OUKA
Osaka University Knowledge Archive

Title	小規模な計算機資源で実現可能な高品質規則音声合成システムの研究-LSPパラメータにベクトル量子化を適用したVCV素片接続型音声合成方式の開発-
Author(s)	清水, 忠昭
Citation	大阪大学, 2002, 博士論文
Version Type	VoR
URL	https://hdl.handle.net/11094/2800
rights	
Note	

Osaka University Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

Osaka University

小規模な計算機資源で実現可能な高品質規則音声合成システムの研究

＿ LSPパラメータにベクトル量子化を適用した ＿
VCV素片接続型音声合成方式の開発

2001年11月

清 水 忠 昭

2002 年 1 月 22 日 PDF 作成

論文要旨

日本語における規則音声合成の研究は1970年代後半から本格化し，現在に至るまで高品質な合成音声を目指して音声合成システムの規模は拡大を続けてきた. しかし, 医療現場での患者とのコミュニュケーション補助装置などの医療機器や, カーナビゲーション装置や個人向けの小型端末であるPDA(Personal Digital Assistants)などのように, 音声による情報伝達を必要とするが, そのために多くの計算機資源を割けないような応用は現在でも多く存在する. このような応用においては, 高品質な音声合成システムを小規模な計算機資源で構成できる規則音声合成方式の研究開発が必須である. 本論文では, このような目的を達成する枠組みとして，1Mから4Mバイト程度の小規模な合成単位辞書によって高品質な合成音声を得ることを目標としたLSPベクトルVCV規則音声合成方式を提案した. 本手法では，VCV素片の記憶のためにベクトル量子化されたLSPパラメータを用いることにより，合成単位辞書の容量を小さく抑えながら，多数のVCV素片を格納することを可能にしている. これにより, 小規模な音声合成システムでも合成音声の品質を向上できる可能性が高いのが本方式の特長である.

本研究では, 提案したLSPベクトルVCV規則音声合成方式に必要な仕様を実験的に決定するとともに，その性能を評価した．特に合成音声の品質を左右する合成単位素片(VCV素片)の選択手法として，音韻環境を素片選択基準とするPER選択法と，接続歪みを最小化するMLD選択法を提案し，素片選択実験と合成音声の品質評価実験を行った. 本研究で得られた主だった成果を下記に列挙する.

1） 本方式では，VCV素片の収録数にして14,000個程度の規模の合成単位辞書を用いればよい.

2） PER選択法とMLD選択法では，選択されるVCV素片は一致しないが，聴感上の品質は同程度である.

3） PER選択法とMLD選択法の選択基準には，強い関係が存在する.

4) 先行音韻環境として考慮する音韻の個数は2個, 後続音韻環境として考慮する音韻の

個数は1個で十分である.

5） 本法におけるベクトル量子化のコードブックサイズは128程度で十分であること

以上の結果により, 合成単位辞書とVQコードブック, 残差波形辞書を合わせても当初の目的よりも少ない500Kバイト程度の記憶容量で高品質な音声合成システムを構成できることを示し，本手法の有効性を示すことができた.

また，LSPベクトルVCV規則音声合成方式による合成音声の明瞭性と自然性の一層の向上のために, LSP音声合成フィルタの駆動音源のための残差信号の符号化法として, 少ない容量で破裂子音部の残差信号を符号化するPEC法と，ウェーブレット変換を用いる手法を提案した. 評価実験により, これらの手法を用いて合成音声の明瞭性と自然性を向上できることを示した.

以上の研究成果により, 高品質な音声合成システムを小規模な計算機資源で構成できる規則音声合成方式の開発という目的を達成できたものと考える. 今後は, 本手法による音声合成システムの実用機器への組み込み実証試験を行いたい. また, 本手法の応用研究も広がってゆくことを願っている.

目次

目　次

目　次

図目次

目　次

表目次

略語一覧

本論文では, 便宜上多くの略語を用いている. 略語による混乱を防ぐために, 以下に主な略語の一覧を示し，簡単な説明と本文中の参照ページ及び参考文献番号を付記した.

音声分析に関する用語

PARCOR 係数　**PAR**tial auto-**COR**relation coefficient:
偏自己相関係数

線形予測系の音声分析手法であるPARCOR分析のパラメータ. (本文 p.11, 参考文献[5] 〜 [7])

LSP パラメータ　**L**ine **S**pectrum **P**air parameter:
線スペクトル対パラメータ:

線形予測系の音声分析手法である LSP 分析のパラメータ. PARCOR 分析と LSP 分析は，線形予測法による音声の符号化特性を改善するために，1970 年代後半から 1980 年代にかけて相次いで開発された. (本文 p.11, 参考文献[8] 〜[10])

規則音声合成の合成単位

CV 単位，CV-VC 単位，VCV 単位，CVC 単位，CVCV 単位

日本語の規則音声合成のための合成単位は一般に母音 **V**(Vowel)と子音 **C**(Consonant)の組み合わせの単位が用いられる. (本文 p.16，参考文献[11] 〜[21])

素片選択法に関する用語

LD　**L**SP **D**istance
LSP パラメータ距離

VCV 素片選択時に素片の接続歪みを評価する目的で定義した LSP パラメータ間の距離. (本文 p.41)

PER スコア **P**honemic **E**nvironmental **R**esemblance Score
音韻環境類似度スコア

VCV 素片選択時に素片の音韻環境の適合度を評価する目的で定義した指標. (本文 p.40)

LD-PER 平面 **L**SP **D**istance - **P**honemic **E**nvironmental **R**esemblance plane

接続歪み指標と音韻環境指標を 2 軸とする平面. 素片選択結果の評価に用いる. (本文 p.62)

MLD 選択法 **M**inimal **L**SP **D**istance method
LSP 距離最小化選択法 (本文 p.41)

PER 選択法 **P**honemic **E**nvironmental **R**esemblance method
音韻環境類似度選択法 (本文 p.39)

DTL 選択法 **D**istance **T**able **L**ook-up Method
距離テーブル参照法 (本文 p.78)

DP **D**ynamic **P**rogramming
動的計画法 (本文 p.42，参考文献[43])

ベクトル量子化に関する用語

VQ コードブック **V**ector **Q**uantization Code Book
ベクトル量子化コードブック (本文 p.27，参考文献[32])

VQ インデックス **V**ector **Q**uantization Index
ベクトル量子化インデックス (本文 p.27，参考文献[32])

LBG アルゴリズム **L**inde-**B**uzo-**G**ray algorithm

ベクトル量子化のコードブックを作成するためのアルゴリズム．命名は開発者の名前より. (本文 p.82，参考文献[32])

その他

PEC 法 **P**ower **E**nvelope **C**oding 法
パワ・エンベロープ符号化法 (本文 p.96)

PDA **P**ersonal **D**igital **A**ssistants

携帯して用いることを前提とした小型携帯端末. キーボードを有する場合もあるが，ペン入力を持つ電子手帳型のものが一般的. 携帯性を優先させた設計のため, 計算処理速度やメモリ量など計算機資源の点では小規模である．(本文 p.1)

第1章　序論

本論文では, 主たる応用分野として医療機器やカーナビゲーション装置, PDA(Personal Digital Assistants)といった機器への組み込み利用を想定した規則音声合成手法について論じる. このような応用においては, 高品質な音声合成システムを小規模な計算機資源で構成できる規則音声合成方式の研究開発が必須である. 本稿では, 音声合成システムの小規模化を達成するために LSP ベクトル VCV 規則音声合成方式を提案し，その研究開発を行った.

以後の章では, 提案した方式において, 適切な合成単位辞書の規模や合成単位素片の選択方式, ベクトル量子化の代表ベクトル数など音声合成システムの性能を決定する仕様について, 聴覚的な主観評価を含む実験により詳細に検討した結果を報告する. また, 合成音声の子音部の明瞭度向上のために, 音声合成フィルタを駆動するための音源に子音のパワ・エンベロープを用いる残差信号符号化法を提案し，その性能評価を行った. 本章では, これらの議論に先立ち本研究の背景と目的について述べ, 本論文の構成について述べる.

1.1 本研究の背景

1.1.1 音声の特質

人間同士の日常的な情報交換では, 視覚や他の感覚器官を用いるどんな情報伝達手段と比べても, 音声が最も大きな役割を担っている. 音声による情報伝達の優位性は, 日常的な経験からは良く理解できるが, 一度に受容できる情報量では人間の聴覚は視覚に全く及ばない事を考えると, 音声による情報伝達の優位性について単純には論じられないことが判る. このことは, 音声による情報伝達は, 文字や絵, 動画など視覚による情報伝達に比べ, 単純に情報量だけといった一面的な評価では尽くせない利点を持っていることを示している.

音声による情報伝達の最大の利点は, 情報の発信者による発話から情報の受信者による受聴までの各プロセスにおいて, 人間にとって負担が少ないことである. 我々は, 必要以上に意識しなくても楽に発話でき, 特別な場合を除けば他人の言葉を聞き取るにも神経を集中する必要はない. 仕事などの作業中に, 隣の人と会話を楽しむことさえ可能である.

音声による情報伝達の第ニの利点は, 人間にとって負担が軽いわりに, 意味伝達が比較的に高速で確実に行われるという点である. 音声による情報伝達では, 特別な道具を必要とせず, 発話した音声がすぐに相手に理解されるために, 話者間での情報伝達は非常に速い. また, 現代では通信手段として電話が良く用いられるが, 多くの場合電話によって伝えられる音声だけで情報交換はほぼ確実に行われる. このような利点を持つ音声言語による情報伝達は, 人間にとって日常的にごく自然な情報伝達手段である.

人間にとって自然な情報伝達手段である音声を人間と機械の間の情報伝達に取り入れることができれば, 利用者に多大な利便をもたらすことができる. 特に, 近年ではコンピュータやネットワークを中心とした情報の蓄積, 検索, 伝達等の技術の進歩により, 人間と機械の間の情報伝達はより複雑になる傾向があり, 音声の導入の必要性はますます高まっている. また, これらの技術の進歩は, 音声情報処理をより高度に行うことを可能にし, 音声による人間と機械の間の情報伝達を実用化する原動力ともなっている. 本研究では, 音声による人間と機械の間の情報伝達のうち, 機械から人間に向けての情報発信にあたる音声合成に関する研究を行った.

1.1.2 音声合成技術の進歩

人間と機械の間の情報伝達のための音声合成技術として, **録音編集方式**が早い時期から実用化されている. 録音編集方式は, あらかじめ録音した音声を必要に応じて繋ぎ合わせて文を作成する方式であり, 成功した例として JR の列車アナウンス等がある. JR の列車アナウンスの場合は, 定型の発話文を録音しておき, 駅名や時刻等を置き換える方法で音声を生成する. 録音編集方式は, 編集する単位が単語や文単位のように長い場合は, 生成された音声は明瞭で自然性も失わない. しかし, 設計時に想定された定型の文しか生成することができず, 自由度に欠けるという欠点がある.

一方で, テキストデータの読み上げなどを目的とし, 任意のテキストから音声を合成する技術を**テキスト音声合成**と呼ぶ. テキスト音声合成は, 録音編集方式での実現は不可能であり，規則によって任意の文音声を合成可能な**規則音声合成方式**[1]の開発が必要である．テキスト音声合成は，英語では比較的古くから研究されており[2][3]，日本語における規則音声合成も，1970 年代後半から本格的に研究が開始された[4]．その中でも，音声の合成単位素片を接続することにより任意の音声を生成する**単位素片接続型の規則音声合成方式**が数多く提案された. 単位素片接続型の規則音声合成方式では, 合成に用いる合成単位を人間の発話から採取し，音声合成システム内の合成単位辞書に保持する必要がある．合成単位辞書の規模を小さくするために，合成単位素片は **PARCOR 分析**[5]〜[7]や **LSP 分析**[8]〜[10]等による分析パラメータの形で記録する手法が主流を占めていた.

日本語は子音と母音の組み合わせからなるCV音節を単位として発声されることから，子音と母音を組み合わせた音声素片を合成単位とする単位素片接続型の規則音声合成方式が各種提案されてきた．最も単純な CV 単位に基づいた方式[11]〜[16]では合成単位の数が少なく，1985 年当時のハードウェアでも十分に実現可能であった．さらに高品質な合成音声を得るために, 音韻から音韻に至る音声の動的な部分を含むより長い合成単位を用いる方式として，CV-VC 単位を用いる方式[17]や，より長い単位である VCV 単位[18]を用いる方式，さらにはCVC単位[19]を用いる方式も提案された．これらは，CV 単位に比べて簡単な制御でより良い品質の合成音声を生成できるが, 合成単位辞書に記憶する合成単位数が非常に多くなる.

1980 年代後半以後，コンピュータの技術進歩にともなって，音声合成の手法も大規模化した. より長い合成単位を用いる方式として, CVCV単位にまで選択範囲を広げて合成単位のセットを検討する研究[20]や，より長い合成単位を選択的に用いる手法[21]が提案された. これらによる合成音声の品質はCV単位による合成音声を大きく上回ることが報告されている. 一方で, クラスタリング技術によって音韻環境を基準にして合成単位を自動的に生成する手法[22][23]も提案された．さらに，近年における計算機の記憶容量の増大にともない, 合成単位素片を分析パラメータではなく音声の時間波形の形で記憶し, それを接続する方法で音声合成を行なう**波形編集方式**[24]〜[31]あるいは**波形重畳方式**が提

案されている.

波形重畳方式は, 分析パラメータを用いる方式に比べて合成音声の自然性が高く, 品質の高い合成音声を得ることができる. しかし, 合成単位辞書に音声の時間波形を記憶しておくために非常に大きな記憶容量が必要である．単位素片接続型の規則音声合成の研究は, 合成音声の高品質化を目指して, より長い合成単位を用いる方向へ, またより多くの合成単位素片を保持する方向へと発展し, 音声合成システムは大規模化の一途を辿ったといえる.

1.2 本研究の目的

前節で述べたように, 音声合成の研究分野ではより高品質な合成音声を目指して, 合成音声を生成するために合成システムが保持すべき合成単位辞書の記憶容量は増加してきた. 現在, 実用されている音声合成システムの多くはパーソナル・コンピュータ上で動作しており, 比較的多くのメモリと計算速度を必要とする. 一方で, 医療現場での患者とのコミュニュケーション補助装置などの医療機器や, カーナビゲーション装置や個人向けの小型端末である PDA などのように，音声による情報伝達を必要とするが，そのために多くの計算機資源を割けないような応用は現在でも多く存在する. このような用途では, 合成単位辞書の記憶容量を小さくし, 音声合成の処理を簡便にする必要があるが, 合成音声の品質を大きく引き下げる事は許されない. これに対し, 少ない情報量で高品質な音声合成を行なうために, 麻生らによって音声素片データを記録するパラメータに**ベクトル量子化**[32]を適用する手法が特許出願されている[33].

我々は, 規則音声合成システムのROM化などの小規模な応用に向けて, 1Mから4Mバイト程度の小規模な合成単位辞書によって高品質な合成音声を得ることを目標に, **LSPベクトル VCV 規則音声合成方式**[34]〜[39]の開発研究を続けてきた．本手法では，VCV 素片の記憶のためにベクトル量子化されたLSPパラメータを用いることにより, 合成単位辞書の容量を小さく抑えながら，多数の VCV 素片を格納することを可能にしている．これにより, 小規模な音声合成システムでも合成音声の品質を向上できる可能性が高いのが本

方式の特長である．本研究では，提案したLSPベクトルVCV規則音声合成方式に必要な仕様を実験的に決定するとともに，その性能を評価し，本手法の有効性を示した．

1.3 本論文の構成

本論文は, 高品質な音声合成を小規模な計算機資源でも実現できる規則音声合成方式の実現を目的として続けてきたLSPベクトルVCV規則音声合成方式の開発研究の研究成果をまとめたものであり，本章を含む10章からなる．

第1章は，序論として，本研究の背景と目的について述べる．音声の特質や，日本におけるテキスト音声合成の研究の歴史について簡単にふれ, 本研究の位置付けと目的を明らかにする．さらに本論文の構成について述べる．

第2章では, 音声合成について基本的な概念について解説する．音声合成の大きな枠組みの中で, 本研究が目指すテキスト音声合成の位置付けを明らかにする．また, 日本語の規則音声合成の先行研究の流れを, 合成単位の長さに着目して分類するという視点から概観する．

第3章では，最初に本研究における音韻やVCV素片の取り扱い方法について述べる．続いて，本論文で提案するLSPベクトルVCV規則音声合成方式について詳述する．

第4章では，LSPベクトルVCV規則音声合成方式における，合成単位辞書の大きさを決める収録素片数と素片選択方式について議論する. 素片選択方式として音韻環境を基準としたPER選択法とVCV素片の接続歪みを基準としたMLD選択法の2つを提案した．提案した音声合成システムによる合成音声の主観評価実験により, 合成単位辞書の規模の評価と，2種類の素片選択方式の比較を行った実験結果について報告する. [34]

第5章では，PER選択法とMLD選択法による素片選択結果について，両者の選択基準の関係を実験的に検証した．実験の結果, 両者の選択基準には, 一方を最適にすると他方は準最適になるという強い関係があることを示した. [36]

第6章では，PER選択法において考慮すべき音韻環境の長さについて改めて議論した．

その結果，先行音韻環境として2音韻，後続音韻環境として1音韻を考慮すれば十分であることが示された. [37]

第7章では，LSPベクトルVCV規則音声合成方式における，ベクトル量子化の量子化数について検討を行った. また，ベクトル量子化の特性を利用して，距離テーブルを用いることにより，MLD選択法の素片選択速度を改善するDTL選択法を提案する. [35]

第8章では，LSPベクトルVCV規則音声合成方式の合成音声において，破裂子音部の明瞭性を改善するための残差信号の符号化法を提案する. 裂子音を含む合成音声による主観評価実験により，提案した符号化手法の評価を行った. [38]

第9章では，合成音声の自然性の改善のために，ウェーブレット変換を用いた残差信号の符号化法について検討する. [39]

第10章は，本論文の結論である．第2章から第9章で議論した結果を総括し，今後の課題についても簡単にふれる.

第2章　音声合成の概要

音声に関する研究の歴史は長く, 音声合成についても多くの研究がなされてきた. 音声合成に関する多彩な研究の数々は, 音声合成技術が非常に多くの課題を含み, 様々な視点からのアプローチが可能なことを示している. この点を踏まえ, 本論文の主要テーマであるLSPベクトルVCV規則音声合成方式に言及する前に，本章では音声合成の概要について述べる. まず, 本研究が目指すテキスト音声合成あるいは規則音声合成が, 音声合成の大きな枠組みの中でどのような位置付けを持つか考察する. 次に, テキスト音声合成の主要技術である単位素片接続型の規則音声合成技術について述べる. さらに, 日本語を対象とした規則音声合成の過去の研究の流れを, 合成に用いる合成単位の長さの視点から概観することにより，本研究の位置付けを明らかにする.

2.1 音声合成のためのモデル

2.1.1 音声生成過程と音声合成系のモデル

人間が情報伝達手段として音声を用いるとき, 相手に伝えたい**意味概念**を音声信号に変換する作業を無意識に行っている. 意味概念が人間が脳の中でどのような形でコーディングされているかは明らかになっていないが, その伝達において**言語**が用いられていることは確かである. **音声**は, 言語を音波というキャリアに乗せて相手に伝達する手段であるといえる. このように考えると, 人間が音声を発話する際の**音声生成過程**は, 概略的には図2.1のようにモデル化できるであろう.

人間の音声生成過程において, 相手に伝えるべき意味概念は**言語符号**に変換され, さらに発声器官を制御する**生体制御信号**へと変換される. 生体制御信号は神経系を介して発声器官を制御する. 与えられた制御信号に従った発声器官の働きによって, 物理的に音声が発声される. このようにモデル化された人間の音声生成過程を, 機械による音声合成過程に結び付けて考えると，図2.1のように音声合成技術を階層的に考えることができる.

まだアイデアのみで実用には遠い道のりが残されているが, 最も高度な音声合成は**概念からの音声合成**である. 概念からの音声合成では, 意味概念からテキスト生成系によりテキストを生成する. テキストは言語を文字で表したものであり, 人間の音声生成過程では言語符号に対応する. 人間の音声生成過程において発話器官を制御する身体制御信号は, 機械による音声合成過程では音声の**音響モデル**のパラメータに対応する. 人間の音声生成過程において言語符号が身体制御信号に変換されるのに対応して, 音声合成の過程においてテキストはパラメータ生成系によって**音響パラメータ**に変換される. この音響パラメータにより制御された音響モデルの働きによって，音声信号が生成される.

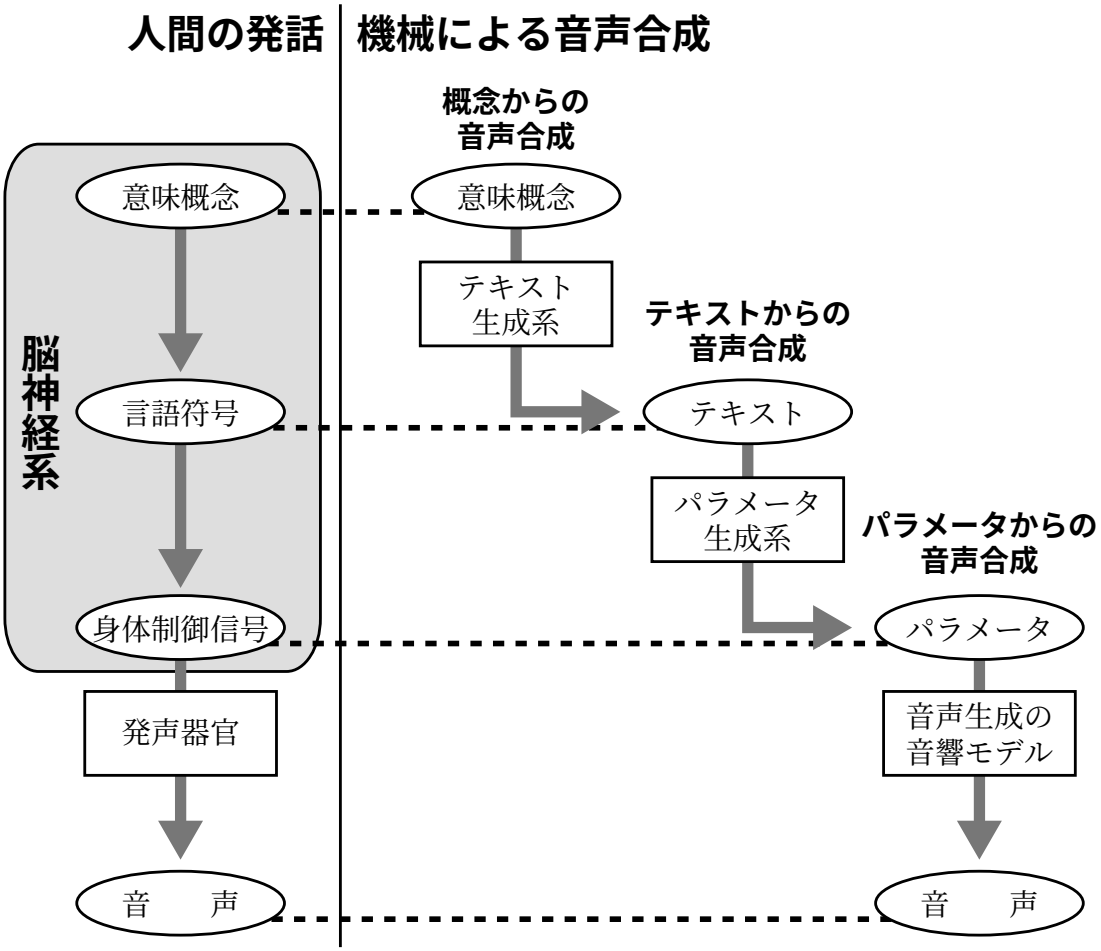


図 2.1　人間の音声生成過程と音声合成の階層構造

図2.1に示すように概念からの音声合成は, 階層的な形で下位にテキストからの音声合成(以下，単に**テキスト音声合成**と呼ぶ)を含んでいる．さらに，テキスト音声合成は，その下位にパラメータからの音声合成(以下，単に**パラメータ音声合成**と呼ぶ)を含んでいる. パラメータ音声合成は, 音声の音響分析の結果として比較的早くから研究され実用化も進められてきた．詳しくは次項で述べるが，**線形予測法**に代表される分析手法により，人間の発話器官の音響的な特性を近似するモデルが数多く提案されている.

概念からの音声合成は, 現在でも解決の目処が立たない課題が多く, まだ先の未来の技術である．一方，パラメータ音声合成は，すでに多くの実用技術を生み出している．テキスト音声合成は，それらの中間に立つ技術であり, 現在盛んに研究が行われている．テキスト音声合成の目標は, テキストに即して適切な音響パラメータを生成し, 音響モデルを制御して高い品質の合成音声を生成することである．このために，テキスト音声合成では，音素等を規則に従って接続したり, 韻律等を規則により制御することにより, 適切な音響パラメータを生成し音声を合成する. このように, 何らかの規則に従って音声合成を行う方式を**規則音声合成方式**と呼ぶ.

2.1.2 音声生成の音響モデル

音声生成の音響モデルは, 人間の音声生成過程の中で物理的な音波を発生させる発話器官の部分のモデルである．人間の発話器官における音声生成は，図2.2に示すように音源の生成，調音，放射の3つの過程からなる．母音などの有声音の場合，音源の生成は，肺からの空気流によって声帯が振動することによる．生成された空気振動は，喉頭や口腔，鼻腔などからなる共鳴管である声道で音色を付与され, 口唇から音波として放射される．無声音の場合は, 様々な音源が用いられるが, 代表的なものは声道の一部を狭くすることによって発生する乱流音である.

人間の発声器官を音声の分析や合成に利用出来るようにモデル化したものが, 音声生成の音響モデルである. 現在, 主流となっているモデルは, 音源の生成と調音を完全に分離した形でモデル化した**線形分離等価モデル**である. 線形分離等価モデルでは, 音源の伝達関数を $G(z)$，調音フィルタの伝達関数を $H(z)$として，音声 $S(z)$を次の式で表すことがで

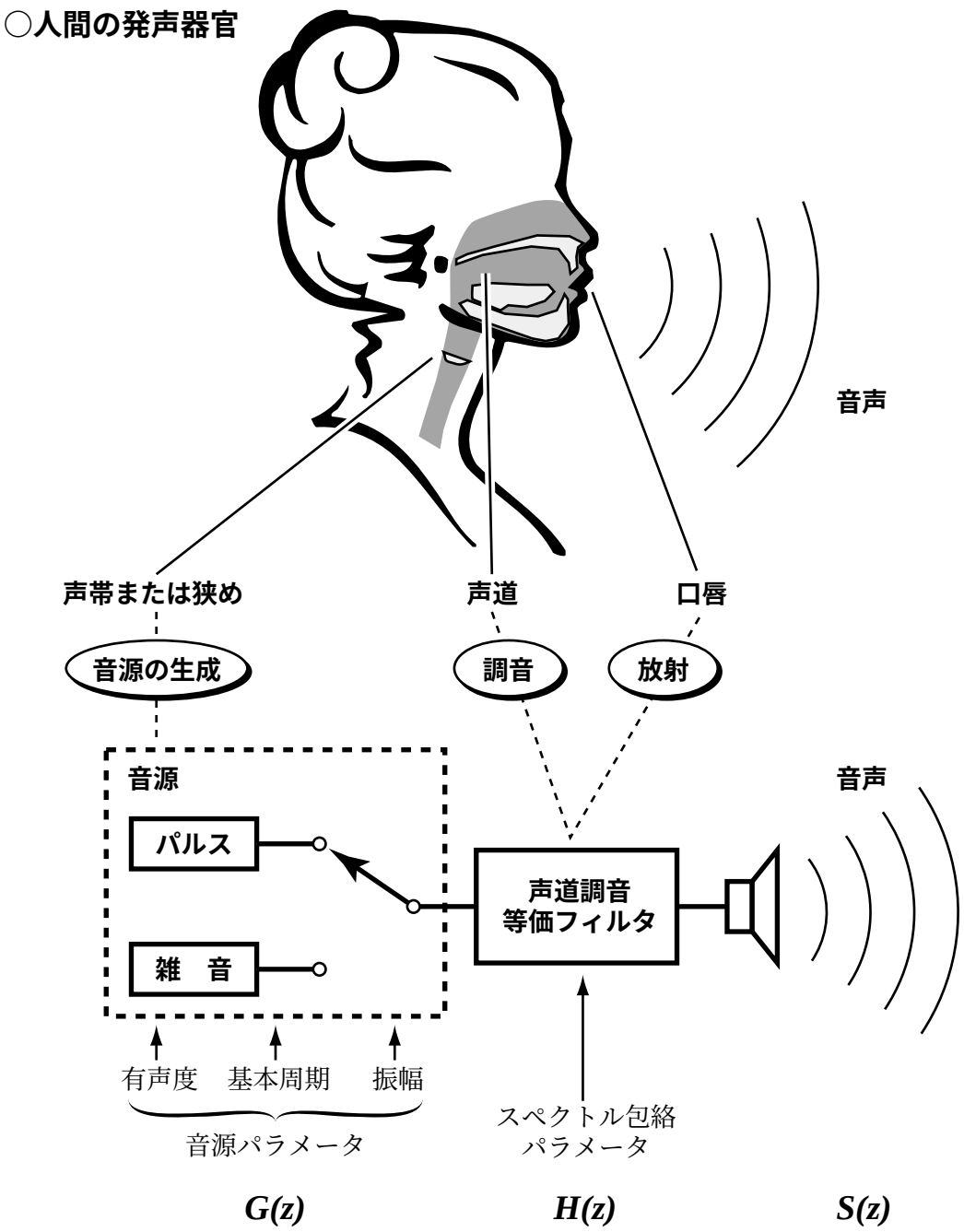


図 2.2　人間の発声器官と音響モデル

きる.

$$S(z) = H(z)G(z) \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (2.1)$$

音源の生成を行う**音源部**は, 最も簡単なモデルでは有声音に対応するパルス列と, 無声音に対応する雑音源でモデル化される. しかし, 明瞭で自然な合成音声を生成するためには音源部をより高度なモデルにする必要がある. 本論文でも, 後の章で, 合成音声の明瞭性の向上を目標に規則音声合成に適した音源部のモデルについて検討している.

線形分離等価モデルの**声道調音等価フィルタ部**には, 調音の特性だけでなく, 口唇における放射特性も含めて簡易に扱うことが多い. 声道調音等価フィルタの実現方法には, 非常に多くの方式が提案されている. 代表的なものとしては, 線形予測によって調音特性を推定しフィルタを構成する線形予測フィルタがある. 線形予測系の分析合成技術は良く研究され，音声合成により適した PARCOR 分析[5]～[7]や LSP 分析[8]～[10]が開発されている. また, 音響パラメータとしてメルケプストラムパラメータを用いるメル対数スペクトル近似(MLSA)フィルタを用いる方式[12]なども提案されている．本研究では，パラメータの接続性や音響フィルタの安定性の点から規則音声合成系に適用しやすいLSP分析による音響モデルを用いた.

2.1.3 LSP 分析合成系

LSP 分析は，線形予測分析と等価な音声分析法である．線形予測法では，式(2.2)に示す全極型のフィルタで声道調音フィルタを模擬する. このモデルにおいて, 分母の多項式にあらわれる係数 a_i を**線形予測係数**と呼ぶ.

$$H(z) = \frac{1}{A_p(z)} = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_p z^{-p}} \qquad \cdots\cdots\cdots\cdots\cdots \quad (2.2)$$

式(2.2)の分母の多項式 $A_p(z)$ は，次の漸化式を満たす.

$$\begin{aligned} A_n(z) &= A_{n-1}(z) - k_n z^{-1} B_{n-1}(z) \\ B_n(z) &= z^{-1} B_{n-1}(z) - k_n A_{n-1}(z) \end{aligned} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (2.3)$$

但し, $B_n(z) = z^{-n} A_n(z)$ であり，初期条件は $A_0(z) = 1$，$B_0(z) = 1$ である．式(2.3)中の k_i

は，**PARCOR 係数**(偏自己相関係数: **par**tial auto-**cor**relation coefficient)とよばれる係数であり，この係数を用いて線形予測による声道調音等価フィルタと等価な**格子型のフィルタ**を構成することができる．

式(2.2)の $A_p(z)$ が与えられたとき，$k_{p+1} = 1$ および $k_{p+1} = -1$ としたときの $A_p(z)$ を，それぞれ $P(z)$ ，$Q(z)$ とする．$k_{p+1} = 1$ および $k_{p+1} = -1$ とすることは，PARCOR 係数による声道調音等価フィルタの声門の境界条件を完全開端または完全閉端とすることに相当する．

$$\begin{aligned} P(z) &= A_p(z) - z^{-1} B_p(z) \\ Q(z) &= A_p(z) + z^{-1} B_p(z) \end{aligned} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (2.4)$$

このとき，声道調音等価フィルタ $H(z)$ が安定であれば，$P(z) = 0$，$Q(z) = 0$ の根は複素平面の単位円周上に互いに他を隔離するように配置することが証明されている．従って，p が偶数の場合，$P(z)$ ，$Q(z)$ は以下のように因数分解できる．

$$\begin{aligned} P(z) &= (1 - z^{-1}) \prod_{i = 2,4,\cdots,p} (1 - 2\, z^{-1} \cos \omega_i + z^{-2}) \\ Q(z) &= (1 + z^{-1}) \prod_{i = 1,2,\cdots,p-1} (1 - 2\, z^{-1} \cos \omega_i + z^{-2}) \end{aligned} \qquad \cdots\cdots\cdots\cdots\cdots\cdots \quad (2.5)$$

但し，$\{\omega_i\}$を，$0 < \omega_1 < \omega_2 < \cdots < \omega_{p-1} < \omega_p < \pi$の条件で順番付けるものとする．このようにして得られた$\{\omega_i\}$は，調音等価フィルタの境界条件を完全反射とし無損失にすることで生じる線スペクトルの角周波数であり，**LSP(線スペクトル対: Line Spectrum Pair)パラメータ**と名付けられている．LSPパラメータによって，$P(z)$ ，$Q(z)$ が決まり，$P(z)$ ，$Q(z)$ から以下の式(2.6)によって $A_p(z)$ を再構成できる．

$$A_p(z) = \frac{P(z) + Q(z)}{2} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (2.6)$$

式(2.6)からも判るように，LSP パラメータから線形予測による声道調音等価フィルタと等価なフィルタを構成できる．p が偶数の場合について，LSP パラメータを用いて構成した音声合成フィルタ(声道調音等価フィルタ)を，図 2.3 に示す．

これまでに述べたように，LSP パラメータは，PARCOR 係数と並んで，線形予測係数と等価なパラメータと言って良い．しかし，LSPパラメータは，線形予測係数やPARCOR

係数と比べて, 量子化特性に優れ少ないビット割り当てでも合成音声の劣化をまねきにくい. 自然な音声を分析して求めたLSPパラメータの例を図2.4に示す. 局所的な微細な動きを除けば, LSPパラメータは発話中の音韻の推移に従って滑らかに変化する. この性質は, 音声素片を滑らかに接続するために非常に有利であり, 本論文で述べる単位素片接続型の規則音声合成の単位素片の符号化に適している.

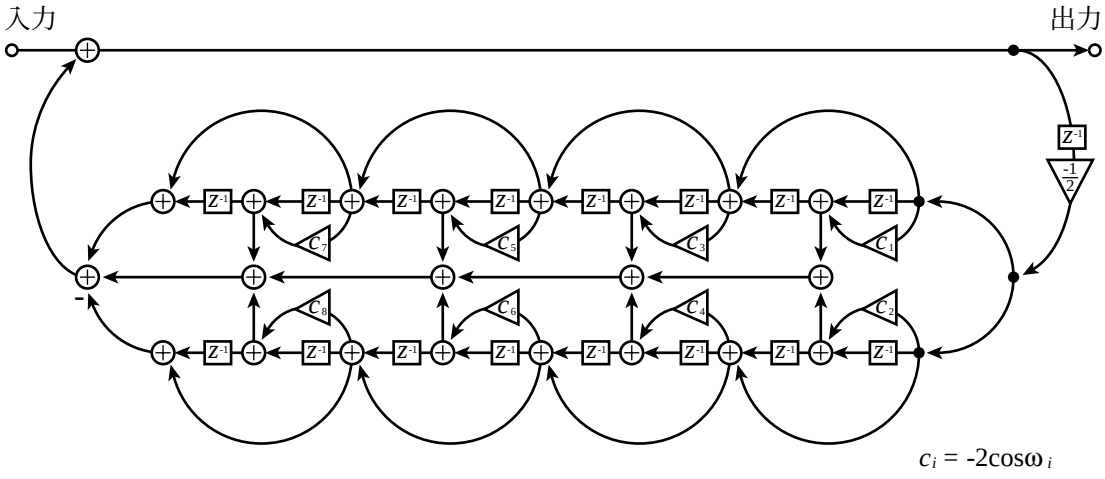


図 2.3　LSP 音声合成フィルタ

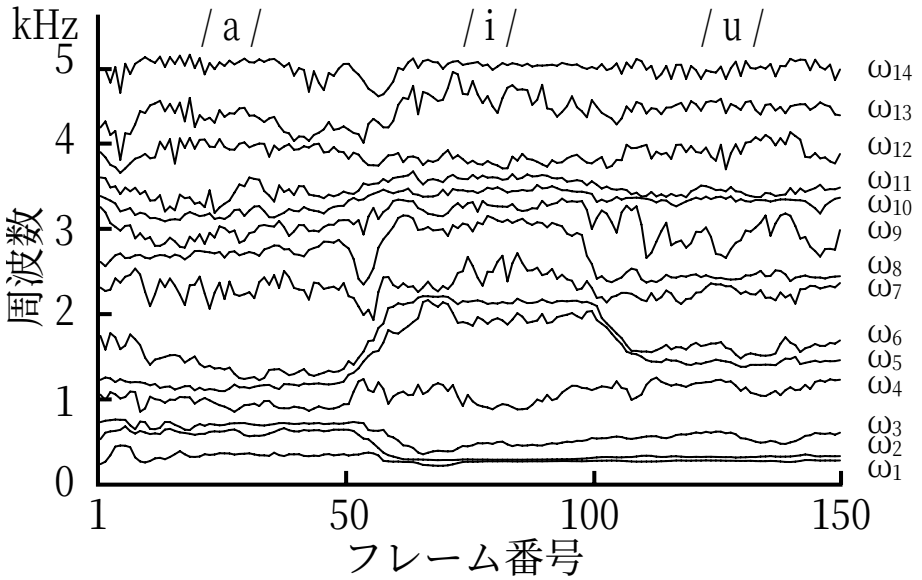


図 2.4　LSP パラメータの例

2.1.4 単位素片接続型の規則音声合成

音声合成技術として，録音編集方式が早い時期から実用化されている．録音編集方式は, あらかじめ録音した音声を必要に応じて繋ぎ合わせて文を作成する方式であり, 成功した例としてJRの列車アナウンス等がある．JRの列車アナウンスの場合は，定型の発話文を録音しておき, 駅名や時刻等の部分を置き換える方法で音声を生成する. 録音編集方式は, 編集する単位が単語や文単位のように長い場合は, 生成された音声は明瞭で自然性も失わない. しかし, 設計時に想定された定型の文しか生成することができず, 自由度に欠けるという欠点がある.

録音編集方式には制限が多く, 任意の文を合成することを目的としたテキスト音声合成を実現することは不可能である. テキスト音声合成の実現には, 規則によって任意の文音声を合成可能な規則音声合成方式[1]の開発が必要である．テキスト音声合成は，英語では比較的古くから研究されており[2]，**MITalk**[3]は良く知られている．日本語における規則音声合成も，1970年代後半から本格的に研究が開始された[4]．その中でも，音声の合成単位素片を接続することにより任意の音声を生成する単位素片接続型の規則音声合成方式が数多く提案された.

単位素片接続型の規則音声合成の合成単位として最も基本的なものは**音素**である．図2.5に，合成単位を音素とした場合の規則音声合成手法の概略図を示す．図中の**合成単位辞書**には, 音声合成用いる合成単位を人間の発話から採取した合成単位素片が登録されている．合成音声を生成する手順は以下の通りである.

1) 合成したい合成目的文を合成単位の系列に変換する
2) 合成単位の系列に従って，合成単位辞書から合成単位素片を選択する.
3) 合成音声の品質を高めるために，規則に従って，合成単位素片を変型，補間し接続して合成音声の音響パラメータを得る.

この手順の中で, 合成音声の品質を左右するのは, 合成単位素片の選択規則と, 合成単位素片の変型, 補間, 接続規則である．またこれらの規則は，合成単位をどのように選ぶかに影響される.

例えば，日本語の場合 30 から 50 種類の合成単位(音素)により任意の文を生成できるため, 合成単位素片を記憶しておくための記憶容量は少なくてすむ. しかし, 人間が発話する文音声において, 音素どうしの接続規則は複雑であり, 十分に高品質な合成音声を生成するために, 選択, 変型, 接続の規則を定めることは困難を極める. この困難を緩和する方法として, 合成単位を音素より長くとることにより, 合成単位自体に音声の動的特性を持たせ, 接続規則を簡易化することが考えられる. 一方で, 合成単位を音素より長くとれば，合成単位の数が増え，合成単位辞書は大規模化してしまう. この関係を図2.6にまとめておく. 一般に, 合成単位の長さを短くとれば, 合成単位辞書は小規模ですむが合成規則は複雑になり, 合成音声の品質は低くなりやすい. 逆に, 合成単位の長さを長くすれば, 合成単位辞書は大規模になるが合成規則は単純にでき, 合成音声の品質を高くしやすい. 次節では, 主として合成単位の長さに着目して, 日本語のテキスト合成のこれまでの先行研究を概観する.

