東京外国語大学『語学研究所論集』第 18 号（2013.3）, 1-18

〈論　文〉

言語音声の聴知覚研究のためのツール構築

佐藤 大和，益子 幸江

1. はじめに

言語音声を音声学的に記述するためには，内省に基づく考察ばかりでなく，音声スペクトルや波形の基本周波数など，音声の音響的諸特性を表示するツールによる分析が必要である．しかしながら，精密な音響分析が可能になったとしても，それだけで言葉としての音声の特徴を明らかにすることができる訳ではない．音響的諸特性の言語音声に果たす役割，すなわち音韻や韻律の知覚に及ぼす影響や役割を検討してはじめて言葉と音響特性との関連を明らかにすることができる．

上記のような研究目的のため，音声分析の他に音声を再合成する機能を有するソフトウェア・ツールとして，近年 praat が広く利用されるようになった [1]．筆者らは，アクセントや声調，イントネーションなど，音声の超分節的特性の研究に資することを目的として，音声の基本周波数パタンを変形して，その音声を再合成できるツールの作成を進めた．このツールは，一般に広く利用されている表計算ソフト：エクセルを援用するなど，知覚実験に使用する音声刺激の作成が，誰にでも簡単にできるよう配慮したものとなっている．このツールを利用すれば，例えば日本語のようなアクセント言語におけるアクセント知覚や，各種声調言語における声調知覚などの聴覚実験に容易に適用することが可能となる．本論は，このようなソフトウェア・ツール（以後，韻律制御エディター(SpitEditor)と呼ぶ）に関して，その内容を述べたものである．

2. 韻律制御エディター(SpitEditor)の概要

韻律制御エディターは，音声波形の基本周波数（ここではピッチ周波数と呼ぶ）を抽出し，その周波数や時間長情報を変更したデータに基づいて，新たな音声を合成し，言語音声の超分節的特徴に及ぼすピッチ周波数や時間情報の役割を検討するためのツールである．その機能と処理の概要は，以下のとおりである．

（1）音声波形の表示
（2）音声波形のピッチマークの自動設定と手動修正
（3）声音部，無声音部，無音部等のラベリング

（４）ピッチ周波数データ等のエクスポート
（５）エクセル上でのピッチデータ等の編集
（６）編集して変更されたピッチデータのインポート
（７）編集ピッチデータに基づく音声合成
（８）音声合成データの確認と保存

処理の概要を図１に示す．以下，これらの内容と使用方法について詳述する．

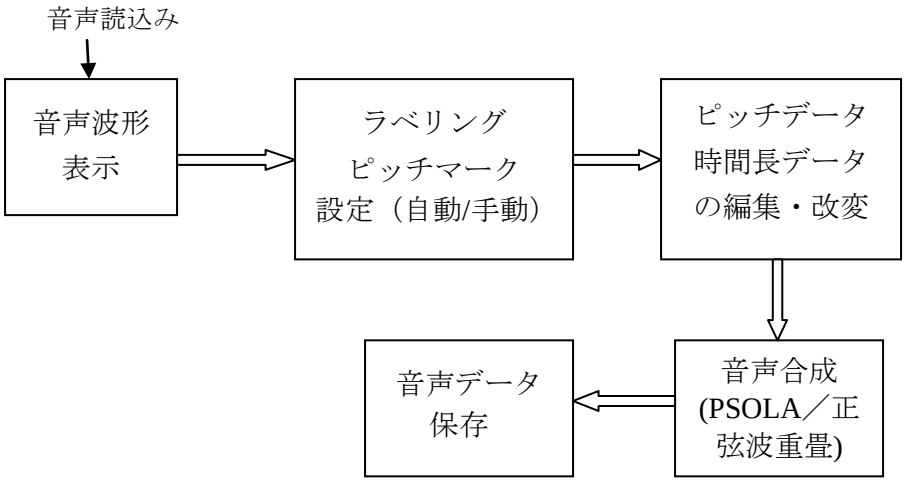

図１　SpitEditor の処理内容の概要

3. 手法の内容

3.1. 音声波形のゼロ交差点の検出と基本周波数

ピッチ周波数の抽出には，（１）音声波形の周期性を相関関数から求める方法（自己相関法），（２）声道特性の影響を除去した音源波形を用いて相関法によって求める方法（変形相関法），（３）波形のピークなど特定の時点を定めて波形周期を求める方法，などがある[2]．

本ツールにおいては，（３）の音声波形から直接基本周期を求める手法を採用している．これは，その後の音声合成において，周期を定める時間位置の情報（以後これをピッチマークと呼ぶ）が必要となるためである．また，ピッチ周波数は厳密な周期性を示

すものではなく，“ゆらぎ”や“欠落”などのため，相関法などの平均処理による求め方ではエラーの生じることがしばしばある．聴覚実験に使用するなど厳密性が求められる場合には，声帯の excitation の位置をあらかじめきちんと定めておかなくてはならない．

それでは音声波形のどの部分に着目して周期を求めるのがよいであろうか？波形の大きなピーク位置に着目するのもひとつの方法であるが，そのピークの位置は高調波成分の影響により微妙にゆらぐ可能性がある．そのため，ここでは波形の零交差位置に着目し，波形の値が負から正に変る時点，あるいは逆に正から負に変る時点を抽出する（default は，負→正への零交差時点としている）．この零交差時点の検出により，より精度よくピッチ周期の抽出ができると期待される．音声波形の最初のピッチマークが検出されると，自己相関法により次のピッチマークの位置を予測し，その近傍で零交差位置を検出する．これを繰り返し実施することにより，連続的にピッチマークが設定される．一定周期（フレーム）毎に平均のピッチ周波数を求めたい場合は，一定の窓幅におけるピッチマーク周期の平均値から求める．

