白色ガウス雑音を用いた日本語音声の合成

中尾　睦彦*　稲川　千津**　福原　始***

Speech Synthesis of Japanese Language by White Gaussian Noise

Mutsuhiko NAKAO, Chizu INAGAWA, Hajime FUKUHARA

ABSTRACT

In this experiment, by using white Gaussian noise and Discrete Fourier Transform (DFT), spoken Japanese language was synthesized. Firstly, the spectrum of voice signals was calculated by DFT and multiplied to spectrum of white Gaussian noise. Secondly, through inverse DFT, synthesized voice signals was obtained. Furthermore, to improve this signal performance, several bandpath filters were used. Synthetic speech was evaluated by the human auditory sense and quantification theory I.

KEYWORD : speech synthesis, white Gaussian noise, quantification theory

１．はじめに

近年、対人自動サービスシステムが発達するにつれて、人の声による情報伝達方式が望まれている。従って、今後、明瞭で、個性的な音声を合成することに対するニーズが高まってくることが期待される。

音声の合成方法には、多くの観測事象を統計的に処理するコーパスベース手法及び音声の分析結果に基づいて、合成規則を構築することを主眼にしたルールベース手法がある。

本研究は後者に属し、音声信号のスペクトルを解析し、それに基づいて音声合成、明瞭度の向上、個性の付与等を目指すものである。現時点では、子音/k/,/s/,/h/,/t/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)～7)]。

昨年度までは、多数のフィルタを設計し音声合成を行っていたために合成に時間を要したが、本研究では、白色ガウス雑音と離散フーリエ変換（DFT）を用いて音声信号のスペクトルを周波数軸上で合成したため、短時間で音声の合成が可能になった。

ここでは、過去に合成していない、鼻音/n/,/m/、わたり音/y/,/w/ を含む日本語子音の合成を行う。これらの子音部の音声信号は、それぞれ異なった特徴を持っている[1)]。そのため、まずこれらの音声信号について、その特徴を把握した上で、合成の段階に入る事にした。

２．音声信号

日本語の音声信号は、子音部、フォルマント遷移部、母音部に分類される。子音部は振幅の小さい雑音で構成され、母音部は概周期関数的である。フォルマントとは、発声した母音音声波の周波数スペクトルのうちで、特定の周波数領域にエネルギーが集中して生じる山のことである。フォルマントは、周波数スペクトルにおいて周波数の低いほうから順に第１フォルマントＦ１、第２フォルマントＦ２等と定義される。

日本語音声の一例である日本語「ざ」についての時間波形と各部の構成図を図１に示す。

日本語の音声信号は、はっきりとは子音部・フォルマント遷移部・母音部の境界を示さないものもあるが、図１のような構成になっている事をふまえた上で、フォルマント遷移について述べる。

