子音/ｔ/をもつ日本語音声の合成

中尾 睦彦*　岸本 昌也**　濱田 憲治***

Speech Synthesis of Japanese Language having Consonant /t/

Mutsuhiko NAKAO,Masaya KISHIMOTO,Kenji HAMADA

ABSTRACT

An attempt is made to synthesize Japanese language with the consonant /t/. It belongs to 'stops', and looks like the consonant /s/ in a spectrum figure. On the basis of experimentally obtained spectra, the consonant part is synthesized by filtering white Gaussian noise. Furthermore, two synthesis methods are carried out to decrease the number of filtering times. Synthetic speech is firstly evaluated by human auditory sense, and then is evaluated by quantification theory I.

KEY WORDS: speech synthesis, consonant /t/, white gaussian noise, quantification theory I

１．はじめに

近年、人間に対する自動サービスシステムが発達するにつれて、音声による内容伝達が極めて重要になっている。然るに現在の音声発生装置は、機械的な合成音の域を出ず、十分満足できるものとはいい難い。本研究室のこれまでの一連の研究において、子音/k/,/s/,/h/をもつ日本語音声、また母音/a/,/i/,/u/,/e/,/o/の解析および合成は完了している[4)5)6)] そのため本研究では、子音/t/をもつ日本語音声の合成を試みるが、日本語「た」行音の子音部の音声信号は、持続時間が非常に短く特徴把握が困難である[1)]。そのため十分な解析を行い、特徴を把握した上で合成の段階に入る事にした。本研究では、子音/t/をもつ日本語音声の時間波形およびスペクトルを解析しそれに基づいて合成を試みる。

2. 閉鎖音[1)]

子音/t/は閉鎖音に属している。閉鎖音の極めて重要な調音的特徴は、声道の瞬間的な妨害である。調音的な閉鎖の持続時間はさまざまであるが、通常は 50～100 ms で、それに引き続き阻害の背後に閉じ込められていた気圧が空気の破裂を伴って解放される。典型的な場合、破裂は持続時間にして 5～40ms である。

＊1　電気情報工学科
＊2　明石高専卒業生（現：京都大学　工学部　電気電子工学科）
＊3　明石高専卒業生（現：京都大学　工学部　電気電子工学科）

閉鎖音の閉鎖解放の時間は高々40 ms に過ぎず、大抵の場合はもっと短い。閉鎖音の破裂スペクトルは調音位置によって変わる。このスペクトルの変化は、特定の調音形態によって決定される共鳴特性に起因する。図 1 に、閉鎖音/t/をもつ日本語「た」行音の音声信号の時間波形を示す。いずれも大きい振幅の母音に対して、小さい振幅で、非常に短い閉鎖音が先行しているが、「ち」と「つ」に関しては持続時間が他と比べてやや長いことが見て取れる。

3. 子音/t/をもつ日本語音声の構造

「た」行音の音声信号は、子音部、フォルマント遷移部、母音部に分類される。その一例である日本語「た」についての構成図を図 2 に示す。

一般に、発話中の声道形状の変化は、声道共鳴の変化によって音響的に示される。その音響的変化は、基礎を成す調音変化と同じ持続時間を有する。閉鎖音の調音に関して、閉鎖音から母音、あるいは母音から閉鎖音への遷移が持続時間にして約 50 ms であると言われている。この 50 ms という時間内において、全てのフォルマント周波数は閉鎖音に対する値から母音に対する値へと移行する。これをフォルマント遷移と呼ぶ。

フォルマント遷移は F1 周波数遷移（調音方法に対

する Cue）と F2 と F3 周波数遷移（調音位置に対する Cue）との組み合わせとなる調音結合情報を含む。例えば、「た」という音声を合成する時には閉鎖子音部/t/と母音部/a/の境界にフォルマント遷移部を合成しなければならない。この遷移部によって自然音声により近い合成音声になるのである。

４．日本語「た」行音の子音部の波形及びスペクトル解析

以下に、日本語「た」と「ち」の音声信号の時間波形と子音部のスペクトルを示す。サンプリング周波数は 44.1 kHz とし、時間波形の縦軸は最大値、最小値が ±1 以上にならないように正規化された振幅を表す。波形解析の目的は　音声信号波形の子音部の最大振幅と母音部のそれとの振幅比、および子音継続時間を求めるところにある。スペクトル解析では音声信号のパワースペクトル密度の解析を行う。