図２に自動抽出されたピッチマークの例を示す．

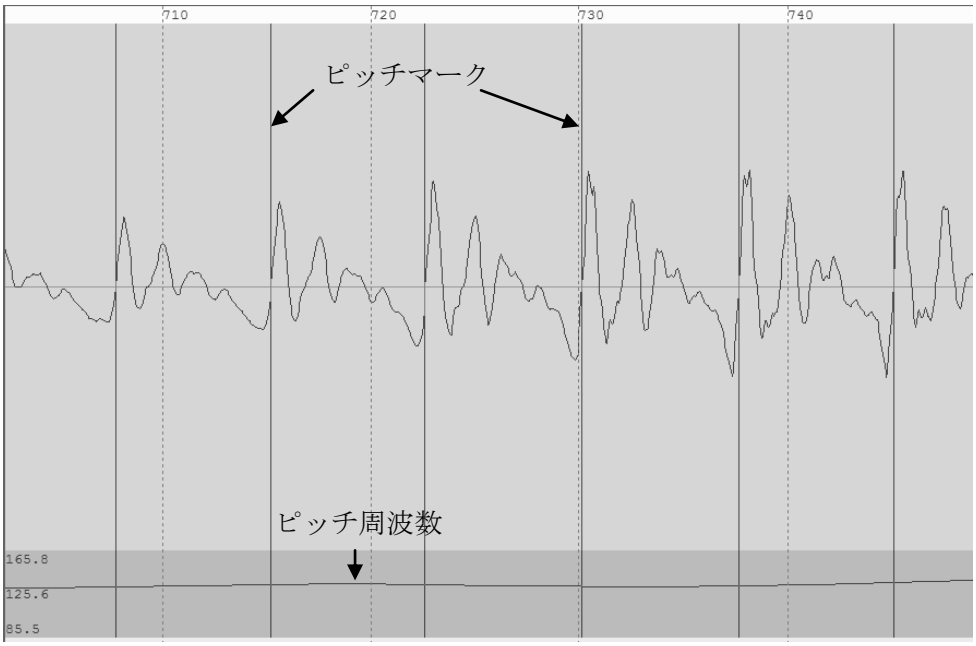

図２　ピッチマーク自動設定の例

3.2. 音声合成の手法

いったん求めたピッチ周波数を変更し（変更の手法は後に述べる），変更されたピッチ周波数で音声を再合成する方法は，以下の二通りの方法が選択できるようになっている．

3.2.1. PSOLA 方式

PSOLA とは，Pitch Synchronous Overlap and Add（ピッチ同期重畳加算方式）の略である．それまでピッチ周波数を可変にする音声合成は，声帯振動の音源と声道のスペクトル特性を分離した“音源－声道モデル”によってなされてきたが，E. Moulines らによって波形を直接処理することによって合成を可能にした手法である [3,4].

PSOLA では，ピッチマークを中心として窓関数 w(n)（2 ピッチ分の時間長（N サンプル)）で切り出し，ピッチマーク毎に得られた音声波形要素を，順次新たなピッチ周期に合わせてずらしつつ重畳して音声を再合成する手法である．窓関数としては，以下のハニング窓が使われる．

$$W(n) = 0.5 - 0.5\cos(2\pi n/(N-1)) \quad , \quad (n=0 \sim N-1)$$

ピッチ周波数が原音声と大きく異ならない限り，原音声の音質に近い合成音が得られる．

3.2.2. 正弦波重畳方式

正弦波重畳方式は，原音声の編集波形を正弦波の重ね合わせによって実現する手法である．まず原音声から順次 1 ピッチ波形を切り出し，これらを FFT（Fast Fourier Transform）にて周波数領域に変換するとともに，線形予測分析によってスペクトル包絡を求める．フーリエ変換した音声の基本周波数成分およびその倍音成分を用いて，目的とする新たな基本周波数成分に基づき，標本化周波数の 1/2 までの周波数成分を正弦波として表現し，それらを重畳して目的音声を合成する．波形のピッチ周期の境界では，位相が連続するように接続される．

この方式は，音声の調波構造を再現するためクリアな音声が得られるが，PSOLA と比べて音質は若干異なる．実験目的に応じて両者を使い分けるのが望ましい．以後この方式を SWS (Sinusoidal Wave Superposition)方式と呼ぶことにする．

