させた（コントロール）。酢臭を 90 秒間嗅いでも

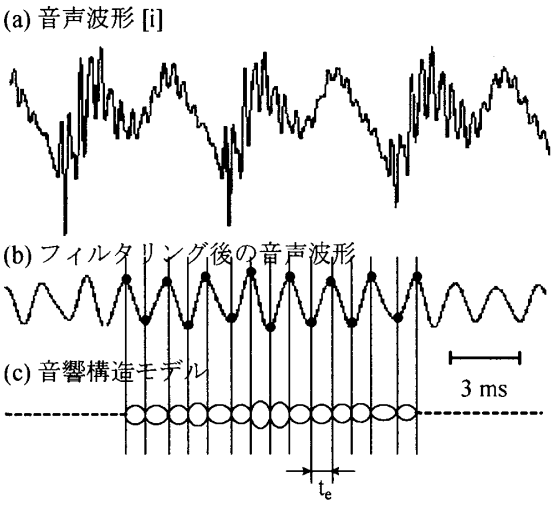

図１　音響構造モデルの生成。(a) 音声波形の一部であり母音の[i]に相当する。(b) フィルタリング後の波形。512 Hz から 1024 Hz を通過帯域幅とするフィルターを使用した。(c) 音響構造モデル。音響エネルギーの基本単位を白丸で表した。持続時間は隣り合う極値から極値までの t_e で表した。音響構造モデルは、音響エネルギーの基本単位を無限に連結したものとして表した。

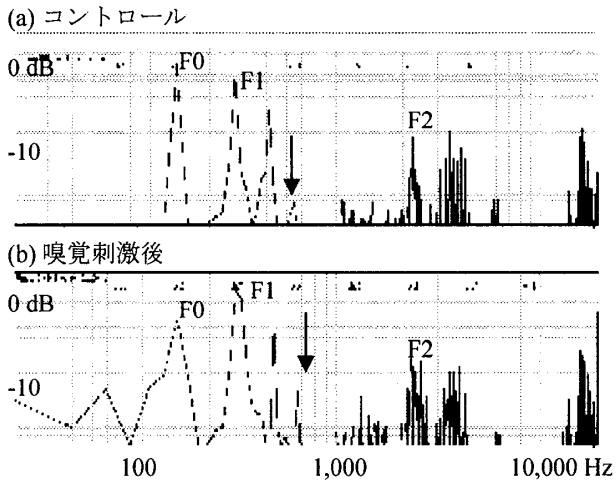

図２　採取した音声「良い香りだ」の中で、母音[i]に相当する箇所を切り出して、パワースペクトルを算出した。窓長 51.2 ms のハミング窓を使用した。(a) コントロール時に採取した音声のパワースペクトル。(b) 嗅覚刺激後に採取した音声のパワースペクトル。