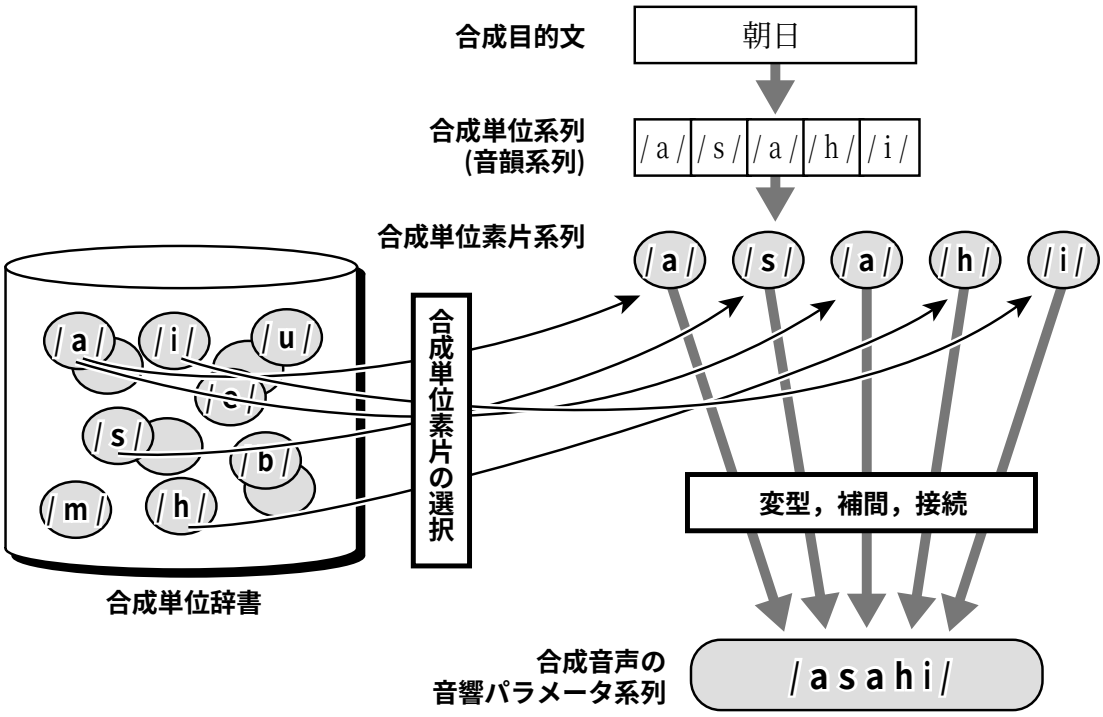


図 2.5　単位素片接続型の規則音声合成の概念図

2.2 日本語を対象としたテキスト音声合成
―合成単位から見た規則音声合成―

2.2.1 比較的短い合成単位を用いる規則音声合成

日本語は**子音 (Consonant)** と**母音 (Vowel)** の組み合わせからなる CV 音節を単位として発声されることから, 子音と母音を組み合わせた音声素片を合成単位とする合成方式が各種提案されてきた．1980 年代前半には，最も単純な **CV 単位**に基づく方法が提案された[11]．古市ら[13]は，CV 単位の規則音声合成において，合成単位素片の記録にメルケプストラムパラメータを用いて，合成時にはメル対数スペクトル近似(MLSA)フィルタ[12]を使用する方式を提案している．さらに，CV 音節データファイル(合成単位辞書)を自動作成する手法[14]を提案し, 人手によるわずかな手直しだけで合成単位辞書を作成できることを示している．1980 年代半ばには，新居ら[15][16]によって，LSP パラメータを用いた CV 単位の規則音声合成が提案され，DSP(Digital Signal Processor)を用いた装置化の検討もなされている．CV 単位の規則音声合成は合成に必要な合成単位の数が少なく，当時のハードウェアでも十分に実現可能であった．この手法による合成音声の品質については，新居は V-C 間接続の滑らかさに若干の不満が残ると報告している．

音韻から音韻に至る音声の動的な特性を，合成単位により多く持たせるために，CV 単位だけでなく VC 単位を加えた **CV-VC 単位**を用いる方式[17]方式が提案されている．CV 単位を用いる方式に比べて, 母音から子音への遷移部の動的情報を持つVC単位を用いるため，合成音声の品質を高めやすい．また，佐藤[18]は 1978 年という早い時期に，より長い合成単位として **VCV 単位**を用いる方式を提案している．佐藤の手法は，接続性の良

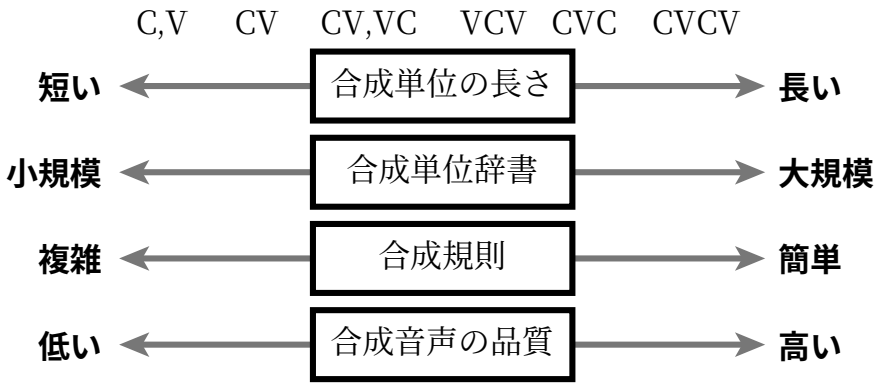


図 2.6　単位の長さと合成手法の性能

いＬＳＰパラメータが開発される前に提案されているため，音響パラメータとしてPARCORパラメータを用いており，母音部の接続での困難を伴っている．しかし，結果として高い単語了解度を示しVCV単位の有効性を示している．CV-VC単位やVCV単位を用いると，CV単位に比べて簡単な制御でより良い品質の合成音声を生成できるが，合成単位辞書に記憶する合成単位数が非常に多くなる．

より高品質な合成音声を目指して，佐藤[19]は**CVC単位**を用いて，子音部で素片を接続する手法を提案した．しかし，CVC単位はVCV単位に比べても合成単位数が非常に多くなるため，CVC単位を出現頻度が高いものだけに限り，他をCV単位やVC単位で補う方法をとっている．佐藤は，この方式により規則音声合成音声でありながら分析合成音声に近い高品質な合成音声が得られたと報告している．

2.2.2 規則音声合成システムの大規模化

1980年代後半以後，コンピュータの技術進歩にともなって，音声合成の手法も大規模化した．より長い合成単位を用いる方式として，**CVCV単位**にまで選択範囲を広げて合成単位のセットを検討する研究[20]が市川らによって行われた．この研究では，ピッチの影響まで含めて合成単位が検討した結果，国語辞典の単語についてCVCV単位を50%カバーし，他をCVC，CV混合でカバーする方法でも，8000種類という膨大な合成単位が必要であることを報告している．また，武田ら[21]は，より長い合成単位を選択的に用いる手法を提案している．この方式は，複数の基準を巧みに用い，長さが異なる合成単位を最適に組み合わせて用いるという非常に複雑な合成方式である．この方式では，音声合成のために最大で7音素程度の長さを持つ合成単位が用いられている．このような複雑さの代償として，合成音声の品質はCV単位による合成音声を大きく上回ることが報告されている．

これまでに述べた方式は，合成単位のバリエーションは先験的な言語知識に基づいて決められている．これに疑問を持った中嶌らは，**クラスタリング技術**によって音韻環境を基準にして合成単位を自動的に生成する手法[22][23]を提案している．一方，合成単位をVCVなどの比較的短い単位とし，同一の合成単位に対して様々な音韻環境を持つ複数の素片を保持することで，合成音声を高品質化する方法[25]もとられている．

多くの合成方式では，合成単位素片の記憶には，PARCOR や LSP，メルケプストラムなどの分析パラメータが用いられてきた. しかし, 近年における計算機の記憶容量の増大にともない, 合成単位素片を分析パラメータではなく音声の時間波形の形で記憶し, それを接続する方法で音声合成を行なう波形編集方式[24]〜[29]あるいは波形重畳方式と呼ばれる方式が提案されている. 波形重畳方式は, 分析パラメータを用いる方式に比べて合成音声の自然性が高く, 品質の高い合成音声を得ることができる. しかし, 合成単位辞書に音声の時間波形を記憶しておくために非常に大きな記憶容量が必要である. 小山らによる VCV 単位を用いる波形規則音声合成[25]の報告では，合成単位辞書に 60M バイトの記憶容量を要することが報告されている.

このように, 合成音声の高品質化を目指した研究は, 一般に音声合成システムの大規模化と一体になって進んできた. 合成音声の高品質化は, 声質[30]や感情を表現できる音声合成[31]の検討が可能な程度まで進んできた. しかし，これらの音声合成手法は，医療機器やカーナビゲーション装置，PDA といった機器への組み込み利用に向かない大規模なものとなってしまった.

2.3 まとめ

本章では音声合成の概要について述べた. その中で論じたように, 本論文で目標としているテキスト音声合成の実現手法として, 単位素片接続型の規則音声合成が主流である. 単位素片接続型の規則音声合成では, 合成単位の長さが合成単位辞書の規模や合成規則の難易度, 合成音声の品質に大きく影響する. 日本語を対象とした規則音声合成の過去の研究の流れを合成に用いる合成単位の長さの視点から概観すると, 合成音声の高品質化のために，合成単位を長くし合成単位辞書を大規模化する方向に進んできた.

一方で, 医療現場での患者とのコミュニュケーション補助装置などの医療機器や, カーナビゲーション装置，PDA などのように，音声による情報伝達を必要とするが，そのために多くの計算機資源を割けないような応用は現在でも多く存在する. このような用途では, 合成単位辞書の記憶容量を小さくし, 音声合成の処理を簡便にする必要があるが, 合

成音声の品質を大きく引き下げる事は許されない. このような観点に立ち, 本研究では規則音声合成システムの ROM 化などの小規模な応用に向けた研究開発を行った.

具体的な数値目標としては，1M から 4M バイト程度の小規模な合成単位辞書によって高品質な合成音声を得ることを目標とし, LSPベクトルVCV規則音声合成方式提案した. 次章では，提案した LSP ベクトル VCV 規則音声合成方式の構成の詳細について述べる.

第3章　LSPベクトルVCV規則音声合成方式

前章までに詳述したように, 本研究の目的は, 高品質な音声合成システムを小規模な計算機資源で構成できる規則音声合成方式の開発である. このような規則音声合成方式は, 医療機器やカーナビゲーション装置, PDA(Personal Digital Assistants)といった機器への組み込み利用の場面で必要とされている. 本章では, 音声合成システムの小規模化の具体的な目標として，1Mから4Mバイト程度の小規模な合成単位辞書によって高品質な合成音声を得ることを目標に，LSPベクトルVCV規則音声合成方式を提案する.

単位素片接続型の規則音声合成方式において高品質な合成音声を得るためには, 音韻環境の上で広いバリエーションを持つ合成単位素片を合成単位辞書に多数保持することが必要である．本方式では，合成単位素片(VCV素片)の記憶のためにベクトル量子化されたLSPパラメータを用いることにより, 合成単位辞書の容量を小さく抑えながら多数のVCV素片を格納することを可能にしている. 本章では, 以後の議論のために, まず本論文における音韻や音節の取り扱いを説明する．続いて，本研究で提案したLSPベクトルVCV規則音声合成方式[34]の構成について述べる.

3.1 VCV合成単位

3.1.1 音韻の取り扱いとVCV単位

本論文で提案するLSPベクトルVCV規則音声合成方式は，基本的には母音・子音・母音の組み合わせからなるVCV単位(Vowel-Consonant-Vowel Unit)を合成単位とする単位素片接続型の規則音声合成方式である．VCV型の規則音声合成手法は，合成単位を決定するにあたり母音や子音といった言語学上の先験的な知識を元にしているが, 日本語における**音韻**の種類や数には諸説がある. 本研究では音韻の種類と表記に関しては原則的に斎藤[40]の分類に従い, 外来語の音節表記(ウィ, ヴァ等)に対応する外来語音韻は取り扱わなかった.

表3.1 に，日本語の**音節**の一覧を示す．日本語の音節は，通常音節と特殊音節に分けられる．**通常音節**は，1ないし3音韻の接続によって構成される．1音韻からなるものは，**母音音節**であり仮名表記で「あ，い，う，え，お」にあたる音節である．2 音韻からなる音節は，子音 C と母音 V からなる CV 音節か，半母音 S と母音 V からなる SV 音節であり，これらを**標準音節**と呼ぶ. 標準音節は, 仮名表記で1文字となるほとんどの音節を含んでいる．3 音韻からなる音節は，子音 C と半母音 S，母音 V からなる CSV 音節であり，**拗音節**と呼ばれる．拗音節は，仮名表記では「きゃ，にょ」のように2文字の表記となる．**特殊音節**は，**撥音**「ん」と**促音**「っ」，**長音**「ー」であり，これらの音節は語頭には使用できないという制限があり．促音と長音は単独では発話できない.

本研究では，VCV 音声合成を簡便に行うため，半母音，拗音，撥音については特に以下に述べる取り扱いをした．半母音 S(/ j /, / w /)は子音 C の一種として扱った．また，拗音節を作る CS(/ kj / , / sj / 等)を一つの子音として扱い変則的な表記法を用いた．具体的には,「きゃ：kja」のように表記されるべき音節を,「きゃ：Ka」のように，CS の表記を S を省略して，C を表す**音韻記号**の大文字1文字で表記した．このような取り扱いのため，表 3.2 の音韻分類表に示すように，子音は 26 種類である.

また，撥音 / X / は母音の一種として扱ったため，標準的な日本語 5 母音に撥音を加えて母音は 6 種類である．撥音 / X / は，次に続く音韻によって様々に発音されることから，VCV 規則音声合成システムでは 4 つ程度に分類して扱うことが多い．しかし，本研究の音声合成系では, 同じ VCV 合成単位に属する素片を様々な音韻環境から複数切り出して使用するため, 撥音を複数に分類することはしなかった. これは, 音声合成時の合成単位素片の選択において適切な撥音を含む合成単位素片が選択されることを期待しているためである.

日本語の母音は, 音韻環境によって声帯の振動を失う無声化を起こす場合がある. 一般的な標準語では，無声化を起こすのは / i / と / u / であるが，表 3.2 では，形式的に全ての母音に無声化を想定して音韻記号を割り当てている．但し，実験システムにおける VCV 素片収集に当たっては，母音の無声化については考慮しなかった.

このような音韻の取り扱いをしたため，合成単位としては母音で子音を挟む形の VCV

表 3.1　日本語の音節一覧

通常音節（自立音節）

V：母音音節

あ：**a**　い：**i**　う：**u**　え/へ：**e**　お/を：**o**

CV：標準音節

か：**ka**	き：**ki**	く：**ku**	け：**ke**	こ：**ko**
が：**ga**	ぎ：**gi**	ぐ：**gu**	げ：**ge**	ご：**go**
さ：**sa**	し：**si**	す：**su**	せ：**se**	そ：**so**
ざ：**za**	じ：**zi**	ず：**zu**	ぜ：**ze**	ぞ：**zo**
た：**ta**	ち：**ci**	つ：**cu**	て：**te**	と：**to**
だ：**da**	ぢ：**zi**	づ：**zu**	で：**de**	ど：**do**
な：**na**	に：**ni**	ぬ：**nu**	ね：**ne**	の：**no**
は：**ha**	ひ：**hi**	ふ：**hu**	へ：**he**	ほ：**ho**
ぱ：**pa**	ぴ：**pi**	ぷ：**pu**	ぺ：**pe**	ぽ：**po**
ば：**ba**	び：**bi**	ぶ：**bu**	べ：**be**	ぼ：**bo**
ま：**ma**	み：**mi**	む：**mu**	め：**me**	も：**mo**
ら：**ra**	り：**ri**	る：**ru**	れ：**re**	ろ：**ro**

CSV：拗音節

きゃ：**Ka**	きゅ：**Ku**	きょ：**Ko**
ぎゃ：**Ga**	ぎゅ：**Gu**	ぎょ：**Go**
しゃ：**Sa**	しゅ：**Su**	しょ：**So**
じゃ：**Za**	じゅ：**Zu**	じょ：**Zo**
ちゃ：**Ca**	ちゅ：**Cu**	ちょ：**Co**
にゃ：**Na**	にゅ：**Nu**	にょ：**No**
ひゃ：**Ha**	ひゅ：**Hu**	ひょ：**Ho**
ぴゃ：**Pa**	ぴゅ：**Pu**	ぴょ：**Po**
びゃ：**Ba**	びゅ：**Bu**	びょ：**Bo**
みゃ：**Ma**	みゅ：**Mu**	みょ：**Mo**
りゃ：**Ra**	りゅ：**Ru**	りょ：**Ro**

SV：標準音節

や：**ja**　ゆ：**ju**　よ：**jo**
わ/は：**wa**

特殊音節(付属音節)

ん（撥音）：**X**　っ（促音）：**+**　ー（長音）：**-**

表 3.2　音韻の分類

音韻種別	音　　韻
母音・撥音	**/ a /, / i /, / u /, / e /, / o /, / X /**
無声化母音	**/ A /, / I /, / U /, / E /, / O /**
無声破裂子音	**/ k /, / t /, / p /, / K /, / P /**
有声破裂子音	**/ g /, / d /, / b /, / G /, / B /**
無声破擦子音	**/ c /, / C /**
有声破擦子音	**/ Z /**
無声摩擦子音	**/ s /, / h /, / S /, / H /**
有声摩擦子音	**/ z /**
鼻音	**/ n /, / m /, / N /, / M /**
流音	**/ r /, / R /**
拗音	**/ j /, / w /**
無音区間	#

型 570 種類，子音を挟まない VV 型 35 種類，語頭用の #CV 型 95 種類と #V 型 5 種類，語尾用の V# 型 6 種類がある．以後，これらの型の合成単位を総称して**VCV 合成単位**(VCV unit)と呼び，それぞれの VCV 合成単位の素片データを**VCV 素片**(VCV instance)と呼ぶ．

3.1.2 VCV 素片収集の方針

人間の自然な発話では, 同一音韻であっても, その前後に発話された音韻の影響で音響的な性質がかなり変化する．このような現象を**調音結合**と呼ぶ. 規則音声合成では, 調音結合の影響をうまく再現することが, 自然で高品質な合成音声を得るための一つのポイントである．単位素片接続型の規則音声合成において，調音結合の影響を再現するためには，以下の 2 方法が考えられる．

1) 代表的な合成単位素片だけを持っておき, それに対して何らかのルールに従った変型操作を加えることで調音結合の影響を再現する．
2) 調音結合の影響により様々に変型した合成単位素片をバリエーション豊かに保持しておき，その中から合成時に適切なものを選択する．

前者は, 合成単位素片の変型ルールをうまく決定できれば, 音声合成システムを効率的に構成できる．しかし実際には, 調音結合の影響が非常に複雑で, そのルール化が難しいために, 前者の方法では合成音声の品質が低くなる．後者は, 合成単位素片を格納する合成単位辞書が大きくなるなどの欠点があるが, 高品質な合成音声を生成しやすいという利点があり, 現在多く用いられている手法である．本研究で提案する方式では, 合成音声の品質の向上を優先して後者の方法をとり, 合成単位辞書が大きくなる欠点をベクトル量子化などの手法を用いて克服する．

ここまでに述べた観点に立つと，VCV 規則音声合成方式で高品質な合成音声を得るためには, 1種類のVCV単位に対して, 採取時の音韻環境が異なるVCV素片をバリェーション豊かに数多く保持する必要がある．VCV 素片は，人間の発話から収集するが，この際に大きく分けて次の 2 つの考え方がある．

i)　VCV素片のバラエティが片寄らないように, 人工的に音韻をバランスさせた文章

発話から VCV 素片を採取する.

ii) 自然な人間の発話の性質に合わせて最適な VCV 素片を収集するために，自然な文発話から VCV 素片を採取する.

前者は，VCV 素片をバランス良く採取できるため，どのような文音声でも一定の品質を持って合成出来る利点がある. しかし, 日常の人間の発話では非常に出現頻度の低い音韻環境の VCV 素片も保持することになるため，平均的な合成音声品質の割には合成単位辞書が大きくなる. 後者は, 人間の発話で頻度が高い音韻環境を持つ文音声には高い品質を与えられるため, 合成単位辞書の規模の割りには平均的な合成音声品質を高くできる. しかし, 日常で出現頻度が低い音韻並びの文を合成する場合には, 合成音声品質が低下する恐れがある．両者の一長一短を考慮した上，本研究では VCV 素片収集のために人工的に作成した発話文ではなく, 通常のニュース発話を用いた. 合成単位辞書を小規模化する場合には, 全ての文に対して一定の合成音声品質を保証して全体的に低い品質にするよりも，頻出する文の合成音声品質を向上させる方が有利だと考えたからである.

本研究では, カーナビゲーション等の音声案内や医療器具等での文書の読み上げなどを考慮して, 標準的な日本語発話の合成を対象とした. このため, 標準的な日本語と考えられる NHK アナウンサの発話を合成単位辞書の作成に用いた．実験では読み上げ文書として新聞の報道記事を用いるため, 分野が近い音声資料という意味で NHK の FM ラジオ・ニュースを合成単位の収集に用いた.

本項でのべた VCV 素片収集の方針をまとめると，以下の通りである.

a) 標準的な日本語音声発話(NHK アナウンサの発話)で

b) 音韻バランスにおいて日本語における自然な偏りを持った普通発話の資料から

c) 調音結合の影響により様々に変型した VCV 素片をバリエーション豊かに採取する.

3.1.3 VCV 素片の収集

合成単位辞書に収録する VCV 素片は，NHK の FM ラジオ・ニュースの男性アナウンサの発話部分に視察で音韻マーキングした資料を用いて，母音部の中間点で切り出す方法で自動的に生成した．VCV 素片の収集に使用した音声資料は，1997 年 8 月 6 日から 9 月 25 日までに放送された 7 日分の FM ラジオ・ニュースから切り出した同一の男性アナウンサの発話部分である．男性アナウンサの発話部分は，FM ラジオ・ニュース 1 日分で約 10 分間あり，合計 70 分間の音声資料となる．音声資料は，標本化周波数 11.025kHz，量子化数 16 ビットでサンプリングした．

表 3.3 に，音声資料から採取した VCV 合成単位の種類数と，各々の VCV 合成単位に属する VCV 素片の平均個数を示す．音声資料に含まれる VCV 素片が全ての VCV 合成単位を網羅していないため，表 3.3 の VCV 合成単位の種類数は，3.1.1 項で述べた VCV 合成単位の種類数に達していない．参考のために表中のカッコ内に論理的に可能な VCV 単位の種類数を示している．

音声合成時には，合成単位辞書に収録されていない VCV 合成単位は，子音と後続母音の一致する他の VCV 合成単位から先行母音の部分を除いて作成した CV 合成単位によって代用する．この際，先行母音部は補間によって作成する．収集した VCV 素片が全ての VCV 合成単位を網羅していない事が，素片選択あるいは合成音声品質に与える影響については第 4 章で詳しく議論する．

表 3.3　収集した VCV 素片

合成単位の型	合成単位の種類数	収録素片の平均個数
VCV型	470(570)	38.4
VV型	35(35)	177.1
#CV型	78(95)	21.5
#V型	5(5)	101.2
V#型	6(6)	362.7
合　計	594(711)	26,517

注）#は無音を表している．
　　合成単位の種類数のカッコ内は，論理的に可能な種類数である．
　　収録素片の平均個数の合計欄は素片の総数である．

3.2 VCV 素片の LSP パラメータのベクトル量子化

本論文で提案する LSP ベクトル VCV 規則音声合成方式は，合成単位素片である VCV 素片の記録にベクトル量子化されたLSPパラメータを用いる方式である. 本方式では, ベクトル量子化の高い情報圧縮能力により，1 種類の VCV 単位に対して採取時の音韻環境が異なる VCV 素片をバリエーション豊かに数多く保持しても，合成単位辞書(VCV unit dictionary)の記憶容量を小さく抑えることができる. 小規模な計算機資源しか利用できない環境下でも, 高品質な合成音声を与える音声合成システムを実現できる可能性が高いのが特長である.

音声資料から収集されたVCV素片の例と, その符号化手順を図3.1に模式的に示す. 図では VCV 素片 / asa / を例にとってある．最上段に示した音声信号はサンプリングされた時間波形であり, ピッチ周期を持ち比較的振幅が大きい母音部と, 周期性が見られず比較的振幅が小さい無声子音部の区別がよく判る．この例では，/ asa / の音声信号の長さは 1728 サンプルあり，量子化は 16 (bits / サンプル) = 2 (bytes / サンプル)である．従って，この 1 単位素片の記憶容量は 1728(サンプル)× 2(bytes / サンプル) = 約 3.4Kbytes である.

VCV素片の符号化の第一段階として, 音声資料から切り出したVCV素片の音声信号を分析フレーム単位でLSP分析してLSPパラメータの系列に変換する. 1フレームのLSPパラメータは，LSP 分析の次数と同じ数のパラメータの組である．図 3.1 中段には，256 サンプルの長さのフレームを，64 サンプルずつずらしながら，12 次の LSP 分析を行った例を示してある．分析フレームは 25 フレームあり，1 フレームの LSP パラメータは ω_1 から ω_{12} の 12 個のパラメータからなる．図中では LSP パラメータ系列は，時間軸(フレーム番号)を横軸とする 12 本の線として描かれている．このとき，ω_1 から ω_{12} の各パラメータに 4 (bits) = 0.5 (bytes)ずつ割り当てたとすると，LSP パラメータで表現された VCV 素片の記憶容量は，25(フレーム) × 12(次) × 0.5 (bytes) = 150bytes である.

VCV 素片の符号化の第二段階では，各フレームの LSP パラメータを，分析次数の次元を持つベクトルとみなして, ベクトル量子化を行う. LSPパラメータの系列として表現された VCV 素片は，**VQ コードブック**を用いてベクトル量子化することにより，**代表ベクトル**を表す **VQ インデックス**の系列として符号化される．従って，本方式では VCV 素片

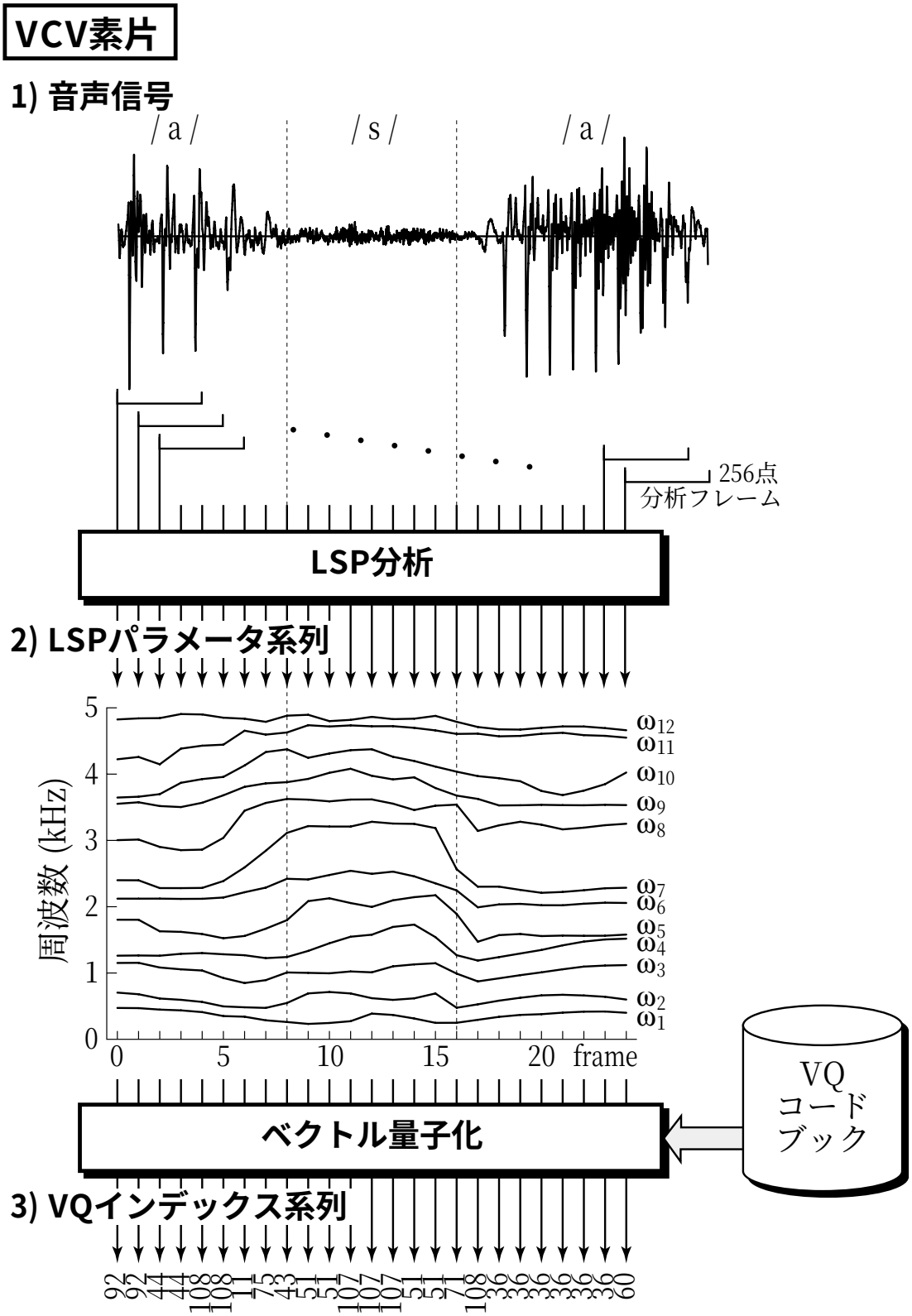


図 3.1　VCV 素片の符号化

はベクトル量子化の代表ベクトルのVQインデックスの系列として合成単位辞書に収録される．代表ベクトルが128個で，そのVQインデックスを7(bits)で表現できたと仮定すると，図3.1のVCV素片の記憶容量は25(フレーム) × 7(bits) / 8 = 約22bytesの容量となる．この例では，VCV素片を音声信号として記憶するのに比べて，LSPパラメータでは23分の1の記憶容量ですみ，さらにベクトル量子化を用いれば，160分の1の記憶容量となる．

一方, VQインデックスからのLSPパラメータを復号化は, 簡単に言うとVQコードブックからVQインデックスに対応する代表ベクトル取り出す操作である．図3.1に例示したVCV素片のVQインデックスから，LSPパラメータを復号化した例を図3.2に図示する．ベクトル量子化による符号化が, 非可逆的な情報圧縮にあたるため, 図3.1のLSPパラメータ系列と図3.2のLSPパラメータ系列を比較すると，LSPパラメータ系列は完全には再現されていない．本方式においては，合成音声の品質の観点から，VQコードブックの代表ベクトル数を決定し本手法にとって適切な情報圧縮量を定めることが重要な課題である．

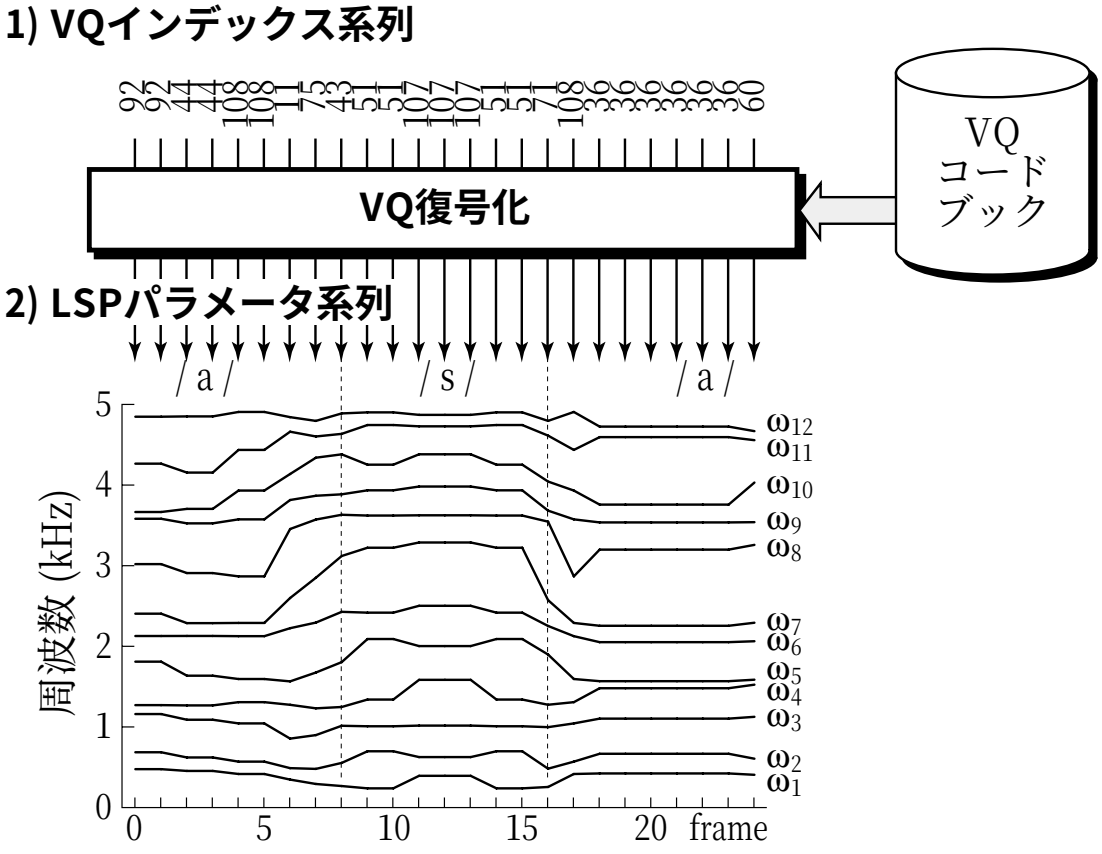


図 3.2　VQ インデックス系列からの VCV 素片の復号化

3.3 LSP ベクトル VCV 規則音声合成方式

3.3.1 LSP ベクトル VCV 規則音声合成方式の概要

LSPベクトルVCV規則音声合成方式の音声合成システムのブロック図を図3.3に示す. 音声合成システムは，以下の手順で合成音声を生成する.

1) 入力された所望の文章の文字列(**テキスト**:text)を，**音韻列**(phonemes)に変換する.
2) 音韻列を VCV 合成単位の系列(VCV sequence)に変換する.
3) VCV 合成単位の系列に従って合成単位辞書の中から VCV 素片(VCV instance)を選択し接続することにより，代表ベクトルの VQ インデックス系列(VQ index sequence)を得る.
4) 得られた代表ベクトルのVQインデックス系列を, VQコードブック(VQ code-book)を参照することにより復号化し, 所望の文章を合成するためのLSPパラメータ系列(LSP parameters sequence)を得る.
5) 一方，VCV 合成単位の系列に従って，**代表残差波形辞書**(residual signal waveform dictionary)を用いて駆動音源信号を生成する.
6) 生成された LSP パラメータ系列と駆動音源信号を用いて LSP 合成を行い，合成音声を得る.

上記の 1)から 2)の過程には，本来高度な言語処理を必要とする．本研究では，規則音声合成の音響特性の制御に関わる 3)から 5)の過程に焦点をあてるため，テキストとしてカナ文章を入力するものとし, 形態素解析や読み付与といった高度な言語処理を省略した簡易システムとした.

3.3.2 VCV 素片選択法

前項で概観したLSPベクトルVCV規則音声合成方式において, 合成単位辞書には同一のVCV合成単位に属するVCV素片が多数収録されており, 同一の文章を作成する場合でも，可能な VCV 素片の組み合わせが多数存在する．高品質な合成音声を得るためには, 適切なVCV素片を選択し接続することが必要である. 本論文では, まずLSPベクトルVCV

規則音声合成方式のための VCV 素片の選択方法として以下の 2 つの選択法を提案した.

(i) VCV素片を収集した際の音韻環境と合成する文章中でのVCV素片の音韻環境の類似度を素片選択の基準にする方法(**PER 選択法**).

(ii) VCV 素片の接続部での接続歪みを素片選択の基準にする方法(**MLD 選択法**).

一般的な性質として，PER 選択法は合成単位辞書の記憶容量は大きくなり不利であるが，素片選択の速度が速い．一方 MLD 選択法は合成単位辞書の記憶容量は小さくでき有利であるが，素片選択の速度が遅い．第 4 章では，2 つの選択法の詳細について説明し，両者による合成音声の聞き取り実験による性能の比較検討を行った結果について述べる．

また，素片選択の速度が遅いという MLD 選択法の欠点を解決するために，距離テーブル参照法(Distance Table Look-up Method: **DTL 選択法**)を提案した．DTL 選択法では，べ

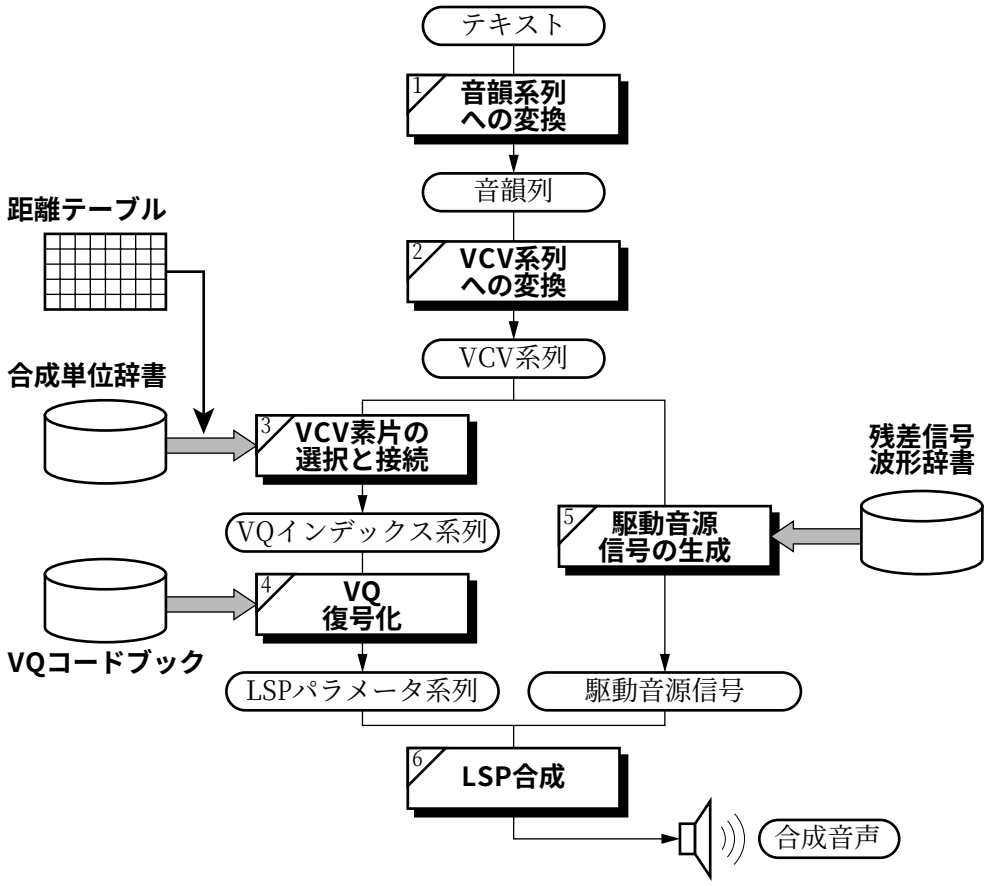


図 3.3　LSP ベクトル VCV 方式による音声合成システムのブロック図

クトル量子化の特長を生かし, VQコードブックの代表ベクトル間の距離を予め計算して作成した距離テーブル(distance table)を参照することで VCV 素片の接続歪みの計算を高速化する．この手法ついては，第 7 章で詳細に述べる．

3.3.3 駆動音源信号の生成

VCV素片を選択し接続することで生成されたLSPパラメータ系列から合成音声を得るための駆動音源は, 音韻毎に代表残差波形辞書に用意した代表残差波形を接続して生成する．代表残差波形として音韻の種類別に以下の波形を準備した.

(i)　撥音 / X / を含む各母音については残差信号の 1 ピッチ分の波形

(ii)　有声子音については後続母音別に残差信号の 1 ピッチ分の波形

(iii)　無声子音については, 後続母音別に無声子音部の残差信号をLSP分析のインターバル長の整数倍の単位で切り出した波形

これらは，VCV 素片を採取した音声資料から得られたLSP分析の残差信号から視察によって切り出した．図 3.4 に，母音 / e / の代表残差波形と，後続母音として / o / を持つ無声子音 / k(o) / の代表残差波形を示す．母音 / e / の代表残差波形は，ピッチ周期 7.26ms の長さで切り出している．無声子音 / k(o) / の代表残差波形は，LSP 分析の 4 インターバル分 23.22ms の長さで，/ k / の破裂点を中心にして切り出している．

音声合成時には，VCV 素片に付加された子音開始点と後部母音の開始点の情報に従って代表残差波形を接続する. 有声音の代表残差波形は, 所望のピッチ間隔で並べて接続する. この時, 代表残差波形の長さよりピッチ間隔が長い場合には代表残差波形の後に必要な長さだけ振幅0のデータを挿入する．また，ピッチ間隔が短い場合には代表残差波形をピッチ間隔に合わせて打ち切って接続する．無声子音の代表残差は，VCV 素片に付加された子音開始点の情報に従って適切な位置に挿入する．

代表残差信号によって音源を生成する手法は簡便だが, パルスと白色雑音を用いる単純な方法に比べると，自然性の上で品質の良い合成音声を得ることができる．本論文の第8

章では，特に子音の明瞭度を向上させるためのより高度な駆動音源生成法について述べる．

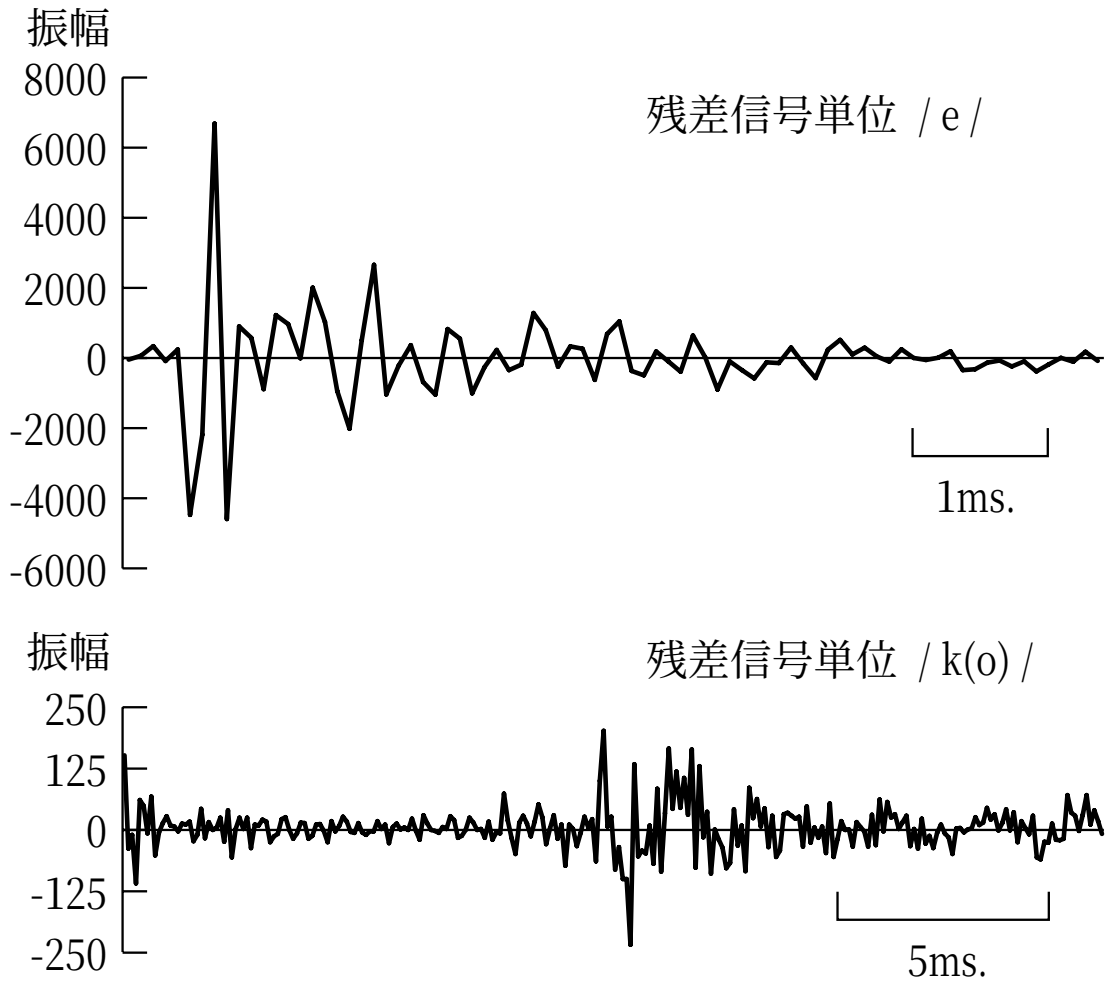


図 3.4　代表残差信号の例

3.3 まとめ

本章では, 以後の議論のために, まず本論文における音韻や音節の取り扱いについて説明した．LSP ベクトル VCV 規則音声合成方式では，日本語の言語学的な先見的知識に基づきながら，VCV 素片接続型の音声合成に適用しやすいように簡略化した音韻取り扱いルールを用いている．また，LSP ベクトル VCV 規則音声合成方式における，VCV 素片の符号化法について述べ, 合成単位辞書を小さな記憶容量に抑える原理について説明した.