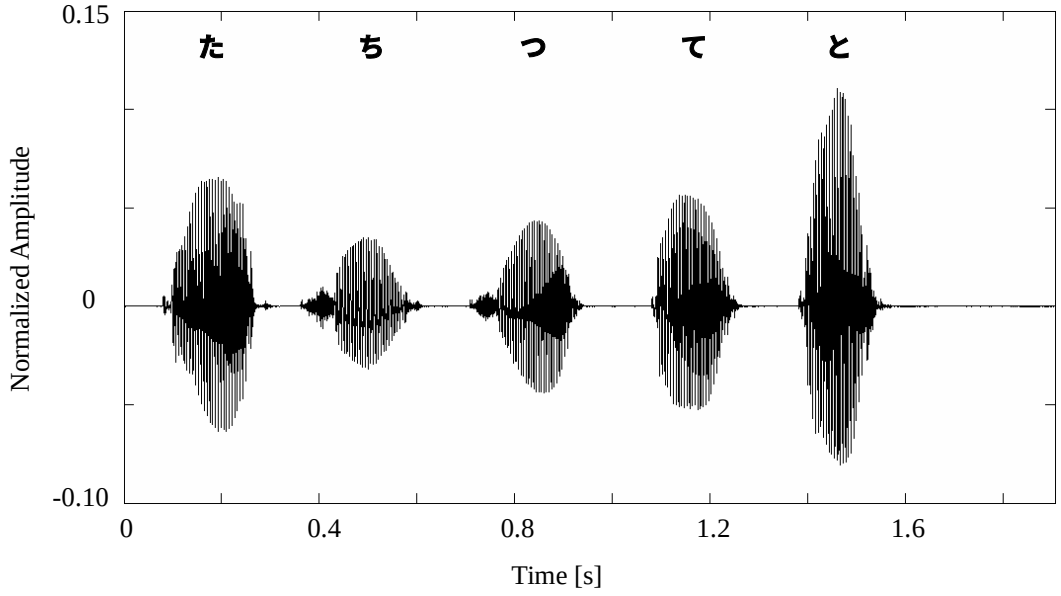

図１　日本語「た」行音の音声信号の時間波形

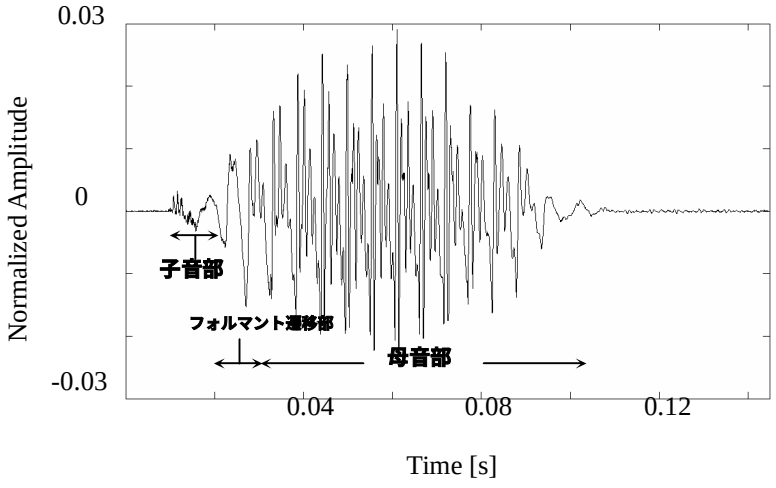

図２　日本語「た」の音声信号の構成

４･１　日本語「た」と「ち」の子音部の波形及びスペクトル

図 3 に子音部の短い日本語「た」の子音部のスペクトルを示す。図 4 に子音部の比較的長い日本語「ち」の音声信号波形を、図 5 に日本語「ち」の子音部のスペクトルを示す。

