日本語音声におけるパワースペクトル因子の音声知覚上の役割

岸田, 拓也

https://doi.org/10.15017/1931919

出版情報：九州大学, 2017, 博士（芸術工学）, 課程博士
バージョン：
権利関係：

日本語音声におけるパワースペクトル因子の音声知覚上の役割

Perceptual roles of spectral-change factors in Japanese speech

岸田　拓也

Takuya Kishida

2018年3月

目 次

第1章　序論

はじめに

生物どうしによるコミュニケーションは視覚、聴覚、触覚、嗅覚など、様々な感覚を通して行われている。言葉を用いるというヒト特有の性質から、様々な感覚の中でも聴覚を用いたコミュニケーションは我々にとって特に重要であると言えるだろう。言葉は主として音声によって伝えられる。音は振動が空気中を伝搬する物理的な現象に過ぎないが、ヒトは聴覚器官と発声器官とを巧みに利用することで、音声を使って複雑なコミュニケーションができるように進化した。この複雑なコミュニケーションを支える音声知覚の仕組みを解明することは、音声にかかわる研究分野における最も重要なテーマの一つであると言える。

本論文では、聴覚系末梢における音声のスペクトル表現に対し統計的手法を用いることでその主要な成分を取り出し、音声知覚のための手がかりとしてそれらの成分がどのように利用されているのかを聴取実験によって調べた。音声の音響的特徴に含まれる音声知覚の手がかりに冗長性がどれだけあるのかを、合成音声を用いた聴取実験によって調べる研究と、統計的手法による分析を通して調べる研究とを結びつける研究である。

本章では、まず、音声知覚に関する研究背景として、本論文で取り扱った内容と特に関連の深い先行研究を紹介する。先行研究によって明らかとなったことを整理する中で、検討が不十分な点を取り上げる。そして、本論文の目的を示す。最後に、本論文全体の構成を説明する。

1.1 研究背景

言語音としての音声を用いた聴覚コミュニケーションの枠組みには、いくつかの段階があると考えられている (de Saussure, 1959)。図 1.1 は、その聴覚コミュニケーションの様々な段階を模式的に示したものである。話者が自身の考えや感情を他者に伝えようとするとき、まずはその考えや感情が話者の脳内で言語化される。言語は話者の脳内では聴覚イメージとして表現されている。この聴覚イメージを実際の物理的現象としての音波として実現するために、脳から調音器官へ運動指令が送られ、発声が起きる。発声によって生じた音波が聴取者の耳に入力されると、空気の振動であった音波が聴覚系の処理を経て聴取者に聴覚イメージを引き起こさせる。聴取者の脳内でもまた、聴覚イメージは言語そのものと結びついている。よって、聴取者に話者が伝えようとした言語内容が伝わる。以上が聴覚コミュニケーションの基本的枠組みである。Denes and Pinson (1993) はこの枠組みの中に、話者が発話した音声を話者自身が聴取するという段階もあることを述べており、一連の段階が鎖のように連なる様子から、聴覚コミュニケーションの枠組みを言葉の鎖と呼称している。この言葉の鎖において、音波を聴覚イメージと対応づける段階が音声知覚である。

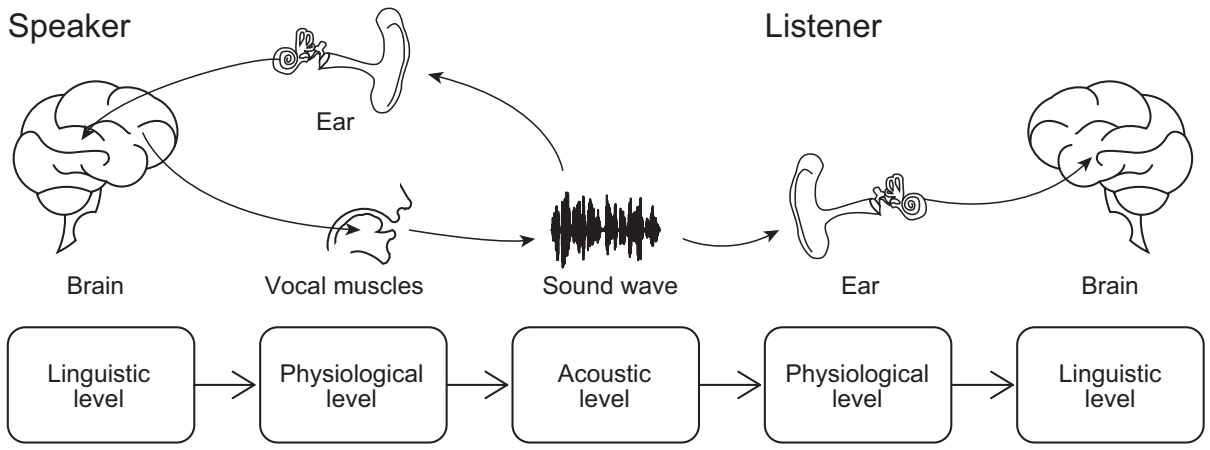

図 1.1 聴覚コミュニケーションの様々な段階 (言葉の鎖)。Denes and Pinson (1993) に掲載の図を参考に筆者が作成。

音声知覚は実験心理学・言語学・聴覚研究・電気工学・人工知能研究などの様々な分野の研究者が関心を寄せる、学際的な領域である (Pisoni, 1985)。様々な立場の研究から得られた知見は、外国語の習得、効率的な音声信号の伝送技術、音声強調技術、自動音声認識技術、難聴

者の装用する補聴器・人工内耳の開発など、豊かな暮らしを設計するために利用できる。

音声には、言語内容以外にも話者の性質･状態･感情といったものを伝える役割がある (Schuller et al., 2013) が、言語内容の伝達に限って言えば、音声知覚の研究はおおよそ 2 つの研究領域に分けられるだろう (Plomp, 2002; Samuel, 2011)。 1 つ目は、音声におけるある音響的特徴が、音素、音節、単語といった音声の構成単位の知覚とどのように関係しているのかを調べる領域である。そしてもう 1 つは、日常会話のように連続的に発話された音声を聴いて、どのようにしてその音声を言語として認識し、処理しているのかを調べる領域である。本研究では、文単位の音声を対象として、音声信号の分析と聴取実験を行った。連続的に発話された音声を扱うという点では本研究は後者の研究領域であるとも言えるが、音声の音響的特徴について知覚と対応づけるという点では前者の領域とも関連が深い。そこで研究背景として、前者と後者の研究領域の違いにこだわらず、本研究と関連が深い先行研究を紹介する。

1.1.1 音素の知覚

音素は音声の最小構成単位である。音素は母音と子音とに大別でき、共通する部分はあるものの言語ごとにその種類と体系は少しずつ異なっている。音声知覚の最初期の研究では、音素を同定するのにどのような音響的特徴が必要であるのかが詳しく調べられた。音を視覚的にとらえることを可能とする、サウンドスペクトログラフ (Potter, 1945) という装置がある。この装置に音声信号を入力すると，信号が解析され、含まれている周波数成分の時間変動が紙上に濃淡で描かれたもの (スペクトログラム) を出力する。図 1.2 は音声の時間波形とその時間波形を解析することで得られるスペクトログラムの例である。一方パターンプレイバック (Cooper, Liberman, & Borst, 1951) は、サウンドスペクトログラフとは逆の発想によるもので、スペクトログラムを模擬して描かれたパターンを読み込み、音を再生するという装置である。パターンプレイバックで音声を模擬した音響信号が作られ、これを使った音素の知覚実験が行われた。ハスキンス研究所 (Haskins Laboratories) の研究者によって行われたこれら最初期の研究から、音素の知覚にはフォルマントが重要であることが明らかにされた (Delattre, Liberman, Cooper, & Gerstman, 1952; Liberman, 1957)。声帯の振動で生じた音波は口唇から放射されるまでの間に周波数ごとに振幅が強められたり弱められたりされる。どの周波数においてどの程度振幅の強弱の変化が起きるのかは、声帯から口唇までの形状、すなわち声道形状による共振特性で決まる。フォルマントはそのようにしてできる音声のスペクトル包絡上の山のことである。 そのピークとなる周波数がフォルマント周波数であり、値が低い順に第 1 フォルマ

ント、第 2 フォルマント、第 3 フォルマントというように呼ばれる。母音の発声時はフォルマントが定常的に観察されるため、フォルマント周波数の分布のパターンを聴き分けることで母音を知覚することができ、一方子音はフォルマント周波数の遷移パターンが特徴的であり、これを手がかりに知覚できると考えられている。図 1.3 に日本語の母音、/a/と/i/のフォルマントのパターンの違いを見ることができる。母音、/a/と/i/をそれぞれ発声する際は声道の狭められる位置が異なるため、共振特性が変化し、それがフォルマントパターン (スペクトル包絡) として現れる。一方スペクトルの細かな変動パターン (微細構造) は、声帯振動によって作られるため、同じ声の高さで発声しようとすれば、その変動の間隔は二つの母音の間でもほとんど変わらない。

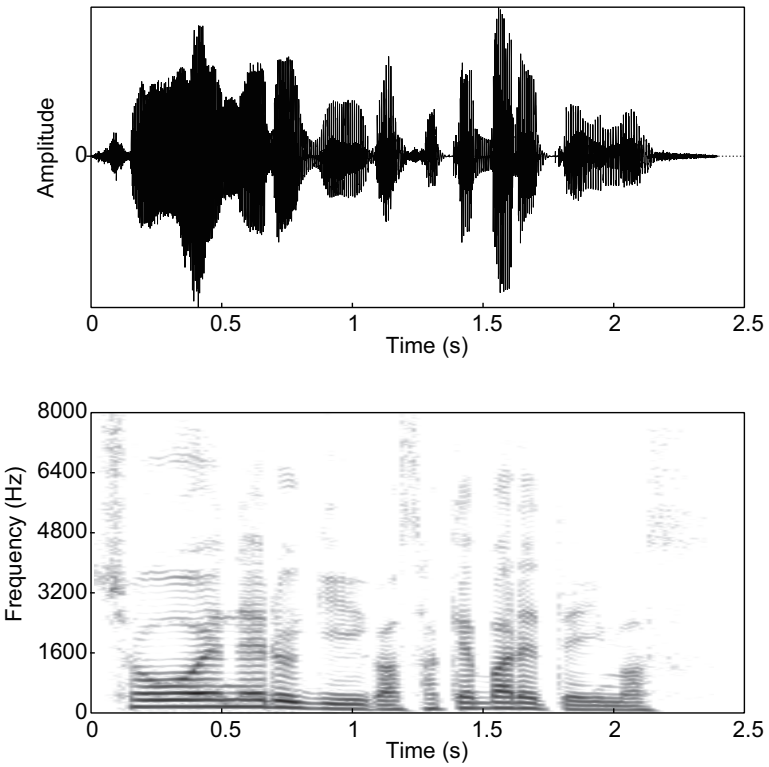

図 1.2 音声の時間波形 (上) およびそのスペクトログラム (下) の例。「省エネルギーが叫ばれています。」という文内容を男性が発話したもの。NTT-AT 多言語音声データベース 2002 に収録。

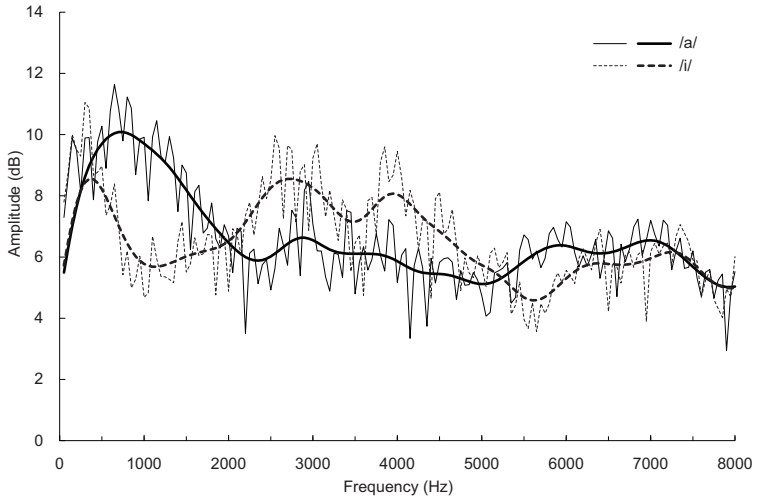

図 1.3 日本語の母音のスペクトル (細い実線が/a/、細い破線が/i/) とそのスペクトル包絡 (太い実線が/a/、太い破線が/i/)。筆者が発声したものを録音、分析した。

音素の同定にはフォルマントが重要であることが分かったが、後続する母音の種類が変わると、同じ子音であってもそのフォルマント周波数の時間遷移パターンが非常に異なることが同時に観察され、特定の子音であると同定できる不変的な音響的特徴を見出すことは困難であることが明らかとなった。そこで、Liberman らは、音素と調音運動との間の一貫した対応関係に着目して、音響的特徴ではなく、調音器官を動かす筋肉への運動指令に子音を同定する不変的特徴を見出すことができるだろうと主張した (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967)。これが音声知覚の運動理論 (Motor Theory) である。さらに運動理論では、ヒトが音声を知覚するときはそれ以外の音の知覚とは別に、専用の機構を利用していると主張している。運動理論を支持する研究者は、例えば、二重知覚 (Rand, 1974) やカテゴリー知覚 (Liberman, Harris, Hoffman, & Griffith, 1957) といった現象を発見し、これが音声刺激を用いたときにだけ特有に起こる現象であるとして運動理論を証明しようとした。音声知覚の運動理論は、後の音声知覚研究に与えた影響の大きさゆえに、最も重要な理論の一つであると言える。

運動理論に対抗する形で、音声知覚の理論が複数提唱されている。ここでは、Blumstein と Stevens が主張する、音響的不変性理論について触れる。この理論はスペクトルの全体構造が手がかりとなるという点で本論文の内容と関連する。運動理論の支持者らは、音声の音響的特徴の中に音素を同定できる不変的な特徴を見出すことを諦めたが、Blumstein and Stevens (1979, 1980) は音素を二項対立素性を使って分類するという考え方に従って、音声の音響的

特徴から音素を同定する不変的特徴を見つけようとした。例えば閉鎖子音の閉鎖の開放から20–30 ms までの区間のスペクトルにおいて、エネルギーが周波数軸上の中心に集約しているか全体的に拡散しているか (集約性–拡散性の対立)、そして拡散している場合は、高域に向うにつれてフォルマントのピークの振幅が増加しているか減少しているか (高音調性–低音調性の対立) を観察することによって、閉鎖子音の調音位置による 3 種類の分類が可能であることを示した。しかし一方で、不変的な特徴と同時に後続母音の違いによって変わる特徴が与えられたときは、聴取者は変わる特徴の方を基にして音素を判断する傾向があることが示されており (Blumstein, Isaacs, & Mertus, 1982; Walley & Carrell, 1983)、音響的不変性理論は完全というわけではない。

他にも音声知覚の理論には代表的なものとして、直接実在理論 (Direct Realism Theory; Fowler, 1991), 一般アプローチ (General Approach; Diehl, Lotto, & Holt, 2004) などがある。これらの理論は、二重知覚現象が非音声でも生じること (Fowler & Rosenblum, 1990) や、カテゴリー知覚がヒト以外の生物にもみられる (Kluender, Diehl, Killeen, et al., 1987) という実験報告を基に立てられたものである。音声知覚の理論におけるそれぞれの考え方は、知覚の対象が調音運動であるかどうか、音声知覚が専用の機構によるものかどうかという立場で分類できる (Diehl et al., 2004)。しかし、誰しもが納得する理論はなく、現在でも活発な議論が続いている。

1.1.2 知覚の手がかりの冗長性

実環境には音声以外にも様々な音があふれており、会話中に無関係の音を音声と同時に聴く状況はしばしばある。さらには、音はすぐに立ち消えてしまうものである。よって音声を使って安定したコミュニケーションを行うためには、音声にある程度余剰に知覚の手がかりとなる情報が含まれていてしかるべきである。実際に、音声における知覚の手がかりの冗長性を示す様々な報告がなされている。ここでは周波数情報に関する冗長性について紹介する。

内耳の組織である蝸牛に音が入力されたとき、異なる周波数ごとに蝸牛の基底膜上で強く振動する場所が異なることから、我々の耳は周波数分析器としてはたらくことが明らかにされている (Schnupp, Nelken, & King, 2011; Plack, 2014)。この聴覚系の周波数分析機能は中心周波数と帯域幅のことなるフィルタが複数並んだフィルタバンクとしてモデル化できる。このフィルタが聴覚フィルタ (Patterson, 1974; Unoki, Irino, Glasberg, Moore, & Patterson, 2006; Moore, 2013) と呼ばれるものである。臨界帯域 (Fletcher, 1940; Zwicker & Terhardt,

1980; Greenwood, 1990; Schneider, Morrongiello, & Trehub, 1990) は聴覚フィルタを矩形に近似したものである。同時マスキングを利用した聴取実験から、臨界帯域の帯域幅が求められる (Fletcher, 1940; Zwicker & Terhardt, 1980)。臨界帯域幅は中心周波数が約 500 Hz までは一定の 100 Hz 程度であるが、500 Hz 以上ではおよそ中心周波数の 20%の広さになるという特徴をもつ。しかし聴覚フィルタの形状は実際には矩形ではない。Patterson (1974) によって聴覚フィルタの形状を求める実験手法が提案され、聴覚フィルタの形状が明らかになっていった。聴覚フィルタは中心周波数に対して対称な形状ではなく、低域側はなだらかにフィルタ出力が低下し、高域側は急峻にフィルタ出力が低下するという特徴がある。さらに、低域において一定の帯域幅と考えられていた聴覚フィルタは実際には低域ほど狭くなるということも明らかとなった。また、聴覚フィルタの形状は入力レベルによっても変化することも分かっている。中程度のレベルにおいては、フィルタ出力は対数周波数軸上で対称であると考えてよい。聴覚フィルタの帯域幅を等価矩形帯域幅に換算することは有用である。等価矩形帯域幅とは、聴覚フィルタが通す白色雑音のパワーと同量のパワーを通すような矩形フィルタの帯域幅のことである (Moore, 2013)。矩形フィルタの高さは聴覚フィルタの中心周波数における高さにそろえられる。図 1.4 で中程度のレベルの音に対する聴覚フィルタの等価矩形帯域幅と臨界帯域幅とを比較することができる。比較的高帯域においては、聴覚フィルタの等価矩形帯域幅と臨界帯域幅は同程度であることがわかる。 臨界帯域の帯域幅は基底膜の 1.3 mm 分の長さに対応し (Fastl & Zwicker, 2006)、これは帯域の中心周波数のおよそ 1/4〜1/3 オクターブ程度である (Plomp, 2002)。これが聴覚の周波数分解能であるが、次に示すいくつかの例のように、音声の知覚においてはこの分解能は十分すぎる。

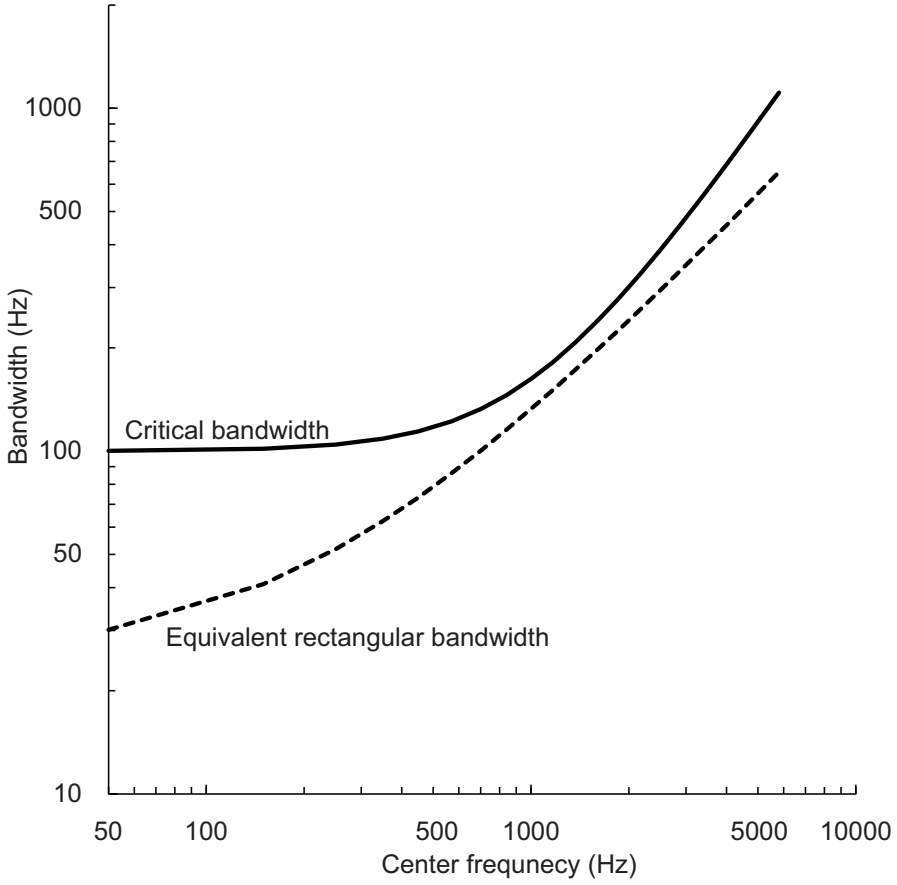

図 1.4 臨界帯域幅と聴覚フィルタの等価矩形帯域幅の比較。両軸ともに対数軸。Zwicker and Terhardt (1980) および、Moore (2013) を参考に作成した。

音声はおよそ 50–8000 Hz の周波数帯域にエネルギーが分布しており、その範囲に音声知覚のための様々な手がかりが与えられている。同じ遮断周波数において低域通過フィルタに通された音声と高域通過フィルタに通された音声との 2 つの条件で、音声の明瞭度が遮断周波数の変化とともにどのように変化するかを調べた研究が複数ある (French & Steinberg, 1947; Hirsh, Reynolds, & Joseph, 1954; Miller & Nicely, 1955; Studebaker, Pavlovic, & Sherbecoe, 1987)。遮断周波数が低い場合は、高域通過フィルタに通された音声の方が明瞭度が高く、その逆の場合は低域通過フィルタに通された音声の方が明瞭度が高くなるわけだが、2 つの条件で同じ明瞭度となるところの遮断周波数 (およそ 1700 Hz 付近) において、多くの研究で音節正答率や単語正答率が 50%以上となることが報告されている。このことは相補的な音響的情報それぞれだけを用いて、音節あるいは単語の知覚がある程度可能であることを示しており、知覚の手がかりの冗長性を示す一つの例である。また関連する研究として、1/3 オクターブの広さの狭帯域フィルタに通された音声信号でも、非常に高い明瞭度が得られることが、さら

に 1/20 オクターブのより狭い帯域フィルタに通された音声信号であっても、フィルタの中心周波数が 1500 Hz 付近であれば相当高い明瞭度が得られることが Warren, Riener, Bashford, and Brubaker (1995) によって報告されている。

上にあげた周波数帯域を制限する研究の他に、スペクトル全体に含まれる情報を劣化させた音声による聴取実験でも、音声知覚の手がかりに冗長性があることが報告されている。Ter Keurs, Festen, and Plomp (1992, 1993) は、音声のスペクトル包絡の変化をガウスフィルタを使って鈍らせ、スペクトル包絡上の山が低く、谷が浅くなった音声が定常雑音下でどれだけ明瞭に聴きとれるかを調べた。ガウスフィルタの帯域幅を変えて、スペクトル包絡の変化の鈍さの異なる音声で比較したところ、ガウスフィルタの帯域幅が 1/3 オクターブまでの音声は、処理を行わない条件の音声と同等に明瞭であることが分かった。それだけでなく、4 オクターブの帯域幅でスペクトル包絡を鈍らせても、雑音のレベルに対して十分に音声のレベルを大きくすれば、明瞭に音声を知覚できることも分かった。ここで彼らは、音声が男性のものであっても女性のものであっても結果が変わらないということから、スペクトルに含まれる微細構造よりもスペクトル全体の包絡構造が音声の明瞭度を決定づける要因であると考察している。

スペクトル全体の情報を劣化させた音声の別の例として、チャンネルボコーダ音声 (Dudley, 1939) がある。チャンネルボコーダ音声は、音声をいくつかの周波数帯域 (チャンネル) に分けて処理をすることで、そのチャンネルにおける振幅包絡のみを取り出し、その取り出された振幅包絡で別の信号 (搬送信号) の対応するチャンネルをそれぞれ駆動することで合成される音声信号である (図 1.5)。 搬送信号が帯域雑音の場合は雑音駆動音声 (noise-vocoded speech)、正弦波の場合は正弦波駆動音声 (sine-vocoded speech) と呼ばれる。チャンネルボコーダ音声はチャンネル数を変化させることで段階的にスペクトル情報を劣化させることができる。チャンネルボコーダ音声を用いた研究の成果は人工内耳の周波数チャンネルの設定などに利用されている (Xu & Pfingst, 2008)。

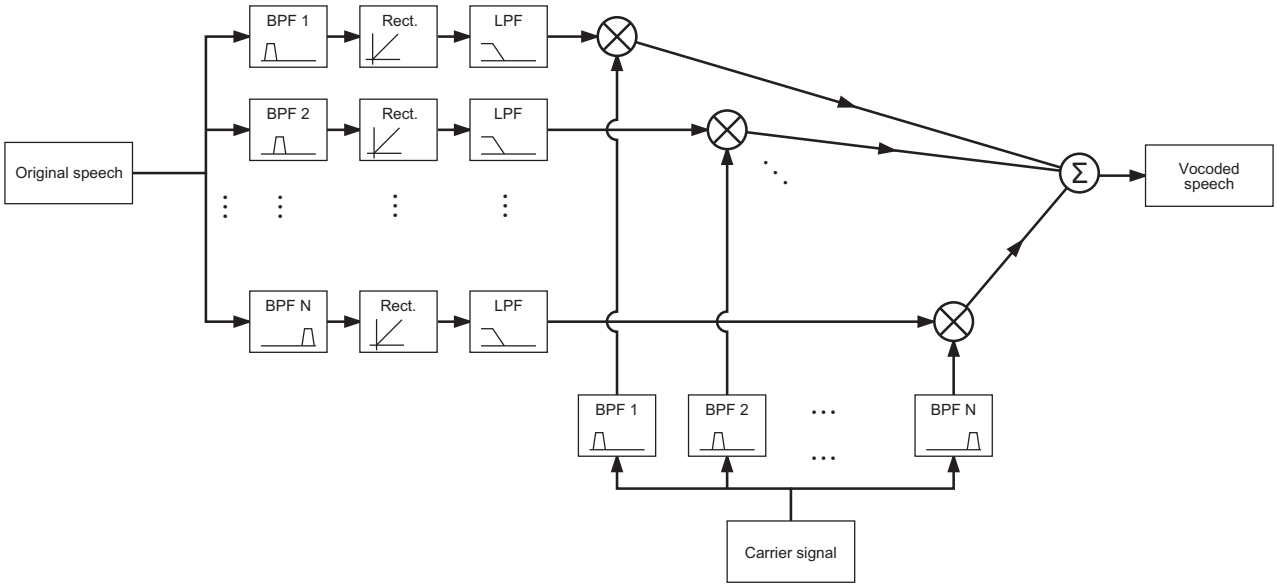

図 1.5 N チャンネルからなるチャンネルボコーダ音声の作成手順を示す流れ図。中心周波数の異なる N 個の帯域通過フィルタ (図中の BPF) に通されたそれぞれの音声信号を、さらに整流 (図中の Rect.) し、低域通過フィルタ (図中の LPF) に通すことで各周波数帯域における音声信号の振幅包絡が得られる。この振幅包絡で搬送信号 (雑音または正弦波等) の対応する周波数帯域を振幅変調し、各帯域の信号を足し合わせることでチャンネルボコーダ音声が合成される。

Shannon, Zeng, Kamath, Wygonski, and Ekelid (1995) の実験では、音声を 4000 Hz 以下に帯域制限したうえで、4 チャンネルの雑音駆動音声として合成した場合でも、文音声であれば単語正答率が 90%を超えることが報告されている。また、雑音駆動音声と正弦波駆動音声とを比較した実験では、どちらの条件でも 4 チャンネルで文音声の単語正答率が 90%を超える結果が得られている (Dorman, Loizou, & Rainey, 1997)。さらに同様の実験は、複数の話者を使う (Loizou, Dorman, & Tu, 1999)、若年者と高齢者の 2 つの実験参加者グループに分ける (Sheldon, Pichora-Fuller, & Schneider, 2008)、チャンネル内の時間情報を変化させる (Souza & Rosen, 2009), 言語やチャンネルの遮断周波数を変える (Ellermeier, Kattner, Ueda, Doumoto, & Nakajima, 2015) など、様々な条件で行われてきたが、いずれの実験においてもチャンネルボコーダ音声は 4 〜 6 帯域程度あれば十分に明瞭になることが分かっている。このように少ない帯域数のチャンネルボコーダ音声でこれだけ明瞭に音声を知覚できるのは、チャンネル内の時間変化の情報が第 1 の知覚の手がかりであるからだと考察されている (Shannon et al., 1995) が、チャンネル間のレベル差によって与えられるスペクトルの大まかな構造が手がかり

となっていると報告する研究もある (Roberts, Summers, & Bailey, 2010)。チャンネルボコーダ音声を用いた研究は、どれだけ音声に冗長性があるのかを示してきた。しかしながら、なぜチャンネルボコーダ音声で明瞭に音声を知覚することができるのかという問いに対しては十分に答えることはできていない。また、どの周波数帯域の情報の寄与が明瞭度に与える効果が大きいかについても十分に検討されていない。

1.1.3 スペクトルの全体構造がもつ手がかり

ここまで、音声知覚においては、スペクトルの全体構造が手がかりであるという可能性を繰り返し示してきた。このようなスペクトルの全体構造がもつ手がかりについて別の角度から調べる方法として、音声の音響的特徴を統計的手法によって分析する研究がある。ここでは、統計的手法による音声の分析についての先行研究を紹介し、本論文のテーマとなるパワースペクトル因子について導入する。

Plomp, Pols, and Geer (1967) はオランダ語の 15 の母音のスペクトルを臨界帯域幅に近い、1/3 オクターブバンドで 18 帯域に分割し、それぞれの帯域のパワーを基に主成分分析を行った。彼らは主成分分析によって、スペクトルの全体的な特徴がどのような単純なパターンによって構成されるのかを確かめたのである。その結果、第 2 主成分まででデータのおよそ 70%が説明できると分かり、15 の母音を第 1 、第 2 主成分空間で十分に区別可能であることが示された。また、この第 1 、第 2 主成分空間上における母音の配置は、第 1 フォルマント周波数と第 2 フォルマント周波数の対数値を軸とする平面上における母音の配置と対応付けられることが分かった (Pols, Tromp, & Plomp, 1973; Plomp, 1976, 2002)。このようなデータを背景に、母音の識別において、フォルマント周波数よりもむしろスペクトル全体の形状が手がかりとして有用であることを Zahorian and Jagharghi (1993) が示している。

Ueda and Nakajima (2017) は Plomp et al. (1967); Pols et al. (1973); Plomp (1976, 2002) の分析手法を、 8 つの異なる言語・方言における連続的に発話された文音声を対象に拡張した。彼らは、Zwicker and Terhardt (1980) を参考として設定した 20 の臨界帯域で、文音声のパワースペクトルの時間変動を分割し、1 ms 毎に 20 帯域のパワー値として取り出されたものを因子分析にかけた。この分析によって 3 つないし 4 つの因子を取り出すと、 8 つ全ての言語において共通するパターンの因子が得られることを彼らは発見した。この結果は言語を超えた普遍的な音響的特徴が音声に含まれていることを示すものである。本論文ではこの因子を、パワースペクトルの時間変動を構成することから、「パワースペクトル因子」と呼ぶこととする。

これまでにおいて、音声のスペクトルの全体的な形状の統計的分析から得られるスペクトルを構成する主要な特徴が、音声知覚の手がかりとしてどのように機能するかについて調べた研究は筆者が探す範囲では見当たらない。このような統計的手法を用いた音声のスペクトルの分析結果を、音声知覚の仕組みと直接結びつけて考察するというよりも、取り出された主成分空間を用いて母音を自動的に識別する技術に利用する方向に研究が発展しているようである。Ueda and Nakajima (2017) が取り出した 3 つまたは 4 つのパワースペクトル因子は、音声のスペクトルを 4 帯域に分割するような特徴を持っている。このことは、先に述べた 4 帯域のチャンネルボコーダ音声が高い明瞭度をもつことと何らかの関連があることをうかがわせる。Ellermeier et al. (2015) は、Ueda and Nakajima (2017) が取り出した 3 つないし 4 つのパワースペクトル因子によって分割される 4 帯域に従ってドイツ語・日本語の 4 帯域雑音駆動音声を合成し、聴取実験によってこの 4 帯域雑音駆動音声が高い明瞭度をもつことを確かめている。

