TR-A-0005

Spectrogram Reading

スペクトログラムリーディング

Shigeru Katagiri

片桐　滋

1987.6.9

ATR視聴覚機構研究所

概　要 (Abstract)

本報告は、音声のスペクトログラムリーディングに関する技術的紹介を行っている。音声のスペクトログラムは、音声波形が持つエネルギーをある時点－周波数における濃淡表示として視覚表現したものである。スペクトログラムリーディングは、このスペクトログラムを音声や言語に関する様々な知識を用いて人間が「読み取る」技術であり、この読み取り過程で用いられる人間の能力を音声認識システム等の構築に応用することを目的としている。

本報告では、初めに、スペクトログラムリーディングの方法と、人間が達成できる「読み取り能力」に関する研究成果を紹介する。次に、スペクトログラムリーディングを工学的に応用するために行われている研究アプローチの中から、知識処理技術を用いた音声認識システム構築の試み、音声信号処理と記号処理との双方を効率よく実現しようとする研究環境の整備、スペクトログラム上の音響的特徴を把握するために必要な音声データベースの整備について、現状を整理する。最後に、スペクトログラムリーディングに基づく音声研究において、今後解決すべき課題を提起する。

発行時　配付先 (Initial Distribution Specifications)

聴覚研究室員	筧リーダ（NTT 基礎研）
淀川社長	古井室長（　〃　）
樽松社長	

備考 (Notes)

スペクトログラムリーディング

ＡＴＲ視聴覚機構研究所
片桐　滋

１．　まえがき

音声信号処理研究の分野において新しい展開が期待されている研究手法の一つに、スペクトログラムリーディング［１］がある。本手法は、視覚表現された音声のスペクトログラムを専門家が読み取る際に用いる経験的知識に基づいて、音韻特徴の発見とその音声情報処理技術への応用を行うものである。米国のＭＩＴの研究に代表される本手法は、ＭＩＴにおける夏期特別講座や先頃開かれたＡＴＲ基礎セミナーの盛況が示す様に、音声研究者の広い関心を集めている。本論では、このスペクトログラムリーディングの技術的紹介と研究手法として抱えている技術的課題の紹介を行う。

２．　スペクトログラムリーディングの概要

スペクトログラムリーディングの手法の概要を紹介するため、実際にスペクトログラムを「読む」過程を示す。なお、本論の目的が、リーディング技術の獲得ではなく、その概要の理解にあることから、リーディング過程に登場する用語の詳細には立ち入らないこととする。

２．１．音声のスペクトログラム

図１に音声スペクトログラムの例を示す。図中、上から音声波形、スペクトル変化量、短区間パワー、スペクトログラムである。スペクトル変化量は、値が大きい程その時点付近のスペクトログラムの変化が大きいことを示している。スペクトログラムは、音声波形が持つある時点ー周波数におけるエネルギーを濃淡表示したものであり、表示が濃い程その部分のエネルギーが大きいことを表している。

これまで、音声のスペクトログラムを得るためには、アナログ的なスペクトル分析装置であるソナグラフが広く利用されてきた。この装置は、図１の様な濃淡表示と、この濃淡表示を地図の等高線図の様に表現した等高線表示、いわゆる「声紋」とを作ることができ、さらに、３００［Ｈｚ］（広帯域）と４５［Ｈｚ］（狭帯域）とのスペクトル分析用帯域通過フィルタを切り替えることにより、２種類の時間及び周波数分解能をも実現できる。広帯域分析は、高い時間分解能を持つため、ホルマント等の時間的な動きの観察に適している。また、狭帯域分析は、ホルマント等の時間的な変化がぼやけるものの、声帯振動に起因するピッチの高調波の観察に適している。

しかし最近では、計算機の発達によりディジタル的なスペクトログラムの作成が盛んに行われる様になっている。図１も、図２に示す様なディジタル処理によって作成したものである。音声波形は、標本化周波数１６ｋＨｚのＡ／Ｄ変換によってディジタル量に変換され、続いて、２．５ｍｓｅｃずつ移動する５ｍｓｅｃのハミング型時間窓によって分析区間毎に切り出される。切り出された音声波形は、５１２点のＤＦＴによって分析され、さらに対数パワースペクトルに変換される。対数