４･２　日本語「た」行音の子音部/ｔ/の特徴

5 人のサンプル音声（A~E で区別する）から、表 1

と表2に日本語「た」と「ち」の子音継続時間および振幅比を求めた。

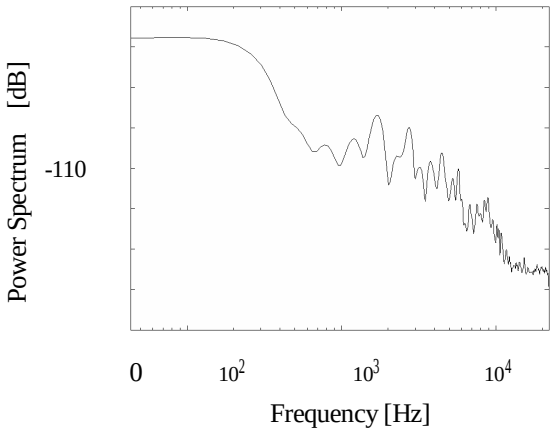

図3　日本語「た」の子音部のスペクトル

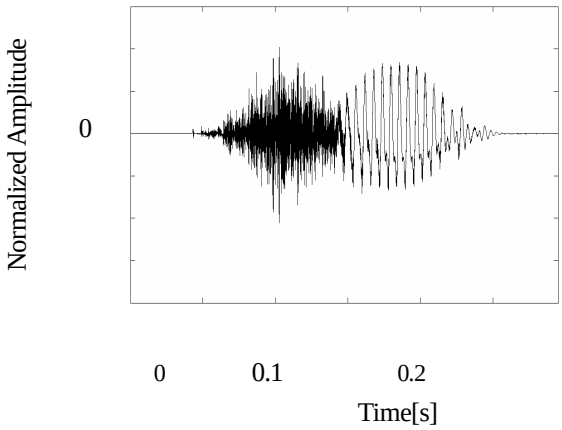

図4　日本語「ち」の音声信号波形

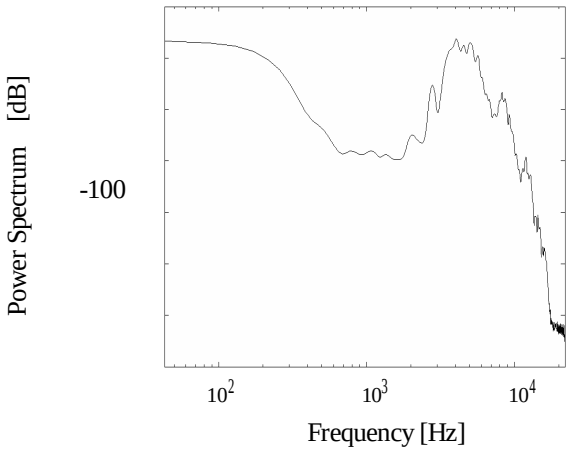

図5　日本語「ち」の子音部のスペクトル

表1　日本語「た」の子音継続時間および振幅比

サンプル	継続時間（ms）	振　幅　比
たA	4.25	0.342
たB	6.23	0.271
たC	5.36	0.189
たD	6.75	0.223
たE	4.97	0.322
平均	5.51	0.269

表2　日本語「ち」の子音継続時間および振幅比

サンプル	継続時間（ms）	振　幅　比
ちA	92.25	1.020
ちB	98.69	1.225
ちC	84.54	0.998
ちD	80.40	0.956
ちE	90.20	1.105
平均	89.22	1.061

これらの表から「た」の子音継続時間は短く母音部の10分の1程度であり、「ち」の子音継続時間は長く「た」の15倍以上あることが分かる。また母音部からみた子音部の振幅比も「ち」のほうが大きいことが分かる。

一方、「た」の子音部のスペクトルは高周波に向かって振動しながら減衰していくのに対し、「ち」の子音部のスペクトルは4.5kHzにピークを持ち以後高周波に向かって急速に減衰している。

ここでは「た」と「ち」についてその特徴を記述したが、他の「た」行音の「つ」は「ち」に近く、「て」と「と」は「た」に近い特徴を持っている。

5．ガウス雑音を用いた音声信号の合成方法

日本語「た」行音の子音部を合成するにあたって、本研究では振幅がガウス分布（正規分布）に従い、相互に無相関なランダム系列を用いる。

すなわち、系列における各サンプル値の振幅はそれぞれ独立で、各サンプル値の振幅は、

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \qquad (1)$$

で与えられる確率密度関数に従って発生する。この式では平均値が0で、分散は1であるとしている。式(1)より発生させる時系列は無相関であり、したがってガウス雑音となる。