発話中の声道形状の変化は、声道共鳴の変化を伴い、その音響的変化は、調音変化と同じ時間を有する。

* 電気情報工学科、** 明石高専専攻科、***京都大学工学部

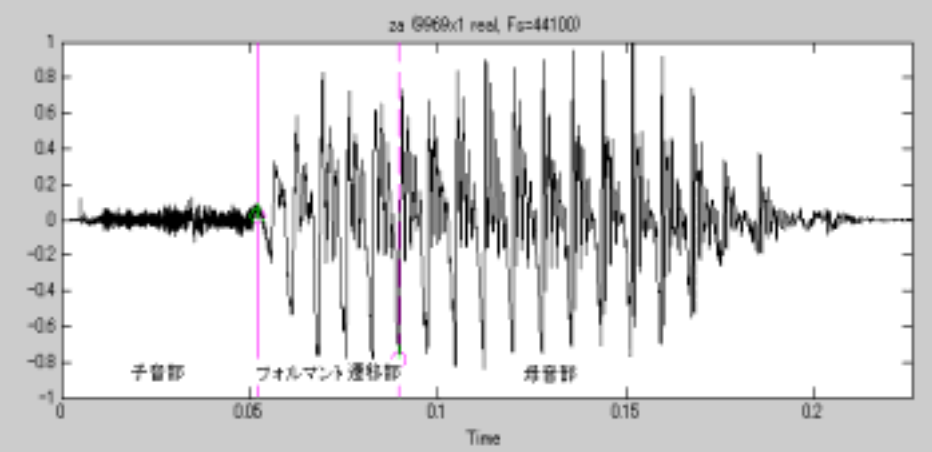


図１　ざの時間波形

その時間内において、全てのフォルマント周波数は子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（調音方法に対するキュー）とＦ２とＦ３周波数遷移（調音位置に対するキュー）との組み合わせとなる調音結合情報を含む。例えば、「ざ」という音声を合成する時には摩擦子音部/z/ 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合成音声がより自然音声に近くなる。

2･1　子音の分類

子音はそれぞれの音響的特徴にかなり違いがある [1)]。雑音がかなりあるものもあれば、ほとんどないものもある。声道を一定時間完全に阻害して生成されるものもあれば、単に声道を狭めるだけで生成されるものもある。また、音の伝達の際に口腔のみを通過する子音もあれば、音響エネルギーの伝達に鼻腔を必要とするものもある。このような違いが存在する結果、子音は音響的特徴において弁別的ないくつかのグループに分けられる。一般的なグループ分けを以下(1)～(6)に示す。

(1)　閉鎖音　　/k/,/t/,/g/,/d/,/b/,/p/
(2)　摩擦音　　/s/,/h/,/z/
(3)　破擦音　　英語では/dʒ/,/t ʃ/ （日本語では発声される事はまれである）
(4)　鼻音　　/m/,/n/
(5)　わたり音　/w/,/y/
(6)　流音　　/r/

2･2　鼻音

鼻音/m/,/n/ を子音とする日本語音声は、それぞれ「ま」行音と「な」行音である。

鼻音は、口腔を閉鎖し鼻腔から音を放射することで生成される。この場合、声道が二又に分かれるので、そのシステムのエネルギー伝達関数は零点を含む。すなわち、鼻音にはアンチフォルマントが存在する。

鼻音を生成するために口腔内のどこか１点に閉鎖が作られると、アンチフォルマントの周波数は口腔が鼻からの伝送を短絡させるような周波数になり、その周波数のエネルギーは鼻腔を通過しない。鼻音/m/,/n/ はそれぞれ 750～1250 Hz、1450～2200 Hz のアンチフォルマントによって特徴付けられるが、口腔も鼻音の共鳴特性に寄与するためフォルマントも存在する。図２に鼻音「ま」と「な」の時間波形を示す。

2･3　わたり音

わたり音/w/,/y/ を子音とする日本語音声は、それぞれ「わ」行音と「や」行音である。

わたり音は半母音とも呼ばれ、徐々に変化するフォルマント構造を持っている。その調音は、著しい狭めをもつ声道の形から次の母音の声道の形への比較的緩やかな移行として考えられる。

わたり音/w/ は、唇でつくられる狭めと、舌と軟口蓋あるいは硬口蓋でつくられる狭めによって調音されるため、その声道の形は母音 /u/と酷似する [1)]。

また、わたり音/y/ は母音 /i/とよく似た形の声道である [1)]。図３に「ば」、「わ」、「うあ」 と発話した音声信号のスペクトログラムを示す。「わ」の場合、「母音＋母音」として発話したものに比べ、周波数の遷移の仕方は同じだが、フォルマント周波数が遷移する時間が短い。そして、「わ」よりも「ば」と発話したものの方が、フォルマント周波数が遷移する時間がより短い。つまり、これらの音声信号の区別は、遷移する時間の長さであることがわかる。