図３に PSOLA 方式と正弦波重畳方式(SWS)の処理の概要を示した．

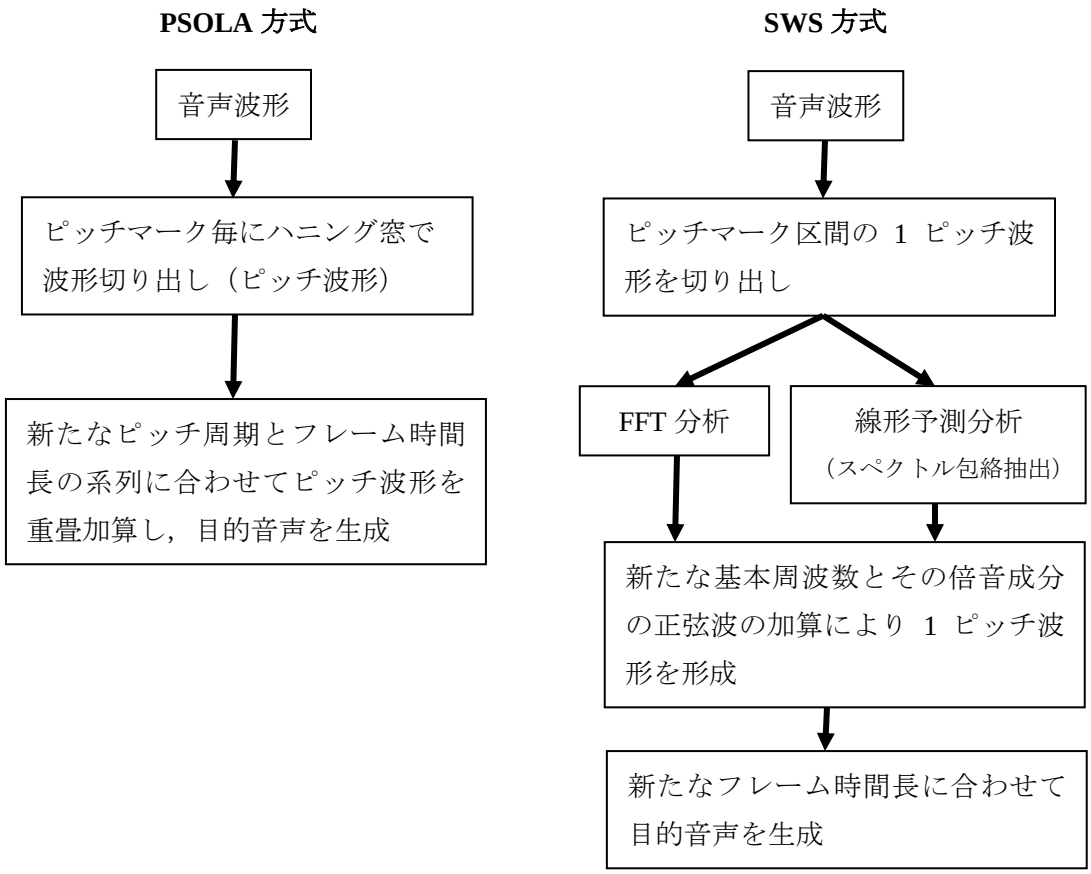

図３　音声合成の 2 方法（PSOLA 方式と SWS 方式）の処理の概要

4. SpitEditor の使用方法

4.1. 音声波形の表示

プログラム SpitEditor.exe を起動し,「ファイル」メニューから「音声ファイルを開く」を選択すると，図４のような音声波形が表示された画面になる．画面は，(時間表示領域), (音声波形領域), (ピッチ周波数表示部), (ラベリング部１), (同２), および(SWS, PSOLA 両合成合成のためのピッチ周波数と時間情報）の領域から成っている．

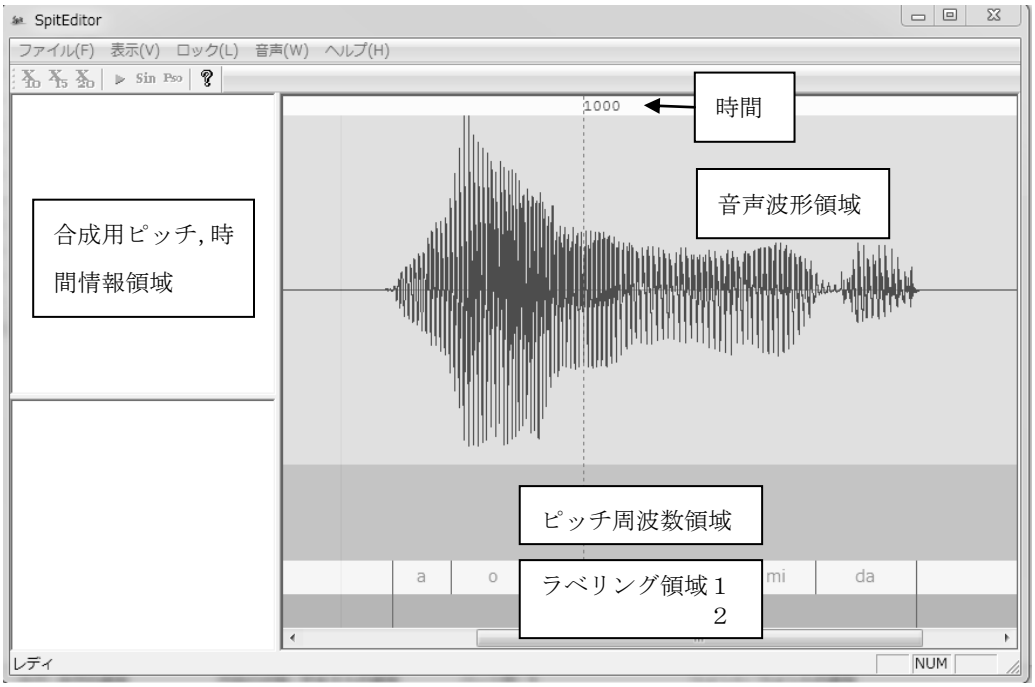

図 4　SpitEditor の最初の画面例

音声波形の水平方向（時間軸），垂直方向（振幅軸）の拡大・縮小は，シフトキーと矢印キーを使って，以下のように行う.

