平成 10 年 10 月 1 日

研究テーマ
「人間の聴覚特性を考慮した音声信号処理」

本文で紹介する研究は、最終ページに示すように、筆者が NTT 基礎研究所（1984.4-1986.9, 1990.11-1992.3）、ATR 視聴覚機構研究所（1986.10-1990.10）、北陸先端科学技術大学院大学（1992.4-）で行った研究の一部をまとめたものである。

まえがき

人間が相互にコミュニケーションを行う場合、言葉を発して相手に自分の考え、感情などを伝えようとする一方で、相手が伝えてきた考え、感情などの情報を受け取り、理解して、そして、適切な応答を行う手段が必要である。自分自身の中でこのサイクルが上手くまわることによって、コミュニケーションが保たれる。このサイクルのことを“ことばの鎖”（図1）と呼んでいる。機械による音声認識は、「ことばの鎖の中の音声知覚過程を工学的に実現する一つの応用問題」、と言うことができる。将来的に“ことばの鎖”がすべて機械の上に実現され、機械と違和感なくコミュニケーションが行われる状況が来るためには、この応用問題を解かなければならない。

そこで、「言葉を話す、言葉（音）を聞くは人間の営みである」という原点に立ち返り、機械による音声認識において問題となっている様々な点について、人間の優れた特性から問題解決のヒントを得て、これを解決しようとする研究が行われている。これは、人間には簡単であるが機械は苦手な次のような問題について、音響心理／生理の知見に基づいて解決を試みるアプローチである。

(1) 調音結合の補正、
(2) 雑音環境下での信号音の抽出、
(3) 個人性の制御

このような研究を行うためには、まず、(a) 解剖学、生理学、心理学から得られた聴覚におけるの知見を整理し、工学的に応用できるかどうかを見極める必要がある。また、知見が不足している場合は自らが測定を行う。そして、(b) 工学的に応用できるものが見つかれば、これを基にして働きを機能的に模擬するモデルを構築する。この場合に、聴覚特性の関数解析的モデル化とこれを計算機上に実装するディジタル信号処理技術が必要である。最後に、(c) 構築したモデルを音声認識、音声分析・合成へのシステムに適用し、その有用性を検討する、ことが必要である。すなわち、このような研究を行うためには、図2に示すように、工学、心理学、生理学にまたがった分野での総合的な研究が必要なのである。

本文では、このような研究手法で実際に行った研究のいくつかを研究期間もまじえて紹介する。

１．調音結合からの回復

人間は音声を知覚する場合、音声信号から巧みに特徴を抽出し、利用している。しかし、この過程を機械に行わせようとすると、様々な問題が生じる。例えば、連続音声の生成過程において、前後の音韻

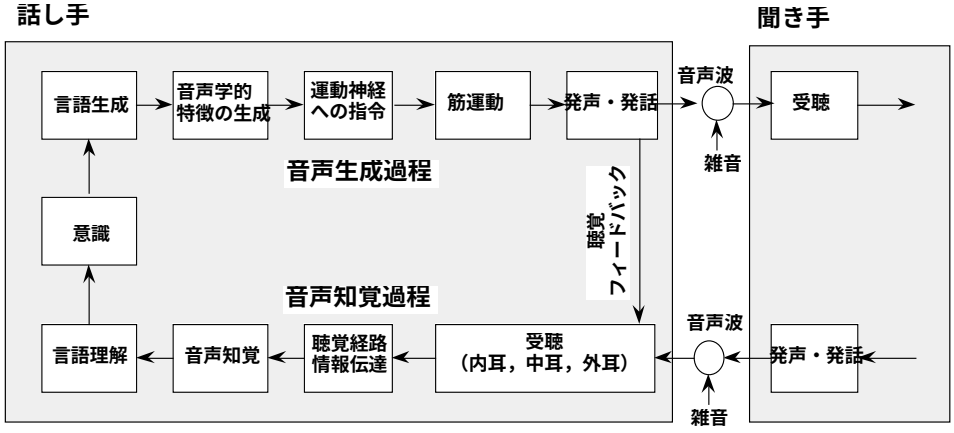

図1　　ことばの鎖

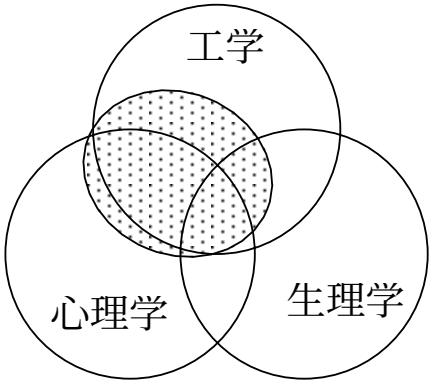

図2　　研究の関連分野

から影響が及ぼされる調音結合がその一つである。

連続音声中では、ある音韻を発話する際その音韻の調音の目標まで達しないうちに次の音韻の発話に移る。従って、実現された音響的特徴は、単独で発話されたときの音韻固有の音響的特徴に達していない(なまけ音)。また、連続的に変化するために特徴が完全でない部分が生じる(わたり音)（図3)。機械はこのような音韻が不完全な部分を誤認識してしまう。しかしながら、人間には不完全な音響的特徴から目標値の推定を行なうという知覚機構が存在すると言われている。従って、この機構がモデル化ができ、これを認識装置に組み込むことができれば認識率の向上が期待出来る。そこで、補正機構の一つと考えられているーゲット予測と文脈効果のモデル化について次説で考える。