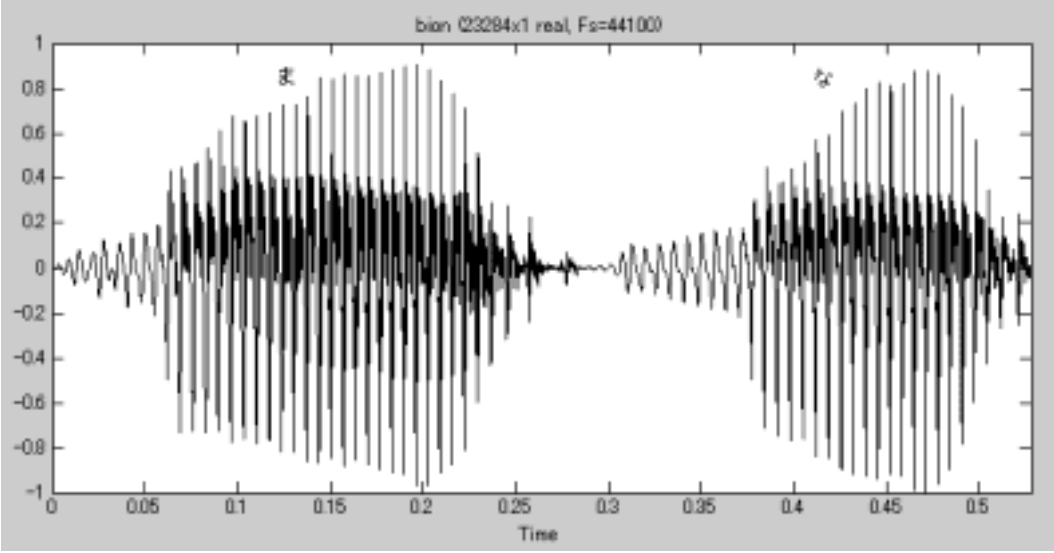
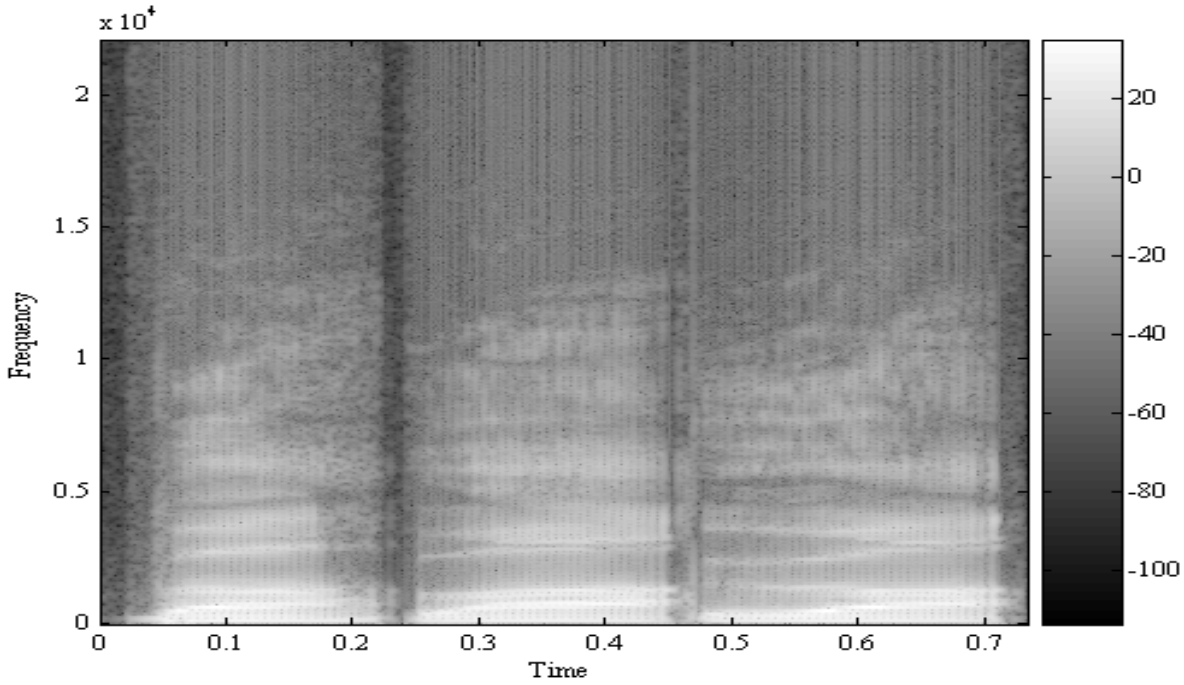