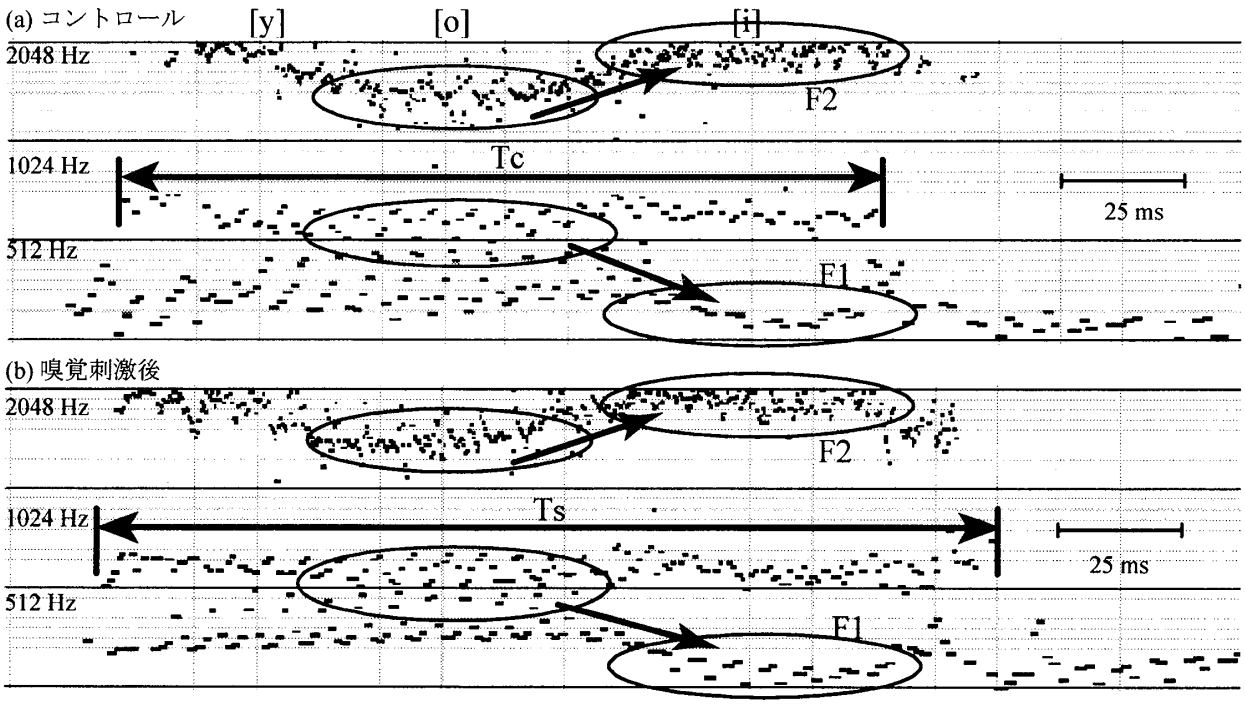

図３　採取した音声「良い香りだ」の中で、「良い」に相当する箇所を切り出して、提案した周波数分析方法を使用して分析した。黒い小さな点の１つ１つが、想定した音響エネルギーの基本単位を表している。(a) コントロール時に採取した音声。(b) 嗅覚刺激後に採取した音声。着目した 500 Hz から 800 Hz の周波数成分は、コントロール時には Tc 期間、嗅覚刺激後には Ts 期間で観察された。

らい、同じく「良い香りだ」と３回繰返して発話させた。中立な発話と発話速度を心掛けてもらった。収集した音声データには、主観的ながら、相当注意して聴けば聴き取れる程度の僅かな印象の相違を得た。非言語的な情報を可視化するために、窓長 51.2 ms のハミング窓を使用した高速フーリエ変換結果と、提案手法を使用した周波数分析結果を算出して比較を試みた。

5. 結果

図２に高速フーリエ変換の結果を示した。図２(a)がコントロール、図２(b)が嗅覚刺激後の音声のパワースペクトルである。分析に使用した音素は「良い香りだ」の中の[i]を選んだ。ここで価値判断を表す言葉「良い」の中に変化が観察されていたことは、先の報告に記した通りである。高速フーリエ変換と、これを基礎にして改良が施された様々な解析手法を駆使して発見された特徴パラメータは、基本周波数(F0)と、調音器官を通じてエネルギーが強められたフォルマント(F1, F2)と呼ばれる成分であった。ここで母音[i] の第１フォルマントは 180 Hz から 550Hz の範囲に、第２フォルマントは 1,900 Hz から 3,300 Hz にあることが知られており[7]、図２の結果は順当なものと考えられる。

図３には、提案した周波数分析方法を用いて作成した散点図を示した。図３(a)がコントロール、図３(b)が嗅覚刺激後の音声「良い」について算出した、第一段階の音響構造モデルとなる。母音[o]と[i]の第 1、第２フォルマントに相当する周波数成分の時間変化が観察することのできる、３つのチャンネル（256－512 Hz、512－1024 Hz、1024－2048 Hz）が選ばれた。ここで母音[o] の第１フォルマントは 450 Hz から 850Hz の範囲に、第２フォルマントは 500 Hz から 1,400 Hz にあることも知られている。併せて考えると、図３の中で、楕円で囲まれた点（音響エネルギーの最小単位）の集中領域が、フォルマントに相当していると考えるのが自然である。母音[o]から[i]への音素の遷移は、点の集中領域の滑らかな時間変化として表されていた。