子音部の合成では、その波形のスペクトルがサンプル音声の子音部のそれに近くなるように、ガウス雑音のスペクトルを調整する必要がある。ここでは以下のようにして行う。

（1） ガウ雑音の振幅を変化させそのスペクトル強度を変化させる。

（2） これをサンプル音声の子音部のスペクトルと比較し、強度が一致する周波数帯域を求める。

（3） それを通過させる帯域フィルタを設計し、ガウス雑音から必要な部分を抜き出す。

（4） この操作を繰り返し、抜き出した信号を加え合わせて子音部を合成する。

最後に、この合成信号の振幅及び継続時間を調整したものを、フォルマント遷移させた母音部に付加し、日本語「た」行音を合成する。図６に、上記のガウス雑音を用いた音声信号の合成方法の一例を示す。

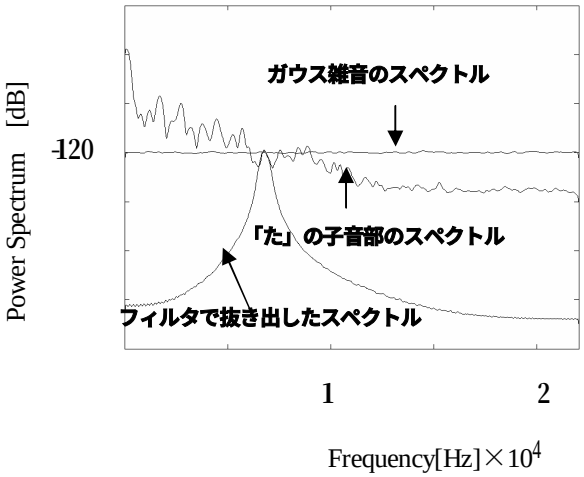

図６　ガウス雑音を用いたスペクトルの合成

６．日本語「た」行音の子音部の合成法１

図７に、５．で述べた方法による日本語「た」行音の子音部の合成結果を示す。

ここで、行った合成には、一つの子音あたりに平均すると 15 個程度のフィルタを用いている。この場合、フィルタの個数は図６に示すようにスペクトルの形状における比較的に大きい峰の部分の個数に依存している。また、その帯域は峰の部分の帯域に対応している。

７．日本語「た」行音の子音部の合成方法２

６．で述べた合成方法には多くのフィルタが必要である。その理由として実現すべきスペクトルに多くの凹凸が存在するからである。

そこで音響学的にみて、「た」行と「さ」行は発音時の口の動きがよく似ている点に着目した。

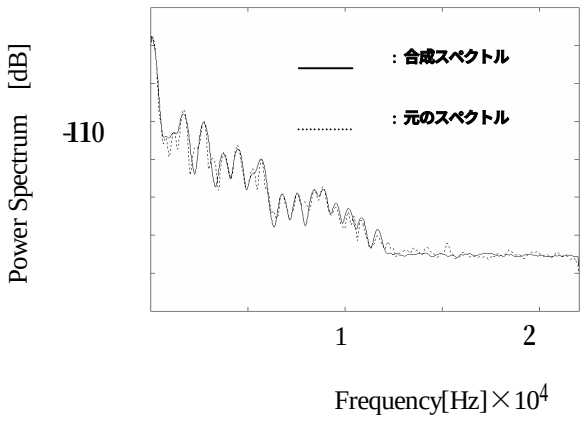

図７　日本語「た」の子音部の合成スペクトル

図８に日本語「さ」の子音の後半部分（子音継続時間を「た」のそれと同じにしたもの）および「た」の子音部のスペクトルを示す。

このように、日本語「た」行音の子音部と日本語「さ」行音のそれはスペクトルにおいて類似しているので、この子音を長時間発音すればスペクトルの平均化が可能になる。その結果スペクトルの凹凸が減少することになる。そこで子音/s/を用いて子音/t/の合成が期待できる。図９に、日本語「さ」の子音部の合成結果を示す。ただし、元の日本語「さ」行音の子音部は、スペクトルの平均化を行うために声道の形を変化させずに５～６秒間発音し続けたものである。