この研究はパワースペクトル因子がもつ音声知覚の手がかりについて考察するための重要な報告である。しかしながら、より直接的な考察を行うためには、パワースペクトル因子が表現しうる情報だけを持つ音声を合成し、それを用いた聴取実験を行うことが必要であろう。Zahorian and Rothenberg (1981) においてそのような試みがなされている。彼らは Plomp et al. (1967) が行った主成分分析と同様の方法で取り出された主成分から音声を再合成し、その音声の明瞭度を測定しているが、彼らの研究は分析における最適な条件の探索に主眼が置かれてあり、音声知覚において主成分がもつ意味についての考察は十分になされていなかった。

1.1.4 言語のリズムと鳴音性

音声知覚の研究は単独で発話された音素、音節、単語の知覚を調べる領域と、会話時のように連続的に発話された音声の知覚を調べる領域に分かれると最初に述べた。日常の中で発話される音声は音響的には切れ目のない連続体である。この連続的な音を聴いて言語として正しく認識するためには、切れ目のない音声をある単位に分節するという処理が行われなければならない。では、どのような単位に分節されて音声は知覚されているのであろうか。ここでは、言語のリズムに焦点を当ててこの問題について取り上げる。

言語にはリズムがある。単語・文節・文によってリズムは階層的に作られ、リズムは話者の感情を伝える目的や、特定の語を強調する目的で用いられることもある (Handel, 1989)。異なる言語の音声どうしを聴き比べてみれば、そのリズムが異なることに気づくだろう。実際に言語はそのリズム構造によっていくつかの代表的なグループに分類される。Ramus, Nespor, and

Mehler (1999) は様々な言語の音声を分析し、母音の区間の割合と 1 文内の子音の区間の割合の標準偏差とで表される平面に各言語を配置すると、3 つのグループに分かれて配置されることを示した。この 3 つのグループは言語のリズム構造の代表的なグループとされる、ストレスタイミング言語 (stress-timed language)、音節タイミング言語 (syllable-timed language)、モーラタイミング言語 (mora-timed language) にそれぞれ対応している。例えば英語・ドイツ語はストレスタイミング言語、フランス語・イタリア語は音節タイミング言語 (Ladefoged & Johnson, 2011)、そして日本語・タミル語はモーラタイミング言語である (Port, Dalby, & O' Dell, 1987; Ramus et al., 1999)。言語におけるリズムの役割の一つは知覚の単位を形成することであると言われている (Cutler, 1994)。知覚実験によって、英語音声がストレスの単位で、フランス語音声が音節の単位で知覚されていることを (Cutler, Mehler, Norris, & Segui, 1986)、日本語音声においてはモーラを単位にして知覚されていることを (Otake, Hatano, Cutler, & Mehler, 1993) 示すデータが得られている。

言語のリズムとパワースペクトル因子に関連があることを示す研究がある。Yamashita et al. (2013) は、英語と日本語のそれぞれの言語環境下で育てられた乳幼児が自然に発声した声を継続的に録音し、乳幼児の成長の過程において音声の音響的特徴にどのような変化が見られるのかを観察した。彼女らは月齢 15、20、24 か月の 3 つの時期の音声に対して Ueda and Nakajima (2017) と同じ方法で因子分析を行い、月齢が高い乳幼児の音声ほど、パワースペクトル因子のパターンが成人のものに近いことを見つけた。さらに、因子分析で 3 因子を取り出したうちの一つである 1100 Hz 付近の中帯域に大きい因子負荷量をもつ因子の因子得点について、その自己相関関数を求めることでその因子得点の時間変動においてリズムパターンのようなものが見られるのかを調べた。自己相関関数の値にはっきりとしたピークが現れたときに、そのピークができる時間を時間間隔とするリズムが形成されていると考えることができる。この分析によって、月齢の高い乳幼児の音声のリズムは成人の音声のリズムに近いことが分かった。

1100 Hz 付近の帯域に大きい因子負荷量をもつパワースペクトル因子は言語のリズムを調べることに利用できることを示したが、この因子についてさらに詳しく分析した Nakajima, Ueda, Fujimaru, Motomura, and Ohsaka (2017) の研究について触れる。Nakajima et al. (2017) はイギリス英語音声に対して Ueda and Nakajima (2017) で用いられた音声のスペクトル構造に対する因子分析を用いて 3 つのパワースペクトル因子を取り出し、各音素の因子得点を観察した。彼らは各音素を因子得点にしたがって 3 因子の空間上に配置させると、各音素があ

る曲線上を、鳴音性 (sonority) の尺度の順に従うように分布する傾向があることを見つけた。さらにそこから、1100 Hz 周辺の中帯域において因子負荷量が大きいパワースペクトル因子に鳴音性の尺度と正の相関があることが、そして 3300 Hz 以上の高帯域において因子負荷量が大きいパワースペクトル因子に鳴音性の尺度と負の相関があることを見つけた。鳴音性とは、それぞれの音素について、それらをどれだけ大きく響かせて発話できるかを示す順序尺度であり、言語学や音声学の研究者らによって提唱されたものである (Selkirk, 1984; Harris, 1994; Spencer, 1996)。de Saussure (1959) は、発話の際にどれくらい調音器官が開いているか、それによってどれだけ音が響くかという観点で音素を開口度 (aperture) という順序尺度で分類している。開口度と呼称されているが、調音と聴覚とが切り離せないものという考えに基づいており、音の聴こえとの関係に重きを置いて考察が進められている。開口度も鳴音性と同様のものであると考えることができる。Spencer (1996) による鳴音性の尺度では、母音、渡り音、流音、鼻音、摩擦音・破擦音、破裂音の順に鳴音性が低くなるとしている。音節は音素が連結されることによって構成されるが、基本的に鳴音性が低い音素から高い音素へとつながり、そしてまた低い音素につながるようになっている。これは鳴音性連続原理 (sonority sequencing principle; Rahilly, 2016) と呼ばれるもので、この規則に従って音素が連なると、鳴音性の山ができる場所に音節の核が形成される。Nakajima et al. (2017) の研究の特筆すべき点は、鳴音性の精神物理学的実体を提案したことであり、この方法で鳴音性を定義すれば、単語 *stop* のような英語において頻出する、摩擦音/s/と破裂音/t/の頭子音連結において、/s/が音節の核とはならないことを鳴音性連続原理に矛盾せずに説明することができる。

音声を聴いたときに感ぜられるリズムは強弱の要素が時間的に規則性をもって並んでいるのを知覚することで形成される。この音声の強弱の要素が鳴音性であると考えられている (Handel, 1989)。Galves, Garcia, Duarte, and Galves (2002) も鳴音性が言語のリズムと関連があることを異なるリズム構造をもつ言語の音声を音響的に分析することで確かめている。 言語のリズムをとらえることが音声の知覚に重要であることから、鳴音性が音声の知覚にどのように影響を与えるのかについて調べることもまた重要であろう。Ueda and Nakajima (2017)、Nakajima et al. (2017) の研究によって鳴音性がパワースペクトル因子という測定できるものでとらえることができるようになった。よってパワースペクトル因子から直接音声を再合成すれば、音声知覚と鳴音性の関係を調べる聴取実験を行うことができる。

1.2　本論文の目的

本論文では、音声の臨界帯域ごとのパワー変動を構成するパワースペクトル因子について、音声知覚におけるその役割を聴取実験によって調べることを目的とする。これを実現するために、パワースペクトル因子から音声を再合成する手法を確立する。合成音声を用いた音声の明瞭度を測定する聴取実験と、統計的手法による音声のスペクトルの構造分析とを結びつける研究に位置づけられる。

1.3　本論文の構成

第 1 章では、本研究の背景として、聴覚系末梢によって得られる音声のスペクトル表現が音声の知覚の際にどのように利用されているのかについて、聴取実験を通して調べた先行研究および音声の統計的分析を通して調べた先行研究を紹介した。その中で問題点を整理し、本研究の目的を示した。

以降、第 2 章から第 4 章において、本研究で行った 2 つの分析および 4 つの実験について報告する。第 2 章の分析 1 では Ueda and Nakajima (2017) による音声の臨界帯域ごとのパワー変動に対する因子分析によって得られるパワースペクトル因子から、臨界帯域ごとのパワー変動を再構成し、聴取実験に用いるための刺激音を作成するのに適した因子分析法を提案する。この因子分析法によって日本語・イギリス英語・中国語 (普通話) の音声を分析し、得られたパワースペクトル因子が Ueda and Nakajima (2017) の分析のものと同等の因子であるのかを確認する。続く実験 1 では、パワースペクトル因子によって表現できる音声のパワースペクトルの時間変化の情報によって、日本語音声がどの程度正確に伝えられるのかを聴取実験で調べる。この実験によってパワースペクトル因子をいくつまで用いれば、十分に明瞭な音声を合成することができるのかを確かめる。第 3 章では第 2 章のパワースペクトル因子を用いた音声の再合成の際に起きる問題点に注目し、この問題を回避する方法としてパワースペクトル因子の直交性を維持したまま非負値化したものに修正する方法を提案する。分析 2 として、パワースペクトル因子による、音声のパワースペクトル変化の説明率がこの修正によってどれだけ影響を受けるのかを調べる。実験 2 では実験 1 と同様の方法を用いて、この非負値化を行ったパワースペクトル因子を用いて合成された音声の明瞭度を測定する。実験 1 と実験 2 の結果を比較することによって、第 2 章における音声の再合成の際に生じていた問題が本研究の目的を達成する上で重要な問題であるのかを検討する。第 4 章では第 2 章および第 3 章で分かっ

た音声の明瞭な知覚において重要となるパワースペクトル因子について、個々の因子の役割に注目する。音声をパワースペクトル因子から再合成する際に、いくつかのパワースペクトル因子によって与えられるパワースペクトルの時間変化の情報を取り除き、再合成された音声の明瞭度がどれだけ低下するのかを調べる。取り除く因子が異なると、明瞭度がどれだけ異なるのかを比較することで、因子のもつ個々の役割について考察する。

第５章では総合考察を行う。まず、第２章から４章において行った分析および実験の結果をまとめ、パワースペクトル因子が音声の知覚にどのような役割をもつのかについて結論を述べる。次に導かれた結論が研究史の中でどのように位置づけられるのか、または先行研究に対してどのような新しい解釈を与えるのかについて考察する。その中で、わずかな帯域数のチャンネルボコーダ音声でなぜ明瞭に内容を知覚することができるかについて、本研究の結論から説明を試みる。最後に研究の問題点について触れ、今後の展望を述べる。

第2章　文音声の明瞭な知覚に要するパワースペクトル因子の個数

2.1　第２章の目的

我々が音声を聴取すると、聴覚の周波数分析機能によって聴覚系末梢において音声のスペクトル表現が得られている。臨界帯域フィルタを用いることで、聴覚系末梢における音声のスペクトル表現を模擬することができる。Ueda and Nakajima (2017) が導出したパワースペクトル因子は 20 個の臨界帯域フィルタによって得られた音声のパワースペクトルをより少ない個数の因子の線形結合によって近似するものである。すなわち、取り出す因子数が 20 個に近いほど臨界帯域フィルタの出力を忠実に再現できる。Ueda and Nakajima (2017) が分析対象とした 8 言語間で共通した因子構造が得られたのは 4 因子までであった。彼らの因子分析は主成分分析を基礎としていることから、この分析結果は、音声のパワースペクトル構造のうちの主要な特徴を構成するための 4 因子が各言語において共通しているということを示している。それでは、言語を超えて共通する特徴を持った音声のパワースペクトル因子は音声を明瞭に知覚するためにどれだけの情報を与えうるのだろうか。

本章では、音声を再合成するのに適したパワースペクトル因子を得ることができる新しい主成分分析、「起点移動主成分分析 (origin-shifted principal component analysis)」を提案する。分析 1 として、提案手法によって音声を分析した場合、先行研究で得られた因子と同等の因子が得られるかを確認する。次に実験 1 として、パワースペクトル因子から日本語音声を雑音駆動音声として再合成した時、因子をいくつまで用いれば再合成された雑音駆動音声を十分明瞭に聴き取ることができるのかを調べる。

2.2 分析１：起点移動主成分分析によるパワースペクトル因子の抽出

従来の主成分分析によって得られるパワースペクトル因子から音声を再合成する場合、後述する定常雑音成分が発生するという問題があり、聴取実験に用いるのが不適切であると考えられる。そこで主成分分析の変法を提案し、この問題を回避することとした。この節では、新しく提案する「起点移動主成分分析」を通して得られるパワースペクトル因子と従来の主成分分析を通して得られるパワースペクトル因子とを比較し、分析法の変更が、結果に対して本質的な影響を与えていないかを確かめる。

2.2.1 分析試料

NTT-AT 社の「多言語音声データベース 2002 (NTT-AT, 2002)」にディジタル収録 (16-bit 量子化、16000 Hz サンプリング) された、日本語、イギリス英語、中国語 (普通話) 音声を分析試料として用いた。日本語音声、イギリス英語音声はそれぞれ 200 文からなり、各言語の母語話者である男性５名がそれぞれの文を発話した。中国語 (普通話) については、78 文からなり、母語話者である男性５名がそれぞれの文を発話した。各文は平均２ s程度の長さで発話されている。すべての分析対象の音声の総発話時間は、日本語、イギリス英語、中国語 (普通話) の順に、2484 s、1979 s、870 s であった。これは、安定した分析結果が得られるのに必要な長さであるとされる 30 s (Li, Hughes, & House, 1969; Zahorian & Rothenberg, 1981) を十分に超えている。音声の平均基本周波数は日本語、イギリス英語、中国語 (普通話) でそれぞれ 136 Hz（SD = 31 Hz)、126 Hz（SD = 30 Hz)、164 Hz（SD = 38 Hz）であった。これらの３つの言語は、異なる言語グループの代表として、Ueda and Nakajima (2017) で分析された８言語の中から選出した。日本語・イギリス英語・中国語 (普通話) は互いに異なる言語リズムを持っており、イギリス英語はストレスタイミング言語、日本語はモーラタイミング言語、中国語は音節タイミング言語である (Cutler, 1994; Ramus et al., 1999)。Ueda and Nakajima (2017) の分析では、男性話者だけでなく女性話者の音声も分析に用いられていた。本論文では、聴取実験で男性話者の音声を原音声に用いた。男性話者の音声は女性話者の音声よりも基本周波数が低い。基本周波数が低い音声はパワースペクトルの包絡形状の情報が取り出しやすい。そのため、男性話者の音声を聴取実験に用いる原音声とした。聴取実験で用いるパワースペクトル因子を得るために、男性話者の音声のみを分析することにした。

2.2.2　手続き

分析手続きを図 2.1 に示す。分析は各音声信号を処理して臨界帯域ごとのパワー変動を得る部分と、臨界帯域ごとのパワー変動を起点移動主成分分析にかけてパワースペクトル因子を得る部分とに分けられる。

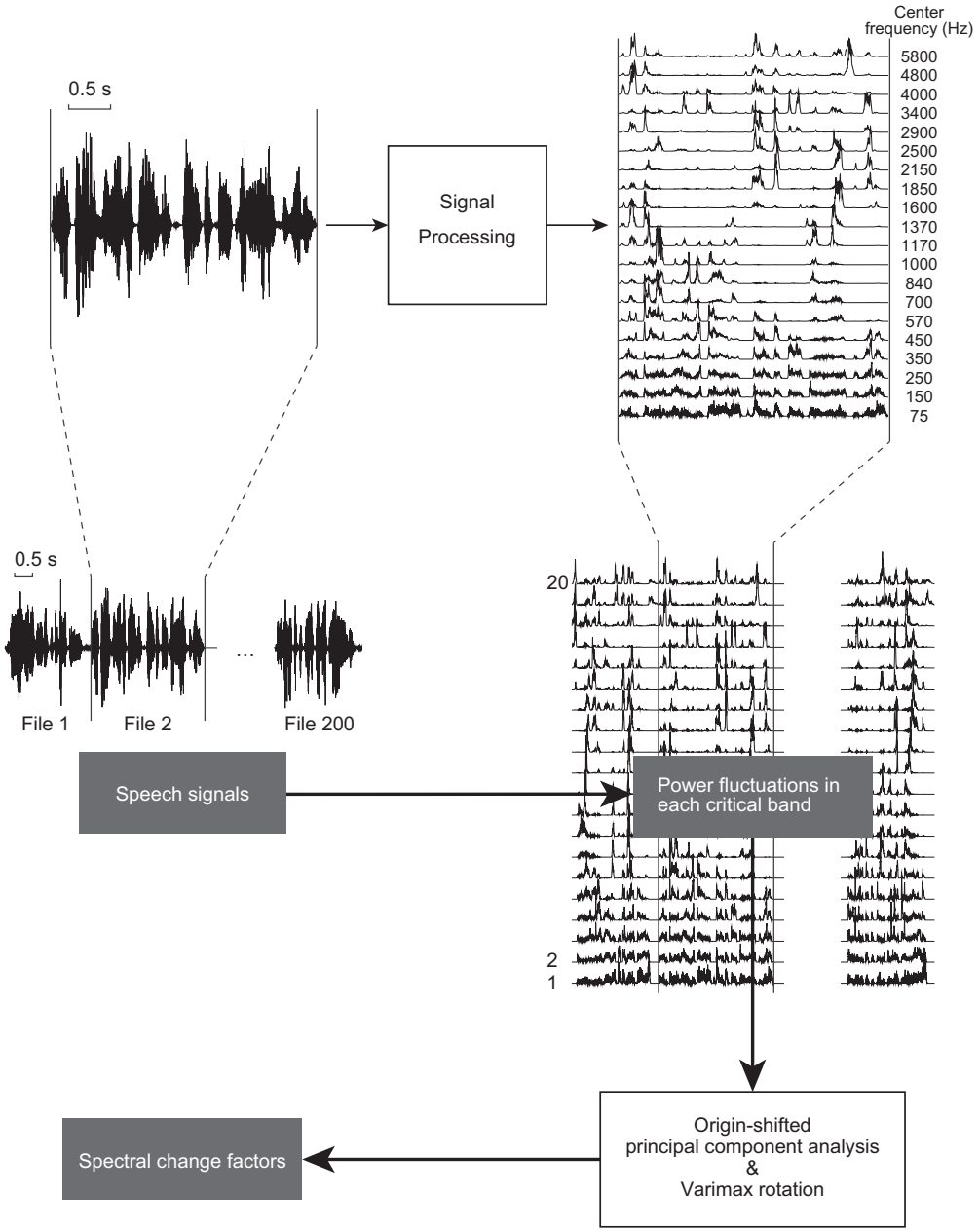

図 2.1 分析手続きの流れ図。ディジタル録音された原音声は信号処理を経て 20 の臨界帯域ごとのパワー変動となる。20 の臨界帯域ごとのパワー変動を起点移動主成分分析とバリマックス回転にかけて、パワースペクトル因子が得られる。

まずは、因子分析にかけるデータとなる臨界帯域ごとのパワー変動を得るための分析手続きについて説明する。ここで行った信号処理の流れは図 2.2 のようにまとめられる。

分析対象の音声を 20 の臨界帯域に分割し、各帯域のパワー変動を 1 ms 間隔で得るための処理を行った。分析対象音声のある時点でのパワースペクトルを得るために、その時点を中心とする 30 ms の区間の時間波形を窓関数で部分的に切り取った。窓関数にはハミング窓を用いた。次に高速フーリエ変換を行い、ハミング窓で切り取られた 30 ms 分の短時間信号から振幅スペクトルを得た。さらにその振幅スペクトルを 2 乗してパワースペクトルを算出した。

ここまでの処理で得られたパワースペクトルは、周波数軸上でパワーが微細に変動している。このパワースペクトルの微細な成分は主に声帯振動に起因し、音声の基本周波数に関連する成分である。一方パワースペクトルの包絡の特徴は発声時の声道形状によって決まる。音声のパワースペクトルは声帯振動 (音源) が作る周波数特性と声道形状 (フィルタ) によって決まる共鳴の周波数特性との積でモデル化される。この考え方は音源フィルタ理論 (Fant, 1960) と呼ばれている。声道形状に起因するパワースペクトル包絡の特徴を分析することがここでの目的である。Ueda and Nakajima (2017) の分析結果では、4 因子を取り出す分析において、中心周波数の低い臨界帯域において因子負荷量が臨界帯域一つ飛ばしで大きな値となるような因子が取り出されている。パワースペクトルの微細構造に見られる一つひとつの山の間隔は声帯振動の周期、すなわち基本周波数とおおむね一致する。基本周波数が臨界帯域幅を超えている場合は、ある臨界帯域でパワースペクトルの微細構造の山ができているときにその隣の臨界帯域では谷ができるようなことがある。これが、因子負荷量が臨界帯域一つ飛ばしで大きな値となった原因であろう。そこで本研究ではパワースペクトルの包絡形状を推定し、これを分析することとした。 パワースペクトル包絡を推定するのに音響工学の分野でよく用いられている、ケプストラム分析 (e.g., Rabiner & Schafer, 1978) を採用した。パワースペクトルの対数を取り、それをフーリエ変換することで得られるケプストラムにはその高次の成分 (高ケフレンシ成分) に音源の周波数特性が、低次の成分 (低ケフレンシ成分) に声道の周波数特性が現れていることが音源フィルタ理論を基に示されている。ある値以上の高ケフレンシ成分をケプストラムから取り除いたうえで、パワースペクトルに戻すことでパワースペクトルの振幅包絡を推定することができる。

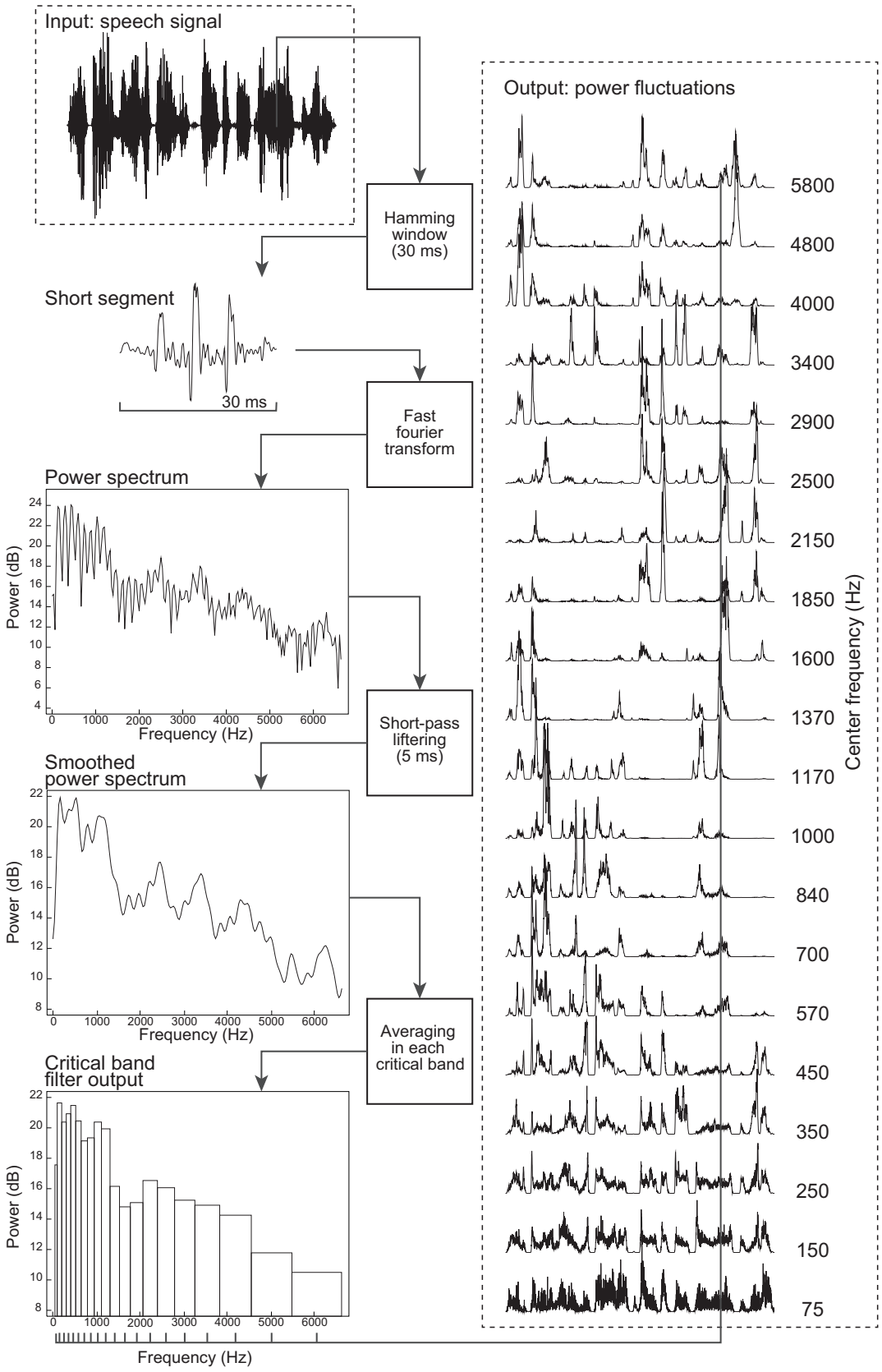

図 2.2 音声信号から 20 の臨界帯域ごとのパワー変動を得るための信号処理手順。

そこで本研究では、ケフレンシ軸上で 5 ms 以上の高ケフレンシ成分を除去 (リフタリング) した。これによって、微細な構造が平滑化されたパワースペクトルを得た。 リフタリングによって平滑化されたパワースペクトルを 20 の臨界帯域に周波数分解し、各臨界帯域で帯域内の平均のパワー値を求めた。以上の操作で音声信号のある時点における 20 の臨界帯域のパワー値が得られる。この操作をハミング窓の位置を 1 ms ずつずらして (フレーム周期 1 ms で) 行うことで、20 の臨界帯域のパワー変動を 1 ms 間隔で得た。20 の臨界帯域の中心周波数および遮断周波数は、Ueda and Nakajima (2017); Nakajima et al. (2017) と同じ条件で行うために、 Zwicker and Terhardt (1980) を参考とした。ただし、50 Hz 以下の帯域は音声とは関連が小さく、データベース収録の際に除去されているため、第 1 番目の臨界帯域が 0–100 Hz となっているところを、 50–100 Hz に変更した。よって 50–6400 Hz の範囲の周波数に 20 の臨界帯域が配置された (表 2.1)。分析対象の原音声がもともと 8000 Hz まで収録されていたのに対して、6400 Hz の周波数成分までしか分析しないこととなる。原音声を 6400 Hz 以下に帯域制限しても音声の言語内容を正しく聴きとることが可能であることは事前に確認している。また、サンプリング周波数 16000 Hz で録音可能な周波数範囲の上限付近の周波数は、たとえ録音できていたとしても本来とは異なる特徴に歪められているおそれもある。以上のような理由により、6400 Hz までを分析対象の周波数範囲と決めた。

また、分析に臨界帯域を用いることの是非について述べておく。 本研究の目的は、音声のパワースペクトル包絡にみられる特徴を因子として抽出することである。分析対象とする音声の基本周波数は 150 Hz 程度である。低域において 100 Hz 以下の帯域幅となる聴覚フィルタを必ずしも用いる必要はない。臨界帯域幅で分割された 20 チャンネルの雑音駆動音声がほぼ完全に明瞭である。雑音駆動音声を合成する場合においても一つのチャンネル内でスペクトルが平坦である方が簡便である。以上のようなことから臨界帯域を議論の出発点にすることが妥当であると考えられるであろう。

表 2.1 臨界帯域フィルタの中心周波数と通過帯域

Band no.	Center frequency (Hz)	Passband (Hz)
1	75	50–100
2	150	100–200
3	250	200–300
4	350	300–400
5	450	400–510
6	570	510–630
7	700	630–770
8	840	770–920
9	1000	920–1080
10	1170	1080–1270
11	1370	1270–1480
12	1600	1480–1720
13	1850	1720–2000
14	2150	2000–2320
15	2500	2320–2700
16	2900	2700–3150
17	3400	3150–3700
18	4000	3700–4400
19	4800	4400–5300
20	5800	5300–6400

図 2.2 の手続きで得られた 20 のパワー変動は 20 変量からなる多変量データとしてみることができる。この多変量データを本研究で新しく提案する「起点移動主成分分析」にかけ、主成分を取り出し、主成分の因子負荷量をバリマックス回転 (Kaiser, 1958) することでパワースペクトル因子を抽出した。

本来の主成分分析は、多次元空間中に表現された多変量データに対して、そのデータの分散が最大となるような多次元空間上の方向を主成分として順次求めていく分析手法である (e.g., Jolliffe, 2002)。そのため、主成分を決定づける固有ベクトルはデータの重心を起点として求められる。このようにして求められた主成分が表現しうる情報だけで元の多変量データを再構成するということは、元の多変量データを主成分空間に正射影したものに置き換えるということである。ここでもし、データの零点、本研究の場合は無音を表す点が主成分分析による部分空間に含まれなかった場合、無音を表す点はデータを再構成する際に歪められ、臨界帯域のどこかにパワーを持った点に移る。こうして再構成されたデータを基に再合成した音声は定常的な

雑音を含むこととなる。これを聴取者が聴けば、再合成された音声の中で雑音が鳴り続いているように感じるであろう。本来意図しない定常的な雑音成分が生じる音声を聴取実験に用いるのは適切ではない。おそらく、同様の定常雑音が Zahorian and Rothenberg (1981) の再合成音声においても生じていたと考えられるが、このことについて特に言及はされていなかった。

これに対して起点移動主成分分析は、主成分分析によって求められる部分空間を定義するベクトルがデータの重心ではなく、全ての変量の値が零となる点、本論文では無音を表す点を起点とするように変形した手法である[1]。こうすることによって、無音を表す点は、データの再構成をしても必ず無音のままとなり、上述の問題に起因する、意図しない定常的な雑音成分が発生するということはなくなる。図 2.3 は、2 変量からなる非負のデータに対して、通常の主成分分析と起点移動主成分分析をそれぞれ行った場合とで、算出される主成分がどのように異なるのかを示す概念図である。

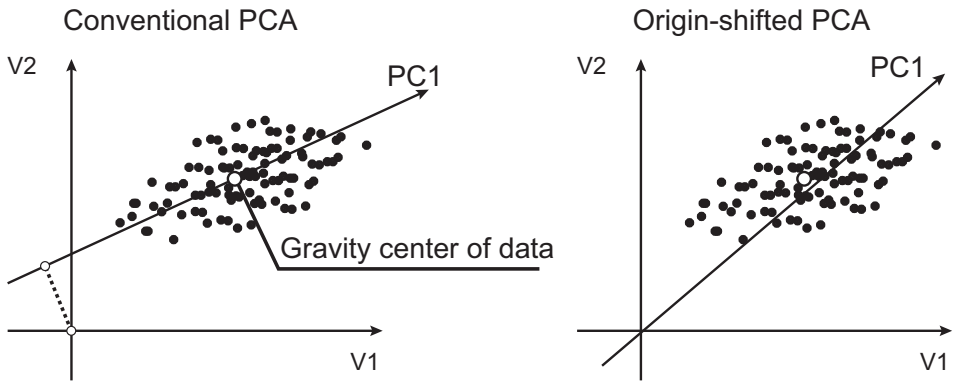

図 2.3 通常の主成分分析 (左) と起点移動主成分分析 (右) における主成分算出の概念図。通常の主成分分析では、多次元空間の原点が含まれるように主成分が算出されておらず、原点を主成分軸に正射影するとずれが生じる。このずれが音声の再合成においては定常雑音の原因となる。

起点移動主成分分析で得られた主成分空間を定義するベクトル、すなわち因子負荷量をバリマックス回転することで、起点移動主成分分析によって得られたパワースペクトル因子をより解釈しやすい、つまり各因子がどの臨界帯域と関連が強いのかが分かりやすい形にすることができる。バリマックス回転を第何主成分まで含めて行うかを変えることで、9 種類のパワースペクトル因子の組を得た。例えば、3 因子からなるパワースペクトル因子を得たい場合は、

[1] 主成分空間を定義するベクトルの起点を全ての変量の値が零となる点にするための実際的な方法はいくつか考えられる。本論文では、主成分分析にかける多変量データに符号を逆転させた多変量データをつなげて、データの重心を変量の値が零となる点に変えるという方法を用いた。