図2　鼻子音をもつ日本語音声信号の時間波形



ば　　わ　　うぁ

図3　日本語音声「ば」「わ」「うあ」のスペクトログラム

3．　線形システムの時間軸と周波数軸における入出力の対応

インパルス応答 ***h[n]*** をもつ線形システムに入力される信号 ***x[n]*** と出力信号 ***y[n]*** は図４に示すように、時間軸から見るとたたみ込みの関係であるが、それぞれを DFT（離散フーリエ変換）し周波数軸から見ると、単純な積の関係である。

データ点数を制限し DFT ではなく FFT（高速フーリエ変換）を用いても、たたみ込みを行うより余分な計算をするように考えられるが、周波数領域に変換して積を計算し再び時間領域へと変換する方が、たたみ込みを計算して ***y[n]*** を求めるより短時間で計算できることになる。

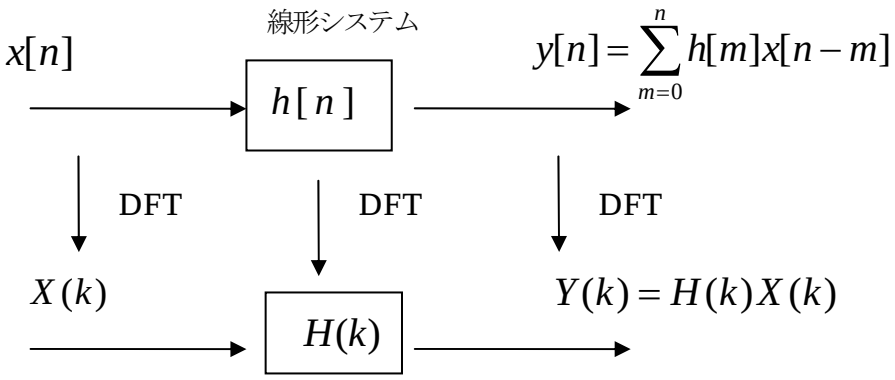


図４　線形システムとそのＤＦＴとの関係

４．音声信号の合成

白色ガウス雑音を用いた日本語子音の合成について述べる。

４･１　白色ガウス雑音を用いた音声信号の合成方法

日本語子音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すなわち白色ガウス雑音を用いる。

この雑音を用いるのは、発生が容易でスペクトルが周波数に対して一定であるためスペクトル形成がしやすいためである。

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、正規分布を表す

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

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、したがって白色ガウス雑音となる。

本論文における音声の合成は、/a/を母音として持つ日本語のみを対象とし、以下の手順で行う。

(1) サンプル音声信号を子音部、フォルマント遷移部、母音部に分割する。

(2) それぞれの離散フーリエ変換を求め、その絶対値を計算し、サンプル音声信号の振幅スペクトルを求める。

(3) 白色ガウス雑音を離散フーリエ変換し、サンプル音声信号の振幅スペクトルと同じ周波数成分同士を掛ける。

(4) (3) の計算結果を逆離散フーリエ変換する。(1)～(3)をブロック線図で示すと図５のようになる。

(5) 帯域フィルタを設計しそれを適用して、特定の周波数の振幅成分の強調または除去を行う。この手順を加工と呼び、加工は子音部のみに行うものとする。1 種類の合成音声信号から、加工方法によって異なる 4 種類の音声信号を作成する。従ってサンプル音声信号は計 5 種類となる。加工の方法は表 1 に示す 4 つとする。