6. 考察

我々は本研究に先立って、コントロール時と嗅覚刺激後の音声の相違として、500 Hz から 800 Hz の周波数成分に着目していた。同成分は発話に同期して出現していることから、実験室の雑音とは考え難かった。同成分の出現期間を、図３の中で、両端に矢印を付けた期間として表した。図３に記された小さな点の１つ１つが、音響エネルギーの最小単位を表現していることから、エネルギーは小さいながらも、確かに 500 Hz から 800 Hz の周波数帯域に存在していたことになる。観察した周波数帯域は、母音[o]の第１フォルマント周波数と重なり合うが、母音[i]を特徴付けているフォルマント周波数からは外れていた。実験で得られたそれぞれのデータについて、「良い」の発声時間 Tyoi に対する、500 Hz から 800 Hz の周波数成分の出現時間（図３(a)には Tc、図３(b)には Ts で表記）の比をとってみた。すると、比の値は、嗅覚刺激後で有意な増大が観察された（「良い」の平均発音時間 208±24 ms, 比の値 Tc/Tyoi=0.416±0.064（コントロール時の平均と s.d.），Ts/Tyoi=0.535±0.087（嗅覚刺激後），t=4.19, $p<0.01$, student t-test）。図２には相当するパワースペクトルが、フォルマントに比べると小さく出現していた（下向きの矢印を参照）。音声認識は、フォルマントを特徴パラメータとして実現されることから考えても、図２の下向きの矢印で示した信号成分が検討される必要性は薄く、これまでノイズとして無視されてきたと考えても不思議ではない。

本研究でも使用したフィルターバンクシステムは、従来はフィルター処理後、半波整流、平滑化し、パワースペクトル包絡の良好な近似として利用されてきたものである。歴史的にみて、半波整流作用は、内耳の蝸牛の中にある内有毛細胞の電気生理学的な特性を[8]、ダイオードの電気特性と等価とみなしたことによる。平滑作用の方は、神経に固有の膜キャパシタンスに由来したものと考えられる。内有毛細胞の作り出す電位の報告[9]によると、聴覚刺激に使用した音波と等しい周波数の交流成分が観察された。しかしながら、聴覚刺激の周波数が約 3,000 Hz 以上となると、恐らく膜キャパシタンスの影響が無視できなくなり、平滑されてスローウエーブ様に変化していた点が興味深い。本研究では、フィルタリング後の出力波形を敢えて観察することで、音響のモデル化を試みた。

7. まとめ

音響エネルギーの基本単位を仮定した。積み木の様なイメージとして表現される音響エネルギーが、無限に連結するとした音響構造モデルを提案した。モデルは、従来のフォルマントに相当するエネルギーの集中状態の他、非定常なエネルギーの分散状態も可視化した。例えば、500 Hz から 800 Hz の周波数成分とは、２つのフォルマントの間に存在する微弱信号の存在を示唆しており、ヒトの聴覚心理にどのような作用を及ぼすかについては解明が待たれる。

参考文献

[1] Cooley, J.W. and Tukey J.W. An algorithm for the machine computation of complex Fourier series, Math. Comput. 19, 297-301, 1965.

[2] Wiener, N. The Fourier integral and certain of its application, Dover, NewYork, 1958.

[3] Grossmann A. and Morlet J. Decomposition of Hardy functions into square integrable wavelets of constant shape, SIAM J. Math. Anal. 15, 723-736, 1984.

[4] Kawahara, H., Katayose, H., de Cheveigne, A. and Paterson, R.D. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and Periodicity, Proc. Eurospeech 99, 2781-2784, 1999.

[5] Nguyen, P.C., Takao, O. and Akagi M. Modified restricted temporal decomposition and its application to low rate speech coding, IEICE Trans. Inf. & Syst. E86-D(3), 397-405, 2003.

[6] 吉田秀樹、後藤晃一、岡田信一郎、藤原祥隆　匂い刺激が音声の非言語的な側面に与える影響の可視化 FIT2003 一般講演論文集第３分冊 565-567, 2003.

[7] Nakagawa S. et al., Diffrences in feature parameters of Japanese vowels with sex and age, Studia Phonologica, XIV:33-52, 1980.

[8] Klatt, D.H., Speech processing strategies based on auditory models. In The representation of speech in the peripheral auditory system, R. Carlson and B. Granstrom (eds.). Amsterdam: Elsevier, 1982.

[9] Palmer, A.R. and Russel, I. J., Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. Hearing Research 24, 1-15, 1986.