第 3 主成分までをバリマックス回転した。

2.2.3 結果と考察

図 2.4、2.5、2.6 は 3 つの言語から得られた 9 組のパワースペクトル因子について、臨界帯域ごとの因子負荷量を示したものである。まずは、得られたパワースペクトル因子の特徴が 3 つの言語の間で似ているかどうかを見ていく。4 因子を抽出したところまで互いに似た因子の配置が得られているのが分かる。5 因子を超えてパワースペクトル因子を抽出すると、言語間で共通していると考えられる因子を見つけるのが困難になった。以上の結果は Ueda and Nakajima (2017) で報告された内容と一致している。

次に、言語間で共通した特徴のパワースペクトル因子が得られた因子数で、どのような特徴を持った因子が得られたのかを見ていく。1 因子分析では、すべての臨界帯域において因子負荷量が正の値であった。これは起点移動主成分分析によって得られる第 1 主成分 (因子) が必ずもつ特徴である。2 因子分析では、約 1000 Hz を中心とする中帯域に大きい因子負荷量を持つ因子とその両側の帯域で因子負荷量が大きい因子とが得られている。 さらに、3 因子分析では、約 1000 Hz を中心とする中帯域に大きい因子負荷量を持つ因子、約 3000 Hz 以上の高帯域に大きい因子負荷量をもつ因子、そして約 500 Hz 以下の低域および、中帯域と高帯域の間の帯域 (約 1500–3000 Hz) に分かれて大きい因子負荷量を持つ二峰性の因子がそれぞれ得られている。二峰性の因子が現れるという特徴は、Ueda and Nakajima (2017) の分析でも同様に報告されている。本分析が本質的にはUeda and Nakajima (2017) の分析と同じ結果を導いていることを示す指標であると言えるであろう。そして 4 因子分析では、3 因子分析のときに得られた中帯域に大きい因子負荷量を持つ因子と高帯域に大きい因子負荷量を持つ因子に加え、二峰性の因子が二つの因子に分かれたような因子が得られている。この結果もまた、Ueda and Nakajima (2017) の分析結果と一致している。したがって、今回導入した起点移動主成分分析によっても、先行研究と同等のパワースペクトル因子が得られたと判断できる。

ある臨界帯域において、特定の因子の因子負荷量の絶対値が大きく、それに比べてそれ以外の因子の因子負荷量の絶対値が小さい場合は、その帯域のパワーの変化が因子負荷量の絶対値が大きな因子によって説明される割合が大きいことを意味する。また別の臨界帯域においても同じような因子負荷量の関係になっている場合、それらの複数の帯域においてパワーが同じように変化することとなる。例えば 3 因子分析の 1000 Hz 付近のいくつかの臨界帯域においては、一つの因子 (白抜きの丸で表したもの) の因子負荷量が正の値で大きく、それ以外の因子

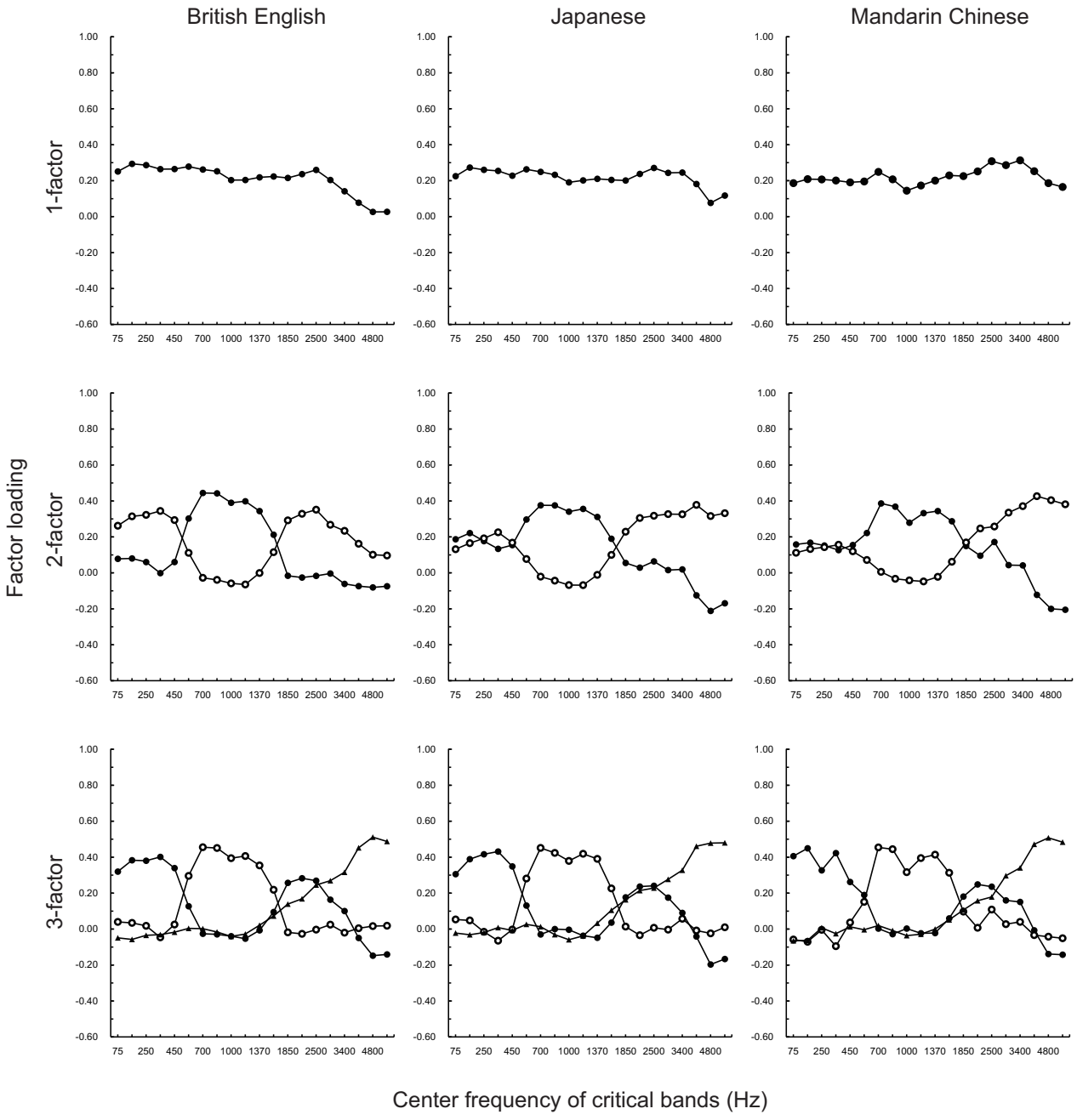

図 2.4 起点移動主成分分析に基づいて得られたパワースペクトル因子の、臨界帯域ごとの因子負荷量。横軸の数値は 50–6400 Hz の範囲の周波数に配置された 20 の臨界帯域の各中心周波数を示している。臨界帯域の中心周波数およびその帯域幅については、表 2.1 を参照のこと。左から、イギリス英語母語話者、日本語母語話者、中国語 (普通話) 母語話者の結果を示す。上段から、1 因子、2 因子、3 因子、をそれぞれ抽出した場合の結果である。

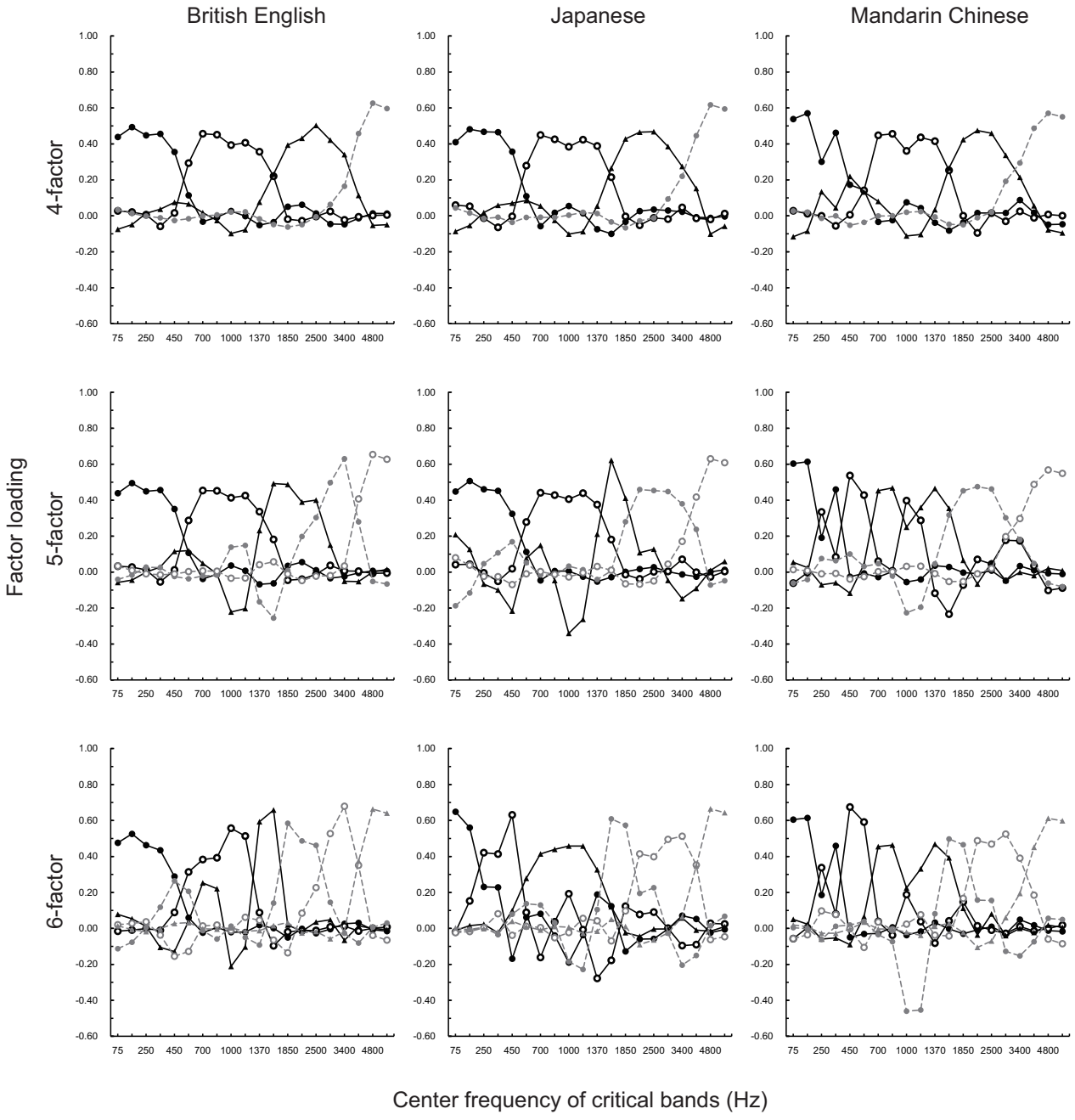

図 2.5 起点移動主成分分析に基づいて得られたパワースペクトル因子の、臨界帯域ごとの因子負荷量。横軸の数値は 50–6400 Hz の範囲の周波数に配置された 20 の臨界帯域の各中心周波数を示している。臨界帯域の中心周波数およびその帯域幅については、表 2.1 を参照のこと。左から、イギリス英語母語話者、日本語母語話者、中国語 (普通話) 母語話者の結果を示す。上段から、4 因子、5 因子、6 因子、をそれぞれ抽出した場合の結果である。

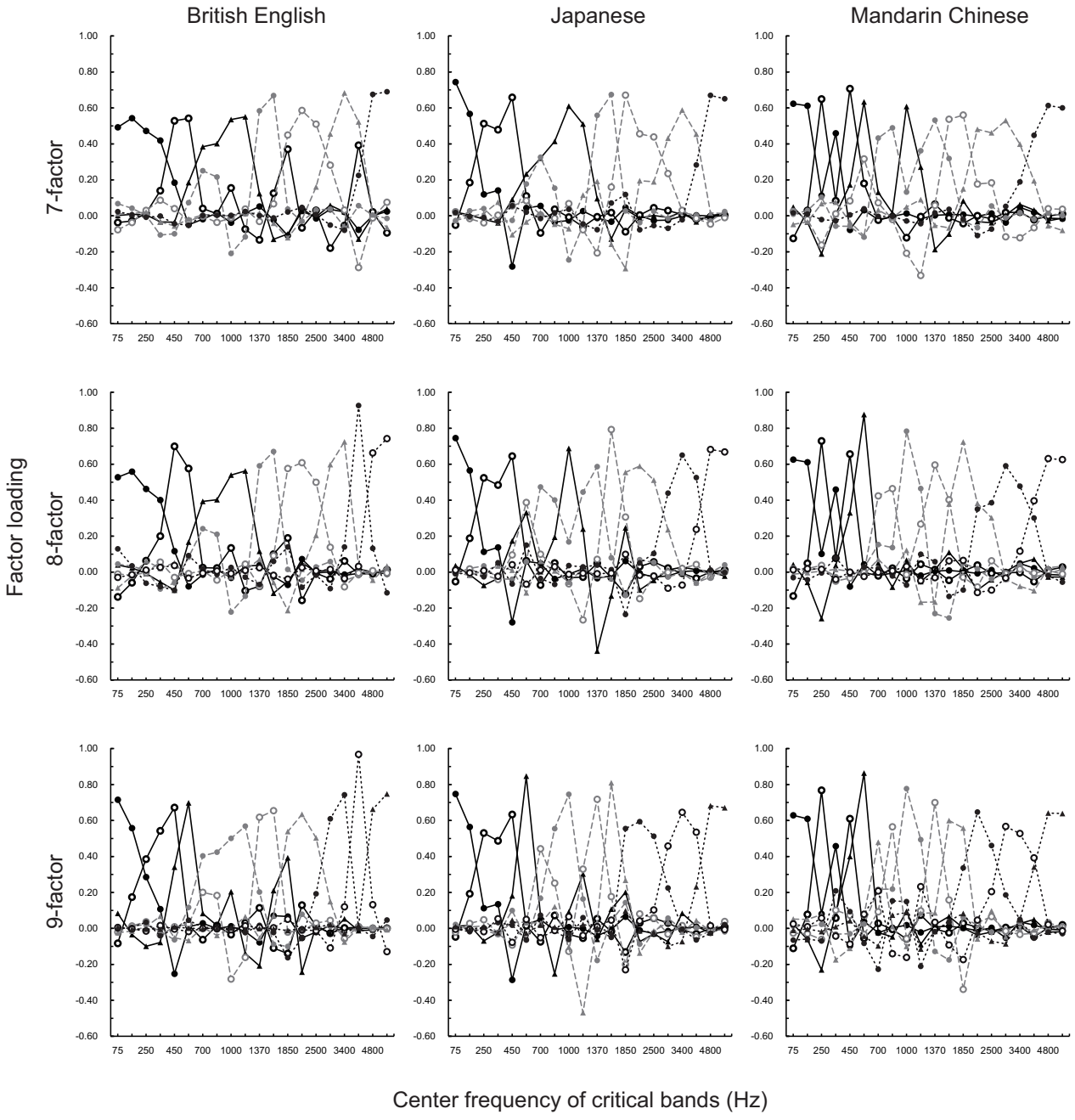

図 2.6 起点移動主成分分析に基づいて得られたパワースペクトル因子の、臨界帯域ごとの因子負荷量。横軸の数値は 50–6400 Hz の範囲の周波数に配置された 20 の臨界帯域の各中心周波数を示している。臨界帯域の中心周波数およびその帯域幅については、表 2.1 を参照のこと。左から、イギリス英語母語話者、日本語母語話者、中国語 (普通話) 母語話者の結果を示す。上段から、7 因子、8 因子、9 因子、をそれぞれ抽出した場合の結果である。

の因子負荷量は零に近いという特徴が共通している (図 2.4)。よってこれらの帯域のパワーはまとまって変動していると解釈することができる。

さらに 4 因子分析において、約 500 Hz 以下の 5 つの臨界帯域で因子負荷量の大きい因子は、それらの 5 つの帯域において同程度の因子負荷量であった。ケプストラム分析によってパワースペクトルの音源特性が適切に取り除かれたためであろう。

分析する多変量データの構造 (多次元空間上でのデータの分布) によっては、通常の主成分分析と起点移動主成分分析とでは大きく異なる因子を導く可能性がある。3 因子分析と 4 因子分析で得られたパワースペクトル因子が、通常の主成分分析に基づいて得られた対応するパワースペクトル因子と似たよう特徴を持っているということは、臨界帯域ごとのパワースペクトルの変動のデータを多次元空間上に表現した際に、データの重心と無音点 (多次元空間の原点) とを結ぶ直線上付近にデータが分布していたということになる。

表 2.2 に起点移動主成分分析に基づいて得られた因子、および通常の主成分分析に基づいて得られた因子の累積寄与率を示す。寄与率とは、ある主成分または因子が元の多変量データの分散をどれだけの割合で保持しているのかを示すものである。それぞれの主成分または因子が、元の多変量データの情報をどれだけ説明しているかを表すものとも考えることができる。累積寄与率はその寄与率をある主成分数、因子数までで累積したものである。起点移動主成分分析に基づいて得られた第 1 因子の寄与率は 17–22%程度であり、因子数が増加するにしたがって累積寄与率は緩やかに上昇し、第 9 因子までで 74–77%程度まで上昇した。各言語でパワースペクトル因子の特徴が共通している第 4 因子までの累積寄与率は 49–56%程度であった。音声のパワースペクトル変動の特徴のおよそ半分が第 4 因子までで説明でき、さらには 3 つの言語間でその特徴が共通しているということを示している。

言語ごとに、通常の主成分分析と起点移動主成分分析の累積寄与率を同じ因子数の間で比較すると、その差は 2 %以下であった。累積寄与率という観点でも、起点移動主成分分析に基づいて得られたパワースペクトル因子が通常の主成分分析に基づいて得られたパワースペクトル因子と同等であると言えるだろう。

連続的に発話した音声のパワースペクトル変動のデータは、起点移動主成分分析を用いるのに適していると言える。一方で、母音の定常部のパワースペクトルを母音ごとに一つ一つ観測し、それらのパワースペクトルの周波数帯域ごとのレベルを変量に用いる場合は、起点移動主成分分析は適さないと考えられる。ちょうどPlomp et al. (1967) が分析対象とした音声がそれにあたる。母音の定常部はパワーの変化が安定しており、母音を一音ずつ発話した場合には、

話者が敢えてそうしようとしない限り、母音ごとにパワーが大きく異なるということはないであろう。よって観測データの多次元空間上の分布はデータの重心と空間の原点とを結ぶ直線上付近には分布しない。このようなデータに対して通常の主成分分析を行えば、第 1 主成分の因子負荷量のいくつかが負の値[2]となることが予測されるが、同じデータに起点移動主成分分析を行えば、第 1 主成分の因子負荷量はすべて正の値となり、分析結果が大きく異なるおそれがある。

通常の主成分分析によって得られたパワースペクトル因子から音声を再合成した場合に生じる定常雑音の例を図 2.7 に見ることができる。原音声が雑音駆動音声として再合成されたものである。実際の再合成の方法については次節で詳しく述べる。この図で示す例の原音声は 1.9–2.0 s 付近がほぼ無音状態である。これを通常の主成分分析で得られた 4 因子から再合成した場合、およそ 1000–1500 Hz の帯域に定常的な雑音が生じているのが分かる。一方、起点移動主成分分析で得られた因子から再合成した場合は、そのような定常的な雑音は生じていない。

[2]因子負荷量の符号自体には意味はない。すべての符号を入れ替えても主成分または因子が表すものは同じである。符号の違いが意味をもつのは、一つの主成分または因子内において、変量間で因子負荷量を比較する場合である。よって、起点移動主成分分析で得られる第 1 主成分の因子負荷量はすべて負の値となると言い換えることもできる。本論文では分析結果を見やすくするために、それぞれの因子について、因子負荷量の絶対値が最も大きいものが正の値となるように符号をそろえて表示した。

表 2.2 起点移動主成分分析と通常の主成分分析で得た因子の累積寄与率。

Japanese		
	Cumulative contribution (%)	
Number of factors	Conventional	Proposal
1	23.3	22.9
2	38.0	37.3
3	49.9	47.9
4	55.9	55.6
5	61.7	61.3
6	66.5	66.3
7	70.7	70.5
8	74.3	74.1
9	77.7	77.6
British English		
	Cumulative contribution (%)	
Number of factors	Conventional	Proposal
1	22.4	21.5
2	35.3	34.5
3	48.7	45.6
4	54.0	53.7
5	59.9	59.7
6	65.2	64.5
7	69.2	68.8
8	73.3	72.3
9	76.5	76.3
Mandarin Chinese		
	Cumulative contribution (%)	
Number of factors	Conventional	Proposal
1	17.9	17.3
2	32.0	31.5
3	43.1	40.7
4	48.8	48.6
5	55.7	55.3
6	61.2	60.3
7	65.8	65.4
8	70.3	69.3
9	73.8	73.7

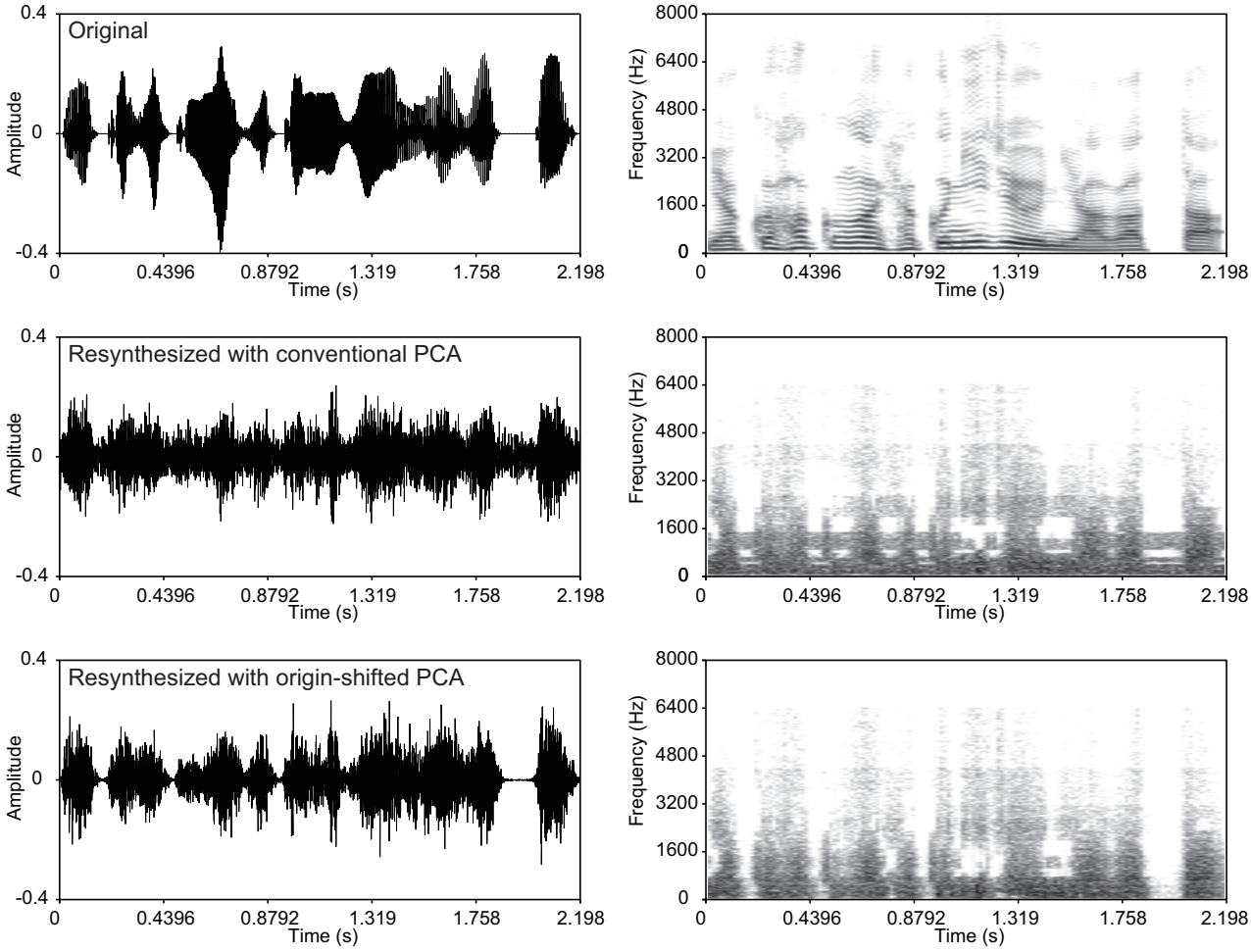

図 2.7 原音声 (上段)、通常の主成分分析による因子から再合成した雑音駆動音声 (中段)、起点移動主成分分析による因子から再合成した雑音駆動音声 (下段) の時間波形 (左) とスペクトログラム (右)。原音声の 1.9〜2.0 s 付近には無音区間があるが、通常の主成分分析による因子から再合成した雑音駆動音声ではその区間に定常的な雑音が発生している。起点移動主成分分析による因子から再合成した雑音駆動音声にはそのような雑音は生じていない。

2.3 実験 1 ： パワースペクトル因子の個数と文音声の明瞭度の関係

起点移動主成分分析によって得られたパワースペクトル因子が先行研究 (Yamashita et al., 2013; Ueda & Nakajima, 2017; Nakajima et al., 2017) で得られたパワースペクトル因子と同等のものであることが確認できたことによって、パワースペクトル因子から音声を再合成し、聴取実験に用いる準備ができた。パワースペクトル因子のもつ音声知覚上の役割を調べる最初の段階として、実験 1 では十分に明瞭な音声の知覚に必要な因子数を確かめる。因子分析で得られた多言語に共通する因子に、音声知覚の手がかりがどれだけ含まれているのかが分かる。明瞭な音声の知覚に必要な因子数が 4 つ以下であれば、多言語で共通する特徴をもつパワースペクトル因子が音声知覚においても共通の役割を持っていると考えることができるだろう。

2.3.1 実験参加者

19〜24 歳 (平均 = 21.5 歳、$SD = 1.6$ 歳) の 6 名の男性および 6 名の女性が実験に参加した。実験参加者はすべて日本語母語話者であった。いずれの参加者も両耳ともに、純音聴力レベルが 25 ｄ B HL 以下であることを 125–8000 Hz の帯域で確認した。

本実験は九州大学大学院芸術工学研究院の倫理審査委員会の承認の下、参加同意書に実験参加者の署名を得た上で実施した。

2.3.2 実験装置

実験は背景雑音のレベルが 25 ｄ B A 以下である防音ブース内で行われた。刺激音はオーディオボード (E-MU 0404) を搭載したコンピュータ (Frontier, KZFM71/N) にディジタル信号 (16-bit 量子化、16000 Hz サンプリング) として保存されており、D/A コンバータ (ONKYO, SE-U55GX)、低域通過フィルタ (NF 回路設計, DV-04 DV8FL 遮断周波数 7000 Hz)、グラフィックイコライザ (Roland, RDQ-2031)、ヘッドフォンアンプ (STAX, SRM-323S)、ヘッドフォン (STAX, SR-307) の順に通って実験参加者の両耳に呈示された (図 2.8)。ローパスフィルタはエイリアシング防止のために、グラフィックイコライザは再生系の周波数応答を平坦にするためにそれぞれ用いられた。ヘッドフォンアンプの出力レベルは、刺激音と同程度の音圧レベルで作成した帯域雑音を再生した時に、78 ｄ B A となるように調整した。よって各刺激音も同

程度のレベルで実験参加者に呈示された。刺激音の呈示レベルの測定には、人工耳 (Brüel & Kjær, Type 4153)、ハンドヘルドアナライザ (Aco, Type 6240) を用いた。

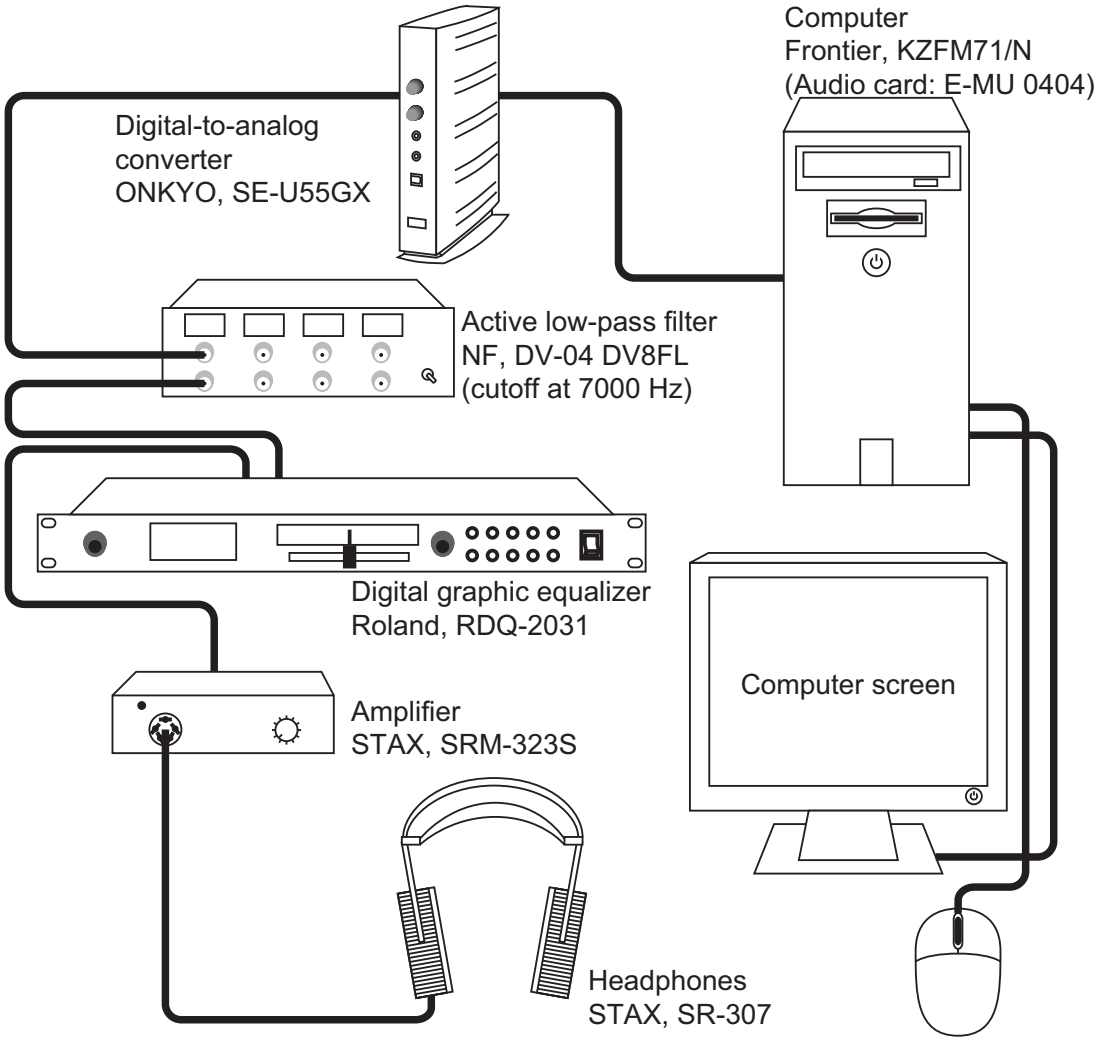

図 2.8 実験 1 および 2 に用いた装置一式。

2.3.3 刺激音

刺激音の合成に用いた原音声は、NTT-AT 社の「多言語音声データベース 2002」に収録されている、日本語母語話者による日本語音声のうち、NJ01M と登録されている男性 1 名の音声から 57 文を選出した。そのうち 12 文は練習試行およびウォーミングアップ試行用に、残りの 45 文は結果の整理に用いるための本試行用とした。刺激音の原音声として日本語を用いたのは、日本語はモーラに分解することができ、ひらがなで記述した際に、1 文字が 1 モーラ

に対応するという特徴をもつために、実験参加者の回答を採点するのに適しているためである。本試行用の 45 文は 9 つのリストに 5 文ずつ振り分けられ、1 つのリスト内に含まれる文のモーラ数は 17〜19 モーラとし、リスト内の平均モーラ数が 18 モーラとなるようにした (表 2.3)。

表 2.3: 刺激音の原音声として用いた文音声。NTT-AT 社の「多言語音声データベース 2002」よりモーラ数が 17〜19 モーラのもの 45 文を用いた。45 文は 5 文ずつのリストに分けられ、実験参加者ごとに異なる処理条件を割り当てて刺激音を合成した。

