概要

音声合成の手法の 1 つとして音節波形接続型音声合成 [1] が提案されている．この手法は録音音声から音節を単位とし，条件が一致する音節素片を切り出し，信号処理を加えずに接続することで自然性の高い音声を合成できる．音節波形接続型音声合成において，音声波形の音節開始位置・終了位置は，音節境界位置が記載されたラベルを利用している．ラベルは人手で作成されているが，波形接続用ではないため音節境界位置の精度が低い．そのため，音声合成時に波形の接続点が不連続になり，音声品質が劣化する．そこで波形を滑らかに接続するように波形接続時に人手で音節開始位置・終了位置に修正を加えている．しかし，この修正作業にはコストがかかる [2].

そこで，音節の精密な開始位置・終了位置を自動的に決める方法を提案した [3]．具体的には，音節素片のパワーが最大となる周波数を求め，その初期位相が “-$\frac{\pi}{2}$” となる時間を音節開始位置にする．しかし，離散フーリェ変換の窓長を音節素片の音節開始位置における 1 周期の整数倍にしなければ，初期位相に誤差が生じることがわかった.

本研究では，離散フーリエ変換の窓長を音節素片の開始時から 1 周期の整数倍にするために，窓長の最大 ±0.5ms(約 0.0625ms 刻み，計 17 種類) に対して離散フーリエ変換を行い，振幅が最も 0 に近い音節開始位置を選択することで誤差の修正を試みた.

提案方法を用いて合成音声を作成し，音声品質を調査する．音声品質を評価するために，聴覚実験ではオピニオン評価実験と対比較実験を行った．実験に用いた音声は，女性話者 2 名と男性話者 2 名である.

聴覚実験における対比較実験の結果において，女性話者の場合，提案方法で作成した合成音声は，ATR ラベルによる音節境界位置から作成した合成音声と比較して平均 76.0%，人手で音節境界位置の調整を行い作成した合成音声と比較して 50.6%の音声の品質が良いと判断された．また，男性話者の場合，提案方法で作成した合成音声は，ATR ラベルによる音節境界位置から作成した合成音声と比較して平均 55.5%，人手で音節境界位置の調整を行い作成した合成音声と比較して 39.7%の音声の品質が良いと判断された.

実験結果より，女性話者において，提案方法で作成した合成音声は人手で音節境界位置の調整を行い作成した合成音声と同等の音声品質を得ることができた．男性話者にお

いて，提案方法で作成した合成音声は ATR ラベルによる音節境界位置から作成した合成音声より音声品質を向上させることができた．したがって，本研究の有効性が証明された．

目 次

図 目 次

表 目 次

第1章　はじめに

現在，カーナビゲーションシステムや電車の車内アナウンスなどのように，音声ガイダンスを利用したシステムやサービスが様々な場面において利用されている．この音声ガイダンスの作成には，録音編集方式が広く使われている．録音編集方式は，ユーザに依存しない比較的長い文音声 (以下，固定部) と，ユーザに依存する比較的短い単語・文音節音声 (以下，可変部) を別々に録音しておき，必要に応じて組み合わせることで，目的となる出力音声を作成する方法である．

例えば,「次の交差点を左折です.」という音声ガイダンスを作成する場合,「次の○○左折です.」という固定部に,「交差点を」という可変部を挿入して作成する．

録音編集方式を用いた音声合成においては，可変部と固定部を接続した場合の違和感を軽減するために，一般に同一話者の音声が必要となる．可変部と固定部を分離して録音することにより，必要となるすべての音声を録音する場合に比べて話者に対する負担は若干軽減されるが，可変部に挿入する単語が増大した場合，同一話者から全ての音声を録音することは困難となる．そこで，固定部は録音音声，可変部は合成音声を用いる方式がとられている．その合成音声を作成する方法の１つとして，音節波形接続方式 [1] が提案されている．

音節波形接続方式は，音響的なパラメータを使用せず，言語的なパラメータのみで合成音声を作成する方式であり，信号処理を加えないで接続することにより，自然性の高い合成音声を作成できる．この方式の過去の研究として，固有名詞，普通名詞，文節 (短文節)，フレーズ (長文節) を対象として行われた．その結果，品質の高い合成音声が得られたことが報告されている．

節波形接続型音声合成において，音声波形の音節開始位置・終了位置は，音節境界位置が記載されたラベルを利用している．ラベルは人手で作成されているが，波形接続用ではないため音節境界位置の精度が低い．そのため，音声合成時に波形の接続点が不連続になり，音声品質が劣化する．そこで波形を滑らかに接続するように波形接続時に人手で音節開始位置・終了位置に修正を加えている．しかし，この修正作業にはコストがかかる [2].

そこで，音節の精密な開始位置・終了位置を自動的に決める方法を提案した [3]．具体的には，音節素片のパワーが最大となる周波数を求め，その初期位相が “-$\frac{\pi}{2}$” となる時間を音節開始位置にする．しかし，離散フーリェ変換の窓長を音節素片の音節開始位置における 1 周期の整数倍にしなければ，初期位相に誤差が生じることがわかった．

本研究では，離散フーリエ変換の窓長を音節素片の開始時から 1 周期の整数倍にするために，窓長の最大 ±0.5ms(約 0.0625ms 刻み，計 17 種類) に対して離散フーリエ変換を行い，振幅が最も 0 に近い音節開始位置を選択することで誤差の修正を試みた．

提案方法を用いて合成音声を作成し，音声品質を調査する．音声品質を評価するために，聴覚実験ではオピニオン評価実験と対比較実験を行った．実験に用いた音声は，女性話者 2 名と男性話者 2 名である．

聴覚実験における対比較実験の結果において，女性話者の場合，提案方法で作成した合成音声は，ATR ラベルによる音節境界位置から作成した合成音声と比較して平均 76.0%，人手で音節境界位置の調整を行い作成した合成音声と比較して 50.6%の音声の品質が良いと判断された．また，男性話者の場合，提案方法で作成した合成音声は，ATR ラベルによる音節境界位置から作成した合成音声と比較して平均 55.5%，人手で音節境界位置の調整を行い作成した合成音声と比較して 39.7%の音声の品質が良いと判断された．

実験結果より，女性話者において，提案方法で作成した合成音声は人手で音節境界位置の調整を行い作成した合成音声と同等の音声品質を得ることができた．男性話者において，提案方法で作成した合成音声は ATR ラベルによる音節境界位置から作成した合成音声より音声品質を向上させることができた．したがって，本研究の有効性が証明された．

以降，第 2 章で音節波形接続型音声合成の説明をする．そして第 3 章で位相を用いた音節境界位置の修正方法について説明を行い，第 4 章で従来用いられている方法について述べる．第 5 章で実験方法について説明し，第 6 章で実験結果を示し，第 7 章で実験結果に対する考察を章で述べる．

第2章　音節波形接続型音声合成

2.1　音節素片の選択条件

音節波形接続方式は，波形編集型の音声合成方式の一種で，音響パラメータを使用しないで，言語的なパラメータのみで音声合成を作成する．具体的には，音節波形接続方式で音声を合成する際，収録された大量のデータベースから表 2.1 の言語的なパラメータの条件が一致する音節素片を選択する．

表 2.1: 音節素片を選択する際の条件

1. 中心の音節
2. 直前の音素 (前音素環境)
3. 直後の音素 (後音素環境)
4. 単語のモーラ数
5. 単語のモーラ位置
6. 単語のモーラ数
7. 単語のアクセント型

最後に，音節の開始時間と終了時間に基づいて波形データを切り出し，接続して合成音声を作成する．

音節波形接続型音声合成の例として「発音 /ha/tsu/o/N/」の合成音声を作成する場合に選択される音節素片を示す.「発音 /ha/tsu/o/N/」の「tsu」の音節に用いる音節素片は,「雑音 /za/tsu/o/N/」の「tsu」の音節素片を選択する．これは中心の音節が「tsu」，直前の音素が「a」，直後の音素が「n」，単語のモーラ位置が「2」，単語のモーラ数が「4」，単語のアクセント型が「0111 型」となっている．したがって表 2.1 に示した条件が「発音 /ha/tsu/o/N/」の「tsu」の音節素片と一致するためである．

2.2 音節波形接続方式の問題点

2.2.1 ラベルの利用

波形接続型音声合成では，音声合成する音節の接続点を音節境界位置と定義している．音節境界位置は，音節開始時刻と終了時刻が記載されたラベルに基づいて決定される．図 2.1 から図 2.3 にラベルの例を示す．図 2.1 は「診察 /shi/N/sa/tsu/」のラベルの例を示している．左から音節開始時刻，音節終了時刻，音節の名前が記載されており，時刻の単位はミリ秒である．例えば，開始の音節「pau」は無音を示しており，時刻 0 ミリ秒から時刻 285 ミリ秒に含まれている．同様に，図 2.2 は「威厳 /i/ge/N/」のラベルの例，図 2.3 は「発音/ha/tsu/o/N/」のラベルの例を示している．ラベルは，人手で作成されている．しかし，波形接続型音声合成用ではなくパラメータ合成用であるため，正確に記載されていないという問題点がある．具体的には，ラベルは人手で作成されており，5 ミリ秒間隔で記載されているため，パラメータ合成では問題ないが，波形接続方式では精度が足りず，音声品質に問題が生じる．

図 2.1: 「診察」のラベル (ms)

0	285	pau
285	555	shi
555	730	N
730	1010	sa
1010	1340	tsu
1340	1625	pau

図 2.2: 「威厳」のラベル (ms)

0	295	pau
295	495	i
405	675	ge
675	925	N
925	121	pau

図 2.3: 「発音」のラベル (ms)

0	300	pau
300	550	ha
550	865	tsu
865	1035	o
1035	1295	N
1295	1587	pau

2.2.2　接続部の修正

波形接続型音声合成では，接続部の違和感の発生が音声の自然性に大きく影響する．しかし，ラベルから得た音節境界位置で音節素片を切り出し，そのまま接続すると接続部に違和感が生じる．図 2.4 に例を示す．図 2.4 は例として「診察 /shi/N/sa/tsu/」の shi/N 間の接続部を示している．縦線部が接続部となっており，接続部より左部の波形が「真空 /shi/N/ku/u/」の「shi」の音節を用いて，右部の波形が「申請/shi/N/se/i/」の「N」の音節を用いて作成されている．縦線部に歪みが生じていることがわかる．

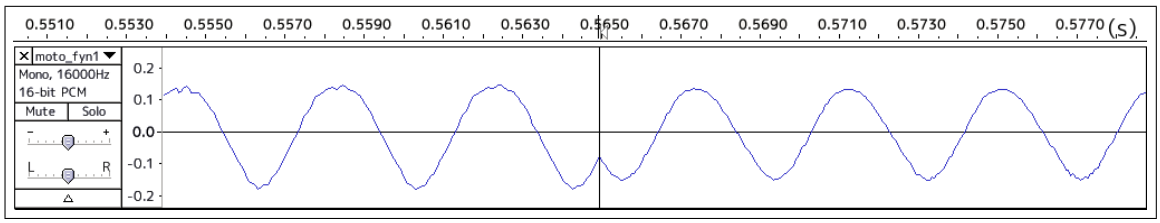

図 2.4: 「診察 /shi/N/sa/tsu/」の shi/N をラベル通りに接続した音声波形

上記の問題があるため，ラベルから得た音節境界位置で切り出した音節素片を接続する場合，2 素片間の接続部を滑らかに接続する必要がある．以下に人手で修正する方法を示す．

ラベルから得た素片開始時間と素片終了時間をもとに，振幅が負から正に変わる部分を，波形が短くなる方向 (開始時間は進む方向，終了時間は戻る方向) に探し，音節素片を切り出す位置を修正する [1]．図 2.5 は図 2.4 の波形を人手で修正した結果を示している．「診断 /shi/N/sa/tsu/」の shi/N 間の接続部を示している．縦線部で滑らかに接続していることがわかる．

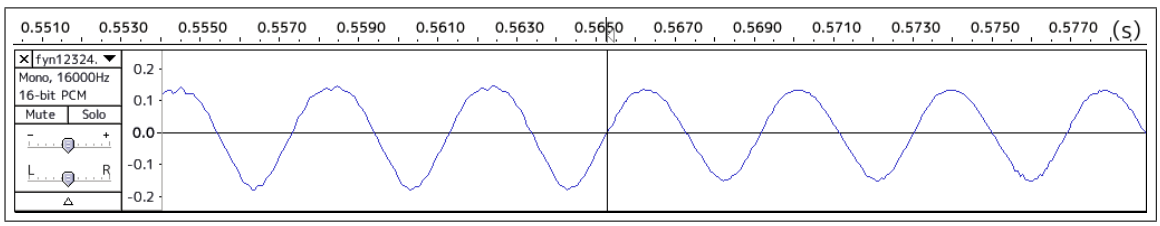

図 2.5: 「診察 /shi/N/sa/tsu/」の shi/N を人手で修正して接続した音声波形

しかし，大量の合成音声作成時に人手の修正を行うには非常にコストがかかる．具体的には，人手で合成音声を作成する平均時間は，合成音声一つにつき約 5 分である．

2.3 音節波形接続型音声合成の例

音節波形接続型音声合成の例として「威厳 /i/ge/N/」,「発音 /ha/tsu/o/N/」「対話 /ta/i/wa/」を表 2.6 に示す．なお,「‾」「‾」は音の強弱 (アクセント) を表している．() 内強調部は，実際に選択される部分を示している．