パワースペクトルは、３６階調の濃淡表示に変換され、各分析区間毎に時間軸にそって表示される。以上の分析条件で得られるスペクトログラムは、広帯域のソナグラフの出力とほぼ等しい時間分解能を実現している。

２．２．スペクトログラムの読み方

図１のスペクトログラムを用いて、スペクトログラムリーディングを行う。まず、図中で強く目を引く部分から始める。０．３ｓから０．４ｓ付近、さらに０．５ｓから１．０６ｓ付近の低域のエネルギーが大きく、この付近が有声音であることが予想される。０．３ｓから０．４ｓ付近では、第１ホルマントが５００Ｈｚ付近、第２ホルマントが２０００Ｈｚ以上にあることから、母音／ｅ／である可能性が大

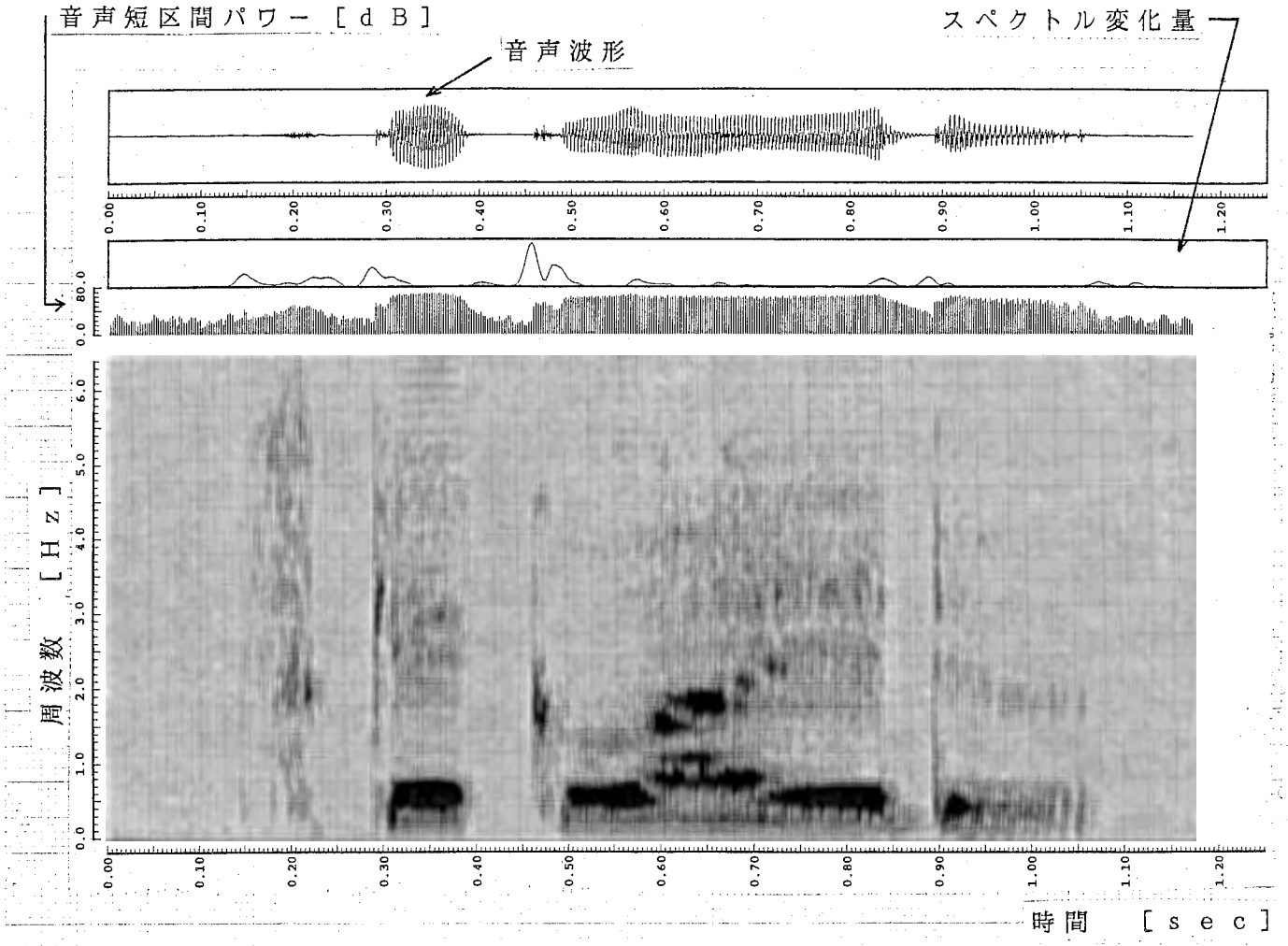

図１．　音声スペクトログラムの例.

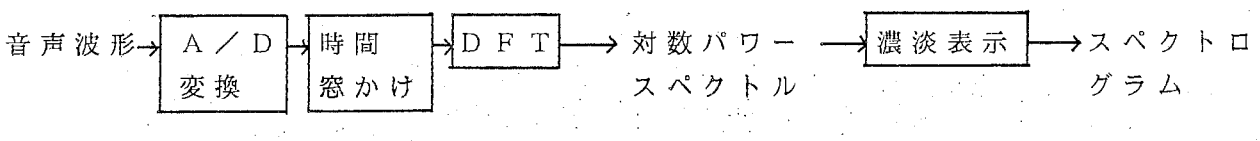

図２．　ディジタル処理による音声のスペクトログラム作成手順.

きい。０．５ｓから０．５８ｓ付近は、第１、第２ホルマントとも低く、／ｏ／または／ｕ／であることが考えられるが、特に第１ホルマントのエネルギーが第２ホルマントのそれよりもかなり大きいことから、／ｕ／の可能性が大きい。次に、０．６ｓから０．６８ｓ付近は、第１、第２ホルマントとも高い位置にあり、／ａ／である可能性が大きい。これに続く０．７ｓから０．８４ｓ付近は、先の０．３ｓから０．４ｓ付近と似ており、／ｅ／と考えられる。最後の有声音区間の０．９ｓから１．０６ｓ付近は、やはり第１ホルマントのみが強く、しかも第１、第２ホルマントとも先行する／ｅ／と考えられる部分より低い。これは、中舌母音、すなわち／ｕ／であることが予想される。