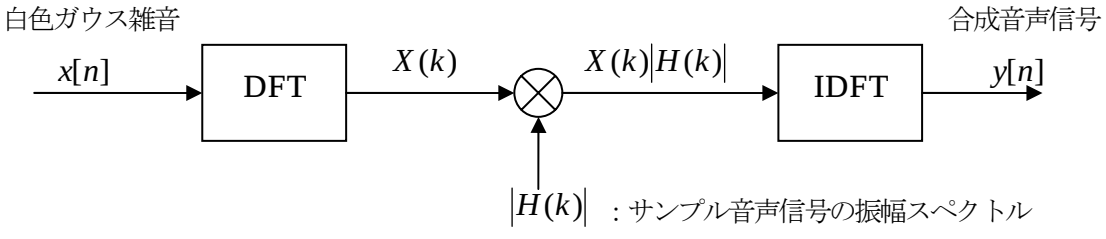


図５　音声合成手順のブロック線図

表 1　音声信号の加工方法

加工方法	内容
1	0〜3000 Hz の振幅成分を強調
2	200〜3000 Hz の振幅成分を強調
3	200 Hz 以下の振幅成分を除去
4	7000〜11000 Hz の振幅成分を強調

加工方法 1 は、わたり音などの比較的低周波の振幅成分を多く含む子音の明瞭度の向上を目的とする。

加工方法 2 は、100〜200 Hz の間に存在する声帯の振動の周波数（基本周波数）を強調せず、比較的低周波の振幅成分を強調することによって子音の明瞭度を向上させることを目的とする。

加工方法 3 は、基本周波数の振幅成分を除去することによって子音の明瞭度を向上させることを目的とする。

加工方法 4 は、摩擦音などの比較的高周波の振幅成分を多く含む子音の明瞭度の向上を目的とする。

また、合成した音声信号は、音声文字の後ろに加工方法を表す番号を付与することで区別する。

以上の手順で子音部、フォルマント遷移部及び母音部をそれぞれ作成し、最後に結合して合成音声とする。フォルマント遷移部は、フォルマント周波数の遷移を実現するために、この部分を時間的に２〜５段階に分割し合成する。

５．　合成音声信号の評価

４．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。評価の手法を**５・１**及び**５・２**で述べ、**５・３及び５・４**では合成した音声信号の評価結果について述べる。

５･１　人間の聴覚による合成音声の評価

評価方法は、合成音声を聴いてもらい、目的の日本語に明瞭に聞こえる場合を 5 点とし、全く聞こえない場合を 1 点、というように 5 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とる。

５･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。

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

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に対応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。表２に加工方法と説明変数の対応を示す。

表２　加工方法と説明変数の対応

	説明変数			
加工方法	X_{11}	X_{12}	X_{13}	X_{21}
（未加工）	0	0	0	0
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

説明変数の値は、特定の加工方法を使用する場合は１、使用しない場合は 0 となるため、加工を施さない未加工の音声信号の説明変数はすべて 0 である。

加工方法の４は周波数の高い成分に対する加工であるので２１という添え字を付けて区別している。

5･3　合成した日本語「ま」の評価

合成音声「ま0」～「ま4」を評価した７９人のアンケート結果を表３に示す。

表３のデータに基づいて、表１で示した 4 つの加工方法を**5・2**で示した 4 つの説明変数とし、目的変数 Y_{11} を「ま」の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。

$$Y_{11} = 2.91 + 0.37X_{11} + 0.20X_{12} + 0.23X_{13} + 0.05X_{12} \quad (3)$$

スコアの値（式（３）の説明変数の係数）を表４及び図６に示す。このスコアは、対応する説明変数 X_{ij} が目的変数 Y_{11} 値に対してどの程度影響しているかを示している。