No.	List	Database No.	Sentence	Number of mora	Average
1		44	こきょうを離れてみるのもいいでしょう	17	
2		60	さりげなく光っているのは何ですか	18	
3	A	91	やっとのことで座ることができました	18	18
4		133	新幹線で逆に京都へ行きます	18	
5		102	用件を録音できる装置が欲しい	19	
6		69	空に入道雲が広がっている	17	
7		123	テレビに時刻が映し出されています	18	
8	B	138	野球の審判員に選ばれました	18	18
9		114	テーブルに白い布がかかっています	18	
10		131	敬語の使い方はむずかしいものです	19	
11		17	昨日は本棚の整理をしました	17	
12		193	スピードの出しすぎには注意しましょう	18	
13	C	125	写真では微妙な違いが分からない	18	18
14		105	パルプは木材から作られています	18	
15		22	ネギは中国から到来したものです	19	
16		29	ボールペンでサインをお願いします	17	
17		71	動物に餌を与えないで下さい	18	
18	D	200	車の価格は三百万円です	18	18
19		6	その地方の名産品は何ですか	18	
20		57	そこには小さい展示ホールがあります	19	
21		75	曇りの日は気分も沈みがちです	17	
22		100	駅へ行く近道を教えて下さい	18	
23	E	77	浜辺の人影も少なくなりました	18	18
24		27	今年も冷害の心配があります	18	
25		11	彼女は手の込んだごちそうを作ります	19	
26		70	使用後はもとの場所に戻しましょう	17	

表は次ページに続く

前ページからの続き

No.	List	Database No.	Sentence	Number of mora	Average
27		120	のどかな田園は素朴な世界です	18	
28	F	34	日記を毎日つけることにしました	18	18
29		66	かくし味に砂糖を少し加えます	18	
30		108	いつもの喫茶店でまっていてください	19	
31		92	ぽっかりと満月が浮かんでいます	17	
32		58	急いでカメラのシャッターを切りました	18	
33	G	159	ついつい長電話になってしまいます	18	18
34		85	絵本をながめると心がなごみます	18	
35		7	制服のデザインを考えてください	19	
36		89	塩分の取りすぎに注意しましょう	17	
37		181	少年は家を飛び出して行きました	18	
38	H	50	路地から急に子供が飛び出しました	18	18
39		192	合唱コンクールで優勝しました	18	
40		74	トンネル工事が着実に進んでいる	19	
41		65	丘の上に奇妙な小屋が見えます	17	
42		39	この流行は当分続きそうです	18	
43	I	124	大小さまざまなテーブルがあります	18	18
44		173	私の趣味は随筆を書くことです	18	
45		82	麦畑を横切ってここまで来ました	19	

これで終わり

図 2.9 に刺激音の合成手続きの概要を示す。図中の流れに沿って 1 つの文音声につき、 9 種類の雑音駆動音声を再合成した。

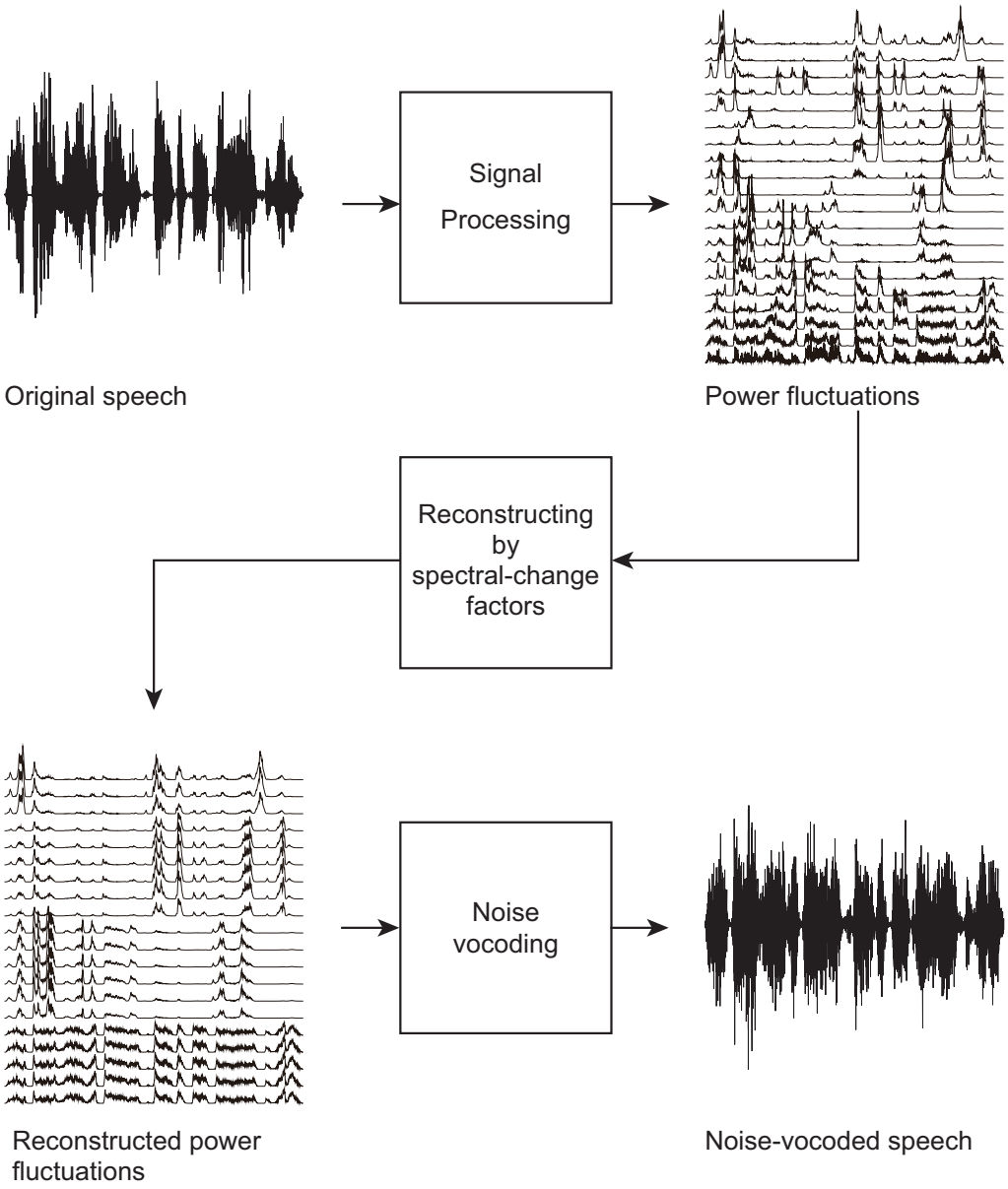

図 2.9 刺激音合成の流れ図。処理は大きく分けて 3 工程からなる。まず、音声信号から 20 臨界帯域ごとのパワー変動が算出される。次に、パワー変動がパワースペクトル因子によって縮約され、再構成される。最後に再構成されたパワー変動を用いて、雑音駆動音声が合成される。

まず、分析 1 で行った方法を用いて、原音声から臨界帯域ごとのパワー変動を得た。得られた 20 系列のパワー変動はパワースペクトル因子によって情報が圧縮され、因子によって説明される情報だけをもったパワー変動に再構成された (図 2.10)。 まず、20 系列のパワー変動は

パワースペクトル因子によって刺激音の条件に応じて 1 〜 9 系列の因子得点に次式によって変換された。

$$X_{k,t} = \sum_{n=1}^{20} W_{k,n} Y_{n,t}, \tag{2.1}$$

ここで、$X_{k,t}$ は第 k 因子の t フレーム目の因子得点、$W_{k,n}$ は第 k 因子の因子負荷量の内の、第 n 臨界帯域での値、$Y_{n,t}$ は第 n 臨界帯域の t フレーム目のパワーである。

次に、20 臨界帯域のパワー変動が次式から再構成された。

$$\hat{Y}_{n,t} = \sum_{k=1}^{K} W_{k,n} X_{k,t}, \tag{2.2}$$

ここで、$\hat{Y}_{n,t}$ はパワースペクトル因子によって再構成された第 n 臨界帯域の t フレーム目のパワーである。K は因子数 $(K = 1, 2, \cdots, 9)$ である。

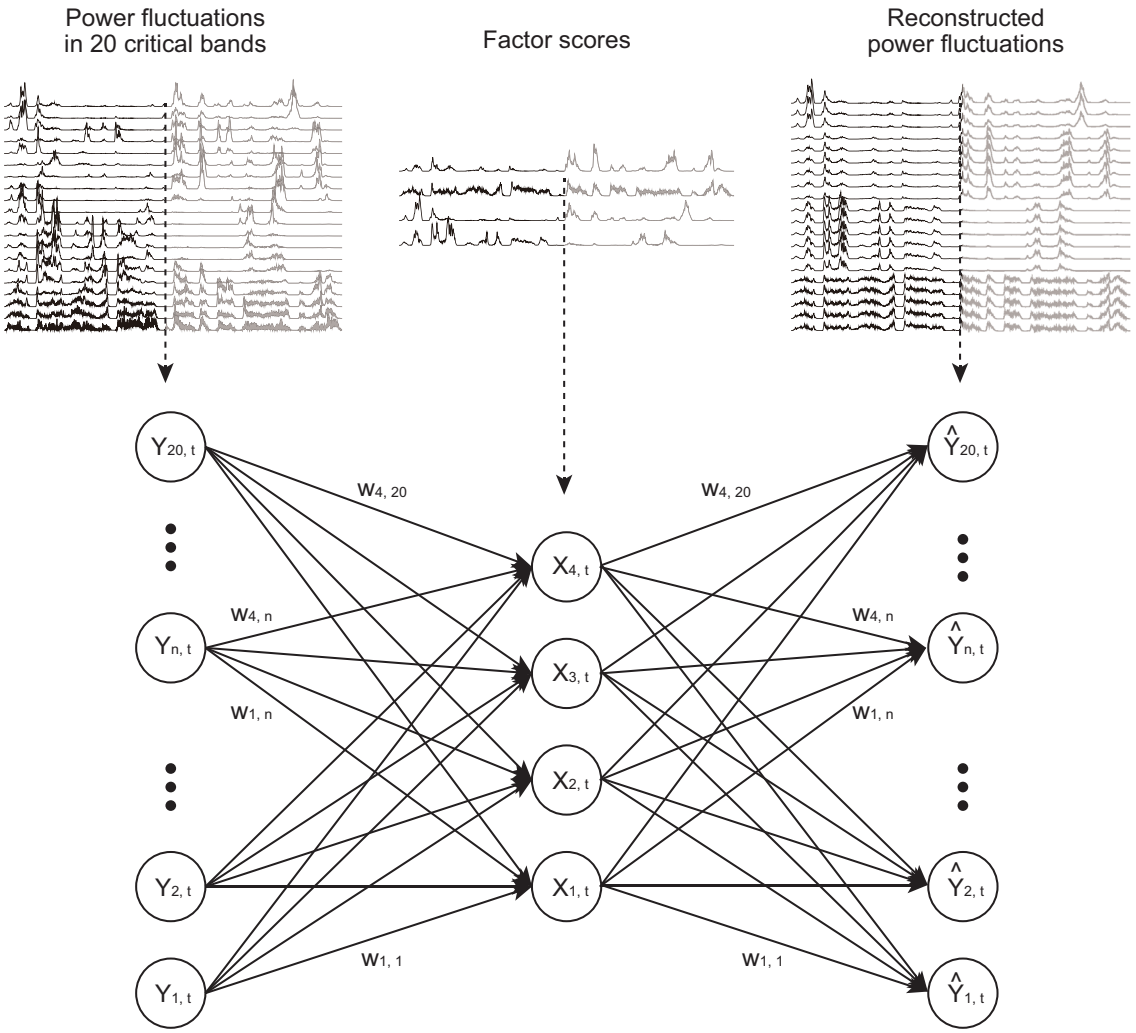

図 2.10 臨界帯域ごとのパワー変動を因子得点に一度変換し、因子得点から臨界帯域ごとのパワー変動を再構成する過程。この図では 4 因子から再構成する例を示している。

この処理によって再構成されたパワー変動には、負の値が含まれていることがあった。本来音声のパワー (振幅の 2 乗) は正の値でしか実現できない。よって負の値として再構成されたデータについては零に修正することとした。この便宜上の処理の影響については第 3 章で取り上げる。

次にこの再構成されたパワー変動を実現する雑音駆動音声の合成について説明する (図 2.11 参照)。白色雑音を生成し、表 2.1 で示す遮断周波数となるディジタルフィルタによって 20 臨界帯域に雑音を分割した。さらに、各帯域の出力を 2 乗したのちに $\sigma = 20$ ms のガウス窓で移動平均をして、雑音のパワー変動の包絡を得た。このガウス窓の移動平均の操作は各臨界帯

域のパワー変動に対する遮断周波数 7 Hz、斜度 9.5 dB/oct. の低域通過フィルタに相当する[3]。

この雑音のパワーと、式 2.2 によって再構成された音声のパワーとの比を各帯域と時間サンプルごとに計算し、帯域雑音のパワー変動が再構成された音声のパワー変動となるように変調した。変調を行うとそれぞれの帯域には含まれない周波数成分が生じるため、各帯域を再びディジタルフィルタに通して、各帯域からはみ出した周波数成分を除去した。最終的に 20 個の変調された帯域雑音が足し合わされて、刺激音となる雑音駆動音声が完成した。

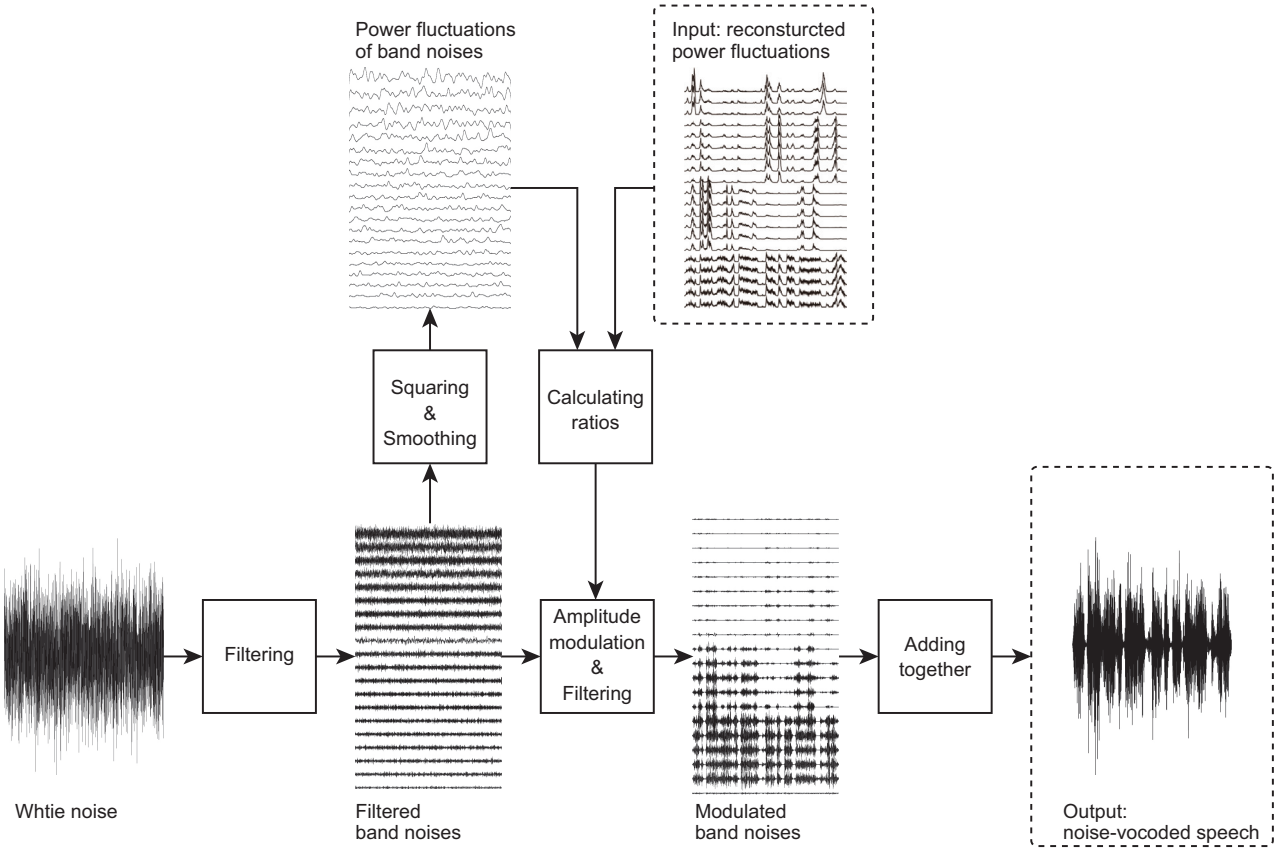

図 2.11 雑音駆動音声の合成手順。パワースペクトル因子によって再構成されたパワー変動と、白色雑音の対応する帯域におけるパワー変動との比が求められ、その比の平方根で白色雑音の各帯域を振幅変調することで雑音駆動音声が合成される。

実験参加者は 9 つの文リストの音声全てを聴取するが、それぞれの文リストがいくつの因子数で再合成された音声であるかが、全ての実験参加者で異なるようにした。例えば、実験参加者 1 が文リスト A に対して因子数 1 の条件で合成した音声を聴き、実験参加者 2 は同じ文リ

[3]ガウス窓で移動平均をかけると、信号にスペクトルの広がりが生じる。ガウス窓の σ の値は小さいほどその広がりが大きくなる。スペクトルの広がりが臨界帯域幅を超えないようにするために $\sigma = 20$ ms を採用した。

スト A に対して因子数 9 の条件で合成した音声を聴いた (表 2.4)。このようにして、文の違いが与える結果への影響が最小限となるようにした。

表 2.4 実験参加者毎の因子数条件と文リストの対応。

	Sentence list								
	A	B	C	D	E	F	G	H	I
Participant I	1	2	3	4	5	6	7	8	9
Participant II	9	1	2	3	4	5	6	7	8
Participant III	8	9	1	2	3	4	5	6	7
Participant IV	7	8	9	1	2	3	4	5	6
Participant V	6	7	8	9	1	2	3	4	5
Participant VI	5	6	7	8	9	1	2	3	4
Participant VII	4	5	6	7	8	9	1	2	3
Participant VIII	3	4	5	6	7	8	9	1	2
Participant IX	2	3	4	5	6	7	8	9	1
Participant X	9	8	7	6	5	4	3	2	1
Participant XI	8	7	6	5	4	3	2	1	9
Participant XII	7	6	5	4	3	2	1	9	8

2.3.4 手続き

実験参加者は、練習試行ブロックで、9 個の刺激に回答した後、3 ブロックに分かれた本試行ブロックで、48 個の刺激に回答した。練習試行で実験参加者が聴いた 9 個の刺激は、それぞれ 9 個の条件に対応しており、すべての実験参加者が同じ 9 種類の刺激音を聴いた。刺激音はすべて無作為な順に呈示された。本試行ブロックの初めの各 1 試行はウォーミングアップ試行として、結果の分析には用いなかった。

実験参加者が、ディスプレイ上に表示される再生ボタンをクリックすると、刺激音は 2 秒後に実験参加者の両耳に呈示された。1 回の再生で、刺激は 3 度繰り返して呈示された。刺激間時間間隔は 1.5 秒であった。音声聴取後、聴き取れた音声をひらがなでワープロソフトを使って入力させるように実験参加者に求め、聴き取れなかった箇所を推測して回答することを避けるように教示した。なお、ひらがなで入力をさせる場合、「は」と「へ」についてはそれぞれ 2 通りの音韻があるため、実験者が区別できるように表記させた。練習試行を含むすべての試行において、実験参加者に正答の文内容を教えるなどの、回答に対するフィードバックは与えなかった。1 ブロック (16 試行) あたりの所要時間は約 7 分であった。

2.3.5 結果と考察

実験参加者の回答と正答を比較し、モーラ正答率を算出した。図 2.12 に再合成に用いたパワースペクトル因子の個数ごとのモーラ正答率の実験参加者平均を示す。1 因子のみから再合成された音声のモーラ正答率は 0.83%と、実験参加者はほとんど文の内容を聴き取ることができていなかった。因子数が 1 つ増えて、2 因子の再合成音声のモーラ正答率は 6.9%であった。これは 1 文中におよそ 1 モーラを正しく知覚できるようになったことを意味し、ほとんど正答率の上昇は見られなかった。しかし、3 因子に増えたときモーラ正答率は 69.2%まで急激に上昇した。さらに 4 因子条件で 83.7%に到達し、それ以降の上昇は頭打ちとなった。

モーラ正答率に逆正弦変換を施した上で、対応のある一元配置分散分析を行ったところ、因子数の変化によってモーラ正答率が統計的に有意に異なることが確かめられた、$F(8, 88) = 315.4$、$p < 0.001$。さらに、Tukey 法による多重比較検定ですべての条件対における正答率の差の検定を行った。因子数が 1 つ増えたときに、正答率が統計的に有意に上昇することが確かめられたのは、4 因子条件までであり (1 – 2 因子条件間: $p < 0.05$, 2 – 3 , 3 – 4 因子条件間 : $p < 0.001$)、4 因子条件と 5 因子条件の正答率には、統計的に有意な差はなかった ($p = 0.96, n.s.$)。4 因子以上の条件間で正答率に統計的に有意な差があったのは、4 因子条件に対して 6 〜 9 因子条件 (4 – 6 因子条件間: $p < 0.05$, 4 – 7 , 8 , 9 因子条件間: $p < 0.01$)、5 因子条件に対して 7 〜 9 因子条件 (5 – 7 , 9 因子条件間 : $p < 0.01$, 5 – 8 因子条件間: $p < 0.05$) であった。6 因子以上の条件間ではいずれも正答率に統計的に有意な差はなかった (6 – 9 因子条件間: $p = 0.066, n.s.$)。以上の結果から、パワースペクトル因子から日本語の雑音駆動音声を再合成する場合、2 因子まででは音声の内容はほとんど聴き取れないが、3 因子になると急に内容が聴き取りやすくなり、4 因子で十分に明瞭に、6 因子でほぼ完全に聴き取れるようになるということが分かった。

パワースペクトル因子による累積寄与率の変化は、その上昇量が次第に減っていく緩やかな曲線を描いて上昇するのに対して、モーラ正答率は 2 因子条件と 3 因子条件の間で急峻に上昇するという結果になった。音声の明瞭度の上昇を決定づけたのは累積寄与率ではなく、因子の個数であったと考えられる。3 因子条件における累積寄与率は 47.9%である中、3 因子条件のモーラ正答率は 69.2%であった。このことは、この 3 因子がもつ全体の半分に満たない物理的情報の方が、残りの情報よりも音声の知覚にとってより有用であることを示している。相補的な音響的特徴においてともに高い明瞭度が得られることが知られている (French & Steinberg, 1947; Hirsh et al., 1954; Miller & Nicely, 1955; Studebaker et al., 1987) ため、3 因子によっ

て説明されない残りのパワースペクトルの変動の情報だけでも音声の知覚ができる可能性はある。実際にそのように合成した音声を筆者を含む数名で試聴してみたが、積極的に音声として聴こうとしなければ、明瞭に内容を聴きとることは困難であった。

パワースペクトル因子から臨界帯域の各帯域のパワー変動を再構成したとき、負の値として再構成されたデータがあった。本研究ではパワーは振幅の 2 乗値として定義しているので、負のパワー値をそのまま使って音声を再合成することはできない。そのため、負の値を 0 に修正せざる得なかった。しかしこのような修正を行ってしまえば、実現される再合成音声はパワースペクトル因子が与える情報をそのまま反映しているとは言えない。次章ではこの問題を取り上げる。

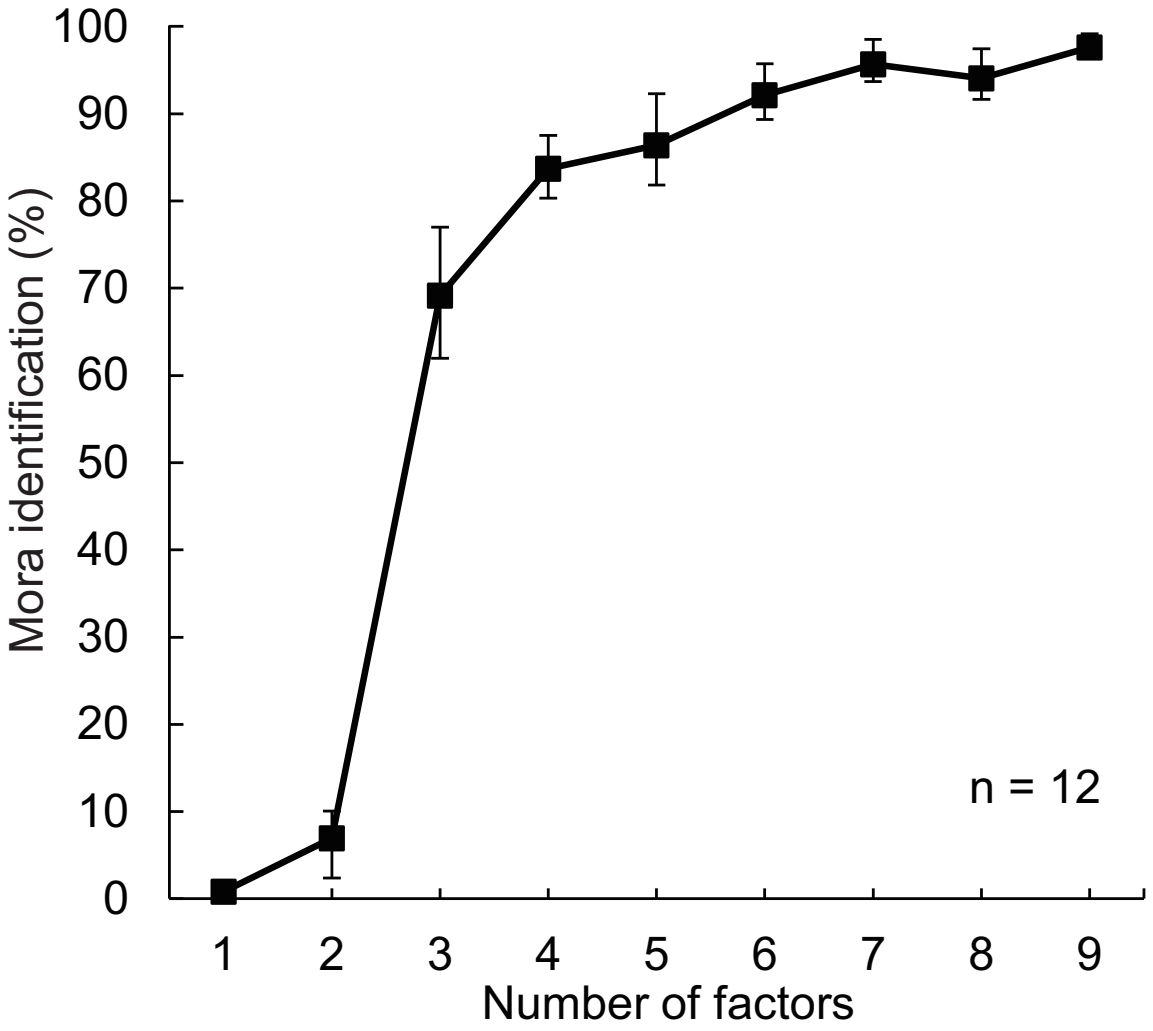

図 2.12 パワースペクトル因子から再合成した音声の平均モーラ正答率。エラーバーは 95%信頼区間を示す。

第２章のまとめ

本章では因子から元のデータを再合成するのに適した、起点移動主成分分析を新しく提案し、この分析法を用いて取り出した日本語音声のパワースペクトル因子がいくつ用いられれば、明瞭に知覚できる音声を再合成できるのかを調べた。分析および実験結果は次のようにまとめられる：

(i) 起点移動主成分分析の分析結果は従来の主成分分析のものと本質的に違いはなく、Ueda and Nakajima (2017) の分析で得られたパワースペクトル因子と同等の因子を取り出すことができる。

(ii) 元の音声信号における無音を無音として再合成できるという点で、起点移動主成分分析から得られたパワースペクトル因子を用いた音声の再合成法は優れている。

(iii) パワースペクトル因子から再合成した音声の明瞭度は、因子の累積寄与率によって決まるというよりも、因子の個数が決定的な要因のようであり、３つ以上用いると急激に明瞭度が高くなり、４つで十分明瞭に、６つ以上でほぼ完全に明瞭になる。

(iv) パワースペクトル因子から音声の再合成を行う過程の中で、因子と再合成された音声との直接的な関係を幾分歪めてしまう処理が避けられなかったので、この影響についての検討が必要である。

第3章　パワースペクトル因子の非負直交基底化

3.1　第３章の目的

第２章では、パワースペクトル因子から臨界帯域ごとのパワー変動を再構成する際に、次元圧縮による影響で、再構成されたパワーに負の値が含まれることがあった。振幅の２乗値として定義しているパワー値は本来負の値になることはない。そのため雑音駆動音声を合成するときは便宜上、そのような負のパワー値を零にするという修正をせざる得なかった。しかしながらこのような修正を行ってしまえば、パワースペクトル因子と再合成された音声との間の因果関係が多少なりとも崩れてしまうこととなる。そのため、実験１で得られた結果がパワースペクトル因子によって表現される音響的特徴によってもたらされたものであるかについて疑いがもたれるだろう。そこで本章では、パワースペクトル因子の因子負荷量をすべて非負の値に修正することで、次元圧縮を行っても負のパワー値を生じさせ得ないようにしたうえで音声を再合成した。このようにして再合成した音声を用いて実験１と同様の聴取実験を行えば、その結果を直接に因子と結びつけて考察を行うことができる。また、実験１の結果と比較することで実験１で得られた結果の妥当性について吟味することができるだろう。

本章では、まずパワースペクトル因子について、その直交性を維持した状態で非負値化する方法を説明する。パワースペクトル因子を非負値化することで因子の累積寄与率が減少することが想定されるため、これがどの程度であるかを確認する (分析２)。そして、実験２として実験１と同じ実験を非負値化したパワースペクトル因子から再合成した雑音駆動音声を用いて行い、結果を実験１と比較する。実験１で得られた結果に対する、再合成の際の便宜上の処理の影響が大きいかどうかを検討する。

3.2　分析２： パワースペクトル因子の非負値化に際する影響の量的な検討

パワー変動を再構成する際に、負のパワー値を生じさせないようにするためには、再構成に用いる因子の因子負荷量が非負の値からなっていればよい。この節では、パワースペクトル因子からどのようにして非負の因子を導出するのかを説明し、その非負の因子はパワースペクトル因子とどのような量的な違いがあるのかを確認する。

3.2.1　非負直交基底化の方法

第２章の分析１で得られたパワースペクトル因子から、直交性が維持されたまま非負値化された因子を導出するために次の手続きを行った：

1. 臨界帯域ごとに与えられている各因子の因子負荷量を比較し、０以上で最大の因子負荷量のみを保持し、残りを０に修正する。

2. 各因子ベクトルの大きさが１となるように正規化する。

分析１で得られた日本語音声の９組のパワースペクトル因子を非負値化した。各臨界帯域における非負値化したパワースペクトル因子 (以降非負直交基底因子とする) の因子負荷量を図 3.1 に示す。１つの臨界帯域のパワー変動を構成するのはただ１つの非負直交基底因子となっている。そのため、パワースペクトルの時間変動の情報が単純化されており、因子の累積寄与率はある程度減少することが想定される。

3.2.2　非負直交基底化による累積寄与率の変化

表 3.1 に分析１で得られたパワースペクトル因子と本節で導出した非負直交基底因子の累積寄与率を示す。１因子分析の場合は、元のパワースペクトル因子がすべて正の因子負荷量からなるために非負直交基底因子との間に差はないが、２因子分析以上から累積寄与率に差が現れる。パワースペクトル因子に非負直交基底化の処理を施すことによって 2.7–5.6%程、累積寄与率が低下することが分かる。この累積寄与率の低下が再合成された音声の明瞭度にどの程度の変化を与えるのかを次の実験２で確かめる。