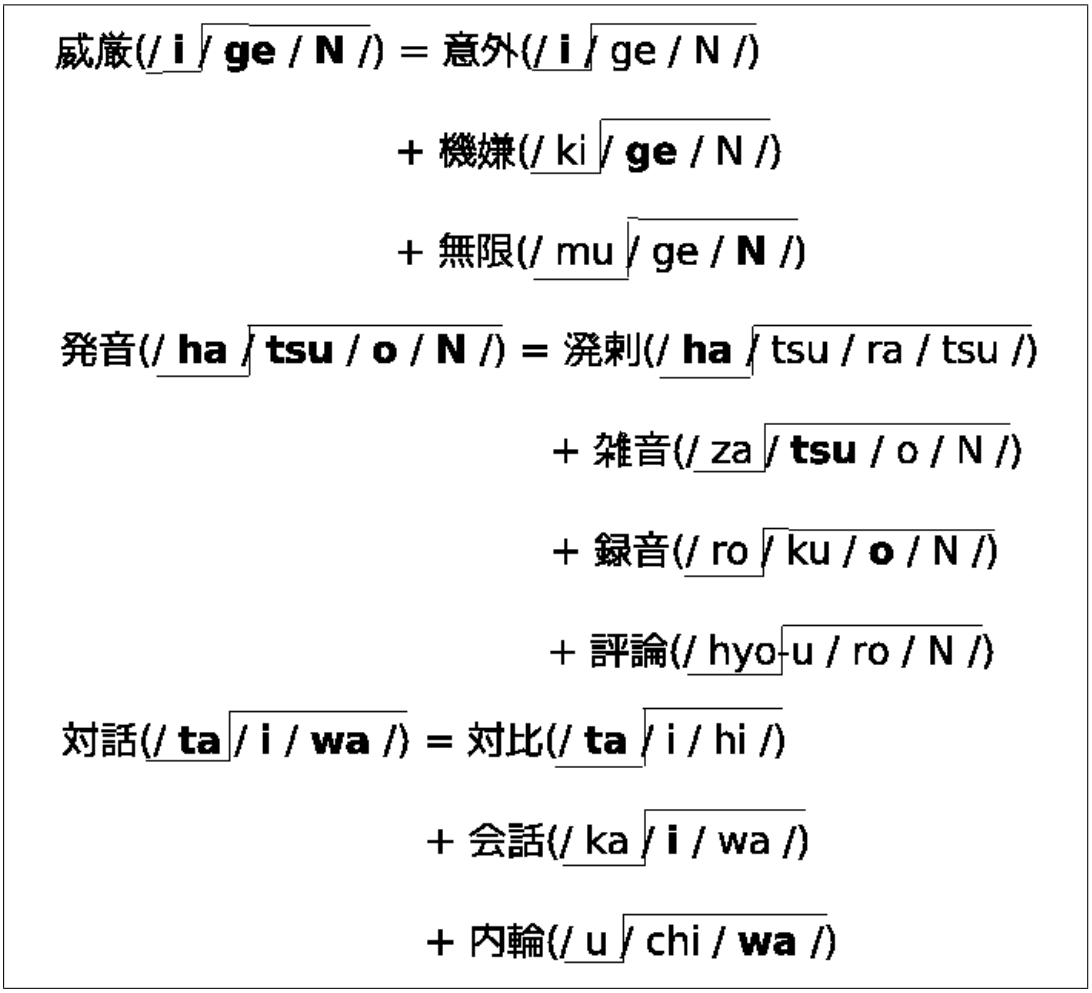

図 2.6: 音節波形接続型音声合成における音節素片選択の例

図 2.6 に示した 3 通りの合成音声の接続部の波形データを次ページから示す.

2.3.1 「威厳 i/ge/N/」

音節波形接続型音声合成の例として「威厳 /i/ge/N/」の合成音声を作成する場合に選択される音節素片を示す.「威厳 /i/ge/N/」の「i」の音節に用いる音節素片は,「意外 /i/ga/i/」の「i」の音節素片を選択する．これは中心の音節が「i」，直前の音素が「pau」，直後の音素が「g」，文節のモーラ位置が「1」，文節のモーラ数が「3」，文節のアクセント型が「0111 型」となっている．したがって表 2.1 に示した条件が「威厳 /i/ge/N/」の「i」の音節素片と一致するためである．同様に,「威厳 /i/ge/N/」の「ge」の音節に用いる音節素片は,「機嫌 /ki/ge/N/」の「ge」の音節素片を選択する．これは中心の音節が「ge」，直前の音素が「i」，直後の音素が「N」，文節のモーラ位置が「2」，文節のモーラ数が「3」，文節のアクセント型が「0111 型」である．また,「威厳 /i/ge/N/」の「N」の音節に用いる音節素片は,「無限 /mu/ge/N/」の「N」の音節素片を選択する．これは中心の音節が「N」，直前の音素が「e」，直後の音素が「pau」，文節のモーラ位置が「3」，文節のモーラ数が「3」，文節のアクセント型が「0111 型」である．それぞれ表 2.1 に示した条件が「威厳 /i/ge/N/」の各音節素片と一致する．表 2.7 から表 2.10 にラベル通りに接続した場合と人手で修正を加えた場合の各接続部の波形を示す.

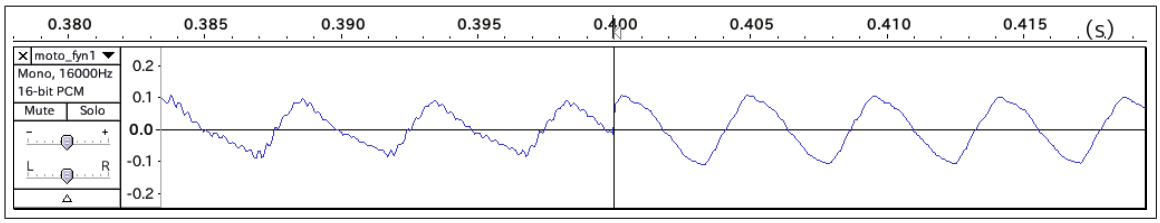

図 2.7: 「威厳 /i/ge/N/」の i/ge をラベル通りに接続した音声波形

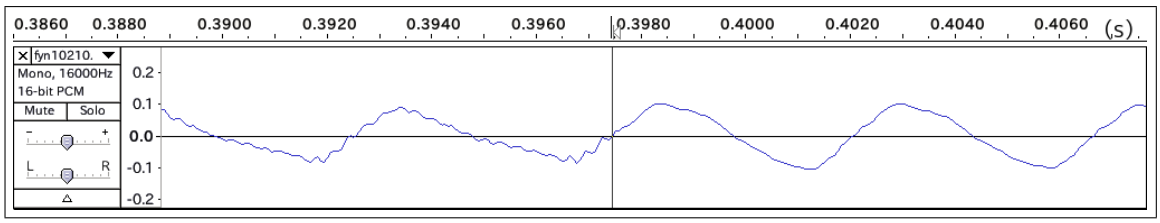

図 2.8: 「威厳 /i/ge/N/」の i/ge を人手で修正して接続した音声波形

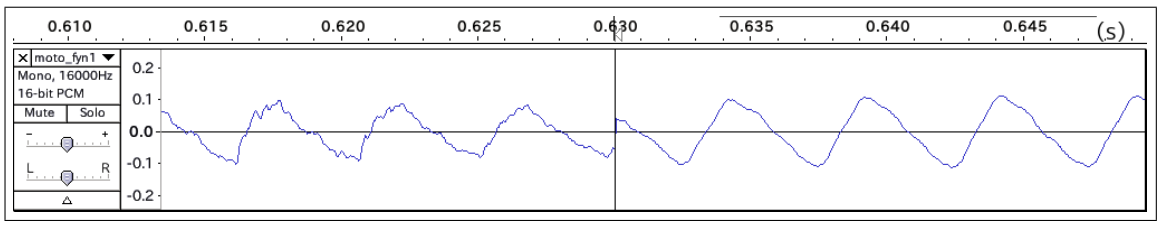

図 2.9: 「威厳 /i/ge/N/」の ge/N をラベル通りに接続した音声波形

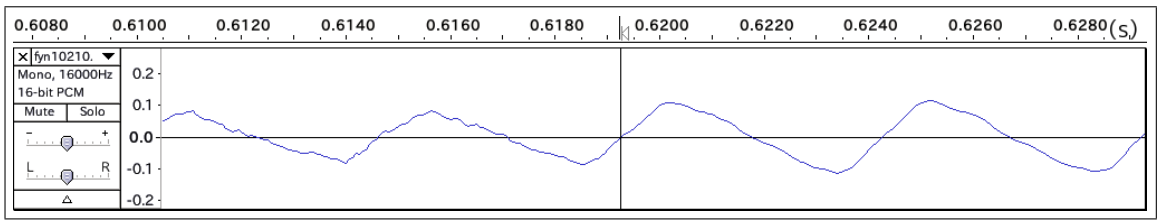

図 2.10: 「威厳 /i/ge/N/」の ge/N を人手で修正して接続した音声波形

2.3.2 「発音 /ha/tsu/o/N/」

音節波形接続型音声合成の例として「発音 /ha/tsu/o/N/」の合成音声を作成する場合に選択される音節素片を示す.「発音 /ha/tsu/o/N/」の「ha」の音節に用いる音節素片は,「溌剌 /ha/tsu/ra/tsu/」の「ha」の音節素片を選択する．これは中心の音節が「ha」，直前の音素が「pau」，直後の音素が「ts」，文節のモーラ位置が「1」，文節のモーラ数が「4」，文節のアクセント型が「0111 型」となっている．したがって表 2.1 に示した条件が「発音 /ha/tsu/o/N/」の「ha」の音節素片と一致するためである．同様に,「発音 /ha/tsu/o/N/」の「tsu」の音節に用いる音節素片は,「雑音 /za/tsu/o/N/」の「tsu」の音節素片を選択する．これは中心の音節が「tsu」，直前の音素が「a」，直後の音素が「o」，文節のモーラ位置が「2」，文節のモーラ数が「4」，文節のアクセント型が「0111 型」である．また,「発音 /ha/tsu/o/N/」の「o」の音節に用いる音節素片は,「録音 /ro/ku/o/N/」の「o」の音節素片を選択する．これは中心の音節が「o」，直前の音素が「u」，直後の音素が「N」，文節のモーラ位置が「3」，文節のモーラ数が「4」，文節のアクセント型が「0111 型」である．最後に,「発音 /ha/tsu/o/N/」の「N」の音節に用いる音節素片は,「評論/hyo-u/ro/N/」の「N」の音節素片を選択する．これは中心の音節が「N」，直前の音素が「o」，直後の音素が「pau」，文節のモーラ位置が「4」，文節のモーラ数が「4」，文節のアクセント型が「0111 型」である．それぞれ表 2.1 に示した条件が「発音 /ha/tsu/o/N/」の各音節素片と一致する．表 2.11 から表 2.16 にラベル通りに接続した場合と人手で修正を加えた場合の各接続部の波形を示す.