5･3・1　合成した日本語「ま」について

(1) 式（３）より、合成した「ま」の中で加工方法１(0～3000 Hz の振幅成分を強調)を用いた「ま１」の予測値が最も高くなり、その値は Y_{11} ＝３．２８である。これは「ま1」を 57%の人が明瞭に聞こえると判定したことに対応している。

(2) 図６より、b_{11} が最も大きく、b_{12}, b_{13} は b_{11} に比べ小さいことから、「だ」「ば」及び「な」と同じく「ま」の子音部 /m/ の 3 kHz 以下の振幅成分の強調が「ま」の明瞭化に適した方法であり、明瞭な「ま」の合成には３ｋＨｚ以下の振幅成分の相対的な大きさを考慮することが必要であると考えられる。

(3) 図６より、$b_{21} > 0$ であるから、7kHz～11 kHz の振幅成分の強調は、明瞭な「ま」の合成に有効であると考えられるが、b_{11}, b_{12}, b_{13} よりも値が小さいため、これらに対応する加工方法よりも効果が小さいといえる。

表４　合成音声「ま」のスコア

スコア	スコアの値
b_{11}	0.37
b_{12}	0.20
b_{13}	0.23
b_{21}	0.05

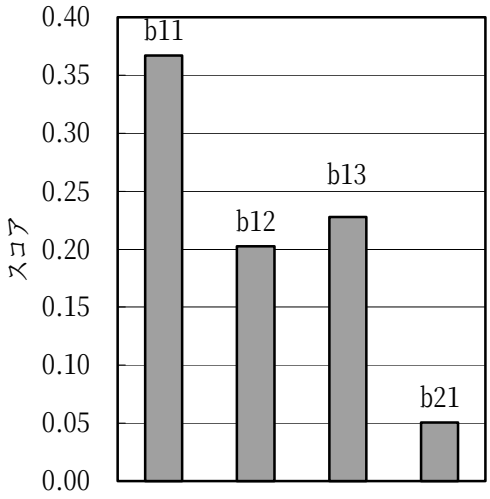


図６　合成音声「ま」のスコアの比較

5･4　合成した日本語「わ」の評価

合成音声「わ0」～「わ4」を評価したアンケート結果を表５に示す。

表５を元に、加工方法を**5・2**で示した 4 つの説明変数とし、目的変数 Y_{13} を「わ」の得点の予測値とおくと、多変量解析により式（４）が得られる。

表３　合成音声「ま0」～「ま4」を評価したアンケート結果

	人数 ［人］					
音声信号	5 点	4 点	3 点	2 点	1 点	平均得点
ま 0	7	14	26	29	3	2.91
ま 1	13	23	22	15	6	3.28
ま 2	11	20	25	13	10	3.11
ま 3	14	22	16	15	12	3.14
ま 4	9	18	23	19	10	2.96

表5　合成音声「わ0」～「わ4」を評価したアンケート結果

	人数 ［人］					
音声信号	5 点	4 点	3 点	2 点	1 点	平均得点
わ0	7	20	30	17	5	3.09
わ1	11	24	21	15	8	3.19
わ2	3	16	24	21	15	2.63
わ3	17	12	21	14	15	3.03
わ4	12	14	21	14	18	2.85

$$Y_{13} = 3.09 + 0.10X_{11} + 0.46X_{12} + 0.06X_{12} + 0.24X_{21} \quad (4)$$

スコアの値（式（4）の説明変数の係数）を表6及び図7に示す。このスコアは、対応する説明変数X_{ij}が目的変数Y_{13}に対してどの程度影響しているかを示している。