水平方向（時間軸）拡大：Shift ＋ → ，　縮小：Shift ＋ ←

垂直方向（振幅）　拡大：Shift ＋ ↑ ，　縮小：Shift ＋ ↓

4.2. ピッチマークの自動設定

波形上にピッチマークを自動設定するために，まず自動設定する音声範囲を定める. 音声波形の始端部に合わせてマウスを左クリックし, そのまま右にドラッグすると黄色い領域として音声範囲が定まる（長い音声の場合，あるいは波形先頭を精密に定めたい場合などは，時間軸を伸長（拡大）して波形の始端を定めて先頭部分を選択したあと，逆に時間軸を縮小して音声全体を表示する. 先頭部の選択された黄色い区分の右端をマウスの左ボタンでドラッグして音声区間の終わりまで含めると, 音声区間全体範囲を定

めることができる.）

次に，音声波形全体を選択した状態で,「音声」メニューから「ピッチマークの自動設定」を選択すると，ゼロ交差位置にピッチマークが表示される．ゼロ交差位置は，波形の（負→正）に変わるゼロ交差を default とするが,「音声」メニューの「ゼロクロス設定」で,（正→負），あるいはその（両方）のゼロクロス位置を選択することが可能である．

ピッチマークの自動設定では,その探索のため男声や女声などによってピッチ探索の範囲をあらかじめ定めておくと抽出精度がよくなる．「音声」メニューの中の「F0 推定範囲設定」を選択すると，図５のような条件設定ウインドウが表示され，これを使ってピッチ周波数の上限値と下限値を設定する．

図５　ピッチ周波数推定範囲設定のウインドウ

ピッチマークの自動設定を行うと,「音声波形領域」に縦棒（赤）でピッチマークが表示される．また，このピッチマークに基づいて,「ピッチ周波数領域」にピッチ周波数パタンが表示される．

また，ピッチマークの設定に伴い，画面の左部分の「合成用ピッチ，時間情報領域」にも自動的に情報が書き込まれる．「合成用ピッチ，時間情報領域」は上下二つに区分されており，上部は SWS，下部は PSOLA の合成に関わる合成時間長とそのピッチ周波数が示されている．この情報によって，そのとき表示されているピッチマークに基づいた合成による音声を聞くことができる.画面上部にある３つのボタン(▶),(Sin),(Pso)

をクリックすると，（▶）は原音声，（Sin）は SWS 方式合成音，（Pso） は PSOLA 方式合成音が再生される．二つの合成音は，「音声」メニューの「音声合成」で合成再生することも可能である．

図 6 にピッチマークの表示された SpitEditor の画面例を示す．音声波形と重ねて表示されている縦棒の列がピッチマークである．その下にピッチ周波数パタンが表示されている．

4.3. ピッチマークの修正

ピッチマークの自動設定は，必ずしもいつも正しい位置に設定されるとは限らない．ミスピッチ抽出が生ずることはしばしばある．そのため，いったん抽出されたピッチマークの位置を修正したり，追加／削除することが可能となっている．

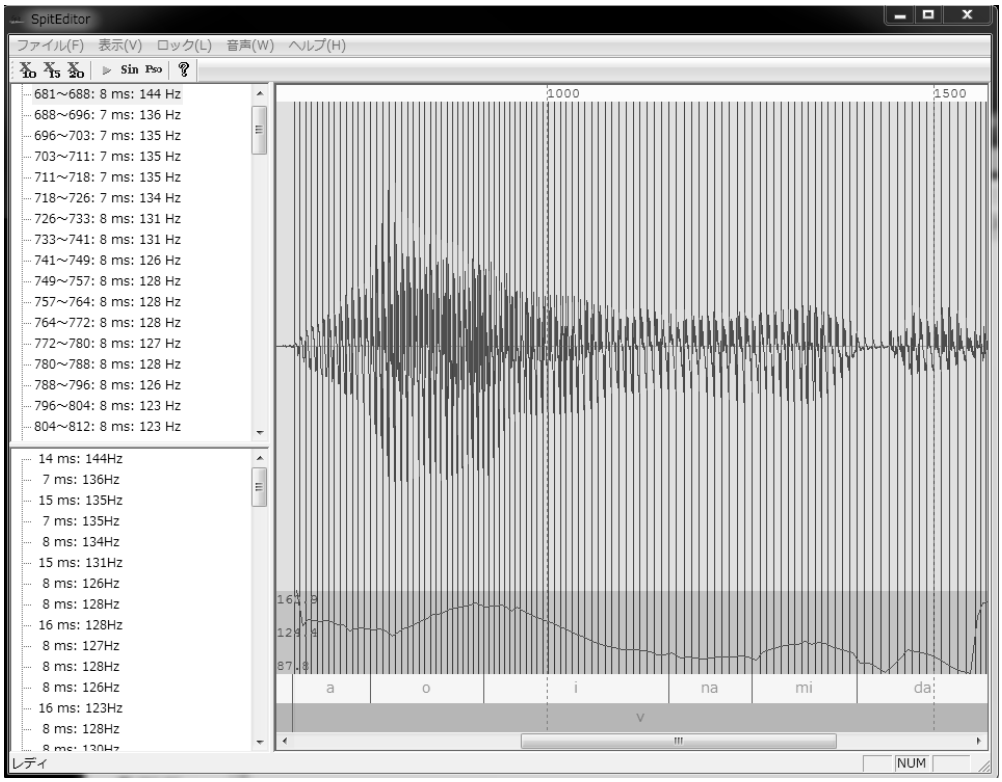

図 6　ピッチマークの表示された SpitEditor の画面例

ピッチマークの修正等は，以下のようにして行う．

（1）ピッチマークの修正

マウスの左ボタンでピッチマークを選択し，ドラッグで動かすことができる．但し，隣接するピッチマークを超えて動かすことはできない．ドラッグして隣のピッチマークと重ねると，そのピッチマークは削除されることになる．

ピッチマークをドラッグするとき，Shift キーを押して行うと，波形のゼロ交差位置に合うように位置を移動することができる．

（2）ピッチマークの追加

（Ctrl ＋ 左クリック）で新たなピッチマークを追加することができる．

このとき, Shift キーも同時に押していると, つまり（Ctrl＋Shift＋左クリック）のとき，クリックした位置に近いゼロ交差位置にピッチマークを追加できる．

（3）ピッチマークの削除

ピッチマーク上で（Ctrl ＋ 右クリック）するとそのピッチマークを削除できる．

（4）ピッチマーク全体の削除

音声波形の範囲を選択した状態で,「音声」メニューから「ピッチマークの削除」を選択すると，範囲内のすべてのピッチマークが削除される．

発話末におけるピッチマーク修正の例を図 7 に示す.