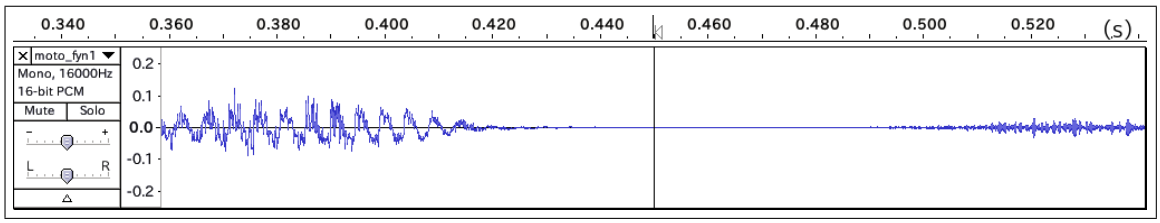

図 2.11: 「発音 /ha/tsu/o/N/」の ha/tsu をラベル通りに接続した音声波形

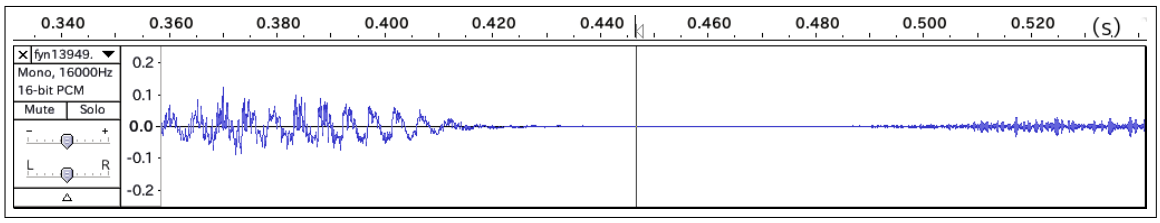

図 2.12: 「発音 /ha/tsu/o/N/」の ha/tsu を人手で修正して接続した音声波形

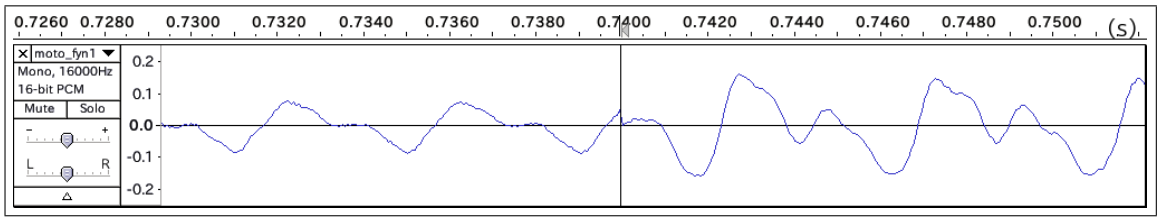

図 2.13: 「発音 /ha/tsu/o/N/」の tsu/o をラベル通りに接続した音声波形

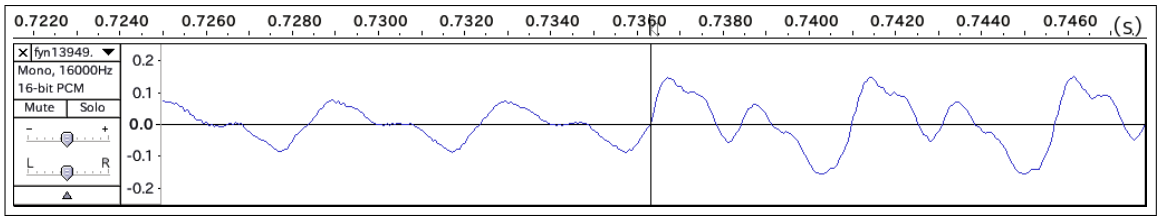

図 2.14: 「発音 /ha/tsu/o/N/」の tsu/o を人手で修正して接続した音声波形

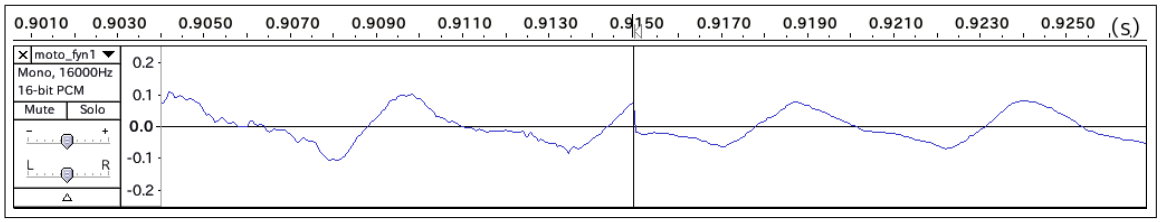

図 2.15: 「発音 /ha/tsu/o/N/」の o/N をラベル通りに接続した音声波形

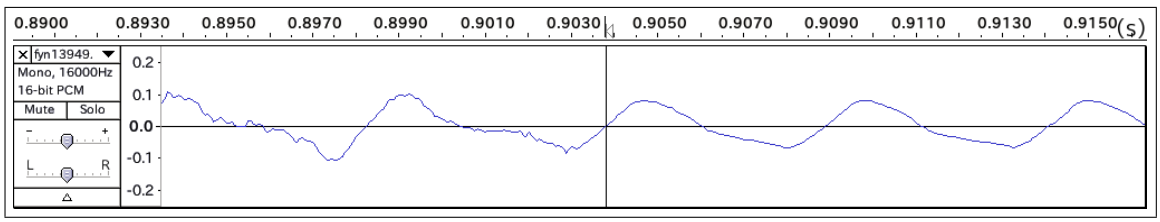

図 2.16: 「発音 /ha/tsu/o/N/」の o/N を人手で修正して接続した音声波形

2.3.3 「対話 /ta/i/wa/」

音節波形接続型音声合成の例として「対話 /ta/i/wa/」の合成音声を作成する場合に選択される音節素片を示す.「対話 /ta/i/wa/」の「ta」の音節に用いる音節素片は,「対比 /ta/i/hi/」の「ta」の音節素片を選択する．これは中心の音節が「ta」，直前の音素が「pau」，直後の音素が「i」，文節のモーラ位置が「1」，文節のモーラ数が「3」，文節のアクセント型が「0111 型」となっている．したがって表 2.1 に示した条件が「対話 /ta/i/wa/」の「ta」の音節素片と一致するためである．同様に,「対話 /ta/i/wa/」の「i」の音節に用いる音節素片は,「会話 /ka/i/wa/」の「i」の音節素片を選択する．これは中心の音節が「i」，直前の音素が「a」，直後の音素が「w」，文節のモーラ位置が「2」，文節のモーラ数が「3」，文節のアクセント型が「0111 型」である．また,「対話 /ta/i/wa/」の「wa」の音節に用いる音節素片は,「内輪/u/chi/wa/」の「wa」の音節素片を選択する．これは中心の音節が「wa」，直前の音素が「i」，直後の音素が「pau」，文節のモーラ位置が「3」，文節のモーラ数が「3」，文節のアクセント型が「0111 型」である．それぞれ表 2.1 に示した条件が「対話 /ta/i/wa/」の各音節素片と一致する．表 2.17 から表 2.20 にラベル通りに接続した場合と人手で修正を加えた場合の各接続部の波形を示す.

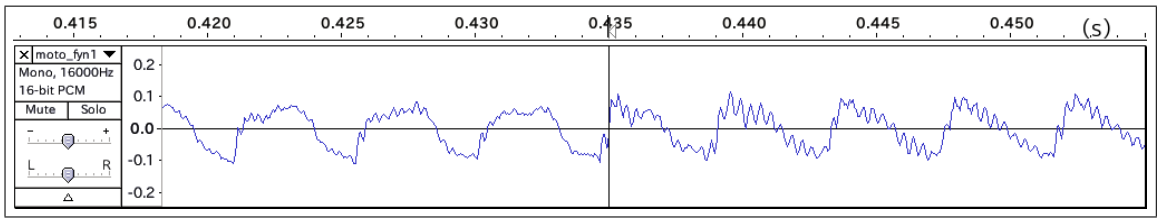

図 2.17: 「対話 /ta/i/wa/」の ta/i をラベル通りに接続した音声波形

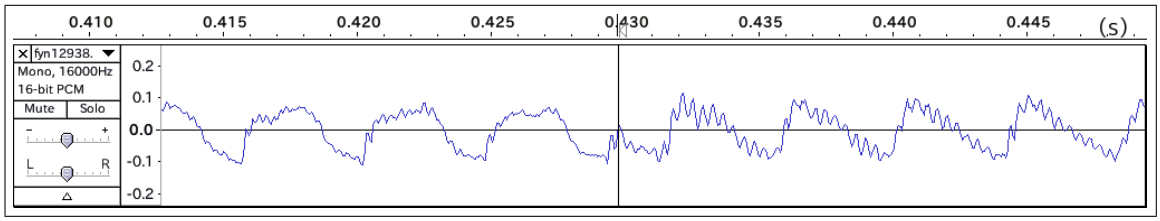

図 2.18: 「対話 /ta/i/wa/」の ta/i を人手で修正して接続した音声波形

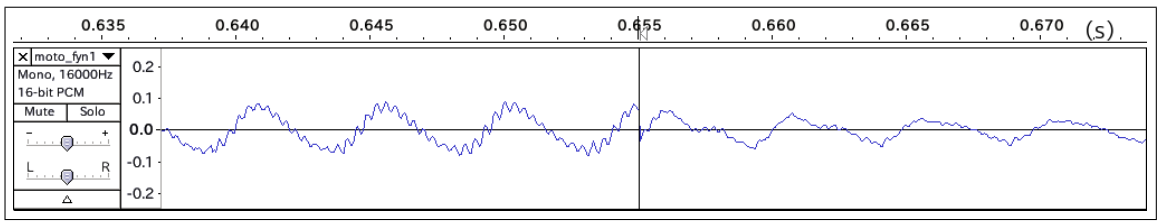

図 2.19: 「対話 /ta/i/wa/」の i/wa をラベル通りに接続した音声波形

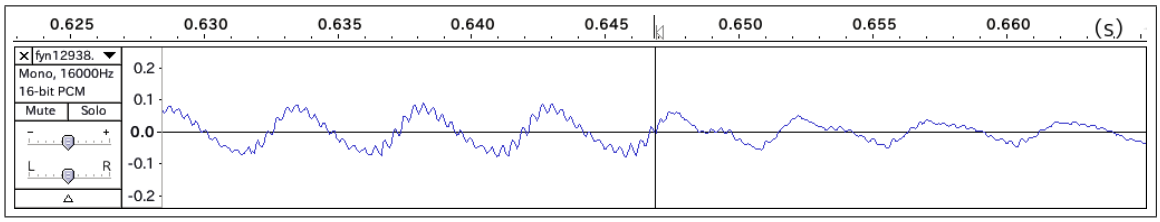

図 2.20: 「対話 /ta/i/wa/」の i/wa を人手で修正して接続した音声波形

第3章　音節境界位置変更方法

3.1　自動化方法

第 2.2.2 節の問題を解決するために，音節の精密な開始位置・終了位置を自動的に決める方法を提案した [3]．具体的には，音節素片のパワーが最大となる周波数を求め，その初期位相が “-$\frac{\pi}{2}$” となる時間を音節開始位置にする．

提案方法は，人手で作成した合成音声と同様の音声波形を得ることを目的として，位相情報を用いて音節境界位置の変更を行っている．

図 3.1 から図 3.3 に，自動化方法を行った合成音声の接続部の例を示す．図 3.1 は,「威厳 /i/ge/N/」の i/ge 間の接続部の音声波形を示している．図 3.2 は,「対立/ta/i/ri/tsu/」の ta/i 間の接続部の音声波形を示している．図 3.3 は,「乗り物 /no/ri/mo/no/」の no/ri 間の接続部の音声波形を示している．それぞれの図は縦線部が接続位置を示しており，接続部の波形に注目すると，人手で修正した音声と同様に振幅が “-” から “+” に変わる点で接続していることがわかる．接続部の波形は滑らかに接続されていることがわかる．

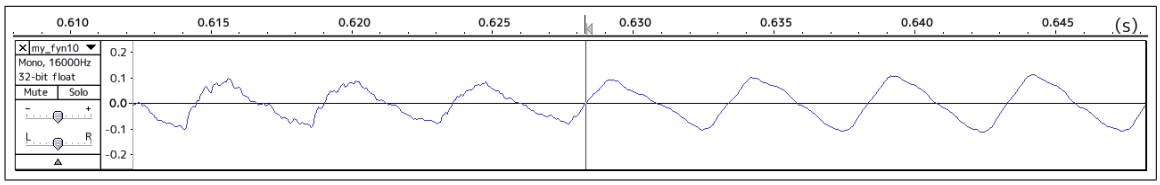

図 3.1:「威厳 /i/ge/N/」の ge/N 間の音節境界位置変更後の音声波形

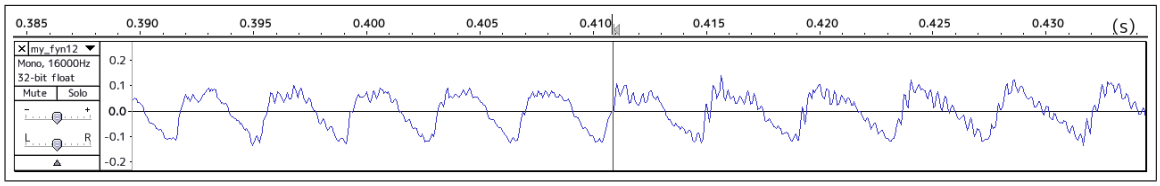

図 3.2:「対立 /ta/i/ri/tsu/」の ta/i 間の音節境界位置変更後の音声波形

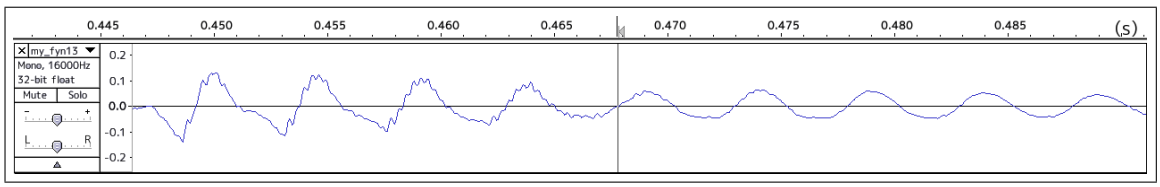

図 3.3: 「乗り物 /no/ri/mo/no/」の no/ri 間の音節境界位置変更後の音声波形

3.2 音節境界位置変更手順

図 3.4 に音節境界位置変更方法のフローチャートを示す.

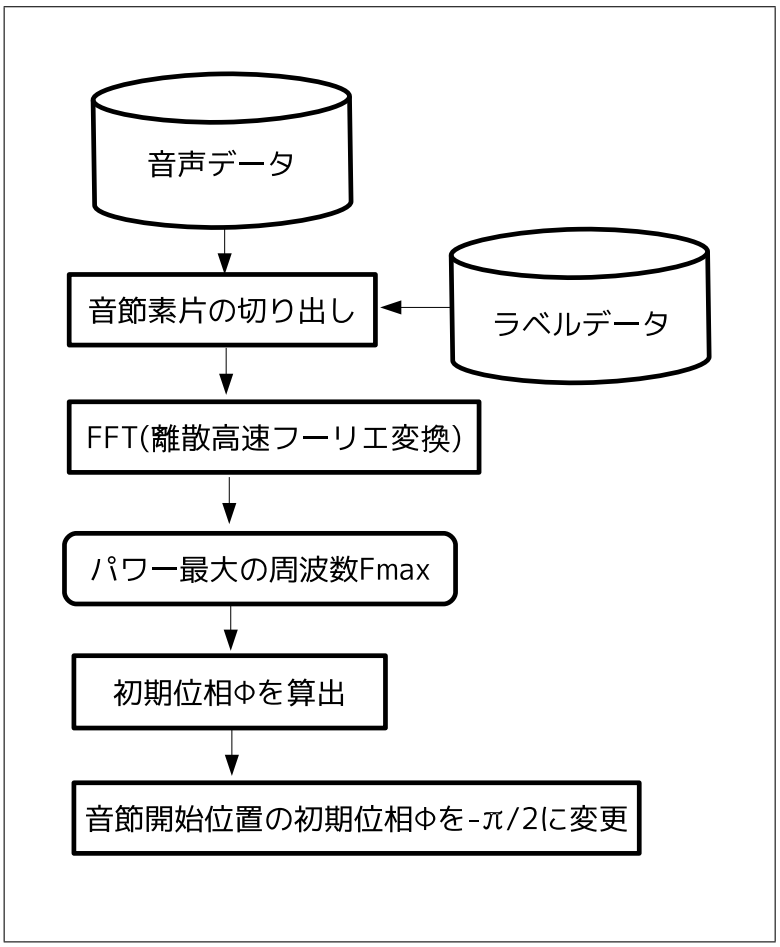

図 3.4: 音節境界位置変更方法のフローチャート

音節境界位置を変更する手順を具体的に以下に示す.