1.1　わたり音の回復　－ターゲット予測－ [1]-[4]
（研究期間：1984 - 1989）

調音結合からの補正機構のモデルの一つとして、心理学的知見のみならず聴覚末梢系の知見を取り入れたターゲット予測モデルが提案されている[1][2]。

モデルでは、補正は「聴覚系内に音韻ターゲット

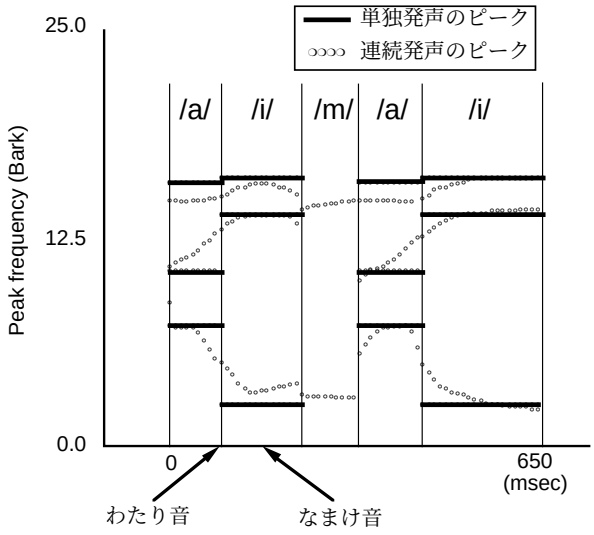

図3　　わたり音、なまけ音

予測機構が存在し、この予測値を人間は知覚している」ために生じると仮定し、この機構を計算機上に実現している。モデル化にあたって用いた知見および手法は次のものである。

Klatt が提唱した聴覚末梢系の工学的モデルのうち、

(1) 有毛細胞をモデル化するための半波整流器
(2) ラウドネス尺度近似のための対数変換
(3) 基底膜をモデル化するためのメル尺度
(4) 側抑制をモデル化するための周波数領域での重み付け

を採用している。また、心理学的知見として

(1) 人間はホルマントの変化だけではなくスペクトル全体の変化を聴いている
(2) スペクトル変化極大点が前出音と後続音の知覚を分ける時点である
(3) 変化量と変化速度には相補的関係がある
(4) 予測には短い時間（約50 ms）の時間幅の情報だけを用いる

さらに、音声生成側からの制約として

(1) 音声の物理的特徴量の変化は臨界制動二次系で近似できる

を用いて、次式によりスペクトルの変化を記述している。

$$\begin{cases} x_n = -ky_n + \dot{y}_n \\ \dot{x}_n = \lambda x_n \end{cases}$$

ここで、y_nは時刻nのスペクトル、$\dot{y}_n$はスペクトルの変化、kはフィードバック係数、λは変化時定数である。

上式において$k = \lambda$とすれば、臨界制動二次系となり、一般解は

$$y = a(1 - \lambda t)\exp(\lambda t) + b$$

となるので、この式のbが推定できれば、ターゲットの予測が可能となる。文献[1]では、bの推定を指数関数のパラメータ推定問題に置き換えることによって、短時間の情報だけでターゲットを予測している。結果の一例を図4に示す。音声 /kiai/ において、わたり音である [e] の時間幅が減少していることがわかる。筆者はこの研究 [1] で電子情報通信学会論文賞を受賞した。

ターゲット予測モデルは音声認識装置の前処理としても有効であり [3][4]、線形判別のための前処理 [3]、あるいは、LVQのための前処理 [4] として有効に働く。

1.2　なまけ音の回復　－文脈効果モデル－ [5]-[8]
（研究期間：1988 - 1996）

音響レベルでの文脈効果モデルとして、心理物理実験から得られた知見を基に、スペクトルピーク間の相互作用による文脈効果モデルが提案されている [5][6]。モデルは、「音響レベルの文脈効果はスペク

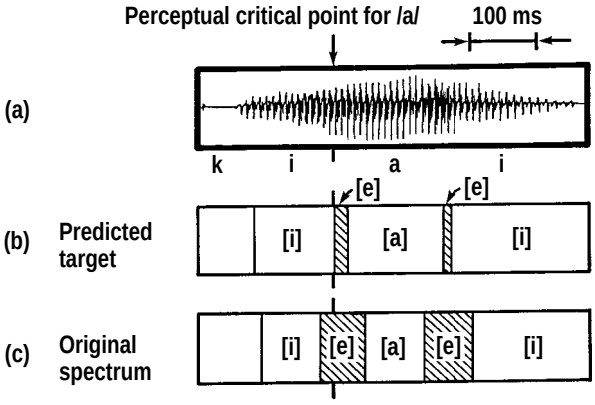

図４　わたり音区間の減少、（上段）原音声/kiai/、（中段）ターゲット予測あり、（下段）予測なし

トルピーク対の相互作用の和としてモデル化できる」ことを仮定して、次式により定式化されている。

$$\begin{cases} T_p(t) = R_p(t) + D_p(t) \\ D_p(t) = \dfrac{1}{2N+1}\sum_t \sum_f g(\Delta t, \Delta f) \end{cases}$$