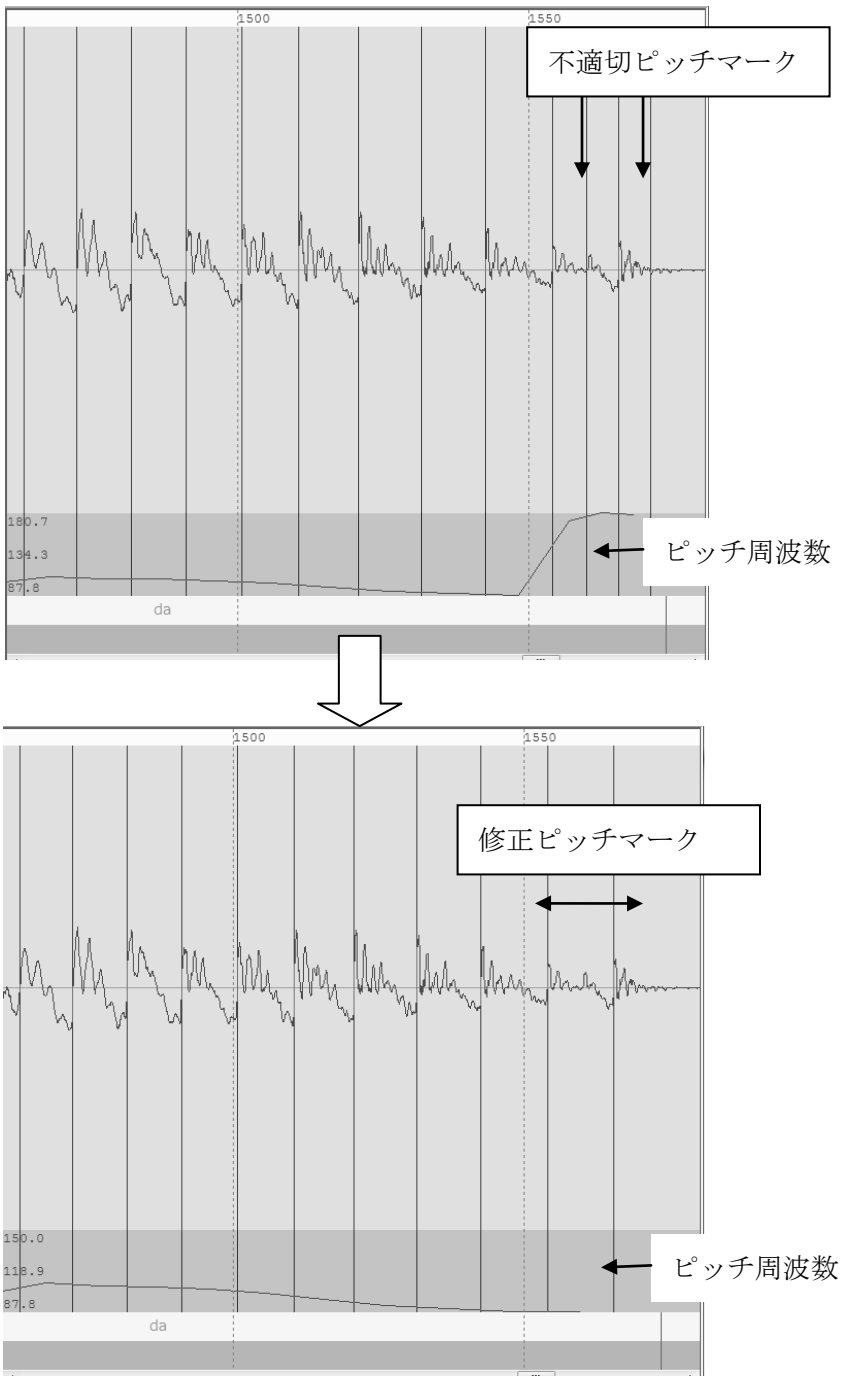

図７　ピッチマーク修正の例

なお,「音声」メニューから「ピッチマークのロック」を選択すると，ピッチマークの変更・追加・削除が行えなくなる．ピッチマークの位置が確定し，誤って変更してしまうことを避ける場合に使用する．その解除も同様にして行う．

4.4. ラベリング

音声情報のラベリング領域はピッチ周波数パタンの表示部の下にあり，（領域１）と（領域２）の 2 層のラベリング層からなっている．

（１）ラベリング境界線の設定

各層とも，ラベリングの区分（領域）を定めるラベリング境界線は，ピッチマークのときと同じように（Ctrl ＋ 左クリック）で挿入することができる．このとき，Shift キーを同時に押してドラッグする（Ctrl＋Shift＋左クリック）の場合は，波形のゼロ交差点に合うように境界線を設定することができる．

（Ctrl ＋ 右クリック）で境界線は削除される．

（２）ラベリング領域１（音素等ラベル）

境界線が定まると，境界線に囲まれた区間にラベリングをすることができる．領域１には，音素，音節など任意のラベリングをキーボードから入力できる．

（３）ラベリング境界２（セグメントラベル）

領域２は，音声セグメントの素性を入力する領域である．次の２素性は必須項目であり，これは必ずその記号でラベリングしなければならない．この 2 記号は，あとの音声合成の際に利用するからである．

V ：Voiced を表す．ピッチマークが付き，音声合成の対象となる区間.