ATR のラベルから得られる音節境界位置を用いて音節開始位置・終了位置を得る．音節素片全体の波長を窓長として FFT(離散高速フーリェ変換) を行う．FFT によって得られるパワーが最大時の周波数 F_{max} における 1 周期の波長 (ms) を求める．DFT の結果から得られるパワーが最大時の周波数 F_{max} における初期位相を求める．求めた初期位相 ϕ を “-$\frac{\pi}{2}$” にするように音節開始位置を変更する．初期位相 “-$\frac{\pi}{2}$” は音声波形の振幅が “-” から “+” に変わる点である．変更する時間幅 T は $T = \frac{\frac{1}{F_{max}} * P_0}{2\pi}$ になる．また，初期位相 ϕ は連続値で扱うので，サンプリング周波数 16kHz の音声で音節開始位置を変更するため，音節開始位置の変更値は約 0.0625ms 毎の離散値になる．

3.3 周波数の誤差修正

第 3.2 節の方法において，音節素片全体の窓長に対して高速フーリエ変換を行うと，パワー最大の周波数を算出すると誤差が含まれる [3]．フーリエ変換では，窓長が，1 周期の整数倍の波長でなければ誤差が発生する．図 3.5 と図 3.6 に具体例として「免除 /me/N/jo/」の/me/ の音声波形を示す．図 3.5 において，□で囲まれた波形は「免除/me/N/jo/」の/me/における音声波形 1 周期を示している．1 周期ちょうどの波形に対して離散フーリエ変換を行う場合，周波数が正しく算出される．しかし，図 3.5 において，□で囲まれた波形は「免除 /me/N/jo/」の/me/における音声波形 1 周期ではない．1 周期ではない波形に対して離散フーリエ変換を行う場合，周波数に誤差が生じる．そのため，窓長を 1 周期の整数倍にする必要がある．そこで，誤差修正方法として，求めた波長の 1 周期の近傍の波長を窓長として DFT(離散フーリェ変換) を行う．具体的には，求めた波長を約 0.0625ms 刻みで最大 0.5ms まで増減させ，計 17 種類の窓長で，離散フーリエ変換を行い，振幅が最も 0 に近い音節開始位置を選択することで誤差の修正を試みる．この修正方法を含めた音節境界位置変更手順を提案方法とする．具体的な修正方法は，第 3.4 に記載する．

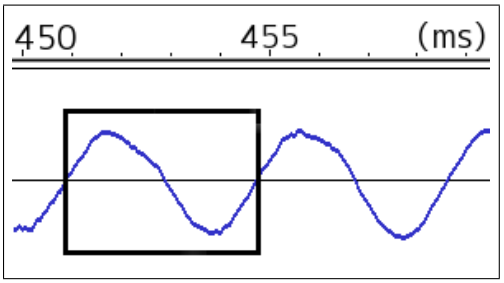

図 3.5: 「免除 /me/N/jo/」の/me/

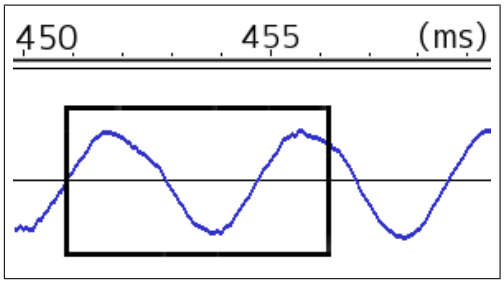

図 3.6: 「免除 /me/N/jo/」の/me/(誤差)

3.4 提案方法

図 3.7 に誤差修正方法を組み込んだ提案方法のフローチャートを示す.

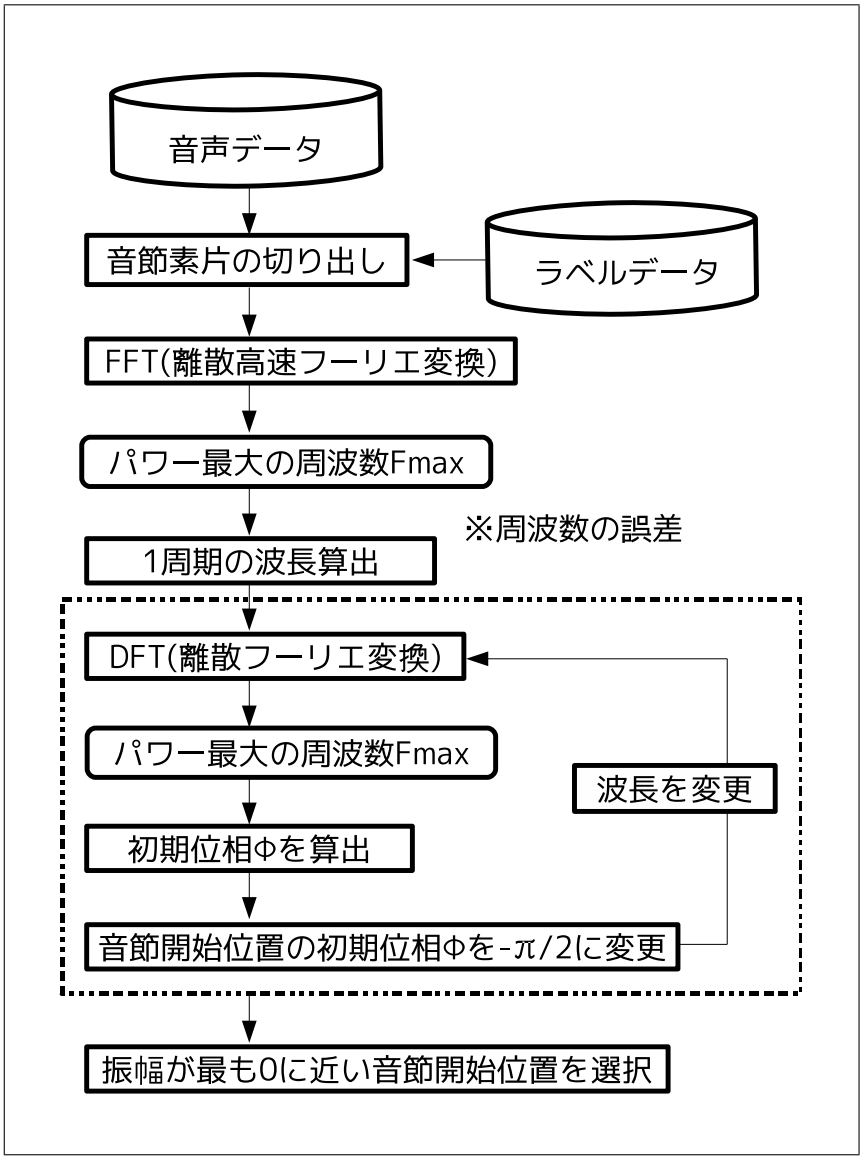

図 3.7: 提案方法のフローチャート

提案方法を具体的に以下に示す．ATR のラベルから得られる音節境界位置を用いて音節開始位置・終了位置を得る．音節素片全体の波長を窓長として FFT(離散高速フーリェ変換) を行う．FFT によって得られるパワーが最大時の周波数 F_{max} における 1 周期の波長 (ms) を求める．求めた波長には誤差が含まれるため，1 周期の近傍の波長を窓長として DFT(離散フーリェ変換) を行う．具体的には，求めた波長を約 0.0625ms 刻みで

最大 0.5ms まで増減させ，計 17 種類の窓長で，音節素片の音節開始位からDFT を行う．DFT の結果から得られるパワーが最大時の周波数 F_{max} における初期位相をそれぞれの窓長に対して求める．それぞれの初期位相 ϕ を “-$\frac{\pi}{2}$” にするように音節開始位置を変更する．初期位相 “-$\frac{\pi}{2}$” は音声波形の振幅が “-” から “+” に変わる点である．変更する時間幅 T は $T = \frac{\frac{1}{F_{max}} * P_0}{2\pi}$ になる．また，初期位相 ϕ は連続値で扱うので，サンプリング周波数 16kHz の音声で音節開始位置を変更するため，音節開始位置の変更値は約 0.0625ms 毎の離散値になる．それぞれの音節開始位置の中から振幅が最も 0 に近い音節開始位置を選択する．

3.5 FFT(高速フーリエ変換)

FFT は離散フーリエ変換を計算機上で高速に計算するアルゴリズムである．離散フーリエ変換は，時間軸上でサンプリング (離散化) して得られたデータ列に対するフーリエ変換である．N の 2 乗個のデータ数しか，扱えない (N は任意)．離散フーリエ変換の式を (3.1) 式に示す．

$$f_j = \sum_{k=0}^{N-1} x_k \exp\left(-\frac{2\pi ijk}{N}\right) \qquad j = 0, \ldots, N-1 \tag{3.1}$$

離散フーリェ変換の時間計算量は O (N2) である．高速フーリエ変換では，時間計算量を O(N log N) に減らすことが可能である．

3.6 初期位相修正

第 3.2 節で求めた初期位相から初期位相が 0 になる音節素片の切り出し位置を決める．計算式を (3.1) 式に示す．求める初期位相からのずれを時間 x(ms) とする．

$$x = \frac{\frac{1}{\text{パワーがピーク時の周波数 } (Hz)} * (\text{初期位相 } (rad/s) + \frac{\pi}{2})}{2\pi} \tag{3.2}$$

(3.1) 式で求めた初期位相からずれている時間 x(ms) の値だけ音節素片を切り出す音節境界位置を変更する．

3.7　提案方法実行例

提案方法を行うプログラムの具体的な動作例を以下に示す．例として「威厳 /i/ge/N/」の合成音声作成に用いられる「無限 /mu/ge/N/」の「N」の音節の音節境界位置を変更し，切り出す過程を示す．

3.7.1　ラベルの音節境界位置による音節素片の切り出し

最初に，ラベルに記載されている音節開始時間，音節終了時間により「N」の音節素片を切り出す．図 3.8 に切り出した音節素片の波形全体を示す．y 軸が振幅を示し，x 軸が時間 (s) を示している．図 3.9 は音節素片開始部分を拡大した図である．どちらの図も時刻 0 が音節開始位置である．表 3.1 は音節素片を切り出す時に使用したデータである．

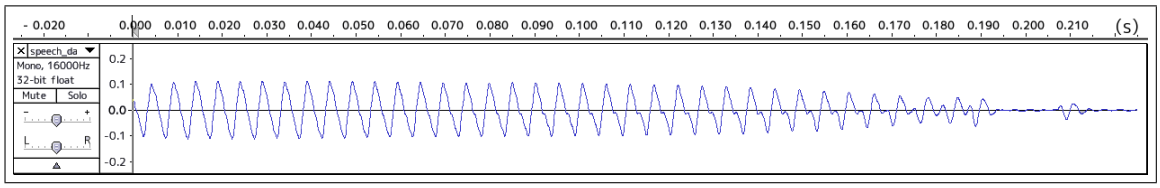

図 3.8:「無限 /mu/ge/N/」の「N」の音節素片を切り出した波形

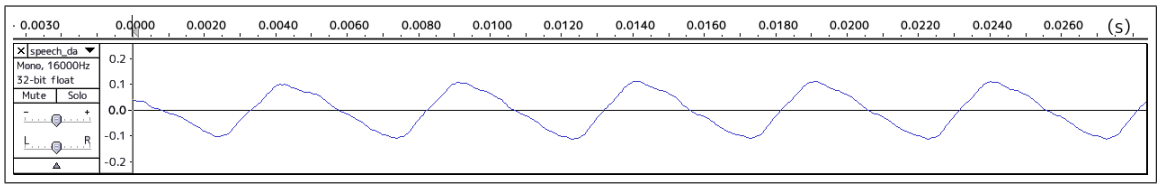

図 3.9:「無限 /mu/ge/N/」の「N」の音節素片を切り出した波形 (開始部拡大)

表 3.1:「無限 /mu/ge/N/」の「N」の音節素片を切り出す時に用いた情報

start time(ms) :	745.000000
end time(ms) :	970.000000
total time(ms) :	225.000000
start point :	11920.0
end point :	15520.0
total data number :	3601

表 3.1 において，start time(ms) と end time(ms) はラベルに記載された音節開始・終了時刻を示している．また，total time(ms) が音節の持続時間．そして，start point と end point は，音声波形の離散値における音節の開始ポイント数，終了ポイント数を示している．total data number は，音節の離散値のデータ数である．

3.7.2 パワー最大時の周波数における初期位相算出

次に，切り出した音節素片に対し FFT(離散高速フーリエ変換) を行い，パワーを求める．図 3.10 にそれぞれの周波数のパワースペクトルを示す．y 軸がパワーを示し，x 軸が周波数を示している．図 3.10 より，パワーが最大時の周波数を得ることができる．表 3.2 は求めた最大パワーとその時の周波数を求める時に用いたデータである．

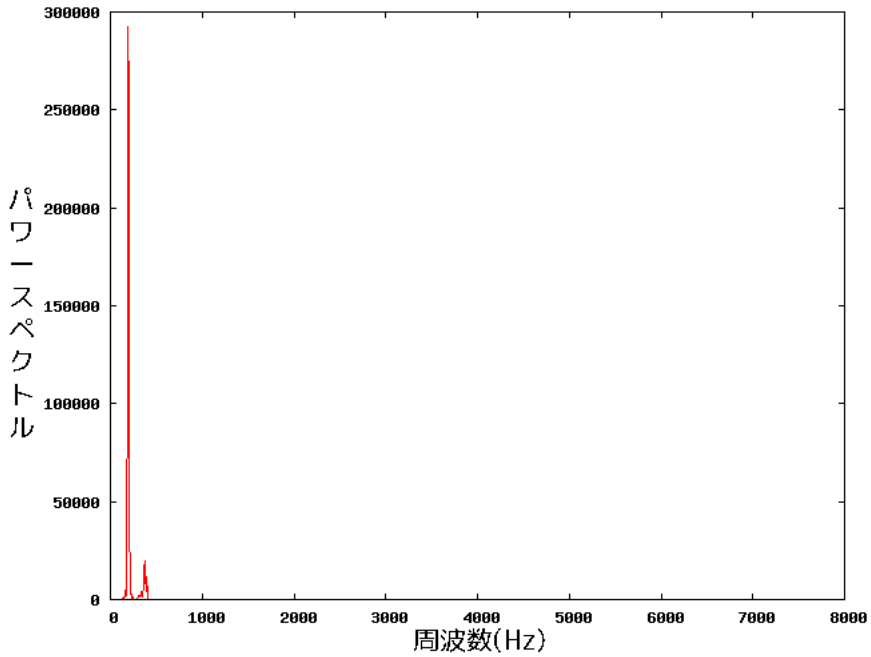

図 3.10: 「無限 /mu/ge/N/」の「N」の音節のパワースペクトル

表 3.2: 「無限 /mu/ge/N/」の「N」のパワーが最大時の周波数算出時に用いた情報

data number :	2 exp 12 = 4096
frequency resolution(hz) :	3.906250
frequency(hz) :	195.312500
max power :	282121.442077
initial phase(rad) :	1.595435
1 period length(ms) :	5.120000
1 period point :	81.920000