ここで、$T_p(t)$：単独発話した場合のスペクトルピークの周波数、$R_p(t)$：連続発話した場合のスペクトルピークの周波数、$D_p(t)$：スペクトルピークの知覚的移動量、N：サンプリング数である。また、$g(\Delta t, \Delta f)$は相互作用関数と呼ばれ、時間Δt、周波数Δf離れたスペクトルピークから受ける知覚的影響量を規定している。そして、その値はモデルの定式化において極めて重要である。なお、ここで用いる周波数は人間の聴覚特性を考慮したERB rate[24]である。

$$\text{ERB rate} = 21.4\log_{10}\left(4.37 f[\text{kHz}] + 1\right)$$

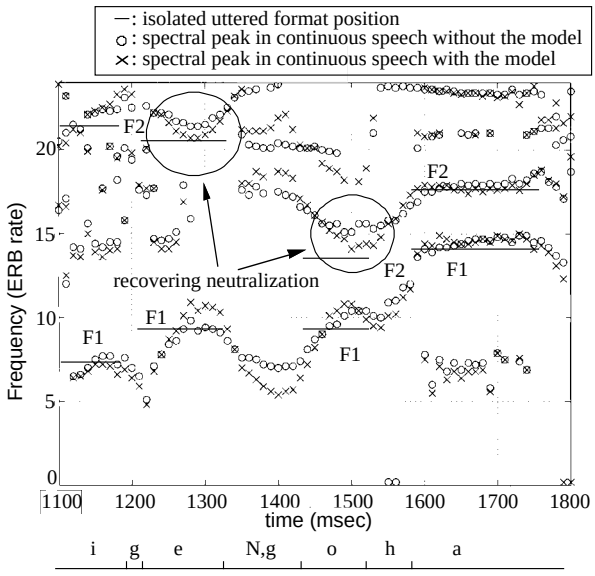

図５　なまけ音の回復結果

相互作用関数は、現在までに、心理物理実験による推定[5]-[7]、および$R_p(t)+D_p(t)$を$T_p(t)$にできるだけ近付けるという規範の下での一般逆行列による推定[6]により得られているが、これらの関数が調音結合のある音声の認識に有効である保証はない。そこで、相互作用関数をスペクトルピーク間の相互作用だけではなくスペクトル全体に拡張し、次に示すように、最小分類誤り学習の理論を用いて認識誤りが最小となるように決定する[8]。

第１段階として、学習時の計算量の軽減および過学習の防止のために、相互作用関数を簡単な形式で近似する。これまでの研究から、スペクトルピークによる文脈効果は、

(1) 時間差Δtが小さいスペクトルピークからは同化効果を受け、時間差が大きいスペクトルピークからは対比効果を受ける

(2) 時間差が非常に小さい或は非常に大きいスペクトルピークから受ける文脈効果は小さいと考えられる。これらのことを踏まえ、相互作用関数を

$$g(\Delta t, \Delta f) = Ae^{B|\Delta f|}\sin(C\Delta f)\cdot |\Delta t|^D e^{E|\Delta t|}\cos(F\Delta t)$$

ただし、$|\Delta t| > 3\pi/2F$ および $|\Delta f| > \pi/C$ では$g(\Delta t, \Delta f) = 0$、という形式で近似する。

第２段階として、連続発話データの母音中心のスペクトルをモデルにより変形し、変形後のスペクトルと単独発話の母音のスペクトルとのユークリッド距離を識別関数として、補正後のスペクトルを識別する。次に、最小分類誤り基準に従い、式の係数AからFの値を学習することにより、識別誤り率が最小になるような相互作用関数を求める。

結果を次式に示す。

$$g(\Delta t, \Delta f) = 4.28e^{0.03|\Delta f|}\sin(0.30\Delta f)\cdot |\Delta t|^{0.50} e^{-0.019|\Delta t|}\cos(0.031\Delta t)$$

相互作用関数の学習結果は、周波数の差が5 ERB Rateのスペクトルピークから受ける影響が最も大きい、また時間差が50 msec以内では同化効果、50～150 msecでは対比効果が優勢であることを示している。これらは、心理実験から得られた結果[5][6]とほぼ一致する。

モデルによるスペクトル変形の効果を確かめるために、連続音声中のスペクトル系列に対してモデルを適用した（図5）。この図は怠けているスペクトルピークの軌跡が単独発話時のピーク位置に近づいていることを示している。また、モデルにより変形を加えた連続音声中の母音中心（最もなまけが大きいと思われる部分）のスペクトルと変形を加えていないスペクトルの識別率の違い、およびワードスポッティングの前処理としてモデルを用いた場合のスポッティング率を調べた。この結果、モデルを用いた場合に有意な性能の向上が認められた[8]。

２．雑音の除去

かつて聖徳太子は同時に１０人の訴えを聞きそれを処理した、と言われている。我々一般人がこれを

真似ようとしても旨くは行かないだろうが、１０人の中の一人の話す内容に注目して聞き取ることは、我々にとってもさして難しいことではない。このように、二つ以上のメッセージが混在していても一方を選択的に聴取可能であるような聴覚上の効果を「カクテルパーティ効果」と呼んでいる。もし、必要な音だけを選択し他の音を除外するというような音源分離問題を解くことができれば、実環境におけるロバストな音声認識システムの実現が期待できる。