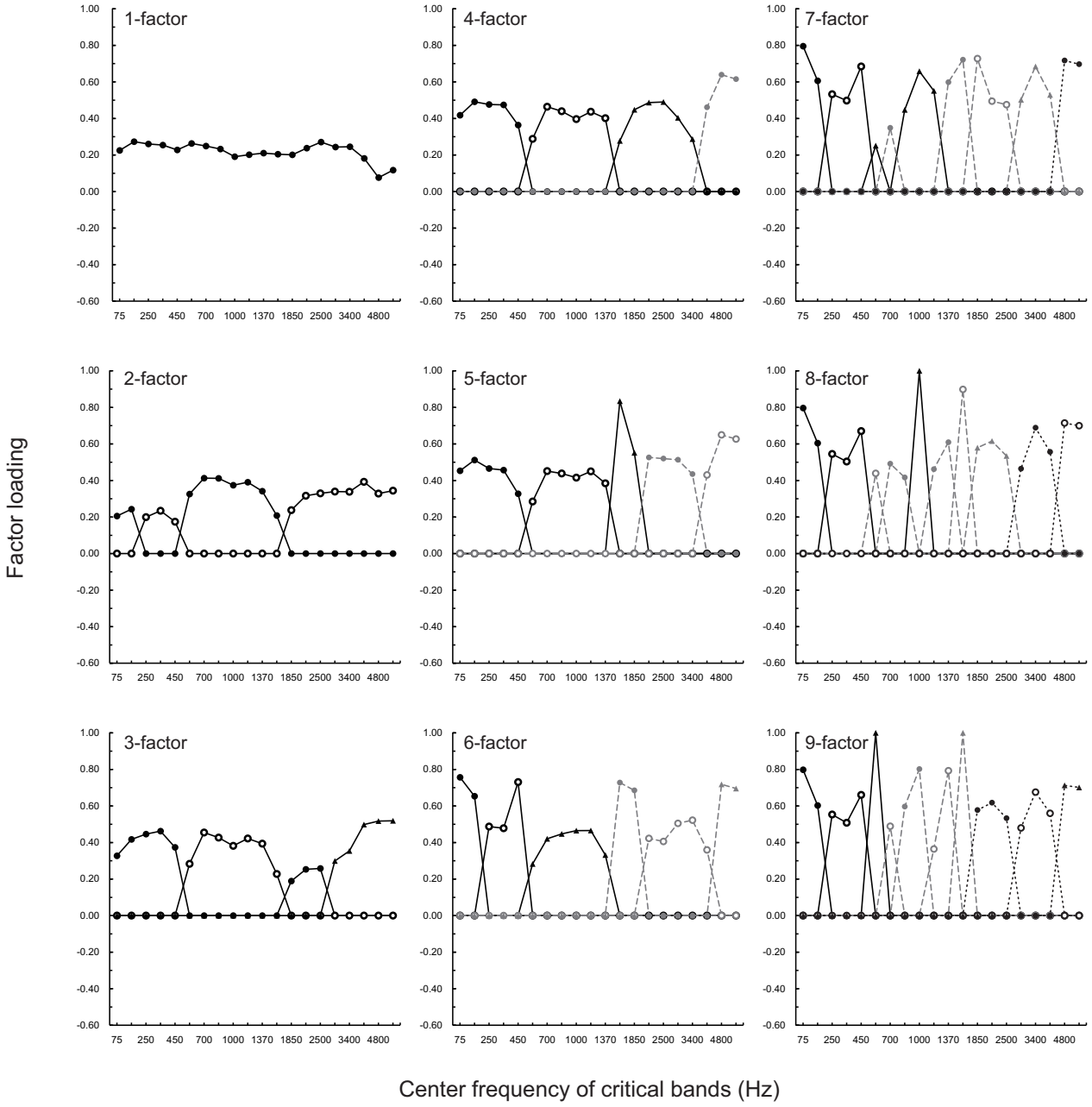

図 3.1 パワースペクトル因子を非負値化することで得られた非負直交基底因子の因子負荷量。

表 3.1 パワースペクトル因子と非負直交基底因子の累積寄与率

	Cumulative contribution (%)	
Number of factors	Spectral-change factors	Non-negative bases
1	22.9	22.9
2	37.3	33.5
3	47.9	43.4
4	55.6	52.9
5	61.3	57.0
6	66.3	61.3
7	70.5	64.9
8	74.1	68.8
9	77.6	72.3

3.3 実験２： 非負直交基底因子を用いた文音声明瞭度の測定

非負直交基底因子を用いて日本語の雑音駆動音声を再合成し、実験 1 と同じ条件で聴取実験を行った。この実験の目的は、実験 1 と同様の結果が得られるかどうかを確認することである。そのため、実験 2 において再合成に用いる因子以外の条件は可能な限り実験 1 にそろえた。実験 2 の結果が実験 1 の結果と本質的に違いがなかったならば、パワースペクトル因子を用いて音声を再合成する際に生じる便宜上の処理の影響は無視してよいとみなすことができるだろう。

3.3.1 実験参加者

実験 1 には参加していない、19〜25 歳 (平均 = 21.3 歳、$SD = 1.5$ 歳) の 6 名の男性および 6 名の女性が実験に参加した。実験参加者はすべて日本語母語話者であった。いずれの参加者も両耳ともに、純音聴力レベルが 25 ｄ B HL 以下であることを 125–8000 Hz の帯域で確認した。

本実験は九州大学大学院芸術工学研究院の倫理審査委員会の承認の下、参加同意書に実験参加者の署名を得た上で実施した。

3.3.2 実験装置

実験装置は実験 1 と同一であった (図 2.8)。

3.3.3 刺激音

刺激音の合成に用いた原音声は、実験 1 と同じく、NTT-AT 社の「多言語音声データベース 2002」にディジタル収録 (16-bit 量子化、16000 Hz サンプリング) された日本語音声 57 文を用いた。この 57 文の練習試行、ウォーミングアップ試行、本試行への割り当てもすべて実験 1 と同一にした。非負直交基底因子を用いたという点を除いて実験 1 と同一の方法で雑音駆動音声を再合成したものを刺激音とした。

3.3.4 手続き

実験手続きは実験 1 と同一であった。 1 ブロック (16 試行) あたりの所要時間は約 7 分であった。

3.3.5 結果と考察

図 3.2 に実験 2 の結果を示す。比較のために、実験 1 の結果も同時に示している。モーラ正答率は因子数が増えるごとに上昇していき、4 因子以降でその上昇は頭打ちになり、93.1%に達した。因子数の変化によってモーラ正答率が統計的に有意に異なることを、モーラ正答率に逆正弦変換を施したうえで行った一元配置分散分析によって確かめた、$F(8, 88) = 285.4$、$p < 0.001$。さらに、Tukey 法による多重比較検定によって因子数が 4 つになるまでモーラ正答率が統計的に有意に上昇していることが分かり (例えば、3 因子条件と 4 因子条件の間で、$p < 0.01$)、4 因子以上の条件間でモーラ正答率に統計的に有意な差がないことが分かった (例えば、4 因子条件と 5 因子条件の間で、$p = 0.99, n.s.$)。

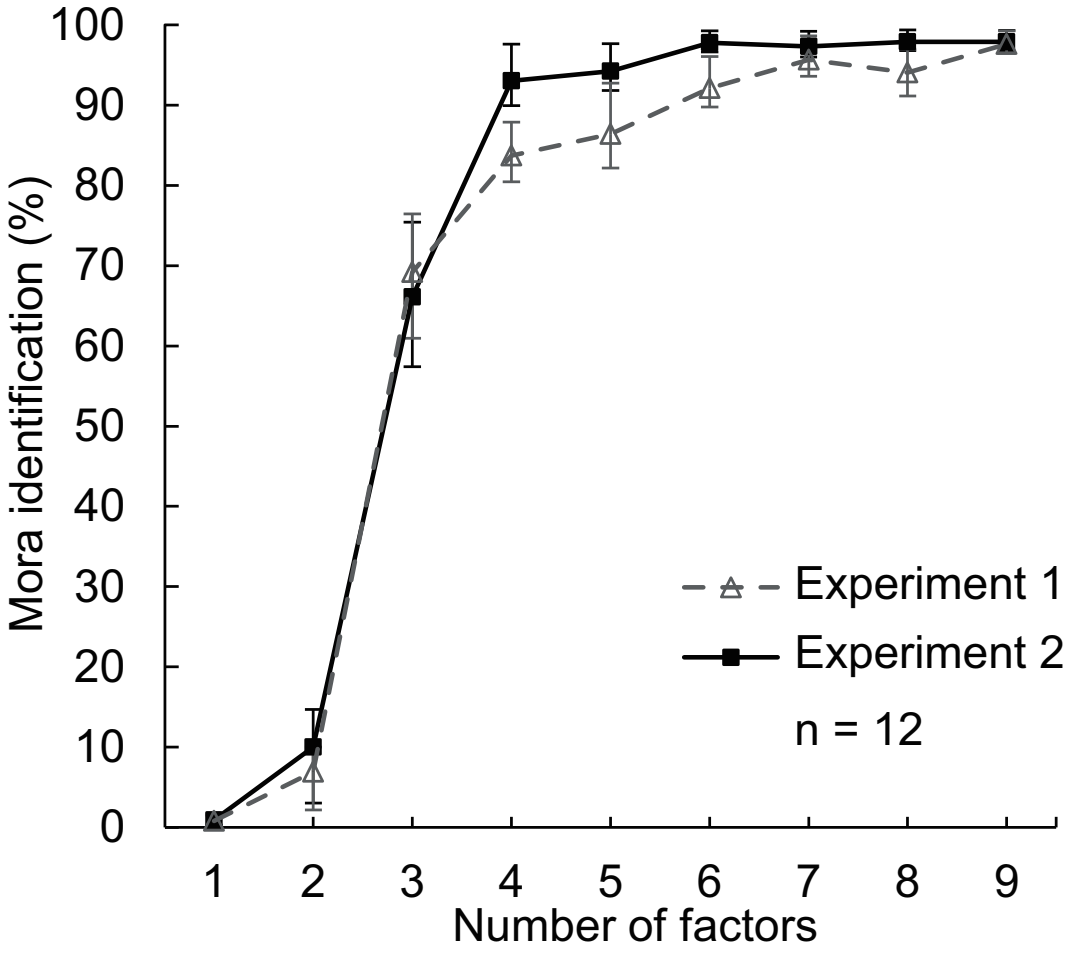

図 3.2 非負直交基底因子から再合成した音声の平均モーラ正答率。エラーバーは 95%信頼区間を示す。比較のために、実験 1 の結果も同時に示す。

実験 1 と同様に、2 因子条件と 3 因子条件の間でモーラ正答率の大きな上昇が起きた (10.0%から 66.1%)。この結果は、3 因子以上を用いることがパワースペクトル因子から明瞭な音声を再合成するのに決定的な条件であるという実験 1 で示されたことを再び支持している。また、モーラ正答率が初めて 80%を超えたのは、実験 1 と同じく 4 因子条件からであったが、モーラ正答率は 93.1%と非常に高い値であった。実験 1 と実験 2 とでそれぞれ得られたモーラ正答率の平均値が、3 因子条件間、4 因子条件間で統計的に有意な差があるかどうかをモーラ正答率に逆正弦変換を施したうえで、t 検定により調べた。棄却限界値は Bonferroni 法によって調整した。3 因子条件間においては統計的に有意な差はなかった ($t = 0.57, p = 0.58, n.s.$) が、4 因子条件間においては、実験 2 の結果の方がモーラ正答率が統計的に有意に高いことが分かった ($t = 3.90, p < 0.01$)。 非負直交基底因子の方が、累積寄与率が低い、つまりスペクトルの再現率が低いにも関わらずより高い明瞭度が得られたというのは驚きである。図 3.3 に実験に

用いられたある刺激音の原音声、それを４因子からなるパワースペクトル因子から再合成した雑音駆動音声、４因子からなる非負直交基底因子から再合成した雑音駆動音声それぞれのスペクトログラムを示す。スペクトログラムの観察を基にすると、非負直交基底因子で再合成した音声の方が明瞭度が高かった理由として次の２つが考えられる。まず第１に、負のパワーが生じなくなったために、再合成されたスペクトル上に不自然にパワー値が零となる部分がなくなり、より音声らしく聴こえるようになったからという理由である。そして第２に、各臨界帯域のパワー変動が１つの因子だけで構成されるようになったため、帯域間のパワーのコントラストをよりはっきりと感じられるようになったからという理由である。

実験２でも実験１と本質的に同じ結果が得られたことから、実験１の結果の妥当性が示された。つまり、パワースペクトル因子が２つまででは雑音駆動音声の言語内容を聴きとることができないが、３つまで用いられれば言語内容の全体の 70%程度を正確に聴きることができるようになり、４つまで用いられれば、80%以上正確に聴きとることができると分かった。２因子条件と３因子条件の間で正答率が急激に上昇するという現象から、３因子の重要性が分かる。実験で用いられた刺激音の原音声は標準的な速度で発話されたものである。また文の内容も日常会話で使われる単語を中心に構成されている。より速く発話された音声や、より内容の難しい文内容の音声に代えて実験を行った場合は、明瞭度が下がることが予想される。よって場合によっては３因子だけでは不十分であることも考えられる。標準的な原音声を４因子から再合成した場合は、80%以上正確に聴きとることができる。通常のコミュニケーションにはほとんど支障がないであろう。そこで本研究では、十分明瞭に音声を知覚するためには４因子まで必要であると強く推察する。また、実験結果の本質が変わらないのであれば、パワースペクトル因子の音声知覚上の役割を明らかにするという本論文の本来の目的に合わせて、次章以降は非負直交基底因子ではなくパワースペクトル因子を用いて議論を進めることとする。これらの４因子に関して、個々の因子がもつ音声知覚における役割は明らかにはなっていない。第４章では、個々の因子がもつ役割について確かめる実験を行う。

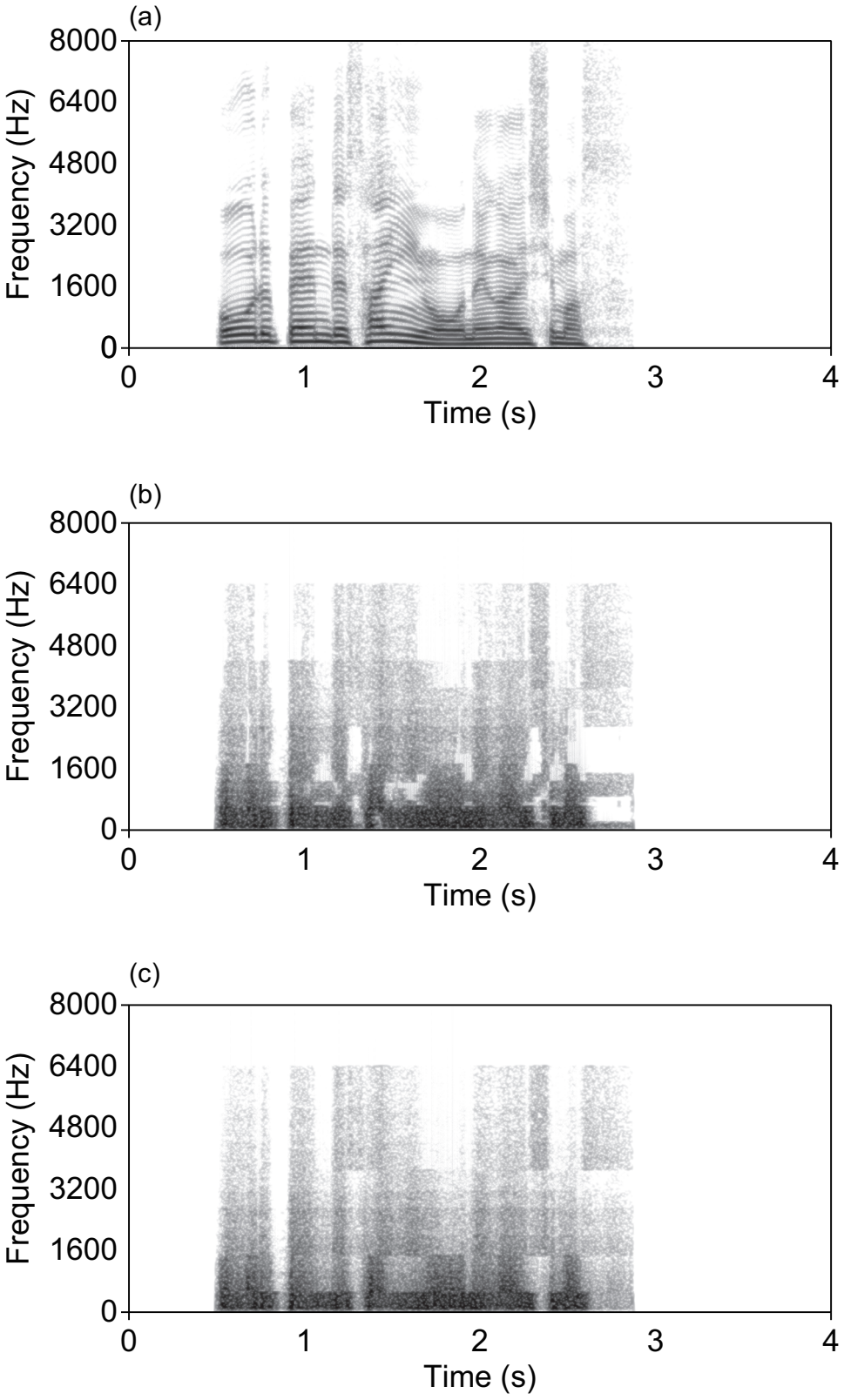

図 3.3 (a) 実験 2 に用いられたある刺激音の原音声のスペクトログラム。「ボールペンでサインをお願いします。」と発話されたもの。 (b) これを 4 因子からなるパワースペクトル因子から再合成した雑音駆動音声のスペクトログラム。(c) 4 因子からなる非負直交基底因子から再合成した雑音駆動音声のスペクトログラム。

第３章のまとめ

本章では、パワースペクトル因子を直交性を保ったまま非負値化するという方法で再合成の際に負のパワーが生じない因子を提案した。これによって因子と再合成した音声との直接的な関係を保つことができる。非負値化の影響について調べた分析２と実験２の結果は以下のようにまとめられる：

(i) 非負直交基底因子は、分析データの再合成の際に負の値を生じないという利点があるが、累積寄与率は非負値化前に比べて 2.7–5.6%程度低下する。

(ii) 非負直交基底因子から再合成した雑音駆動音声も、３因子以上用いた時に明瞭度が急激に上がった。この結果は実験１の結果から示された、３因子用いることが明瞭な日本語の雑音駆動音声の合成には決定的に重要であるということを支持する。

(iii) 音声の再合成に４因子以上を用いればほぼ完全に明瞭な音声が得られたため、非負直交基底因子から再合成した音声の方が非負値化前のパワースペクトル因子から再合成した音声よりも明瞭であると分かる。

第4章　パワースペクトル因子の個々の役割

4.1　第４章の目的

実験１および２によって、４つのパワースペクトル因子によって構成されるスペクトル変化を手がかりにして、音声を十分明瞭に知覚することができるということが分かった。それでは、パワースペクトル因子それぞれがもつ特徴や音声知覚に対する役割の違いはどのようになっているのだろうか。パワースペクトル因子の音声知覚における役割が異なるのであれば、パワースペクトル因子の内の１つの因子がもつ情報を除去したときに、音声の明瞭度に与えられる影響が、どの因子を取り除いたかによって異なるはずである。そこで本章では、パワースペクトル因子から日本語音声を再合成する際に、一部の因子が有する時間変動を除去し因子の情報が失われたときに、音声の明瞭度がどの程度変化するのかを調べることを目的とした実験３および実験４を行った。

実験３では、４因子からなるパワースペクトル因子から１つの因子の情報を取り除いて合成した音声を用いて条件ごとの明瞭度の違いを比較した。４因子からなるパワースペクトル因子がそれぞれ音声知覚において異なる役割を持っているならば、情報を取り除く因子を変えたときに、合成した音声の明瞭度に違いが現れることが予想される。明瞭度が最も低くなったときに取り除かれた因子が、４つのパワースペクトル因子の中で最も重要であり、その因子の因子負荷量が大きい周波数帯域に、音声知覚において重要な情報が含まれているということを意味するだろう。 さらに実験４では、２因子、３因子、４因子のそれぞれの因子分析の結果に対して、２つ以上の因子を用いて (それ以外の因子を除去して) 合成した音声を用いて条件ごとの明瞭度の違いを比較した。この実験によって、明瞭な音声の知覚には、音声の再合成にどのような因子を用いる必要があるのかと、因子数をいくつ以上用いる必要があるのかとを分けて考えることができる。

4.2　実験３：４因子からなるパワースペクトル因子の個々の役割

4.2.1　実験参加者

実験１・２には参加していない、20〜24 歳 (平均 = 21.6 歳、$SD = 1.3$ 歳) の５名の男性および５名の女性が実験に参加した。実験参加者はすべて日本語母語話者であり、両耳ともに純音の聴力レベルが 25 ｄ B HL 以下であることを 125–8000 Hz の範囲で確かめた。

本実験は九州大学大学院芸術工学研究院の倫理審査委員会の承認の下、参加同意書に実験参加者の署名を得た上で実施した。

4.2.2　実験装置

実験は背景雑音のレベルが 25 ｄ B A 以下である防音ブース内で行われた。刺激音はコンピュータ (BTO) にディジタル信号 (16-bit 量子化、44100 Hz サンプリング) として保存されており、オーディオインターフェース (Roland, OCTA-CAPTURE) によってアナログ信号に変換され、低域通過フィルタ (NF, DV-04 DV8FL, 遮断周波数: 15000 Hz)、グラフィックイコライザ (Roland, RDQ-2031)、ヘッドフォンアンプ (STAX, SRM-323S)、ヘッドフォン (STAX, SR-307) の順に通って実験参加者の両耳に呈示された (図 4.1)。ローパスフィルタはエイリアシング防止のために、グラフィックイコライザは再生系の周波数応答を平坦にするためにそれぞれ用いられた。ヘッドフォンアンプの出力レベルは、刺激音と同程度の音圧レベルで作成した白色雑音を再生した時に、78 ｄ B A となるように調整した。刺激音の呈示レベルの測定には、人工耳 (Brüel & Kjær, Type 4153)、ハンドヘルドアナライザ (Aco, Type 6240) を用いた。

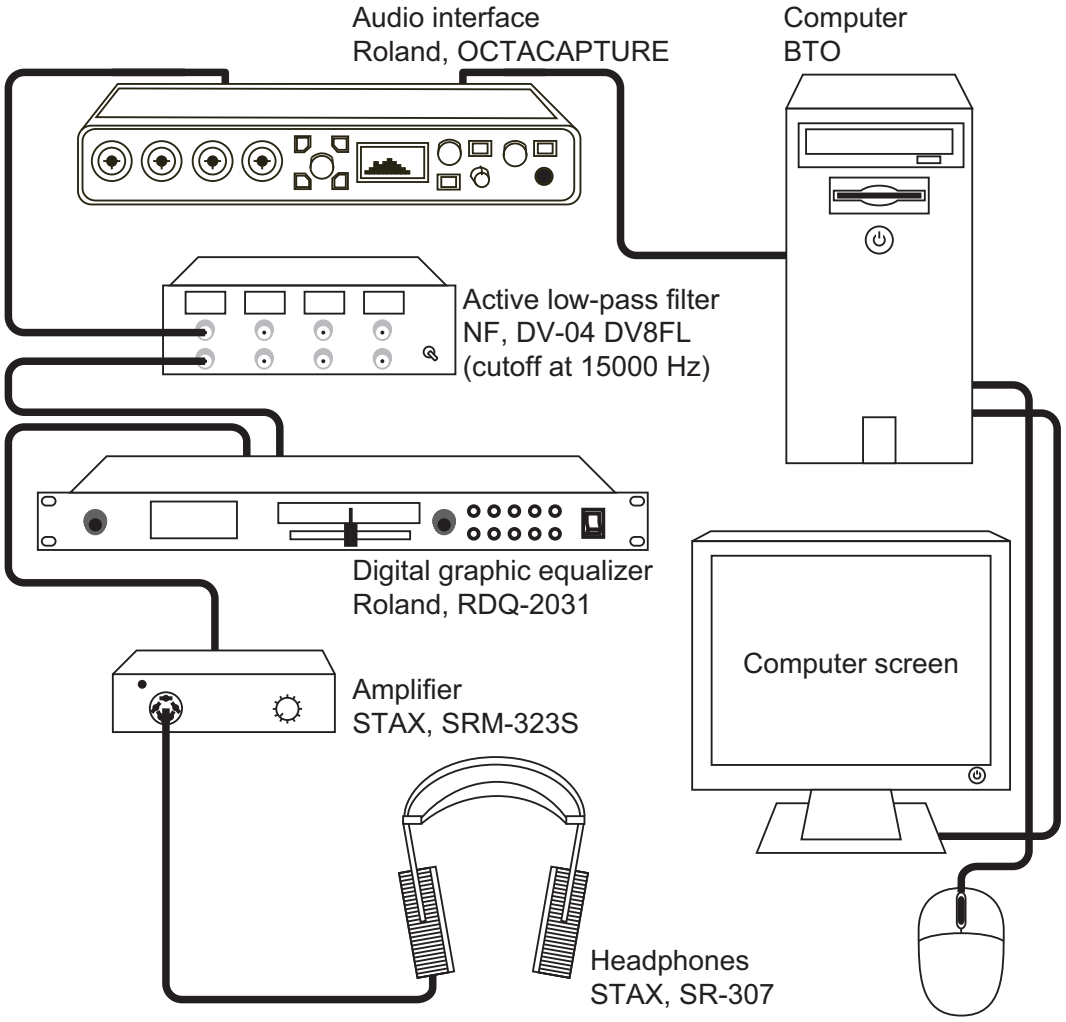

図 4.1 実験 3 および 4 に用いた装置一式。

4.2.3 刺激音

刺激音の元となる音声は、NTT-AT 社の「多言語音声データベース 2002」にディジタル収録 (16-bit 量子化、16000 Hz サンプリング) された、日本語を母語とする男性話者が発話した 40 文を利用した。この話者は実験 1・2 と同じであった。このうち 10 文は練習試行に、5 文はウォーミングアップ試行に、残りの 25 文を本試行に用いた。本試行に用いた 25 文は実験 1・2 で用いた 45 文中のリスト A〜E の文であった (表 2.3)。また、実験装置の都合上、サンプリング周波数を音声編集ソフトウェアの Praat (Boersma & Weenink, 2014) を用いて、44100 Hz に変更した。

実験 1・2 で用いられた方法と同じ方法で 4 因子からなるパワースペクトル因子 (図 4.2) を

用いて、雑音駆動音声を再合成した。以降、便宜上この４つのパワースペクトル因子を、その因子負荷量の値が最大となる臨界帯域の中心周波数が低い順に、因子 4-1、因子 4-2、因子 4-3、因子 4-4 と呼ぶこととする。例えば、510–1480 Hz に設定された６つの臨界帯域において大きい因子負荷量をもつ因子は因子 4-2 である。

実験３では５つの条件で雑音駆動音声を再合成した。４因子からなるパワースペクトル因子のすべてを用いて再合成する条件と、４因子のうち一つが与える臨界帯域におけるパワーの時間変動の情報を取り除いて再合成する条件 (因子除去条件) とからなる５条件である。因子除去条件では、まず除去をしない３つの因子だけから再構成される各臨界帯域のパワーの時間変動を計算した。次に、残り一つの因子が与える各臨界帯域におけるパワーの時間変動の情報を除去するために、その因子によって与えられるパワー変動を一定の値に置き換えた。具体的には、除去する因子によって再構成されるパワー変動を計算した後、それぞれの帯域内でパワー変動の最大値を求め、その最大値の２倍の値に帯域内のすべてのパワー値を変換した。除去する因子によって再構成されうるパワー変動よりも十分に大きい一定のパワーの値を作ることで、その因子が与える情報を除去した。このようにしてできた定常的なパワーを、３つの因子によって再構成された臨界帯域ごとのパワー変動に足し合わせた。

以上の処理を経て算出された各臨界帯域のパワーの時間変動を実現する雑音駆動音声が再合成された。刺激音は実験参加者に３回繰り返して聴かせるため、1.5 s の間隔をあけて３回繰り返したものを１つの音声ファイルとして保存した。なお、各繰り返しにおいて駆動された帯域雑音はその都度生成しているため、完全に同一の刺激音が３回繰り返されるわけではない。また、定常的なパワーによる雑音成分は音声の開始の２ s 前から始まり、３回目の繰り返しの２ s 後まで続くようにした。このようにしたのは、実験参加者に音声部分と定常雑音とが分離して聴かれるようにするためである。定常雑音には振幅の急な立ち上がりと立ち下がりによるスペクトルの広がりを防ぐ目的で、振幅包絡がコサイン形状となる 66.2 ms の立ち上がり区間と立ち下がり区間を付けた。刺激音のスペクトログラムの例を図 4.3 に示す。

実験参加者は５個の文リストの音声全てを聴取するが、それぞれの文リストがどの条件で再合成された音声であるかが、全ての実験参加者で異なるようにした。例えば、実験参加者１が文リスト A に対して因子 4-1 除去の条件で合成した音声を聴き、実験参加者２は同じ文リスト A に対して因子除去なし条件で合成した音声を聴いた (表 4.1)。このようにして、文の違いが与える結果への影響が最小限となるようにした。

表 4.1 実験参加者毎の除去因子数条件と文リストの対応。

	Sentence list				
	A	B	C	D	E
Participant I	4-1	4-2	4-3	4-4	none
Participant II	none	4-1	4-2	4-3	4-4
Participant III	4-4	none	4-1	4-2	4-3
Participant IV	4-3	4-4	none	4-1	4-2
Participant V	4-2	4-3	4-4	none	4-1
Participant VI	none	4-4	4-3	4-2	4-1
Participant VII	4-4	4-3	4-2	4-1	none
Participant VIII	4-3	4-2	4-1	none	4-4
Participant IX	4-2	4-1	none	4-4	4-3
Participant X	4-1	none	4-4	4-3	4-2

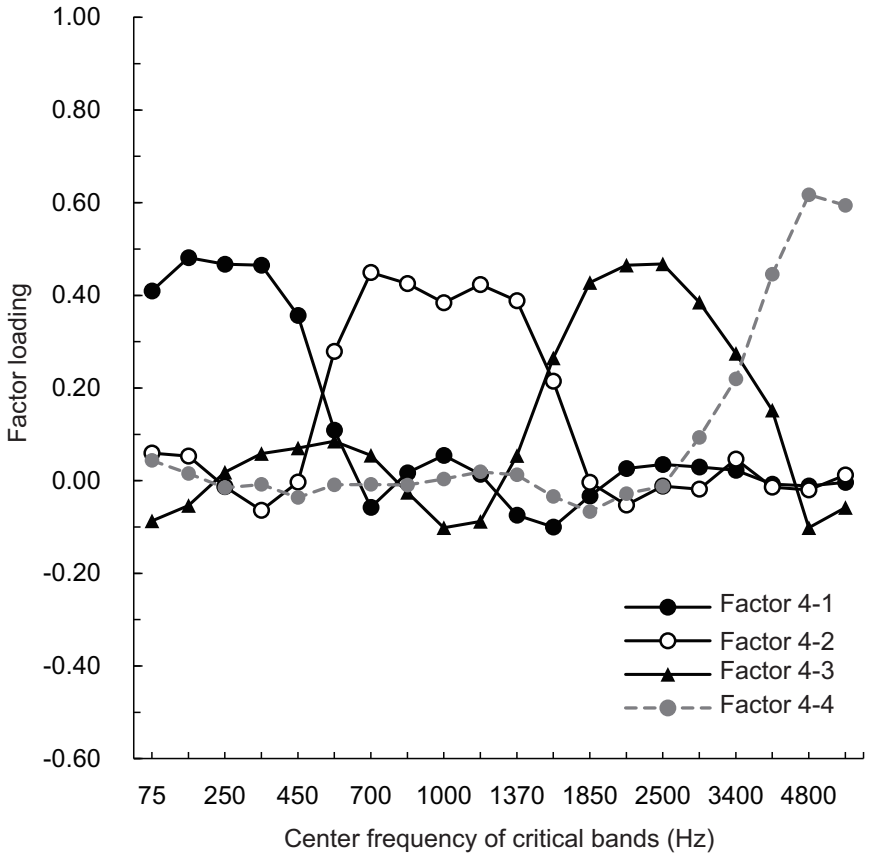

図 4.2 音声の再合成に用いられた 4 因子分析のパワースペクトル因子。便宜上、各因子の因子負荷量が高い帯域が低い順に、因子 4-1、4-2、4-3、4-4 とする。