Si：無音区間を表す.

後で示すように，この２種類の区分の音声は持続時間長を変えることができる．

無声摩擦音などその他のセグメント記号は，UV など任意で構わないが，上記 2 記号以外のラベリング区間は, 合成に際して原音声の波形がそのまま使用されることになる．

ラベリングの例を図８に示す.

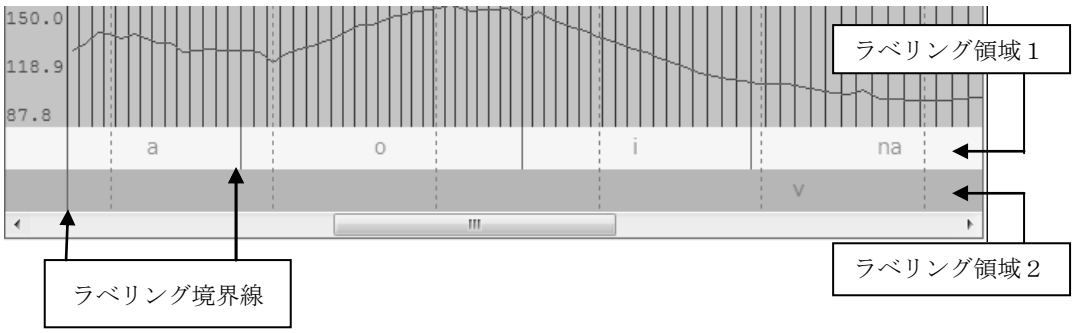

図８　ラベリング境界とラベリングの例（音声「青い波だ」）

4.5. 韻律データ編集と音声合成

ピッチマークの設定された音声データに基づき, そのピッチ周波数パタンや音声セグメントの持続時間を変更して合成音を作成することができる. エクセル表を用いた変更の手順を以下に示す.

ピッチ周波数や持続時間の韻律データの変更は, 一定時間長のフレーム周期毎に音声を区分化したデータに基づいてなされる. 画面の左上部にある（X10),（X15),（X20）の３つのボタンのひとつをクリックすると，それぞれ 10 ms，15 ms，20 ms の 3 種類のフレーム周期に相当する音声データの系列がエクセル表として表示される. フレーム周期 10 ms の場合の例を，図 9 に示した.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Time(ms)	Power	StartPos	EndPos	Phoneme	Segment	Duration(m	Pitch(Hz)		Original dui	Original pitch(Hz)	
2	0	1588.38	14838	15181	a	v	15.6	167.18		15.6	128.6	
3	10	2214.43	15181	15343	a	v	7.3	176.93		7.3	136.1	
4	20	2607.54	15343	15506	a	v	7.4	175.89		7.4	135.3	
5	30	4316.05	15506	15834	a	v	14.9	174.85		14.9	134.5	
6	40	3553.24	15834	15999	a	v	7.5	173.68		7.5	133.6	
7	50	3926.06	15999	16167	a	v	7.6	170.69		7.6	131.3	
8	60	6622.98	16167	16510	a	v	15.6	167.18		15.6	128.6	
9	70	5180.72	16510	16683	a	v	7.8	165.75		7.8	127.5	
10	80	5513.19	16683	16855	a	v	7.8	166.66		7.8	128.2	
11	90	9387.72	16855	17202	a	v	15.7	165.23		15.7	127.1	
12	100	8015.18	17202	17375	o	v	7.8	165.75		7.8	127.5	
13	110	9342.4	17375	17550	o	v	7.9	163.8		7.9	126	
14	120	10480.9	17550	17732	o	v	8.3	157.56		8.3	121.2	
15	130	14421.1	17732	18080	o	v	15.8	164.71		15.8	126.7	
16	140	9214.53	18080	18250	o	v	7.7	168.61		7.7	129.7	
17	150	9049.92	18250	18417	o	v	7.6	171.6		7.6	132	
18	160	12422.5	18417	18742	o	v	14.7	176.41		14.7	135.7	
19	170	8358.51	18742	18899	o	v	7.1	182.52		7.1	140.4	
20	180	9387.21	18899	19056	o	v	7.1	182.52		7.1	140.4	
21	190	12493.1	19056	19363	o	v	13.9	186.68		13.9	143.6	
22	200	8523.44	19363	19514	o	v	6.8	189.8		6.8	146	
23	210	12151.7	19514	19813	o	v	13.6	191.75		13.6	147.5	
24	220	7761.22	19813	19961	o	v	6.7	193.7		6.7	149	
25	230	11498.2	19961	20256	o	v	13.4	194.35		13.4	149.5	
26	240	8161.82	20256	20406	o	v	6.8	191.1		6.8	147	
27	250	10214	20406	20704	o	v	13.5	192.4		13.5	148	
28	260	6029.55	20704	20852	o	v	6.7	193.7		6.7	149	
29	270	7929.15	20852	21156	i	v	13.8	188.63		13.8	145.1	
30	280	4911.13	21156	21306	i	v	6.8	191.1		6.8	147	
31	290	4783.28	21306	21460	i	v	7	186.16		7	143.2	
32	300	6639.57	21460	21775	i	v	14.3	182		14.3	140	
33	310	4710.03	21775	21936	i	v	7.3	178.1		7.3	137	
34	320	6796.59	21936	22266	i	v	15	173.68		15	133.6	
35	330	4931.49	22266	22435	i	v	7.7	169.65		7.7	130.5	
36	340	4989.33	22435	22607	i	v	7.8	166.66		7.8	128.2	
37	350	5131.45	22607	22782	i	v	7.9	163.8		7.9	126	
38	360	7320.95	22782	23143	i	v	16.4	158.86		16.4	122.2	
39	370	5043.77	23143	23329	i	v	8.4	154.05		8.4	118.5	
40	380	4766.2	23329	23520	i	v	8.7	150.02		8.7	115.4	