カクテルパーティ効果が生じる原因としては、音の到達方向の違い、音源のピッチの違い、音色の違い、また音声の場合には言語的知識、経験などが関係していると見られているが、未だにはっきりしたことはわかっていない。しかし最近、音による情景理解（auditory scene analysis:ASA）に関する研究から新たな知見が報告されつつある [9]。

本文では、カクテルパーティ効果（あるいは、カクテルパーティ効果の重要な側面の一つである音源分離）のモデルとして、音知覚過程からのアプローチである ASA を基としたモデル [10]-[12] と両耳聴の知見を取り入れた雑音除去モデル[13][14]を紹介する。

2.1　ASA のモデル [10]-[12]（研究期間：1992 - ）

近年、心理物理学からの知見を取り入れてカクテルパーティ効果のモデル化を試みる研究が見られる。Bregmanによる音による情景理解に関する研究から、人間が音声あるいは音楽を聞く場合、個々の物理的特徴の分離（segregation）／群化（grouping）が起こり、群化された物理的特徴から一連の流れ（stream）を形成した上で聞き取っていることがわかってきた。

Bregmanは、音を通じて環境を把握する情景解析の問題を解くために聴覚がストリーム形成に利用している制約条件のいくつかを音響事象に関係する４つの発見的規則：

(I) 共通の立ち上がり／立ち下がりに関する規則
(II) 漸近的変化に関する規則
(III) 調波関係に関する規則
(IV) １つの音響事象に生じる変化に関する規則

としてまとめている。そして、これらの発見的規則を物理的制約条件としてとらえ直すことにより、計算論的な聴覚の情景解析の問題を解くことが可能である。

例として筆者らが行った二波形分離問題[10]-[12]について紹介する。

図6に二波形分離の概略図を示す。本システムは Auditory filterbank 部、Segregation 部、Grouping 部の３つの場面からなっている。

今、信号 $f_1(t)$ と雑音 $f_2(t)$ が加算された波形 $f(t)$ が観測されたとする。Auditory filterbank 部では、観測された波形を聴覚特性を考慮したフィルタであるガンマトーンフィルタ [24] を基底関数とする Wavelet 分析系に入力し、帯域ごとに分割する。このとき、ガンマトーンフィルタのヒルベルト変換対を考えて瞬時振幅 $S_k(t)$ と瞬時位相 $\phi_k(t)$ を計算しておく。次に、Separation 部では発見的規則 (II), (IV) を考慮して、各フィルタ内で信号の振幅 $A_k(t)$ と雑音の振幅 $B_k(t)$、それに信号と雑音の位相差を推定する。最後に、Grouping 部において、発見的規則 (I), (III) を用いて、信号と雑音それぞれに関係する振幅成分と位相成分を集め、逆 Wavelet 変換により信号と雑音の波形 $\hat{f}_1(t)$, $\hat{f}_2(t)$ を推定する。この処理を行えば、同一周波数帯域に信号と雑音が存在していても分離可能である。

この一連の処理によって求められた波形は、対数スペクトル歪みが約 20 dB 改善されている。

2.2　雑音除去モデル　－NORPAM－ [13][14]（研究期間：1993 - ）

機械による音声認識においては、雑音等で汚れて

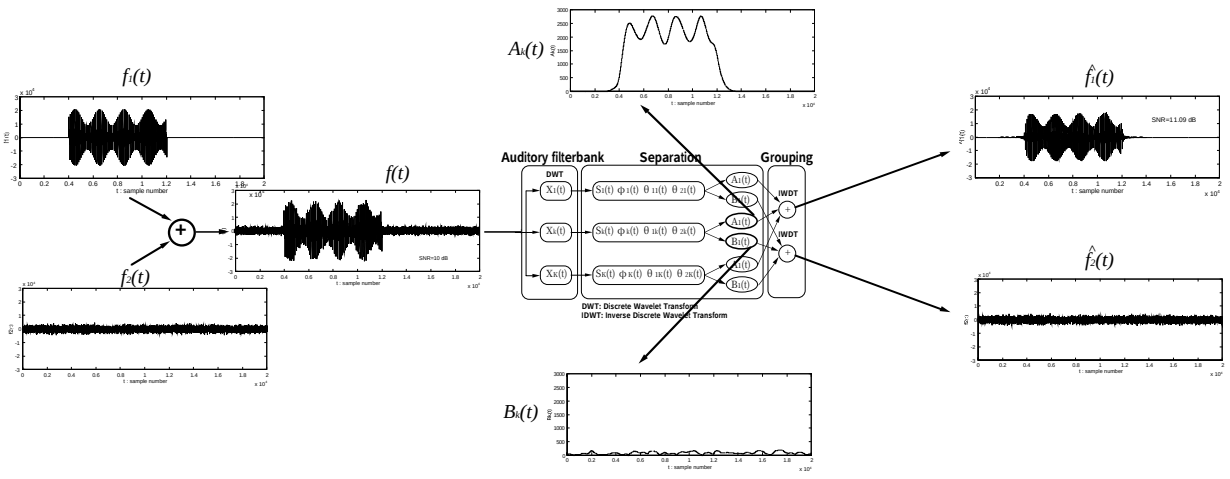

図６　　二波形分離問題の概略図

いないきれいな音声ではほぼ実用レベルに達してはいるものの、周囲雑音が存在する場合には認識率の著しい低下は免れない。一方、人間は周囲雑音が大きく、複数の話者が存在しているような状況、あるいは、残響のある環境においてさえも、左右の耳で着目する話者の音声を選択的に聴取することができる。この能力は、音環境にほとんど影響を受けず頑健である。