次に、０．２ｓの付近を見ると、全帯域にわたってエネルギーが存在しており、摩擦性の音韻が存在することがわかる。この区間のエネルギーの分布は帯域全体にわたってほぼ均等であり、エネルギー自体小さい。そこで、／ｓ／や／ｈ／、あるいは／ｔｓ／が候補として考えられる。詳細に観察すると、０．１５ｓ付近にやや不鮮明ながらも、破裂性のスリットが観察される。以上より、この区間は摩擦音でなく破擦音／ｔｓ／である可能性が大きい。そこで、この摩擦性区間の後半部、すなわち０．２２ｓ付近に注目すると、５００Ｈｚ以下の低域に比較的明らかなエネルギーの集中が認められ、さらに２０００Ｈｚ付近にはかなりはっきりとホルマントが見いだされる。無声化した母音／ｕ／がここに存在する可能性がある。

０．２５ｓから０．２９ｓは、どの周波数成分のエネルギーも存在しない。すなわちこの部分は閉鎖区間であり、有声音に先行する閉鎖区間に存在するバズバーが無いことや０．３ｓ付近の母音の立ち上がり時間が長いことを考慮すると、無声破裂音である。さらに０．２９ｓから始まる破裂部分に注目すると、破裂直後のバーストが明確に存在し、しかも後続母音の第２、第３ホルマントが寄り集まる付近のみのエネルギーが大きい。これは口蓋破裂音特有の現象である。以上の観察より、この区間に存在するものは、無声の口蓋破裂音／ｋ／と考えられる。０．４７ｓ付近に関してもこれとほぼ同様の観察が行え、さらに口蓋破裂音に固有のダブルバーストも見える。以上より、この区間にも／ｋ／が存在することが予想される。