表6　合成音声「わ」のスコア

スコア	スコアの値
b_{11}	0.10
b_{12}	−0.46
b_{13}	−0.06
b_{21}	−0.24

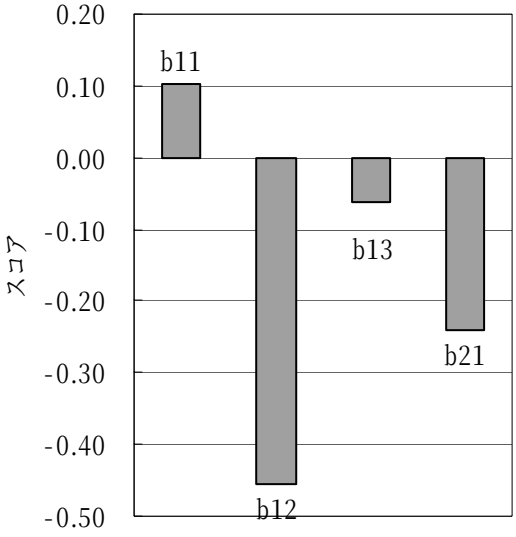


図7　合成音声「わ」のスコアの比較

(1) 式（3）より、合成した「ま」の中で加工方法1を用いた「ま1」の予測値が最も高くなり、その値はY_{11} ＝３．２８である。これは「ま1」を57%の人が明瞭に聞こえると判定したことに対応している。

(2) 図6より、b_{11}が最も大きく、b_{12}, b_{13}はb_{11}に比べ小さいことから、「だ」「ば」及び「な」と同じく「ま」の子音部/m/の3 kHz以下の振幅成分の強調が「ま」の明瞭化に適した方法であり、明瞭な「ま」の合成には3 kHz以下の振幅成分の相対的な大きさを考慮することが必要であると考えられる。

(3) 図6より、$b_{21} > 0$ であるから、7 kHz～11kHzの振幅成分の強調は、明瞭な「ま」の合成に有効であると考えられるが、b_{11}, b_{12}, b_{13}よりも値が小さいため、これらに対応する加工方法よりも効果が小さいといえる。

5･4･1　合成した日本語「わ」について

(1) 式（4）より、合成した「わ」の中で加工方法1を用いた「わ1」の予測値が最も高くなり、その値はY_{13} ＝３．１９である。これは「わ1」を55%の人が明瞭に聞こえると判定したことに対応している。

(2) 図7より、b_{11}が正の値で最も大きく、b_{12}, b_{13}が負の値であることから、「だ」「ば」及び「や」と同じく「わ」の子音部/w/の3kHz以下の振幅成分の強調が「わ」の明瞭化に適した方法であり、明瞭な「わ」の合成には3kHz以下の振幅成分の相対的な大きさを考慮することが必要であると思われる。

(3) 図7より、$b_{21} < 0$ であるから、7 kHz～11 kHzの振幅成分の強調は、その強調の度合いあるいは周波数帯域が明瞭な「わ」の合成には適当でないと考えられる。

6．結び

本論文では、音声信号を合成するための方法として、従来用いていた音声信号のスペクトル成分をフィルタを用いて白色雑音から抜き出し、振幅を調整して合成するという方法を採用せず、音声信号の振幅スペクトルを直接白色雑音に乗ずることによりスペクトルを合成し、これをフーリエ逆変換して合成するという方法を採用した。

この方法により、短時間で正確なスペクトルをもつ音声信号を合成することが可能となった。本論文ではこの方法により、これまで合成していなかった鼻音とわたり音の合成を試みた。得られた信号に種々の加工を加え、その結果を評価したところ加工による聞き取り性能の向上が認められた。他に流音についても合成を試みており、良好な結果が得られているが、紙幅の都合で割愛した。

今回の報告により、白色ガウス雑音をベースにして日本語の単音レベルの合成が可能なことが判明した。今後は更なる明瞭度の向上と単語レベルの合成に取り掛かる予定である。

基本的なスペクトルと加工知見が蓄積されれば、明瞭さをもつ音声の合成や性質の異なる音声への変換が可能になると考えられる。

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).

7） 中尾睦彦,岸本昌也,濱田憲治：“子音/t/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 48 号,pp.11‐18,(2005).