SPT11D7.tmp

図９　エクセル表による音声パラメータの表示と変更

（上記は，「ファイル」メニューから，「エクセルでピッチ情報を編集（10ms)」「　同（15ms)」「　同（20ms)」を選択することでも可能である．）

表中，A 列は（フレーム単位の時間)，B 列は（そのフレームのパワー)，C，D 列はそれぞれ（音声情報の始点と終点のサンプル点)，E と F 列は（ラベリング情報）を表

す．またGとH列，およびJとK列は，それぞれその（フレームの合成時の時間長と平均ピッチ周波数）を表すが，前2列は合成用データとして編集可能であるが，後２列は原音声のオリジナルな時間とピッチ周波数であって変更できない. 最初にエクセル表を開いたときには，編集部分とオリジナル部分は全く同じデータが入っている．図９では，編集部分のピッチ周波数のセルの値を，オリジナル部の値の1.3倍とし，高い値に設定した値が示されている．

このエクセル表を閉じて保存すると，画面の合成ボタンによって，PSOLA と SWS の2種類の合成音を聞くことができる．また，編集されたエクセルデータは，エクスポートして保存し，必要に応じてインポートして使用することができる．

SWS合成では，FFTと線形予測（LPC）分析のときのパラメータを，「音声」メニューの「合成パラメータの設定」で変更することができる．音声の帯域，標本化周波数などに合わせて調整する（図１０）．

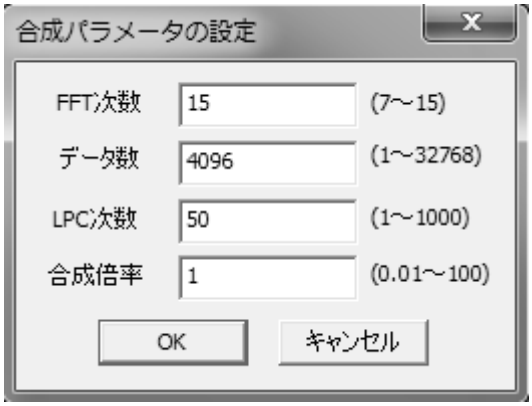

図１０　SWS合成におけるFFTとLPC分析のパラメータ設定

また，PSOLA 合成において，ピッチマークの位置はゼロ交差位置に通常は設定されるが，この合成を「波形のピークの位置」にピッチマークを合わせて合成したい場合には，4.3.で述べたピッチマーク位置の修正手法によってピッチマーク位置を変えることによって同様に合成することができる．