このようにして合成した子音/s/を加工し、子音継続時間を子音/t/のそれと同じにすれば、子音/t/の合成が期待できる。図１０に日本語「た」の子音部の合成結果を示す。

加工後のスペクトルと「た」のスペクトルとは、図のように、形がずれているが、その原因は、子音/s/の時系列の局所的な部分のスペクトルを求めたことによると思われる。この時系列の有効性は聞き取りテストで判定することにする。

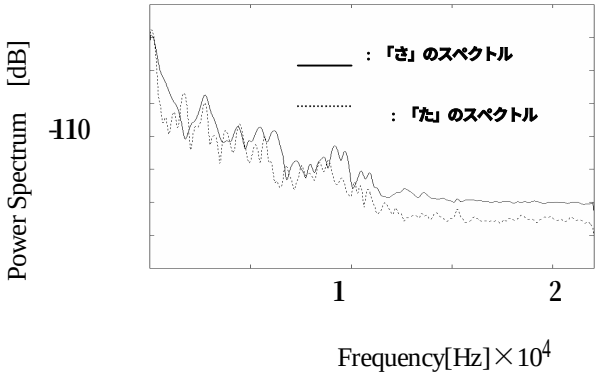

図８　日本語「さ」と「た」の子音部のスペクトル

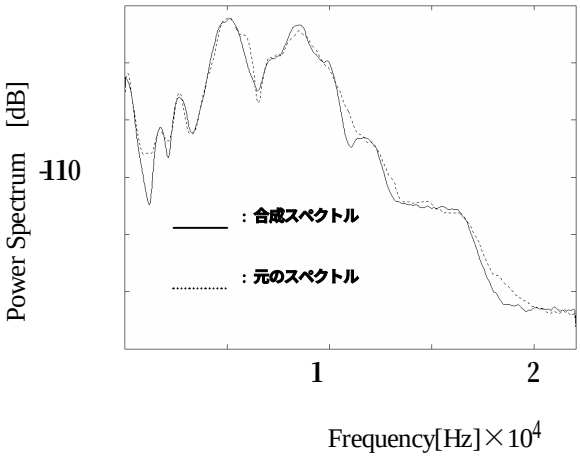

図９　日本語「さ」の子音部の合成結果

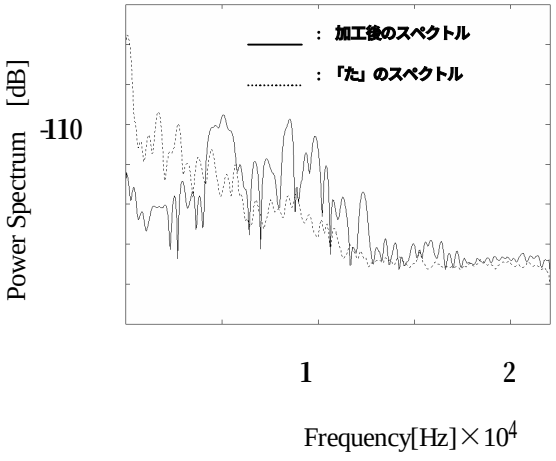

図 10　日本語「た」の子音部の合成結果

８．日本語「た」行音の子音部の合成法３

フィルタを適用する回数をできるだけ少なくするには、合成法２のようにスペクトルの平均化が可能な子音を用いるという方法以外に、元の子音のスペクトルそのものを近似するという方法がある。

図１１と図１２に、日本語「た」行音の子音部の元のスペクトルおよびこのスペクトルを近似して合成した結果を示す。なお、図１１に示した子音は、低周波をＬＰＦで再現し、10^3Ｈｚより高周波域を７個のフィルタで再現している。それに対して、図１２に示した子音は、低周波域をＬＰＦで再現し、10^3Ｈｚより高周波域を２，３個のフィルタで再現している。