続いて LSP ベクトル VCV 規則音声合成方式の構成について述べた．本研究では，規則音声合成の音響特性の制御に焦点をあてるため, 形態素解析や読み付与といった高度な言語処理を省略した簡易な音声合成システムを作成して実験を行っている. 従って, 以後の議論の中心となる主題は，合成単位辞書の規模と VCV 素片の選択法や駆動音源の生成法である.

第4章　合成単位辞書の規模と素片選択法の検討

本章では，本論文で提案したLSPベクトルVCV規則音声合成方式による音声合成システムの実現に向けて，合成単位辞書に収録する合成単位素片の適正な数とその選択方法を実験的に検証した結果を報告する．合成単位素片の選択法として，(i)本論文で定義した音韻環境類似度スコア(PERスコア)を用いて合成単位素片の音韻環境を最適化するPER選択法と，(ii)合成単位素片接続部における接続歪みを最小化するMLD選択法について検討した．

ニュースの読み上げ音声から採取した様々な音韻環境におけるVCV素片を合成単位素片として用い，上記の2つのVCV素片選択法によって作成した合成音声の品質を主観評価実験により評価した．音声合成に用いるVCV素片の数を増加させると，どちらのVCV素片選択法を用いても合成音声の品質は向上するが，品質の向上はVCV素片の数が14,000個程度になると飽和することが判明した．また，どちらのVCV素片選択法を用いても合成音声の品質には差がないことが判った．[34]

4.1 実験システムと素片選択法の詳細

4.1.1 評価実験の方針と音声合成システム

LSPベクトルVCV規則音声合成方式による音声合成の過程を模式的に図4.1に示す．本研究で構成した音声合成システムは言語処理部分を省略した簡便なシステムであるため，合成目的文からVCV単位の系列を生成するまでの処理は一種の文字列処理とみなしてよい．仮名文章として与えられた合成目的文は，第3章で述べた音韻取り扱いルールに従って音韻系列に変換される．さらに，音韻系列を母音を重複させる形でVCV単位に分割することにより，VCV系列に変換する．このとき得られるVCV系列は，合成目的文に対してユニークに定めることができる．

次の段階では，合成音声を得るために，VCV 単位の系列に従って合成単位辞書から VCV 素片を取り出し，接続する．しかし，図 4.1 に示すように合成単位辞書には同一の VCV 合成単位に属する VCV 素片が多数収録されているため, 同一の文章を作成する場合でも可能な VCV 素片の組み合わせが多数存在し, VCV 素片の取り出し方はユニークではない.

合成単位辞書に同一の VCV 合成単位に属する VCV 素片を多数収録する理由は, 人間の自然な調音特性を音声合成システムで模擬するためである. 人間が発話した自然な音声では，同じ音韻の組み合わせの VCV 素片であっても，いつも同じ音響特性を持つとは限らない．同一 VCV 単位に属する VCV 素片にも様々なバリエーションが生じる．この現象は，主として VCV 素片が前後の音韻からの調音結合の影響により様々な変型を受けることによって起こる. このような特性を反映して高品質な合成音声を得るためには, バラエ

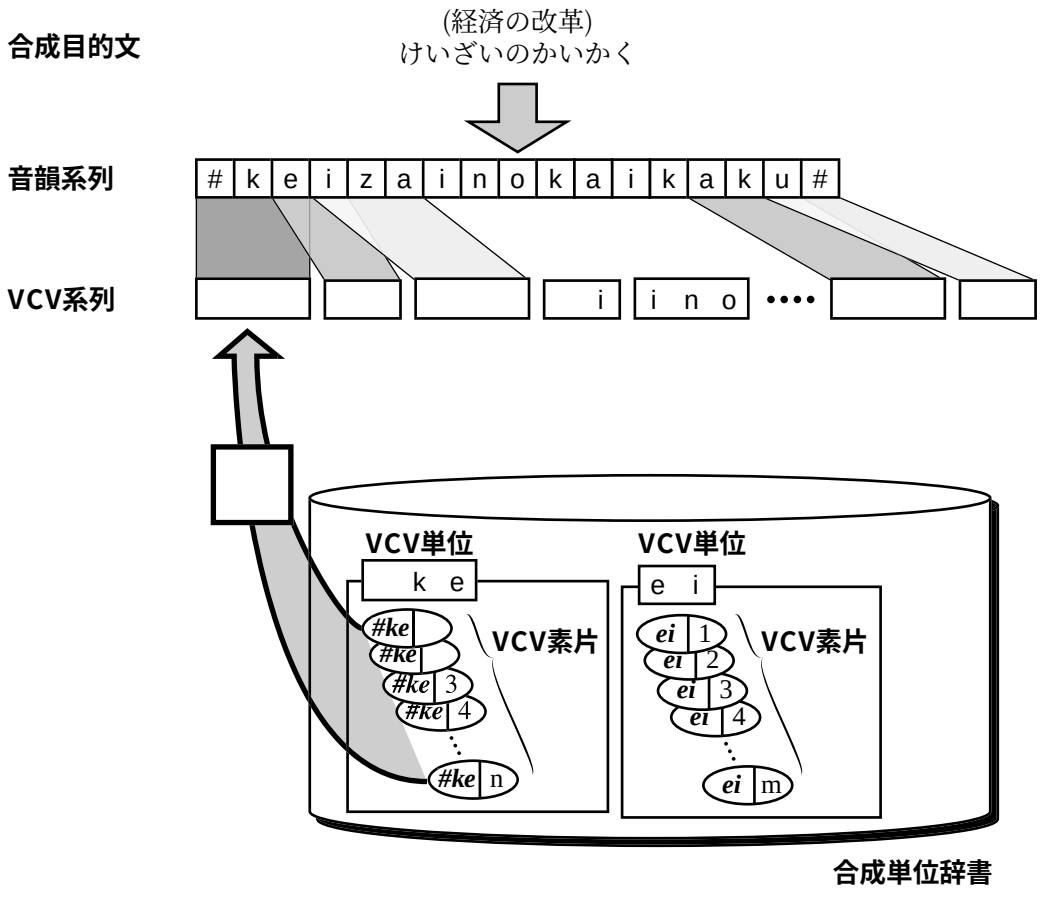


図 4.1　VCV 素片選択の概念図

ティに富んだ十分な数のVCV素片を合成単位辞書に収録する必要がある．また，同一のVCV合成単位に属する多くのVCV素片を保持することから，必然的にVCV素片を適切に選びだすという素片選択の問題が重要になる．

大規模な計算機資源が利用できる場合，合成音声の品質向上のためには十分なVCV素片のバリエーションを尽くせば良い．しかし，小規模な計算機資源によって音声合成システムを構成する場合，合成音声の品質を良好に保つ上で必要最低限のVCV素片数つまり合成単位辞書の規模について検証することは重要な課題となる．いずれの場合でもVCV素片数の選択方法について検証することは重要である．このとき，合成単位辞書の規模とVCV素片数の選択方法は深く関係しており，切り離しては考えられないであろう．また，本方式では合成単位辞書の規模をVCV素片の符号化の観点から小さくするために，VCV素片をLSPパラメータに対してベクトル量子化を適用する手法により符号化する．従って，ベクトル量子の量子化誤差による合成音声の品質劣化についても評価しなければならない．

これまでの議論をまとめると，LSPベクトルVCV規則音声合成方式を合成単位辞書の規模と合成音声の品質の観点から評価する場合，(i) 合成単位辞書に収録するVCV素片数，(ii) VCV素片選択法，(iii) ベクトル量子化の量子化誤差の3点が主要な検討課題となる．しかし，これら3点を同時に評価しようとすると，評価実験が煩雑かつ大規模になりすぎる．そこで，本論文では，以下の手順で議論をすすめる．

1) ベクトル量子化による量子化誤差が無い条件で，(i) 合成単位辞書に収録するVCV素片数，(ii) 素片選択方法についての検討を行う．
2) 1)の結論に従い，合成単位辞書の規模を固定して(iii) ベクトル量子化の量子化誤差の検討を行う．

この方針に従い，本章ではLSPベクトルVCV規則音声合成方式の実現に先立って，ベクトル量子化を行っていない合成単位辞書を用いて，合成単位辞書に収録するVCV素片数とその選択方法について実験的に検証した．従って，実験に用いた音声合成システムは，図4.2に示すように，LSPベクトルVCV規則音声合成方式でベクトル量子化に関わる

部分を省略し，合成単位辞書にLSPパラメータの形式でVCV素片を格納した音声合成システムである．VCV 素片の選択法として, (i)本章 4.1.2 項で定義する音韻環境類似度スコア(PER スコア)を用いて合成単位素片の音韻環境を最適化する PER 選択法と, (ii)合成単位素片接続部における接続歪みを最小化する MLD 選択法の 2 種類を検討した.

本章で展開する議論において, PER選択法による素片選択はベクトル量子化の有無を問わず同じ選択結果となるため, ベクトル量子化をしない条件での実験結果は, ほぼそのままベクトル量子化を行った場合にも適用できると考えられる．MLD 選択法では，ベクトル量子化によって生じる量子化誤差により, ベクトル量子化の有無によって素片の選択結果に違いが出る可能性がある．しかし，本方式において，LSPパラメータのベクトル量子化による合成音声の品質劣化が小さいことを予備実験により確かめている. これは, ベクトル量子化による誤差が, 最適な接続の場合の素片の接続歪みに比べて十分小さいことを

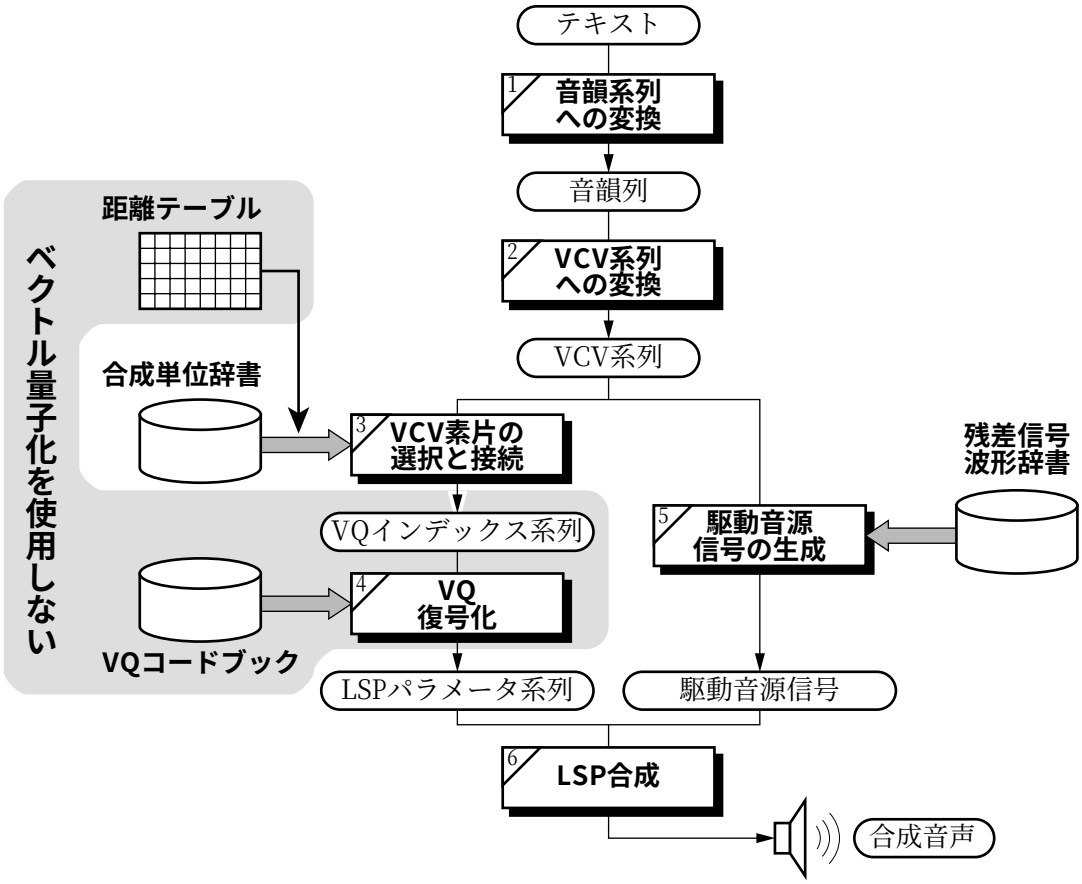


図 4.2　本章の実験に用いる音声合成システムのブロック図

示唆しており,ベクトル量子化を適用することにより選択法に関する実験結果に大きな影響を及ぼさないと考えられる.

4.1.2 音韻環境の類似度による VCV 素片の選択法

4.1.1節で述べた調音結合によるVCV素片の変型は, その前後に発話された音韻が何であったかを示す**音韻環境**に強い影響を受ける. このため, 合成単位素片の選択の際に音韻環境を考慮する方法が多数提案されている[22][23][41][42]. 多くの場合, 複雑な先験的知識によって音韻環境の適応度を算出する方法がとられているが, 本研究では簡便性を考慮して, 音韻環境の類似度を得点化する方法をとった.

本法では, 合成単位辞書に格納された全ての VCV 素片に対し, **音韻環境情報**として, 収集した際に前後にどのような音韻が発話されていたか前後5つずつの音韻名を記録しておく. 合成単位辞書には, 図 4.3 に模式的に示すように同じ VCV 単位に属する VCV 素片が複数存在するが, 付加されている音韻環境情報(前後の音韻の並び)は異なる. このようなVCV素片の中から, 合成したい目的文中でのVCV単位の音韻環境と最もよく一致するものを選び出すことが本法の主眼である.

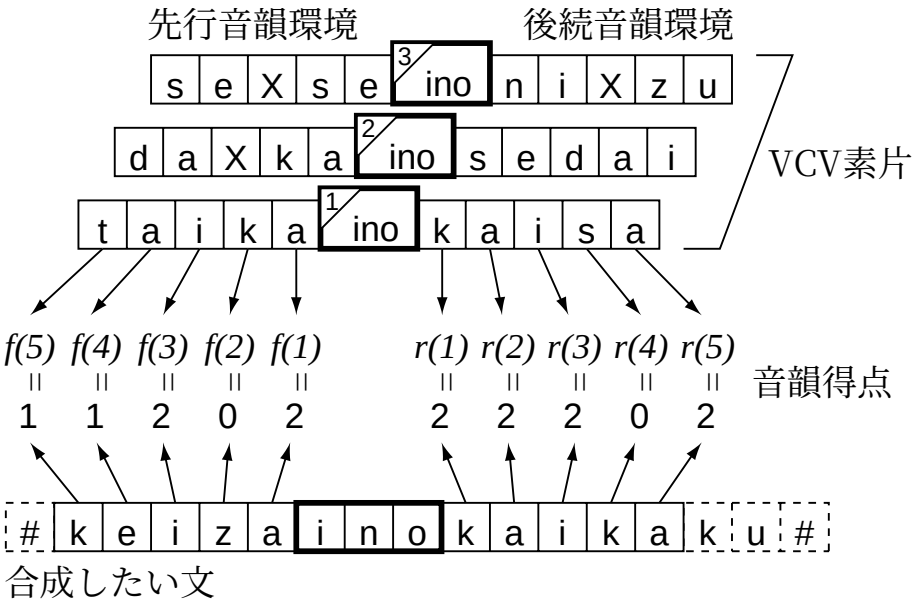


図 4.3　PER スコアによる音韻環境の得点化

VCV素片を収集した際の音韻環境と所望の文章中でのVCV単位の音韻環境の類似度を評価するために，図 4.3 に示すように VCV 素片の前後 5 つずつの音韻を考慮し，式(4.1)の**音韻環境類似度**(Phonemic Environmental Resemblance Score: **PER スコア**)を定義した．

$$PER = \frac{1}{2}\sum_{i=1}^{5}\frac{1}{3^i}\Big(f(i)+r(i)\Big) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (4.1)$$

ここで，$f(i)$ はVCV素片の先行する i 番目の音韻について，VCV素片を収集した際の音韻と合成する文章中での音韻の一致度を表す**音韻得点**である．$f(i)$ には，音韻が一致すれば2点，母音，摩擦子音，破裂子音等の**音韻種別**が一致すれば1点を与え，どちらも一致しない場合には0点を与える．$r(i)$ はVCV素片の後続する i 番目の音韻についての $f(i)$ と同様な得点である．音韻の種別については，表4.1のように言語学的な先見知識に基づいて定めた. 破裂子音や摩擦子音といった音韻分類が一致するものは, 有声／無声の別を考慮せずに，同一種別の音韻として扱った．このようにして求めた音韻得点 $f(i)$ と $r(i)$ をVCV素片の前後それぞれ5つの音韻について時間経過を考慮した重み付きで合計したものをPERスコアとしている．音韻得点の重みは，VCV素片により近い音韻得点の差が，

表 4.1　音韻種別の一覧表

音韻種別	音　　韻
母音・撥音	/ a /, / i /, / u /, / e /, / o /, / X /
無声化母音	/ A /, / I /, / U /, / E /, / O /
無声破裂子音	/ k /, / t /, / p /, / K /, / P /
有声破裂子音	/ g /, / d /, / b /, / G /, / B /
無声破擦子音	/ c /, / C /
有声破擦子音	/ Z /
無声摩擦子音	/ s /, / h /, / S /, / H /
有声摩擦子音	/ z /
鼻音	/ n /, / m /, / N /, / M /
流音	/ r /, / R /
拗音	/ j /, / w /
無音区間	#

より遠い音韻得点の差によって逆転されないように, 3の指数の逆数としている．音声合成時には，音韻環境の情報と所望の文章中でのVCV素片の音韻環境を用いて式(4.1)で定義したPERスコアを計算し，PERスコアが最大となるVCV素片を選択する．本選択法を以後**PER選択法**(Phonemic Environmental Resemblance method)と呼ぶ.

4.1.3 接続歪みの最小化によるVCV素片の選択法

単位素片接続型の規則音声合成において, 合成音声の品質を低下させる原因の一つとして, 素片の接続点における不連続性から生じる**接続歪み**があげられる．このため, 音韻環境と並んで, 素片選択において素片の接続歪みを最小化する基準が良く用いられる. 本研究では, VCV素片選択の基準として素片の接続部における接続歪みを評価するためにLSPパラメータの距離を用いる．図4.4に示すように，VCV素片の接続部における先行VCV素片の最終フレームのLSPパラメータが $\boldsymbol{\omega}^f=(\omega_1^f, \omega_2^f, \cdots, \omega_p^f)$，後続VCV素片の先頭フレームのLSPパラメータが $\boldsymbol{\omega}^r=(\omega_1^r, \omega_2^r, \cdots, \omega_p^r)$ であるとき，VCV素片の接続部における**LSPパラメータの距離**(LSP Distance: LD)を式(4.2)で定義する.

$$LD(\boldsymbol{\omega}^f, \boldsymbol{\omega}^r)=\sqrt{\sum_{i=1}^{p}\left(\omega_i^f-\omega_i^r\right)^2} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (4.2)$$

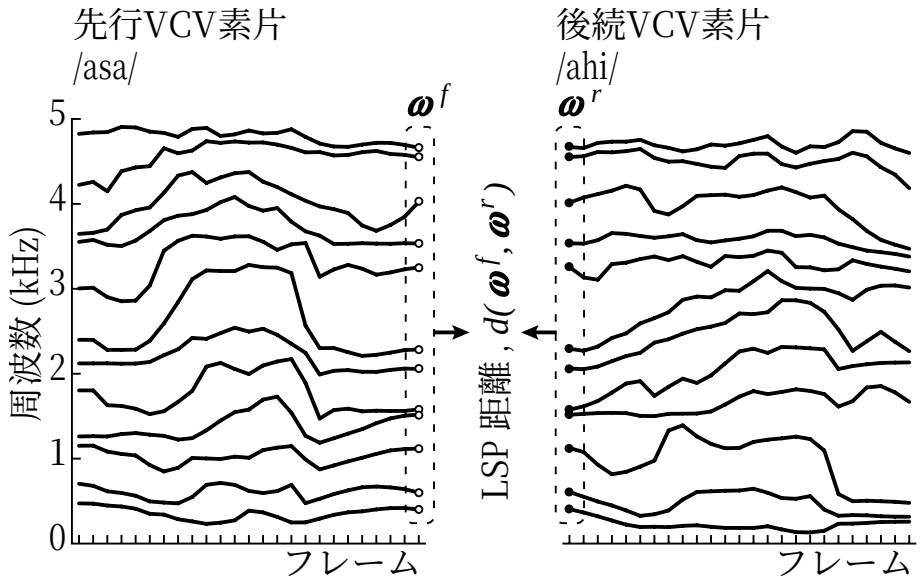


図4.4　VCV素片の接続点におけるLSP距離

接続歪みは，スペクトルの距離(スペクトル歪み)やケプストラム距離(ケプストラム歪み)など様々なパラメータによる評価が可能であるが，本方式ではVCV 素片をLSP パラメータの系列として扱うため，LSP パラメータの距離を用いることが実用的である．

式(4.2)で定義した $LD(\boldsymbol{\omega}^{f},\boldsymbol{\omega}^{r})$ を用いて，**ダイナミック・プログラミング(DP)** [43]の手法によりVCV 素片選択を行なう **LSP 距離最小化選択法**(Minimal LSP Distance method: MLD 選択法)の概念図を図4.5 に示す．同一のVCV 合成単位に属するVCV 素片が複数存在するため，図中において破線で示すようなVCV 素片の様々な接続が可能である．このように，VCV 素片の接続が可能であることを示す経路を接続可能経路と呼ぶ．MLD 選択法は，接続可能経路に式(4.2)で定義した $LD(\boldsymbol{\omega}^{f},\boldsymbol{\omega}^{r})$ をコストとして割り振り，合成音声の開始点から終了点までのコストの総和が最小となる経路をDPの手法によって探索する方法である．探索の結果得られたLSP 距離最小の経路上にあるVCV 素片を，最適なVCV 素片として選択する．

MLD 選択法では，式(4.2)の距離計算を多数回行ない，DP による膨大な探索を行う必要があり計算量がかかる．選択処理に要する計算量の点では，PER 選択法がMLD 選択法

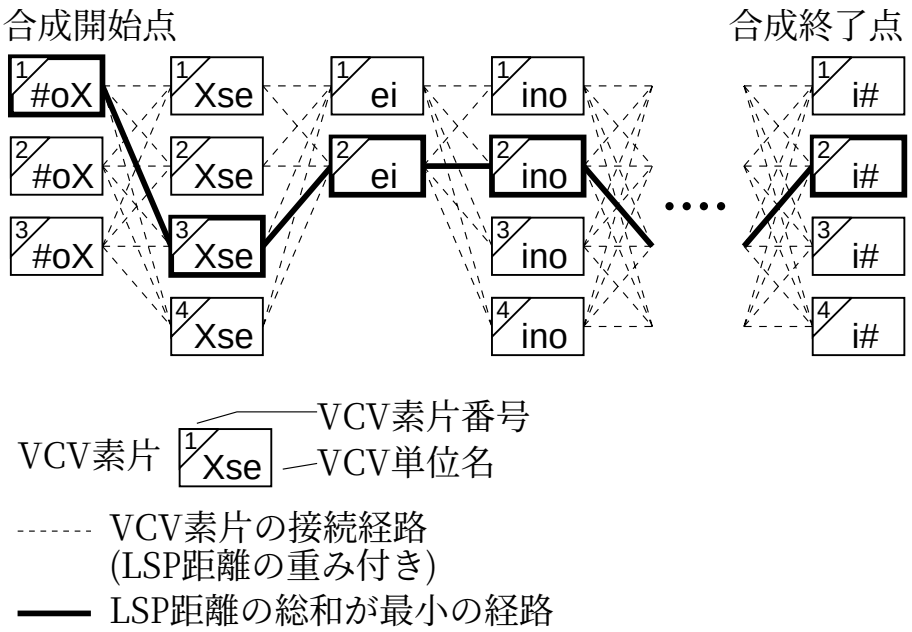


図 4.5　MLD 選択法による VCV 素片の選択

に比べて有利である．一方，MLD選択法では，合成単位辞書中のVCV素片データに音韻環境情報などの付加的な情報を加えておく必要がない. 従って, 合成単位辞書の記憶容量の点ではMLD選択法がPERスコアによる選択法に比べて優れている．

4.2 VCV素片の選択実験　－客観評価

4.2.1 合成単位辞書の作成と合成目的文

合成単位辞書に収録するVCV素片の適正な数の検証と，VCV素片の選択法の評価のために，PER選択法またはMLD選択法を用い，合成単位辞書の大きさを変えてVCV素片を選択する実験を行った．実験に用いたVCV素片は，3.1.3項で述べた男性アナウンサの発話によるニュース音声を収録した約70分の音声資料から収集した．

本実験では，VCV素片の収集に用いる音声資料の長さを変えることにより，合成単位辞書の大きさを変えた．具体的には，10分から10分きざみで70分までの長さの音声資料からVCV素片を収集して作成した7種類の大きさの合成単位辞書を作成した. 以後, 各々の合成単位辞書を「合成単位辞書(10)」～「合成単位辞書(70)」のようにVCV素片を収集した音声資料の長さを付して区別する．各々の合成単位辞書が収録しているVCV素片の数は，表4.2にVCV素片の総数として示した通り，約4,000個から26,500個である．

音声合成の対象である合成目的文として, 見出しを除く新聞記事の本文を用いた. 実験に用いた新聞記事の長さは，VCV合成単位の個数にして45,269個分の長さである．

4.2.2 評価指標

自然発話の音声資料から収集したVCV素片によって，音韻の組み合わせで可能な全てのVCV合成単位を網羅することは困難である．音声合成時に，合成単位辞書に収録されていないVCV合成単位が必要になった場合には，後部のCVの部分だけが一致する他のVCV合成単位の素片からCV素片を作成して代用する．この際，前部のV(母音)部分は補間によって作成する．一般的には，このようなVCV単位の代用，置換が起こると，合成音声の品質は低下する．従って，合成単位辞書は，出来る限りVCV合成単位を網羅す

表 4.2　音声資料の長さと採取された VCV 単位の種類数，VCV 素片の個数

音声資料の長さ	VCV型		VV型		#CV型		#V型		V#型		総種類	素片総数
	種類	平均個数	種類	平均個数	種類	平均個数	種類	平均個数	種類	平均個数		
(分)	(種)	(個)	(種)	(個)	(種)	(個)	(種)	(個)	(種)	(個)	(種)	(個)
10	341	7.7	29	19.3	52	6.5	5	20.8	6	73.5	433	4,050
20	400	12.5	31	36.7	64	9.7	5	35.8	6	134.0	506	7,759
30	429	17.2	35	48.9	65	11.0	5	41.6	6	152.8	540	10,926
40	446	21.4	35	61.4	67	11.6	5	48.2	6	169.1	559	13,713
50	455	27.4	35	81.2	73	15.6	5	68.6	6	246.5	574	18,271
60	466	32.7	35	98.2	76	20.2	5	88.6	6	328.5	588	22,622
70	470	38.4	35	177.1	78	21.5	5	101.2	6	362.7	594	26,517
可能な種類	570		35		95		5		6		711	

注）#は無音を表しており，#CV型と#V型は発話開始点に，V#型は発話終了点に用いる

べきであり，音声合成時に VCV 単位の置換が起こらないようにしなくてはならない．

上記の議論の観点から合成単位辞書に収録されたVCV合成単位の数を評価するために，**VCV単位網羅率**と**VCV素片置換率**を次のように定義する．VCV単位網羅率は，音韻の組み合わせで可能な VCV 合成単位の総数を N，VCV 素片の収集で得られた VCV 合成単位の数を n として，以下のように定義する．

VCV 単位網羅率：　$\gamma = n / N$　…… (4.3)

VCV素片置換率は，合成目的文中に含まれるVCV合成単位の総数をM，そのうちで合成の際に CV 素片に置換された VCV 素片の数を m として，以下のように定義する．

VCV 素片置換率：　$\rho = m / M$　…… (4.4)

本実験では，音韻の組み合わせで可能な VCV 合成単位の総数 N = 711，合成目的文中に含まれる VCV 合成単位の総数 M = 45,269 である．ここで，VCV 単位網羅率と VCV 素片置換率が合成単位辞書の規模についてだけ評価を行う指標であり，VCV 選択法によらない指標であることに注意が必要である．これらの指標は，合成単位辞書に登録されているVCV単位の種類と合成目的分に含まれるVCV単位の関係だけで決まり，素片選択方法に関係しない．

PER 選択法と MLD 選択法による素片選択の結果を評価するために，全ての VCV 素片選択結果について，(i)選択結果の VCV 素片系列の平均 PER スコアと，(ii)選択結果の VCV 素片系列中の VCV 素片の接続部での平均 LSP 距離の 2 つの指標を求めて，両者の関係を調べた．

4.2.3 実験結果

図 4.6 に，実験の結果求まった合成単位辞書の規模と VCV 単位網羅率，VCV 素片置換率の関係を示す．合成単位辞書の規模を大きくすると VCV 単位網羅率は向上するが，合成単位辞書のVCV素片の収録数が10,000個から15,000個程度から向上率は徐々に低下する．しかし，VCV 単位網羅率は，約 26,500 個の VCV 素片を収録した「合成単位辞書(70)」でも 83.5%であり，音韻組み合わせで可能なVCV合成単位の16.5%が収録されていない．

これは，3章の3.1.2項で述べたVCV素片収集の方針に従い，「日本語における自然な偏りを持った普通発話の資料」からVCV素片を収集したために，日本語の文章に現れにくいVCV単位が音声資料中に含まれなかったためである．

一方，合成単位辞書の規模を大きくするとVCV素片置換率は急激に減少し，合成単位辞書のVCV素片の収録数が14,000個以上ではあまり変化がない．約14,000個のVCV素片を収録した「合成単位辞書(40)」を用いた場合，VCV素片置換率は1.7%以下で非常に小さい値である．

「合成単位辞書(40)」を用いた場合，VCV単位網羅率は78.9%と低く，音韻組み合わせで可能なVCV合成単位のうち約2割が合成単位辞書に含まれていない．それにもかかわらず，VCV素片選択実験の結果ではVCV素片置換率は1.7%以下であり，合成単位辞書に収録されていないVCV単位が使用されたのは2%未満であることが示された．この結果は，VCV素片の収集にもれたVCV合成単位が音声合成時に使用される頻度は非常に小さいことを示しており，合成単位辞書のVCV素片収録数が14,000個以上の場合はVCV単位網羅率の低さが合成音声の品質低下に与える影響はごく小さいと考えられる．このことは，本研究におけるVCV素片収集の方針が適切であったことをも示唆している．

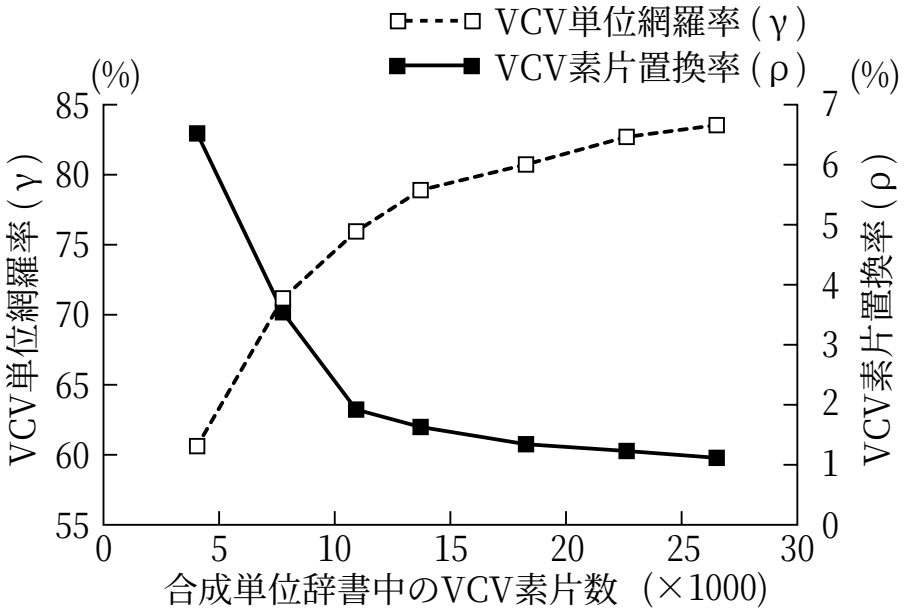
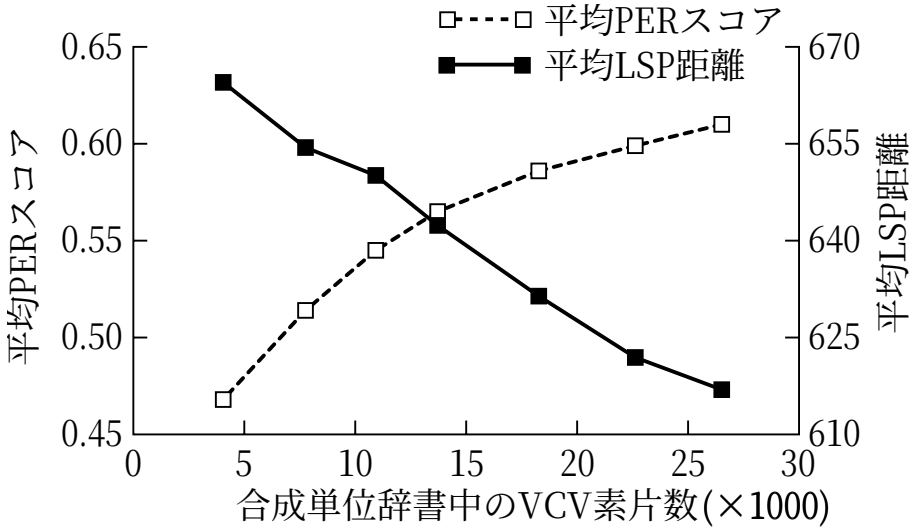
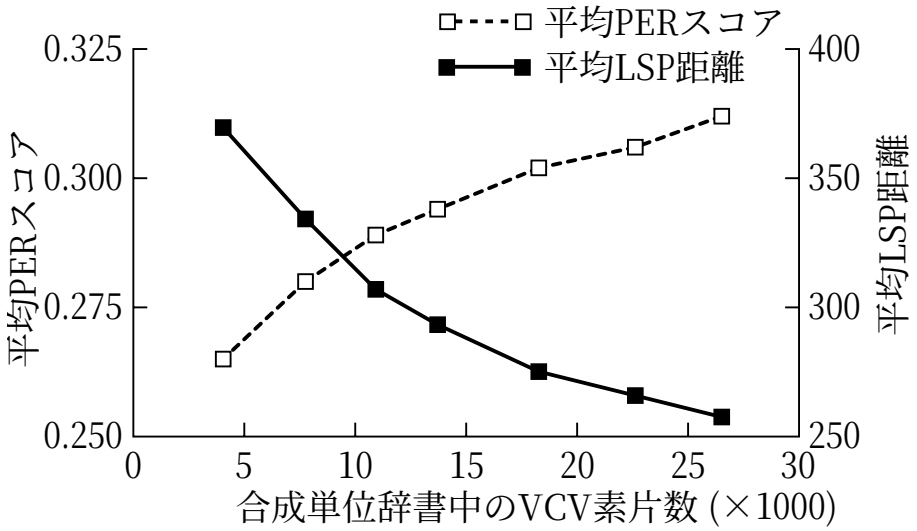


図 4.6　合成単位辞書の規模と VCV 合成単位網羅率，VCV 素片置換率の関係

PERスコアによる選択法とMLD選択法で選択されたVCV素片について，平均PERスコアと平均LSP距離を求めた結果を図4.7に示す．PERスコアによる選択方法を用いた場合，合成単位辞書の規模が大きくなると平均PERスコアは上昇し平均LSP距離が減少した．LSP距離最小選択化法を用いた場合，合成単位辞書の規模が大きくなると平均LSP距離は減少し平均PERスコアは上昇した．



a) PER選択法によるVCV選択



b) MLD選択法によるVCV選択

図 4.7　合成単位辞書の規模と
選択結果における平均 PER スコア，平均 LSP 距離

この結果は, PERスコアによる選択方法はVCVの接続歪みを小さく抑える傾向があり, LSP距離最小選択化法はPERスコアの高いVCV素片を選択する傾向があることを示している．しかし，2つの選択法で，平均PERスコアと平均LSP距離の絶対値は一致していない．両選択法の選択基準の関係については第 5 章で，より詳しく検討する．

4.3 合成音声の主観評価

4.3.1 合成単位辞書の大きさと合成音声の品質

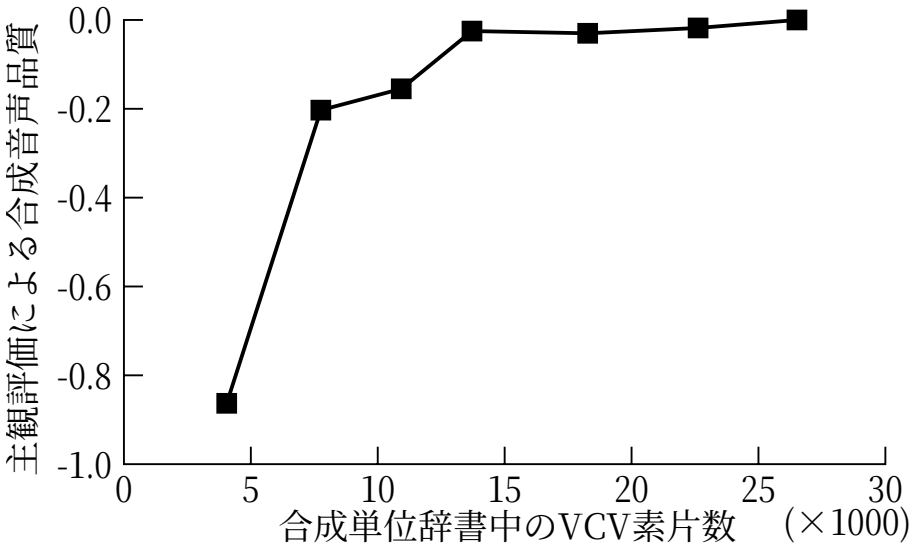
前節で述べた実験により，VCV単位網羅率とVCV素片置換率による評価の上では，約14,000 個の VCV 素片を収録した「合成単位辞書(40)」を用いれば十分であることが示された．しかし，音声合成システムの最終的な評価は，合成音声の聴覚的な品質評価によらなければならない．合成単位辞書に収録するVCV素片の適正な数を，合成音声の聴感上の品質の点から検証するために, 合成単位辞書の大きさをかえて合成した1対の合成音声のうち「どちらの合成音声が聞き取りやすいか」の判定を行う**1対比較**による**主観評価実験**を行った．

実験には，3 秒程度の 4 つの短文について，「合成単位辞書(10)」から「合成単位辞書(70)」を用いた7種類の合成音声を合成して用いた．被験者には，7回の練習比較の後，合成単位辞書が異なる 7 種類の合成音声の組み合わせ 21 対について順序の入れ替えを含めて 8 回ずつ 168 回の 1 対比較を課した．比較対の提示順はランダムとし，練習比較については，それが練習であることを被験者に知らせていない．

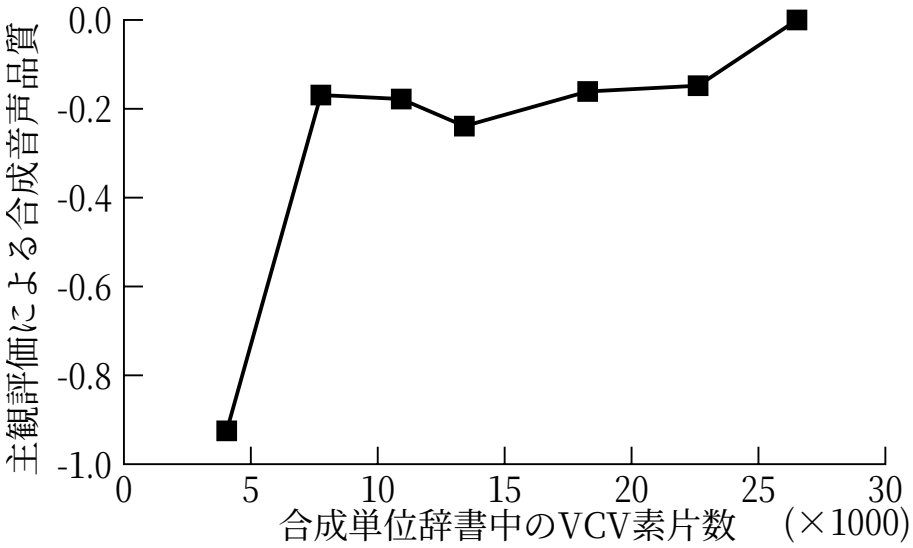
上記の 1 対比較実験を，PERスコアによる選択法とMLD選択法による合成音声について行った．PERスコアによる選択法について被験者は健康な 20 代の男女 11 名，MLD 選択法について被験者は健康な 20 代の男女 10 名で実験を行った．1 対比較実験で得られた判定結果から，**Thurstone の比較判定の法則**[44]を用いて，「合成単位辞書(70)」による合成音声を基準として，合成音声の**品質尺度値**を求めた．このとき，練習比較のデータは，合成音声の品質尺度値の算出には用いていない．