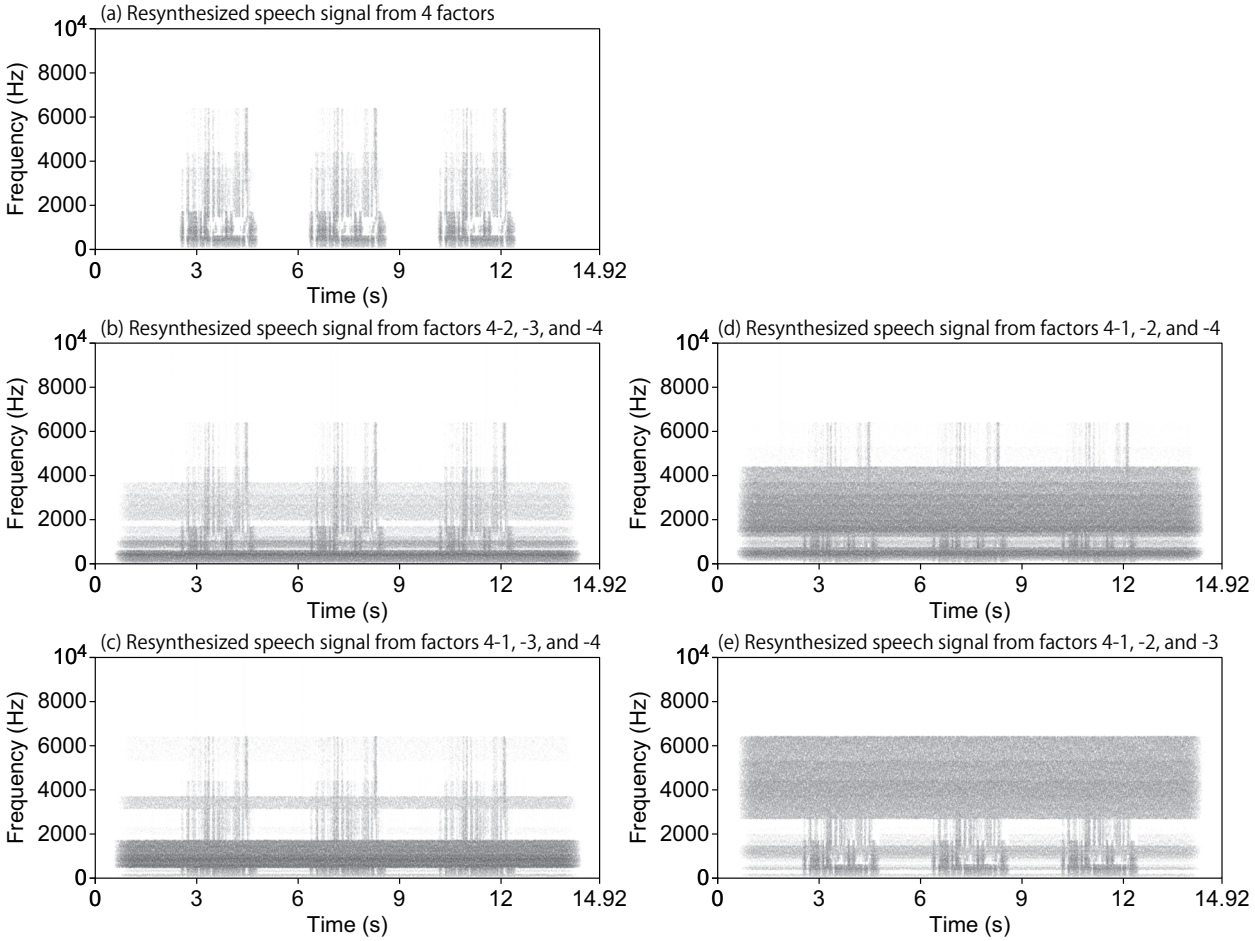

図 4.3 刺激音のスペクトログラムの例。ここでは、日本語母語の男性が発話した、「こきょうを離れてみるのもいいでしょう。」という音声を抽出因子数を 4 つにした因子分析で得られたパワースペクトル因子を用いて、5 種類の雑音駆動音声として再合成したものを載せる。原音声が 1.5 s の間隔をあけて 3 回繰り返し再合成されたものが一つの刺激音として用いられた。(a) 4 つのパワースペクトル因子すべてを用いて再合成したもの。(b)〜(e) は 4 つの因子のうちのひとつの因子によって与えられる、臨界帯域ごとのパワーの時間変動の情報を取り除いて再合成したもの。(b) は因子 4-1 を、(c) は因子 4-2 を、(d) は因子 4-3 を、(e) は因子 4-4 をそれぞれ除去して再合成したものである。

4.2.4 手続き

実験はまず実験参加者を回答に慣れさせるための練習試行ブロックから始まり、続いて本試行ブロックが 2 ブロック行われた。練習ブロックでは、10 試行で 5 条件を 2 試行ずつ実験参加者に聴かせた。本試行ブロックの刺激数は第 1 ブロックが 13 刺激、第 2 ブロックが 14 刺激であり、各ブロックの第 1 番目の試行はウォーミングアップ試行であった。練習試行およびウォーミングアップ試行で得られたデータは結果の分析には用いなかった。

実験参加者は、ヘッドフォンを装着した状態で椅子に座り、目の前のコンピュータディスプレイに表示される刺激音再生ボタンを 1 試行ごとにクリックするように求められた。再生ボタンがクリックされた 0.5 s 後に刺激音の再生が始まり、実験参加者は刺激音の聴取後、聴こえた内容をひらがなで回答用の別のコンピュータに入力した。実験参加者にははっきりと聴き取れなかった箇所を推測して回答することを避けるように教示した。刺激音の呈示順序は練習ブロック、本試行ブロックそれぞれで無作為な順序にした。ウォーミングアップ試行では、ウォーミングアップ試行用の刺激音 5 つのうち、2 つが無作為に選ばれ、実験参加者に呈示された。1 ブロック (13 または 14 試行) あたりの所要時間は約 10 分であった。

4.2.5 結果と考察

実験参加者から得られた回答と正答の文とを比較し、モーラ単位で正答率を算出した。図 4.4 に各条件におけるモーラ正答率の全実験参加者の平均値およびその 95%信頼区間を示す。4 つの因子すべてを用いて再合成した雑音駆動音声のモーラ正答率は 83.6%であった。4 つの因子のうち、1 つの因子の時間変動を除去した条件では、モーラ正答率の低下がみられた。これらの条件でのモーラ正答率は、因子 4-1 除去条件では、64.1%、因子 4-2 除去条件では、36.8%、因子 4-3 除去条件では、60.1%、そして因子 4-4 除去条件では、61.4%であった。

モーラ正答率を逆正弦変換したうえで統計的検定を行った。一元配置分散分析によって、5 つの条件間で、モーラ正答率の平均値に統計的に有意な差があることが示された、$F(4, 36) = 22.29$、$p < 0.0001$。また、Tukey 法を用いて、各条件間におけるモーラ正答率の平均値の対比較を行ったところ、因子除去なし条件と、その他の 4 つの条件とがそれぞれ 1 %水準でモーラ正答率の平均値に統計的に有意な差があることが分かり、また因子 4-2 除去条件と、因子 4-1、4-3、4-4 除去条件との間にそれぞれ 1 %水準でモーラ正答率の平均値に統計的に有意な差があることが分かった。因子 4-1, 4-3, 4-4 除去条件間ではモーラ正答率に統計的に有意な差はなかった (例えば、因子 4-1 除去条件と因子 4-3 除去条件間：$p = 0.97, n.s.$)。

4 因子すべてを用いて再合成した雑音駆動音声のモーラ正答率は実験 1 の結果とほぼ一致した。やはり日本語音声において、4 因子用いることで十分に明瞭な雑音駆動音声を再合成できることが確かめられた。

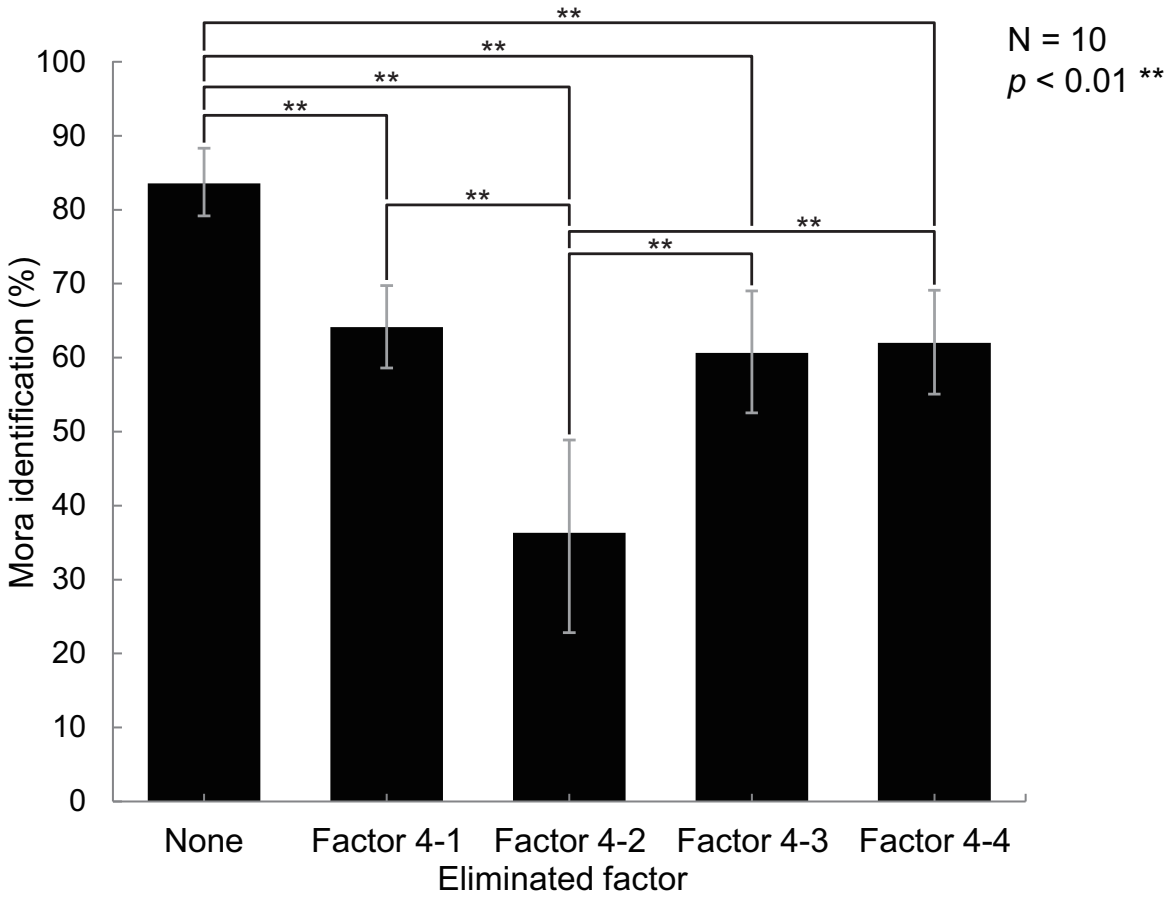

図 4.4 4 因子からなるパワースペクトル因子から 1 つの因子を除去して合成した雑音駆動音声の明瞭度。エラーバーは 95%信頼区間。

因子の時間変動を除去することにより、雑音駆動音声の明瞭度が低下することが示され、特に因子 4-2 の除去が明瞭度の低下に与える影響が大きいことが示された。よってこの実験の結果から、4 因子からなるパワースペクトル因子の中の、因子 4-1、4-3、4-4 によって与えられる情報は同程度の量であるが、因子 4-2 によって与えられる情報、すなわち 510–1480 Hz に設定された 6 つの臨界帯域において大きい因子負荷量をもつ因子が与えるスペクトルの時間変動の情報は日本語音声の知覚に特に重要であることが明らかとなった。

4 因子の中で最も多くの音声知覚の手がかりを与えると考えられる因子 4-2 の構造に注目すると、3 因子からなるパワースペクトル因子の中にも、ほぼ同じ構造の因子があることが見つかるが、2 因子からなるパワースペクトル因子の中には因子 4-2 と対応するような構造の因子は現れていない (図 2.4)。実験 1 ・ 2 では、再合成に用いるパワースペクトル因子の個数が 3

つ以上となることが明瞭な雑音駆動音声を再合成する条件であると結論づけたが、単に因子の個数ではなく、因子 4-2 のような構造をした因子が現れるかどうかがより重要であるとも考えられる。

次節では、実験３の条件を拡張した実験を行う。２因子からなるパワースペクトル因子、３因子からなるパワースペクトル因子、４因子からなるパワースペクトル因子を用いて音声を再合成し、510–1480 Hz に設定された６つの臨界帯域において大きい因子負荷量をもつ因子が含まれる条件とそうでない条件とで再合成音声の明瞭度がどれだけ異なるのかを検討する。

4.3 実験４： ２因子、３因子、４因子からなるパワースペクトル因子の個々の役割

実験３で、510–1480 Hz に設定された６つの臨界帯域において大きい因子負荷量をもつパワースペクトル因子によって与えられる情報が音声の知覚に重要であることが示された。この因子は、取り出す因子数を３因子にした—３つの主成分を取り出して、それをバリマックス回転して得た—ときから初めて現れる (図 4.5)。よってまだこの因子が現れない２因子からなるパワースペクトル因子よりも、３因子からまたは４因子からなるパワースペクトル因子の内の、510–1480 Hz に設定された６つの臨界帯域において大きい因子負荷量をもつ因子を含んだ２つの因子の方が音声を明瞭に知覚するのに役立つのではないかと考えられる。実験４では、２因子から、３因子から、そして４因子からなるパワースペクトル因子のそれぞれ中から、２因子以上を用いて再合成した音声の明瞭度を測定し、上述の考えを確かめることを目的とする。

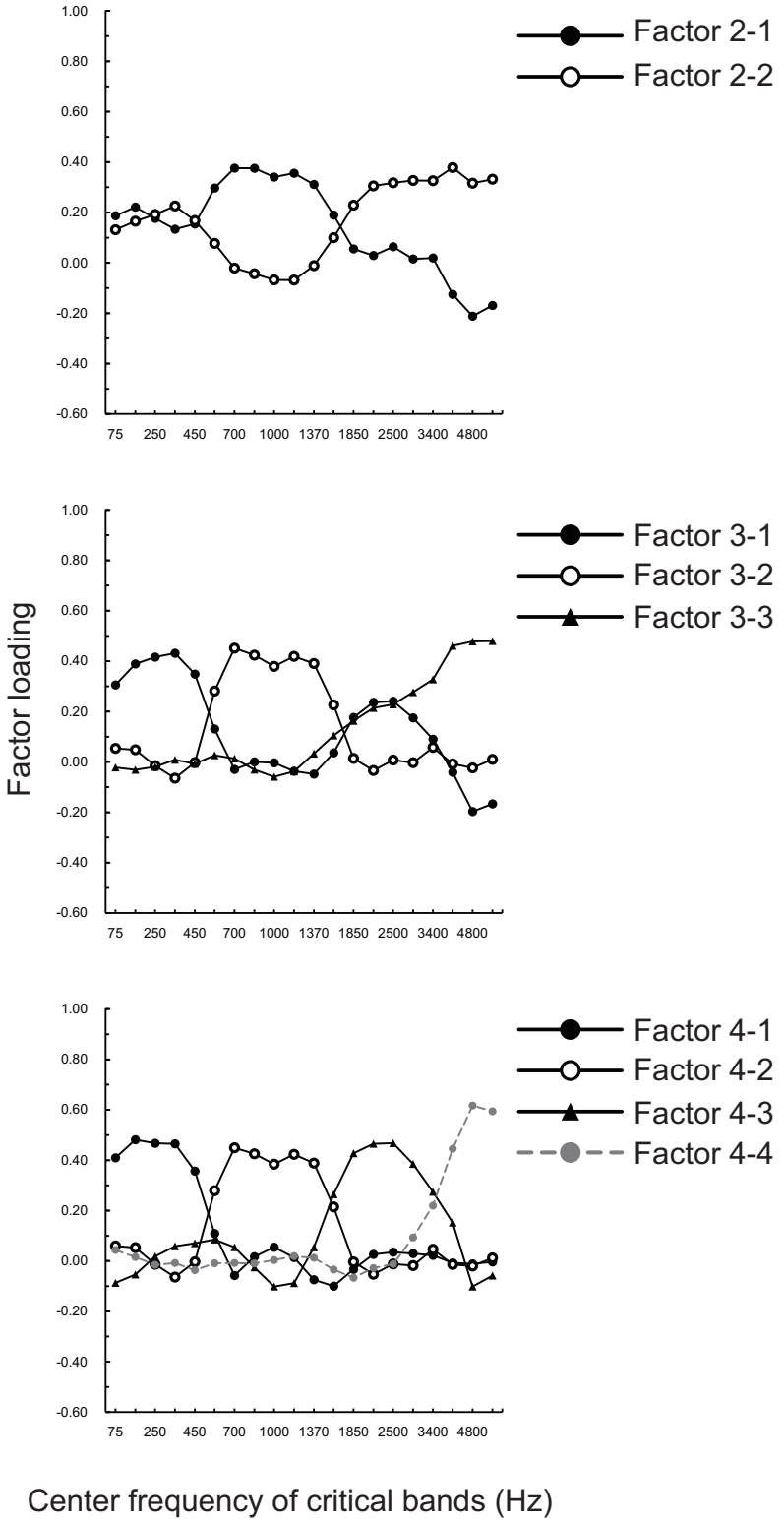

図 4.5 起点移動主成分分析を行って得られたパワースペクトル因子の因子負荷量。上段から、２因子、３因子、４因子をそれぞれ抽出した場合の結果である。

4.3.1 実験参加者

実験 1 ・ 2 ・ 3 に参加していない、19〜23 歳（平均＝ 20.6 歳、$SD = 1.2$ 歳）の 8 名の男性および 8 名の女性が実験に参加した。実験参加者はすべて日本語母語話者であり、両耳ともに純音の聴力レベルが 25 ｄ B HL 以下であることを 125–8000 Hz の範囲で確かめた。

本実験は九州大学大学院芸術工学研究院の倫理審査委員会の承認の下、参加同意書に実験参加者の署名を得た上で実施した。

4.3.2 実験装置

実験装置は実験 3 と同一のものを用いた (図 4.1)。

4.3.3 刺激音

刺激音の元となる音声は、実験 1 で用いた音声と同じく、NTT-AT 社の「多言語音声データベース 2002」にディジタル収録 (16-bit 量子化、16000 Hz サンプリング) されたものから選出した。本試行のために 80 文、練習試行のために 16 文、そしてウォーミングアップ試行のために 16 文が選ばれた。本試行のための 80 文のうちの 45 文は実験 1 ・ 2 で用いられたものと同一のものであった。新たに追加した 35 文を表 4.2 に載せる。

表 4.2: 実験 4 から新たに追加された 35 文。

No.	List	Database No.	Sentence	Number of mora	Average
46		80	たばこの煙で喉を痛めました	17	
47		149	公園で駐車をしてはいけません	17	
48	J	13	小刀やハサミは素朴な道具です	18	18.2
49		53	友人は温和で誠実な人柄です	20	
50		33	お城の前で記念撮影をしました	19	
51		81	相変わらず熱帯夜が続きます	17	
52		154	洗剤で河川が汚れてしまった	17	
53	K	199	来年こそスキーに行くつもりです	17	18.2
54		47	今のところ乗客の足は順調です	20	
55		28	古墳時代の遺跡が発見されました	20	
56		94	毎朝犬をつれて散歩をします	17	
57		157	今夜台風が上陸しそうです	17	
58	L	141	部屋には机の他なにもありません	18	18.2
59		163	読者の意見に共感を覚えました	19	
60		9	幼稚園児を連れて芋掘りに出かけます	20	
61		99	最近街はゴミで汚れています	17	
62		168	真新しい畳の匂いがします	17	
63	M	62	当時このあたりはへんぴな田舎でした	19	18.4
64		185	空調設備の点検をして下さい	19	
65		31	審査の結果第二位に入賞しました	20	
66		121	老若男女が祭りに参加した	17	
67		174	デビューとともにトップスターとなった	17	
68	N	109	マヒの原因はコンピューターの故障だ	19	18.4
69		188	編集記事の題材を考えましょう	19	
70		45	もうすぐ宇宙に行ける時代が来るでしょう	20	
71		129	今にも夕立が降ってきそうです	17	
72		176	甲子園の砂を持ち帰りました	17	
73	O	117	雑誌も多様化の時代に入りました	19	18.2
74		197	動物園は家族連れでいっぱいです	19	
75		155	この先三キロメートルの渋滞です	19	
76		140	沼のほとりに大きな木があります	17	
77		184	池のまわりを走るのは危険です	17	
78	P	136	すき焼きの材料を買ってきて下さい	19	18.2
79		198	運動会は雨天のため延期します	19	
80		5	満天の星が夜空にまたたいている	19	

実験３で用いた方法と同じ方法で原音声から雑音駆動音声を再合成した。音声の再合成に用いられたパワースペクトル因子は、２因子からなるパワースペクトル因子、３因子からなるパワースペクトル因子、４因子からなるパワースペクトル因子の３種類とした (図 4.5)。２因子からなるパワースペクトル因子と３因子からなるパワースペクトル因子についても、以降は便宜上因子負荷量の値が最大となる臨界帯域の中心周波数が低い順に、因子 2-1、因子 2-2、因子 3-1、因子 3-2、因子 3-3 と呼ぶこととする。

３種類の因子の組の中から２個以上の因子を組み合わせて用いることによって、16 条件が設定された。因子の組み合わせの 16 種類は表 4.3 に示すとおりである。因子の除去が行われる条件はそのうちの 13 条件である。さらに 13 条件中６条件は４因子からなるパワースペクトル因子から２因子を除去する条件である。この条件においては除去される２つの因子それぞれによって生じうる各臨界帯域のパワーの時間変動が帯域ごとに合計された後に、パワーの最大値を求め、その最大値の２倍の値にすべてのパワーを置換した。このようにしてできた定常的なパワーを、残り２つの因子によって再構成された臨界帯域のパワー変動に足し合わせた。１つの因子を除去する７条件と因子の除去をしない３条件では、実験３と同一の方法で臨界帯域ごとのパワーの時間変動が作られた。

以上の処理を経て算出された各臨界帯域のパワーの時間変動を実現する雑音駆動音声が再合成された。１つの音声ファイルの構成は実験３と同様で、因子が除去される条件では音声部分の開始２ s 前に定常的なパワーによる雑音成分が 66.2 ms の立ち上がりで始まり、雑音成分が生じている中で音声部分が 1.5 s の時間間隔で３回繰り返され、音声終了の２ s 後に 66.2 ms の立ち下がりで雑音成分が終了するというものであった。雑音成分の立ち上がり立ち下がりの振幅包絡はコサイン形状となるようにした。

実験参加者は 16 の文リストの音声全てを聴取するが、それぞれの文リストがどの条件で再合成された音声であるかが、全ての実験参加者で異なるようにした。例えば、実験参加者１が文リスト A に対して表 4.3 の条件１で合成した音声を聴き、実験参加者２は同じ文リスト A に対して表 4.3 の条件９で合成した音声を聴いた (表 4.4)。このようにして、文の違いが与える結果への影響が最小限となるようにした。

表 4.3 刺激音の合成に用いられたパワースペクトル因子の組み合わせ。

Condition no.	Factors used in resynthesizing			
1	2–1	2–2	–	–
2	3–1	3–2	–	–
3	3–1	3–3	–	–
4	3–2	3–3	–	–
5	4–1	4–2	–	–
6	4–1	4–3	–	–
7	4–1	4–4	–	–
8	4–2	4–3	–	–
9	4–2	4–4	–	–
10	4–3	4–4	–	–
11	3–1	3–2	3–3	–
12	4–1	4–2	4–3	–
13	4–1	4–2	4–4	–
14	4–1	4–3	4–4	–
15	4–2	4–3	4–4	–
16	4–1	4–2	4–3	4–4

表 4.4 実験参加者毎の条件と文リストの対応。表内の数字は表 4.3 の条件番号を示す。

	Sentence list															
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Participant I	1	11	4	3	2	16	15	14	13	12	10	7	5	6	8	9
Participant II	9	1	11	4	3	2	16	15	14	13	12	10	7	5	6	8
Participant III	8	9	1	11	4	3	2	16	15	14	13	12	10	7	5	6
Participant IV	6	8	9	1	11	4	3	2	16	15	14	13	12	10	7	5
Participant V	5	6	8	9	1	11	4	3	2	16	15	14	13	12	10	7
Participant VI	7	5	6	8	9	1	11	4	3	2	16	15	14	13	12	10
Participant VII	10	7	5	6	8	9	1	11	4	3	2	16	15	14	13	12
Participant VIII	12	10	7	5	6	8	9	1	11	4	3	2	16	15	14	13
Participant IX	13	12	10	7	5	6	8	9	1	11	4	3	2	16	15	14
Participant X	14	13	12	10	7	5	6	8	9	1	11	4	3	2	16	15
Participant XI	15	14	13	12	10	7	5	6	8	9	1	11	4	3	2	16
Participant XII	16	15	14	13	12	10	7	5	6	8	9	1	11	4	3	2
Participant XIII	2	16	15	14	13	12	10	7	5	6	8	9	1	11	4	3
Participant XIV	3	2	16	15	14	13	12	10	7	5	6	8	9	1	11	4
Participant XV	4	3	2	16	15	14	13	12	10	7	5	6	8	9	1	11
Participant XVI	11	4	3	2	16	15	14	13	12	10	7	5	6	8	9	1

4.3.4 手続き

実験参加者を回答に慣れさせるための 16 試行からなる練習ブロックと、各 21 試行からなる４つの本ブロックで実験を構成した。練習ブロックでは、16 試行で 16 条件を 1 試行ずつ実験参加者に聴かせた。本ブロックの各第 1 試行目はウォーミングアップ試行であり、ウォーミングアップ試行の為に用意された 16 条件に対応する 16 種類の刺激音のうちから無作為に選ばれた４つの刺激音をそれぞれのブロックで呈示した。全ての刺激は無作為な順番で呈示された。上記で説明した試行数およびブロック数を除いて、全ての手続きは実験 3 と同一の方法を用いた。 1 ブロック (21 試行) あたりの所要時間は約 10 分であった。

4.3.5 結果と考察

実験参加者から得られた回答と正答の文とを比較し、モーラ単位で正答率を算出した。図 4.6 に、各条件におけるモーラ正答率の全実験参加者の平均値及びその 95%信頼区間を示す。モーラ正答率は因子を 2 個用いた条件で 8.7–31.8%の範囲に、因子を 3 個用いた条件で 38.7–60.8%の

範囲に、そして因子を４個用いた条件で81.8%となり、合成に用いられる因子の個数が多いほど高かった。

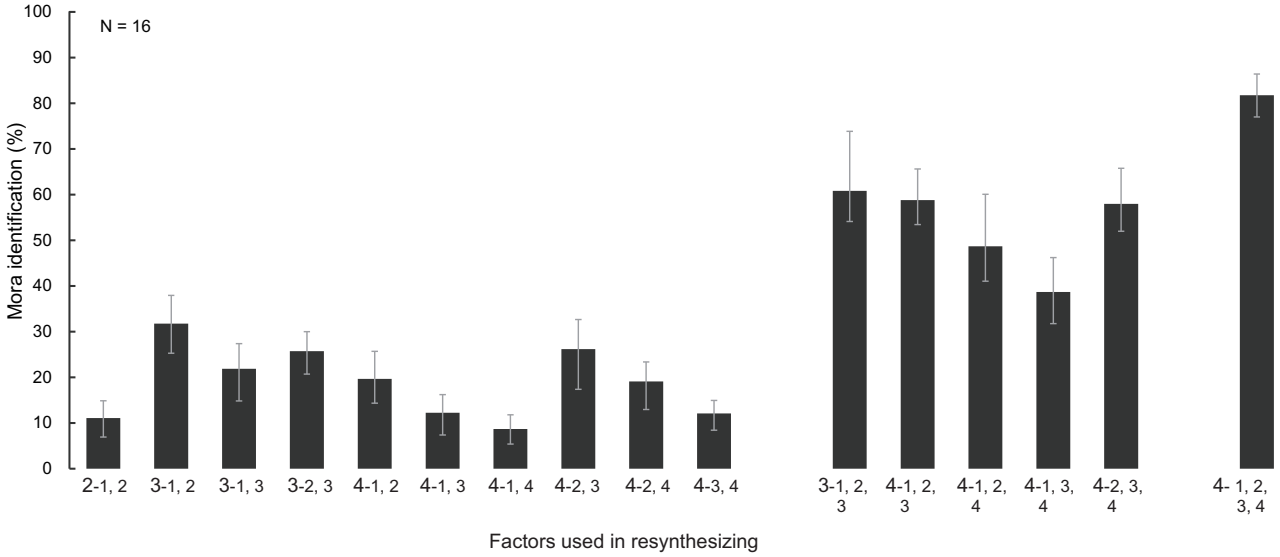

図 4.6 ３組のパワースペクトル因子 (２因子分析、３因子分析、４因子分析) 中の２因子以上を用いて再合成した雑音駆動音声の明瞭度。エラーバーは95%信頼区間。

モーラ正答率を逆正弦変換したうえで統計的検定を行った。対応のある一元配置分散分析を行った結果、モーラ正答率に条件間で統計的に有意な差があることが示された、$F(15, 225) = 71.3$、$p < 0.001$。また、Tukey 法による多重比較検定を因子数 2 の条件間のすべての組み合わせで行ったところ、45 個の組み合わせの内、13 個の組み合わせにおいて１%水準でモーラ正答率に統計的に有意な差がある組み合わせがあることが、２個の組み合わせにおいて５%水準でモーラ正答率に統計的に有意な差がある組み合わせがあることが分かった (図 4.7)。１%水準でモーラ正答率に統計的に有意な差があった組み合わせは、因子 2-1, -2 vs. 因子 3-1, -2、因子 2-1, -2 vs. 因子 3-2, -3、因子 2-1, -2 vs. 因子 4-2, -3、因子 3-1, -2 vs. 因子 4-1, -3、因子 3-1, -2 vs. 因子 4-1, -4、因子 3-1, -2 vs. 因子 4-3, -4、因子 3-1, -3 vs. 因子 4-1, -3、因子 3-2, -3 vs. 因子 4-1, -3、因子 3-1, -3 vs. 因子 4-1, -4、因子 3-2, -3 vs. 因子 4-3, -4、因子 4-1, -3 vs. 因子 4-2, -3、因子 4-1, -4 vs. 因子 4-2, -3、因子 4-2, -3 vs. 因子 4-3, -4 であった。５%水準でモーラ正答率に統計的に有意な差があった組み合わせは、因子 3-1, -2 vs. 因子 4-2, -4、因子 4-1, -2 vs. 因子 4-2, -4 であった。また、モーラ正答率の差に有意傾向があったのは因子 2-1, -2 vs. 因子 3-1, -3 ($p = 0.061$) と因子 4-1, -4 vs. 因子 4-2, -4 ($p = 0.054$) の２つの組み合わせであった。

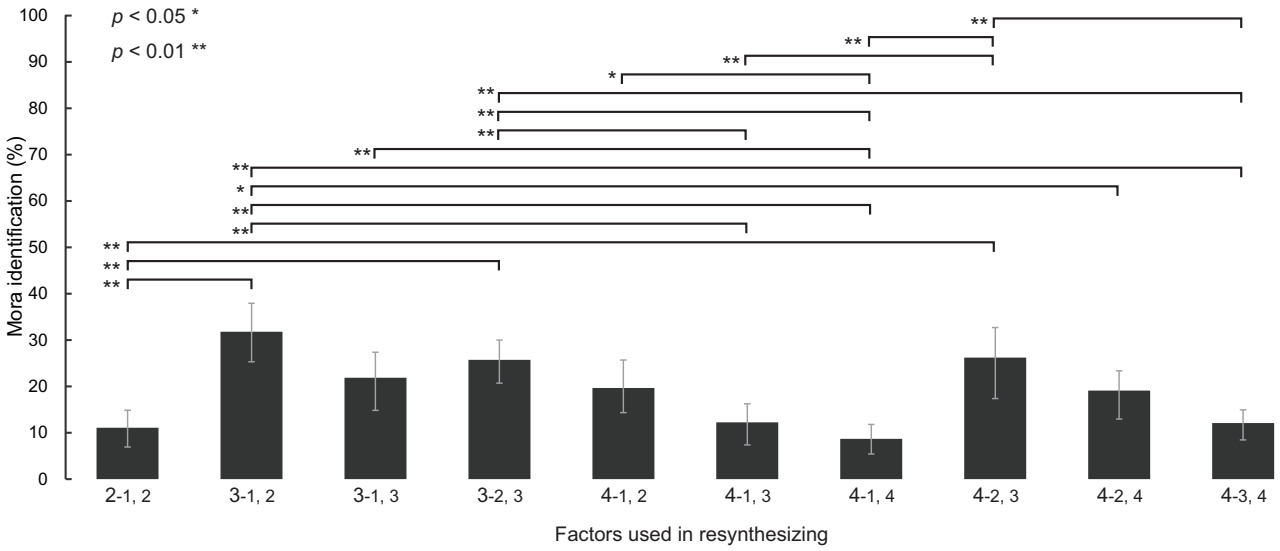

図 4.7 3 組のパワースペクトル因子 (2 因子分析、 3 因子分析、 4 因子分析) 中の 2 因子を用いて再合成した雑音駆動音声の明瞭度。エラーバーは 95%信頼区間。