０．７ｓ付近の／ａ／から／ｅ／に向かう遷移部の区間はかなり長く、０．７ｓの前後は母音連続が起こっていると考えられる。これに対し、０．６ｓ付近の／ｕ／から／ａ／に対する遷移部は、母音が連続しているにしては短い。ここになんらかの子音が存在するものと予想される。０．５ｓから０．５８ｓまでの区間は、仮に母音とすれば／ｕ／の可能性が大きいが、／ｕ／にしては第２ホルマントが低すぎる。この区間における／ｕ／以外の候補として考えられるのは、鼻音か半母音／ｗ／である。しかし第１ホルマントの位置から鼻音ではない。一方、／ｕ／に／ｗａ／が続く場合、唇の丸めが起こるためホルマント位置は低域に移行する。以上のことを考慮すると、この区間には／ｗ／が存在する可能性が大きいことになる。

最後に、０．８４ｓから０．８９ｓまでの短時間のエネルギーギャップ、かつ低域にある有声性のエネルギーから／ｒ／の存在が推測される。１．０４ｓ以降の部分がグロッタルストップであることは、スペクトル及び波形の乱れから容易に判断できる。

以上の過程より、図が示す発声内容を知ることは容易である。すなわち、「付け加える」である。

ここで例示した様に、スペクトログラムリーディングにおいて用いられる知識は、音声生成過程に関する知識に支えられている。また、英語のスペクトログラムリーディング［１］で示される様に、音声学的知識［２］や単語中の音韻の生起頻度等の知識も使われる。スペクトログラムリーディングの背景には、音声に関する様々な研究成果が利用されていることがわかる。

２．３．スペクトログラムリーディングの正確度

ここでは、スペクトログラムリーディングによって人間が達成できる音声認識力について紹介する。ＭＩＴの実験［１］によれば、音声研究に関する豊かな経験を持つ専門家の音韻認識率が８０％を越すことが報告されている。この結果は、人間の聴覚と比較してもかなり高い認識率であり、スペクトログラムが豊富な音韻情報を提供していること、さらにその情報も、訓練による経験的知識として十分に利用可能なことが示された。

しかし、こうした実験においてはスペクトログラムを読む実験者の能力やスペクトログラムの表示法が実験結果に影響することが考えられる。Ｍ．Ａ．Ｂｕｓｈら［３］は、ＤＦＴスペクトログラムとＬＰＣスペクトログラム、調音上重要と考えられる３時点のＬＰＣスペクトルスライス、さらにこの３時点ＬＰＣスペクトルスライスから得た特徴パラメータの数値表を用いた認識率の比較を行っている。その結果、ＬＰＣスペクトログラムを用いた結果が最も優れているものの、この４種の表示法による結果間の差異は小さく、しかも単に特徴パラメータを並べただけの数値表を用いた方がむしろ３時点ＬＰＣスペクトルスライスを用いた結果より良いことが明らかにされた。Ｍ．Ａ．Ｂｕｓｈらは、この実験の結果から、数値表として表現された情報がスペクトログラムリーディングにおける経験的な判断の材料として有効であることは、こうした経験的知識を計算機上に実現し得ることを示すものであると指摘している。