実験の結果として，合成単位辞書のVCV素片の収録数と合成音声の品質尺度値の関係

を図4.8に示す．PERスコアによる選択法を用いた場合，合成単位辞書に収録するVCV素片を14,000個以上に増やしても，合成音声の品質尺度値は向上していない．また，MLD選択法を用いた場合，合成単位辞書に収録するVCV素片を8,000個以上に増やしても，合成音声の品質尺度値は向上していない．ここに述べた2つのVCV素片選択法を用いる場合，多くても14,000個程度のVCV素片を収録した合成単位辞書を用いて音声合成システ



a) PER選択法によるVCV選択



b) MLD選択法によるVCV選択

図 4.8　主観評価による合成音声の品質

ムを構築すれば良いといえる．この結果は，前節で述べたVCV単位網羅率とVCV素片置換率による評価の結果と一致する．

4.3.2 VCV 素片選択法の比較

PERスコアによる選択法とMLD選択法による合成音声の品質の比較のために，1対比較による主観評価実験を行った．実験には，3秒程度の4つの短文について，「合成単位辞書(70)」を用いて，PERスコアによる選択法とMLD選択法によって合成した合成音声を用いた．被験者には，10回の練習比較の後，20回の1対比較を課した．練習比較についてはそれが練習であることを被験者に知らせていない．被験者には，比較対の「どちらの合成音声が聞き取りやすいか」を「同程度である」という評価を許して判定させた．被験者は健康な20代の男女11名である．1対比較実験でより聞き取りやすいと判定された合成音声に2点，他方に0点を与え，同程度と判定された場合には両方の合成音声に1点ずつを与えて，被験者の判定結果を得点化した．

上記の実験の結果，図4.9に示すようにPERスコアによる選択法を用いた合成音声の得点率は53.1%，MLD選択法を用いた合成音声は46.9%となった．両者の得点について，両側二項検定を行った結果，有意水準5%で有意な差はみられなかった．このことから，PERスコアによる選択法による合成音声とMLD選択法による合成音声の品質には聴感上の差はないことが判明した．

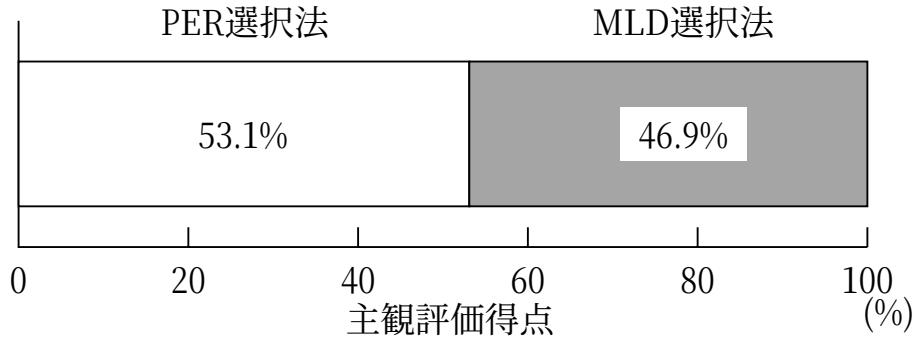


図 4.9　主観評価による PER 選択法と MLD 選択法の合成音声の比較

4.4 まとめ

本章では，LSPベクトルVCV規則音声合成方式の実現のために，合成単位辞書に収録する合成単位素片の適正な数とその選択方法を実験的に検証した. 新聞記事を用いたVCV素片の選択実験では，26,500個のVCV素片を収録した合成単位辞書を用いた場合でもVCV単位網羅率は83.5%と低い値であった．しかし，14,000個以上のVCV素片を収録した合成単位辞書を用いた場合, VCV素片置換率はほぼ一定の1%程度と非常に小さい値となった．この結果は，VCV素片の収集にもれたVCV合成単位が音声合成時に使用される頻度は非常に小さいことを示しており，合成単位辞書のVCV素片収録数が14,000個以上の場合はVCV単位網羅率の低さが合成音声の品質低下に与える影響はごく小さいと考えられる．

また，主観評価実験では，14,000個以上のVCV素片を収録した合成単位辞書を用いた合成音声には，聴感上の品質の差がなかった．これより，合成単位辞書に収録するVCV素片の数を増加させると合成音声の品質は向上するが，品質の向上はVCV素片の数が14,000個程度になると飽和することが判明した．この結果は，VCV素片置換率による評価と一致する．

PER選択法とMLD選択法の比較において，VCV素片の選択実験では，両選択法による選択結果について平均PERスコアと平均LSP距離は一致していない．しかし，主観評価実験では，PER選択法を用いた合成音声の得点率は53.1%，MLD選択法による合成音声は46.9%となり，両側二項検定を行った結果，有意水準5%で有意な差はなかった．これは，PER選択法とMLD選択法では，選択されるVCV素片は一致しないが，聴感上の品質は同程度であることを示している．

PER選択法は処理速度は速いが合成単位素片の音韻環境を記憶しておく必要があるために合成単位辞書の記憶容量が増える. 一方, MLD選択法はDPの手法を使用するため速度は遅いが, 合成単位辞書の記憶容量は小さい. 両者とも合成音声の品質に大きな差がないことが判明したため，記憶容量の削減を重要な目標とするLSPベクトルVCV規則音声合成方式では，MLD選択法を用いてVCV素片の選択を行えば良いことが明らかになった．

第 4 章　合成単位辞書の規模と素片選択法の検討

従来提案されてきた音韻環境を考慮する方法[16][20][28][29]は, 多くの場合, 複雑な先験的知識によって音韻環境の適応度を算出する方法がとられている. 本研究で述べたPER選択法は簡便性を考慮して，音韻環境の類似度を得点化する方法をとった．実験の結果は, 本研究で扱った合成単位辞書の規模では, このような簡便な方法でも素片選択の基準としては十分であることを示している．一方，MLD 選択法と同様な手法は，他の研究でも多く用いられている[20]．しかし，音韻環境基準と接続歪みによる基準の関係は明らかではない．本章では，PER 選択法と MLD 選択法を比較することで，VCV 素片選択における，音韻環境基準と接続歪みによる基準が，異なる選択結果をもたらすが，合成音声品質には聴感上の差がないことを示した．

第5章　VCV規則音声合成における音韻環境指標と接続歪み指標の関係

前章では，LSPベクトルVCV規則音声合成方式の実現に向けて，合成単位辞書に収録する合成単位素片の適正な数とその選択方法を実験的に検証した. 合成単位素片の選択法としてPER選択法とMLD選択法を提案し，両者の選択結果を合成単位辞書の規模との関係で評価した. 合成音声の主観評価実験の結果, 両者による選択結果は一般には一致しないが，合成音声には聴感上の品質の差がないことを示した.

上記の結果について，筆者は，PER選択法における素片選択基準(音韻環境類似度の最大化)とMLD選択法における選択基準(合成単位素片接続部における接続歪みの最小化)の間には, 一方を改善すれば他方も改善されるといった強い関係があると考えている. しかし, 全く異なる2つの素片選択基準を合成音声の品質の面からのみ比較するだけでは十分とは言えない．両選択基準の間にはそれ程の関係がなく，合成する個々の文については, その文の性質によって2つの素片選択基準で合成音声の品質に差があっても, 平均的には両者が同品質になることも十分ありうる.

本章では, 2つの素片選択基準に強い関係があることを実証するために，音韻環境類似度と合成単位素片接続部における接続歪みの間の関係を実験的に調べた. 実験の結果, 一方の選択基準を最適化するようにVCV素片を選択すると，他方の基準で見ても非常に良い選択が行なわれることを，評価指標値の平面上で明確に示すことができた. [36]

5.1　素片選択方法の比較のための指標

5.1.1　PER選択法とMLD選択法

本章で議論する音声合成システムは, 第4章で議論したものと同じである. 音声合成システムの詳細は省略するが，VCV素片の選択法については簡単に説明しておく.

1) PER 選択法

合成素片の選択の際に音韻環境を考慮する方法は多数提案されているが, 多くの場合, 複雑な先験的知識によって音韻環境の適応度を算出する方法がとられている．本研究では, 小規模な計算機資源で実現できることを目標とし, 音韻環境の類似度を得点化する簡便な方法をとった. VCV素片を収集した際の音韻環境と所望の文章中でのVCV素片の音韻環境の類似度を評価するために，式((5.1)の音韻環境類似度(Phonemic Environmental Resemblance Score: PER スコア)を定義した.

$$PER = \frac{1}{2}\sum_{i=1}^{5}\frac{1}{3^i}\Big(f(i)+r(i)\Big) \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (5.1)$$

ここで，$f(i)$ はVCV 素片の先行する i 番目の音韻について，VCV 素片を収集した際の音韻と合成する文章中での音韻の一致度を表す音韻得点である. $f(i)$ には, 音韻が一致すれば2点，母音，摩擦子音，破裂子音等の音韻種別が一致すれば1点を与え，どちらも一致しない場合には0点を与える. $r(i)$ はVCV素片の後続する i番目の音韻についての $f(i)$ と同様な得点である．音韻種別などの詳細については，第3章を参照されたい．PER選択法では，VCV 単位毎に(5.1)式で定義したPERスコアを計算し，その値が最大となるVCV素片を選択する.

2) MLD 選択法

VCV 素片の接続部における接続歪みを評価するために，LSP パラメータの距離を用いる．VCV 素片の接続部における先行 VCV 素片の最終フレームの LSP パラメータが $\boldsymbol{\omega}^f=(\omega_1^f,\omega_2^f,\cdots,\omega_p^f)$，後続 VCV 素片の先頭フレームの LSP パラメータが $\boldsymbol{\omega}^r=(\omega_1^r,\omega_2^r,\cdots,\omega_p^r)$ であるとき，VCV 素片の接続部における LSP パラメータの自乗距離 (LSP distance: LD)を(5.2)式で定義する.

$$LD(\boldsymbol{\omega}^f,\boldsymbol{\omega}^r)=\sqrt{\sum_{i=1}^{p}\left(\omega_i^f-\omega_i^r\right)^2} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (5.2)$$

式(5.2)で定義した $LD(\boldsymbol{\omega}^f,\boldsymbol{\omega}^r)$を選択可能な探索経路上のコストとして，DP の手法により経路探索を行うことで素片選択を行う手法を，LSP 距離最小化選択法(minimal LSP distance method: MLD 選択法)と呼ぶ.

5.1.2　音韻環境指標と接続歪み指標

VCV規則音声合成方式において，ある文章を合成しようとした場合の合成単位辞書からのVCV素片選択は，図5.1に示すようにVCV素片をノードとする経路選択とみなすことができる．図5.1では，1つのVCV単位に属するVCV素片を数個程度として簡単に模式化してある．しかし，実際には同じVCV単位に属するVCV素片は平均で25個程度，多いものでは数百に達する．このため，含まれるVCV単位が少なく比較的短い文でも，VCV素片選択の際に可能な経路は膨大な数にのぼる．

可能な経路の中から適当にある経路を1つ選択すると，それがどのような選択方法に

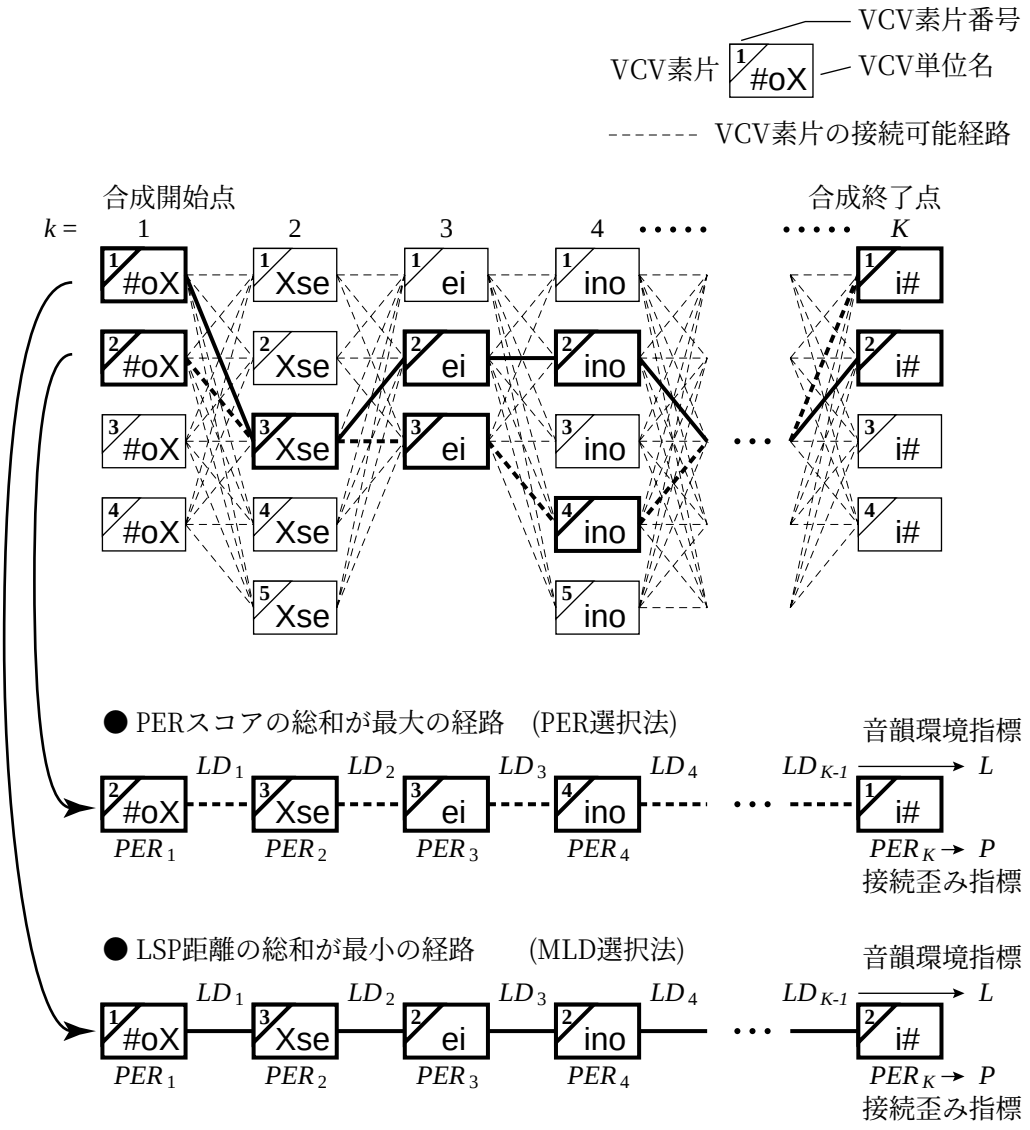


図 5.1　VCV 素片選択経路と指標の計算

よって選択された経路でも，選択した経路上の全てのVCV素片にPERスコアを付けることができる. 合成する文章の長さによらず, 選択した経路の良し悪しを音韻環境の適合度によって評価するための指標として，PERスコアを経路中に含まれるVCV素片について平均したものを考える．以後，この指標を，選択された経路に対する**音韻環境指標**(phonemic environment index)と呼ぶ．合成目的文中に K 個の VCV 単位が含まれており，その k 番目に相当する VCV 素片の PER スコアを PER_k として，音韻環境指標 P を，式(5.3)のように定義する．

$$P = \frac{1}{K} \sum_{k=1}^{K} PER_k \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (5.3)$$

同様に, いかなる方法で選択された経路でも, 経路上のVCV素片の全ての接続点でLSP自乗距離を計算することができる. 合成する文章の長さによらず, 選択した経路の良し悪しをVCV素片の接続歪みの大小によって評価するため, VCV素片の選択経路上の各接続点でのLSP自乗距離を平均したものを考える. 以後, この指標を, **接続歪み指標**(connective distortion index)と呼ぶ．合成目的文中に K 個の VCV 単位が含まれており，その k 番目と k +1 番目の VCV 素片の接続点における LSP 自乗距離が LD_k であるとき，接続歪み指標 L を，式(5.4)のように定義する．

$$L = \frac{1}{K-1} \sum_{k=1}^{K-1} LD_k \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (5.4)$$

先に述べた PER 選択法の最適選択は，音韻環境指標を最大にする選択である．また，MLD 選択法の最適選択は，接続歪み指標を最小にする選択である．

5.2　音韻環境指標と接続歪み指標の分布

5.2.1　実験の目的

本章では，PER 選択法と MLD 選択法による VCV 素片選択の結果を比較し，両者の関係を調べることを目的としている．この目的のために，5.1 節では，音韻環境指標と接続歪み指標を定義した. 具体的には, 次の点を素片選択実験により検証し, 両選択法の比較を行う.

1) PER選択法では音韻環境指標を最大化するようなVCV素片選択がなされるが, その選択結果について接続歪み指標がどのような値をとるか.

2) MLD選択法では接続歪み指標を最小化するようなVCV素片選択がなされるが, その選択結果について音韻環境指標がどのような値をとるか.

このような実験に先立ち, 本節では選択法の評価に用いようとしている音韻環境指標と接続歪み指標の性質について論じておく.

先にも述べたように, 合成単位辞書中に同一のVCV単位に属するVCV素片が多数存在するため, 一つの合成目的文についても VCV 素片選択における可能な選択経路の数は膨大である. その全ての可能な選択経路に対して, 先に定義した音韻環境指標Pと接続歪み指標Lを計算することができる. もし, 定義した指標の上で可能な選択経路が示す分布形がいびつな形をしていれば, 選択基準としての指標自体を見直す必要がある. 本節では, 膨大な数の可能な選択経路の中からランダム･サンプリングした経路を用いて, 可能な選択経路が示す分布形を推定する実験を行った.

5.2.2　音韻環境指標と接続歪み指標の分布の例

音声資料

実験に用いるVCV単位辞書の作成のためのVCV素片の収集には, 標準的な日本語と考えられる NHK アナウンサの発話を用いた．VCV 素片の収集は以下の手順で行なった.

1) FMラジオ･ニュースから切り出した同一の男性アナウンサの40分間の発話部分を表 5.1 に示す条件で，サンプリングした.

2) 1)の音声資料に，視察で音韻マーキングを行なった．

3) 母音の中間点で切り出す方法で自動的に VCV 素片を生成した．

4) 各 VCV 素片に対し，表 5.1 の仕様で LSP 分析を行なった．また，収集された VCV 素片には，子音開始点と後部母音の開始点を示す情報を付加した．

合成単位辞書に収録した VCV 素片の種類と個数について，表 5.2 にまとめた．実験に用いた合成単位辞書の規模は，第 4 章の実験結果に従って決定した．

表 5.1　音声資料の分析条件

サンプリング条件	
標本化周波数	11.025kHz
量子化数	16bits
LSP分析条件	
分析次数	14次
フレーム長	256点
インターバル長	64点

表 5.2　合成単位辞書に収録した VCV 単位の種類と VCV 素片数

合成単位の型	合成単位の種類数	収録素片の平均個数
VCV型	446(570)	21.4
VV型	35(35)	61.4
#CV型	67(95)	11.6
#V型	5(5)	48.2
V#型	6(6)	169.1
合　計	559(711)	13,713

注）#は無音を表している．
合成単位の種類数のカッコ内は，論理的に可能な種類数である．
収録素片の平均個数の合計欄は素片の総数である．

VCV 素片のランダム選択

先に述べたように，VCV 素片選択における可能な選択経路の数は膨大である．全ての経路について，音韻環境指標*P*と接続歪み指標*L*を調べることは不可能なので，全ての経路の中からランダムに取り出した1,000個の経路について音韻環境指標*P*と接続歪み指標*L*を調べた．実験は，新聞記事から次の条件を満たす 100 文を採取して行なった．

1) 文の長さが VCV 単位の数にして 10 から 30 程度であること

2) 文に使用されている全ての VCV 単位について，VCV 単位辞書に 5 個以上の VCV 素片が登録されていること

1)の条件は，人間が一息で発話できる範囲の文を実験対象としたためである．2)の条件は，文中に極端に選択肢の少ない VCV 単位が含まれると 2 つの素片選択方式の選択結果が一致しやすくなり，実験結果が信頼できなくなることをさけるためである．

ランダムに選択した経路における音韻環境指標 *P* の分布と接続歪み指標 *L* の分布の一例として，文「生産改革の動きは国内メーカーにとどまらない.」の例を以下の図にあげる．図 5.2 にランダム選択した場合の音韻環境指標 *P* の分布を，図 5.3 には接続歪み指標 *L* の分布をヒストグラムとして示している．両者とも，実線の曲線は分布に正規分布をあ

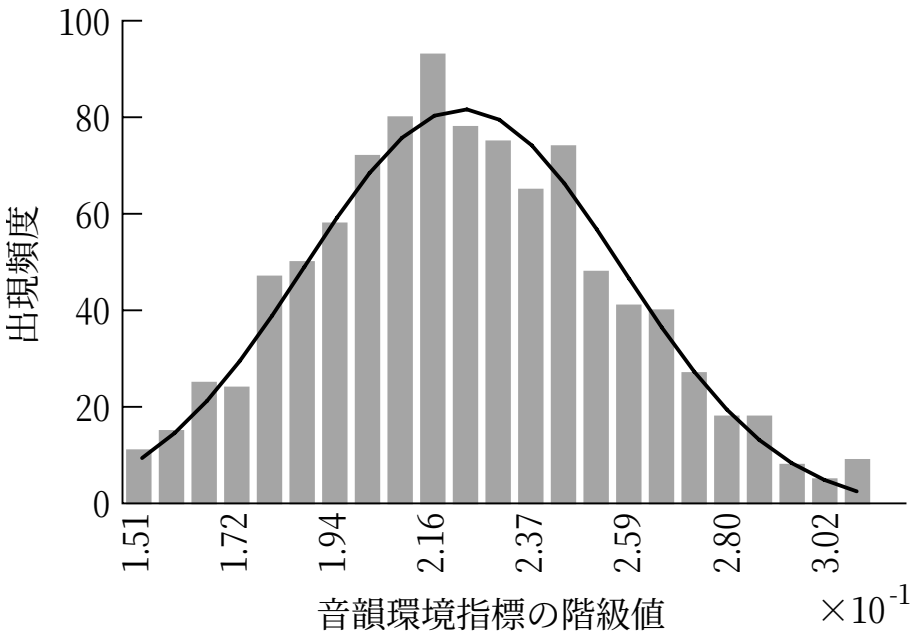


図 5.2　ランダム選択における音韻環境指標の分布例

てはめた場合の理論値を示し, 棒グラフはランダムに選択した経路のデータから得たヒストグラムを示している．図 5.2，図 5.3 ともに，正規分布と非常によく一致している．

5.2.3　音韻環境指標と接続歪み指標の分布の正規分布への適合度

100 文の実験データ全てについて, 音韻環境指標 *P* の分布に正規分布をあてはめた場合の**適合度検定**を次の方法で行なった.

1) 全ての可能な選択経路からランダムに 1,000 個の経路を選びだす. 以後これを**ランダム接続経路**と呼ぶ.

2) 音韻環境指標 *P* のヒストグラムを作成

2-1) ランダム接続経路のうち, 音韻環境指標 *P* について, 下位 5 位の値と上位 5 位の値の範囲をヒストグラムの値の範囲とする.

2-2) ヒストグラムの値の範囲を 25 等分してカテゴリを別け，階級値を定める

2-3) ヒストグラムの各カテゴリについて度数が 25 より小さい場合には, 隣のカテゴリと合併する．合併後のカテゴリの総数を *M* とする

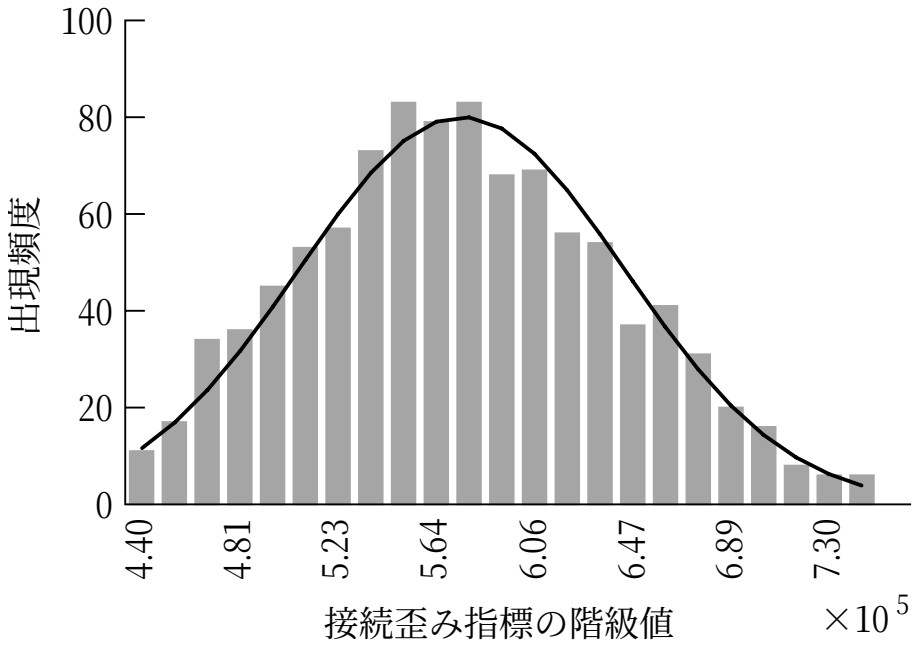


図 5.3　ランダム選択における接続歪み指標の分布例

2-4) i 番目のカテゴリの度数を f_i とする.

3) 正規分布の適合度検定

3-1) ランダム接続経路について音韻環境指標 P の平均 μ と分散 σ^2 を求める

3-2) 音韻環境指標 P の分布が平均 μ と分散 σ^2 の正規分布であると仮定して，各カテゴリの理論上の度数の期待値 e_i を求める. $(1 \leqq i \leqq M)$

3-3) 全てのカテゴリに対して以下の仮説を立てる.

H0: $\{ f_i \}$が正規分布 $N(\mu, \sigma^2)$にあてはまる

H1: $\{ f_i \}$が正規分布 $N(\mu, \sigma^2)$にあてはまらない

3-4) 仮説 H0 の元で(5.5)式が，近似的に χ 自乗分布に従うことを利用して χ 自乗検定を行なう.

$$X^2 = \sum_{i=1}^{M} \frac{(f_i - e_i)^2}{e_i} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \qquad (5.5)$$

以上の方法で，実験資料100文について，有意水準1%で検定した結果，全ての文について仮説 H0 を採択した. 全ての文で，音韻環境指標Pは正規分布とみなして良い. また，接続歪み指標Lについても，同様な検定を行なった. この結果も，全ての文で有意水準1%で仮説 H0 を採択し，接続歪み指標 L は正規分布とみなして良いことがわかった.

5.3　VCV 素片選択における音韻環境指標と接続歪み指標の関係

5.3.1　LD-PER 平面

本節では，PER 選択法と MLD 選択法による選択結果について，音韻環境指標 P と接続歪み指標 L の関係を調べた．同じ手法によって選択を行なっても，音韻環境指標 P と接続歪み指標 L は，合成する目的文によって大きく異なる．もし，合成単位辞書に合成目的文を作るために都合の良い素片が多く含まれていれば，音韻環境指標 P の分布は値が大きな方に片寄り，接続歪み指標 L の分布は値が小さな方に片寄る．もし，合成単位辞書に合成目的文を作るために都合の良い素片が含まれていなければ，音韻環境指標 P の分布は値が小さな方に片寄り，接続歪み指標 L の分布は値が大きな方に片寄る．このため，様々な文についての選択結果の良し悪しを，音韻環境指標 P と接続歪み指標 L の値をそのまま直接比較することができない．

5.2 節で，合成目的文によらず，音韻環境指標 P と接続歪み指標 L の分布は正規分布とみなして良いことを確かめた．ある合成目的文について，ランダムに選んだ 1,000 個の接続経路の音韻環境指標 P の平均値を μ_P，標準偏差を σ_P とする．同様に，接続歪み指標 L の平均値を μ_L，標準偏差を σ_L とする．これらを用いて音韻環境指標 P と接続歪み指標 L を式(5.6)によって**標準化**し，**z- スコア**として扱う．

$$z_P = \frac{P - \mu_P}{\sigma_P}$$
$$z_L = \frac{L - \mu_L}{\sigma_L} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (5.6)$$

ここで，P は評価したい接続経路の音韻環境指標であり，これを標準化して z_P を得る．また，L は評価したい接続経路の接続歪み指標あり，これを標準化して z_L を得る．

以後，特にことわらない限り，音韻環境指標は標準化された z_P を指し，接続歪み指標は z_L を指すものとする．音韻環境指標 z_P は値が大きいほど選択結果が良いことを示しており，接続歪み指標 z_L は値が小さいほど選択結果が良いことを示している．混乱を避けるため，両指標の評価は値の大小でなく選択結果の良悪で表現する．例えば，最大，最小と表記する代わりに最良，最悪のように表記する．

また, 選択された接続経路の接続歪み指標 z_L を横軸に, 音韻環境指標 z_P を縦軸に取った平面を **LD-PER 平面**と呼ぶ.

5.3.2 最良選択と最悪選択の分布

5.2節の実験と同じ条件で，同じ資料100文について，PER選択法とMLD選択法でおのおの最適なVCV素片選択を行なった．また，望ましくない素片選択が行なわれた場合について調べるために，音韻環境指標と接続歪み指標が各々最悪になる選択を行なった．100文全てについて上記4種類の選択結果の場合の接続歪み指標 z_L と音韻環境指標 z_P をLD-PER平面にプロットした結果を図5.4に示す．

図5.4において，音韻環境指標最良はPER選択法での最適なVCV素片選択を示し，音韻環境指標の値は各文についての最良値となる．このとき，接続歪み指標 z_L もz-スコア値で-1.56±0.93と良い方に片寄ることが読み取れる．音韻環境指標を最良にするようにVCV素片選択を行なうと，接続歪み指標は非常に良い値となっている．

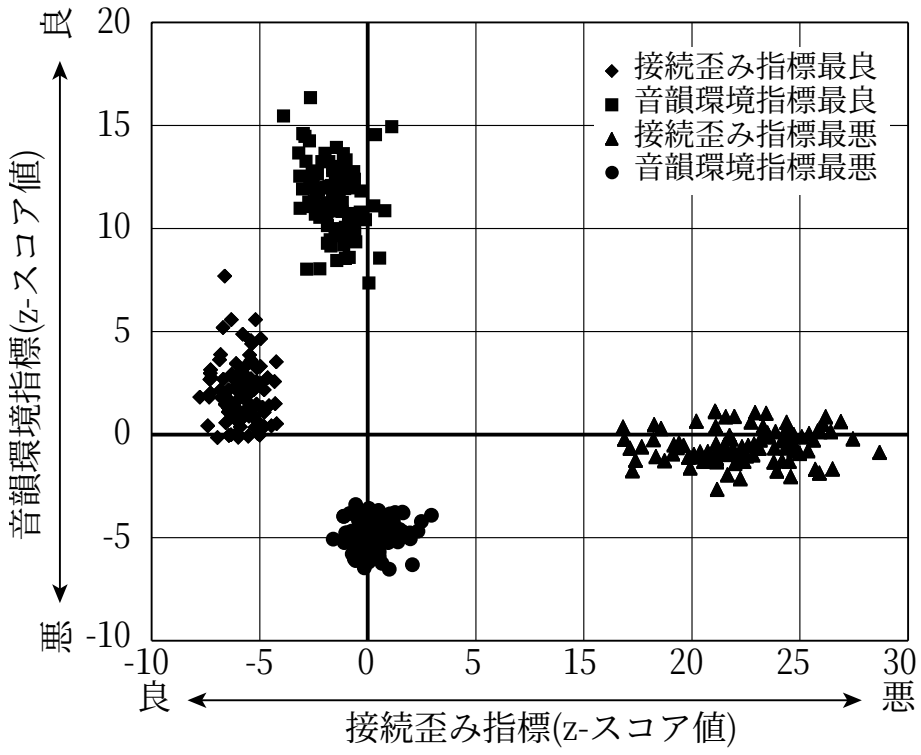


図 5.4 LD-PER 平面上での最良選択と最悪選択の分布

同様に，接続歪み指標最良は MLD 選択法での最適な VCV 素片選択を示し，どの文についても接続歪み指標は非常に良い値を取っている．この場合，音韻環境指標 z_P が z-スコア値で 2.08 ± 1.40 と良い方に片寄ることが判る．

一方，音韻環境指標最悪の選択を行なった場合の接続歪み指標 z_L は 0.27 ± 0.86 と平均値付近に分布している. 同様に, 接続歪み指標最悪の選択を行なった場合の音韻環境指標 z_P は -0.52 ± 0.80 であり平均値付近に分布している．

5.3.3　音韻環境指標と接続歪み指標の関係

音韻環境指標と接続歪み指標の関係をより詳しく調べるために，VCV 素片選択において選択された経路が全ての可能な経路の中でどの程度良い選択になっているかを評価する. この目的のために, ある合成目的文について可能な選択経路を音韻環境指標が良い順に並べたとき､着目している選択経路が上位何パーセントに位置するかを示す**音韻環境順位**を考える. 音韻環境順位は, 音韻環境指標 z_P が正規分布していることを仮定すると, 正規分布の上側累積確率(百分率)を計算することで求めることができる．同様に，接続歪み指標 z_L を正規分布の下側累積確率(百分率)に変換した**接続歪み順位**を考える．

資料100文についてPER選択法によって音韻環境指標が最良となるVCV素片選択を行なった場合，それらの選択結果の接続歪み順位についての度数分布と，累積度数を図5.5に示す．図5.5では棒グラフが度数分布を，折れ線グラフが累積度数を示している．音韻環境指標が最良となるVCV素片選択を行なった結果, 27の文例で接続歪み順位が上位2%以内となっており, 約70の文例で上位10%以内となっていることが判る．この結果から，PER選択法で音韻環境指標が最良となるVCV素片選択を行なった場合には，接続歪み指標の基準でも準最適といえる上位の選択となっているといえる．

同様に, 資料100文についてMLD選択法によって接続歪み指標が最良となるVCV素片選択を行なった場合, それらの選択結果の音韻環境順位についての度数分布と, 累積度数を図5.6に示す．約半数の文例で音韻環境順位が上位2%以内となっており，約75の文例で上位10%以内となっていることが判る．この結果から，MLD選択法で接続歪み指標が

最良となる VCV 素片選択を行なった場合には音韻環境指標の基準でも準最適といえる上位の選択となっていることが示された．音韻環境指標と接続歪み指標が上位の部分では、両者には強い関係があると言える．

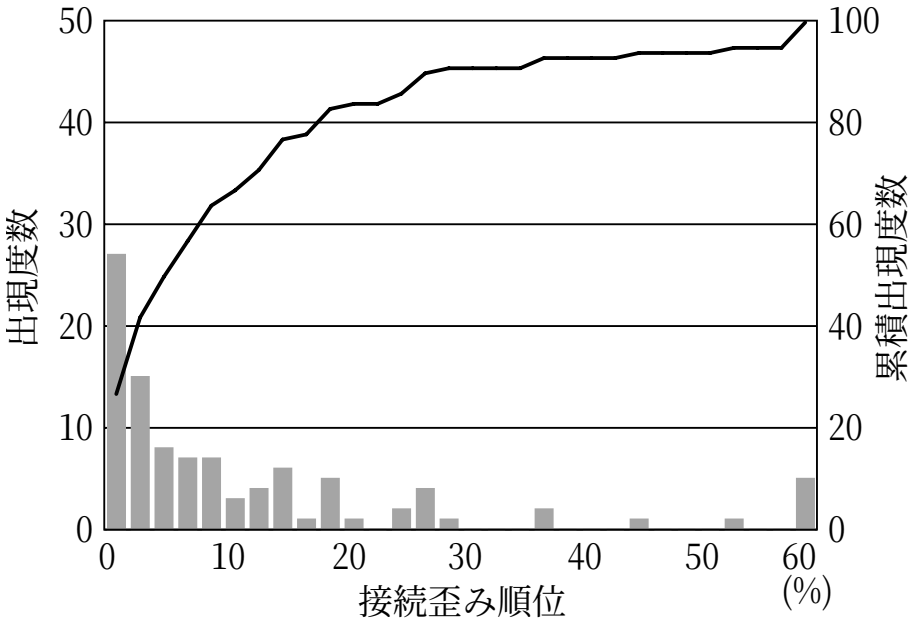


図 5.5　音韻環境指標について最良選択した場合の接続歪み順位の分布

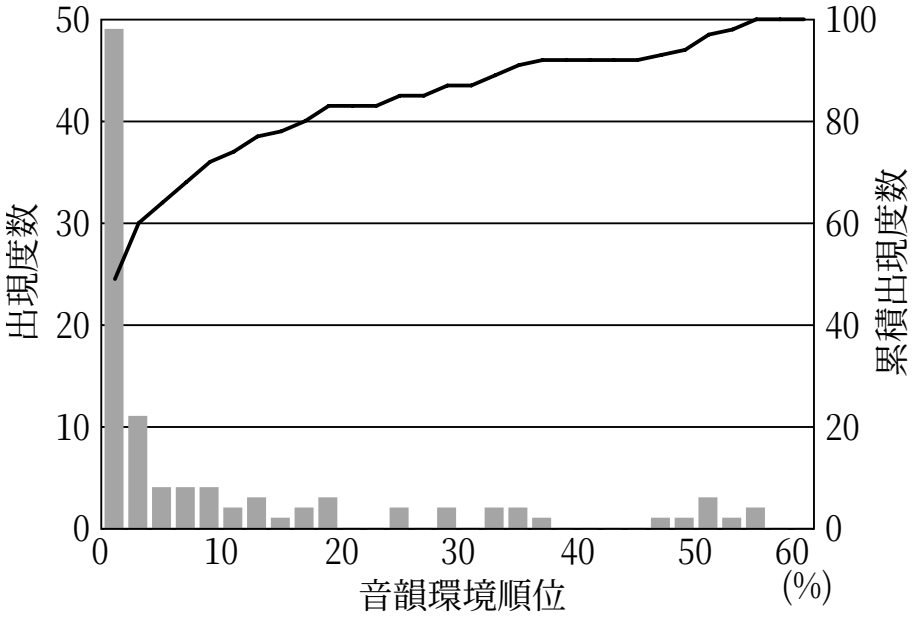


図 5.6　接続歪み指標について最良選択した場合の音韻環境順位の分布

一方, 音韻環境指標が最悪になる選択と接続歪み指標が最悪になる選択を行なった例について，図5.5，図5.6と同様なグラフを図5.7と図5.8に示す．図5.7の分布は，わずかな片寄りは見られるものの，ほぼ一様な分布となった．音韻環境指標が最悪値の付近では、接続歪み指標は特に片寄りが無いと言える．図5.8の分布は，音韻環境順位が悪い方に片

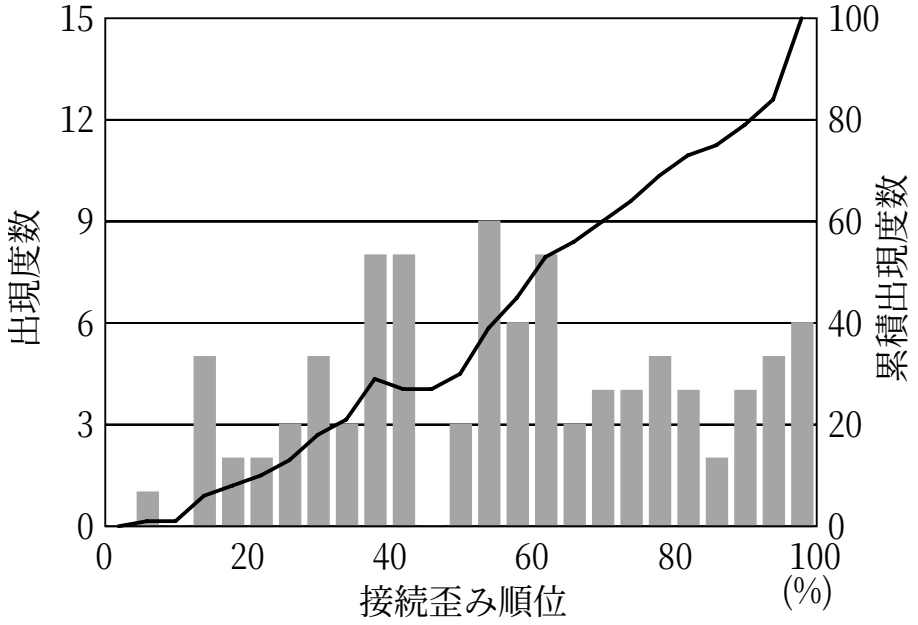


図 5.7　音韻環境指標について最悪選択した場合の接続歪み順位の分布

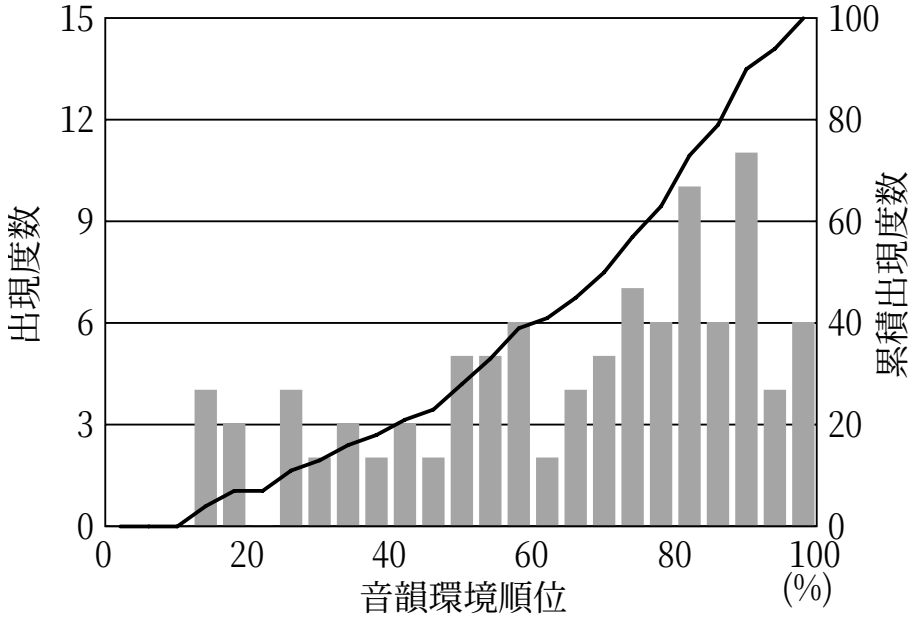


図 5.8　接続歪み指標について最悪選択した場合の音韻環境順位の分布

寄った分布となった. 接続歪み指標が最悪値の付近では、音韻環境指標はわずかだが悪い方に片寄る傾向がみられた.