表 3.2 において，data number は FFT の窓長の離散値におけるポイント数を示している．FFT は，2 の X 乗 (X=0，1，2，⋯) のデータ数しか入力値として扱うことができない．そのため，音声のデータ数が 2 の X 乗でなければ，2 の X 乗に足りないデータを 0 で補うことにより 2 の X 乗として入力する．frequency resolution(hz) は，周波数分解能を示している．frequency(hz) は，図 3.10 におけるパワーが最大となる周波数を示している．そして，最大のパワーの値は，max power で示されている．パワー最大の周波数から初期位相を求めることができる．初期位相は，initial phase(rad) で示されている．そして，最後に 1 周期の波長 (ms) と，1 周期の波長の離散値がそれぞれ 1 period length と 1 period point に示されている．

3.7.3 周波数の誤差修正

周波数の誤差を修正を目的として，離散フーリェ変換の窓長を音節素片の開始時から 1 周期の整数倍にするために，窓長の ±0.5ms まで (約 0.0625ms 刻み，計 17 種類) に対して離散フーリエ変換を行い，振幅が最も 0 に近い音節開始位置を選択することで誤差の修正を試みる．次ページの表 3.3 に位相修正を行う時に用いたデータを示す．

表 3.3: 「無限 /mu/ge/N/」の「N」に対してフーリエ変換を行った結果

data number	74	75	76	77	78
frequency resolution(hz)	216.216216	213.333333	210.526316	207.792208	205.128205
frequency(hz)	216.216216	213.333333	210.526316	207.792208	205.128205
max power	1371.825709	1382.304121	1390.597346	1396.709986	1400.937979
initial phase(rad)	0.309486	0.353311	0.395940	0.437506	0.478055
1 period length(ms)	4.625000	4.687500	4.750000	4.812500	4.875000
shift time(ms)	1.384060	1.435459	1.486825	1.538225	1.589663
shift point	-22	-23	-24	-25	-25
initial amplitude	1518	1283	1248	963	963

data number	79	80	81	82	83
frequency resolution(hz)	202.531646	200.000000	197.530864	195.121951	192.771084
frequency(hz)	202.531646	200.000000	197.530864	195.121951	192.771084
max power	1404.109487	1406.285700	1407.525564	1407.013845	1405.076525
initial phase(rad)	0.517430	0.555702	0.592932	0.629561	0.665546
1 period length(ms)	4.937500	5.000000	5.062500	5.125000	5.187500
shift time(ms)	1.640986	1.692214	1.743363	1.794763	1.846361
shift point	-26	-27	-28	-29	-30
initial amplitude	728	450	88	60	-253

data number	84	85	86	87	88
frequency resolution(hz)	190.476190	188.235294	186.046512	183.908046	181.818182
frequency(hz)	190.476190	188.235294	186.046512	183.908046	181.818182
max power	1401.455628	1395.906178	1388.003210	1377.908184	1366.033870
initial phase(rad)	0.701122	0.736541	0.772195	0.808151	0.844301
1 period length(ms)	5.250000	5.312500	5.375000	5.437500	5.500000
shift time(ms)	1.898332	1.950878	2.004330	2.058753	2.114061
shift point	-30	-31	-32	-33	-34
initial amplitude	-253	-454	-916	-1267	-1773

data number	88	89	90
frequency resolution(hz)	181.818182	179.775281	177.777778
frequency(hz)	181.818182	179.775281	177.777778
max power	1366.033870	1352.002315	1336.457808
initial phase(rad)	0.844301	0.881115	0.918287
1 period length(ms)	5.500000	5.562500	5.625000
shift time(ms)	2.114061	2.170676	2.228343
shift point	-34	-35	-36
initial amplitude	-1773	-2098	-2220

表 3.3 は，表 3.2 において，1 周期の波長の離散値が 1 period point に示されており，81.920000 ポイントだった．そこで，位相を修正する目的でその ±8 ポイント (±0.5ms) の範囲で 1 ポイント (0.0625ms) 刻みで波長を増減させ，離散フーリエ変換を行った結果を示している．表 3.3 では，data number で示されている波長で，それぞれ，74 ポイントの窓長から 90 ポイントまでの窓長で離散フーリエ変換を行っている．表における 2frequency resolution(hz) から，1 period length(ms) までは，表 3.2 と同じ数値を示している．shift time(ms) は初期位相が，“-π/2” から何 ms ずれているから示している．これに対して，

離散値で shift point の値だけ音声波形の位相をずらせば，初期位相が “-π/2” になる．そして，最後に initial amplitude は，音声波形の初期位相を “-π/2” にずらした時の音節開始位置における振幅の値である．本研究では，計 17 ポイントの窓長から求まる音節開始位置の振幅の中から最も 0 に近い値の音節開始位置を音節境界位置として選択する．

3.7.4 提案方法を用いて作成した合成音声の音声波形例

提案方法を用いて音節境界位置の変更を行い，作成した合成音声の音声波形を図 3.11 から 3.13 に示す．図 3.11 に，例として「威厳/i/ge/N/」の i/ge 間の接続部の音声波形を示す. 縦線部は接続部を示している．接続部に注目すると滑らかに接続できていることがわかる．また同様に，図 3.12 に「発音/ha/tsu/o/N/」の o/N 間の接続部の音声波形を示す．また，図 3.13 に「対話 /ta/i/wa/」の ta/i 間の接続部の音声波形を示す．

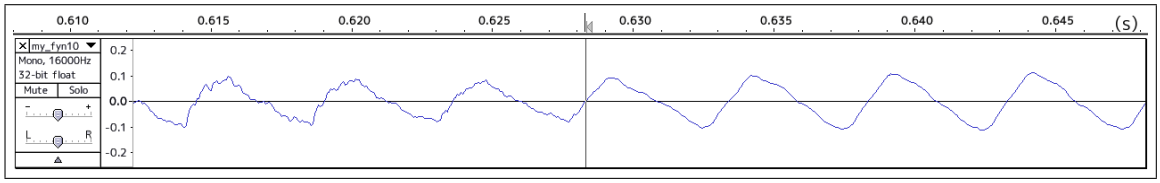

図 3.11: 「威厳 /i/ge/N/」の ge/N 間を提案方法を用いて接続した音声波形

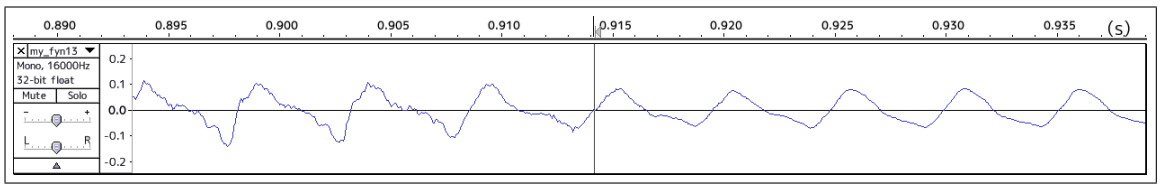

図 3.12: 「発音 /ha/tsu/o/N/」の o/N 間を提案方法を用いて接続した音声波形

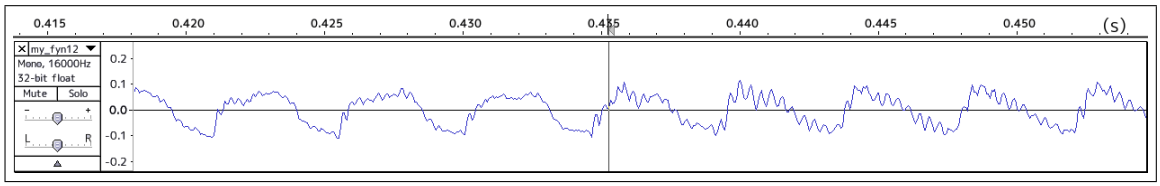

図 3.13: 「対話 /ta/i/wa/」の ta/i 間を提案方法を用いて接続した音声波形

第4章　従来方法

4.1　クロスフェード方法

音節波形接続型音声合成における従来の音節接続方法として，Hirai らが用いているクロスフェード法が挙げられる [5]．図 4.1 に Hirai らが，用いた方法のフローチャートを示す.

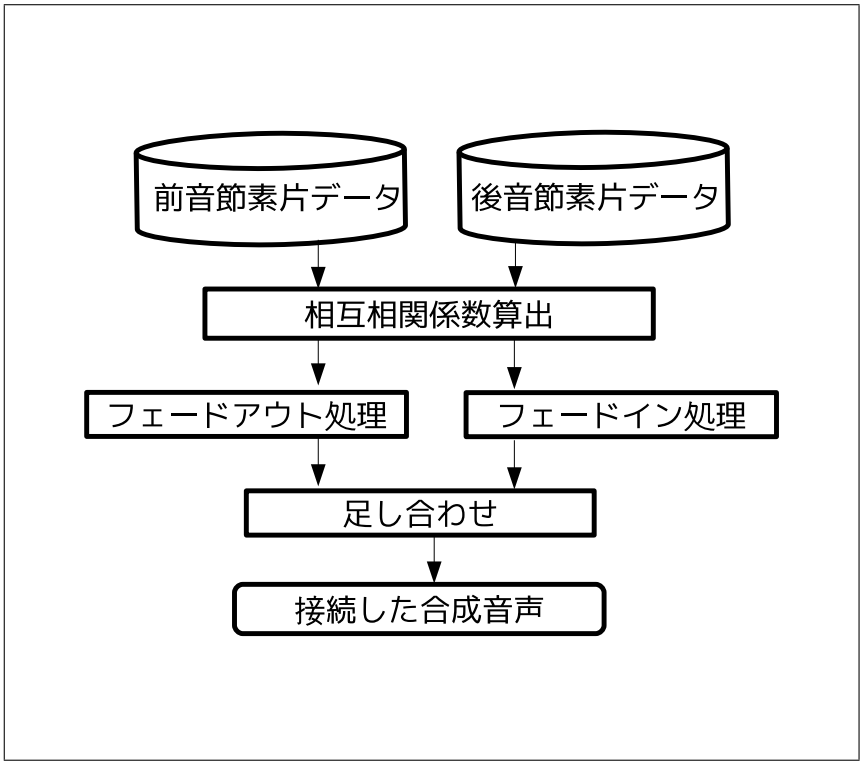

図 4.1: 従来方法のフローチャート

従来方法を具体的に以下に示す．接続したい 2 つの音節素片のデータを入力とする．それぞれの音節素片の相互相関係数を算出することで，2 つの音節素片を接続する位置を決める．そして，接続する前音節データには，振幅が終端に向けて小さくなるように線形近似を行う (フェードアウト処理)．また，接続する後音節データには，振幅が開始か

ら大きくなるように線形近似を行う (フェードイン処理)．最後に，それぞれの音声波形をフェードイン・アウト処理を施した部分で重ね合わせ，振幅を足し合わせることで 2 つの音節素片を接続する．

4.2 従来方法実行手順

従来方法の実行手順を図 4.2 に示す．

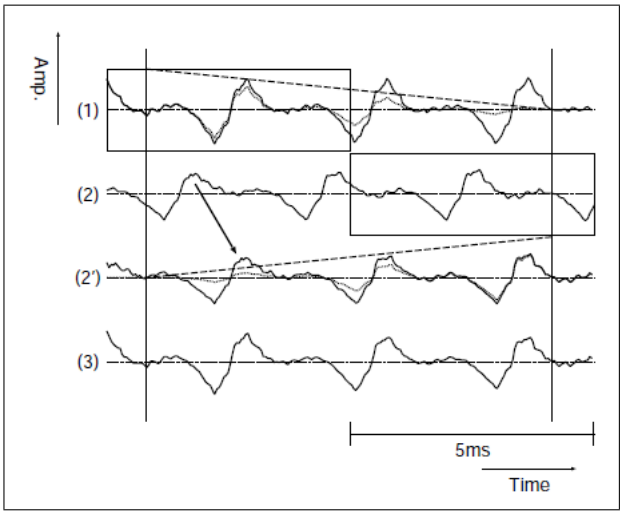

図 4.2: 従来方法の波形接続方法

図 4.2 において接続したい前の音節素片 (1)，接続したい後の音節素片 (2) の 2 つの音節素片の波形が示されている．波形における□で囲まれた部分がそれぞれの接続する波形である．

1. 後の音節素片の波形を相互相関係数を用いることによって，2 つの音節素片の相関が最も高い位置に波形をシフトする (2')．これにより，接続する位置が決定される．
2. それぞれの音声波形を点線で示された直線に近似することにより，それぞれの波形の振幅を足しあわせた時に元の波形と同様の振幅になるように重み付ける．
3. 最後にそれぞれ線形近似した (1)，(2) の波形を足しあわせることにより，(3) の波形となる．

4.3　従来方法実行例

4.3.1　相関係数を用いた音声波形シフト

まず，相互相関係数を用いて 2 つの接続する音節素片の波形を最も相関の高い位置にシフトさせる．図 4.3 と図 4.4 に接続する 2 つの音節素片を示す．前音節として,「意外 /i/ga/i/」の「i」と「機嫌/ki/ge/N/」の「ge」を接続して作成した「威厳/i/ge/N/」の「/i/ge/」．後音節として,「無限 /mu/ge/N/」の「N」の音節素片を示している．