上記の方法で合成した音声の波形例を図１１に示す.（A）は原音声波形（「青い波だ」の冒頭部）である．この原音声のピッチ周波数を0.8倍にして低い音声として合成した

ものが（B）と（C）であり，前者は PSOLA 合成音，後者は SWS 合成音である．

合成された音声は，順次 WAV ファイルとして保存し，聴知覚実験に利用することができる．

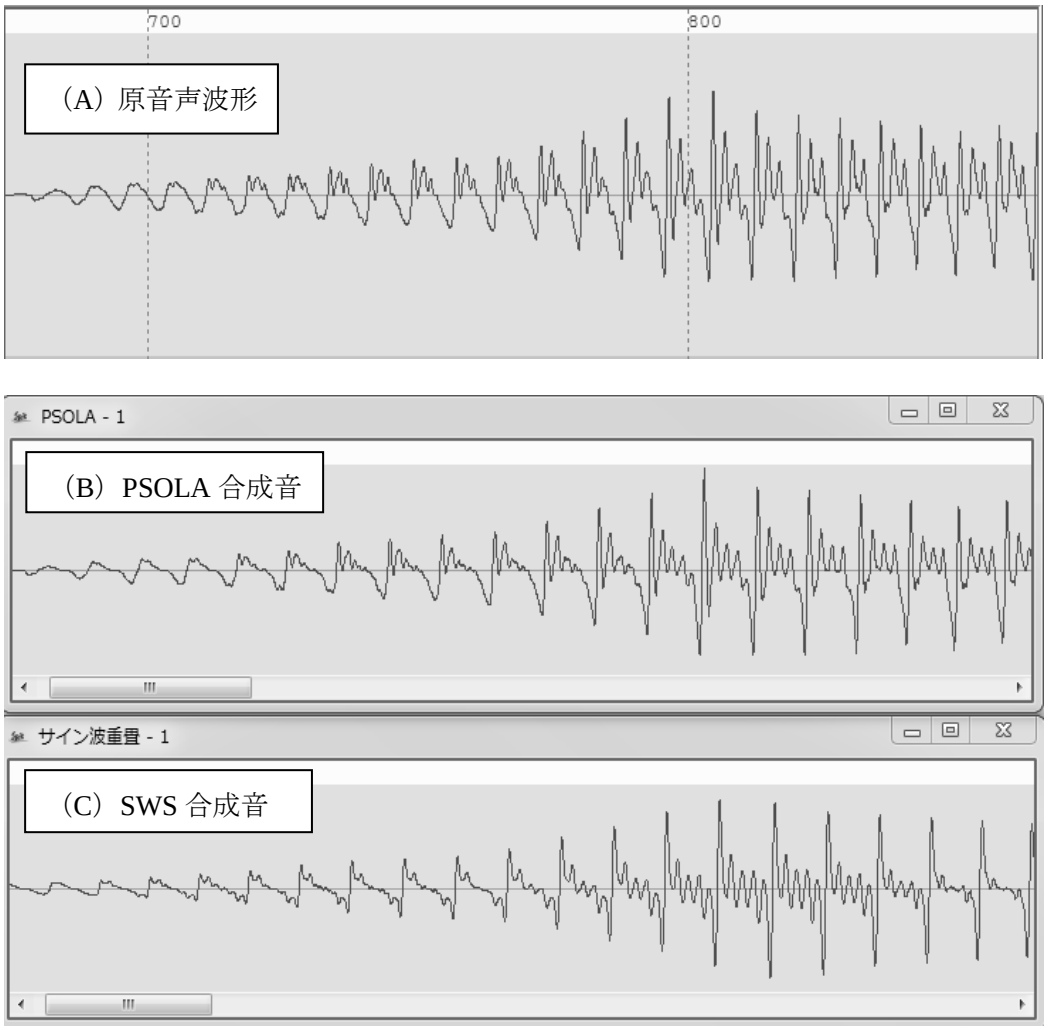

図１１　原音声と合成音音声波形（音声「青い波だ」の冒頭部分）
（合成音は，いずれもピッチ周波数が原音声の 0.8 倍になっている）

5. まとめ

言語音声のピッチ周波数パタンや音韻持続時間の特性など，超分節的特性を種々に変

形した合成音を作成する実験ツールについて報告した．合成音の作成では，1 ピッチ波形の編集に基づく PSOLA 方式と，音声の調和構造を実現する正弦波重畳（SWS）方式の 2 つの方法による合成が可能である. いずれも自動もしくは手動編集によって設定されたピッチマークに基づいて合成がなされる．ピッチ周波数や持続時間の変更は，広く利用されて馴染みのある表計算ソフト：エクセルを使用した.

本ソフトウェア・ツールの利用によって，アクセント，声調，イントネーションなどの超分節的特性を様々に変更した音声の合成が可能になる．こうした変更は，表計算ソフト上で直接数値を書き換えたり, 内部関数を使ってピッチ周波数特性を生成させたりできるため, 聴知覚実験で使用する多くの刺激音声の作成も容易にできるようになるであろう.

6. 今後の課題

今後本ツールは，日本語アクセントや東南アジア諸言語の声調，イントネーションの研究等に実際に利用していく予定である. また, 本論で述べた SpitEditor とリンクして, 再合成された音声刺激を被験者にランダム提示し, 判断結果を集計する受聴実験ツールの作成を進めており, これによって知覚実験用のツールとして利用しやすいものとなるであろう.

なお，本ツールは，今後利用しながら，使いやすいように変更することがあるため, 画面表示,「メニュー」の構成など本論の説明とは若干異なってくる可能生がある．最終的な形式はマニュアル等を参照していただきたい.

＜謝 辞＞

本研究とツール作成にあたり, 超分節研究プロジェクトの共同研究者としてご討論いただいた本学：峰岸真琴教授（アジア・アフリカ言語文化研究所），岡野賢二准教授，降幡正志准教授，神田外語大学：春日淳准教授に感謝申し上げる．また本ツール作成に尽力いただいた杉浦功一氏に深謝する.

なお，本研究は，科学研究費（基盤 B）「東アジアと東南アジア言語における超分節特性の比較対照に関する研究」（研究代表者：佐藤大和，課題番号：23300093）の補助金によってなされたものである.

参考文献

1. Boersma, Paul and Weenink, David:　praat version: 5.3.41
URL: http://www.fon.hum.uva.nl/praat/
2. 板倉文忠, 東倉洋一. 1978.「音声の特徴抽出と情報圧縮」, 情報処理, 35, pp.644-656, など参照
3. Moulines, Eric and Charpentier, Francis. 1990. "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphone," Speech Communication 9, pp.453-467
4. Hirokawa, Tomohisa, Itoh, Kenzo and Sato, Hirokazu. 1992. "High quality speech synthesis based on wavelet compilation of phoneme segments," Proceedings of ICSLP 92, pp.567-570

An experimentation tool for perceptual researches of spoken languages

Hirokazu SATO and Yukie MASUKO

A software tool (SpitEditor) of supra-segmental modification for spoken language perception experiments is proposed. Two speech synthesis methods are available in the system. One is PSOLA (Pitch Synchronous Overlap and Add) method, and the other is SWS (Sinusoidal Wave Superposition) method, which are both based on pitch marks automatically placed at zero-crossover points of the speech wave.

Microsoft Excel, which is an extensively used spreadsheet software, is applied to the system to edit or modify original speech parameters such as pitch frequencies and segmental durations on the spreadsheet.

A number of required speech stimuli for perceptual evaluation tests can be compiled by the software tool. This paper describes how to make the tool work on a computer screen, how to detect zero-crossing points of original speech, how to edit extracted pitch frequencies and segmental durations, how to newly synthesize speech on these modified parameters, and how to save the resynthesized speech.