図 5.5 と図 5.6，図 5.7 と図 5.8 を比較すると，接続歪み指標によって最良及び最悪選択を行なった場合の方が，音韻環境指標によって最良及び最悪選択を行なった場合に比べて, 他方の指標の値に片寄りが強くあらわれることが読み取れる. これは, 接続歪み指標の最適化により音韻環境指標が最適化される度合いの方が, その逆よりも高いことを示している.

本研究では，音韻環境の評価に音韻種別の一致だけを考える簡易な方法を用いたために，2 つの指標の間に上記のような非対称な関係が生じたものと考えられる. しかし，特に最良選択の場合には, 簡易な方法による粗い音韻環境の評価を行なった場合でも, 音韻環境指標と接続歪み指標は非常に強い関係を示した. これは, 音韻環境を粗く評価しても関係が読み取れるほど, 音韻環境と接続歪みに非常に強い関係があることを示している. 音韻環境指標の算出に音韻の音響的な特徴などを用いた精密な方法を用いれば, 両指標の関係の非対称さは減少すると考えられる.

5.4 まとめ

筆者は，VCV 規則音声合成方式において，素片選択を行なう際の音韻環境類似度と合成単位素片接続部における接続歪みの間には一方を改善すれば, 他方も改善されるといった強い関係があると考えてきた. これを実証するために, 本章では音韻環境類似度とVCV 素片接続部における接続歪みの間の関係を実験的に調べた.

実験の結果，PER 選択法で最適な VCV 素片選択を行なった場合，約 70% の文例で接続歪み指標の基準でも上位 10% 以内の選択となっており，MLD 選択法で最適な VCV 素片選択を行なった場合，約 75% の文例で音韻環境指標の基準でも上位 10% 以内の選択となることが示された. これにより, 両者の選択基準には強い関係があり, 一方の素片選択基準で最適に素片選択を行なえば, 他方の素片選択基準でも準最適な選択となる事を示すことができた. これは，1）どちらか一方の選択基準だけで素片選択を行えば十分であるこ

と，2）両選択基準による合成音声に聴感上の差が無いというこれまでの研究結果に論理的な裏付けを与えることができたと云える．また，今回の研究結果は，規則的音声合成法のシステム開発をする際に，どのような素片選択法を採用すべきかについての適切な指針を得るのに有効である．つまり，メモリ規模に重点を置く音声合成システムにはMLD法に準じた方法で素片選択を行い，処理時間に重点を置く音声合成システムにはPER法に準じた方法で素片選択を行えば品質を損なうことなく良好な合成音が得られることが判った．

第6章　VCV素片選択において考慮すべき音韻環境の長さ

第4章と第5章で，LSPベクトルVCV規則音声合成方式に関する検討課題の中で素片選択法を中心に議論をすすめてきた. 提案した2つの素片選択法のうちPER選択法はMLD選択法に比べて選択速度の上では有利であるが，合成単位辞書が大きくなる欠点があった．前章までの議論では，音韻環境を十分反映させるために，VCV素片に先行する5音韻と後続する5音韻を考慮してPERスコアを計算する方式をとった．本章では，PER選択法に関する議論をさらに深め, 本法において考慮すべき**音韻環境の長さ**について改めて実験的に検証した．その結果，先行2音韻，後続1音韻を考慮するだけで，十分な精度でVCV素片選択が行われることが示された. [37]

6.1　音韻環境の長さを変化させたPER選択法

6.1.1　部分PERスコア

本論文で提案したPER選択法は，前後5音韻の長さの音韻環境を考慮してPERスコアを計算することにより，音韻環境の適合度を評価してVCV素片を選択する素片選択法である．PER選択法の有効性は第4章と第5章で示すことができたが，PER選択法において考慮する音韻環境の範囲をある程度狭めても, 合成音声の品質劣化はほとんど起こらないのではないかという議論も行われてきた．本章では，PER選択法によるVCV素片選択の際に考慮すべき音韻環境の長さを検証した. これは, 人間の発話過程において物理的な発話器官の動特性のために生じる調音結合の影響範囲を, 音声合成システムの構築という視点から検証することに相当する.

PER選択法において考慮すべき音韻環境の長さを検証するために, 先行音韻環境と後続音韻環境の長さを変えて音韻得点を集計する新たな素片選択基準を式(6.1)で定義する.

$$PER(F,R)=\sum_{i=1}^{F}\frac{1}{3^{i-1}}f(i)+\sum_{i=1}^{R}\frac{1}{3^{j-1}}r(j) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (6.1)$$

式(6.1)中で，F は**先行音韻環境**として考慮する音韻の個数であり，R は**後続音韻環境**として考慮する音韻の個数である. 前後5音韻の長さの音韻環境を考慮したPER選択法が VCV 素片選択に有効であることが示されているため，実験で用いた F の値の範囲は 0　F　5 とし，R の値の範囲は 0　R　5 とした. 但し，$F = R = 0$ では，音韻環境が全く考慮されず, 単に合成単位辞書中での素片登録順に依存した選択となるので実験条件から除外した.

式(6.1)の素片選択基準は図 6.1 に示すように，基本的には第 4 章で定義した PER スコアの計算を限定された音韻環境の範囲内で打ち切ったものである. 以後これを, **部分PERスコア**と呼ぶ.

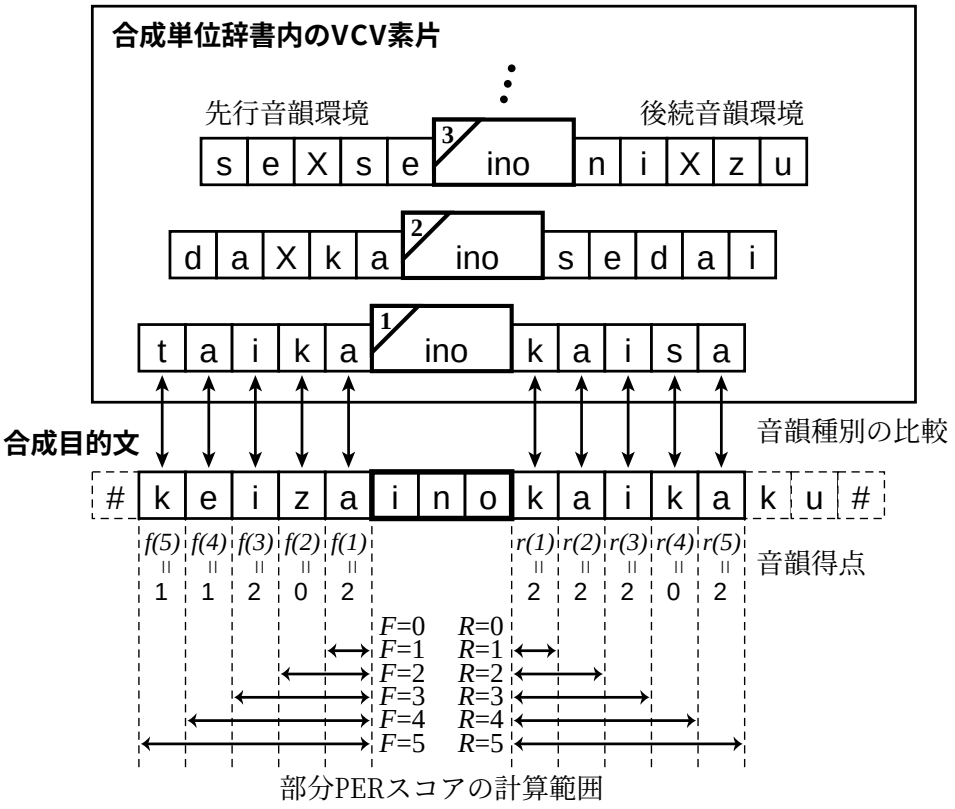


図 6.1　部分 PER スコアの概念図

6.1.2　部分 PER スコアによる素片選択実験

PER 選択法において考慮すべき音韻環境の長さを検証するために，従来の PER スコアに変えて(6.1)式で定義した部分 PER スコアを素片選択基準として用いた PER 選択法により，VCV 素片を選択する実験を行った．素片選択実験に用いる合成単位辞書は，第 5 章の実験で用いたものと同じ辞書であり，登録素片数は約 14,000 個である．また，素片選択実験に用いた合成目的文は，第5章で述べた素片選択実験に用いたものと同じく，以下の条件を満たす新聞記事文章 100 文である．

1) 文の長さが VCV 単位の数にして 10 から 30 程度であること
2) 文に使用されている全ての VCV 単位について，VCV 単位辞書に 5 個以上の VCV 素片が登録されていること

1 つの合成目的文に対し，部分 PER スコア *PER*(*F*,*R*)の計算時に先行音韻環境として考慮する音韻の個数 *F* と，後続音韻環境として考慮する音韻の個数 *R* を 6.1.1 項の条件で変えて VCV 素片選択を行う．従って，1つの合成目的文に対して，選択時に用いた部分 PER スコア *PER*(*F*,*R*) の *F* と *R* が異なる 35 種類の選択結果が得られる．

選択結果の比較を行うために，第5章で定義した音韻環境指標 *P* と接続歪み指標 *L* を用いる．上記の全ての選択結果に対し，改めて音韻環境指標 *P* と接続歪み指標 *L* を計算し，さらに標準化して z_P と z_L を求めた．標準化のために，それぞれの合成目的文について，可能な VCV 素片選択経路の中から 10,000 個の経路をランダムに取り出し，音韻環境指標の平均 μ_P と分散 σ_P^2 と，接続歪み指標の平均 μ_L と分散 σ_L^2 を求めて用いた．

このようにして求めた標準化した音韻環境指標 z_P と接続歪み指標 z_L を，合成目的文 100 文について平均し選択結果の評価指標とした．この指標を簡単のために，平均音韻環境指標と平均接続歪み指標と呼ぶ．

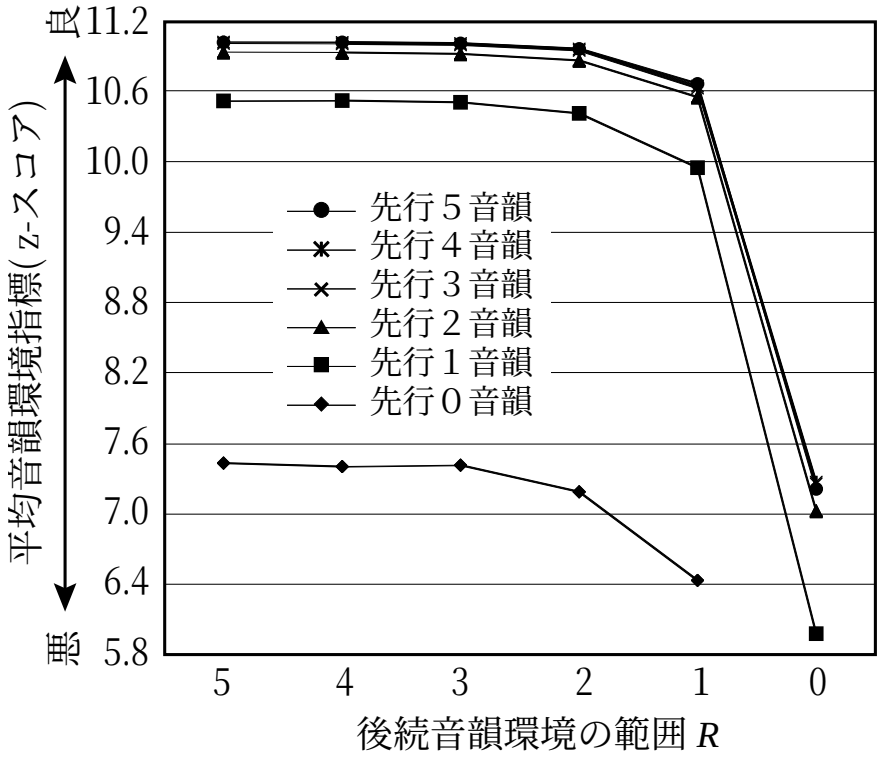
6.1.3　PER 選択法に対する音韻環境の長さの影響

部分 PER スコア $PER(F,R)$ の F と R を変えて行った VCV 素片選択実験の結果の平均音韻環境指標による評価を図 6.2 に示す．図 6.2a)は，先行音韻環境として考慮する音韻の個数 F を固定して，横軸に後続音韻環境として考慮する音韻の個数 R，縦軸に平均音韻環境指標をとったグラフである．また，図 6.2b)は，後続音韻環境として考慮する音韻の個数 R を固定して，横軸に先行音韻環境として考慮する音韻の個数 F，縦軸に平均音韻環境指標をとったグラフである．同様の評価を，平均接続歪み指標によって行った結果を図 6.3 に示す．

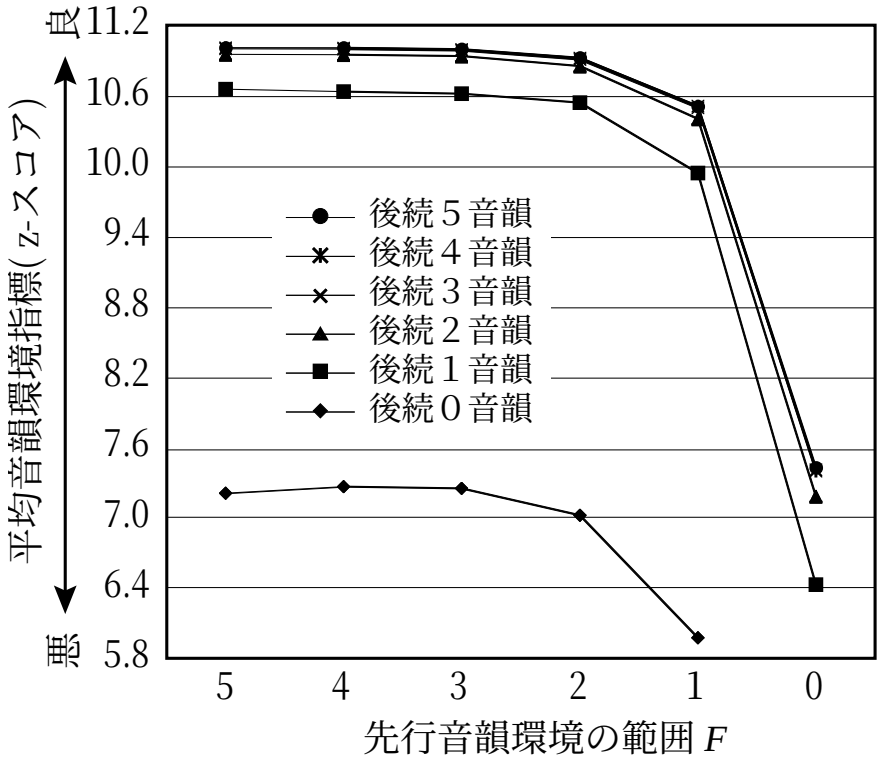
図 6.2 と図 6.3 に共通して，先行音韻環境または後続音韻環境のどちらか一方を全く考慮しない条件では($F = 0$ または $R = 0$)，平均音韻環境指標も平均接続歪み指標も極端に悪くなることが判る．このことは，VCV 素片が，先行音韻と後続音韻のいずれからも無視できない大きさで調音結合の影響を受けていることを示している．これは，人間の発話器官の動作において，前の音の発話の構えから連続的に推移してくるために現在の音の発話の構えが影響を受け，同時に次の音の発話の構えの準備のためにその影響を受けるという発声機構上の相互影響の関係から考えても納得できる結果である．

VCV 素片選択実験の結果の平均音韻環境指標による評価では，図 6.2 より，先行音韻環境として考慮する音韻の個数 F と後続音韻環境として考慮する音韻の個数 R が共に 2 以上であれば評価指標の値はほとんど変わらないことが読み取れる．$F = 1$ または $R = 1$ のとき，評価指標の値が悪化するが，その程度はわずかである．平均音韻環境指標による評価では，$F = 1$ ，$R = 1$ で十分であるとみて良い．

一方，VCV 素片選択実験の結果の平均接続歪み指標による評価では，図 6.3 より，先行音韻環境として考慮する音韻の個数 F が 1 のとき，明らかに評価指標の値が悪化することが読み取れる．$F = 2$ であれば，後続音韻環境として考慮する音韻の個数 R を 1 音韻まで減らしても平均音韻環境指標にあまり変化が無いことが判る．平均接続歪み指標による評価は，平均音韻環境指標による評価より厳しく $F = 2$ ，$R = 1$ が必要であることを示している．

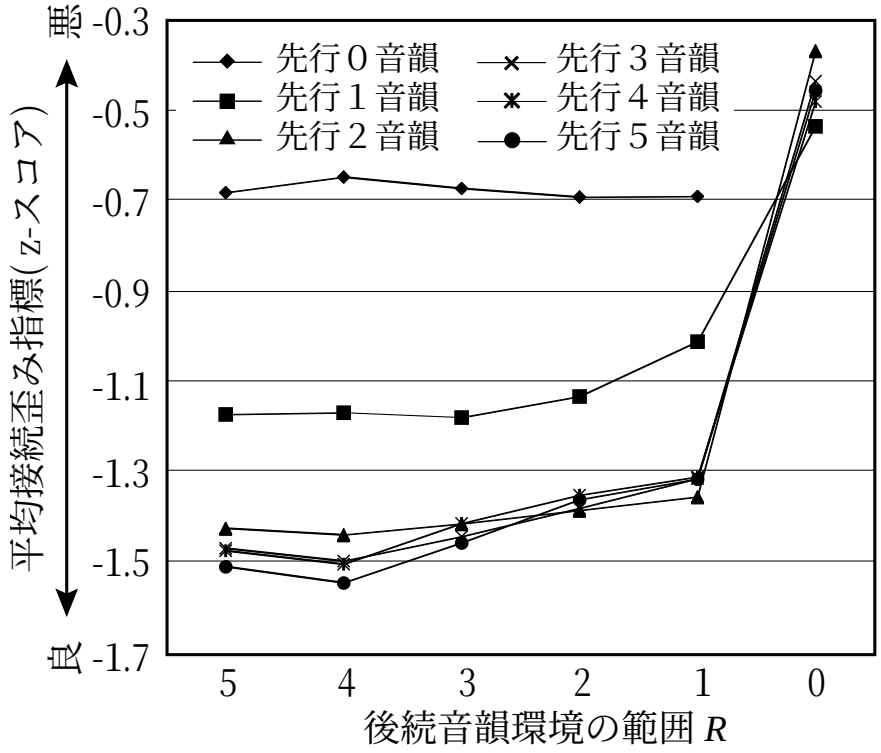


a) 先行音韻環境の範囲固定時の平均音韻環境指標

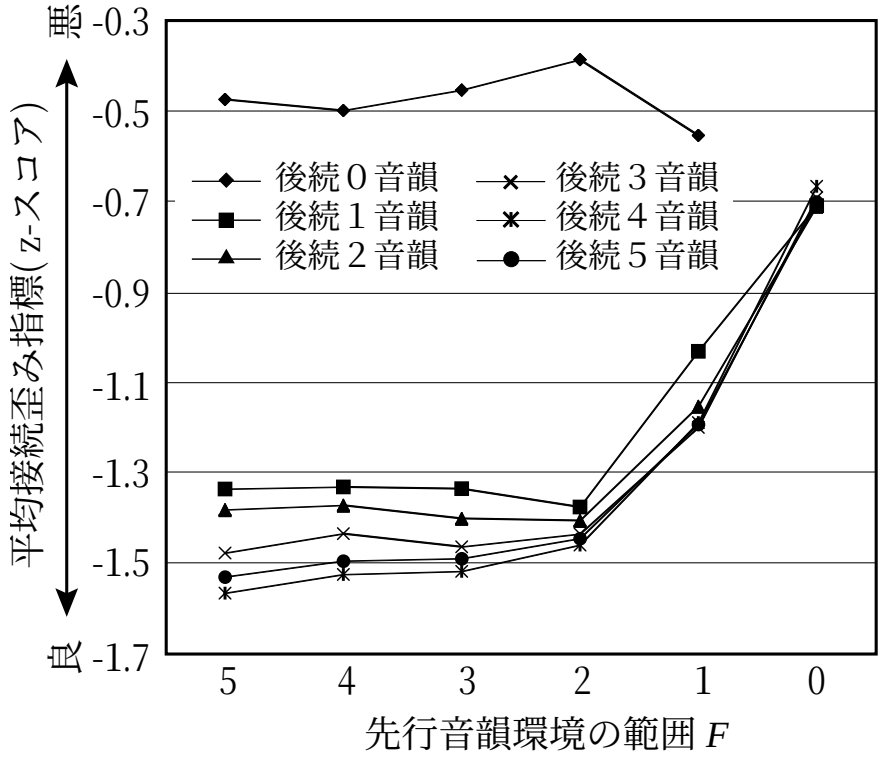


b) 後続音韻環境の範囲固定時の平均音韻環境指標

図 6.2　VCV 素片選択結果の平均音韻環境指標による評価



a) 先行音韻環境の範囲固定時の平均接続歪み指標



b) 後続音韻環境の範囲固定時の平均接続歪み指標

図 6.3　VCV 素片選択結果の平均接続歪み指標による評価

以上より, PER選択法を音声合成システムに採用する場合, 平均音韻環境指標による評価と平均接続歪み指標による評価を良く保つためには，合成単位辞書に登録するVCV素片に付加する音韻環境情報は先行2音韻・後続1音韻が必要であることが判った. また, 音韻環境の長さをこれ以上に増やしても，両指標とも向上は見られない.

6.2 まとめ

本章ではPER選択法において，VCV素片選択の際に考慮すべき音韻環境の長さを検証した．PERスコアの計算に算入する音韻環境の長さを制限した部分PERスコアを定義して，素片選択実験を行った結果以下の結論を得た.

1) VCV素片に対する調音結合の影響は素片の前後両方向から及んでおり，先行音韻環境情報，後続音韻環境情報のいずれも省略できない.
2) 先行音韻環境として考慮する音韻の個数は2個, 後続音韻環境として考慮する音韻の個数は1個で十分である.

この結論より，PER選択法において，合成単位辞書中のVCV素片に付加する音韻環境情報を先行2音韻と後続1音韻に抑えることができ，合成単位辞書を大幅に小規模化できる．また，素片選択の処理速度も高速化できることが判った.

第7章　VCV規則音声合成に適用する ベクトル量子化の検討

小規模な計算機資源で実現可能な高品質音声合成の手法として，LSPベクトルVCV規則音声合成方式を提案した．前章まで，本方式に必要な合成単位辞書の規模とVCV素片選択法について詳細に検討してきた．本章では，本方式において合成単位辞書中のVCV素片(LSPパラメータ系列)に適用するベクトル量子化について検討を行う．また，ベクトル量子化を用いて合成単位辞書の容量を削減するだけでなく,ベクトル量子化の特長を生かしたVCV素片選択法である距離テーブル参照法(Distance Table Look-up Method: DTL選択法)を提案し，その性能について検証した．DTL選択法では，ベクトル量子化の代表ベクトル間の距離テーブルを参照することにより,接続歪みの評価を高速に行い,合成単位辞書から適切なVCV素片を効率的に選択する,

本章における検証の結果, i) 規則音声合成系にベクトル量子化を適用した場合, 対象話者が1名に限られるため128個程度の非常に少ない代表ベクトル数で量子化が可能であること，ii) 距離テーブルとして全ての代表ベクトル間の距離を記憶する必要はなく，各代表ベクトルに対して距離が近い順に1位から8位ないし16位程度までの代表ベクトルについて距離テーブルを作成すれば十分であることが判明した. [35]

7.1　LSPベクトルVCV規則音声合成方式とDTL選択法

7.1.1　LSPベクトルVCV規則音声合成方式

LSPベクトルVCV規則音声合成方式の構成については，第3章で述べた．その最大の特長は，音声合成の基本単位であるVCV素片の記録にベクトル量子化されたLSPパラメータを用いることにより合成単位辞書(VCV unit dictionary)の記憶容量を小さく抑える点である．本方式では，VCV素片はVQコードブックを用いてベクトル量子化することにより，代表ベクトルを表すVQインデックスの系列として符号化され，合成単位辞書に

収録される.

第 4 章から第 6 章に渡り，LSP ベクトル VCV 規則音声合成方式のために提案した以下の 2 つの素片選択法について論じてきた.

(i) VCV素片を収集した際の音韻環境と合成する文章中でのVCV素片の音韻環境の類似度を素片選択の基準にする方法(PER 選択法).

(ii) VCV 素片の接続部での接続歪みを素片選択の基準にする方法(MLD 選択法).

実験的に比較検討した結果, 両者によって生成した合成音声に聴感上の差はなく, 合成単位辞書の記憶容量を小さく抑えたいような応用では，MLD選択法を用いればよいことを報告した．MLD選択法は，合成単位辞書の記憶容量は少なくできるが，VCV 素片の選択に多数回に渡る接続歪みの計算を伴い計算時間がかかるという問題点があった.

本章では，LSP ベクトル VCV 規則音声合成方式におけるベクトル量子化について検討を行う．また，ベクトル量子化の特長を生かし，VQ コードブックの代表ベクトル間の距離を予め計算して作成した距離テーブル(distance table)を参照することで VCV 素片の接続歪みの計算を高速化する距離テーブル参照法(Distance Table Look-up Method: DTL 選択法)を提案する.

7.1.2　距離テーブルを用いる VCV 素片選択法

前章まで検討してきた LSP 距離最小化選択法(Minimal LSP Distance method: MLD 選択法)の概念図を改めて図 7.1 に示す．同一の VCV 合成単位に属する VCV 素片が複数存在するため，図 7.1 において破線で示すような VCV 素片の様々な接続が可能である．このように，VCV 素片の接続が可能であることを示す経路を接続可能経路と呼ぶ．MLD選択法は, 接続可能経路に接続歪みをコストとして割り振り, 合成音声の開始点から終了点までのコストの総和が最小となる経路をDPの手法によって探索する方法である．探索の結果得られたLSP距離最小の経路上にあるVCV素片を，最適なVCV素片として選択する.

MLD選択法では, VCV素片の接続部における接続歪みを評価するために, LSPパラメータの距離を用いる．図 7.2a)に示すように，VCV 素片の接続部における先行 VCV 素片の

最終フレームのLSPパラメータが $\boldsymbol{\omega}^f=(\omega_1^f,\omega_2^f,\cdots,\omega_p^f)$，後続VCV素片の先頭フレームのLSPパラメータが $\boldsymbol{\omega}^r=(\omega_1^r,\omega_2^r,\cdots,\omega_p^r)$ であるとき，VCV素片の接続部におけるLSP距離(LSP distance)を式(7.1)で定義する．

$$d(\boldsymbol{\omega}^f,\boldsymbol{\omega}^r)=\sqrt{\sum_{i=1}^{p}\left(\omega_i^f-\omega_i^r\right)^2} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (7.1)$$

式(7.1)で定義した $d(\boldsymbol{\omega}^f,\boldsymbol{\omega}^r)$ を用いて，ダイナミック・プログラミング(DP)の手法によりVCV素片選択を行なう．DPによる経路探索の過程で，式(7.1)の距離計算を多数回繰り返すことは，素片選択の処理速度の点で非常に不利である．

LSPベクトルVCV規則音声合成方式では，VCV素片の記録にLSPパラメータを直接用いるのではなく，ベクトル量子化を行ったVQインデックスを用いる．この場合，VCV素片の接続とは，図7.2b)に模式的に示すようにVQインデックス系列の接続に他ならない．ベクトル量子化の代表ベクトルの数は有限であるから，代表ベクトル間の距離を予め計算した表として，**距離テーブル**を作ることができる．VCV素片の選択時には，先行VCV素片の最終フレームのVQインデックスと後続VCV素片の先頭フレームのVQインデック

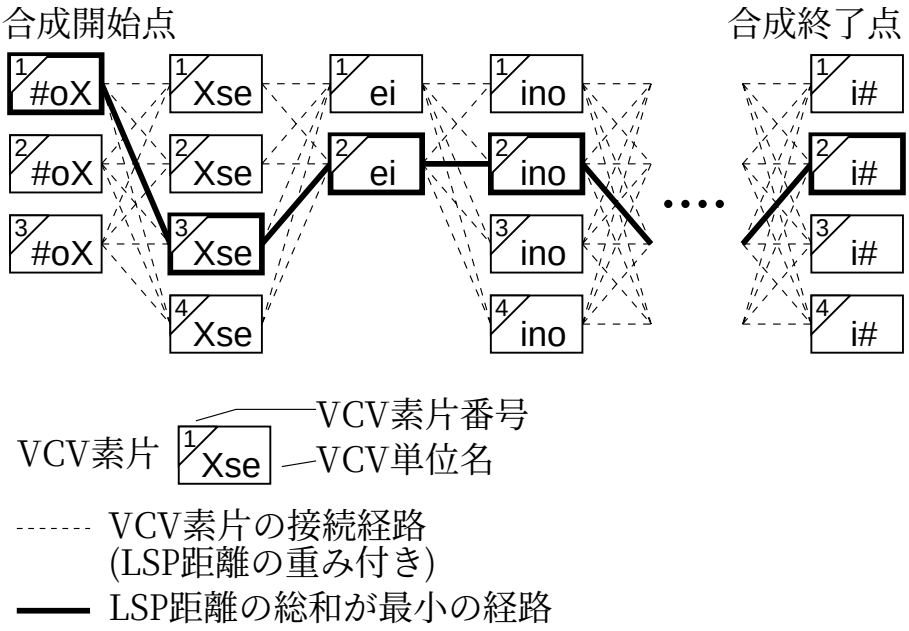


図 7.1　MLD 選択法による VCV 素片の選択

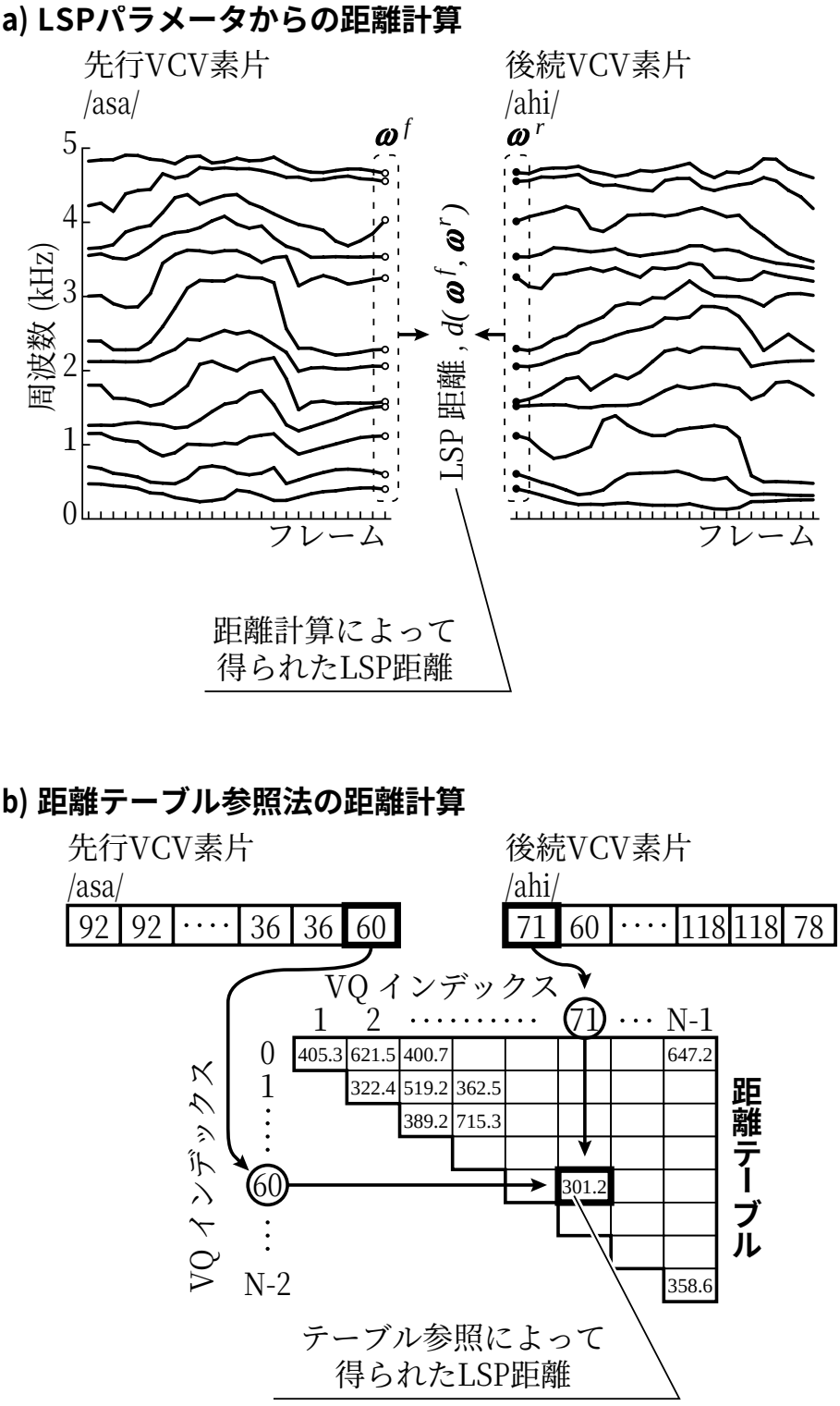


図 7.2　VCV 素片の接続部における LSP 距離と距離テーブル参照法の概念図

スによって距離テーブルを参照し, 接続部におけるLSP距離を得ることができる. 距離計算をテーブル参照は, 多数の積和を含む距離計算に比べて計算量が少なく, 非常に高速に実行できる．このような手法により，接続歪みの評価を高速化し，MLD 選択法を処理速度の点で改良する方式として，**距離テーブル参照法**(Distance Table Look-up Method: **DTL 選択法**)を提案する.

7.1.3　音声資料と合成単位辞書

本章の実験で用いた音声資料は，NHK の FM ラジオ・ニュースの同一の男性アナウンサの発話部分を音声資料として用いた．音声資料は，標本化周波数 11.025kHz，量子化数 16 ビットでサンプリングした．VCV 素片は，視察で音韻マーキングした音声資料から母音継続時間の中心点で切り出す方法で自動的に生成した. 第5章の実験結果に従って, 約 40 分間の音声資料から約 14,000 個の VCV 素片を採取した．音声資料より採取した VCV 素片を，ハニング窓で切り出し，フレーム長256点，インターバル64点，次数12次でLSP分析した. 得られたLSPパラメータにベクトル量子化を施すことにより, 代表ベクトルの VQ インデックスの系列として表現された VCV 素片を合成単位辞書に収録した．このようにして作成した合成単位辞書の仕様を表 7.1 に示す.

以下 7.2 節では，LSP ベクトル VCV 規則音声合成方式におけるベクトル量子化について検討し，7.3 節では，新たに提案した DTL 選択法の性能に関して検証を行う.

表 7.1　合成単位辞書に収録した VCV 単位の種類と VCV 素片数

合成単位の型	合成単位の種類数	収録素片の平均個数
VCV型	446(570)	21.4
VV型	35(35)	61.4
#CV型	67(95)	11.6
#V型	5(5)	48.2
V#型	6(6)	169.1
合　計	559(711)	13,713

注）#は無音を表している.
　　合成単位の種類数のカッコ内は，論理的に可能な種類数である.
　　収録素片の平均個数の合計欄は素片の総数である.

7.2　ベクトル量子化の代表ベクトル数

7.2.1　VQ コードブック作成のアルゴリズム

LSP パラメータのベクトル量子化のための VQ コードブック作成には，**LBG アルゴリズム**(Linde-Buzo-Gray alogrithm)[32]と**2分割繰り返しアルゴリズム**を用いた．この方法では，人間の発話から得た多数の LSP パラメータ(LSP ベクトル)を訓練用のサンプルとして，2分割手法により代表ベクトルを増やしながら，繰り返し手法で代表ベクトルの最適化を行うことで，VQ コードブックを作成してゆく．以下に，VQ コードブック作成のアルゴリズムの詳細について述べる．

VQ コードブックの**訓練サンプル**となる LSP ベクトルの総数を M，分析次数を p として訓練サンプル集合を $\boldsymbol{\Omega} = \{\boldsymbol{\omega}_i = (\omega_{i,1}, \cdots, \omega_{i,p}) ; i = 0, \cdots, M-1\}$ と表記する．LBG アルゴリズムでは，繰り返し処理により代表ベクトルを変更してゆく．VQ コードブックの大きさが N で，LBG アルゴリズムの繰り返しが r 回目のときの VQ コードブックを $\boldsymbol{V}^{(N,r)} = \{\boldsymbol{v}_j^{(N,r)} = (v_{j,1}^{(N,r)}, \cdots, v_{j,p}^{(N,r)}) ; j = 0, \cdots, N-1\}$ と表記する．このとき，訓練サンプルの LSP パラメータ $\boldsymbol{\omega}_i$ と，VQ コードブックの代表ベクトル $\boldsymbol{v}_j^{(N,r)}$ の LSP 2 乗距離を $d^2(\boldsymbol{\omega}_i, \boldsymbol{v}_j^{(N,r)})$と表記する．また，ベクトル集合の重心を求める演算を $C(\bullet)$ と表記する．

a) LBG アルゴリズム

(処理 1a)　初期化

$r = 0$，VQ コードブックの初期集合を $\boldsymbol{V}^{(N,0)}$ 歪み閾値を ε，平均歪みを $D_{-1} = \infty$ とする．

(処理 2a)　訓練サンプル集合の分割

「もし，$t \neq j$ なる全ての $\boldsymbol{v}_t^{(N,r)} \in \boldsymbol{V}^{(N,r)}$ について $d^2(\boldsymbol{\omega}_i, \boldsymbol{v}_j^{(N,r)}) < d^2(\boldsymbol{\omega}_i, \boldsymbol{v}_t^{(N,r)})$ なら $\boldsymbol{\omega}_i \in \boldsymbol{S}_j$ 」なる規則により，$\boldsymbol{\Omega}$を $\boldsymbol{V}^{(N,r)}$によって N 個の互いに素な部分集合 $\boldsymbol{S}_j$ $(j = 0, \cdots, N-1)$ に分割する．

(処理 3a)　平均歪みの計算

式(7.2)に示す平均歪みを計算する．

$$D_r = \frac{1}{M} \sum_{j=0}^{N-1} \sum_{\boldsymbol{\omega}_i \in \boldsymbol{S}_j} d^2(\boldsymbol{\omega}_i, \boldsymbol{v}_j^{(N,r)}) \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (7.2)$$

(処理 4a)　終了条件の判定

もし $(D_{r-1}-D_r)/D_r<\varepsilon$ なら $\boldsymbol{V}^{(N,r)}$ を準最適なVQコードブックとして手続きを終了する.

(処理 5a)　VQ コードブックの変更

訓練サンプルの各部分集合 $\boldsymbol{S}_j$ $(j=0,\cdots,N-1)$ に対して，重心計算 $\boldsymbol{v}_j^{(N,r+1)}=C(\boldsymbol{S}_j)$ を行なうことで, 改良されたVQコードブック $\boldsymbol{V}^{(N,r+1)}=\{\boldsymbol{v}_j^{(N,r+1)};j=0,\cdots,N-1\}$を得る．$r=r+1$ として(処理 2a)へ

b) 2 分割繰り返しアルゴリズム

(処理 1b)　初期化

大きさの小さい適当なベクトルを Δ ，$N=1$ とし，初期 VQ コードブックを $\boldsymbol{V}^{(1,0)}=\{C(\boldsymbol{\Omega})\}$　とする.

(処理 2b)　代表ベクトルの分割

VQコードブックの各代表ベクトルを $\boldsymbol{V}^{(N,r)}=\{\boldsymbol{v}_0^{(N,r)},\cdots,\boldsymbol{v}_{N-1}^{(N,r)}\}$, 近接した2つのベクトル $\boldsymbol{v}_j^{(N,r)}+\Delta$ と $\boldsymbol{v}_j^{(N,r)}-\Delta$ に置き換えることにより，2 倍の数の代表ベクトルを持つ VQ コードブック $\boldsymbol{V}^{(2N,0)}=\{\boldsymbol{v}_0^{(2N,0)},\cdots,\boldsymbol{v}_{2N-1}^{(2N,0)}\}$ を作成する.

(処理 3b)　LBG アルゴリズム

$\boldsymbol{V}^{(2N,0)}$ を初期値としてLBGアルゴリズムにより, 準最適なVQコードブック $\boldsymbol{V}^{(2N,r)}$ を得る．$2N$ が求めたいVQ コードブックサイズなら処理を終了．それ以外なら，$N=2N$ として(処理 2b)へ

LBGアルゴリズムによって求められる代表ベクトルが最適であることは保証されていないが，式(7.2)の平均歪みは単調減少し収束してゆくことが保証されている．言い換えると，LBG アルゴリズムは局所的に平均歪みが極小となる準最適な代表ベクトルを与える．しかし，初期値にもよるが，十分な量の訓練サンプルを用いてLBG アルゴリズムを実行することにより, 実用上は十分良好なVQコードブックを得ることができるため, 多くの応用で使用されている.

7.2.2　VQ コードブックの作成

LSPパラメータのベクトル量子化コードブックの作成には, 男性アナウンサの発話による約110分間のニュース音声を用いた. 男性話者は, 7.1.3項で述べた話者と同一話者であり, 約110分間の音声資料にはVCV素片の採取に用いた資料が含まれている. VQコードブックの訓練サンプルは，上記の音声資料を LSP 分析して得られた LSP パラメータであり，訓練サンプル数 M は約 990,000 個である．2 分割アルゴリズムと LBG アルゴリズムにより，VQ コードブックのサイズ $N = 2$ から 2^{14} (16,384)までの VQ コードブックを作成した．VQ コードブックのサイズ(代表ベクトル数)が最大の $N = 2^{14}$ (16,384)の場合を考えても，訓練サンプル数はその 60 倍以上の数があり，十分な個数の訓練サンプルを用いているといえる.