また、Ｂ．Ｇ．Ｇｒｅｅｎｅ［４］らによって、音声に関する深い知識がなくともある程度の訓練を施すことでかなりのスペクトログラムリーディング能力が実現されることも示されている。ＡＴＲでは、大規模音声データベース構築の一環として、訓練された作業員の視察による音韻等の表記付を行っている［５］。ここで表記とは、子音や母音で構成される発声内容や音響物理的な特徴に従って音声波形をセグメント化するものであり、セグメントの始終端の時点とセグメント名からなる。図３は、このデータベースで用いられている表記の構成を示している。この作業を始める際、作業員が付ける表記の正確度を明らかにするため、作業員が付けたイベント表記と専門家が付けたイベント表記との比較実験を行った［６］。図３の中では、イベント表記が最も直接音韻特徴を反映するため、これに注目して比較を行った。実験に参加した５名の作業員は、この実験の最初に行われた２ヶ月間の訓練において初めてスペクトログラムや音声に関する表記法に接した者ばかりである。セグメントの表記名が一致しかつ始終端の時間的なずれがある閾値以下であると言う基準のもとで、作業員が作成した音韻表記は９４．６％の正解率を示した。この実

験においては、作業員が予めスペクトログラムの発声内容を知っているとはいえ、非常に正確なセグメントの指定を実現できることは注目される。

以上、これらの報告から、スペクトログラムリーディングが特定の専門家のみに許される技術ではないこと、スペクトログラム上には普遍的で、かつ発見も比較的容易な音韻特徴が存在することがわかる。

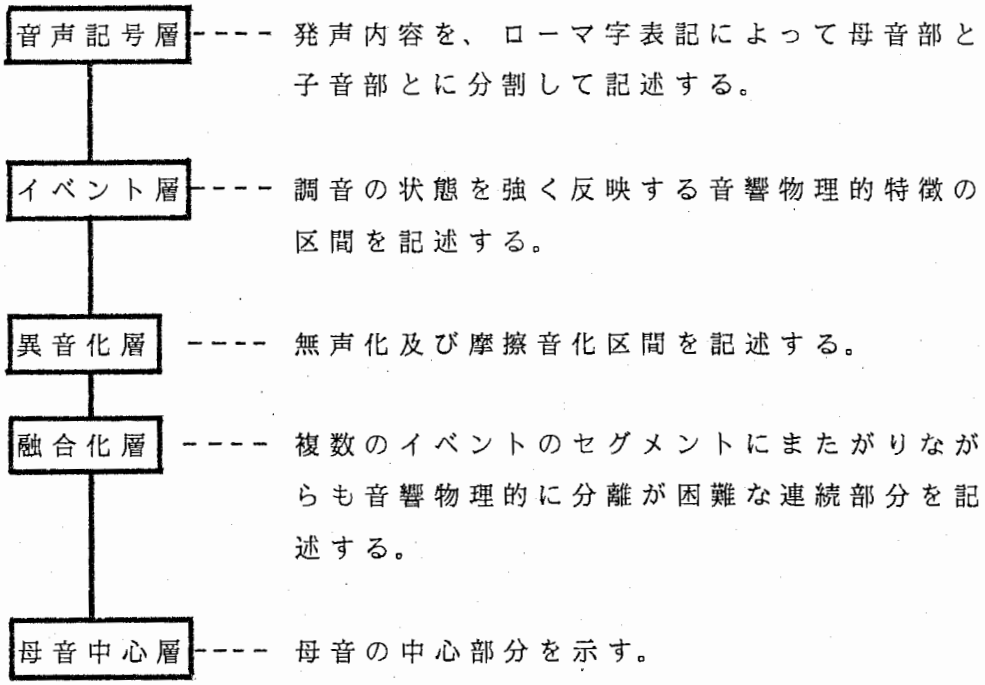

図３．　ＡＴＲにおける音声データベースの表記構造.

表１．　ＡＴＲにおける音声データベースのイベント層表記.