統計的に有意な差があった上記の 15 の組み合わせの内、[因子 3-1, -2 vs. 因子 4-2, -4] と [因子 3-1, -3 vs. 因子 4-1, -4] を除く組み合わせは、片方の条件に 510–1480 Hz に設定された 6 つの臨界帯域において大きい因子負荷量をもつ因子 (因子 3-2 または因子 4-2) を含む組み合わせであった。一方、因子 3-2 または因子 4-2 を含む条件と含まない条件の組み合わせで、正答率に統計的に有意な差がなかったのは、因子 2-1, -2 vs. 因子 4-2, -4 ($p = 0.99, n.s.$)、因子 3-1, -2 vs. 因子 3-1, -3 ($p = 0.18, n.s.$)、因子 3-1, -3 vs. 因子 3-2, -3 ($p = 0.97, n.s.$)、因子 4-1, -2 vs. 因子 4-1, -4 ($p = 0.25, n.s.$)、因子 4-1, -2 vs. 因子 4-3, -4 ($p = 0.26, n.s.$)、因子 4-2, -4 vs. 因子 4-3, -4 ($p = 0.60, n.s.$) の 6 つの組み合わせであった。多重比較検定の結果は、因子 3-2 と因子 4-2 を含む条件がそれらを含まない条件よりも明瞭度が高くなることが多いことを示している。

実験 4 の結果を全体的に観察すると、再合成に用いられる因子の個数が多いほど、明瞭度は高くなる傾向があると分かる。実験 3 で 4 因子からなるパワースペクトル因子の中で最も重要であると分かった因子 4-2 を含んでいたとしても、因子が 2 個除去された条件では音声のモーラ正答率は 20%前後であり、明瞭ではなくなることが分かった。しかし、因子 4-2, -3 条件は、因子 4-2 を含まない条件である因子 4-1, -3 条件、因子 4-1, -4 条件、そして因子 4-3, -4 条件よりも明瞭度が高いことが示された。また、同じく因子 4-2 を含む条件である因子 4-2, -4 条件

とは正答率に統計的に有意な差はなかった ($p = 0.71, n.s.$)。統計的検定では、3 因子からなるパワースペクトル因子の中から 2 因子を用いた条件の間に明瞭度に違いがあることは示されなかった。しかし、510–1480 Hz に設定された 6 つの臨界帯域において大きい因子負荷量を持つ因子である因子 3-2 を含む条件は、この因子と対応する因子 4-2 を含まない 3 つの条件や、2 因子からなるパワースペクトル因子を用いた条件よりも得られた明瞭度が高い。以上のことから、因子の個数だけでなく、510–1480 Hz に設定された 6 つの臨界帯域において大きい因子負荷量をもつ因子が合成に用いられているかどうかも、雑音駆動音声の明瞭度を決定する要因であると考えられる。この 510–1480 Hz に設定された 6 つの臨界帯域において大きい因子負荷量をもつ因子の意味づけについては先行研究の結果と比較しながら次章で考察する。

第 4 章のまとめ

本章ではパワースペクトル因子の個々の役割を明らかにするために、1 つの因子が与えるスペクトルの時間変動の情報を除去して音声の再合成をしたときに、音声の明瞭度にどれだけ影響があるかを調べた。実験 3 では 4 因子からなるパワースペクトルから 1 つの因子の情報を除去して再合成した音声で、実験 4 では 2 因子・3 因子・4 因子からなるパワースペクトル因子から 2 因子以上を用いて再合成した音声でそれぞれ聴取実験を行った。結果は以下のようにまとめられる：

(i) 510–1480 Hz に設定された 6 つの臨界帯域においてに大きい因子負荷量をもつ因子のもつ情報は、4 因子からなるパワースペクトル因子の中で音声知覚に最も貢献する。

(ii) たとえこの 6 つの臨界帯域において大きい因子負荷量をもつ因子が与える情報が音声に含まれていても、合成に用いる因子の数が 2 個では高い明瞭度が得られない。

(iii) しかし、合成に用いる因子の数が 2 個の条件間で比較すると、この 6 つの臨界帯域において大きい因子負荷量をもつ因子が含まれている条件の方がより明瞭度が高くなる傾向が見られた。

第5章　総合考察

はじめに

本章では第 2 章から 4 章まで得られた実験の結果をまとめ、本論文の結論を提示する。次に、本結論が研究史の中でどのように位置づけられるのか、または先行研究にどのような新しい解釈を与えるのかについて論じる。最後に本論文では明らかにすることができなかった問題について触れ、今後の研究の展望を示す。

5.1　結果の概略

本研究で行った 4 つの実験についてそれぞれ結果を簡単にまとめる。第 2 章では、起点移動主成分分析によって取り出された、音声のパワースペクトルの分布とその時間変動を構成するパワースペクトル因子が音声を明瞭に知覚するのにどれだけ貢献するのかを聴取実験によって調べた (実験 1)。パワースペクトル因子から再合成した日本語の雑音駆動音声の明瞭度は、再合成の際に用いる因子の個数が 2 個から 3 個に増やしたときに急上昇することが分かった。また、 4 個の因子によって十分明瞭な (モーラ正答率 80%以上の) 音声が得られ、 6 個の因子によってほぼ完全に明瞭な (モーラ正答率 90%以上の) 音声が得られることが分かった。

第 3 章では、パワースペクトル因子を非負直交基底因子に変換し、再合成した雑音駆動音声の明瞭度を実験 1 と同じ方法で測定した (実験 2)。実験 1 と同じく、再合成に用いる因子の個数が 2 個から 3 個に変化したときに正答率が大きく上昇し、その後 4 因子以上で正答率の上昇が頭打ちになるという結果を得た。非負値化した 4 因子からなるパワースペクトル因子で再合成した音声の明瞭度はほぼ完全に明瞭であった。

第 4 章では、実験 3 として、4 因子からなるパワースペクトル因子のうちの 1 つの因子がもつスペクトルの時間変化の情報を除去したうえで雑音駆動音声を再合成した場合、明瞭度がどれだけ低下するかを調べた。4 因子すべてを用いて再合成した音声のモーラ正答率は 83.6%であったが、50–510 Hz の帯域に大きい因子負荷量をもつ因子が取り除かれたときは 64.1%に、510–1480 Hz の帯域に大きい因子負荷量をもつ因子が取り除かれたときは 36.3%に、1480–3700 Hz

の帯域に大きい因子負荷量をもつ因子が取り除かれたときは60.7%に、3700–6400 Hzの帯域に大きい因子負荷量をもつ因子が取り除かれたときは62%にそれぞれ低下した。510–1480 Hzに設定された6つの臨界帯域において大きい因子負荷量をもつ因子が4因子の中で最も重要であることが分かった。さらに実験4として、2因子からなるパワースペクトル因子、3因子からなるパワースペクトル因子、そして4因子からなるパワースペクトル因子のそれぞれの中から2因子以上を用いて、実験3のときと同様の方法で再合成した雑音駆動音声の明瞭度を調べた。その結果、基本的には再合成に用いる因子の個数が多いほど明瞭度は高くなるが、因子数が同数の条件であっても、この6つの臨界帯域において大きい因子負荷量をもつ因子が再合成に用いられていない場合は明瞭度が特に低くなるということが明らかになった。

5.2 結論

ここまでの考察を基に本節で本論文の結論を述べる。実験1および2によって、3個以上のパワースペクトル因子を用いることが明瞭な日本語の雑音駆動音声を再合成するための決定的な条件であることが示された。さらに、4因子で十分明瞭な音声が再合成できたことから、音声のパワースペクトルの時間変動に含まれる、音声知覚のための手がかりの主要な情報は最初の4個のパワースペクトル因子によって説明できると考えられる。実験3の結果は、4因子からなるパワースペクトル因子のうちの510–1480 Hzに設定された6つの臨界帯域において大きい因子負荷量をもつ因子は音声知覚においてその他の因子よりも重要であることを示している。イギリス英語音声における音素とパワースペクトル因子との関係を調べた先行研究では、この中程度の周波数帯域に大きい因子負荷量をもつ因子は音素の鳴音性の尺度と正の相関があることが指摘されている (Nakajima et al., 2017)。この鳴音性に関連する因子は、3因子からなるパワースペクトル因子の中にも現れるが、2因子からなるパワースペクトル因子の中にははっきりとは現れない。よって実験1および2において、パワースペクトル因子が2個から3個に増えたときに明瞭度が急激に上昇したのは、聴取者にとって鳴音性に関連する音響的特徴が音声を明瞭に知覚するために有用であるためとも解釈できる。しかしながら実験4では、鳴音性に関連する因子が再合成に用いられていたとしても因子数が2個では明瞭度は低いままであることがわかった。それでも、因子数が2個の場合、鳴音性に関連する因子を含まない条件よりも、その因子を含む条件の方が、多くの場合正答率が高かった。よってこのことは、鳴音性に関連する音響的特徴が音声知覚にとって重要であると示している。

実験1から4までの結果を総合すると、パワースペクトル因子から再合成した日本語の雑音

駆動音声が明瞭であるための条件は次のようにまとめられる：

- 鳴音性と正の相関があるパワースペクトル因子によって与えられる、臨界帯域 6 つ分に相当するおよそ 500–1500 Hz の周波数帯域のパワー変動の情報が音声に含まれていること。
- 上記の因子以外の因子によって与えられるパワースペクトルの変動の情報が、２因子分以上含まれていること。

以上の２つの条件が同時に満たされた音声は明瞭に知覚することが可能であると考えられる。本研究では、臨界帯域をZwicker and Terhardt (1980) に従って設定した。聴覚フィルタは本来決まった中心周波数を持つものではないと考えられているため (Moore, 2013)、中心周波数を変えた臨界帯域フィルタを使って音声の分析や合成を行うことができる。よって本研究の分析と実験では、510–1480 Hz に設定された 6 つの臨界帯域に音声知覚に重要な情報が含まれていると分かったが、この数値そのものが厳密であるとは言えない。実際には 510–1480 Hz に臨界帯域幅の半分程度の広さの誤差を含んだ周波数帯域に、音声知覚に重要な情報が含まれていると考えられる。

実験３と４の結果は、鳴音性と正の相関があるパワースペクトル因子が、再合成音声の明瞭度に最も重要であることを示した。では具体的にこの因子は音声知覚においてどのような役割をもつのだろうか。鳴音性と正の相関があるこの因子は、言語のリズムを構成すると先行研究において考察されている (Yamashita et al., 2013)。本研究においても、鳴音性と正の相関があるパワースペクトル因子の因子得点に、言語のリズムに対応づけられるようなリズム構造があるのかを調べてみる。４因子からなるパワースペクトル因子の因子得点について、文ごと分けたものから自己相関間関数を求め、その自己相関関数が正の値でピークとなる最小の時間間隔とその時の相関係数の値を記録した。聴取実験に用いられた原音声の話者が発話した 200 文すべてから得られたそれらの値を散布図にしたものが図 5.1 である。自己相関関数に正の値のピークができる時間間隔の分布の仕方が因子得点ごとに異なる特徴をもつようであるため、Yamashita et al. (2013) と同様に、0.2 s 間隔でヒストグラムを作成した (図 5.2)。鳴音性と正の相関がある因子 4-2 は 0.2–0.4 s の時間間隔において自己相関関数が正の値でピークとなる頻度が最も高いことが示された。この結果はYamashita et al. (2013) が鳴音性と正の相関がある因子が言語のリズムと関連すると考える根拠となった分析結果と一致している。よって本研究で得られた鳴音性と正の相関がある因子は言語のリズムを構成すると考えることができるであろう。日本語はモーラタイミング言語であり、音声の知覚はモーラが単位であると考えら

れている (Otake et al., 1993)。 鳴音性と正の相関があるパワースペクトル因子を用いずに再合成した音声を実験参加者が聴いたとき、明確なリズムを知覚することは困難であったのではないかと予測される。よって実験参加者は音声知覚の単位となるモーラが時間軸上でどのように配置されているのかがとらえられず、再合成音声を明瞭に聴きとることができなかったのであろう。鳴音性と正の相関があるパワースペクトル因子がもつ音声知覚上の役割は、言語のリズムを構成し音声を知覚する上での枠組みを与えることであると本論文では結論づける。

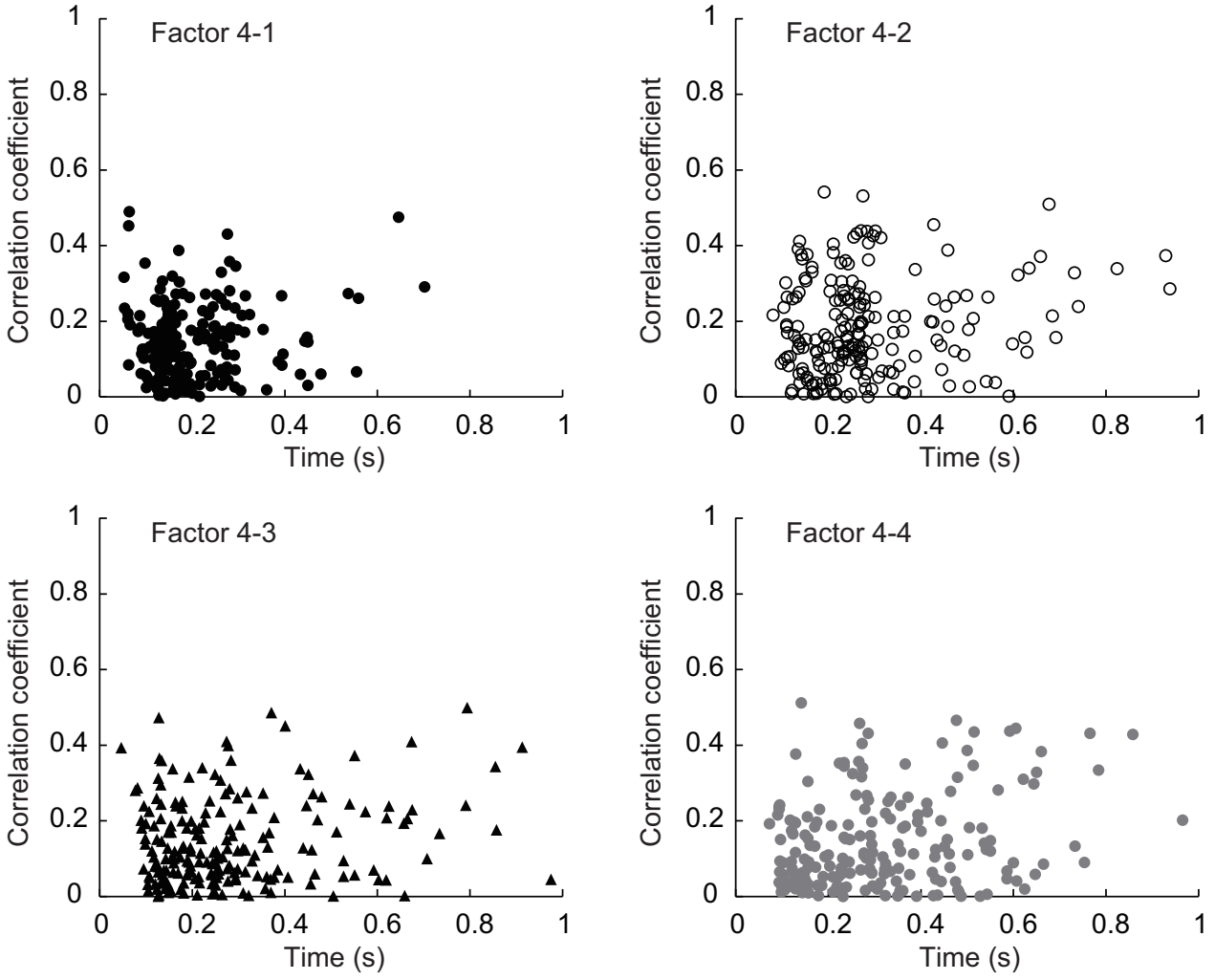

図 5.1 パワースペクトル因子の因子得点の自己相関関数を求め、自己相関関数が正の値でピークとなる最小の時間間隔を横軸に、相関係数を縦軸にした散布図。パワースペクトル因子は 4 因子分析から得られたもの。 1 名の男性日本語母語話者が発話した 200 文の音声 (聴取実験に用いられた刺激音の原音声を含む) から自己相関関数がそれぞれ求められた。

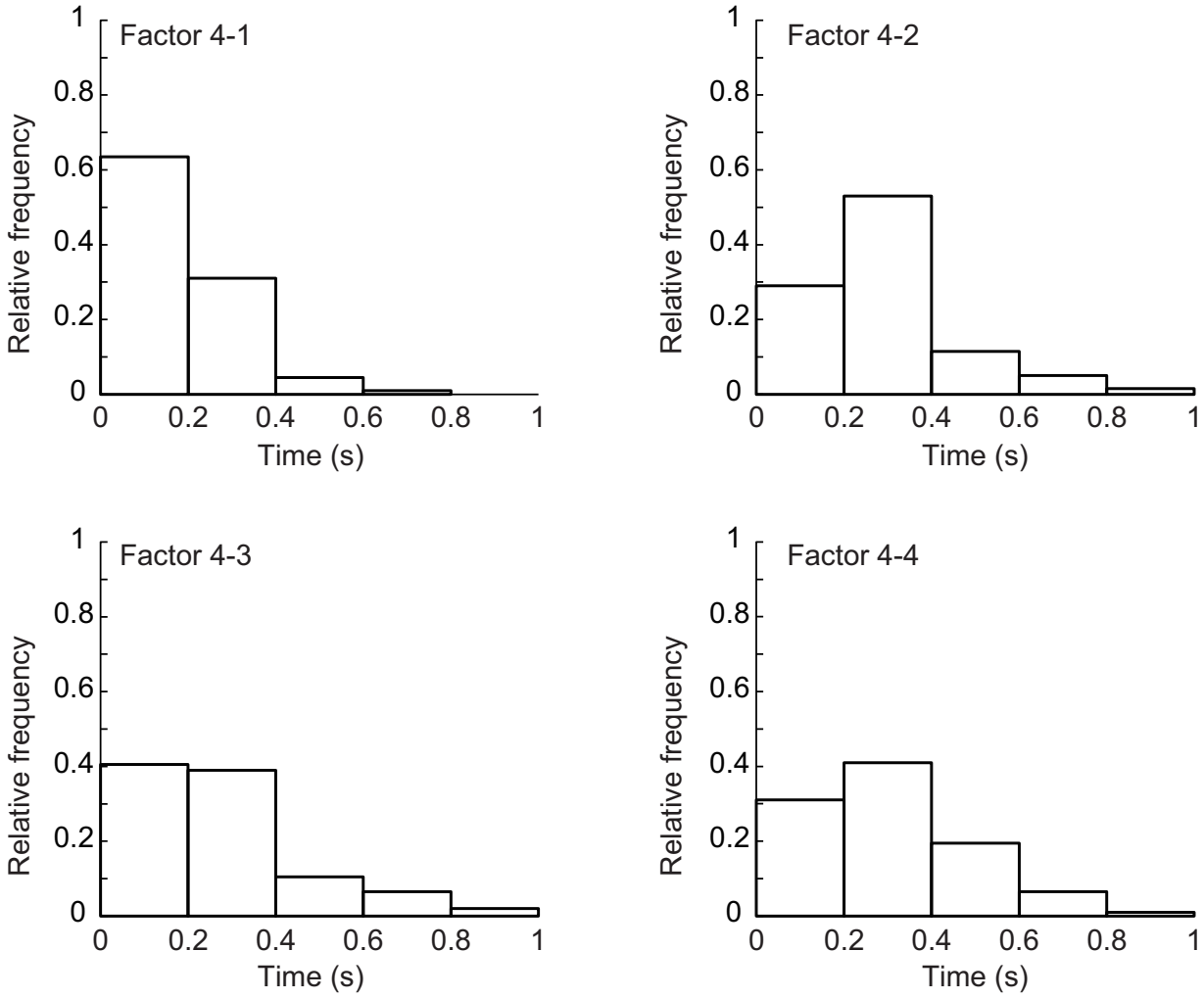

図 5.2 パワースペクトル因子の因子得点の自己相関関数を求め、自己相関関数が正の値でピークとなる最小の時間間隔をヒストグラムにした結果。パワースペクトル因子は 4 因子分析から得られたもの。1 名の男性日本語母語話者が発話した 200 文の音声 (聴取実験に用いられた刺激音の原音声を含む) から自己相関関数がそれぞれ求められた。

5.3 理論的位置づけ

本節では本論文の結論が先行研究の中でどのように位置づけられるかについて考察する。本研究の限界、および応用可能性についてもこの節で述べる。

5.3.1 知覚の手がかりの冗長性を示した先行研究との関係

従来のチャンネルボコーダ (Shannon et al., 1995; Dorman et al., 1997; Loizou et al., 1999; Sheldon et al., 2008; Souza & Rosen, 2009; Roberts et al., 2010; Ellermeier et al., 2015) を用いた音声の明瞭度を測る聴取実験では、4 帯域で正答率 60–90%程度の明瞭度が得られている。本論文の実験における刺激音は、パワースペクトル因子を用いて再合成した 20 の臨界帯域からなる雑音駆動音声であった。雑音駆動音声の帯域数を減らすという方法でなく、各帯域の変動をパワースペクトル因子を使って構成するという方法で音声の周波数情報が劣化しているが、3 または 4 因子で再合成した場合、因子によって音声の周波数帯域が 4 帯域に分けられたような音響的特徴をもっており、4 帯域の雑音駆動音声のようなものととらえることができる。再合成した雑音駆動音声はモーラ正答率 70–90%程度の明瞭度であり、先行研究の結果とよく対応している。本論文で得られた結論から先行研究の結果を解釈すれば、先行研究の 4 帯域のチャンネルボコーダ音声で高い明瞭度が得られていたのは、鳴音性を知覚する手がかりとなる約 500–1500 Hz の帯域のパワーの時間変動の情報が保存されており、かつ帯域数が 3 つ以上あるという条件を満たしていたからであると考えられる。実際に、4 帯域のチャンネルボコーダ音声を扱っている研究 (Shannon et al., 1995; Dorman et al., 1997; Loizou et al., 1999; Sheldon et al., 2008; Souza & Rosen, 2009; Ellermeier et al., 2015) においては、2 番目に低い帯域の下限の境界周波数がおよそ 400–700 Hz の間に、上限の境界周波数がおよそ 1000–1500 Hz の間にあり、鳴音性の情報が含まれると考えられる帯域に対応する。よってこの 2 番目に低い帯域が言語のリズムを構成することで聴取者には知覚の枠組みが形成されたと推察される。知覚の枠組みが形成されれば、各帯域の変動によって構成されるスペクトル全体の構造から音声を明瞭に知覚できるであろう。以上が本論文の結論から説明する、チャンネルボコーダ音声で明瞭に音声を知覚できる理由である。

狭帯域フィルタによって非常に狭い帯域のみに制限された音声でも、フィルタの中心周波数によっては高い明瞭度が得られたというWarren et al. (1995) の実験結果や、ある遮断周波数よりも高域もしくは低域いずれかのみに制限された音声でも遮断周波数によっては高い明瞭度が得られたというFrench and Steinberg (1947); Hirsh et al. (1954); Miller and Nicely (1955); Studebaker et al. (1987) の実験結果については、本論文で得られた結論だけから説明することは難しい。Warren et al. (1995) の実験では臨界帯域幅よりも狭い帯域幅 1 つ分の情報だけで明瞭な音声が得られており、また、鳴音性と正の相関のあるパワースペクトル因子とは関連の低い 1700 Hz 以上の帯域の情報だけである程度明瞭な音声が得られている (French & Steinberg,

1947; Hirsh et al., 1954; Miller & Nicely, 1955; Studebaker et al., 1987)。例に挙げた研究では、本研究では信号処理で除去して取り扱わなかった、スペクトルの微細構造の情報が知覚の手がかりとして利用されていたと予想される。スペクトルの微細構造はスペクトル全体の包絡構造が劣化したときの音声の明瞭度に影響を与えないとする報告はある (Ter Keurs et al., 1992, 1993) が、限られた帯域だけで音声を聴く場合には知覚の手がかりとなるのであろう。しかし、1500 Hz を中心とする狭帯域フィルタに通された音声が高い明瞭度であった (Warren et al., 1995) のは、鳴音性の情報がその帯域に含まれていたからであるという部分的な説明は可能だろう。

5.3.2　スペクトルの統計的分析から得られたものの解釈

Plomp とその同僚によって行われた一連の音声の統計的分析 (Plomp et al., 1967; Pols et al., 1973; Plomp, 1976, 2002) やUeda and Nakajima (2017) による音声の因子分析で得られた、音声のスペクトルを構成する因子の解釈については、Nakajima et al. (2017) やYamashita et al. (2013) の分析によって進められていた。本論文ではパワースペクトル因子の解釈を初めて知覚実験を用いて試みたものである。本研究によって鳴音性と正の相関がある因子は音声知覚においても最も重要な役割をもつことが示された。さらに、これらの先行研究とチャンネルボコーダ音声の研究とを結びつけることができた。

5.3.3　音声知覚の理論との関係

Blumstein、Stevens らの提唱する音響的不変性理論 (Blumstein & Stevens, 1979, 1980) では、例えばスペクトルに集約性の素性があるかどうか、高調音性の素性があるかどうかの 2 つの二項対立素性から 3 種類の閉鎖子音の分類ができるとしていた。閉鎖子音の閉鎖の解放後から 20–30 ms までの区間のスペクトルでは、軟口蓋閉鎖子音の/g/と/k/が 1 – 2 kHz あたりにスペクトルのピークができる集約性の素性がある。このような集約性の素性がない場合は、歯茎閉鎖子音の/d/、/t/か両唇閉鎖子音の/b/、/p/のいずれかの可能性がある。さらにスペクトルに高調音性の素性がある場合は歯茎閉鎖子音に、その素性がない場合は両唇閉鎖子音に分類することができる。ここで 3 因子以上のパワースペクトル因子があれば、スペクトルの集約性–拡散性、高調音性–低調音性の二項対立素性を表現することができそうである。まず集約性の素性は 510–1480 Hz に設定された 6 つの臨界帯域において大きい因子負荷量をもつ因子の因子得点を相対的に高くすることで実現できる。逆に拡散性の素性はこの因子の因子得点を

相対的に低くすることで実現でき、この因子よりも低い側の帯域に大きい因子負荷量をもつ因子の因子得点を相対的に高くすれば低調音性の素性を、高い側の帯域に大きい因子負荷量をもつ因子の因子得点を相対的に高くすれば高調音性の素性をそれぞれ実現することができる。実験 1 や 2 で、2 因子条件から 3 因子条件の間で明瞭度が急上昇したのも、音響的不変性理論における二項対立素性を判断できるようになったことと関連があると推察できる。実際に閉鎖子音発話時に各パワースペクトル因子の因子得点が上に説明したような変化をしているのかを観察することで確かめられるであろう。

音声知覚の運動理論 (Liberman et al., 1967) をはじめとするその他の代表的な音声知覚の理論が本研究とどのように結びつくかについて考察するためには、発声時の調音のデータが必要である。詳しくは次節の今後の展望で述べるが、パワースペクトル因子と調音器官の動きとの関係を調べることで運動理論との関係を調べることができるであろう。

5.3.4 本研究の限界

4 因子からなるパワースペクトル因子について鳴音性と正の相関がある因子が最も重要であることが分かったが、それ以外の因子を個々に意味づけをすることができなかった。これは手続き的な問題で、実験結果をモーラ正答率でしか評価できなかったことが原因の一つである。ある因子の情報が失われたときに音声の知覚にどのような質的な変化が起きたのかを調べるには、あるモーラもしくは音節がどのモーラ、音節として知覚されたかを記録したもの–異聴表–を作成する方法がある (e.g., Miller & Nicely, 1955)。 本研究では実験に用いた原音声に含まれている音節が十分とは言えなかったために、異聴表による結果の整理が行えなかった。十分な量でバランスの取れた音節を含むデータベースを基に実験を行う必要があるだろう。

4 因子からなるパワースペクトル因子のうちの、因子 4-4 や 3 因子からなるパワースペクトル因子の因子 3-3 はNakajima et al. (2017) の研究で、鳴音性と負の相関があることが報告されていた。しかし本研究においては、そのことと関連がありそうな実験結果は得られなかった。これについても、先に述べた実験結果の整理の問題を解決することで新しい見解が得られることが期待できる。

本研究の聴取実験では、実験参加者の語彙や言語能力に依存する要素を取り除いて議論を進めたかったため、回答の際に推測をできる限り避けるように教示したが、文を用いる以上完全に切り離して考えることはできない。どの程度文法的知識や語彙が結果に影響しているのかを検討することは今後の課題である。

5.3.5　本研究の応用可能性

本論文では、起点移動主成分分析を新しく提案した。得られた因子から音声を再合成する際に、無音を無音のままに再合成することができないという通常の主成分分析の問題を起点移動主成分分析では解決することができた。この分析法は非負値行列因子分解法 (Kameoka, 2016) と同様に、音声信号だけでなく非負の値からなる多変量データを次元圧縮する場合に、幅広い分野で適用可能である。

パワースペクトル因子を、聴覚補償の技術や、音声合成の技術、音声強調の技術などに利用することが可能であろう。例えば、因子によって分けられる周波数帯域を参考にして、人工内耳の周波数チャンネル選択の最適化に貢献できる。他にも、約 500–1500 Hz の帯域が目立つように音声信号を強調することで、雑音下での音声信号の聴き取りやすさの改善が期待できる。

パワースペクトル因子を非負化した非負直交基底因子の方では、４因子でほぼ完全に明瞭な雑音駆動音声を再合成することができていた。応用という観点では、非負直交基底因子の方が適用しやすいだろう。非負直交基底因子は因子負荷量が非負の値から構成されているため、因子ひとつひとつを実際の音とみなすことができる。よって非負直交基底因子から合成される音声は、因子に対応する音の部品を重ね合わせて作られたようなものである。直観的な理解が得やすいという点で、聴取実験においても利用価値が高いであろう。

5.4　今後の展望

本論文では、音声の周波数情報に含まれる知覚の手がかりに冗長性がどれだけあるのかを聴取実験で調べた研究と、音声のスペクトルを構成する主要な音響的特徴とはどのようなものであるか統計的手法による分析で調べた研究とを結びつけることができた。今後の展望としては以下に数例を挙げる。

本論文では、Ueda and Nakajima (2017) の見出した８言語にまたがって共通する３因子もしくは４因子からなるパワースペクトル因子によって作られる音声のスペクトルの特徴が、日本語の雑音駆動音声を明瞭に知覚するための重要な手がかりになるということが実験によって確かめられた。これにより、音声のスペクトルを構成するうちの言語に普遍的な特徴が音声知覚の重要な手がかりを担っていることが示唆された。このことを確かめるためには、実際に別の言語で同様の実験を行う必要がある。今回取り扱った日本語はモーラタイミング言語であり (Port et al., 1987)、日本語を知覚するときの単位もモーラ単位であった (Otake et al., 1993)。

リズム構造の類似した別の言語との比較、リズム構造の異なる別の言語との比較をそれぞれ行うことで、パワースペクトル因子が音声知覚において本当に普遍性があるのか、そしてリズム構造の違いが音声知覚に与える影響について明らかにすることができるだろう。

音声の鳴音性の変化が言語のリズムを作る。その鳴音性と正の相関があるパワースペクトル因子によって与えられる情報が音声の明瞭な知覚に重要であることが本論文で示された。リズムは時間の概念と切り離せないものである。 よって音声がもつ時間方向の情報と知覚との関係について調べることも重要である。本論文ではパワースペクトル因子の因子負荷量に注目してきたが、時間方向の情報を担う因子得点の変化を操作して再合成した音声を用いた実験を行うことができる。Xu and Pfingst (2008) のような時間情報と周波数情報の相互関係を調べる研究をパワースペクトル因子の観点から考察できるようになるだろう。

本論文で扱ったパワースペクトル因子が音声を用いた聴覚コミュニケーションにおいて本質的な役割を担っているのであれば、音声生成の際の調音器官の運動もしくはその運動指令や、音声知覚の際の神経活動にパワースペクトル因子と対応するものを見つけることができるかもしれない。パワースペクトル因子は、４つ程度の少ない因子得点で音声の知覚にとって重要な特徴を表現することから、音声生成や音声知覚における神経活動との対応関係を比較的容易に記述することが期待できる。音声生成や音声知覚における神経活動とパワースペクトル因子との関係が明らかにされれば、運動理論をはじめとする音声知覚の理論を検証するのに貢献できるだろう。

第５章のまとめ

本章では、第２章から４章までに行った実験の結果を比較することで、パワースペクトル因子がもつ音声知覚上の役割についての総合考察を行った。考察は以下のようにまとめられる：

(i) 鳴音性を知覚する手がかりとなる、臨界帯域６つ分に相当する約 500–1500 Hz の周波数帯域のパワー変動の情報が含まれ、加えて２つ以上の帯域のパワー変動の情報が含まれていることが、音声を明瞭に知覚するための条件である。

(ii) 510–1480 Hz に設定された６つの臨界帯域において大きい因子負荷量をもつパワースペクトル因子が言語のリズムを構成することにより、聴取者に音声知覚の枠組みが与えられ、それを基に音声の内容が知覚されると考えられる。

(iii) チャンネルボコーダ音声が４～６帯域程度で相当に明瞭であることは、上で述べた音声を明瞭に知覚するための条件を満たしていたからであると説明することができる。

(iv) 実験における手続き的な問題を解消し、実験参加者の回答の質的変化を調べることが今後の課題である。

文献

Blumstein, S. E., Isaacs, E., & Mertus, J. (1982). The role of the gross spectral shape as a perceptual cue to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, *72*, 43–50.