音声認識におけるこの問題を解決するために、適応フィルタを用いて雑音を抑圧する方法、マイクロホンアレイを用いて着目する話者の方向の指向特性を鋭くする方法など様々な音響的前処理方式が研究されている。しかし、これらの方法は高速な信号処理装置、また、アレイを形作るための多数のマイクロホンを必要とするため、装置は大がかりとなり実用的ではない。自動車内での音声による携帯電話・ナビゲーションシステムの制御、混雑した環境での自動販売機への音声入力などへの応用を考えれば、マイクロホンの本数が少なく、しかも簡単な処理で雑音・残響抑圧ができる小規模の前処理装置が必要である。

そこで、マイクロホン対を用いて信号音以外のある一方向の時間・周波数が局在した雑音を推定し、推定した雑音を引きさることによって信号音を浮かび上がらせる手法を提案した[13]。主マイクロホン２本と補助マイクロホン１本を用いた場合の性能評価の結果（図7）、合成波形を用いたシミュレーションの場合 SN 比が 10 ～ 20dB 向上し、実環境では雑音が含まれない信号音との対数スペクトル距離が約 5dB 減少した。

また、音声認識の前処理装置として用いることを前提として、推定した雑音を引き去る場合に、推定した雑音波形そのものを引き去るのではなく、音声認識で良く用いられている振幅スペクトルを引き去る方法（Spectral Subtraction）を用いて雑音除去を試

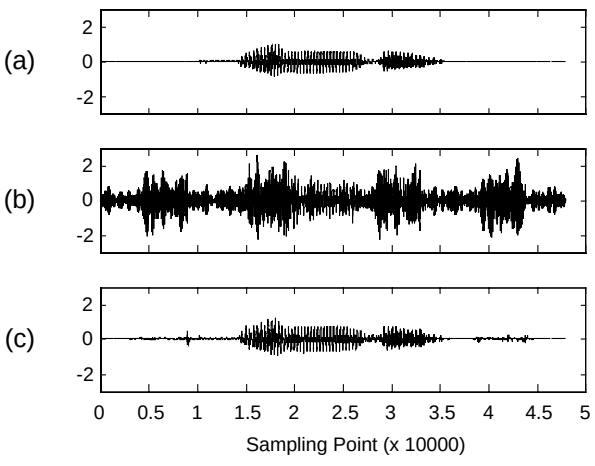

図 7　NORPAM による雑音除去結果、（上段）原音声、（中段）雑音付加音声、（下段）雑音除去音声。sampling 周波数：48 kHz。

みた[14]。その結果、突発雑音、非定常雑音も除去可能であることが明らかとなった。

３．音声の個人性

音声に個人性が含まれるということは周知の事実であるが、ではどのような物理量が音声の個人性を決定する要因になっているのかという問に対する答えは未だ完全ではない。

音声は、声道情報を反映したスペクトルの包絡特性および声帯情報を反映した基本周波数特性の二つの物理量で大まかには記述されると言われているが、個々の物理量に潜む個人性関連量についての細かい議論はあまりなされていないのが現状である。また、個人性知覚に関連する物理量を積極的に制御し、音声認識・合成に応用する研究はほとんど行われていない。

個人性に関連する個々の物理量を抽出しこれを制御する方法を見つけ出すことは、音声の個人性を議論する基礎的な研究に貢献するのみでなく、応用面としての音声認識・合成技術において非常に重要な要素となる。たとえば、個人性制御モデルが構築されれば、個人性を音韻性を損なわない範囲で除去することが可能となり多数話者音声認識の認識率向上が期待できる。また、合成音声に個人性を付与できることとなり多様な合成音声が得られることとなる。さらに、個人性情報を抽出できるようになり話者認識・照合のための特徴量として使用できるようになる。

人間が個人差を知覚する場合に有効な物理量としては、過去の研究から基本周波数と特定の帯域のスペクトル包絡特性が知られている。そこで、個人性の知覚と物理関連量およびこれらの制御法を得るために、スペクトル包絡と基本周波数に埋め込まれた物理関連量を検出することを試みる。この場合、人間が個人差を知覚する時に用いる物理量が最も個人性を表す物理量であるという仮定を設け、個人差の知覚と様々な物理量の関係を心理物理実験を通して明らかにすることとする。

3.1　スペクトル包絡に含まれる個人性 [15]-[19]（研究期間：1992 - 1997）

音声分析合成システムを用いて、実音声から得られたスペクトル包絡を様々に変形した音声を合成する。そして、これを刺激音として用いた心理物理実験から、次のことが明らかとなっている。

(1) スペクトルの分散を調べた結果、22 ERB rate（約2.2 kHz）を境として低域は音韻差による分散が大きく、高域は個人差による分散が大きい（図8）。これは、スペクトルの高域に個人差が含まれていることを示唆する結果である。