表記記号	音声事象
<	無声子音から母音への入り渡り
>	無声子音から母音への出渡り
*>	有声子音から母音への出渡り
tr	スペクトルパターンの乱れ
cl,*cl	破裂(破擦)音内閉鎖区間(*は有声)
/p,t,k,b,d,g/	破裂音内破裂･気音区間
pau	休止区間(pause)
/s,h,sh,z,dj,f/	摩擦音区間
/w,y/	半母音区間
/r/	流音区間
/a,i,u,e,o/	母音区間
/mm/	鼻子音区間
/j/	拗音区間
N	撥音区間
ts,ch	破擦音内摩擦区間

3．　スペクトログラムリーディングの背景にある技術課題

2．において、スペクトログラムリーディングのリーディング過程の例と、この手法によって人間が高い精度で音韻を認識できることを示した。しかし、本手法で獲得される情報が人間の経験的知識として表現されるため、音声情報処理技術に応用するためには解決すべき課題は数多く残されている。ここでは、経験的知識を工学的に応用するために解決が試みられている技術課題の中から１）知識処理的アプローチ、２）研究環境の整備、３）音声データベースの整備に関してその概要を紹介する。

3．1．　知識処理的アプローチ

スペクトログラムリーディングから直接得られる知識が人間の経験的知識であることから、近年盛んに研究されている知識工学的技術［７］の枠組みによって、この知識を利用しようとする試みがなされている。Ｚｕｅら［８］、及びＭｉｚｏｇｕｃｈｉら［９］は、視察によって獲得した経験的知識をプロダクションルールによって利用する音声認識システムの構築を試みており、音声情報処理における知識工学の応用として注目されている。また、ＨＥＡＲＳＡＹ－ＩＩ［１０］やＤｅＭｏｒｉらによる階層モデル［１１、１２］等は、人間の聴覚においても用いられていると考えられる音響物理的特徴や、単語、構文、文脈などの様々な知識を階層的に利用して音声認識を行おうとするシステムであり、連続音声認識等の困難な課題を克服するための強力な枠組みを提供している。

しかし、これらのシステムにおいて用いられている音韻特徴は、ルール表現に適した定性的表現や、数量的なものであっても単純な１次元パラメータに基づいたものが多く、必ずしも十分な音声特徴の表現が行われていない。以下に、Ｚｕｅら［８］が用いているルールの例を示す。

```
If the VOT is short,
   and the following vowel is not a schwa,
then the stop is voiced.
```

3．2．　研究環境の整備

スペクトログラムリーディングにおいては、音声の視察や知識の整理が重要な役割を持つため、音声データを扱う研究環境の優劣が本手法の成否に大きく影響する。望ましい研究環境の条件としては、

１）　会話的に音声データを取り扱えること、
２）　音声の様々な音響物理的側面を容易に観察できること、
３）　獲得した経験的知識を計算機上に容易に表現できること、
４）　獲得した経験的知識の定量化を行うため、大量の音声データを容易に扱えること、

等があげられる。このため、音声の視察や音響入出力、記号処理等を総合的に実現した研究環境の構築が盛んに行われている［１３，１４，１５］。

中でも、ＳＰＩＲＥ［１５］とその上に展開されている研究環境は、スペクトロ

グラムリーディングに基づく研究の推進に大きな役割を果たすことが期待される。このシステムは、記号処理に適したリスプマシン上に音声入出力と高速信号処理環境を実現することにより、様々な側面からの音声観察と経験的知識の表現を容易に行える様にしている。

3．3．　音声データベースの整備

音声研究において音声データベースの整備が不可欠であることは言うまでもないが、音韻特徴を解明しようとする研究においては、特に、音響物理的特徴を反映したできるだけ詳細な表記が必要である。これまでにも様々な研究機関がこうしたデータベースの整備［16、17］を行ってきたが、大量かつ詳細な表記の作成が困難なため、音韻特徴を研究するために十分な質、量を兼ね備えたものは必ずしも作られなかった。

こうした現状を踏まえ、ＡＴＲでは、前述した大規模音声データベースの構築を進めている［5］。特に、このデータベースでは、発声内容や音声の物理的特徴を表す表記を正確に付けることを目指しており、現在も訓練された多数の作業員によるデータ作りが進められている。表1は、図3で示したイベント層で用いられている表記記号を示している。なお筆者らは、このデータベースを基に、ホルマントやバズバー、摩擦部等のさらに詳細な音響物理的特徴の表記をスペクトログラムの2次元空間上に実現する作業も進めている。これは、スペクトログラムリーディングによる経験的知識の評価実験にとって有益なデータとなると考えられる。