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America*, *66*, 1001–1017.

Blumstein, S. E., & Stevens, K. N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America*, *67*, 648–662.

Boersma, P., & Weenink, D. (2014). *Praat: doing phonetics by computer [computer program].* (Version 5.4.04, retrieved 28 December 2014 from http://www.praat.org/)

Cooper, F. S., Liberman, A. M., & Borst, J. M. (1951). The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Sciences*, *37*, 318–325.

Cutler, A. (1994). The perception of rhythm in language. *Cognition*, *50*, 79–81.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, *25*, 385–400.

de Saussure, F. (1959). Course in general linguistics (Baskin, W. Trans.). *New York: Philosophical Library.[JL]*. (Original work published 1916.)

Delattre, P., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel color; observations on one-and two-formant vowels synthesized from spectrographic patterns. *Word*, *8*, 195–210.

Denes, P. B., & Pinson, E. N. (1993). *The Speech Chain* (2nd ed.). New York: W.H. Freeman and Co.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Reviews of Psychology*, *55*, 149–179.

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *Journal of the Acoustical Society of America*, *102*, 2403–2411.

Dudley, H. (1939). Remaking speech. *The Journal of the Acoustical Society of America*, *11*, 169–177.

Ellermeier, W., Kattner, F., Ueda, K., Doumoto, K., & Nakajima, Y. (2015). Memory disruption by irrelevant noise-vocoded speech: Effects of native language and the number

of frequency bands. *Journal of the Acoustical Society of America*, *138*, 1561–1569.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

Fastl, H., & Zwicker, E. (2006). *Psychoacoustics: Facts and Models* (3rd ed.). Heidelberg: Springer.

Fletcher, H. (1940). Auditory patterns. *Reviews of Modern Physics*, *12*, 47-65.

Fowler, C. A. (1991). Auditory perception is not special: We see the world, we feel the world, we hear the world. *Journal of the Acoustical Society of America*, *89*, 2910–2915.

Fowler, C. A., & Rosenblum, L. D. (1990). Duplex perception: a comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 742–754.

French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *Journal of the Acoustical society of America*, *19*, 90–119.

Galves, A., Garcia, J., Duarte, D., & Galves, C. (2002). Sonority as a basis for rhythmic class discrimination. In *Speech Prosody 2002, International Conference.*

Greenwood, D. D. (1990). A cochlear frequency-position function for several species—29 years later. *Journal of the Acoustical Society of America*, *87*, 2592–2605.

Handel, S. (1989). *Listening: An Introduction to the Perception of Auditory Events*. Cambridge, MA: MIT Press.

Harris, J. (1994). *English Sound Structure*. Oxford, UK: Blackwell.

Hirsh, I. J., Reynolds, E. G., & Joseph, M. (1954). Intelligibility of different speech materials. *Journal of the Acoustical Society of America*, *26*, 530–538.

Jolliffe, I. (2002). *Principal Component Analysis*. New York: Springer.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187–200.

Kameoka, H. (2016). Non-negative matrix factorization and its variants for audio signal processing. In S. Toshio (Ed.), *Applied Matrix and Tensor Variate Data Analysis* (pp. 23–50). Springer.

Kluender, K. R., Diehl, R. L., Killeen, P. R., et al. (1987). Japanese quail can learn phonetic categories. *Science*, *237*, 1195–1197.

Ladefoged, P., & Johnson, K. (2011). *A Course in Phonetics* (6th ed.). Boston, MA: Wadsworth, Cengage learning.

Li, K.-P., Hughes, G., & House, A. (1969). Correlation characteristics and dimensionality of speech spectra. *Journal of the Acoustical Society of America*, *46*, 1019–1025.

Liberman, A. M. (1957). Some results of research on speech perception. *Journal of the Acoustical Society of America*, *29*, 117–123.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*, 358–368.

Loizou, P. C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *Journal of the Acoustical Society of America*, *106*, 2097–2103.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *Journal of the Acoustical Society of America*, *27*, 338–352.

Moore, B. C. J. (2013). *An Introduction to the Psychology of Hearing (6th)*. London: Academic Press.

Nakajima, Y., Ueda, K., Fujimaru, S., Motomura, H., & Ohsaka, Y. (2017). English phonology and an acoustic language universal. *Scientific Reports*, *7:46049*.

NTT-AT. (2002). *Multilingual speech database 2002 (NTT Advanced Technology Corporation, Tokyo, Japan).* (`http://www.ntt-at.com/product/speech2002/`(Last viewed December 1, 2015))

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, *32*, 258–278.

Patterson, R. D. (1974). Auditory filter shape. *Journal of the Acoustical Society of America*, *55*, 802–809.

Pisoni, D. B. (1985). Speech perception: Some new directions in research and theory. *Journal of the Acoustical Society of America*, *78*, 381–388.

Plack, C. J. (2014). *The Sense of Hearing* (2nd ed.). New York: Psychology Press.

Plomp, R. (1976). *Aspects of Tone Sensation: A Psychophysical Study*. London: Academic Press.

Plomp, R. (2002). *The Intelligent Ear: On the Nature of Sound Perception*. Mahwah, New Jersey: Lawrence Erlbaum.

Plomp, R., Pols, L. C. W., & Geer, J. P. van de. (1967). Dimensional analysis of vowel spectra. *Journal of the Acoustical Society of America*, *41*, 707-712.

Pols, L. C. W., Tromp, H. R. C., & Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers. *Journal of the Acoustical Society of America*, *53*, 1093-1101.

Port, R. F., Dalby, J., & O ' Dell, M. (1987). Evidence for mora timing in japanese. *Journal of the Acoustical Society of America*, *81*, 1574–1585.

Potter, R. K. (1945). Visible patterns of sound. *Science*, 463-470.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice Hall.

Rahilly, J. (2016). Sonority in natural language: A review. In M. J. Ball & N. Müller (Eds.), *Challenging Sonority: Cross-Linguistic Evidence.* South Yorkshire, UK: Equinox Publishing Ltd.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*, 265-292.

Rand, T. C. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, *55*, 678–680.

Roberts, B., Summers, R. J., & Bailey, P. J. (2010). The intelligibility of noise-vocoded speech: spectral information available from across-channel comparison of amplitude envelopes.

Proceedings of the Royal Society of London B: Biological Sciences, rspb20101554.

Samuel, A. G. (2011). Speech perception. *Annual Reviews of Psychology*, *62*, 49–72.

Schneider, B. A., Morrongiello, B. A., & Trehub, S. E. (1990). Size of critical band in infants, children, and adults. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 642.

Schnupp, J., Nelken, I., & King, A. (2011). *Auditory neuroscience: Making sense of sound*. MIT press.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., MüLler, C., et al. (2013). Paralinguistics in speech and language—State-of-the-art and the challenge. *Computer Speech & Language*, *27*, 4–39.

Selkirk, E. (1984). On the major class features and syllable theory. In M. Aronoff & R. T. Oehrle (Eds.), *Language Sound Structure: Studies in Phonology Presented to Morris Halle by His Teacher and Students* (pp. 107–136). Cambridge, MA: MIT.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303-304.

Sheldon, S., Pichora-Fuller, M. K., & Schneider, B. A. (2008). Effect of age, presentation method, and learning on identification of noise-vocoded words. *Journal of the Acoustical Society of America*, *123*, 476–488.

Souza, P., & Rosen, S. (2009). Effects of envelope bandwidth on the intelligibility of sine-and noise-vocoded speech. *Journal of the Acoustical Society of America*, *126*, 792-805.

Spencer, A. (1996). *Phonology: Theory and Description*. Oxford: Blackwell.

Studebaker, G. A., Pavlovic, C. V., & Sherbecoe, R. L. (1987). A frequency importance function for continuous discourse. *Journal of the Acoustical Society of America*, *81*, 1130–1138.

Ter Keurs, M., Festen, J. M., & Plomp, R. (1992). Effect of spectral envelope smearing on speech reception I. *Journal of the Acoustical Society of America*, *91*, 2872–2880.

Ter Keurs, M., Festen, J. M., & Plomp, R. (1993). Effect of spectral envelope smearing on speech reception Ⅱ. *Journal of the Acoustical Society of America*, *93*, 1547–1552.

Ueda, K., & Nakajima, Y. (2017). An acoustic key to eight languages/dialects: Factor analyses of critical-band-filtered speech. *Scientific Reports*, *7:42468*.

Unoki, M., Irino, T., Glasberg, B., Moore, B. C., & Patterson, R. D. (2006). Comparison of the roex and gammachirp filters as representations of the auditory filter. *Journal of the Acoustical Society of America*, *120*, 1474–1492.

Walley, A. C., & Carrell, T. D. (1983). Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants. *Journal of the Acoustical Society of America*, *73*, 1011–1022.

Warren, R. M., Riener, K. R., Bashford, J. A., & Brubaker, B. S. (1995). Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits. *Perception & Psychophysics*, *57*, 175–182.

Xu, L., & Pfingst, B. E. (2008). Spectral and temporal cues for speech recognition: Implica-

tions for auditory prostheses. *Hearing Research*, *242*, 132–140.

Yamashita, Y., Nakajima, Y., Ueda, K., Shimada, Y., Hirsh, D., Seno, T., et al. (2013). Acoustic analyses of speech sounds and rhythms in Japanese-and English-learning infants. *Frontiers in Psychology*, *4*, 57.

Zahorian, S. A., & Jagharghi, A. J. (1993). Spectral-shape features versus formants as acoustic correlates for vowels. *Journal of the Acoustical Society of America*, *94*, 1966–1982.

Zahorian, S. A., & Rothenberg, M. (1981). Principal-components analysis for low-redundancy encoding of speech spectra. *Journal of the Acoustical Society of America*, *69*, 832–845.

Zwicker, E., & Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America*, *68*, 1523-1525.

謝辞

本論文は筆者が九州大学大学院芸術工学専攻博士後期課程に在籍中の研究成果をまとめたものである。中島祥好先生には学部４年生の時から始まる６年間、研究指導を賜った。研究の相談にはいつも乗ってくださり、その都度的確なご助言をいただいた。上田和夫先生、Gerard B. Remijn 先生、白石君男先生、そして山下友子先生には日ごろからゼミ等で本研究についての報告を聞いていただき、その際には数多くの有益なご意見を下さった。また上田和夫先生、Gerard B. Remijn 先生には副査として本論文を細部に至るまで丁寧に読んでいただき、論文完成のための的確なご指導を賜った。伊藤裕之先生、妹尾武治先生には他領域の立場から本論文に対して貴重なご意見を下さった。鏑木時彦先生には、本論文で行った音声信号処理における理論的背景についてご指導をいただいた。亀岡弘和博士には第３章の非負直交基底因子の導出について多くのご助言をいただいた。藤井芳孝博士には、音声分析のプログラムを整理する際にお手伝いいただいた。Mark A. Elliott 先生には、本論文の執筆の際に、非常に有益なご助言をいただいた。第３章は中尾貫志さんと共同で研究を進めた部分である。実験の計画から実施に至るまで円滑に進めることができたのも、中尾氏のおかげであった。中尾氏には、本論文の執筆の際にも内容に目を通していただいた。また、小野明日香さんには第２章の実験１、梅本晟弥さんには第４章の実験４で、それぞれ実施にご協力いただいた。澤井賢一博士および森本智志博士には、本研究の第２章の部分を投稿論文としてまとめた際に多くのご助言と執筆に対する励ましのお言葉をいただいた。中島研究室、上田研究室、Remijn 研究室、白石研究室、山下研究室、伊藤研究室、妹尾研究室等の皆さんには、研究室ゼミ、サマーキャンプ等で本研究について討論する機会を多くいただき、本論文執筆の刺激となった。皆様に深く感謝申し上げる。

本研究の内容に関して、日本音響学会聴覚研究会、国際精神物理学会、国際心理学会議などの学会、研究会において、非常に有益な討論や情報交換の場を与えていただいたことについて、関係者の皆様に感謝する。特に、たびたび研究発表の練習に付き合っていただき、数多くのご助言と励ましの言葉を頂いた蓮尾絵美博士に感謝申し上げる。

ここには名前を記載することはできないが、数多くの方に、聴取実験に参加していただい

た。皆様に厚く御礼申し上げる。

伊藤浩史先生、長津結一郎先生、江頭優佳先生、村谷つかささん、岡田昌大さんとは毎朝図書館に集まり、それぞれの執筆活動を行うことで励ましあうことができた。本論文の大部分がこの執筆活動中に書かれたものである。また原稿に目を通していただき、有益なご助言も数多くいただいた。この集まりなくして本論文が完成することはなかった。皆様に心より感謝申し上げる。

最後に、日ごろから支えてくれ、精神的な励みとなってくれた家族に心より感謝申し上げる。

付記

- 関連論文および学会発表
- 教示文
- 実験同意書

関連論文および学会発表

本研究の内容を学術雑誌および学会にて発表したものを以下に列挙する。

学術雑誌

- Kishida, T., Nakajima, Y., Ueda, K., & Remijn, G. B. (2016). Three factors are critical in order to synthesize intelligible noise-vocoded Japanese speech. *Frontiers in psychology*, 7:517.

学会発表

- 岸田拓也，中島祥好，上田和夫，Gerard B. Remijn，中尾貫志. (2017). 臨界帯域パワー変動因子を用いた雑音駆動音声の合成—因子の除去が明瞭度に与える効果—. 日本音響学会聴覚研究会，京都，2017 年 3 月.

- 中尾貫志，岸田拓也，中島祥好，亀岡弘和. (2017). 臨界帯域パワー変動から抽出された非負基底の音声再合成における妥当性. 日本音響学会聴覚研究会，京都，2017 年 3 月.

- Kishida, T., Nakajima, Y., Ueda, K., & Remijn G. B. (2016). Effects of factor elimination on intelligibility of noise-vocoded Japanese speech. *The 31st International Congress of Psychology.* Yokohama, Japan, September 2016.

- Kishida, T., Nakajima, Y., & Nakao, K. (2016). Origin shifted principal component analysis: A method suitable for reconstructing non-negative data. *The 31st International Congress of Psychology.* Yokohama, Japan, September 2016.

- Nakao, K., Nakajima, Y., & Kishida, T. (2016). Perceptual validity and analytical advantages of non-negative bases extracted from factor analyses of Japanese speech. *The 31st International Congress of Psychology.* Yokohama, Japan, September 2016.

- Kishida, T., Nakajima, Y., Ueda, K., & Remijn G. B. (2015). A critical number of power-fluctuation factors needed for Japanese noise-vocoded speech perception. *The*

31st Annual Meeting of the International Society for Psychophysics. Quebec, Canada, August 2015.

- 岸田拓也，中島祥好，上田和夫，Gerard B. Remijn. (2015). 主成分分析によって縮約し，再合成した日本語音声の明瞭度について. ヒューマン情報処理研究会，福岡，2015 年 7 月.

- Kishida, T., Nakajima, Y., Ueda, K., & Remijn, G. B. (2015). Perceptual roles of power-fluctuation factors in Japanese speech. 第 48 回知覚コロキウム，大分，2015 年 3 月.

- Kishida, T., Nakajima, Y., Ueda, K., & Remijn, G. B. (2014). Perceptual roles of power-fluctuation factors in speech perception: A new method of factor analysis. 日本音響学会聴覚研究会，福岡，2014 年 12 月.

- Kishida, T., Nakajima, Y., Ueda, K., & Remijn G. B. (2014). Perceptual roles of power fluctuation factors in speech. *The 13th International Conference on Music Perception and Cognition and the 5th Conference for the Asian-Pacific Society for Cognitive Sciences of Music.* Seoul, South Korea, August 2014.

- Kishida, T., Nakajima, Y., & Ueda, K. (2013). Effects of elimination of power-fluctuation factors from critical-band noise-vocoded speech. *The 29th Annual Meeting of the International Society for Psychophysics.* Freiburg i. Br., Germany, October 2013.

- 岸田拓也，中島祥好，上田和夫. (2013). 音声の因子分析と再合成〜因子の除去が明瞭性に与える効果〜. 日本音響学会聴覚研究会，宮城，2013 年 8 月.

実験１・２の教示文

本日は実験に参加いただきありがとうございます。今から、日本語音声の聞き取りに関する実験を行います。実験が開始しましたら、まずヘッドフォンを装着してください。音声は全てヘッドフォンから両耳に呈示されます。図 1.3 のような画面がパソコンのディスプレイに表示されますので、“ Play ”ボタンをクリックして下さい。すると、音声が再生されます。

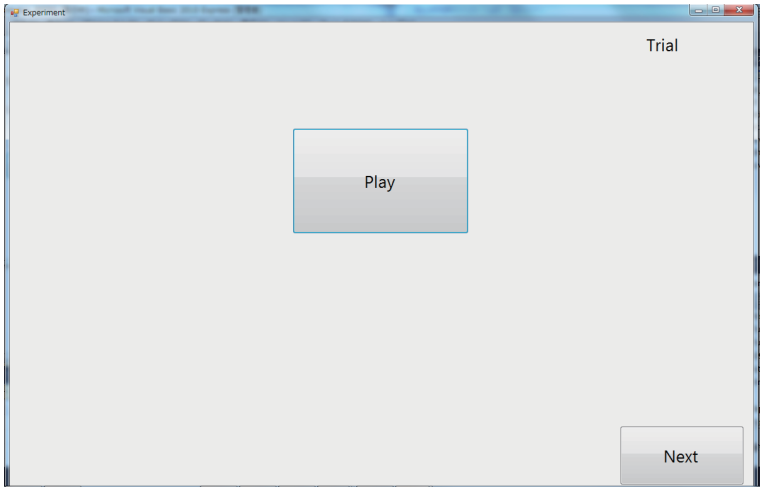

図 1.3 実験画面

回答方法について　実験での回答方法を説明します。“ Play ”ボタンをクリックすると、日本語音声が聞こえてきます。 音声は 3 回繰り返されます。音声の再生終了後、聞こえた音声をコンピュータにひらがなで入力して下さい。その際に、“ は ”と“ へ ”については、助詞として使っているのか、そうでないのかを、その文字の後に続けて、アルファベットを書き加えて区別してください。例えば“ わたしはＷ”、“ はＨだし ”、“ がっこうへＥ”、“ こうへＨいな ”のように書き加えて下さい。音声が聞き取りにくかったとしても, できるだけ推測はせず, 聞こえたとおりに回答するようにつとめてください。もし音声の一部分しか聞き取れなかったとしても、音声のどの部分が聞き取れたかがわかるように、聞き取れなかった部分については、“･” と入力してください。“･” の入力数は、文章全体に対して、聞き取れなかった部分の相対量を大まかに表す程度の数で構いませんが、聞き取れなかった文字数が分かる場合は、その数だけ、“･” を入力してください。回答は音声の呈示が終わってから始めてください。1 つの試行につき、再生は一度きりしか出来ません。回答が終わりましたら“ Next ”ボタンをクリックしてください。“ Play ”ボタンをクリックしてから、回答を入力し、“ Next ”ボタンをクリック

するまでが 1 つの試行になります。“ Next ”ボタンをクリックすると,、次の試行に進みます。“ Play ”ボタンをクリックすると次の音声が呈示されますので、 同様に回答を行ってください。回答は 1 試行につき、1 行ずつ入力し、1 行分間隔を空けて、次の回答を行ってください。

実験ブロックについて　本実験では 9 個の問題からなる練習ブロックの後、休憩をはさんで、1 ブロックにつき 16 個の問題に回答してもらう本試行ブロックをを計 3 ブロック行ってもらいます.1 つのブロックが終了しましたら回答方法に間違いがないかを確認します。ここで休憩をとりますが、 続けて行うこともできます。続けて行いたい場合は、実験者に申し出てください。

注意事項　途中で疲れたり、気分が悪くなったりして休みたい時は、いつでもかまいませんので実験者に申し出てください。休んでも実験全体に支障が出ることはありません。また、実験中は足元の機材に触れないようにお願いします。音が聞こえている間、ヘッドフォンに触れないようにしてください。音声の再生は 1 度きりで、やり直しはききませんので、音声呈示中はその音声に集中してください。

この実験は個人の能力を調べることを目的とはしていません。間違った回答をしているのではないか、などと深く考え込んだりせずに、聴いたまま、感じたままに回答をしてください。この他に分からないことがあれば、いつでも実験者に尋ねてください. それではよろしくお願いします。

実験３の教示文

本日は実験に参加していただき、ありがとうございます。これから、日本語音声の聴き取りに関する実験を行います。この実験では、これから説明する手順に従って、人工的に処理をした音声をヘッドフォンから聴き、聴こえた内容を聴こえた通りにコンピュータに入力するという試行を、複数回行っていただきます。

操作・回答方法

実験が開始しましたら、まずヘッドフォンを装着してください。図 1.4 のような画面が正面のディスプレイに表示されますので、“ Play ”ボタンをクリックして下さい。すると、一文の日本語音声がヘッドフォンの両耳から 3 回繰り返して再生されます。このときに、定常的な雑音が全体を覆うように流れる場合がありますが、再生機器の故障等ではありません。音声として聴こえる部分のみに注意を向けて聴いてください。

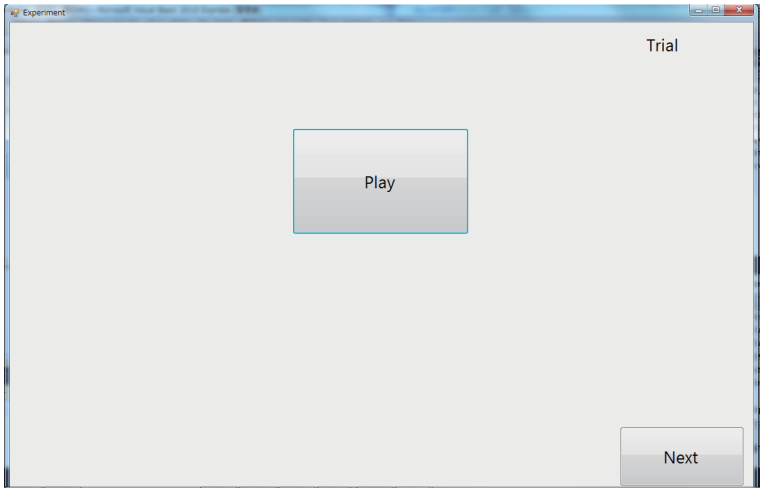

図 1.4 実験画面

音声の再生が終了しましたら、聴こえた文の内容を、机上のノートパソコン (回答用コンピュータ) にひらがなで入力して下さい。その際に、“ は ”と“ へ ”については、助詞として使っているのか、そうでないのかを、その文字の後に続けて、アルファベットを書き加えることで区別してください。例えば“ わたしはＷ ”、“ はＨだし ”、“ がっこうへＥ ”、“ こうへＨいな ”のように書き加えて下さい。音声が聴き取りにくかったとしても, できるだけ推測はせず, 聴こえたとおりに回答するようにつとめてください。文の一部分しか聴き取れなかった場合は、その文の中のどの部分が聴き取れたかがわかるように、聴き取れなかった部分について

は、“·”と入力してください。“·”の入力数は、文全体に対して、聴き取れなかった部分の相対量を大まかに表す程度の数で構いませんが、聴き取れなかった文字数が分かる場合は、その数だけ、“·”を入力してください[1]。回答は音声の再生が終わってから始めてください。1つの試行につき、再生は一度きりしか出来ません。回答が終わりましたら“Next”ボタンをクリックしてください。“Play”ボタンをクリックしてから、回答を入力し、“Next”ボタンをクリックするまでが1つの試行になります。“Next”ボタンをクリックすると、次の試行に進みます。“Play”ボタンをクリックすると次の音声が再生されますので、 同様に回答を行ってください。

回答用コンピュータの画面に表示されている、入力例をご覧ください。この入力例のように、回答は1試行につき1行ずつ入力し、1行分の間隔を空けて、次の回答を行ってください。

実験ブロックについて

本実験では10個の試行からなる練習ブロックの後, 休憩をはさんで、1ブロックにつき13〜14個の試行からなる本ブロックを2ブロック行ってもらいます。1つのブロックが終了しましたら回答方法に間違いがないかを確認しますので、実験ブースから出てきて、実験者に知らせてください。ここで休憩をとりますが、 続けて行うこともできます。続けて行いたい場合は、実験者に申し出てください。

注意事項

途中で疲れたり、気分が悪くなったりして休みたい時は、いつでもかまいませんので実験者に申し出てください。休んでも実験全体に支障が出ることはありません。また、実験中は足元の機材に触れないようにお願いします。音が聴こえている間は、ヘッドフォンに触れないようにしてください。1つの試行につき、“Play”ボタンをクリックできるのは1度きりです (音声の再生は3回繰り返されます)。やり直しはききませんので、音声再生中はその音声に集中してください。

この実験は個人の能力を調べることを目的とはしていません。間違った回答をしているのではないか、などと深く考え込んだりせずに、聴こえたままに回答をしてください。この他に分からないことがあれば、いつでも実験者に尋ねてください. それではよろしくお願いします。

[1]長音 (ー)、撥音 (ん)、促音 (っ) は単独で 1 文字と数えますが、拗音 (ゃ、ゅ、ょ) は直前の文字と合わせて 1 文字と数えます。例えば、“ちょきん”、“せーたー”、“がっこう”の文字数はそれぞれ 3,4,4 となります。

実験４の教示文

本日は実験に参加していただき、ありがとうございます。これから、日本語音声の聴き取りに関する実験を行います。この実験では、これから説明する手順に従って、人工的に処理をした音声をヘッドフォンから聴き、聴こえた内容を聴こえた通りにコンピュータに入力するという試行を、複数回行っていただきます。

操作・回答方法

実験が開始しましたら、まずヘッドフォンを装着してください。図 1.5 のような画面が正面のディスプレイに表示されますので、“ Play ”ボタンをクリックして下さい。すると、一文の日本語音声がヘッドフォンの両耳から 3 回繰り返して再生されます。このときに、定常的な雑音が全体を覆うように流れる場合がありますが、再生機器の故障等ではありません。音声として聴こえる部分のみに注意を向けて聴いてください。

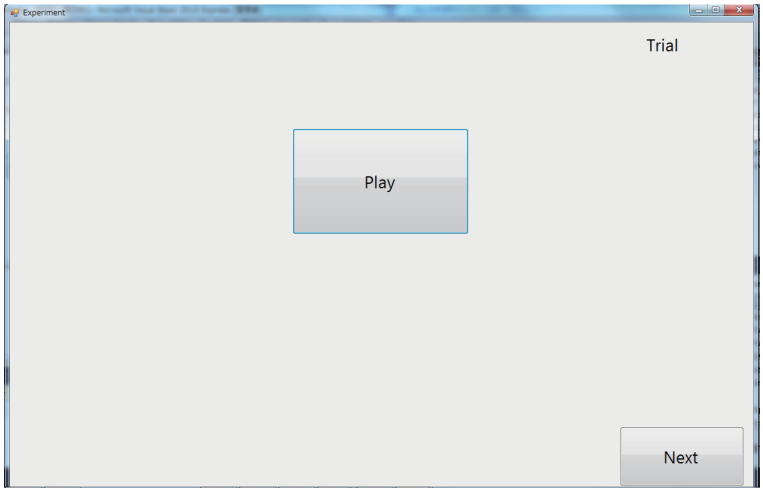

図 1.5 実験画面

音声の再生が終了しましたら、聴こえた文の内容を、机上のノートパソコン (回答用コンピュータ) にひらがなで入力して下さい。その際に、“ は ”と“ へ ”については、助詞として使っているのか、そうでないのかを、その文字の後に続けて、アルファベットを書き加えることで区別してください。例えば“ わたしはＷ ”、“ はＨだし ”、“ がっこうへＥ ”、“ こうへＨいな ”のように書き加えて下さい。音声が聴き取りにくかったとしても, できるだけ推測はせず, 聴こえたとおりに回答するようにつとめてください。文の一部分しか聴き取れなかった場合は、その文の中のどの部分が聴き取れたかがわかるように、聴き取れなかった部分について

は、“·”と入力してください。“·”の入力数は、文全体に対して、聴き取れなかった部分の相対量を大まかに表す程度の数で構いませんが、聴き取れなかった文字数が分かる場合は、その数だけ、“·”を入力してください[2]。回答は音声の再生が終わってから始めてください。1つの試行につき、再生は一度きりしか出来ません。回答が終わりましたら“Next”ボタンをクリックしてください。“Play”ボタンをクリックしてから、回答を入力し、“Next”ボタンをクリックするまでが1つの試行になります。“Next”ボタンをクリックすると、次の試行に進みます。“Play”ボタンをクリックすると次の音声が再生されますので、 同様に回答を行ってください。

回答用コンピュータの画面に表示されている、入力例をご覧ください。この入力例のように、回答は1試行につき1行ずつ入力し、1行分の間隔を空けて、次の回答を行ってください。

実験ブロックについて

本実験では16個の試行からなる練習ブロックの後, 休憩をはさんで、1ブロックにつき18試行以上からなる本ブロックを4ブロック行ってもらいます。1つのブロックが終了しましたら回答方法に間違いがないかを確認しますので、実験ブースから出てきて、実験者に知らせてください。ここで休憩をとりますが、 続けて行うこともできます。続けて行いたい場合は、実験者に申し出てください。

注意事項

途中で疲れたり、気分が悪くなったりして休みたい時は、いつでもかまいませんので実験者に申し出てください。休んでも実験全体に支障が出ることはありません。また、実験中は足元の機材に触れないようにお願いします。音が聴こえている間は、ヘッドフォンに触れないようにしてください。1つの試行につき、“Play”ボタンをクリックできるのは1度きりです (音声の再生は3回繰り返されます)。やり直しはききませんので、音声再生中はその音声に集中してください。

この実験は個人の能力を調べることを目的とはしていません。間違った回答をしているのではないか，などと深く考え込んだりせずに，聴こえたままに回答をしてください。この他に分からないことがあれば、いつでも実験者に尋ねてください。それではよろしくお願いします。

[2]長音 (ー)、撥音 (ん)、促音 (っ) は単独で1文字と数えますが、拗音 (ゃ、ゅ、ょ) は直前の文字と合わせて1文字と数えます。例えば、“ちょきん”、“せーたー”、“がっこう”の文字数はそれぞれ3,4,4となります。

No. _____

実験参加に関する同意書

実験責任者 中島 祥好　　九州大学大学院芸術工学研究院デザイン人間科学部門
実験責任者 上田 和夫　　九州大学大学院芸術工学研究院デザイン人間科学部門
実験責任者 Gerard B. Remijn　九州大学大学院芸術工学研究院デザイン人間科学部門

実験者 _____ _____

1. これからご協力いただきますのは，知覚および認知に関する行動実験および非侵襲の生理実験の一部分，またはすべてです。この実験で得られたデータは「人間の知覚および認知が働く仕組み」について調べる研究に役立てます。実験者は，実験手続き，所要時間の見通し，実験回数について説明を行います。
2. この観察および実験には何ら危険は伴いません。
3. この同意書に署名された後でも，実験を継続できなくなったり，参加の意志がなくなった場合は，いつでも参加を中止していただくことができます。このことによって，参加者の方は，いかなる不利益もこうむることはありません。
4. この実験の結果は，参加者番号を使ったデータとして，個人名を特定できない形で学位論文，学会，学術論文など，学術の分野で発表されます。実験責任者が関係する他の研究機関においても，関連する研究目的で，同じデータを利用させていただく場合があります。また，観察，実験結果を，報道機関等を通じて公表する場合があります。ご本人の許可無く，個人名を特定できる資料（同意書，フェイスシートを含む）を公表することはありません。
5. 上記の説明を読み，内容を充分に理解されたうえで，観察，実験に参加していただける場合のみ，以下の署名および日付欄に自筆による記入をお願いいたします。

参加者氏名： _____ 所属： _____

署名年月日：(西暦)_____ 年_____ 月_____ 日