ここで、10^3Ｈｚより高周波域に適用するフィルタの個数を変えた理由は、後にスペクトル近似の度合いと明瞭度との関連性を調べるためである。

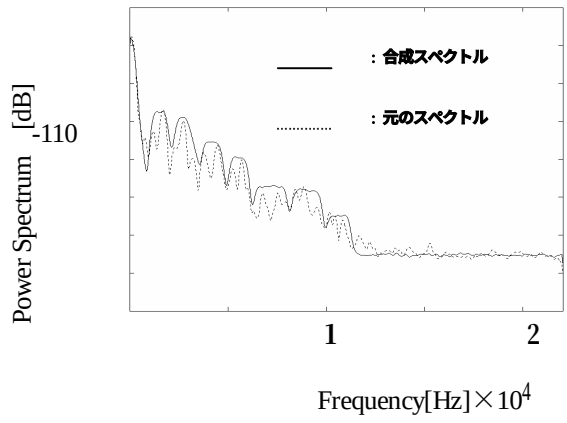

図 11　日本語「た」の子音部のスペクトルとスペクトル近似後の合成結果

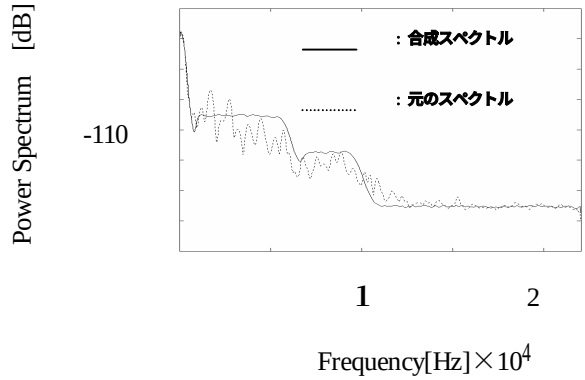

図 12　日本語「た」の子音部のスペクトルとスペクトル近似後の合成結果

９．合成音声信号の評価

６．から８．で合成した音声信号に対して、人間の聴覚による評価を行う必要がある。評価方法は、合成音声を聴いてもらい、目的の日本語「た」行音に聞こえる場合は 2 点、どちらとも言えない場合は 1 点、聞こえない場合は 0 点、というように 3 段階でアンケートに答えてもらい、そのアンケート結果を数量化理論Ｉ類で分析するという方法を採用する。

９･１　アンケート結果の解析

ひとつの日本語「た」行音について、フィルタの個数（スペクトル近似の度合い）と明瞭度との関連性を調べるために以下に述べる(1)~(3)の合成音声を用意し、また、フォルマント遷移部の有無と明瞭度との関連性を調べるために以下に述べる(4)の合成信号を用意した。更に、子音/s/を加工して合成した子音/t/と明瞭度との関連性を調べるために以下に述べる(5)の合成信号を用意する。これらの 5 つの合成信号を評価すれば、

後述の数量化理論Ⅰ類を用いて、アンケートをとっていない実験条件の合成信号の評価をも予測することができる。

(1)　合成方法1を用いたもの
日本語「た」行音の子音部のスペクトルを15個程度のフィルタを用いて忠実に合成し、これにフォルマント遷移させた母音を付加したもの。

(2)　合成方法3を用いたもの(a)
日本語「た」行音の子音部のスペクトルを近似して、8個のフィルタを用いて合成し、フォルマント遷移させた母音を付加したもの。

(3)　合成方法3を用いたもの(b)
日本語「た」行音の子音部のスペクトルを近似して、3～4個程度のフィルタを用いて合成し、これにフォルマント遷移させた母音を付加したもの。

(4)　フォルマント遷移部をなくしたもの
日本語「た」行音の子音部のスペクトルを15個程度のフィルタを用いて忠実に合成し、フォルマント遷移を含まない母音を付加したもの。

(5)　合成方法2を用いるもの
子音/s/を8個のフィルタを用いて合成したものを加工し、これにフォルマント遷移させた母音を付加したもの。

これら合成音を本校機械工学科1年生と電気情報工学科学生75人に試聴して評価してもらった。その集計結果は紙幅の関係で割愛する。

9・2　数量化理論Ⅰ類による評価

数量化理論とは、程度・状態・有無、または、はい・いいえといったような質的データに数量を与え、重回帰分析・主成分分析・判断分析と同じような多次元的解析を行う手法のことである[3)]。数量化理論にはⅠ類、Ⅱ類、Ⅲ類、Ⅳ類があり、質的データから量的に測定される外的基準を予測したり説明したりするⅠ類を、本研究の評価方法として用いる。