4．　スペクトログラムリーディングの技術的展望

音声の研究は、工学のみならず言語学、医学等、実に幅広い分野で長年に渡って行われてきた。しかし、こうした研究から得られた知識の定量的な評価や体系的な集積が必ずしもうまくいっていないことが指摘され、これらの課題を克服する手段の一つとして音声に関する知識処理が期待されている［18］。これまで紹介してきた様に、スペクトログラムリーディングとその関連技術は正にこうした知識処理技術の基礎となるものである。本章では、スペクトログラムリーディングから得られる知識をこれまで以上に効果的に利用するために必要と考えられる課題をまとめる。

広く普及しつつある様々なエキスパートシステム［7］、さらには3．1で紹介した音声情報処理における応用例に見られる様に、これまでの知識処理の多くは記号的表現にとどまる経験的知識あるいは観測されるいくつかの数値列を知識表現の対象としてきた。また、スペクトログラムの様な画像の理解の研究においても、そこで扱われている対象は比較的単純な幾何図形や意味が明確な写真等が多い［19］。これに対し、音声のスペクトログラムは非常に複雑な曲面であり、しかもその変形の幅もかなり大きい。このため、知識処理としての実現を急ぐあまり、この図形的な情報を不適切なパラメータ群に変換してしまうのは望ましくない。先に紹介した知識処理的アプローチに立つ音声認識システム以上に、パターンマッチング的手法［20］を取り込んだ知識処理が望まれる。むしろ、非常に多くの知見が得られているパターンマッチング的手法によって扱いきれないデータの音響物理的特性やその背景にある話者情報、音韻環境、発声速度等の属性を明らかにし、パターンマ

ッチングの結果を補う方向で知識を集積すべきものと考えられる。

また、スペクトログラムリーディングにおける時間－周波数空間を同時に扱えるという長所を生かし、従来の音声の特徴表現に最も欠けていた動的特徴［２１］や特徴間の相互の関係の記述、さらに音韻環境によって特徴の音響物理的性質が変化する調音結合の解明を積極的に行うことが望まれる［２２］。さらに、日本語と英語の様な異なる言語における音響物理的特徴の比較を行うことも、音韻特徴の解明に大いに役立つものと考えられる［２３］。なお、スペクトログラムリーディングの過程で用いられる経験的知識は、人間の聴覚過程におけるそれとは必ずしも一致しない。視覚的に得られた知識と聴覚実験結果との比較も重要である。

大規模音声データベースが重要であることは先にも述べたが、様々な研究手法の正確な評価を行うためにも、これらのデータベースは広く公開されることが望ましい。スペクトログラムリーディングという手法が、音声生成過程の研究や言語学等の広い研究成果に基づいた知識を系統的にまとめようとする様に、そこで扱われる音声データも系統的かつ普遍的なものであることが望まれる。

４．　まとめ

以上、スペクトログラムリーディングとそれに関連する技術的課題とについて概説した。本手法は、「音声を読む」と言う魅惑的かつ挑戦的な名前を持つものの、実は音声研究の極めて正統的かつ地味な手法である。連続音声認識技術や高品質音声合成技術の実現を目指し、これまでの研究成果を整理する一手法として、さらに音韻特徴の解明を試みる手法として広く行われることが望まれる。

参考文献

〔1〕 A.Cole(editor); Perception and Production of Fluent Speech, Chapter 1, pp. 3-50,LAWRENCE ERLBAUM ASSOCIATES PUBLISHERS, 1980.

〔2〕 B.T.Oshika, V.W.Zue, R.V.Weeks, H.Neu and J.Aurbach; The Role of Phonological rules in Speech Understanding Research, IEEE Trans., Vol. ASSP-23, No.1, February 1975.

〔3〕 M.A.Bush, G.E.Kopec, and V.W.Zue; Selecting Acoustic Features for Stop Consonant Identification, IEEE, ICASSP 83, pp. 742-745, April, 1983.