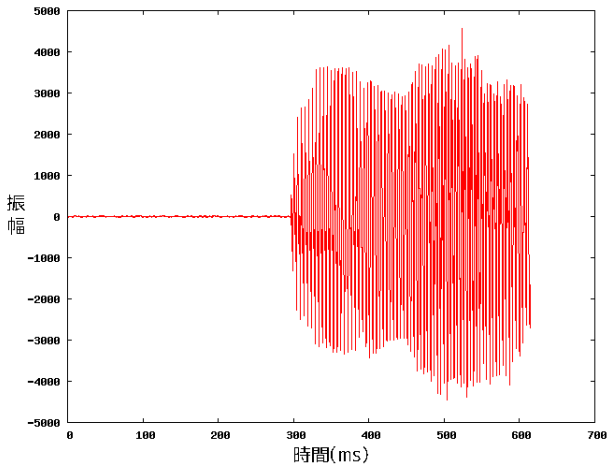

図 4.3: 「威厳 /i/ge/N/」の「/i/ge/」の音声波形

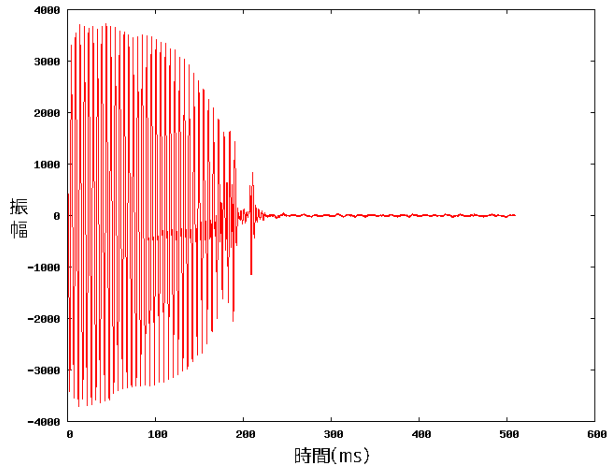

図 4.4: 「無限 /mu/ge/N/」の「/N/」の音声波形

図 4.5 と図 4.6 の 2 つの素片を最も相互相関係数の高い位置にシフトさせる．相関をとる範囲は 8.33ms で，シフト幅 4.17ms である．前音節を固定して，後音節を 4.17ms の範囲でシフトし，最も相互相関係数が高い値を探索する．図 4.5 と図 4.6 に，2 つの素片の相互相関をとった波形を示す．この時相互相関係数は，0.978382 である．2 つの図を見比べると，同じ時刻でほぼ同じ振幅になっていることがわかる．波形の形が同じ場合，第 4.3.2 節でクロスフェード処理を行う際に滑らかに波形を接続することができる．

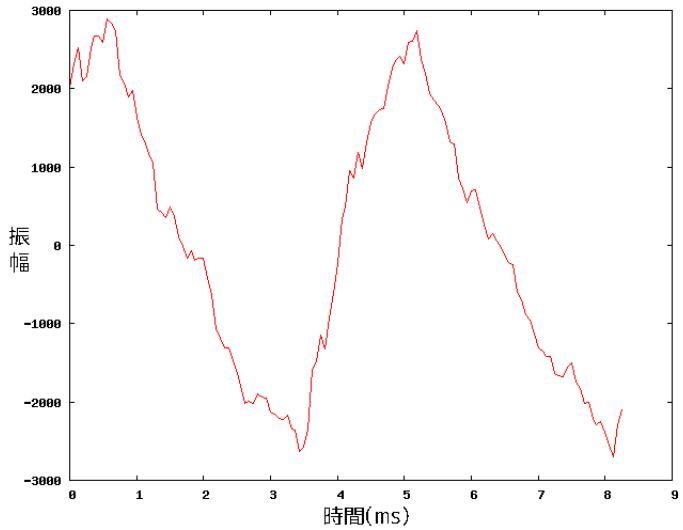

図 4.5: 「威厳 /i/ge/N/」の「/i/ge/」の相互相関をとった音声波形

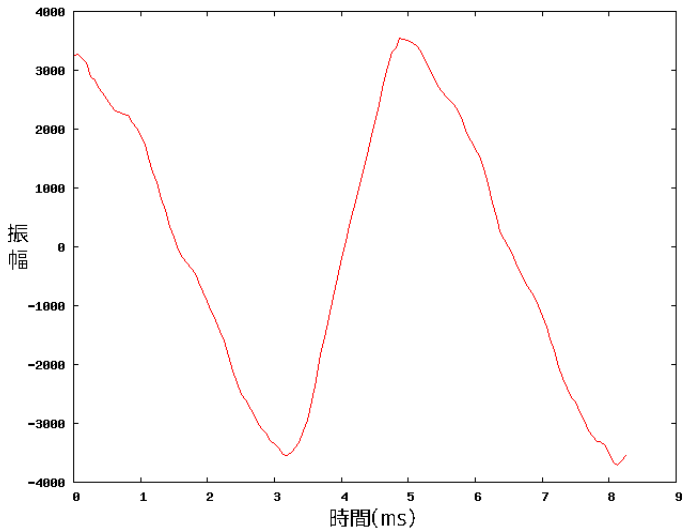

図 4.6: 「無限 /mu/ge/N/」の「/N/」の相互相関をとった音声波形

4.3.2　クロスフェード処理

波形の接続で歪みが生じないように，線形近似して重ね合わせることで，波形を滑らかに接続する．図 4.7 と図 4.8 に線形近似を行った波形を示す．図 4.7 の前音節には，振幅が終端に向かって小さくなるように線形近似を行う．また，図 4.8 の後音節には，振幅が開始から大きくなるように線形近似を行う．

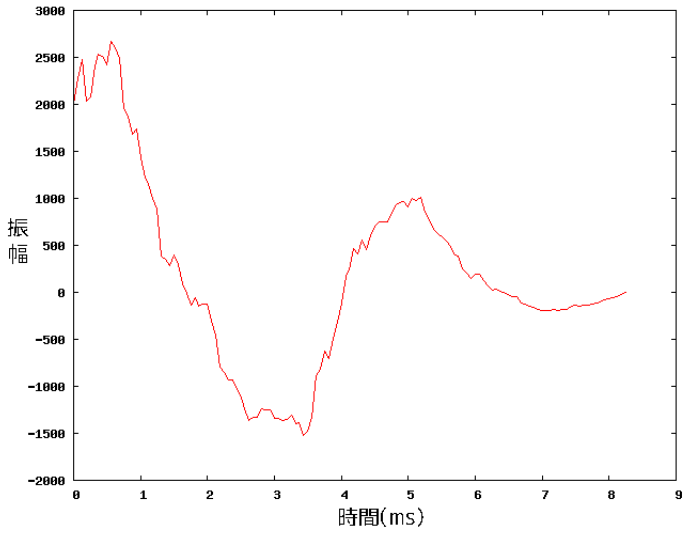

図 4.7: 「威厳 /i/ge/N/」の「/i/ge/」を線形近似した音声波形

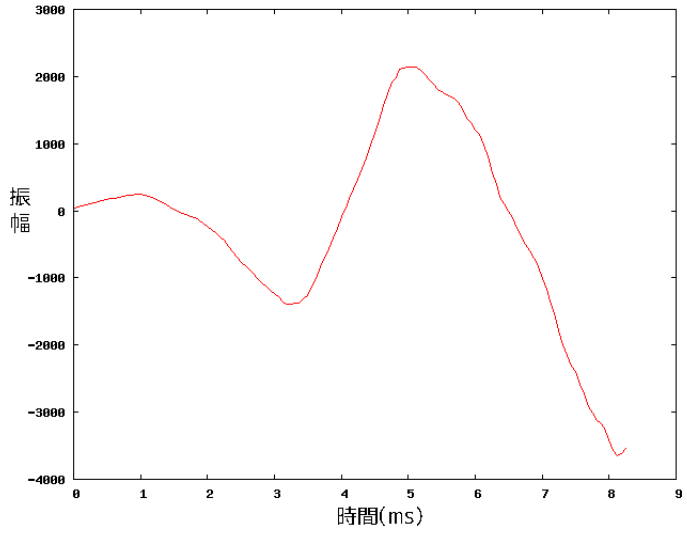

図 4.8: 「無限 /mu/ge/N/」の「/N/」を線形近似した音声波形

4.3.3　従来方法を用いて作成した合成音声の音声波形例

従来手法であるクロスフェード方法を用いて作成した合成音声の音声波形例を示す．作成した合成音声の音声波形を図 4.9 に示す．例として「威厳 /i/ge/N/」の i/ge 間の接続部の音声波形を示す. □内は接続部を示している．□内の接続部を見ると，滑らかに接続されていることがわかる．また，同様に，4.10 に示されている「発音 /ha/tsu/o/N/」の tsu/o 間の音声波形，4.11 に示されている「対話/ta/i/wa/」の i/wa 間の音声波形においても□内の接続部において滑らかに接続されている．

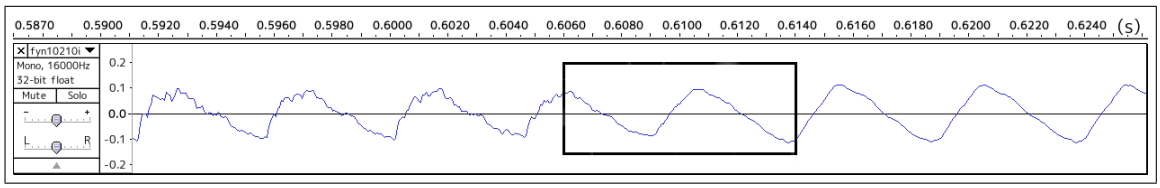

図 4.9: 「威厳 /i/ge/N/」の「ge/N」を従来方法を用いて接続した音声波形

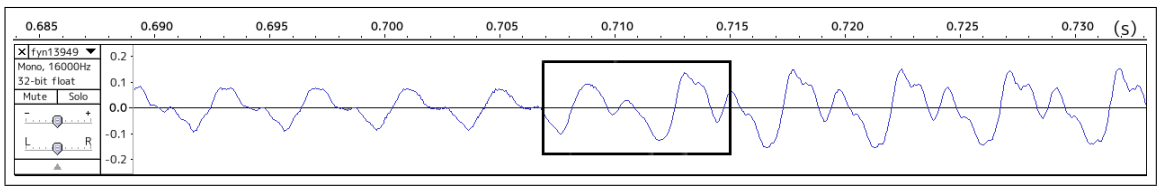

図 4.10: 「発音 /ha/tsu/o/N/」の「tsu/o」を従来方法を用いて接続した音声波形

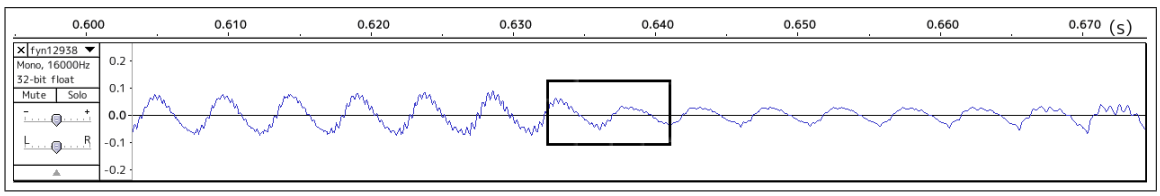

図 4.11: 「対話 /ta/i/wa/」の「i/wa」を従来方法を用いて接続した音声波形

第5章　実験条件

5.1　実験環境

本実験において音声合成に用いる音声素片は ATR 単語発話 DB Aset(1 話者 5,240 件) 内の単語を利用する．話者は女性話者 2 名 (FYN，FTK)，男性話者 2 名 (MAU，MTK) とする．聴覚実験には，以下の 4 種類の音声を用いる．

- 自然音声
  ATR 単語発話 DB Aset(5,240 件) 内の音声．
- 提案合成音声
  提案方法を用いて作成した合成音声．
- ATR 合成音声
  ラベルデータとして ATR のラベルを用い，ATR ラベルから得られる音節境界位置を用いて作成した合成音声．
- 人手合成音声
  人手で正確に作成したラベルを用いて作成した合成音声 [4]．

5.2 評価方法

提案方法で作成した合成音声の音声品質を評価するために，音声研究に関わった経験のない 5 名を対象に，オピニオン評価実験及び対比較実験を行う．評価単語数は各 100 単語とする．また，オピニオン評価は自然音声を含めて行う．

1. オピニオン評価
   音声の自然性を調査するために，オピニオン評価を行う．自然に聞こえた度合を 5 段階 (1 が最も不自然，5 が最も自然) で評価する．以下の 4 種類の音声を対象とする．

   - 自然音声
   - 人手合成音声
   - 提案合成音声
   - ATR 合成音声

2. 対比較実験
   提案方法を用いて作成した合成音声の評価のために，対比較実験を行う．

   対比較実験は，同じ内容の文節発声の 2 種類の音声を連続して聴き，どちらの音声がより自然に聞こえたかを判定する．提案合成音声，従来合成音声，人手合成音声を対象とし，以下の組み合わせで音声品質の比較を行う．

   - 提案合成音声と ATR 合成音声
   - 提案合成音声と人手合成音声

第6章　実験結果

6.1　オピニオン評価実験結果

提案合成音声の音声品質を調査するために，評価者 5 名を対象にオピニオン評価実験を行う．各話者 (女性 2 名，男性 2 名) ごとに，評価者全員のオピニオンスコアと平均を表 6.1 から表 6.4 に示す.

6.1.1　オピニオン評価実験結果:女性話者 fyn

女性話者 fyn の音声を用いて合成音声を作成し，オピニオン評価実験を行った結果を表 6.1 に示す.