(2) スペクトル包絡の22 ERB rateを境とした低域と高域を独立に変形した音声を刺激音として、被験者

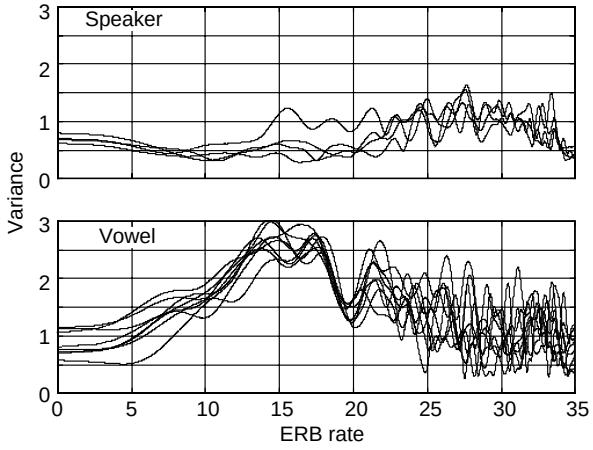

図 8　　各母音における話者間の分散（上段）と各話者における母音間の分散（下段）

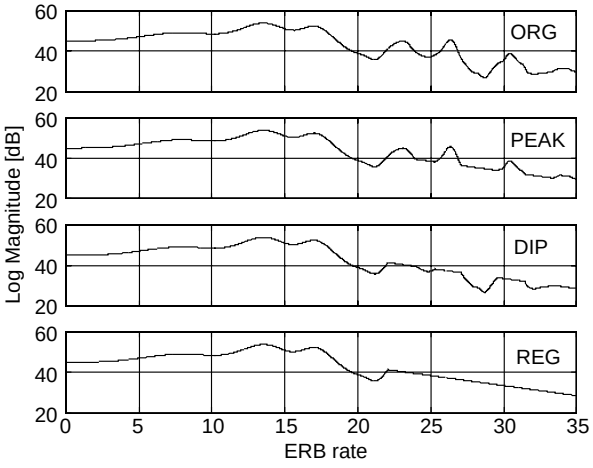

図 9　　刺激音ORG（原スペクトル：母音/a/）、PEAK（回帰直線より上のピークを保存）、DIP（回帰直線より下のディップを保存）、REG（回帰直線で置換）のスペクトル包絡。

に個人性判断を行わせた結果、高域の変形に対して個人性判断が敏感であった。これは、個人性はスペクトル包絡全体に現れるが、高域により多く現れることを示すものである。

(3) また、高域スペクトル中のどのような特徴が個人性判断に関係しているかを、高域のスペクトル包絡を変形した音声（図9）を用いて聴取実験により調べた結果、話者識別率が高い方からORG > PEAK > DIP > REGとなり、話者識別にはスペクトル包絡のディップよりもピークが重要な意味を持っていることが明らかとなった。

(4) これらの知見を基に声質変換を試みた結果、個人性はスペクトル包絡の 20 ERB rate 付近のピーク以上の帯域に顕著に現れ、この帯域を利用して声質変換が可能であることがわかった。

(5) また、この帯域を単純類似度法による話者認識に利用すると高い弁別能力が得られることも明らかとなった。

3.2　基本周波数に含まれる個人性[20][21]（研究期間：1993 - ）

基本周波数包絡を藤崎モデルによって記述し、モデル中のどのパラメータが個人性判断に関係するかを心理物理実験を通して調べた。その結果、

(1) パラメータは、包絡の変化の大きさを表すものとアクセントなどのタイミングを表すものに大別される。それぞれのパラメータの個人差による分散をＦ比によって調べた結果、包絡の変化の大きさを表すパラメータの分散が大きくなっていた。これは、包絡の変化の大きさを表すパラメータが個人性を多く含むことを示唆するものである。

(2) タイミングを表すパラメータはそのまま用いて、包絡の変化の大きさを表すパラメータを他の話者と入れ換えた音声を合成する（図10）。これを被験者に呈示し個人性判断を行わせた。実験結果は、約9割が包絡の変化の大きさを表すパラメータを用いた話者として判断している。この結果は、包絡の変化の大きさを表すパラメータが基本周波数包絡の個人性を表す重要な物理量であることを示している。この結果を用いれば、音声合成において基本周波数包絡を用いた話者変換が可能となる。筆者はこの研究[21]で日本音響学会佐藤論文賞を受賞した。