数量化理論Ⅰ類は、複数個の定性的属性X_{ij}（説明変数）から1つの定量変数Y（目的変数）を式(2)のような重回帰式を用いて予測する方法である。

$$\begin{aligned} Y = b_0 &+ b_{11}X_{11} + b_{12}X_{12} + \cdots\cdots + b_{1p}X_{1p} \\ &+ b_{21}X_{21} + b_{22}X_{22} + \cdots\cdots + b_{2q}X_{2q} \\ &+ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &+ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &+ b_{n1}X_{n1} + b_{n2}X_{n2} + \cdots\cdots + b_{nr}X_{nr} \end{aligned} \qquad (2)$$

ここで、X_{ij}は説明変数、Yは目的変数、b_{ij}はスコアである。

上式において、スコアは対応する説明変数X_{ij}の目的変数Yに対する影響の度合いを示している。すなわち、この値の絶対値が大きければ目的変数への影響が大きくなる。

このような方法により、感性を定量的に扱うことが可能となる。

本論文では、それぞれの実験条件を以下のような説明変数に対応させる。

X_{11}：フィルタを15個程度用いて子音部を作成したかどうかを表す変数

X_{12}：フィルタを8個用いて子音部を作成したかどうかを表す変数

X_{13}：フィルタを3～4個程度用いて子音部を作成したかどうかを表す変数

X_{21}：フォルマント遷移部を含むかどうかを表す変数

X_{22}：フォルマント遷移部を含まないかどうかを表す変数

X_{31}：子音/s/を用いて子音/t/を作成したかどうかを表す変数

X_{32}：子音/s/を用いないで子音/t/を作成したかどうかを表す変数

合成した日本語「た」の数量化理論による解析結果は、フィルタの個数、フォルマントの有無、子音/s/を加工したか否かをさきに示した7つの説明変数とし、目的変数Y_1を得点の予測値とおくと、多変量解析により式(3)が得られる。

$$\begin{aligned} Y_1 = 1.805 &+ 0.088X_{11} + 0.008X_{12} - 0.192X_{13} \\ &+ 0.048X_{21} - 0.192X_{22} \\ &+ 0.032X_{31} - 0.008X_{32} \end{aligned} \qquad (3)$$

合成した日本語「ち」の数量化理論による解析結果は、多変量解析により式(4)が得られる。

$$\begin{aligned} Y_2 = 1.621 &+ 0.373X_{11} - 0.173X_{12} - 0.400X_{13} \\ &+ 0.016X_{21} - 0.064X_{22} \\ &+ 0.416X_{31} - 0.104X_{32} \end{aligned} \qquad (4)$$

合成した日本語「つ」の数量化理論による解析結果は、多変量解析により式(5)が得られる。

$$Y_3 = 1.643 + 0.277X_{11} + 0.011X_{12} - 0.576X_{13} + 0.011X_{21} - 0.043X_{22} + 0.096X_{31} - 0.024X_{32} \quad (5)$$

合成した日本語「て」の数量化理論による解析結果は、多変量解析により式(6)が得られる。

$$Y_4 = 1.877 + 0.016X_{11} + 0.003X_{12} - 0.037X_{13} + 0.040X_{21} - 0.160X_{22} \quad (\text{スコア } b_{31}, b_{32} = 0) \quad (6)$$

合成した日本語「と」の数量化理論による解析結果は、多変量解析により式(7)が得られる。

$$Y_5 = 1.392 + 0.488X_{11} - 0.179X_{12} - 0.619X_{13} + 0.104X_{21} - 0.416X_{22} + 0.309X_{31} - 0.077X_{32} \quad (7)$$

１０．合成結果に対する考察

Y_1 から Y_5 までの式において、各スコア b_{ij} の内、ｉが同じで異なるｊを持つものの最大値と最小値の差（範囲という）は目的変数に対する影響度を表している。同じｉを持つスコアどうしの比較は、その範囲における目的変数への影響度の比較である。