LSP ベクトル VCV 規則音声合成方式で用いる VQ コードブックのサイズ N を決定するために, VQコードブックのサイズ N とベクトル量子化による量子化誤差の関係を実験的に調べた．実験では，VQ コードブックの訓練に用いた訓練サンプルに対して，各サイズの VQ コードブックによりベクトル量子化を施した場合の量子化誤差を，式(7.2)の平均歪みで評価した．式(7.2)は歪み評価のために LSP パラメータ 2 乗距離を用いているため，訓練サンプルと対応する代表ベクトルとの平均LSP 2 乗距離で平均的な量子化誤差を評価

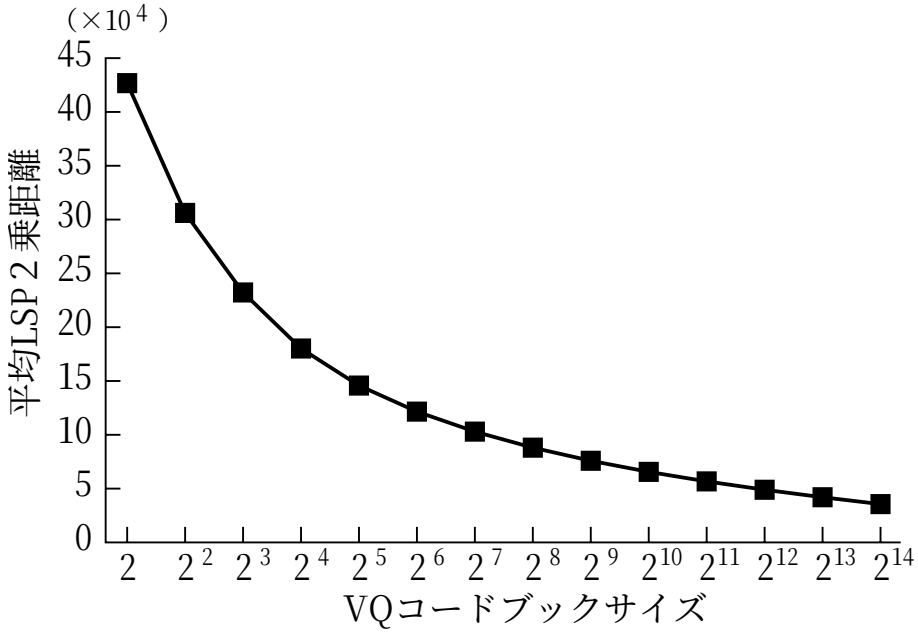


図 7.3　VQ コードブックサイズとベクトル量子化歪みの関係

することになる．

図7.3に，ベクトル量子化のVQコードブックサイズNと，平均LSP 2乗距離の関係を示す．平均LSP 2乗距離は，VQコードブックサイズNの増加とともに減少している．Nが増加すると，平均LSP 2乗距離の減少率はやや低下するが，Nが2から2^{14}の範囲では平均LSP 2乗距離の減少は続いている．しかし，我々が行った予備的な実験では，$N = 2^7$以上のVQコードブックを用いてベクトル量子化を行って作成した合成音声の間では品質の差が聞き取れなかった. 次の項では, 合成音声の聞き取り実験による主観評価によってVQコードブックサイズと量子化歪みの関係を評価した．

7.2.3 合成音声の主観評価によるVQコードブックの評価

VQコードブックサイズNを様々に変えた場合のLSPベクトルVCV音声合成法による合成音声の品質を主観評価実験により評価した. 異なるサイズのVQコードブックを用いて本手法により合成した1対の合成音声のうち「どちらの合成音声が聞き取りやすいか」の判定を行なう1対比較による主観評価実験を行った．実験には，3秒程度の4つの短文について，サイズN = 2,8, 32, 128, 512, 2048のVQコードブックを用いて作成した6種類の合成音声の組み合わせ15対について順序の入れ替えを含めて8回ずつ120回の1対比較を課した．比較対の提示順はランダムとした．被験者は健康な20代の男女10名で実験を行った．

1対比較実験で得られた判定結果から，Thurstoneの比較判定の法則[44]でケースVを適用して，合成音声の品質尺度値を求めた．予備実験では，N =2^7 以上のVQコードブックを用いてベクトル量子化を行って作成した合成音声は, ベクトル量子化を適用していないLSPパラメータを用いたVCV規則音声合成法の合成音声と品質の差が聞き取れなかった．品質尺度値を求めるにあたって，余裕をみて大きめのN=2048のVQコードブックによる合成音声を基準とした．

実験の結果として，VQコードブックサイズと合成音声の品質尺度値の関係を図7.4に示す．図7.4の縦軸は，合成音声の品質尺度値であり，基準となるN = 2048のVQコードブックによる合成音声の尺度値を0としている．VQコードブックのサイズNが32ない

し 128 以上では，品質尺度値に差がないことが図 7.4 から読み取れる．この実験結果は，VQ コードブックのサイズ N が 32 ないし 128 以上とすれば，ベクトル量子化を適用していないLSPパラメータを用いたVCV規則音声合成法の合成音声と同程度の品質の合成音声を作成できることを示している. ベクトル量子化のVQコードブックをこのように小さくできるのは, 規則音声合成で音声提供話者を一人に絞ったことが大きく関与しているものと考えられる.

合成音声品質の安定性を考慮して, VQコードブックサイズを大きめに N =128とし, 本手法の合成単位辞書を作成した場合の記憶容量は以下のように計算できる. 実験に用いた合成単位辞書中の VCV 素片は平均で 20 フレーム程度の長さがあり，VCV 素片の総数は約14,000個である．VQコードブックサイズを N =128 とした場合，代表ベクトルのVQインデックスを 7bits で記録できる．従って，合成単位辞書の大きさは，7bits × 20 × 14,000

256K バイト程度と非常に小さなものにできる．VQ コードブックにおいて，12 次のベクトルの各要素に 10bits の割り当てを行なうと，10bits × 12 × 128　15K バイト程度の大きさとなる．また，残差波形辞書には，残差波形を各母音 6 種類と音節別に分類した子音 95 種類について，平均 2K バイトで記録した．このため，残差波形辞書は，2K バイト ×(6 + 95)　200K バイト程度大きさとなる．従って，合成単位辞書と VQ コードブック，残差波形辞書を合わせても 500K バイト以下で記録できる.

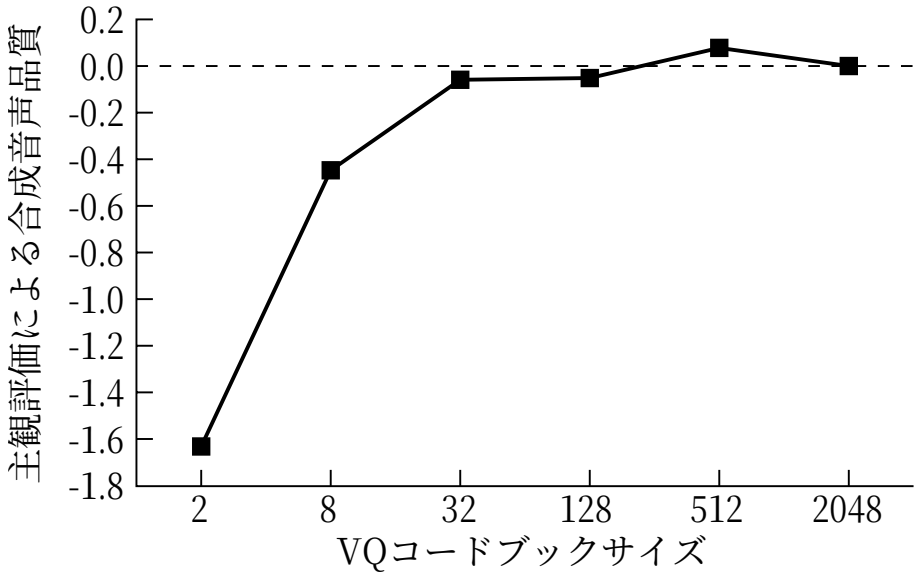


図 7.4　ベクトル量子化コードブックのサイズと合成音声の主観評価

7.3 DTL 選択法

7.3.1 DTL 選択法による VCV 素片の最適選択

VQ コードブック中で VQ インデックス i を持つ代表ベクトル $\boldsymbol{v}_i^{(N,r)}$ と VQ インデックス j を持つ代表ベクトル $\boldsymbol{v}_j^{(N,r)}$ の LSP 2 乗距離 $d^2(\boldsymbol{v}_i^{(N,r)},\boldsymbol{v}_j^{(N,r)})$ を $d^2(i, j)$と略記する．VQ コードブックが決まれば，音声合成システムを作成する時点で，$d^2(i, j)$をあらかじめ計算して距離テーブルとして保持しておくことが可能である．音声合成時の VCV 素片選択時には, 複雑な距離計算を行う代わりに, 距離テーブルを参照することで高速に接続点での LSP 2 乗距離を得ることができる．このように距離テーブルを用いて VCV 素片選択時の距離計算を高速化する素片選択手法を，DTL 選択法と呼ぶ.

距離テーブルとして全ての代表ベクトルの組み合わせについて $d^2(i, j)$を保持しておくと，どのような VCV 素片の接続点でも LSP 2 乗距離を求めることができる．このとき，DTL 選択法による VCV 素片選択結果は，合成文中において VCV 素片の各接続点での LSP 2 乗距離の総和が最小になるという意味で最適な VCV 素片選択となる.

本方式では，各 VCV 単位について音韻環境の異なる多くの VCV 素片を収録するため，ほとんどの場合，どの VCV 単位の接続点でも LSP 2 乗距離が非常に小さくなるような VCV 素片を選択することができる．従って，最適な VCV 系列上の接続点では，距離テーブル中の LSP 2 乗距離 $d^2(i, j)$を距離が小さい順に並べたとき，下位の順位のものはほとんど参照されないことが期待できる.

代表ベクトル間の LSP 2 乗距離 $d^2(i, j)$を各 i 毎に，j について値を昇順に並べた場合の順位を**距離順位**と定義する．もし，上記の仮定が正しければ，距離順位が上位（第 1 位が最上位）となる $d^2(i, j)$だけを格納した距離テーブルを用いても，準最適と考えることができるような VCV 素片選択を行える可能性がある．本項では，DTL 選択法による最適な VCV 素片選択について，VCV 素片の各接続点における $d^2(i, j)$の距離順位を，実際の合成音声について調べた．音声合成の対象には見出しを除く新聞記事の本文を用いた．実験に用いた新聞記事の長さは，VCV 合成単位の個数にして 9,888 個，426 文の長さである．VQ コードブックのサイズに関する前節の結果より，VQ コードブックのサイズは N =128 とした.

実験の結果を，合成音声中の VCV 素片の各接続点における $d^2(i, j)$ の距離順位の累積出現頻度として図 7.5 に示す．距離順位 0 は，VCV 素片の接続点で，同じ代表ベクトルどうしが接続されていることを示している．最適な VCV 素片選択では，全接続点の 90% を距離順位 0 の接続が占めている．また，距離順位 20 位までで全接続点の 98% が含まれている．この実験結果から，距離順位が大きな接続はまれにしか起こらないといえ，距離順位が下位 $d^2(i, j)$はほとんど参照されないという先の仮定が正しいことが判る．

以上の議論より，距離テーブルに全ての $d^2(i, j)$を保持せず，距離順位が上位となる $d^2(i, j)$だけを保持することで，最適選択にごく近い準最適選択を行なえる可能性がある．次の項では，このようなアイデアに基づく順位式距離テーブルを用いた場合の VCV 素片選択の選択結果の検証を行なった．

7.3.2　順位式距離テーブル

$d^2(i, j)$ を $i < j$ の場合について全て格納すると，VQ コードブックのサイズを N，1 つの距離情報を格納するために必要な容量を b_d ビットとして，$b_d N (N - 1) / 2$ ビットが必要である．距離テーブルの容量を削減するために，VQ コードブックの各代表ベクトルに距

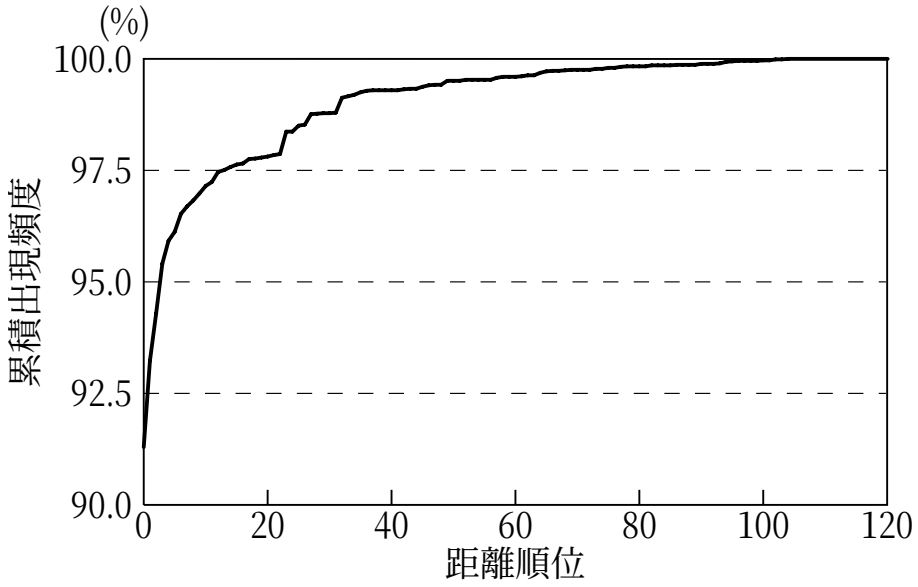


図 7.5　VCV 素片の最適選択における距離順位の出現傾向

離が近い代表ベクトルだけの距離情報を格納しておく**順位式距離テーブル**について検討した．

VQインデックス i を持つ代表ベクトルにLSP 2 乗距離が k 番目に近い代表ベクトルのVQインデックスを $I(i, k)$ と表記する．順位式の距離テーブルには，各代表ベクトル (VQインデックス $i = 0, \cdots, N\text{-}1$) に対してLSP 2 乗距離が近い順に $I(i, k)$ と $d^2(i, I(i, k))$ を M 個($k = 1, \cdots, M$) 格納する．また，距離テーブルに格納されていない距離を概算するために，LSP 2 乗距離が最も遠い代表ベクトルとの距離 $d^2(i, I(i, N\text{-}1))$ も距離テーブルに格納する．距離テーブルに格納されていない代表ベクトル同士の距離は，式(7.3)によって概算する．

$$d^2(i, I(i, k)) = (1 - w)\, d^2(i, I(i, M)) + w\, d^2(i, I(i, N-1)) \quad \cdots\cdots\cdots\cdots \quad (7.3)$$

但し，$M < k \leq N\text{-}1$ であり，w は $0 \leq w \leq 1$ の重み係数である．

このような順位式の距離テーブルを用いた場合，VQインデックスを記録するために b_i ビットが必要とすると，距離テーブルに $(b_d + b_i)MN + b_d N$ ビットが必要である．本手法は，VQコードブックのサイズ N に比べて距離テーブルの順位打ち切り数 M を十分に小さくできれば，距離テーブルの記憶容量の削減が可能である．

順位式距離テーブルの順位打ち切り数 M と，式(3)の重み係数 w を変えてVCV素片を選択する実験を行った．全ての $d^2(i, j)$ を格納した距離テーブルを用いてVCV素片を選択を行った結果を最適選択とし，順位式距離テーブルを用いた場合のVCV素片の選択結果と比較した．音声合成の対象には7.3.1の実験と同じ資料を用いた．VQコードブックのサイズは N =128 とし，M = 32, 16, 8 の場合についてVCV素片選択を行った．

実験で合成した文のうち，最適選択となった文の割合を**最適率**として，図7.6に示した．M の値がどの場合でも，最適率は w の値の増加とともに向上し $w = 0.8$ で最大値に達したが，$w = 0.4$ 以上での最適率の変化はわずかであった．最適率が最大の場合，$M = 32$ では95%の文が最適選択となり，$M = 8$ でも77%以上の文が最適選択となった．

また，w を変えた場合のVCV素片の接続点における平均LSP 2 乗距離の変化を図7.7に示す．図7.7では，最適選択の接続点における平均LSP 2 乗距離を基準とした百分率で

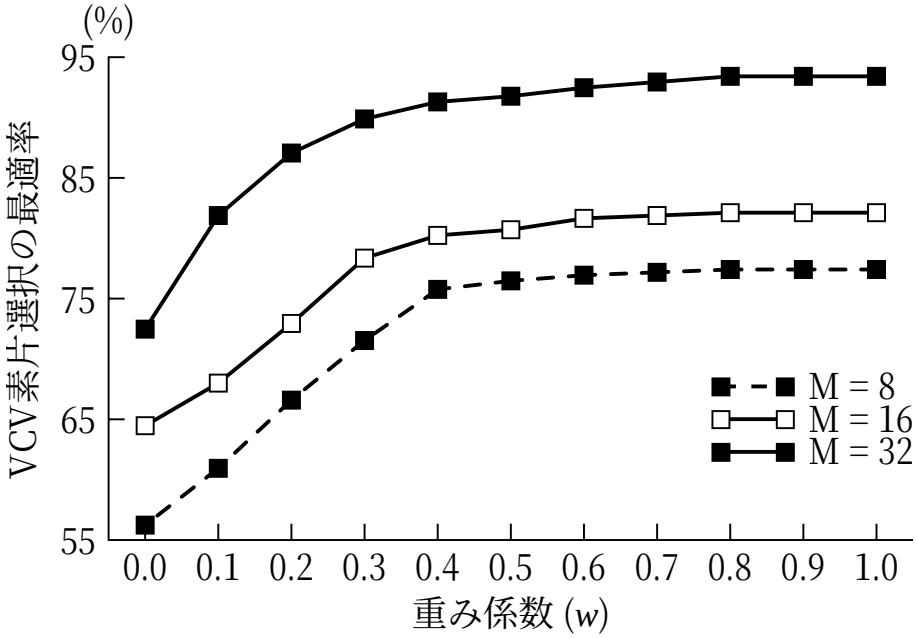


図 7.6　順位式距離テーブルを用いた場合の最適率

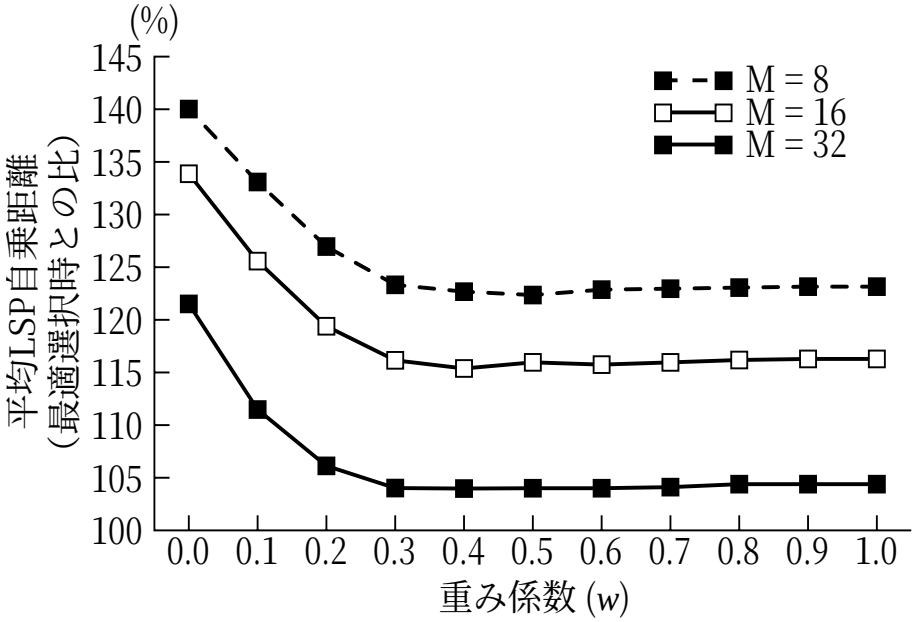


図 7.7　順位式距離テーブルを用いた場合の
VCV 素片接続点における LSP 自乗距離

実験結果の平均LSP ２乗距離を示している．Mの値がどの場合でも，平均LSP ２乗距離はwの増加とともに減少し，$w = 0.3$から$w = 0.5$で最小になった．$w = 0.5$以上では，平均LSP ２乗距離はわずかに増加する．最適選択の場合と比較して，$M = 32$では平均LSP ２乗距離は4%程度，$M = 8$の場合でも24%程度の増加にとどまっている．

最適率を大きくし，VCV素片の接続点における平均LSP ２乗距離を小さくするためには，$w = 0.4$から0.5に設定すれば良い．このとき，距離順位を$M = 8$のように非常に小さな値で打ち切っても, 77%の文が最適選択となり, VCV素片の接続点における平均LSP ２乗距離の増加は24%程度となる．このとき，距離テーブルの大きさは全ての$d^2(i, j)$を記録する場合に比べて1 / 4程度とすることができる．

7.4 まとめ

本章では，LSPベクトルVCV規則音声合成方式におけるベクトル量子化について，適切な代表ベクトル数の検証を行った. また, ベクトル量子化を用いて合成単位辞書の容量を削減するだけでなく, ベクトル量子化の特長を活かして, 代表ベクトル間の距離テーブルを用いて合成単位辞書から適切なVCV素片を効率的に選択するDTL選択法を提案した．これにより素片選択処理の高速化を可能にした．

主観評価実験により本手法のためのベクトル量子化のVQコードブックサイズは128程度で十分であることを示した．このように小さなVQコードブックが可能となったのは, 規則音声合成の性質上, 符号化する音声資料の提供話者が1名に絞られたことが大きく関与していると考えられる．

VQコードブックのサイズを小さく抑えることができたため, 合成単位辞書に14,000個と多くのVCV素片を収録しても，記憶容量は256Kバイト程度と非常に小さなものにでき，VQコードブックと残差波形辞書を合わせても500Kバイト程度で記録できることを示した．また，DTL選択法で順位式の距離テーブルを用いると，距離順位8位でテーブルを打ち切っても，十分な精度でVCV素片の選択が行えることを示した．

以上の結果より，LSPベクトルVCV規則音声合成方式は，多数のVCV素片を収録した

第 7 章　VCV 規則音声合成に適用するベクトル量子化の検討

大規模な VCV 辞書を大幅に情報圧縮でき，DTL 選択法により VCV 素片選択の高速化が行えることを示した．また，本方式によって，ベクトル量子化を適用しない LSP-VCV 規則音声合成方式と比較して，同程度の合成音声品質を得ることができた．

第8章　破裂子音の明瞭性向上のための残差信号の符号化

これまで議論してきたLSPベクトルVCV規則音声合成方式では，合成音声の聞き取りによる評価実験の際に，子音が不明瞭な場合があることが被験者によって指摘されていた．特に，破裂子音を多く含む文例では聞き取りが悪い傾向にある．本章では，記憶情報量の増加を最小限に抑えながら破裂子音の明瞭性を向上させるために，破裂子音残差波形の微少区間パワの包絡形だけを記憶する方法により，少ない容量で破裂子音部の残差信号を符号化するPEC法(Power Envelope Coding法)を提案した．PEC法では，M系列信号に復号化したパワの包絡形をかけることにより破裂子音残差波形を再現する．

PEC法では，破裂子音残差波形の微少区間パワの包絡形をいかに少ないサンプル点数で効率的に符号化するかという点が重要である．我々は，サンプル点数を決めると，サンプル位置制御のための1つのパラメータによりサンプル位置を系統的に決定できる関数を考案した．この関数を使用する事により，実際の破裂子音を用いて，サンプル点数ごとに最適なサンプル位置を決定した．

さらに，サンプル位置を最適化したPEC法による合成音声を用いてVCV型の無意味単語の合成を行い，聞き取り実験により子音の明瞭性を調べた．破裂子音のうち/t/，/g/，/d/，/b/では，PEC法により元の残差信号を用いる場合に近い明瞭性を得ることができ，残差信号の微少区間パワの包絡形を保存するPEC法の有効性が示された．また，/k/，/c/は，M系列音源による合成音声でも，元の残差信号を音源として用いた場合と同程度の明瞭性が得られ，本法によらず改善の必要がないことが示された．[38]

8.1　子音残差信号の符号化

8.1.1　駆動音源としての残差信号

これまで議論してきたLSPベクトルVCV規則音声合成方式では，LSP合成フィルタを駆動するための**駆動音源**を，音韻別に切り出した代表的な**残差波形**を接続することで作成した．この方法は，簡便で記憶情報量が少なく，パルスと白色雑音による完全に人工的な駆動音源より，合成音声の品質がわずかに良い．しかし，合成音声の聞き取りによる評価実験の際に，子音が不明瞭な場合があることが被験者によって指摘されていた．特に，破裂子音を多く含む文例では聞き取りが悪い傾向にある．線形予測分析合成系を用いた音声合成システムにおいて，より合成音声の品質を向上させるために，圧縮した残差信号を保持しておき，それを用いて線形予測フィルタを駆動する方法が報告されている[45][46]．これらの方法は，主に母音など有声音声の品質を改善する目的で残差信号の特徴を用いるものである．

破裂子音の明瞭性を向上させるためには，子音部の駆動音源としてLSP分析で得られた残差駆動をそのまま用いる方法が簡単である．しかし，一種類のVCV単位について多くのVCV素片を持つLSPベクトルVCV規則音声合成方式では，各VCV素片に対して残差信号をそのままの形で保持すると，記憶すべき情報量の極端な増加につながる．このような方法ではLSPベクトルVCV規則音声合成方式によって小規模な音声合成器が実現可能であるという利点が失われる．

筆者は，破裂子音の聴覚的な特徴が，その破裂点での波形の包絡形状，つまり微少区間パワの変化にあると考えた．この考えに基づき，本章では，音声合成システムが保持すべき情報量を極端に増やすことなく破裂子音の明瞭性を向上させるために，残差信号波形の時間的な包絡形状だけを保存することの有効性を実験的に調べた．この目的のために，残差信号の時間的な包絡形状（**パワ・エンベロープ**）をサンプルして符号化する圧縮率の高い符号化法を提案した．また，この符号化手法において，パワ・エンベロープのサンプル位置を決定する手法を8.2節で述べる．これらの手法を含め，本論文で提案する残差信号の符号化／復号化の手法を，**PEC法**(Power Envelope Coding 法)と呼ぶ．

8.1.2　破裂子音部の残差信号

音声信号をLSP分析した際の残差信号は，スペクトル領域で白色化されており，そのスペクトル包絡は平坦である. 音声信号に含まれる情報のうちでスペクトル包絡が担う情報は，LSPパラメータが保持している．残差信号に残された破裂子音の特徴は，時間領域における波形の破裂点付近での包絡形状にあると考えられる．

破裂子音部分の残差信号の例を図8.1に示す．図8.1は，標本化周波数11.025kHz，量子化数16ビットでサンプリングした残差信号を，破裂子音部分の破裂点を中心に256サンプル，約23ms切り出した波形である．残差信号の微細な構造は雑音的であり，子音の種別による違いは見られない. 残差信号の包絡形を考え，その立ち上がりや立ち下がりを観察すると，子音の種類により大きな差異が見られる．図8.1の例では，/ k /の包絡形は鋭く立ち上がり緩やかに立ち下がる. / c /の包絡形は立ち上がりも立ち下がりも非常に緩や

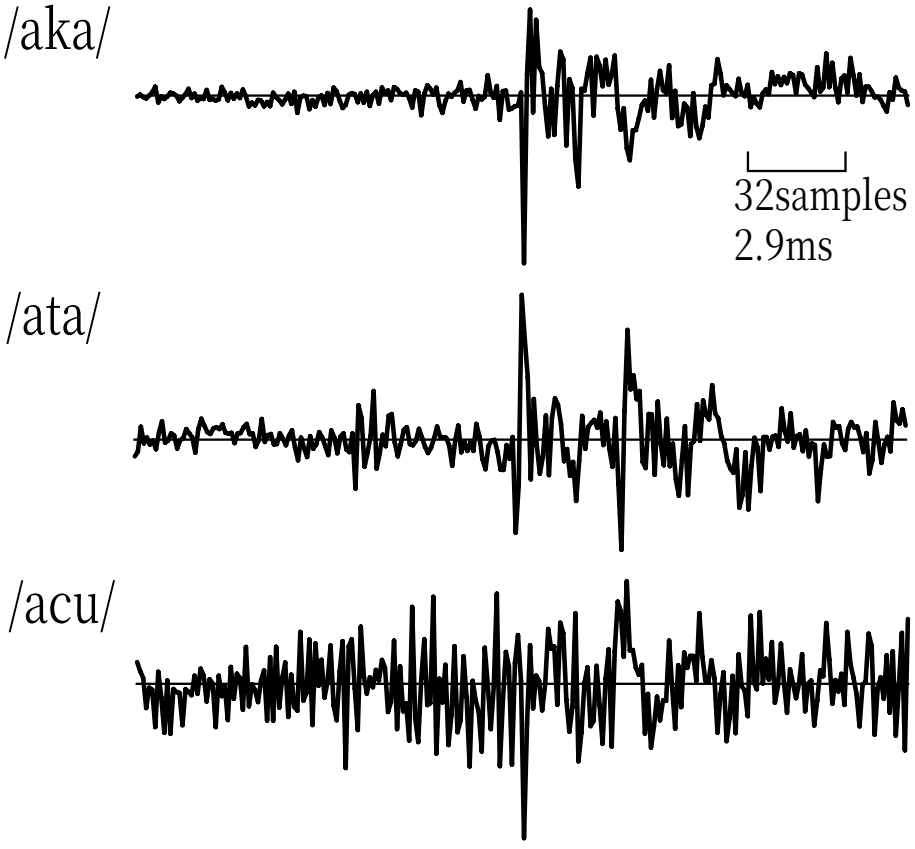


図 8.1　破裂子音の残差信号の例

かである．また，多くの子音を観察すると，同じ子音でも，発話毎に包絡形がかなり異なる場合があることも観察された．

先の章で議論してきた LSP ベクトル VCV 規則音声合成方式では，同じ VCV 単位のどの VCV 素片に対しても，同じ代表残差を用いて合成音声を生成していた．しかし，同じ子音でも発話毎に包絡形がかなり異なる場合があるという事実は，少なくとも破裂子音部に関しては代表残差を用いる方法が不十分であることを示している．これは，合成音声の聞き取りによる評価実験の際に得た子音が不明瞭な場合があるという被験者の指摘とも一致する．

8.1.3　PEC 法による破裂子音残差信号の符号化と復号化

本章で提案した PEC 法(Power Envelope Coding 法)による破裂子音残差信号の符号化および復号化のブロック図を図 8.2 に示す．PEC 法では，残差信号の波形の包絡形を求めるために，残差信号の**微少区間パワ**を用いている．具体的には，破裂点を中心として切り出した N 点の残差信号 $s(n)$, (n= 1, 2, … ,N)の各点で近傍 K 点の微少区間平均パワ $p(n)$ を式(8.1)により求める．以後，$p(n)$ を，平均パワから求めた子音の時間波形の包絡形という意味で，パワ・エンベロープと呼ぶ．以後の実験では，標本化周波数 11.025kHz，量子化数 16 ビットでサンプリングした音声資料に対して，符号化する子音長 N = 256 点とし，微少区間平均パワ計算のための近傍点数 K = 16 点とした．

$$p(n) = \sqrt{\frac{1}{K}\sum_{k=-\frac{K}{2}}^{\frac{K}{2}-1} s(n+k)^2} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (8.1)$$

パワ・エンベロープ $p(n)$ は，元の波形 $s(n)$ の包絡形であり，$s(n)$ に比べて微細な構造を失っており，ゆっくりとした時間変化を示す．このため，$p(n)$ から代表的な点をサンプルすることで情報圧縮符号化を行っても，$p(n)$ の示す概形は失われにくい．PEC 法では，破裂の中心点 1 点と，その前後に M 点ずつ合計 $2M + 1$ 点をサンプルする．M は，(8.1)式によって計算されたパワ・エンベロープ $p(n)$ から符号化のために取り出すサンプル点数を決定するパラメータである．

破裂子音残差信号の復号化では，パワ・エンベロープ$p(n)$からサンプルされた$2M$+1点のデータを線形補間して，パワ・エンベロープを再現する．サンプル点数とサンプル位置が適切ならば，再現されたパワ・エンベロープ$\tilde{p}(n)$は元のパワ・エンベロープ$p(n)$を十分近似できる．サンプル点数とサンプル位置については，次節以降で詳しく議論する．

白色雑音である**M系列**に再現されたパワ・エンベロープ $\tilde{p}(n)$ をかけることで，残差信号を再現する．再現された残差信号 $\tilde{s}(n)$ を合成残差と呼ぶ．

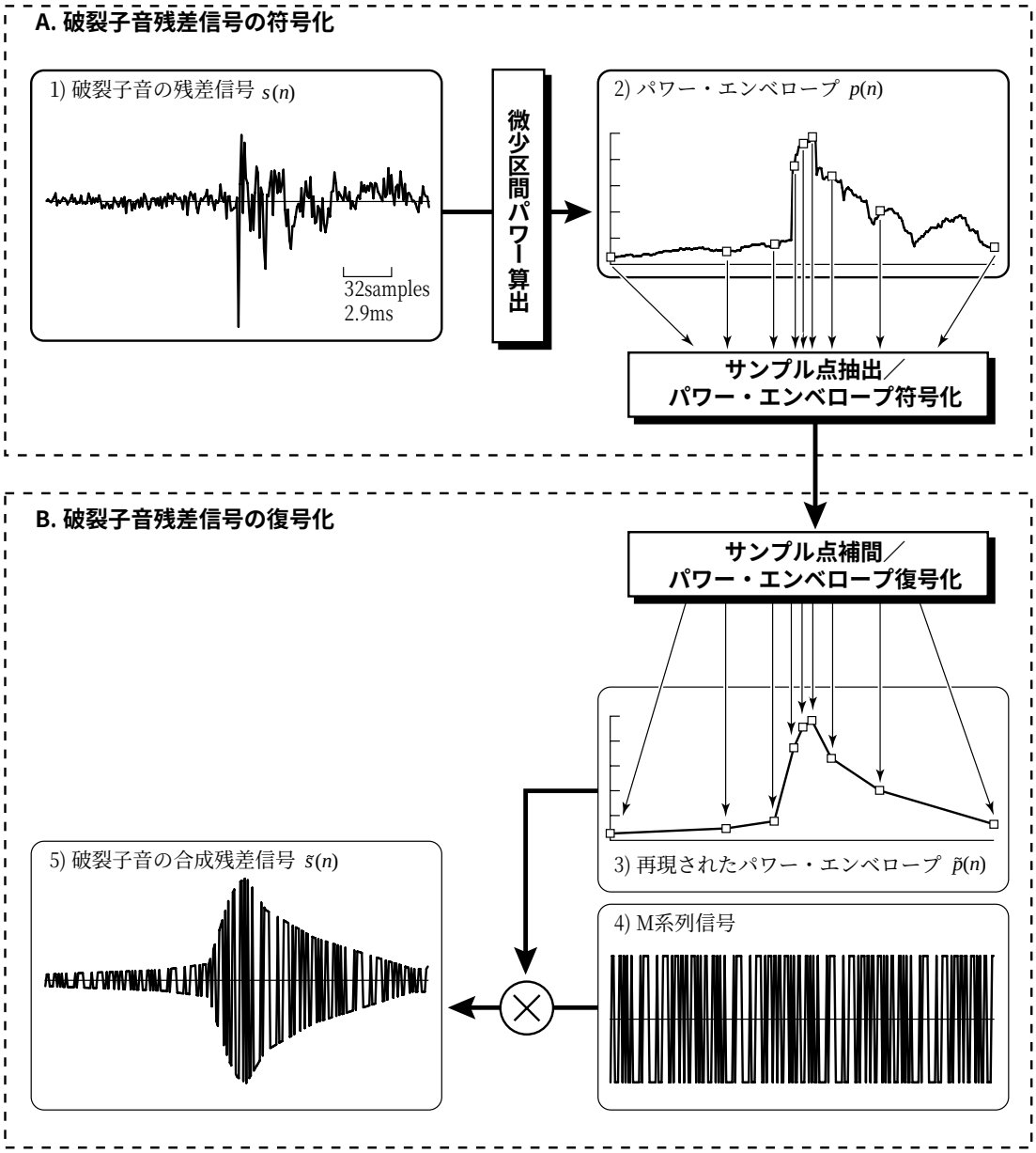


図 8.2　PEC 法による破裂子音残差信号の符号化と復号化

8.1.4　パワ・エンベロープのサンプル位置

PEC 法では，記憶する情報量を減らすために，パワ・エンベロープ $p(n)$ を全ての時点で記憶せず，その中からサンプルした時点の値のみを記憶する．破裂子音の残差信号では，子音の破裂点近傍のパワ・エンベロープの形状が重要であると考えられるため，破裂点に近いほど密にサンプルする事が望ましいと考えられる．しかし，合成音声の聴覚上の品質の点から，破裂点近傍をどの程度密にサンプルすれば良いかは自明ではない．これを知るためには，時間軸上での**サンプル位置**を系統的に制御して，実際の残差信号を用いて実験的に最適なサンプル位置を求める必要がある．

本項では，サンプル位置を決定するために，式(8.2)に示すサンプル位置の制御式を考案した．式(8.2)の制御式は，サンプル位置の破裂点からの相対距離 r_i を与える式となっている．

$$r_i(M,\alpha) = \alpha\left\{ \exp\left(\frac{1}{M}\ln\left(\frac{N}{2\alpha} + 1 \right)\cdot i \right) - 1 \right\},\ (0 \le i \le M) \quad \cdots\cdots\cdots\cdots \quad (8.2)$$

式(8.2)において，N は符号化する子音部の長さ(N = 256)である．M は，破裂点の前後にサンプル点数である．αは，サンプル位置の時間軸上での分布を制御するパラメータで，以後，**符号化パラメータ**と呼ぶ．

図 8.3 に，N = 256, M = 8 としたときの，符号化パラメータ α と式(2)から得られる相対位置 r_i の関係を示した．図 8.3 から読み取れるように，αを大きくすれば，式(8.2)のグラフは直線に近付き，サンプル位置は時間軸上で均等になる．一方，αを小さくすれば，式(8.2)のグラフは下に凸となり，0 に近い方，つまり破裂点の近傍が密に，破裂点から遠い時点は疎にサンプルされる．符号化パラメータαは，破裂点近傍をどれ程重点的にサンプリングするか，あるいは全体的に均等にサンプリングするかを制御するパラメータである．

PEC 法では，式(8.2)の r_i を用いて，式(8.3)でサンプル位置を決定する．

$$j_m = \frac{N}{2} + \mathrm{sign}(m)\cdot r_{|m|},\ (-M \le m \le M) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (8.3)$$

ただし，j_m が整数でない場合は，小数点以下を四捨五入して整数値とする．

図 8.4 に，破裂子音 / k / のパワ・エンベロープ $p(n)$ と，サンプルされた $2M+1$ 点($M=8$)を線形補間して再現したパワ・エンベロープ $\tilde{p}(n)$ の例を示す．図中段は符号化・復号化前のパワ・エンベロープ $p(n)$ であり，上段はパラメータ α を小さく 4 とした場合の再現例，下段はパラメータ α を大きく 200 とした場合の再現例を示している．α を小さくすれば, 破裂点近傍が良く再現されているのに対して破裂点から離れた位置は大雑把な再現となっている．α が大きいと，全体的な形は再現されているが，破裂点近傍の立ち上がりなどは鈍くなっている．次の節では，α をどれほどの値にすれば良いか，パワ・エンベロープの誤差に着目して実験的に検証した．

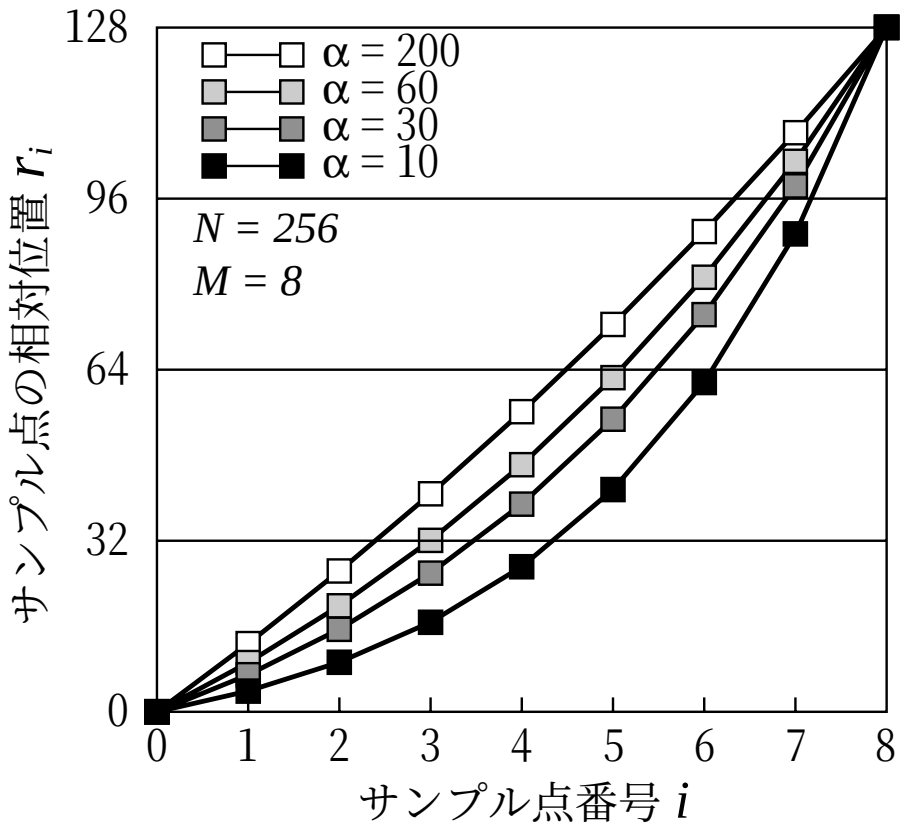
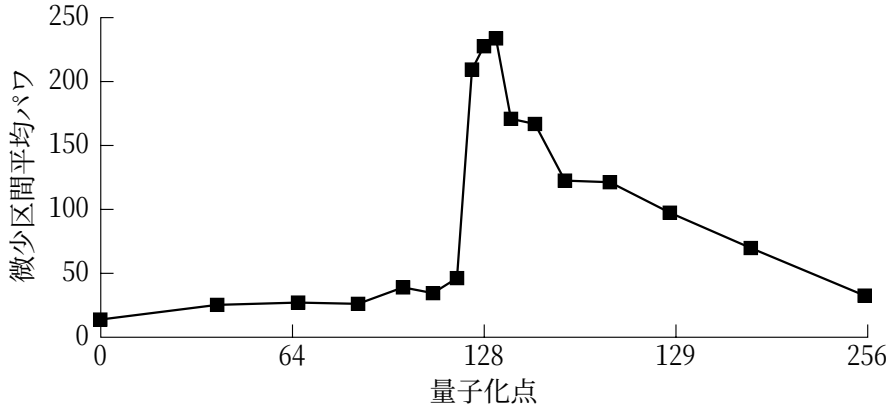
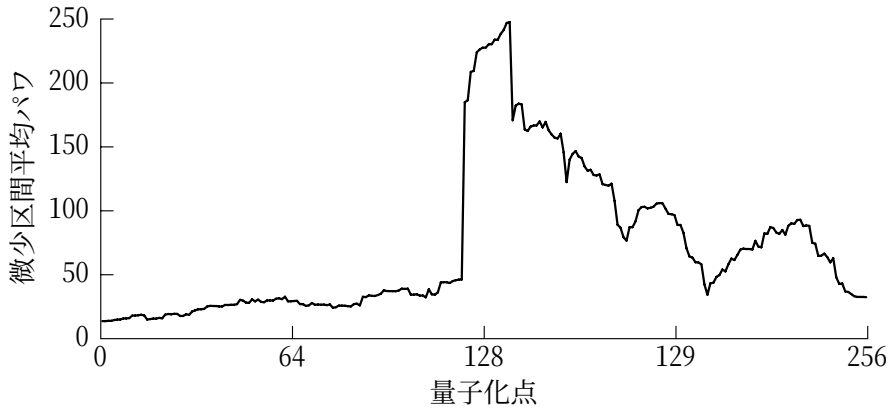


図 8.3　符号化パラメータ α とパワーエンベロープのサンプル点との関係

再現されたパワ・エンベロープ $\tilde{p}(n)$　($M = 8, \alpha = 4$)



パワ・エンベロープ $p(n)$



再現されたパワ・エンベロープ $\tilde{p}(n)$　($M = 8, \alpha = 200$)

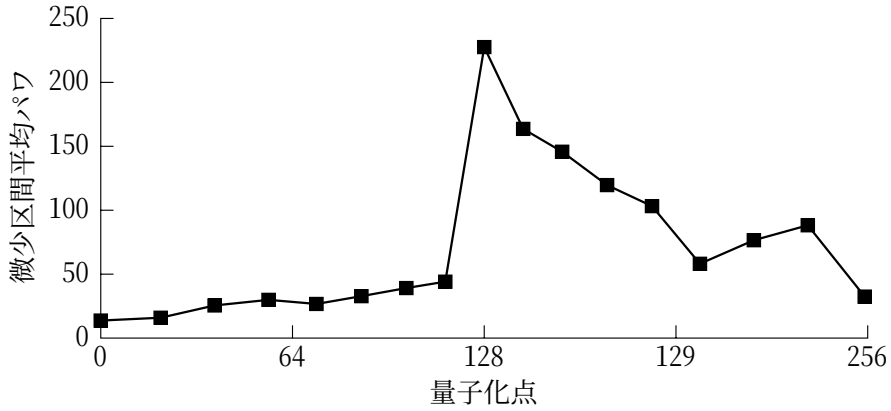


図 8.4　パワ・エンベロープの再現例

8.2　符号化パラメータαの決定

8.2.1　音声資料

PEC法の主観評価に先立ち, 符号化パラメータαの最適値を決定するための数値実験を行った．実験には，20代男性話者1名が読みあげたVCV素片を用いた．話者には，破裂子音/ k /, / t /, / c /, / p /, / g /, / d /, / b /を含む140種類のVCV単位を単独で3回ずつ読みあげるよう依頼し，計420個のVCV素片を採取した.