3.3　異常構音の診断に向けて[22][23]（研究期間：1994 - ）

言語障害による異常構音も、音声の個人性の一分野である。本研究では、昭和大学歯学部と共同で、

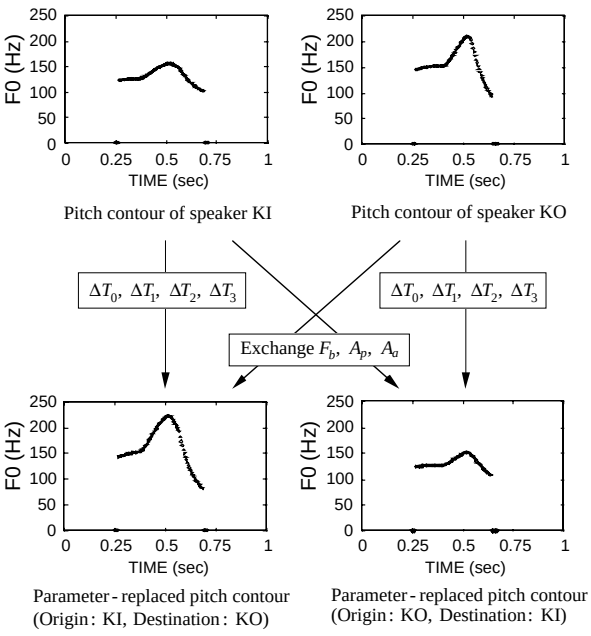

図 10　　パラメータ置換の概略図。音声は単語「青い」。

異常構音の一種である側音化構音を扱っている。

側音化構音とは、舌、顎などに形態的障害がないにもかかわらず、子音 /sh/, /ch/ などを発話する場合、舌を口蓋中央に接触させるために、呼気が臼後部より口腔前庭の側方より出ることにより音が歪むものである。正常構音の場合は呼気は口腔正中より出るが、側音化構音の場合は口腔の側方より出るために独特の歪み音を呈する。構音時に口唇が側方へ変移することもある。舌、顎などの運動機能が未熟な小学校低学年でよく見受けられる。

診断は、ステンレス板による呼気の流出部位検査、エレクトロ・パラトグラフィーによる舌の接触様式の測定、舌造影側方頭部X線規格写真法（X線セファログラフ）を用いて客観的に行なわれなければならないが、膨大な時間がかかるため、現在は言語臨床家による聴覚印象に基づいた臨床診断に頼っているのが現状である。そこで、言語臨床家は異常構音のどの特徴をとらえて診断を下しているのかを、工学的手法および心理物理学的手法を用いて特定し、側音化構音診断の自動化を試みている。

４．聴覚モデル

耳に入ってきた音（音声）が聴覚系内でどのように符号化されているかを調べるために、生理学の知見を取り入れたモデルの構築[25][26]を行なっている。

4.1　聴覚末梢系および蝸牛神経核のモデル化[25][26]（研究期間：1995 - ）

蝸牛神経核は音声などから最初に意味ある情報を取り出している部位として知られているが、このモデルを構築する場合にその高度に特殊化された機能を解明するためには、平均的な情報ではなく、神経発火パルスそのものを入力として用い、出力としても神経発火パルスを出力するモデルを考える必要がある。

そこで、外耳、中耳から蝸牛を通って聴神経に至るまでの聴覚末梢系を生理学の知見に忠実に計算機上に構築し、聴神経における神経発火パルスの再現を試みた[26]。生理学データとの定量的な比較の結果、良い一致を示した。

また、この神経発火パルスを新たに構築した蝸牛神経核モデルへの入力として用い、蝸牛神経核での母音の特徴抽出機構について検討を行なった [27]。

4.2　上オリーブ核のモデル化[27]（研究期間：1996 - ）

音源方向定位に深く関わっている聴覚系内の部位である上オリーブ核のモデルを計算機上に構築している。この場合も、機能を解明するためには神経発火パルスそのものを入力として用い、出力としても神経発火パルスを出力するモデルを考える必要がある、との考えから、ニューロンの活動電位、シナプス伝達機構、神経発火などを生理学的知見に忠実なモデル化を試みている。

５．音色[29][30]（研究期間：1996 - ）

聴覚情報は、視覚情報のように時間に無関係に存在する“対象”の存在が重要なのではなく、対象がどのように変化するかという“事象”の内容が重要である。このため、聴覚は時間的に変化しているものを情報として受け取ることに優れている。たとえば、人間は時間の揺れに対する感度には敏感であり、100 Hzのパルス列に対して1 ms以下の時間ジッターでも検出できる。また、合成用の励振音源として用いるパルス列をオールパスフィルタに通し群遅延を与えた場合、合成音声はより自然な音質となる。これは、位相が揃うという不自然性に対して、人間が敏感である証拠であろう。

一方、永らく人間は時間軸方向の情報の一つである位相に鈍感であると信じられていた。このため、スペクトル構造のみを保存するような符号化方式が多く使われてきた。しかし、これらの音質が“機械くささ”を伴うのも確かである。

そこで、時間軸方向の情報の一つである位相に関して、人間の知覚特性について心理物理実験を行なっている [30]。この研究は、人間の位相知覚特性の測定のみならず、音声の自然性についての議論へも貢献する。

６．音声の符号化[31]（研究期間：1992 - 1997）

音声の低ビットレート符号化を実現するための一つの手段として、時々刻々変化する音声特徴を、音声のイベントとその変化パターンに分解してそれぞれを符号化するTemporal Decomposition法が知られている。この方法は数学的にはきれいに定式化されているものの、処理時間がかかる、結果が分析区間に依存するなど、問題点があった。そこで、 Temporal Decomposition法のための新しい計算法を提案し、処理時間の低減を実現した [31]。

７．関連論文

(a) わたり音の除去（ターゲット予測モデル）
[1] 赤木、古井(1986). ”音声知覚における母音ターゲット予測機構のモデル化”、電子情報通信学会論文誌、J69-A, 10, 1277-1285.（電子情報通信学会論文賞受賞）
[2] Akagi, M. (1990). "Evaluation of a spectrum target prediction model in speech perception", J. of Acoust. Society of America, 87, 2, 858-865.
[3] Akagi, M. and Tohkura, Y. (1990). "Spectrum target presiction model and its application to speech recogni-

tion", Computer Speech and Language, 4, Academic Press 325-344.
[4] Aritsuka, T., Akagi, M. and Katagiri, S. (1991). "Speech recognition using spectrum target prediction model as a front-end processor", Speech Group Tech. Report, IEICEJ, SP91-36