表 6.1: 女性話者 fyn のオピニオン評価実験の結果 (一人当たり各評価単語数：100)

音声の種類	A	B	C	D	E	平均
自然音声	4.49	4.80	4.51	4.66	4.59	4.61
人手合成音声	3.78	3.93	4.08	3.87	3.93	3.92
提案合成音声	3.91	4.03	4.15	3.9	4.03	4.00
ATR 合成音声	2.87	3.37	3.71	2.94	3.19	3.22

表 6.1 より，女性話者 fyn のオピニオン実験結果において，提案合成音声は，人手合成音声と同等の音声品質であった．また，ATR 合成音声より品質が高い結果が得られた．実験結果より，研究の目的を達成できたといえる.

6.1.2 オピニオン評価実験結果:女性話者 ftk

女性話者 ftk の音声を用いて合成音声を作成し，オピニオン評価実験を行った結果を表 6.2 に示す.

表 6.2: 女性話者 ftk のオピニオン評価実験の結果 (一人当たり各評価単語数：100)

音声の種類	A	B	C	D	E	平均
自然音声	4.78	4.96	4.37	4.95	3.90	4.59
人手合成音声	3.70	4.51	3.89	3.69	3.25	3.81
提案合成音声	3.61	4.53	3.99	3.75	3.20	3.82
ATR 合成音声	3.12	3.82	3.69	3.00	2.96	3.32

表 6.2 より，女性話者 ftk において，提案合成音声は，人手合成音声と同等の音声品質であった．また，ATR 合成音声より品質が高い結果が得られた．実験結果より，研究の目的を達成できたといえる.

6.1.3 オピニオン評価実験結果:男性話者 mau

男性話者 mau の音声を用いて合成音声を作成し，オピニオン評価実験を行った結果を表 6.3 に示す.

表 6.3: 男性話者 mau のオピニオン評価実験の結果 (一人当たり各評価単語数：100)

音声の種類	A	B	C	D	E	平均
自然音声	4.77	4.54	4.86	4.34	4.98	4.70
人手合成音声	4.35	4.34	4.51	3.90	4.37	4.29
提案合成音声	4.06	4.07	4.30	3.64	4.19	4.05
ATR 合成音声	3.80	3.97	4.22	3.60	4.05	3.93

表 6.3 より，提案合成音声は，ATR 合成音声より品質が高い結果が得られた．しかし，人手合成音声には及ばなかった．しかし，人手合成音声とのオピニオンスコアの差は少ないため，高い音声品質が得られたといえる.

6.1.4　オピニオン評価実験結果:男性話者 mtk

男性話者 mtk の音声を用いて合成音声を作成し，オピニオン評価実験を行った結果を表 6.4 に示す.

表 6.4: 男性話者 mtk のオピニオン評価実験の結果 (一人当たり各評価単語数：100)

音声の種類	A	B	C	D	E	平均
自然音声	4.89	4.6	4.89	3.85	4.92	4.63
人手合成音声	4.41	4.23	4.73	3.65	4.40	4.28
提案合成音声	4.16	4.14	4.43	3.48	4.35	4.11
ATR 合成音声	3.95	4.1	4.41	3.44	4.25	4.03

表 6.4 より，提案合成音声は，ATR 合成音声より品質が高い結果が得られた．しかし，人手合成音声には及ばなかった．しかし，人手合成音声とのオピニオンスコアの差は少ないため，高い音声品質が得られたといえる.

6.1.5　オピニオン評価実験結果:まとめ

表 6.1 から表 6.4 のオピニオン評価実験の結果，女性話者において，提案合成音声は人手合成音声と同等の音声品質を得ることができた．したがって，本研究の目的が達成できたと言える．また，男性話者において，提案合成音声は ATR 合成音声より高い音声品質を得ることができた．したがって，本研究の有効性が証明された.

6.2 対比較実験結果

提案合成音声の音声品質を調査するために，次の 2 種類の対比較実験を行う．(1) 提案合成音声と人手合成音声，(2) 提案合成音声と ATR 合成音声．各話者 (女性 2 名，男性 2 名) ごとに，評価者全員の対比較実験結果と平均を表 6.5 から表 6.12 に示す．

6.2.1 対比較実験結果:女性話者 fyn

女性話者 fyn の音声を用いて合成音声を作成し，次の 2 種類の対比較実験を行った．(1) 提案合成音声と人手合成音声，(2) 提案合成音声と ATR 合成音声．結果を表 6.5 と表 6.6 に示す．

表 6.5: 女性話者 fyn の対比較実験 (1) の結果 (一人当たり各評価単語対 : 100)

評価者	提案合成音声	人手合成音声
A	51%	49%
B	52%	48%
C	47%	53%
D	53%	47%
E	47%	53%
平均	50.0%	50.0%

表 6.6: 女性話者 fyn の対比較実験 (2) の結果 (一人当たり各評価単語対 : 100)

評価者	提案合成音声	ATR 合成音声
A	74%	26%
B	80%	20%
C	82%	18%
D	84%	16%
E	67%	33%
平均	77.4%	22.6%

表 6.5，表 6.6 の対比較実験結果から，女性話者 fyn において，提案合成音声は人手合成音声と同等の音声品質であった．また，ATR 合成音声より品質が高い結果が得られた．実験結果より，研究の目的を達成できたといえる．

6.2.2 対比較実験結果:女性話者 ftk

女性話者 ftk の音声を用いて合成音声を作成し，次の 2 種類の対比較実験を行った．(1) 提案合成音声と人手合成音声，(2) 提案合成音声と ATR 合成音声．結果を表 6.7 と表 6.8 に示す．

表 6.7: 女性話者 ftk の対比較実験 (1) の結果 (一人当たり各評価単語対：100)

評価者	提案合成音声	人手合成音声
A	48%	52%
B	50%	50%
C	53%	47%
D	49%	51%
E	57%	43%
平均	51.3%	48.7%

表 6.8: 女性話者 ftk の対比較実験 (2) の結果 (一人当たり各評価単語対：100)

評価者	提案合成音声	ATR 合成音声
A	71%	26%
B	77%	20%
C	78%	18%
D	80%	16%
E	67%	33%
平均	74.5%	25.5%

表 6.7，表 6.8 の対比較実験結果から，女性話者 ftk において，提案合成音声は人手合成音声と同等の音声品質であった．また，ATR 合成音声より品質が高い結果が得られた．実験結果より，研究の目的を達成できたといえる．

6.2.3 対比較実験結果:男性話者 mau

男性話者 mau の音声を用いて合成音声を作成し，次の 2 種類の対比較実験を行った．(1) 提案合成音声と人手合成音声，(2) 提案合成音声と ATR 合成音声．結果を表 6.9 と表 6.10 に示す．

表 6.9: 男性話者 mau の対比較実験 (1) の結果 (一人当たり各評価単語対：100)

評価者	提案合成音声	人手合成音声
A	43%	57%
B	44%	46%
C	38%	62%
D	37%	63%
E	41%	59%
平均	40.3%	59.7%

表 6.10: 男性話者 mau の対比較実験 (2) の結果 (一人当たり各評価単語対：100)

評価者	提案合成音声	ATR 合成音声
A	59%	41%
B	56%	44%
C	58%	42%
D	52%	48%
E	59%	41%
平均	56.4%	43.6%

表 6.9，表 6.10 の対比較実験結果から，女性話者 mau において，提案合成音声は ATR 合成音声より品質が高い結果が得られた．しかし，人手合成音声には及ばなかった．

6.2.4 対比較実験結果:男性話者 mtk

男性話者 mtk の音声を用いて合成音声を作成し，次の 2 種類の対比較実験を行った．(1) 提案合成音声と人手合成音声，(2) 提案合成音声と ATR 合成音声．結果を表 6.11 と表 6.12 に示す．

表 6.11: 男性話者 mtk の対比較実験 (1) の結果 (一人当たり各評価単語対：100)

評価者	提案合成音声	人手合成音声
A	39%	61%
B	45%	55%
C	43%	57%
D	36%	64%
E	33%	67%
平均	39.1%	60.9%

表 6.12: 男性話者 mtk の対比較実験 (2) の結果 (一人当たり各評価単語対：100)

評価者	提案合成音声	従来合成音声
A	59%	41%
B	50%	50%
C	48%	52%
D	55%	45%
E	62%	38%
平均	54.6%	45.4%

表 6.11，表 6.12 の対比較実験結果から，女性話者 mau において，提案合成音声は ATR 合成音声より品質が高い結果が得られた．しかし，人手合成音声には及ばなかった．

6.2.5 対比較実験結果:まとめ

表 6.5，表 6.12 の対比較実験結果から，女性話者において，提案合成音声は人手合成音声と同等の音声品質を得ることができた．したがって，本研究の目的が達成できたと言える．また，男性話者において，提案合成音声は ATR 合成音声より高い音声品質を得ることができた．したがって，本研究の有効性が証明された．

第7章　考察

従来合成音声より提案合成音声の音声品質が劣化した合成音声の波形を調査した．その結果から得られた問題点を以下に示す．

7.1　パワーを求める範囲の問題

実験結果では音節素片全体のパワーを求めている．しかし，音節の発話開始時と発話終了時の周波数には差が生じる．そのため，パワー最大の周波数に誤差が生じ，初期位相に誤差が生じる．そこで，発話開始位置近傍の周波数を求める必要がある．

パワー最大の周波数に誤差が生じ，位相が正しく求まらない例を図 7.1 に示す．例として「検挙 /ke/N/kyo/」の「N」の周波数誤差修正を行わない自動化手法を行った結果の音声波形を示す．時刻 0 が音節開始位置である．図 7.1 を見ると音節開始位置が，振幅が “-” から “+” に変わるポイントになっていないことがわかる．これは，パワー最大の周波数に誤差が生じたため，初期位相が正しく算出できなかったためである．

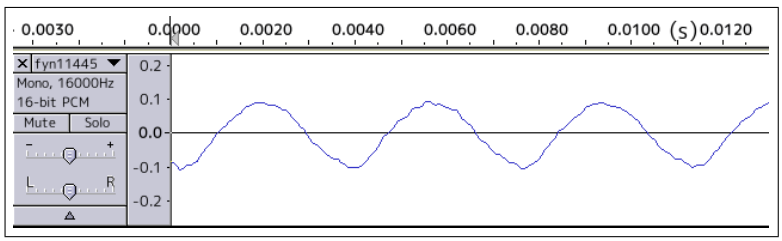

図 7.1: 音声「検挙 /ke/N/kyo/」の「N」に対して離散フーリエ変換を行った音声波形

そこで，解決策として音節の開始位置近傍の波形のみを窓長として離散フーリエ変換を行う．音節開始位置から離散値で 64 ポイント (約 4ms) の窓長で離散フーリエ変換をおこなった結果を図 7.2 に示す．

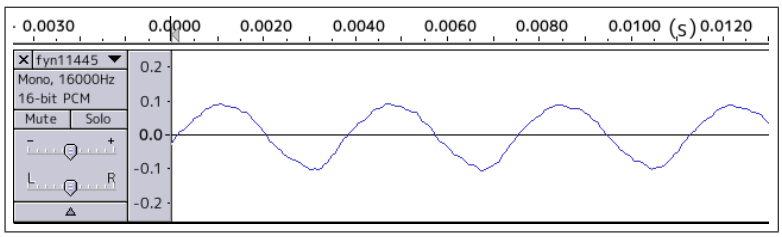

図 7.2: 音声「検挙/ke/N/kyo/」の「N」に対して離散フーリエ変換を行った音声波形 (窓長 64 ポイント)

図 7.2 では，時刻 0 において，音声波形の振幅が “-” から “+” に変わるポイントになっている．したがって，音節開始部の近傍の何周期かの窓長で離散フーリエ変換を行えば，より初期位相の精度が高まると考えられる．

7.2 ラベルの問題

実験結果より，提案方法で作成した合成音声において，音声品質が低い音声の中にラベルの不正確さが原因の音声があった．図 7.3 にラベルが原因で音声品質が劣化した合成音声「反射 /ha/N/sha/」の ha/N 間の接続部を示す．縦線部が接続部を示しており，縦線より左側の音声が「半端/ha/N/pa/」の「ha」を用いて，縦線より右側の音声が「監視 /ka/N/shi/」の「N」を用いて作成された．縦線の左側の□部に注目すると，□部の左側の音声波形とは異なる波形であり，音声「半端 /ha/N/pa/」における「N」の音声の開始部分が含まれている可能性が高い．これは，ラベルの音節境界位置が人手で作成されており，音節開始時刻・終了時刻が 5ms 間隔で記載されており，誤差を含むからと考えられる．

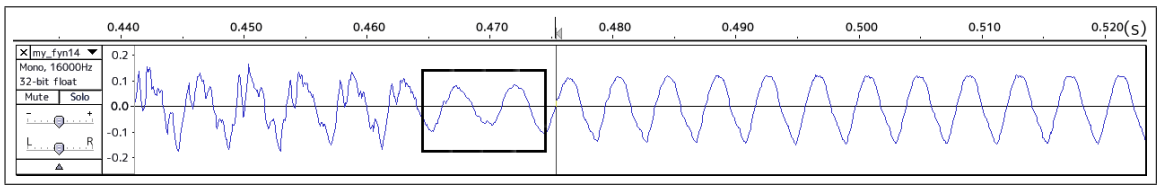

図 7.3: 合成音声「反射 /ha/N/sha/」の ha/N 間の接続部

7.3　子音の問題

本研究では初期位相を求める際に，それぞれの音素に対してフーリエ変換を行い初期位相をずらした．これは，子音と母音のパワーが異なるためである．しかし，ここで，子音に対してフーリエ変換を行った場合，子音は一定の周波数ではないため，求める初期位相に誤差が生じる．

7.4　評価者による評価のばらつき

本実験では，5 人の評価者を対象にオピニオン評価実験，対比較実験を行った．両方の実験においてそれぞれの評価者による評価にばらつきが見られた．オピニオン評価において，表 6.2 では，自然音声の評価は最も高い評価が 4.96，最も低い評価が 3.90 であった．しかし，その他の音声との比率を見るとどの評価者も違いがあまり見られない．対比較実験では評価に差が見られるが，それぞれの評価者が音声を評価する環境が異なるためであると考えられる．したがって，音声評価の実験環境を統一する必要がある．また，対比較実験において音声品質が同等になる音声が多数あったため，判断に偏りがあった場合，誤差であると考えられる．

7.5 対比較実験:従来手法との比較

また，予備実験としてクロスフェード合成音声を用いて対比較実験を行う．提案合成音声と比較し，音声品質の調査を行う．(3) 提案合成音声とクロスフェード合成音声．結果を表 7.1 から表 7.4 に示す.