採取した音声資料は，表8.1に示す仕様でサンプルを行い，破裂子音の破裂点に目視でマーキングを施した．また，同じく表8.1に示す仕様でLSP分析を行い，LSPパラメータと残差信号を算出して以後の実験に用いた.

8.2.2　対数概形誤差

PEC法において, 符号化パラメータαの最適値を決定するための評価指標として, 破裂子音残差信号のパワ・エンベロープ$p(n)$を符号化し復号化した時の誤差を式(8.4)の**対数概形誤差**を定義した.

$$E = 10\log\sum_{n=1}^{N}\left(\frac{\tilde{p}(n)-p(n)}{p(n)}\right)^2 \quad \cdots\cdots (8.4)$$

表 8.1　音声資料の分析条件

サンプリング条件	
標本化周波数	11.025kHz
量子化数	16bits
LSP分析条件	
分析次数	14次
フレーム長	256点
インターバル長	64点

式(8.4)において，$\tilde{p}(n)$ は PEC 法で復号化されたパワ・エンベロープである．対数概形誤差 E は，破裂子音残差信号のパワ・エンベロープ $p(n)$ を時間信号波形に見立てて，SN 比の逆数を計算したものである．対数概形誤差 E は，値が小さいほど，パワ・エンベロープが正しく再現されていることを示している．

前項で述べた音声資料を用い，サンプル点数(2M+1)と符号化パラメータ α を変えて，420 資料ある全ての VCV 素片についてパワ・エンベロープ $p(n)$ を符号化し復号化する実験を行った．サンプル点数(2M+1)と符号化パラメータ α の条件毎に，対数概形誤差 E の全 VCV 素片についての平均値を求めた結果を図 8.5 に示す．

図 8.5 では，様々な M の値について，符号化パラメータ α を変化させた場合の平均対数概形誤差の曲線を示している．M の値が大きくなれば，パワ・エンベロープのサンプル点数が増加するために，パワ・エンベロープがより正しく再現されるようになり，平均対数概形誤差は小さくなっている．また，同じ M では，符号化パラメータ α を横軸として全ての曲線が下に凸の曲線になっており，平均対数概形誤差が最小となる符号化パラメータ α が存在する．これは，以下の 2 つの理由によるものと考えられる．

1) αが小さ過ぎると破裂点近傍だけを密にサンプルすることになり，それ以外の点で誤差が大きくなる．
2) αが大き過ぎると，全体が均等にサンプルされ，破裂形状が複雑な破裂点近傍が再現されにくくなり誤差が大きくなる．

M が大きくなると，サンプル点数が増えるため，パワ・エンベロープを比較的均等にサンプルしても，破裂点近傍の形状を再現しやすい．このため，M が大きいほど上記 2)の影響が少なく，符号化パラメータ α が大きい方での平均対数概形誤差の増大が緩やかになっているものと考えられる．

図 8.5 より，M = 4 の時は α = 30 で平均対数概形誤差は最小値となり，それ以外の M では α = 60 で最小値となっている．この結果より，次節で述べる主観評価実験では，M = 4 の時は α = 30，それ以外の M では α = 60 として，破裂子音残差の符号化・復号化を行った．

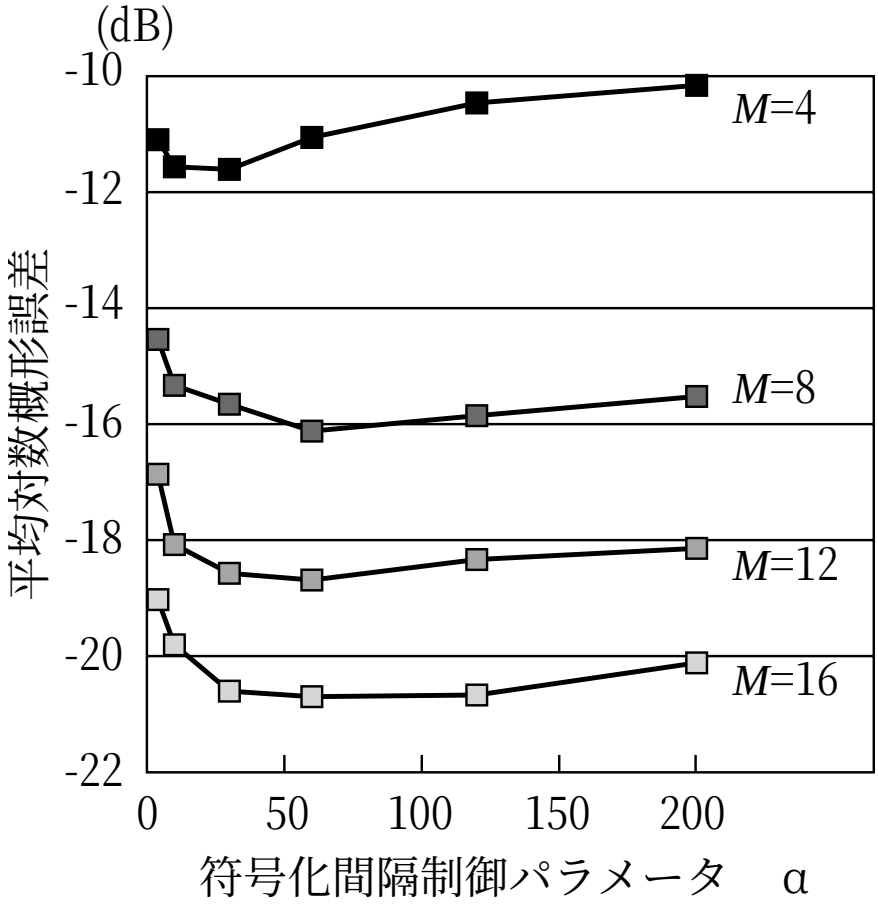


図 8.5　符号化パラメータ α が対数概形誤差に与える効果

8.3　明瞭性の主観評価実験

PEC法の有効性を調べるために, **明瞭性**についての主観評価実験を行った. 本節で述べる主観評価実験は, 第4章や第7章で用いた1対比較による品質評価ではない. 被験者に, 提示した音声資料の発話内容を書き取らせる方法により, 音声資料が正しく聞き取られているかを評価する**明瞭度試験**である. 以下, 実験の詳細について述べてゆく.

8.3.1　音声資料

PEC法による合成音声の明瞭性の主観評価のために, 8.2節の実験で用いた音声資料から, 各VCV単位について, 1素片ずつ140素片を選んで実験に用いた. 被験者が意味を推定して回答するのを避けるために, 無意味単語であるVCV素片を提示資料として選んだ. また, 子音の前後の母音の影響を除くために, 破裂子音/ k /, / t /, / c /, / p /, / g /, / d /, / b /について, 日本語発話で通常用いられる全ての組み合わせのVCV単位合計140種類を実験に用いた. 音声資料のサンプリングとLSP分析の仕様は, 8.2節の実験と同じく表8.1の通りとした.

VCV素片の母音部は, パルス駆動のLSP合成で作成し, 子音部分には, 次の6種類の駆動音源を用いた音声資料を作成した.

a)　M系列(白色ノイズ)

b)　PEC法による音源 ($M = 4, \alpha = 30$)

c)　PEC法による音源 ($M = 8, \alpha = 60$)

d)　PEC法による音源 ($M = 16, \alpha = 60$)

e)　PEC法による音源 (全点)

f)　原残差信号

e) は, パワ・エンベロープをサンプルせず, 全点を保存して, 合成残差を作成したものである. a) と f) は, PEC法との比較のために実験に加えた. 被験者に提示する音声資料は, VCV140素片×6種類の手法 = 840資料である. これらの音声資料は, 原音声資料のサンプリング時と同じ表8.1の仕様で合成した. 従って, M系列は標本化周波数11.025kHzで, 周期255点のものをプログラムにより生成して用いた.

8.3.2　明瞭性の主観評価実験方法

8.3.1 項で述べた 840 資料をランダムに被験者に提示し，発話内容を平仮名で書き取らせる方法で明瞭性の主観評価を行った．

被験者の負担を考えると，840 個の音声資料は，1 回の実験で続けて全て提示するには多すぎる．このため，全資料の 1/10 の 84 資料を 1 セットとし，1 回の実験では 1 セット毎に休憩を挟みながら 5 セットの資料を提示した．1 回の実験の所要時間は，実験要領の説明を含んで約 50 分である．被験者は 2 回の実験で 840 資料全ての音声資料を聞き取ることになる．1 名の被験者について 6 回の実験を行ったため，各被験者は同じ資料を 3 回ずつ聞き取ったことになる．

被験者に音声資料を提示する際, 資料の提示順序が実験結果に影響を与える可能性がある. この影響を軽減するために, 被験者毎に音声資料を提示する順が変わるように提示資料のセットを作成した. また, 同じ被験者が, 同じ資料並びを持つセットを聞き取ることがないようにした.

提示資料のセットに含まれる各音声資料は, ヘッドフォンを用いて被験者に続けて2回ずつ提示した．また，音声資料提示の後に 3 秒間の回答時間を設け，聞き取った内容を回答させた．回答は，この実験用に作成した専用の回答用紙の回答欄に平仮名で記入させた．無回答は許さず，強制回答とした．

被験者には，提示資料が VCV 素片であることは伝えず，提示資料に含まれる音韻数も知らせずに, 無意味単語である事だけを知らせた. このため, 解答欄に記入する文字数にも制限を付けず，自由な記入を行わせた．これらは，聞き取り実験の条件としては，かなり厳しい条件である. また, 実験期間中に, 被験者の間で提示された音声資料について情報交換を行うことを禁止した.

実験に参加した被験者は，音声研究に従事した経験が無い日本語を母国語とする 20 代の男女 13 名である．実験実施にあたっては，被験者に実験内容の十分な説明を行い，書面による実験参加の承諾を得た．

8.3.3　被験者の慣れの効果

本節で述べている明瞭性の主観評価実験において, 実験開始初期の段階では, 被験者が実験に不慣れなことによって実験の結果が不安定になると考えられる，実験を繰り返すと, 被験者が実験に慣れ, 安定した実験結果を示すようになる. 主観評価実験におけるこのような効果を被験者の**慣れの効果**と呼ぶ. 被験者の慣れの効果を検証するために, 全被験者についての実験のセット毎の子音正解率を調べた.

各被験者には，全 6 回の実験で 30 セットの音声資料を提示している．被験者に提示した音声資料のセットに, 提示順にセット番号を付けた. 被験者によって, 音声資料の提示順が異なるため, 被験者が異なれば同じセット番号のセットでも同じ音声資料のセットではない．これは, 前項で述べた通り資料の提示順序の影響を軽減するための措置である. セット毎の子音正解率とは，提示した 1 セット 84 資料中で被験者が子音を正しく答えた正解率である.

図8.6に, 聞き取り資料のセット番号に対する子音正解率の平均とその標準偏差を示す. これは，同じセット番号のセットについて，全被験者の子音正解率を平均したものである．図 8.6 に示した子音正解率には，慣れの効果が出ており，セットを重ねるごとに，正解率が向上していることが判る．また, 実験初期のセットでは, 標準偏差が大きく, 後になる程，標準偏差が小さいことも読み取れる．これは以下の理由によると考えられる.

1) 実験の初期段階では，不慣れのために正解率が低い被験者がいる一方で，最初から高い正解率を示す要領のよい被験者もあり，正解率の標準偏差が大きくなる.
2) 実験の回数を重ねると，全体的に被験者が実験に慣れ，どの被験者も正解率が高くなり，正解率の標準偏差が小さくなる.

正解率の標準偏差のこのような振る舞いは, 被験者ごとに慣れの程度が異なり, 早くから正解率が安定する被験者もあれば, 慣れの効果が長く続く被験者もいることを示している.

各被験者別の各セットでの正解率が安定するまでは, その被験者に慣れの効果が続いている期間と言うことができる. この期間においては, 被験者の解答の信頼性が低いと考え

られる. このため, 慣れの効果が続く期間のセットを破棄し, 以後の結果の統計処理には加えなかった. 前述のように, 慣れの効果が続く期間は被験者ごとに異なるため, 破棄するセットは各被験者ごとに決定した.

8.3.4　子音別の誤聴率と誤答率

音声資料の聞き取り実験における被験者の回答結果から, 子音別に**誤聴率**と**誤答率**を算出した. 誤聴率は, ある子音を含む音声資料を提示したときに, 被験者が誤って提示したものと異なる子音であると回答した割合である. すなわち, 誤聴率は提示した子音別に集計した回答の誤り率である. 一方, 誤答率は, 被験者がある子音であると回答したときに, その回答が提示された音声資料と異なっていた割合である. 言い換えると, 誤答率は被験者の回答音韻別に集計した回答の誤り率である.

一般に, 誤聴率が低い子音は正しく聞き取られており, 明瞭性が高いと言える. しかし, 聞き取れない音声資料を特定の音韻と答える傾向が被験者にある場合, 提示資料の明

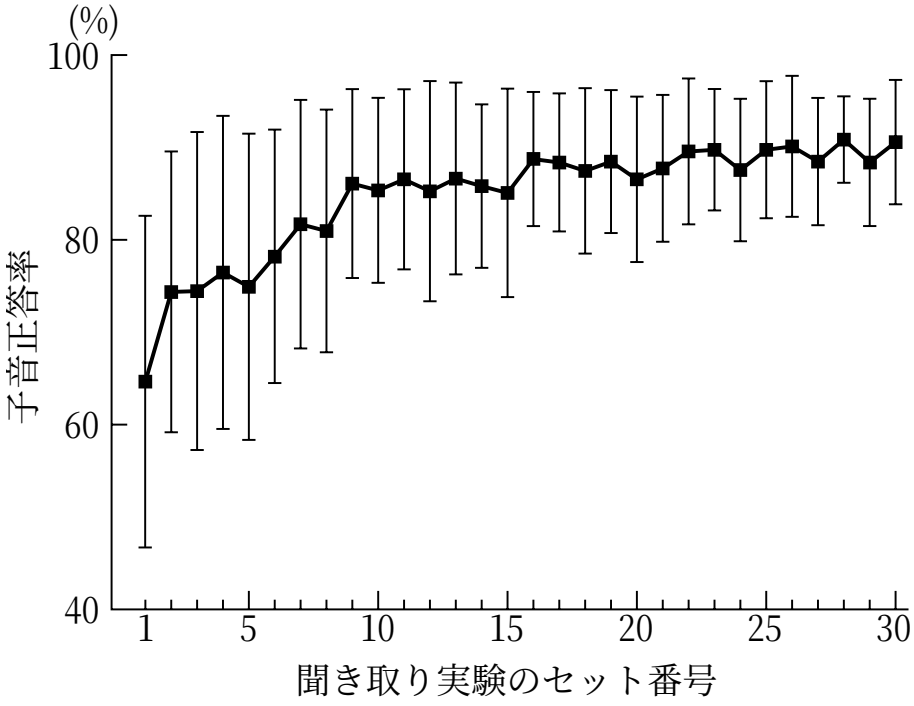


図 8.6　被験者の慣れの効果

瞭性に関わらずその子音の誤聴率が低くなる可能性がある. 誤答率は, 不明瞭な子音がその子音として聞き取られる傾向の強さを示している．従って，誤答率を検証することで，誤聴率による評価の問題点を回避できる．

図8.7に子音別の誤聴率のグラフを示す．PEC法による明瞭性の改善に関して，子音を3つのグループに分類することができる．/ t /，/ g /，/ d /，/ b / の各子音ではM系列音源より，PEC法の誤聴率が低く，残差信号音源に近い値となっている．これらの子音では，残差信号波形の時間的な包絡形状だけを保存するPEC法が有効であると考えられる．これらの音韻の改善の傾向から，PEC法で効果をあげるためには $M = 8$ で十分であると言える．

/ k /，/ c / については，PEC法はM系列音源と同等か，高い誤聴率となっている．しかし，これらの音韻については，M系列音源でも低い誤聴率になっている．従って，PEC法の効果が無いというより, M系列音源で十分な明瞭性が得られており, 改善の余地が無

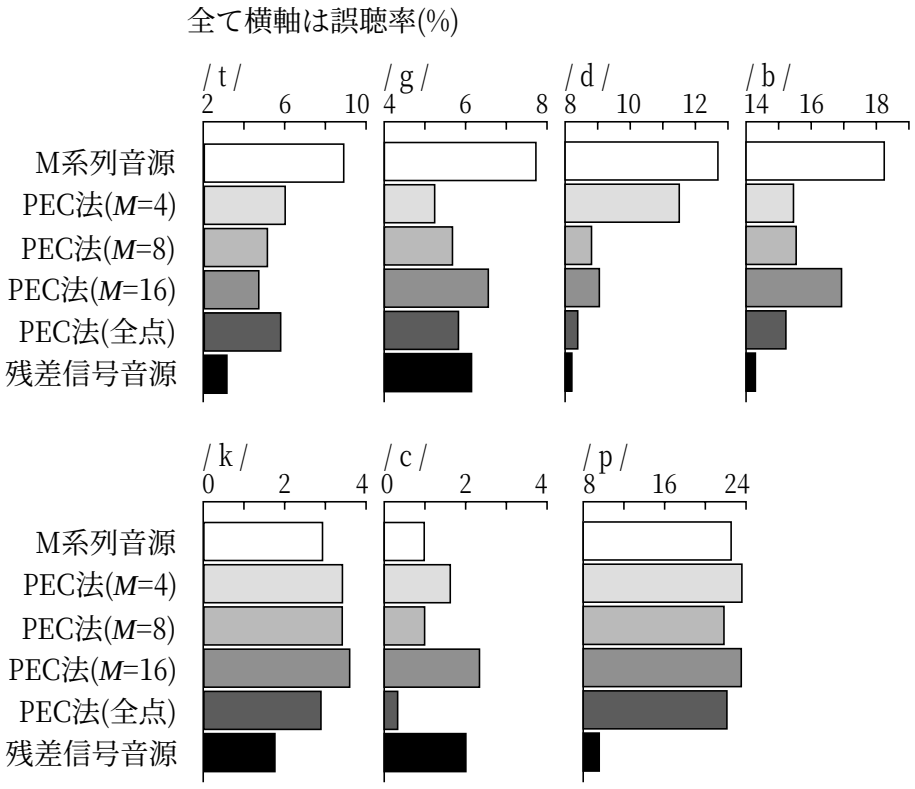


図 8.7　子音の誤聴率

いと考えるべきであろう．/ p /については，M系列音源に比べて残差信号音源の誤聴率が低く改善の余地があるにもかかわらず，PEC法によって明瞭性が全く改善されていない．

図8.8に子音別の誤答率のグラフを示す．誤答率に関しては，/ k /が一際高い値を示している．これは，判断の付きにくい場合には，被験者は/ k /と答える傾向があることを示している．/ d /，/ b /，/ c /については，M系列音源でも誤答率が低く，PEC法による改善の余地が少ない．特に/ c /に関しては，誤聴率，誤答率ともに非常に低く，被験者が確実にこの音韻を識別している事を示している．

程度の差はあるが，/ t /，/ k /では，PEC法によって誤答率が改善している．/ g /，/ p /については，誤答率の改善は見られない．/ c /の場合とは反対に，/ p /については，残差音源の場合を除き，誤聴率，誤答率とも高くなっている．これは，合成音声の品質が少しでも下がると，被験者にとって/ p /を識別することが非常に難しくなることを示している．

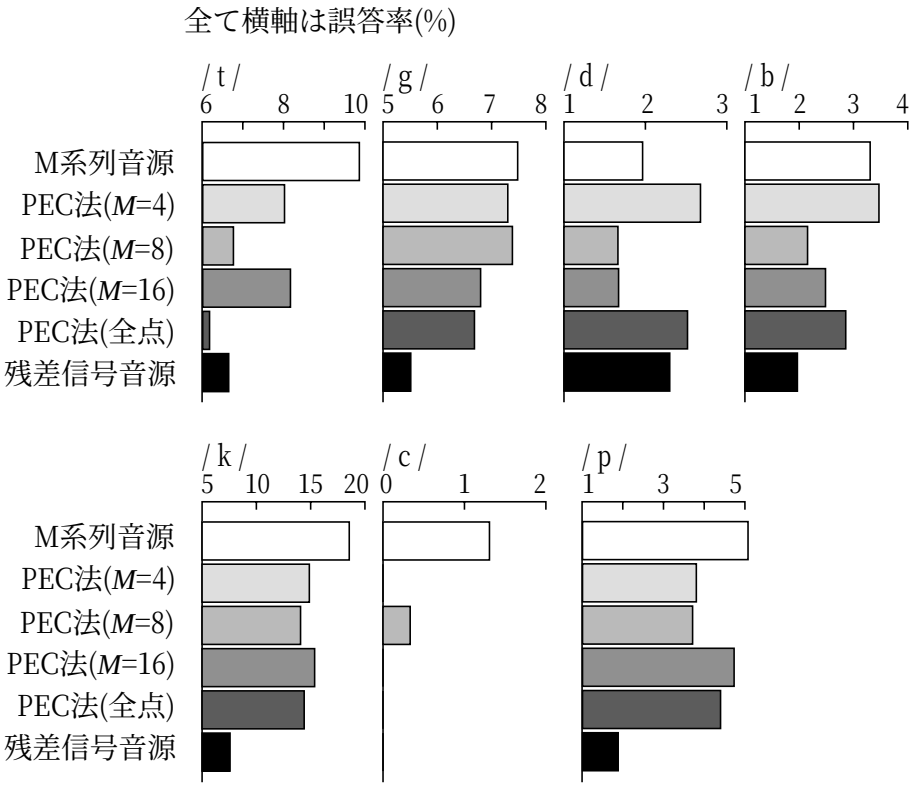


図 8.8　子音の誤答率

8.3.5　子音別の異聴傾向

M系列音源と残差信号音源で明瞭性に差がなく, 改善の余地がない/ k /, / c /を除いて, PEC 法により明瞭性が改善された / t /, / g /, / d /, / b / の各子音と, 明瞭性が全く改善されなかった / p / について, どの子音に間違えて聞き取られているか**異聴傾向**を調べた. 図 8.9 に各子音の異聴傾向を示した. 図表が煩雑になるのを避けるため, PEC 法は $M = 8$ の場合だけを図示した. 図は, 該当音韻に対する回答のうち, それ以外の子音として回答された率を, 回答された子音別に示している. 自由書き取りの方法による実験を行ったため, 小数だがバラエティに富んだ誤答も存在する. これらは個別に分けて評価する程の割合を占めないため, まとめてその他として評価した. その他の中には, 子音の抜けや拗音の挿入などが含まれている.

/ t / は, 残差信号音源を用いた原音声に近い合成音声では, 主として / d / に間違えられる. しかし, M系列音源を用いて, 合成音声の品質が低下すると, / k / やその他への誤答が増える. PEC法では, これらの部分が改善されており, / d / への異聴は3者の間に差が無い. / g / , / d / はそれぞれ / k /, / t / への異聴が起こりやすいが, いずれも PEC 法で改善されている.

/ b /は, 明瞭性が改善されたグループの中では, 残差信号音源を用いた原音声に近い合成音声でも高い誤聴率を示した子音である. 異聴傾向を調べると, / b / は / g /, / p / とともにその他に間違えられる傾向がある. PEC 法によって, / g /, / p / への異聴は改善されている. しかし, その他への異聴は, 残差信号音源を用い場合でも非常に高く, 改善の余地はない. これが / b / の誤聴率が全体として高くなっている原因である.

M 系列音源を用いた合成音声では, / p / は, / k / とその他への異聴が多い. PEC 法では, / k /への異聴はわずかに改善しているが, ほぼ同じ率だけその他への異聴が増加している. これは, PEC法により異聴傾向が変化しただけで, 全体としての誤聴率は改善されなかったことを示している.

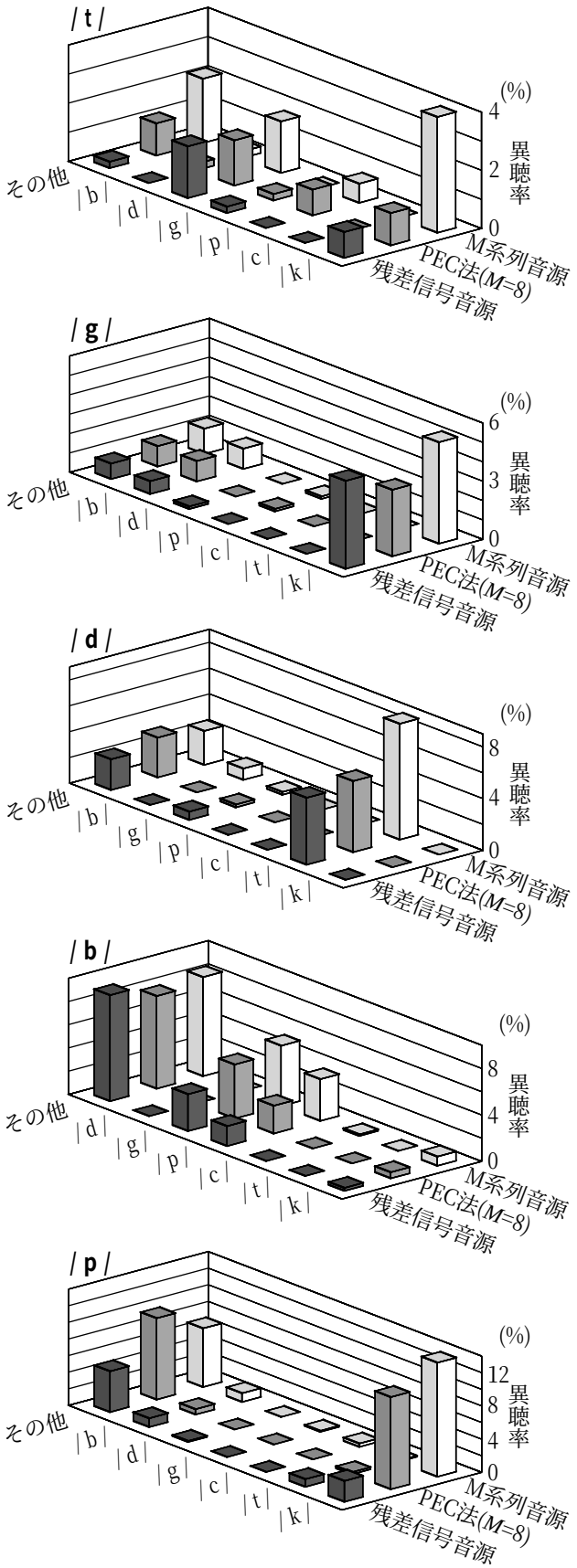

/ t /
(%)
4
2
0
異聴率
その他
b
d
g
p
c
k
M系列音源
PEC法(M=8)
残差信号音源
/ g /
(%)
6
3
0
異聴率
その他
b
d
p
c
t
k
M系列音源
PEC法(M=8)
残差信号音源
/ d /
(%)
8
4
0
異聴率
その他
b
g
p
c
t
k
M系列音源
PEC法(M=8)
残差信号音源
/ b /
(%)
8
4
0
異聴率
その他
d
g
p
c
t
k
M系列音源
PEC法(M=8)
残差信号音源
/ p /
(%)
12
8
4
0
異聴率
その他
b
d
g
c
t
k
M系列音源
PEC法(M=8)
残差信号音源

図 8.9　子音の異聴傾向

8.4　まとめ

LSPパラメータ合成を用いたVCV規則音声合成法の破裂子音部の明瞭性の向上を目的として, 破裂子音部の残差信号波形の時間的な包絡形状だけを保存し, 合成残差を生成する PEC 法を提案し，その有効性を検証した.

破裂子音のうち / t /，/ g /，/ d /，/ b / の各子音については，PEC 法により明瞭性が向上し，残差信号波形の時間的な包絡形状だけを保存する方法の有効性が示された．また，/ k /，/ c / については，M 系列音源による合成音声でも，元の残差信号を音源として用いた場合と同程度の明瞭性を示した. 従って, これらの子音については, M系列音源による簡易な駆動音源を用いても十分な明瞭性を持つ合成音声を生成出来る事が示された.

/ p / については，PEC 法によって明瞭性が全く向上しなかった．異聴傾向を調べると，PEC法によって, / p / の異聴傾向は変化するが, 全体としての誤聴率の向上につながらなかったことが示された．この点については，今後，より詳しい検討が必要である.

第9章　ウェーブレットを用いた LSP分析予測残差の符号化

前章まで，LSPベクトルVCV規則音声合成方式を提案し，小規模な計算機資源で高品質な合成音声を生成する規則音声合成方式について論じてきた. 提案した方法により, 明瞭性の高い合成音声を生成できることを示してきたが, 合成音声の自然性については問題点も残している．現在，数10メガ・バイトから100メガ・バイト以上の波形辞書を用いて, 波形重畳により自然性に優れた音声合成手法が提案されている. しかし，このような方法は，本研究で目指しているような小規模な分野には適用不可能である.

本論文で議論している手法において, 合成音声の自然性が損なわれる主な原因は, 音声合成にLSP分析のような分析合成系を用いていることである. 分析合成系を用いる手法でも, 疑似音響フィルタの駆動に残差信号を直接用いれば高い自然性を保てるが, 情報量は波形を直接用いる方法と同等になってしまう. 記憶する情報量の増加を最小限に抑えて符号化し, 合成音声の品質上において十分な精度で復号化できる高能率な残差信号符号化の手法を考案できれば, 上記の問題の多くが解決する. 本章では，この問題を解決する一手法として，ウェーブレット変換を用いて残差信号を符号化する方式を提案する. [39] 本法は，規則音声合成にとどまらず，広く音声符号化に適用可能な手法である.

9.1　ウェーブレット変換

9.1.1　ウェーブレット変換による情報圧縮

近年では**ウェーブレット変換**が音声信号に適用され様々な成果をあげている[47]. 音声の基礎的な分野では, 音声の駆動源の不連続性の検出と特徴づけのためにウェーブレットを用いる研究が行われている[48]. 音声認識の分野では，音声信号の時間－周波数解析にウェーブレット変換を用いる研究[49]などが行われている. また，音声符号化の分野ではウェーブレットを用いて高品質のオーディオ信号の情報圧縮を行う研究[50]や, ウェーブ

レット変換とベクトル符号化を組み合わせて音声信号の符号化を行う研究[51]などが行われている. ウェーブレット変換の音声分野への適用は, 基礎的な分野から音声認識や符号化まで広がりをみせている.

本章では，音声の LSP 分析における残差信号をウェーブレット変換を用いて符号化する手法を提案する. ウェーブレット変換はフィルタ・バンクによって行うことができ, 複雑な予測等の手続きを必要としないため, 本方式の符号化はマルチパルス符号化[52]などに比べ少ない計算量で容易に実現できる.

提案した手法により，単純な Haar のウェーブレットを用いて LSP 分析における残差信号をウェーブレット係数に分解した場合の各ウェーブレット係数の量子化特性を調べた. その結果, 合成音声のスペクトル歪を小さく抑えながら情報圧縮を行うためには, 高周波数のウェーブレット係数よりも低周波数のウェーブレット係数に多くのビットを割り当てる方が有利であることが判明した. シミュレーション実験では, 残差信号のウェーブレット係数を 11.025kbits/sec および 16.538kbits/sec のビットレートで線形量子化を行って合成音声を作成し，十分な了解度を持つ合成音声を生成できた.

9.1.2　ウェーブレット変換を用いた予測残差の符号化

1段の**離散ウェーブレット変換**と**離散逆ウェーブレット変換**のブロック図を図9.1に示す．レベル j の信号 c_j は，フィルタ H(z)と G(z)により，高周波成分と低周波成分に分解される．これらの信号成分を 1/2 に**ダウン・サンプリング**することにより，レベル j - 1 のウェーブレット係数である d_{j-1} と c_{j-1} が得られる．本研究では，ウェーブレットの中で最も単純な Haar のウェーブレットを用いた．従って，変換と逆変換に用いるフィルタは式

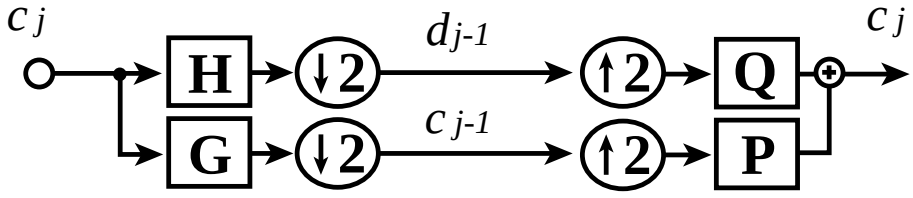


図 9.1　離散ウェーブレット変換と逆変換

(9.1)の伝達関数で表される．

$$H(z)=\frac{1-z}{2} \qquad G(z)=\frac{1+z}{2}$$
$$Q(z)=1-z^{-1} \qquad P(z)=1+z^{-1} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (9.1)$$

本方式では，式(9.1)のフィルタ群を用いて，LSP 分析の残差信号をウェーブレット係数に分解して符号化する．以後，残差信号をウェーブレット変換した係数d_j，c_jをレベルjの残差ウェーブレット係数と呼ぶ．本方式による分析合成の手順を図9.2に示す．本方式の符号化においては，まず音声信号をLSP分析しLSPパラメータと残差信号を算出する．LSPパラメータは適切なビット割り当てを行い伝送するが，この点についてはすでに詳しい研究[10]があるので本論文では扱わない．残差信号は，ウェーブレット変換することによって求めた残差ウェーブレット係数の形で符号化し伝送する．音声信号の復号は，残差ウェーブレット係数を逆ウェーブレット変換することで再生した残差信号でLSP合成フィルタを駆動することによって行う．以下の節では，本方式によるLSP残差信号の符号化特性について検証した結果を報告する．

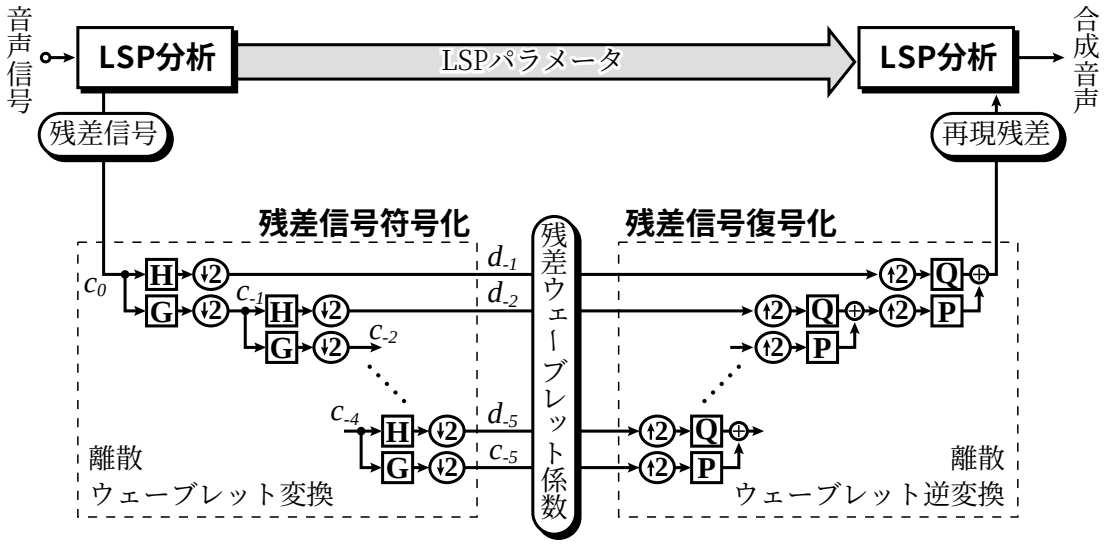


図 9.2　ウェーブレット変換を用いた LSP 予測残差信号の符号化

9.2　残差ウェーブレット係数の性質

9.2.1　音声資料

残差ウェーブレット係数の性質を調べるために, (財)日本情報処理開発協会発行の研究用連続音声データベース中の男性話者の音素バランス文を用いた. 以下の実験では, 音声資料のサンプリング周波数を11.025kHzに変換してから切り出したVCV素片209個, 約40秒分を用いた．LSP分析は，12次で分析フレーム長320サンプル，分析インターバル256サンプルで行った.

9.2.2　残差ウェーブレット係数のパワー

音声資料から切り出した/ sa /の子音部から母音立ち上がりの部分の音声信号をLSP分析し，その残差信号を本手法でウェーブレット変換した例を図9.3に示す．残差ウェーブレット係数は，ダウン・サンプリングの手続きにより，レベルが下がるほど時間軸が粗くなる．図9.3では波形を見やすくするために縦軸のスケールを適当に変化させているが，残差ウェーブレット係数はレベルが下がるほど平均的な振幅は小さくなっている. 無声子音/ s /の部分と有声音である母音/ a /の部分を相対的に比べると，無声子音/ s /の部分が，高いレベルのウェーブレット係数ほど強調されている.

実験に用いたVCV素片から得られた残差信号を有声音部と無声音部に分けて，ウェーブレット変換を行い，残差ウェーブレット係数の平均パワーを求めた．図9.4は，残差ウェーブレット係数のパワーを元の残差信号のパワーで正規化した結果である．有声音部, 無声音部ともにレベルが下がるとウェーブレット係数のパワーは小さくなる. この傾向は無声音部の方がやや強い.