合成した日本語「た」については、b_{11} から b_{13} までの比較により、フィルタの個数は少なくとも 8 個必要である。次に、b_{21} と b_{22} の比較により、明瞭度を上げるためにはフォルマント遷移も必要である。また、範囲は小さいが、b_{31} と b_{32} の比較により、「さ」の子音部からの合成は、直接「た」のスペクトルより合成したものよりも有効である。

合成した日本語「ち」についてはフィルタの個数を８個以下に少なくすると明瞭度が低下する。フォルマント遷移の有無は範囲が小さいので、評価にあまり関係しない。また、「し」からも合成可能である。

合成した日本語「つ」については「ち」と同様の傾向を示している。「す」からも合成可能である。

合成した日本語「て」については、フィルタの個数については範囲が小さいが、「た」と同様の傾向を示している。明瞭度を上げるためにはフォルマント遷移が必要である。「せ」からの合成はスコアが両者０のため同等評価となる。

合成した日本語「と」についてはフィルタの個数を８個以下に減らすと明瞭度が低下する。フォルマント遷移部の存在は明瞭度に関係している。「そ」からも合成可能である。

以上、フィルタの削減はあまり良い結果を生まず、口を大きく動かして発音するものはフォルマントの遷移が必要であり、また子音/ｓ/からの合成も、スペクトルの形状が異なっていたにもかかわらず、可能である等の知見が得られた。最後の知見は「た」行音の子音の継続時間が短いために、直接スペクトルを合成するよりも間接的に合成した方がよい結果を生むという根拠を与えている。

１１．むすび

本論文では、日本語「た」行音の子音部の音声信号の解析、合成及びその評価を行った。日本語「た」行音の子音部の音声信号の解析では、

（1）継続時間、子音と母音の振幅比
（2）時間波形の形状の特徴
（3）スペクトル解析

の 3 項目について行った。

次いで、これらの解析結果に基づいて、ガウス雑音とフィルタを用いて日本語「た」行音の子音部の合成を行った。

数量化理論Ｉ類を用いた合成信号の評価では、フィルタの個数、フォルマント遷移部の有無、子音/s/を加工して子音/t/を合成したか否かと言う合成条件と明瞭度との関連性について考察を行った。

明瞭度を下げずにフィルタの個数をどこまで減らすことができるかを調べるために、数量化理論Ｉ類を用いてその評価を行ったが、詳細にわたる関係性は求められなかった。合成を考えるとき重要な問題であるので、今後の課題とするところである。

また、逆に、スペクトルの細かい凹凸がどこまで必要であるかを見出すまでには至らなかった。今後はその具体的な関連性を見出し、プログラムなどを用いて機械的な合成を可能にすることも考えたい。

日本語「た」行音の音声信号の子音部を取り除いたもの（つまりフォルマント遷移させた母音部のみの音声信号）は、日本語「ぱ」行音に近い音声信号となる音響的特徴があった。これに関しては解析までに至らなかったので、「た」行音が、「ぱ」行音とどのような関係にあるのかを解析するのも、今後の課題とするところである。

本研究の最終的な目標は、明瞭でさまざまな個性あふれる声をもって、任意の文章内容を表現するところにあるが、声のフォントがそろってからの問題と考えている。

参考文献

1) レイ・D・ケント,チャールズ・リード著,荒井隆行,菅原勉監訳：“音声の音響分析”,海文堂(1997).
2) 有馬哲,石村貞夫著：“多変量解析のはなし”,東京図書(1987).
3) 内田治著：“すぐわかる EXCEL による多変量解析”,東京図書(1996).
4) 中尾睦彦,西本宣央,八藤政和：“白色ガウス雑音を用いた子音/s/をもつ日本語音声の合成”,明石工業高等専門学校紀要第 43 号,pp.13 – 18(2000).
5) 中尾睦彦,高松一平：“子音/h/をもつ日本語音声の合成”,明石工業高等専門学校研究紀要第 45 号,pp.21 - 27(2002).
6) 中尾睦彦,坂部太志：“子音/k/をもつ日本語音声の合成”,明石工業高等専門学校研究紀要第 46 号,pp.19 - 24(2003).