7.5.1 対比較実験結果:女性話者 fyn

表 7.1: 女性話者 fyn の対比較実験 (3) の結果 (一人当たり各評価単語対：100)

評価者	提案合成音声	クロスフェード合成音声
A	46%	54%
B	42%	58%
C	37%	63%
D	38%	62%
E	38%	62%
平均	39.9%	60.1%

表 7.1 の対比較実験結果から，提案合成音声よりクロスフェード合成音声の音声品質が高い結果が得られた.

7.5.2 対比較実験結果:女性話者 ftk

表 7.2: 女性話者 ftk の対比較実験 (3) の結果 (一人当たり各評価単語対：100)

評価者	提案合成音声	クロスフェード合成音声
A	44%	56%
B	45%	55%
C	40%	60%
D	33%	67%
E	42%	58%
平均	40.5%	59.5%

表 7.2 の対比較実験結果から，提案合成音声よりクロスフェード合成音声の音声品質が高い結果が得られた.

7.5.3 対比較実験結果:男性話者 mau

表 7.3: 男性話者 mau の対比較実験 (3) の結果 (一人当たり各評価単語対：100)

評価者	提案合成音声	クロスフェード合成音声
A	42%	58%
B	45%	55%
C	40%	60%
D	35%	65%
E	44%	56%
平均	41.0%	59.0%

表 7.3 の対比較実験結果から，提案合成音声よりクロスフェード合成音声の音声品質が高い結果が得られた.

7.5.4 対比較実験結果:男性話者 mtk

表 7.4: 男性話者 mtk 対比較実験 (3) の結果 (一人当たり各評価単語対：100)

評価者	提案合成音声	クロスフェード合成音声
A	40%	60%
B	50%	50%
C	42%	58%
D	23%	77%
E	37%	63%
平均	38.3%	61.7%

表 7.4 の対比較実験結果から，提案合成音声よりクロスフェード合成音声の音声品質が高い結果が得られた.

7.5.5 対比較実験結果:まとめ

表 7.1 から表 7.4 の結果より，女性話者，男性話者共にほぼ同じ割合で提案合成音声よりクロスフェード合成音声の音声品質が高い結果が得られた.

7.6　従来手法に対する考察

7.6.1　音声品質低下の原因調査

表 7.1 から表 7.4 の結果より，提案手法の音声音声品質が従来手法より低かった原因を調査する．本研究手法では，接続点を 1 点に絞っているため，誤差が生じた際に大きな歪みが生じる．しかし，従来手法において，接続するのは，8.33ms の範囲であり，重ね合わせ処理を行うことで振幅に大きなずれが生じにくい．したがって，本研究手法においても同様に，ある程度の範囲で重ね合わせ処理を行うことで音声品質が改善されると考えることができる．

7.6.2　コストの違い

従来方法との対比較実験結果により，提案方法より従来方法の音声品質が高い結果となった．ただし，コストについて違いがあるので，一概に従来方法の性能が良いとは言えない．提案方法では，あらかじめデータベース内の音声に対して処理を行うため，1 度プログラムを実行すれば，それ以降コストは発生せず，音節素片を接続するだけで簡単に合成音声を作成できるようになる．しかし，従来方法では，音節素片を接続する際に処理を行うため，合成音声作成時に毎回コストが発生するという問題がある．そのため，コストにおいては提案方法の性能が良いと言える．

7.7 対比較実験:提案方法の改善

第 7.6 節で述べた改善案を提案手法に対して行い作成した合成音声を用いて対比較実験を行う．改善案は，提案手法+クロスフェード法である．従来手法であるクロスフェード法と比較を行うことで，音声品質の調査を行う．(4) 提案手法+クロスフェード法で作成した合成音声 (以下，提案+クロスフェード合成音声) とクロスフェード合成音声．対比較実験は女性話者 fyn と男性話者 mau に対して行い，評価者は 1 名である．結果を表 7.5 と表 7.6 に示す．

7.7.1 対比較実験結果:女性話者 fyn

表 7.5: 女性話者 fyn の対比較実験 (4) の結果 (一人当たり各評価単語対：100)

評価者	提案+クロスフェード合成音声	クロスフェード合成音声
A	43%	57%

表 7.2 の対比較実験結果から，提案+クロスフェード合成音声よりクロスフェード合成音声の音声品質が高い結果が得られた．

7.7.2 対比較実験結果:男性話者 mau

表 7.6: 男性話者 mau の対比較実験 (4) の結果 (一人当たり各評価単語対：100)

評価者	提案+クロスフェード合成音声	クロスフェード合成音声
A	46%	54%

表 7.1 の対比較実験結果から，提案+クロスフェード合成音声よりクロスフェード合成音声の音声品質が高い結果が得られた．

7.8　提案手法+クロスフェード法に対する考察

7.8.1　音声品質の改善

第 7.5 節の対比較実験結果と比較すると，提案手法とクロスフェード法を組み合わせることで，音声品質が改善されたことがわかる．しかし，相互相関係数を用いたクロスフェード合成音声に対して音声品質が及ばなかった．原因は，位相の誤差にあると考えられる．位相の誤差が生じた時，相互相関をとった場合のように振幅が等しい波形の部分で重ね合わせることができない．そのため，重ね合わせた時に波形に違和感が生じる．しかし，点と点で接続を行うより，ある程度の範囲で重ね合わせて接続する方が歪みが生じにくく，音声品質が低下しにくくなることがわかった．

第8章　おわりに

本研究では，音節素片を滑らかに接続するために，音節素片のパワーが最大となる周波数の初期位相が “-π/2” となる音節境界位置に決める自動化方法を提案した．また，周波数の誤差に対して本研究では，離散フーリェ変換の窓長を音節素片の開始時から 1 周期の整数倍にするために，窓長の最大 ±0.5ms(約 0.0625ms 刻み，計 17 種類) に対して離散フーリェ変換を行い，振幅が最も 0 に近い音節開始位置を選択することで修正を試みた．聴覚実験の結果において，提案方法により作成した合成音声は，オピニオン評価実験，対比較実験の両方の実験において ATR ラベルを用いて作成した合成音声より高い品質であることが確認できた．特に女性話者の実験結果において人手で作成した合成音声と同等の音声品質を得ることができ，本研究の目的を達成することができた．したがって，本研究の有効性が証明された．今後，考察で挙げた問題点を修正することで，男性話者においても，今回の実験結果より高い音声品質を得ることができると考えている．

謝辞

本研究を進めるに当たり，種々の助言を頂きました村田真樹教授に心から御礼申し上げます．3 年間に渡って御指導いただきました村上仁一准教授に心から御礼申し上げます．御多忙の中，助言をいただきました清水忠昭准教授に心から御礼申し上げます．本研究・本論文作成に際して，多大なる検討と様々な御助言をしていただきました徳久雅人講師に心から御礼申し上げます．

論文を執筆するにあたり，参考にさせて頂いた論文の著者，聴覚実験に被験者として協力して下さった計算機 C 工学講座博士前期過程 2 年生の滝川晃司氏，中村健太郎氏，計算機 C 工学講座博士前期過程 1 年生の奥田裕紀氏，東江恵介氏，西村拓哉氏，また学部 4 年生の野口和樹氏，藤原勇氏，古市将仁氏，三浦智氏，三谷宗一郎氏に深く感謝いたします．

参考文献

[1] 村上仁一, 水澤紀子, 東田正信: “音節波形接続による単語音声合成”, 電子情報通信学会論文誌, D-II, Vol.J85-D-II, No.7, pp.1157-1165, 2002.

[2] 居村太介, 村上仁一, 池原悟: “波形接続型音声合成のフレーズへの適用”, 言語処理学会第 14 回年次大会, pp.965-968, 2008.

[3] 橋本浩志, 村上仁一, 池原悟: “波形接続型音声合成における位相情報を利用した音節境界位置の決定方法”, 日本音響学会 2009 年秋季研究発表会, pp.411-412, 2009.

[4] 石田隆浩: “アクセントを考慮した波形接続型単語音声合成”, 鳥取大学大学院工学研究科修士論文，2004.

[5] T. Hirai and S. Tenpaku: “USING 5ms SEGMENTS IN CONCATENATIVE SPEECH SYNTHESIS”, 5th ISCA Speech Synthesis Workshop, pp.37-42, 2004.

[6] 藤尾聡, 村上仁一, 池原悟: “音節波形接続型音声合成における自動セグメンテーションの影響”, 電子情報通信学会技術研究報告, 思考と言語, TL2006-53, pp.73-78, 2007.

[7] 橋本浩志, 村上仁一: “波形接続型音声合成における位相情報を利用した音節境界位置決定方法”, 日本音響学会 2010 年秋季研究発表会, pp.225-226, 2010.

付録

- 各話者におけるオピニオン評価実験に用いた各単語と各評価者の評価値
- 各話者における対比較実験に用いた各単語と各評価者の評価値
- 実験に用いた合成音声と音節素片

オピニオン評価実験結果 (fyn)

※表中の A から E は評価者を示す.

オピニオン評価実験結果 (ftk)

※表中の A から E は評価者を示す.

オピニオン評価実験結果 (mau)

※表中の A から E は評価者を示す.

オピニオン評価実験結果 (mtk)

※表中の A から E は評価者を示す.

対比較実験結果 fyn(1)

評価対象:

- 提案合成音声
- 人手合成音声

※表中の 1 と 0 は，1 なら提案合成音声を選択，0 なら人手合成音声を選択したことを示す．また A から E は評価者を示す.

対比較実験結果 fyn(2)

評価対象:

- 提案合成音声
- ATR 合成音声

※表中の 1 と 0 は，1 なら提案合成音声を選択，0 なら従来合成音声を選択したことを示す．また A から E は評価者を示す.

対比較実験結果 fyn(3)

評価対象:

- 提案合成音声
- クロスフェード合成音声

※表中の 1 と 0 は，1 なら提案合成音声を選択，0 なら自然音声を選択したことを示す．0.5 はまったく同じ音質に聴こえた場合を示す．また A から E は評価者を示す．

対比較実験結果 fyn(4)

評価対象:

- 提案+クロスフェード合成音声
- クロスフェード合成音声

※表中の 1 と 0 は，1 なら提案合成音声を選択，0 なら自然音声を選択したことを示す．0.5 はまったく同じ音質に聴こえた場合を示す．また A は評価者を示す．

対比較実験結果 ftk(1)

評価対象:

- 提案合成音声
- 人手合成音声

※表中の 1 と 0 は，1 なら提案合成音声を選択，0 なら人手合成音声を選択したことを示す．また A から E は評価者を示す.

対比較実験結果 ftk(2)

評価対象:

- 提案合成音声
- ATR 合成音声

※表中の 1 と 0 は，1 なら提案合成音声を選択，0 なら従来合成音声を選択したことを示す．また A から E は評価者を示す.

対比較実験結果 ftk(3)

評価対象:

- 提案合成音声
- クロスフェード合成音声

※表中の 1 と 0 は，1 なら提案合成音声を選択，0 なら自然音声を選択したことを示す．0.5 はまったく同じ音質に聴こえた場合を示す．また A から E は評価者を示す．

対比較実験結果 mau(1)

評価対象:

- 提案合成音声
- 人手合成音声

※表中の 1 と 0 は，1 なら提案合成音声を選択，0 なら人手合成音声を選択したことを示す．0.5 はまったく同じ音質に聴こえた場合を示す．また A から E は評価者を示す.

対比較実験結果 mau(2)

評価対象:

- 提案合成音声
- ATR 合成音声

※表中の 1 と 0 は，1 なら提案合成音声を選択，0 なら従来合成音声を選択したことを示す．0.5 はまったく同じ音質に聴こえた場合を示す．また A から E は評価者を示す.

対比較実験結果 mau(3)

評価対象:

- 提案合成音声
- クロスフェード合成音声

※表中の 1 と 0 は，1 なら提案合成音声を選択，0 なら自然音声を選択したことを示す．0.5 はまったく同じ音質に聴こえた場合を示す．また A から E は評価者を示す．

対比較実験結果 mau(4)

評価対象:

- 提案+クロスフェード合成音声
- クロスフェード合成音声

※表中の 1 と 0 は，1 なら提案合成音声を選択，0 なら自然音声を選択したことを示す．0.5 はまったく同じ音質に聴こえた場合を示す．また A は評価者を示す．

対比較実験結果 mtk(1)

評価対象:

- 提案合成音声
- 人手合成音声

※表中の 1 と 0 は，1 なら提案合成音声を選択，0 なら人手合成音声を選択したことを示す．0.5 はまったく同じ音質に聴こえた場合を示す．また A から E は評価者を示す.

対比較実験結果 mtk(2)

評価対象:

- 提案合成音声
- ATR 合成音声

※表中の 1 と 0 は，1 なら提案合成音声を選択，0 なら従来合成音声を選択したことを示す．0.5 はまったく同じ音質に聴こえた場合を示す．また A から E は評価者を示す.

対比較実験結果 mtk(3)

評価対象:

- 提案合成音声
- クロスフェード合成音声

※表中の 1 と 0 は，1 なら提案合成音声を選択，0 なら自然音声を選択したことを示す．0.5 はまったく同じ音質に聴こえた場合を示す．また A から E は評価者を示す．

実験に用いた合成音声と音節素片

合成音声に用いた音素:

- FYN
- アクセント考慮
- 音節単位
- 音素境界接続

合成音声作成に用いるために選択した音節素片を記載する.