9.2.3　残差ウェーブレット係数へのビット割り当て誤差

残差ウェーブレット係数に対するビット割り当ての目安を得るために, 線形量子化した場合について, 各残差ウェーブレット係数に割り当てるビット数がスペクトル歪に与える影響を調べた．残差ウェーブレット係数 $d_{-1}, d_{-2}, \cdots, d_{-5}, c_{-1}$ の各々の1サンプル当たりに10ビットを割り当てた時の合成音声を基準合成音声として，$d_{-1}, d_{-2}, \cdots, d_{-5}, c_{-1}$ のうち一

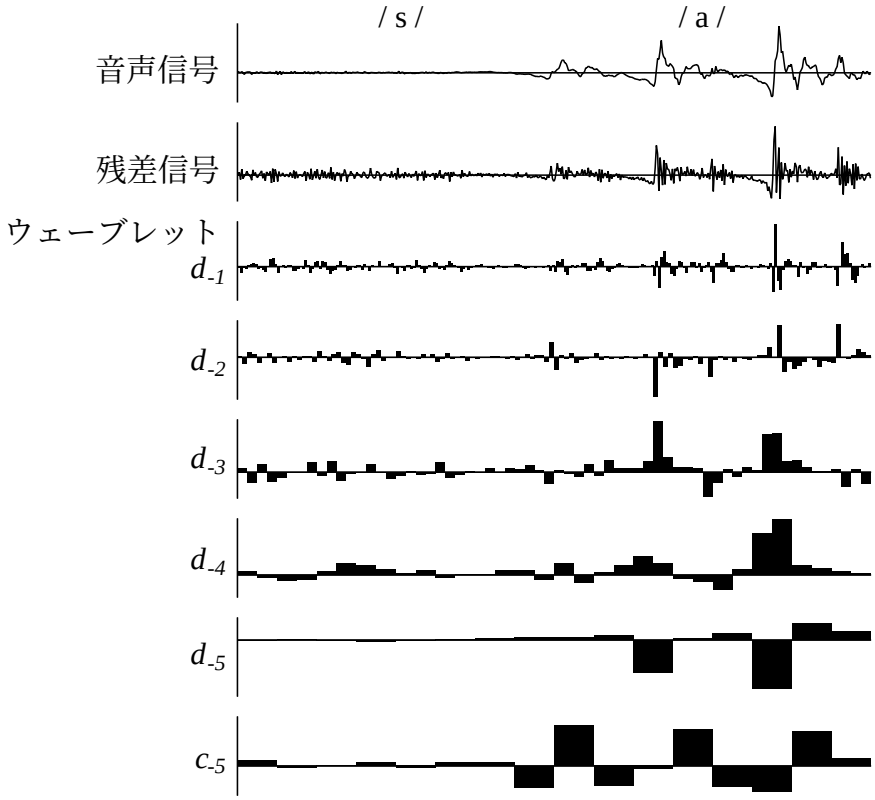


図 9.3　残差信号のウェーブレット変換例

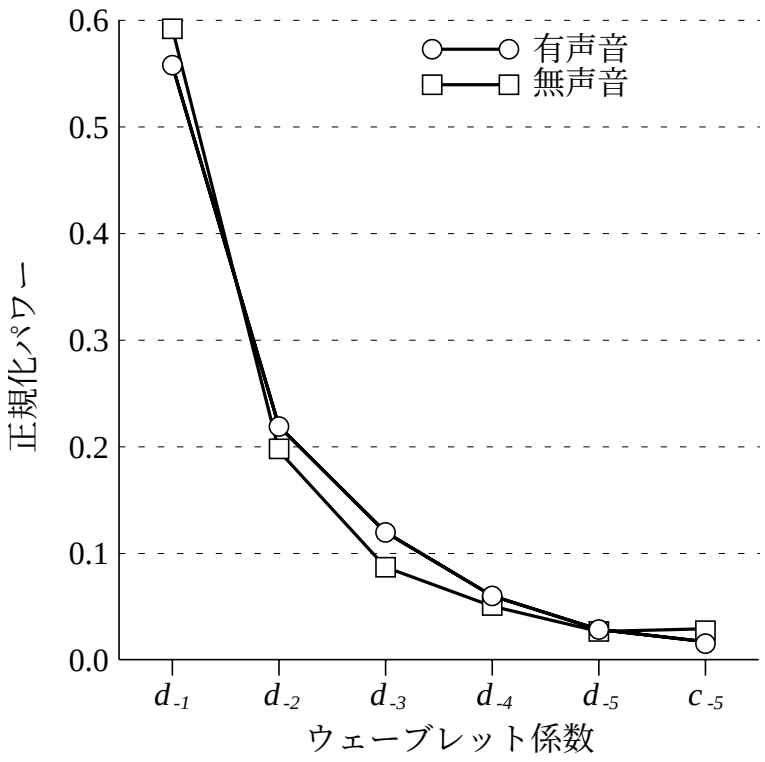


図 9.4　残差ウェーブレット係数のパワ

つの残差ウェーブレット係数への割り当てビット数だけを1サンプルあたり8ビットから0ビットまで変えた合成音声の平均スペクトル歪を求めた．平均スペクトル歪は，式(9.2)で定義する．

$$SD = \sqrt{\frac{1}{N}\sum_{n}\frac{1}{W}\sum_{f}\left\{\hat{S}_n(f) - S_n(f)\right\}^2} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (9.2)$$

ここで，

$\hat{S}_n(f)$ ： 評価対象となる合成音声の対数スペクトル

$S_n(f)$ ： 基準合成音声の対数スペクトル

n ： 分析フレーム番号

N ： 分析フレーム総数

f ： 周波数

W ： 分析周波数帯域

図9.5は，各残差ウェーブレット係数の1サンプル当たりのビット割り当て量とスペクトル歪の関係を表している．有声音，無声音のどちらの場合も，高いレベルの残差ウェーブレット係数のビット割り当て量を削減した場合の方が低いレベルの残差ウェーブレット係数のビット割り当て量を削減した場合よりも，平均スペクトル歪は大きくなる．この傾向は，無声音の方が強く，特にd_{-1}のビット割り当て量を削減したときの平均スペクトル歪の増大は顕著である．

一方，ウェーブレット変換の手続きの中でダウン・サンプリングを行っているため，単位時間あたりのサンプル数は低いレベルの残差ウェーブレット係数の方が少ない．具体的な数値をあげると，音声信号32サンプルあたり(約2.9msあたり)の残差ウェーブレット係数のサンプル数はd_{-1}は16サンプル，d_{-2}は8サンプル，d_{-3}は4サンプル，d_{-4}は2サンプル，d_{-5}, c_{-5}にはそれぞれ1サンプルである．このため，高いレベルの残差ウェーブレット係数に対するビット割り当て量を削減する方が，低いレベルの残差ウェーブレット係数に対するビット割り当て量を削減するよりも，単位時間あたりの情報量削減の効果は大きい．例えば，d_{-1}の1サンプル当たりのビット割り当てを1ビット減らすと，音声信号32

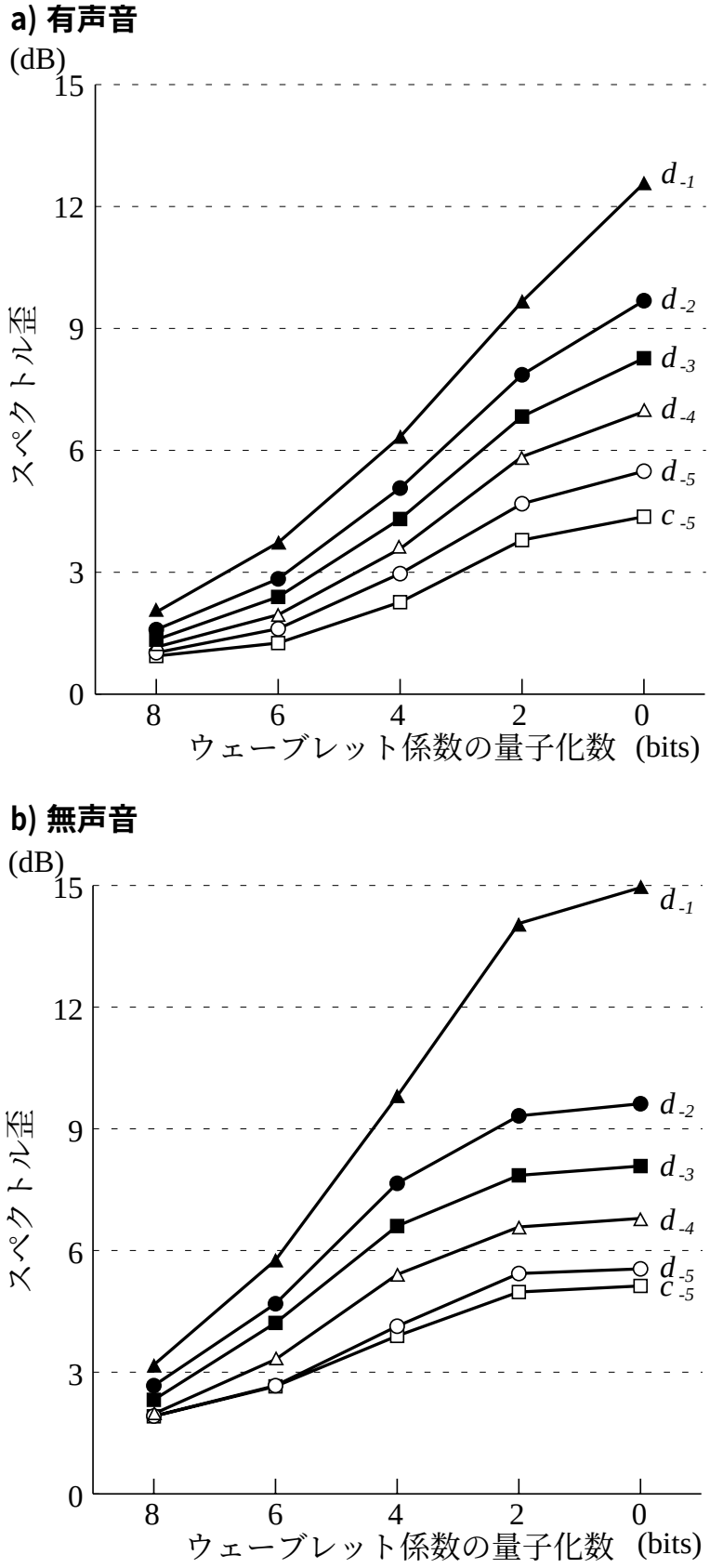
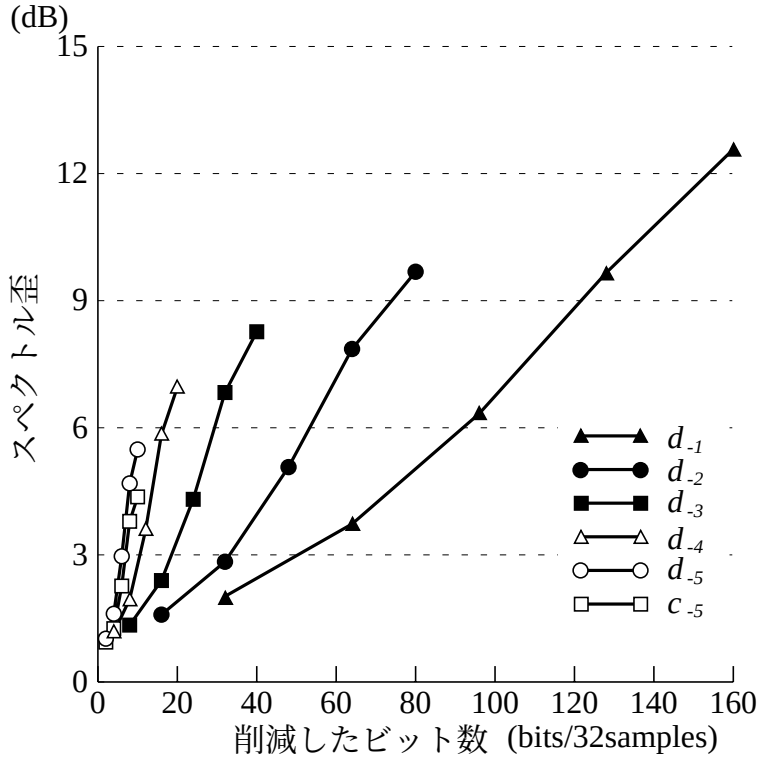


図 9.5　残差ウェーブレット係数に対するビット割り当て量と合成音声の平均スペクトル歪み

a) 有声音



b) 無声音

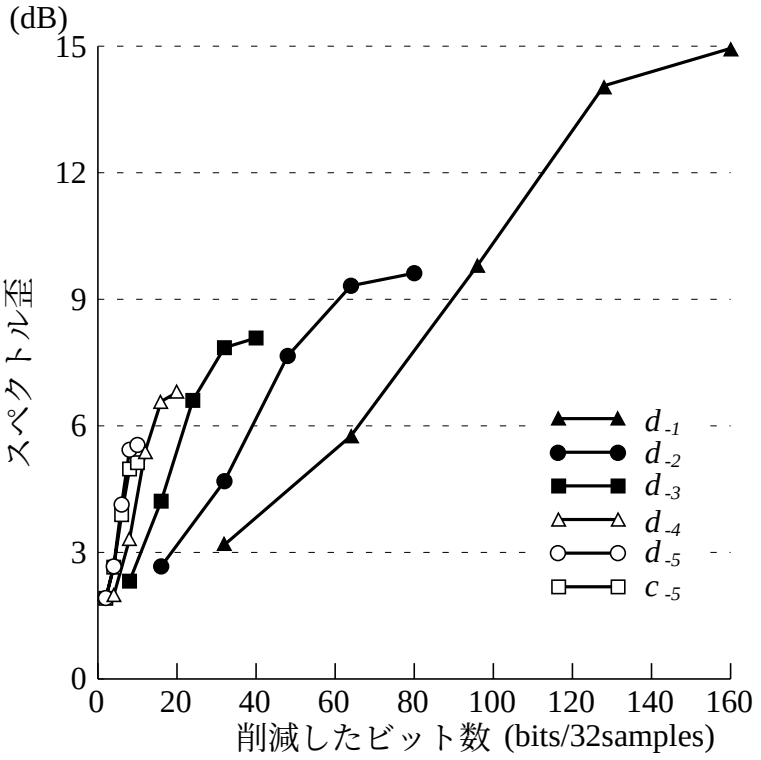


図 9.6　ビット削減量と合成音声の平均スペクトル歪み

サンプルあたり 16 ビットの削減になるが，d_{-5} の場合には音声信号 32 サンプルあたり 1 ビットの削減にしかならない．

図9.6は, 各残差ウェーブレット係数に対するビット割り当ての変更により音声信号32サンプルあたり削減されるビット数を横軸にして, 合成音声の平均スペクトル歪をプロットしたものである．図9.6から，スペクトル歪を大きくせずに低ビットレート化するためには, 高周波数成分を受け持つ高いレベルのウェーブレット係数に少ないビットを割り当てる方が有利であると結論することができる．この結論は，図9.5による先の議論に矛盾しているようにも思えるが, その原因は, 高いレベルのウェーブレット係数におけるビット削減の情報量削減効果が平均スペクトル歪の増大の効果を上回ったためである．

9.2.4　残差ウェーブレット係数へのビット割り当て

スペクトル歪みを評価基準とする前項の評価実験により, 合成音声品質を保ってビット削減を行うには, 高いレベルの残差ウェーブレット係数へのビット割り当てを削減する方が良いといえる．本項では，残差信号に割り当てるビットレートを 11.025kbits/sec と 16.538kbits/sec とし，高いレベルの残差ウェーブレット係数へのビット割り当てが，より低いレベルの残差ウェーブレット係数へのビット割り当てを超えない条件で, 表9.1に示すビット割り当てを行い合成音声の比較を行った．音声資料として, (財)日本情報処理開発協会発行の研究用連続音声データベース中の男性話者1名, 女性話者1名の音素バランス文を用いた．

発話内容は以下の通りである．

A01　「あらゆる現実をすべて自分のほうへねじ曲げたのだ.」

A32　「着用中にダウンやフェザーが飛び出す原因ともなります.」

A47　「日本のエスペラントとしてやはり標準語は必要だ.」

線形量子化の場合, 残差ウェーブレット係数の1サンプル当たりに4ビット以下のビット割り当てを行うと, 合成音声に耳障りなノイズが発生しやすい. 最も合成音声の品質が

良かったビット割り当ては，11.025kbits/sec では C，16.538kbits/sec では d である．16.538kbits/sec の d のビット割り当ての時の合成音声の波形の例を図 9.7 に示す．

C,d の合成音声はどちらも明瞭度は十分であり，16.538kbits/sec の場合，6bits log PCM より雑音が少ない．しかし，高周波数のウェーブレット係数にビットを与えず，削除してしまったため，ややこもった音質になる．

表 9.1　残差ウェーブレット係数に対するビット割り当て

ビット・レート (kbits/sec)		ビット割り当て (bits/sample)					
		d_{-1}	d_{-2}	d_{-3}	d_{-4}	d_{-5}	c_{-5}
11.025	A	0	0	0	6	10	10
	B	0	0	2	2	10	10
	C	0	0	0	8	8	8
	D	0	0	2	4	8	8
	E	0	0	2	6	6	6
	F	0	0	4	4	4	4
16.538	a	0	0	4	6	10	10
	b	0	0	4	8	8	8
	c	0	2	2	4	8	8
	d	0	0	6	6	6	6
	e	0	2	2	6	6	6
	f	0	2	4	4	4	4

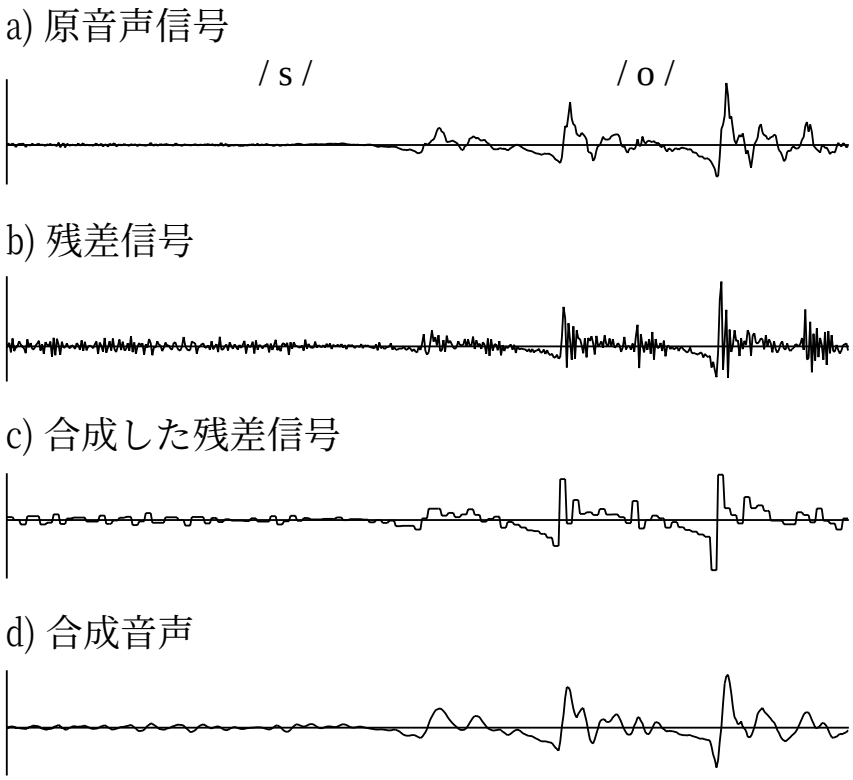


図 9.7　本方式による合成音声の波形例

9.3　まとめ

残差信号をウェーブレット変換を用いて符号化する手法を提案した. 本方式により, 残差信号を 11.025kbits/sec から 16.538kbits/sec 程度のビットレートで符号化して，高品質な合成音声を得ることができた. 本章の議論では, 各残差ウェーブレット係数に対して有声音，無声音とも同じビット割り当てを行った．しかし，9.2 節の結果から有声音と無声音とで各残差ウェーブレット係数に対するビット割り当てを変えることで, より高品質な合成音声を得ることができると考えられる．また，適応的なビット割り当てを行うことで, より低ビットレートでの符号化が可能になる.

第10章　結論

本論文では，医療機器やカーナビゲーション装置，PDA(Personal Digital Assistants)といった機器への組み込み利用を対象とし,高品質な音声合成システムを小規模な計算機資源で構成できる規則音声合成方式の研究開発について論じた.

第3章では，目的を達成する枠組みとして，1Mから4Mバイト程度の小規模な合成単位辞書によって高品質な合成音声を得ることを目標としたLSPベクトルVCV規則音声合成方式を提案した.

第4章では, 提案方式の実現に向けて, 合成単位辞書に収録する合成単位素片の適正な数とその選択方法を実験的に検証した．合成単位素片(VCV 素片)の選択手法として, (i)本論文で定義した音韻環境類似度スコア(PER スコア)を用いて合成単位素片の音韻環境を最適化するPER選択法と, (ii)合成単位素片接続部における接続歪みを最小化するMLD選択法を提案し，素片選択実験と合成音声の品質評価実験を行った.

新聞記事を用いたVCV素片の選択実験では, 14,000個以上のVCV素片を収録した合成単位辞書を用いた場合, 音声合成に必要なVCV素片が合成単位辞書に含まれず, CV素片による代替置換が起こる率であるVCV素片置換率はほぼ一定の1%程度と非常に小さい値である．主観評価実験では，合成単位辞書に収録するVCV素片を増やすと，合成音声の聴感上の品質は向上する．しかし，収録するVCV素片が14,000個を超えると，それ以上では合成音声には聴感上の品質が向上しなくなる．また, 主観評価実験では, PER選択法を用いた合成音声の得点率は53.1%，MLD選択法による合成音声は46.9%となり，両者に有意な差はなかった.

以上より，以下の結論を得た.

1） 本方式では14,000個程度のVCV素片を収録する規模の合成単位辞書を用いれば良い,

2） PER 選択法と MLD 選択法では，選択される VCV 素片は一致しないが，聴感上の品質は同程度である

第 5 章では，第 4 章の議論をもう一歩進め，PER 選択法と MLD 選択法の素片選択基準の関係を実験的に検証した．素片選択実験の結果，PER 選択法で最適な VCV 素片選択を行なった場合，約 70% の文例で接続歪み指標の基準でも上位 10% 以内の選択となっており，MLD 選択法で最適な VCV 素片選択を行なった場合，約 75% の文例で音韻環境指標の基準でも上位 10% 以内の選択となることが示された．これにより，両者の選択基準には強い関係があり, 一方の素片選択基準で最適に素片選択を行なえば, 他方の素片選択基準でも準最適な選択となる事を示すことができた．結論として,

3） PER選択法とMLD選択法のどちらか一方の選択基準だけで素片選択を行えば十分であること,

4） 両選択基準による合成音声に聴感上の差が無いというこれまでの研究結果に論理的な裏付けを与えることができたと云える.

また, 今回の研究結果は, 規則的音声合成法のシステム開発をする際に, どのような素片選択法を採用すべきかについての適切な指針を得るのに有効である. つまり, メモリ規模に重点を置く音声合成システムにはMLD法に準じた方法で素片選択を行い, 処理時間に重点を置く音声合成システムには PER 法に準じた方法で素片選択を行えば品質を損なうことなく良好な合成音が得られることが判った.

第 6 章では，PER 選択法において，VCV 素片選択の際に考慮すべき音韻環境の長さを検証した．PER スコアの計算に算入する音韻環境の長さを制限した部分 PER スコアを定義して，素片選択実験を行った結果以下の結論を得た.

5) VCV素片に対する調音結合の影響は, 素片の前後両方向から及んでおり, 先行／後続のいずれの音韻環境情報も省略できない.

6) 先行音韻環境として考慮する音韻の個数は2個, 後続音韻環境として考慮する音韻の個数は 1 個で十分である.

この結論より，PER 選択法において，合成単位辞書中の VCV 素片に付加する音韻環境

情報を先行2音韻と後続1音韻に抑えることができ，合成単位辞書を大幅に小規模化できる．また，素片選択の処理速度も高速化できることが判った．

第7章では，LSPベクトルVCV規則音声合成方式におけるベクトル量子化について，適切な代表ベクトル数の検証を行った．また，ベクトル量子化を用いて合成単位辞書の容量を削減するだけでなく，ベクトル量子化の特長を活かして，代表ベクトル間の距離テーブルを用いて合成単位辞書から適切なVCV素片を効率的に選択するDTL選択法を提案した．これにより素片選択処理の高速化を可能にした．

7） 主観評価実験により本手法のためのベクトル量子化のコードブックサイズは128程度で十分であることを示した．合成単位辞書に14,000個と多くのVCV素片を収録しても，記憶容量は256Kバイト程度と非常に小さなものにでき，コードブックと残差波形辞書を合わせても500Kバイト程度で記録できることを示した．

8） DTL選択法で順位式の距離テーブルを用いると，距離順位8位でテーブルを打ち切っても，十分な精度でVCV素片の選択が行えることを示した．

以上の結果より，LSPベクトルVCV規則音声合成方式は，多数のVCV素片を収録した大規模なVCV辞書を大幅に情報圧縮でき，TBL選択法によりVCV素片選択の高速化が行えることを示した．また，本方式によって，ベクトル量子化を適用しないLSP-VCV規則音声合成方式と比較して，同程度の合成音声品質を得ることができた．

3章から7章で述べた結果より，LSPベクトルVCV規則音声合成方式では，1Mから4Mバイト程度という当初の目的より小規模な合成単位辞書によって目的を達成できることを示すことができた．また，多くの合成単位素片を持つ単位素片接続型の規則音声合成方式において，素片選択の指標を与えることができた．

第8章と第9章では，LSPベクトルVCV規則音声合成方式で用いる駆動音源の符号化法について議論した．これは，提案法において合成音声の明瞭性だけでなく自然性の向上をも狙った改良を行うためである．第8章では，特に破裂子音の明瞭性を改善するために，少ない容量で破裂子音部の残差信号を符号化するPEC法(Power Envelope Coding法)を提案した．PEC法による合成音声を用いてVCV型の無意味単語の合成を行い，聞き取り実

験により子音の明瞭性を調べた．破裂子音のうち / t /，/ g /，/ d /，/ b / では，PEC 法により元の残差信号を用いる場合に近い明瞭性を得ることができ, 残差信号の微少区間パワの包絡形を保存する PEC 法の有効性が示された．また，/ k /，/ c / は，M 系列音源による合成音声でも, 元の残差信号を音源として用いた場合と同程度の明瞭性が得られ, 本法によらず改善の必要がないことが示された．第 9 章では，ウェーブレット変換を用いて，残差信号を符号化する手法を提案し, 残差信号を11.025kbits/secから16.538kbits/sec程度のビットレートで符号化して，高品質な合成音声が得られることを示した．

本論文では，LSP ベクトル VCV 規則音声合成方式を中心として，小規模な計算機資源で構成できる規則音声合成方式の研究開発について論じた．本方式は，合成単位辞書と VQ コードブック，代表残差辞書を合わせても 500K バイト以下の記憶容量で，明瞭性の高い音声合成を行えることを示した. 提案した方式はLSP分析合成系を含む手法であるため, 波形重畳方式のような高い自然性を持った合成音声を得ることは難しい. 波形重畳方式にせまる自然性を達成するために, 第 8 章と第 9 章で述べたような LSP 合成フィルタの駆動音源の改良を行った. これにより, 本法の明瞭性と自然性をより向上できることが示された．

本研究におけるLSPベクトルVCV規則音声合成方式の研究開発は, 主としてUNIXワークステーション上のシミュレーション実験により行っている. 本手法の有効性を示すことができたので，今後は医療機器やカーナビゲーション装置，PDA といった機器への組み込みの実証試験を行いたい．また，本手法の応用研究が広がってゆくことを願っている．

謝辞

本論文をまとめる機会を与えて頂き, 御指導, 御教示を頂きました大阪大学大学院基礎工学研究科システム人間系専攻生物工学分野の佐藤俊輔教授に深く感謝の意を表します. また, 本論文をまとめるにあたり, 貴重なご助言を頂きました大阪大学大学院基礎工学研究科システム人間系専攻システム科学分野の井口征士教授と藤井隆雄教授に対し, 厚く御礼申し上げます.

この研究をすすめるにあたって, 多くの方々の助力を頂きました. 鳥取大学工学部の菅田一博教授には, 学生時代から熱心な御指導とあたたかい助力を頂いてきました. 鳥取大学工学部の井須尚紀助教授には, 常に熱心な討論と助言を頂きました. 井須助教授より主観評価実験に関する御教授, 御指導を頂かなければ, 本研究は完成しなかったでしょう. 音声合成の研究を始めた当初に, プログラミングや様々な点について熊本電波工業高等専門学校情報工学科の谷口弘教授に多大な御援助を頂きました.

そして, 膨大な音声データと格闘しつつ共に研究を進めてくれた鳥取大学の学生の皆さんに感謝いたします. 合成単位辞書の作成で日夜を問わず頑張ってくれた隅田庸市君と西田博充君のおかげで本研究を開始することができました. 木本雅也君には, いつも実験用のプログラムを改良し管理してもらっています. 山口倫廣君には, 主観評価実験の膨大なデータ処理をお願いしました. また, 主観評価実験の被験者として研究に参加して下さった多くの方々, 全ての方々の名前をあげることは出来ませんが心から感謝しています. その他にも, 多くの学生の皆さんに手伝って頂き, この研究が成立しています.

参考文献

[1] 藤崎博也，広瀬啓吉, “ 規則による音声合成, ” 日本音響学会誌，第 37 巻，第 5 号, 204-209, (1981)

[2] 梅田, “ ベル研究所における英語正書法からの音声合成, ” 信学論（A），vol.J57-A, no.4，318-?, (1974)

[3] J.Allen, et al.,"MITalk-79: The 1979 MIT Text-to-Speech System, 97-Th ASA Meeting, 507, (1979)

[4] 佐藤大和，匂坂芳典，小暮潔，嵯峨山茂樹, “ 日本語テキストからの音声合成, ” 電気通信研究所研究実用化報告，第 32 巻，第 11 号，2243-2252, (1983)

[5] 北脇信彦，板倉文忠，斎藤収三, “ PARCOR 形音声分析合成系における最適符号構成, ” 信学論（A），vol.J61-A，no.2，119-126, (1978)

[6] 東倉洋一, 板倉文忠,“ PARCOR帯域圧縮方式における音声品質向上, ”信学論（A），vol.J61-A，no.3，254-261, (1978)

[7] 北脇信彦，板倉文忠, “ PARCOR 係数の非線形量子化と不均一標本化による音声の能率的符号化, ” 信学論（A），vol.J61-A，no.6，543-550, (1978)

[8] 板倉文忠, “ LSP(線スペクトル対)による音声合成(新時代を開く音声合成技術とその実用装置＜特集＞), ”電子技術(日刊工業新聞社) , vol.22，no.13，14-17, (1980)

[9] 管村昇, 板倉文忠, “ 線形予測計数の線スペクトル表現とその統計的性質, ”信学論（A），vol.J6 ４ -A，no.4，323-330, (1981)

[10] 管村昇，板倉文忠, “ 線スペクトル対(LSP)音声分析合成系による音声情報圧縮, ” 信学論（A），vol.J64-A，no.8，599-606, (1981)

[11] 阿部芳春，今井聖, “ CV 音節のケプストラムパラメータからの音声合成, ” 信学論（D），vol.J64-D，no.9，861-868, (1981)

[12] 今井, 住田, 古市, “ 音声合成のためのメル対数スペクトル近似(MLSA)フィルタ, ” 信学論（A），vol.J66-A，no.2，122-129, (1983)

参考文献

[13] 古市千枝子, 今井聖,“ CV音節のメルケプストラムパラメータの接続に基づく音声の規則合成, ” 信学論（D)，vol.J67-D，no.2，1356-1363, (1984)

[14] 古市千枝子，今井聖, “ 音声の規則合成のためのメルケプストラム CV 音節データファイルの自動作成, ” 信学論（D)，vol.J68-D，no.9，1664-1672, (1985)

[15] T. Minowa and Y. Arai,"The Japanese CV-syllable positioning rule for speech synthesis," Proc. IEEE-IECEJ-ASJ, ICASSP 86, 2031-2034, (1986)

[16] 新居康彦,“ CV音節配置規則を用いたLSP-CV規則音声合成, ”信学論（A)，vol.J70-A，no.5，836-843, (1987)

[17] 伏木田勝信, 三留幸夫, 佐伯猛, “ ホルマント CV-VC方式による規則型音声合成システム, ” 情報処理学会第 31 回全国大会講演論文集，1107-1108, (1985)

[18] 佐藤大和, “ PARCOR-VCV 連鎖を用いた音声合成方式, ” 信学論（D)，vol.J61-D, no.11，858-865, (1978)

[19] 佐藤大和, “ CVC と音源要素に基づく(SYMPLE)音声合成, ” 日本音響学会音声研究会資料，S83-69，541-546, (1984)

[20] 市川昌子，岩田和彦，三留幸夫，伏木田勝信, “ 規則合成における単位音声セットの検討, ” 信学技報，SP87-6，41-48, (1987)

[21] 武田一哉, 安部勝雄, 匂坂芳典, “ 選択的に合成単位を用いる規則音声合成, ”信学論（D-II)，vol.J73-D-II，no.12，1945-1951, (1990)

[22] 中嶌信弥, 浜田洋, “ 音韻環境に基づくクラスタリングによる規則合成法, ”信学論（D-II)，vol.J72-D-II，no.8，1174-1179, (1989)

[23] 中嶌信弥，浜田洋, “ 音韻環境に基づくクラスタリングによる規則合成法, ” NTTR&D，vol.38，No.10，1133-1142, (1989)

[24] 広川智久, 箱田和雄, 中津良平, “ 波形編集型規則合成法における波形選択法, ”信学技報，SP89-114，33-40, (1989)

[25] 小山貴夫，小泉宣夫, “ VCV を基本単位とする波形規則合成方式の検討, ” 信学技報，SP96-8，53-60, (1996)

[26] 望月亮, 西村洋文, 蓑輪利光, “ 波形接続合成に用いるVCV素片データベースの構築方法, ” 信学技報，SP99-73，1-8, (1999)

[27] 小山貴夫, 高橋淳一, 中村太一,“ V･CV波形合成方式のための素片辞書構築方法, ” 信学技報, SP100-97, 41-48, (2000)

[28] 河井恒, 津崎実, 舛田剛志,“ 波形素片接続時の音素環境代替による自然性劣化の知覚的評価, ” 信学技報, SP101-87, 51-57, (2001)

[29] 神谷賢, 雨宮沙織, 有泉均,“ 規則合成における様々な声質の実現の試みについて, ” 信学技報, SP101-86, 7-13, (2001)

[30] 笠松正紀, 西本卓也, 荒木雅弘, “ 適応素片を用いた感情音声の合成, ” 信学技報, SP100-726, 41-46, (2001)

[31] 小山貴夫, 吉岡隆, 高橋淳一, “ ピッチ変形幅を抑えたVCV波形素片生成機構をもつ高品質波形規則合成方式, ”信学論（D-II）, vol.J83-DII, no.11, 2264-2275, (2000)

[32] Gersho A. and Cuperman V. ,"Vector quantization : A pattern-matching technique for speech coding,"IEEE Commun. Mag., 21, 9, 15-21, (1983)

[33] 麻生隆, 大洞恭則, 藤田武, “ 音声合成方法及びその装置, ” 公開特許公報(A), 特開平 5-73100, (1993)

[34] 清水忠昭, 吉村宏紀, 西田博充, 井須尚紀, 菅田一博, “ LSP ベクトル VCV 規則音声合成方式のための合成単位素片数と素片選択法, ” 電気学会論文誌(C), Vol.119-C, No.8/9, 1060-1067(1999)

[35] 清水忠昭, 吉村宏紀, 隅田庸市, 井須尚紀, 菅田一博, “ LSP パラメータにベクトル量子化を適用した小規模応用のためのVCV規則音声合成, ”電気学会論文誌(C), Vol.120-C, No.3, 420-427(2000)

[36] 清水忠昭, 吉村宏紀, 木本雅也, 並木寿枝, 井須尚紀, 菅田一博, “ VCV 規則音声合成における音韻環境指標と接続歪み指標の関係, ” 電気学会論文誌(C), Vol.121-C, No.3, 681-688 (2001)

[37] 木本雅也, 並木寿枝, 清水忠昭, 井須尚紀, 菅田一博, “ VCV 規則音声合成の素片選択において考慮すべき音韻環境の長さ, ” 信学技報, SP2001-49, 31-38, (2001)

[38] 清水忠昭, 吉村宏紀, 並木寿枝, 井須尚紀, 菅田一博, “ LSP-VCV 規則音声合成における破裂子音の明瞭性向上のための残差信号の符号化, ”電気学会論文誌(C), Vol.122-C, No.2, 掲載決定 (2002)

参考文献

[39] 清水忠昭，菅田一博，井須尚紀，吉村宏紀，“音声分析合成のためのウェーブレットを用いた予測残差の符号化，”第11回ディジタル信号処理シンポジウム講演論文集，B4-4，315-320, (1996)

[40] 斎藤由美子，“日本語音声表現法，”桜楓社，東京，82-89, (1990)

[41] 安部勝雄，武田一哉，匂坂芳典，“音韻環境に応じた音声合成素片の接続方法の検討，”信学技報，SP89-66，17-22, (1989)

[42] 岩橋直人，海木延佳，匂坂芳典，“音響的尺度に基づく複合音声単位選択法，”信学技報，SP91-5，33-40, (1991)

[43] 茨木俊秀，“アルゴリズムとデータ構造，”昭晃堂，東京，165-170, (1989)

[44] 難波精一郎, 桑野園子，“音の評価のための心理学的測定法，”コロナ社, 東京, 86-107, (1998)

[45] 武田昌一, 浅川吉章, 市川熹，“残差音源型規則合成における女性音質改善方式の検討，”信学論（A），vol.J73-A，no.4，700-707, (1990)

[46] 伊藤憲三, 佐藤大和，“スペクトル歪最小基準による駆動音源信号の生成と音声合成，”日本音響学会誌，49 巻，10 号，705-710,(1993)

[47] 入野俊夫，“Wavelet 変換による音声信号処理，”信学技報，SP92-81，DSP92-66，pp.59-66 (1992)

[48] 河原英紀, 入野俊夫，“Wavelet変換による音声の駆動源の特徴付けについて，”信学技報，SP91-46，pp.25-32(1991)

[49] 片山喜規, 宮崎明雄, 迫江博昭，“ウェーブレットによる音声信号の解析とその応用について，”信学技報，DSP94-12，RCS94-3，pp.17-24 (1994)

[50] Deepen Sinha, Ahmed H. Tewfik,“Low Bit Rate Transparent Audio Compression using Adapted Wavelets，”IEEE Transactions on Signal Processing，Vol.41，No.12，pp.3463-3479(1993)

[51] E. Mandridake，M. Najim,“Joint Wavelet Transform and Vector Quantization for Speech Coding，”IEEE Int Symp Circuits Syt，Vol.1，pp.699-702(1993)

[52] 李時雨, 高橋寛, 倉橋裕，“無声子音を含む遷移区間の近似合成法を用いたマルチパルス音声符号化システム，”信学論，Vol.J77-A，No.3，pp.293-302(1994)

研究業績一覧

＜学術論文：査読付き論文＞

[1] Hiroki YOSHIMURA, Tadaaki SHIMIZU, Naoki ISU, Kazuhiro SUGATA, "Construction of Noise Reduction Filter by Use of Sandglass-Type Neural Network," IEICE TRANS. FUNDAMENTALS, Vol. E80-A, No.8, 1384-1390(1997-8)

[2] Tadaaki SHIMIZU, Hiroki YOSHIMURA, Yoshihiko SHINDO, Naoki ISU, Kazuhiro SUGATA, "Production of LSP Parameter Sequences for Speech Synthesis Based on Neural Network Approach," IEICE TRANS. FUNDAMENTALS, Vol. E80-A, No.8, 1467-1471(1997-8)

[3] 吉村宏紀，清水忠昭，井須尚紀，菅田一博, “ 砂時計型ニューラルネットワークを用いた雑音除去フィルタの構成, ” 電気学会論文誌(C)，Vol.117-C，No.10，1498-1505(1997)

[4] 吉村宏紀，清水忠昭，佐山卓史，井須尚紀，菅田一博, “ M 系列を用いた BPTT 学習によるリカレントニューラルネットワーク IIR フィルタの構成, ” 電気学会論文誌(C)，Vol.118-C，No.3，411-418(1998)

[5] 吉村宏紀，清水忠昭，井須尚紀，菅田一博, “ 砂時計型ニューラルネットワークを用いたProny法による周波数推定, ”電子情報通信学会論文誌(A)，Vol.J81-A，No.4，799-802(1998-4)

[6] Hiroki YOSHIMURA, Tadaaki SHIMIZU, Naoki ISU, Kazuhiro SUGATA, "An Adaptive Noise Reduction Filter for Discrete Signal by Use of Sandglass-Type Neural Network," Electrical Engineering in Japan, Vol.127, No.4, 39-50 (1999)

[7] 清水忠昭，吉村宏紀，西田博充，井須尚紀，菅田一博, “ LSP ベクトル VCV 規則音声合成方式のための合成単位素片数と素片選択法, ” 電気学会論文誌(C)，Vol.119-C，No.8/9，1060-1067(1999)

[8] 清水忠昭，吉村宏紀，隅田庸市，井須尚紀，菅田一博, “ LSP パラメータにベクトル量子化を適用した小規模応用のためのVCV規則音声合成, ”電気学会論文誌(C)，Vol.120-C，No.3，420-427(2000)

[9] 吉村宏紀，清水忠昭，井須尚紀，菅田一博, “ 多段接続砂時計型ニューラルネットワーク雑音除去フィルタを用いた適応的雑音除去, ” 電気学会論文誌(C)，Vol.120-C，No.4，507-515(2000)

[10] Hiroki YOSHIMURA, Tadaaki SHIMIZU, Naoki ISU, Kazuhiro SUGATA, "Adaptive Noise Reduction by Using Cascaded Sandglass-Type Neural Networks," Electrical Engineering in Japan, Vol.136, No.1, 37-46 (2001)

[11] 清水忠昭，吉村宏紀，木本雅也，並木寿枝，井須尚紀，菅田一博，“VCV 規則音声合成における音韻環境指標と接続歪み指標の関係，”電気学会論文誌(C)，Vol.121-C，No.3，681-688 (2001)

[12] 清水忠昭，吉村宏紀，並木寿枝，井須尚紀，菅田一博，“LSP-VCV 規則音声合成における破裂子音の明瞭性向上のための残差信号の符号化，”電気学会論文誌(C)，Vol.122-C，No.2，掲載決定 (2002)

＜国際会議＞

[1] Isu N, Shimizu T, and Sugata K; Mechanics of Coriolis stimulus and inducing factors of motion sickness, Biol. Sci. Space, 2001 (in print).

＜シンポジウム：査読なし論文＞

[1] 清水忠昭，菅田一博，井須尚紀，吉村宏紀，“音声分析合成のためのウェーブレットを用いた予測残差の符号化，”第11回ディジタル信号処理シンポジウム講演論文集，B4-4，315-320, (1996)

[2] 進藤佳彦，清水忠昭，菅田一博，井須尚紀，吉村宏紀，“ニューラルネットワークを用いたVCV音声合成法，”第11回ディジタル信号処理シンポジウム講演論文集，B4-5，321-326, (1996)

[3] 吉村宏紀，佐山卓史，菅田一博，井須尚紀，清水忠昭，“リカレントニューラルネットワークを用いたBPTT学習法によるディジタルフィルタの構成，”第11回ディジタル信号処理シンポジウム講演論文集，A8-5，665-670, (1996)

[4] 吉村宏紀，菅田一博，井須尚紀，清水忠昭，“砂時計型ニューラルネットワークによる時系列信号の雑音除去，”第 11 回ディジタル信号処理シンポジウム講演論文集，A8-6，671-676, (1996)

＜著書＞

[1] 清水忠昭，菅田一博，“コンピュータ解体新書，”サイエンス社，東京，1992

[2] 清水忠昭，菅田一博，“構造化プログラミング事始め，”サイエンス社，東京，1993

[3] 清水忠昭，菅田一博，“UNIX のススメ，”サイエンス社，東京，1994

[4] 清水忠昭，菅田一博，“C 言語のススメ，”サイエンス社，東京，1994

索引

アルファベット索引

日本語索引

タ

ナ

ハ

マ

ヤ・ラ・ワ