〔4〕 B.G.Greene, D.B.Pisoni and T.D.Carrell; Recognition of speech spectrograms, J.Acoust. Soc. Am. 76(1), July 1984.

〔5〕 武田、匂坂、片桐；音声データベース構築のための音韻ラベリング、音響学会、講論集、３月、１９８７.

〔6〕 片桐、武田、匂坂；視察に基づく音韻ラベルの性質、音響学会、講論集、３月、１９８７.

〔7〕 F. ヘイズーロス、D. A. ウォーターマン、D. B. レナート；エキスパートシステム、産業図書、１２月、１９８５.

〔8〕 V.W.Zue and L.F.Lamel; An Expert Spectrogram Reader: A Knowledge-Based Appraoch to Speech Recognition, ICASSP 86 Proc., Vol. 2, pp. 1197-1200, April 1986.

〔9〕 R.Mizoguchi, K.Tsujino and O.Kakusho;, A Continuous Speech Recognition System Based on Knowledge Engineering Techniques, ICASSP 86 Proc., Vol.2, pp.1221-1224, April 1986.
〔10〕 A.Cole(editor); Perception and Production of Fluent Speech, Chapter 9, pp.215-242, LAWRENCE ERLBAUM ASSOCIATES PUBLISHERS, 1980.
〔11〕 R.De Mori, P.Laface and Y.Mong; Parallel Algorithms for Syllable Recognition in Continuous Speech, IEEE Trans., Vol. PAMI-7, No. 1, pp. 56-69, January 1985.
〔12〕 R.De Mori and L.Lam; Plan Refinement in a Knowledge-Based System for Automatic Speech Recognition, ICASSP 86 Proc., Vol. 2, pp. 1217-1220, April 1986.
〔13〕 G.Kopec; The Integrated Signal Processing System ISP, IEEE TRANS, Vol. ASSP-32, No. 4, pp. 842-851, AUGUST 1984.
〔14〕 片桐、古井；スペクトルリーディングに基づく音韻特徴探索システム、音響学会音声研究会資料　Ｓ８４－５６、１１月、１９８４.
〔15〕 V.W.Zue, D.S.Cyphers, R.H.Kassel, D.H.Kaufman, H.C.Leung, M.Randolph, S.Seneff, J,E,Unverferth,III, and T.Wilson; The Development of the MIT Lisp-Machine Based Speech Research Workstation, ICASSP 86 Proc., Vol.1, pp.329-332, April 1986.
〔16〕 板橋秀一；日本語共通音声データ、信学会情報・システム部門別全国大会論文集、分冊１、Ｓ４－７、ｐｐ．１－３１３　－　１－３１２、１１月　１９８５.
〔17〕 田中、速水、太田；研究用音声データベース構築のためのテキスト生成、ラベリング技法、システム構成、信学会情報・システム部門別全国大会論文集、分冊１、Ｓ４－６、ｐｐ．１－３１１　－　１－３１２、１１月　１９８５.
〔18〕 新美康永；連続音声認識における最近の話題、信学会情報・システム部門別全国大会論文集、分冊１、Ｓ５－１、ｐｐ．１－３２７　－　１－３２８、１１月　１９８５.
〔19〕 長尾　真；コンピューターのパターン認識、東大出版会、ｐｐ．１１３－１５４，１２月、１９８５.
〔20〕 古井貞熙；ディジタル音声処理、ｐｐ．１４９－１９２、東海大学出版会、１９８５.
〔21〕 S.Furui; Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum, IEEE trans., Vol. ASSP-34, No.1, February, 1982.
〔22〕 片桐、古井；音韻特徴探索システムを用いた調音結合の分析、信学会情報・システム部門別全国大会、分冊１、Ｓ４－８、ｐｐ．１－３１５　－　１－３１６、１１月　１９８５.
〔23〕 V.W.Zue, et al.; Textbook of MIT Speech Spectrogram reading, ATR, pp.II-56 - II-96, January, 1982.