(b) なまけ音の回復（文脈効果モデル）
[5] Akagi, M. (1992). "Psychoacoustic Evidence for Contextual Effect Models", In Tohkura, Y., Vatikiotis-Bateson, E. and Sagisaka, Y. Eds., Speech Perception, Production and Linguistic Structure, pp.63-78.
[6] Akagi, M. (1993). "Modeling of contextual effects based on spectral peak interaction", J. of Acoust. Society of America, 93, 2, 1076-1086.
[7] Akagi, M., van Wieringen, A. and Pols, L. C. W. (1994). "Perception of central vowel with pre- and post-anchors", Proc. Int. Conf. Spoken Lang. Process. 94, 503-506.
[8] 米沢、赤木(1997). ”文脈効果のモデル化とそれを用いたワードスポッティング”、電子情報通信学会論文誌、J80-D-II, 1, 36-43.

(c) 音による情景解析のモデル
[9] 赤木正人(1995). ”カクテルパーティ効果とそのモデル化”、電子情報通信学会誌解説, 78, 5, 450-453.
[10] 鵜木、赤木(1997). ”雑音が付加された波形からの信号波形の一抽出法”、電子情報通信学会論文誌、J80-A, 3, 444-453.
[11] Unoki, M. and Akagi, M. (1997). "A method of signal extraction from noisy signal based on auditory scene analysis", Proc. CASA97, IJCAI-97, Nagoya, 93-102.
[12] Unoki, M. and Akagi, M. (1997). "A method of signal extraction from noisy signal", Proc. EUROSPEECH97, 2587-2590.

(d) 雑音除去
[13] Mizumachi, M. and Akagi, M. (1998). "Noise reduction by paired-microphones using spectral subtraction," Proc. ICASSP98, II, 1001-1004
[14] Akagi, M. and Mizumachi, M. (1997). "Noise Reduction by Paired Microphones", Proc. EUROSPEECH97, 335-338.

(e) スペクトル包絡に含まれる個人性
[15] Kitamura, T. and Akagi, M. (1995). "Speaker individualities in speech spectral envelopes", J. Acoust. Soc. Jpn. (E), 16, 5, 283-289.
[16] Kitamura, T. and Akagi, M. (1996). "Relationship between physical characteristics and speaker individualities in speech spectral envelopes", Proc ASA-ASJ Joint Meeting, 833-838.
[17] 北村、赤木(1996). ”連続音声中の母音に含まれる個人性について”、音響学会聴覚研究会資料、H-96-98
[18] 北村、赤木(1997). ”単母音の話者識別に寄与するスペクトル包絡成分”、日本音響学会誌、53, 3, 185-191.
[19] 北村、赤木(1996). ”単純類似度法による話者識別に適した周波数帯域の検討”、平成８年秋季音響学会講演論文、1-6-17.

(f) 基本周波数に含まれる個人性
[20] Akagi, M. and Ienaga, T. (1995). "Speaker individualities in fundamental frequency contours and its control", Proc. EUROSPEECH95, 439-442.
[21] Akagi, M. and Ienaga, T. (1997). "Speaker individuality in fundamental frequency contours and its control", J. Acoust. Soc. Jpn. (E), 18, 2 73-80.（日本音響学会論文賞受賞）

(g) 側音化構音の診断
[22] 赤木正人、高木直子、北村達也、鈴木規子、藤田幸弘、道健一(1996). ”側音化構音の知覚と物理関連量”、電子情報通信学会技術報告、SP96-34.
[23] Akagi, M., Kitamura, T., Suzuki, N. and Michi, K. (1996). "Perception of lateral misarticulation and its physical correlates", Proc ASA-ASJ Joint Meeting, 933-936.

(h) 聴覚モデル
[24] 赤木正人 (1994). ” 聴覚フィルタとそのモデル”、電子情報通信学会誌解説 , 77, 9, 948-956.
[25] Maki, K. and Akagi, M. (1997). "A functional model of the auditory peripheral system", Proc. ASVA97, Tokyo, 703-710.
[26] 牧、赤木、廣田(1998). “ モデルに基づいた前腹側蝸牛神経核における母音に対するチョッパー型応答に関する検討”、音響学会聴覚研究会資料、H-98-50
[27] 伊藤、赤木(1998). ”音源方向定位のための聴覚モデルの検討”、電子情報通信学会技術報告、SP97-138.
[28] 赤木正人(1998). ” 聴覚特性を考慮した波形分析”、日本音響学会誌 , 54, 8, 575-581.

(i) 位相の知覚
[29] 赤木正人(1997). ”位相と知覚　－人間ははたして位相聾か？－”、平成９年秋季音響学会招待講演論文、1-2-2.
[30] 赤木、安武(1998). “ 時間方向情報の知覚の検討　－位相変化の音色知覚に及ぼす影響について－”、電子情報通信学会技術報告、EA98-19.

(j) 符号化
[31] Nandasena, A.C.R. and Akagi, M. (1998). "Spectral stability based event localizing temporal decomposition," Proc. ICASSP98, II, 957-960

研究経歴

1998.10.1