ＤＳＰを用いた音声合成スペクトル・エディタ

DSPを用いた音声合成スペクトル・エディタ

高良富夫*　山城辰也*

The Spectrum Editor System With DSP Speech Synthesizer

Tomio TAKARA* and Tatsuya YAMASHIRO*

Abstract

In order to analyze the phonemic characteristics of speech sounds, we often need to perform a listening test using special kinds of speech data which are synthesized by a computer. Genarally, it takes a long time to prepare speech data for the listening test because a new program is required when we need new speech data.

In this paper, we report on a new speech processing system which provides an easy way to synthesize artificial speech sounds using a personal computer. In this system, the speech sound is first analyzed into pitch, power, and spectrum structure, the three basic factors of sound. Next, the time pattern of these three factors are freely and easily edited using a keyboard and a mouse just as when sentences are edited on a word processor. Last, we synthesize the speech sound from these three edited factors. Because we process the three basic factors of sound, we can synthesize any kind of speech sound on this system.

Key Words: Speech synthesis, Editor, Spectrum, Three basic factor of sound, User interface

1．まえがき

言語音声は，人間に聴取されてはじめて意味を持つ．従って，音声の音韻的特徴の物理的実体を明らかにするためには，音声を音響的に分析するだけでは不十分であり，しばしば種々の音声の聴取実験を行う必要がある．

聴取実験用の音声としては，人間が発声できない特殊なものが必要とされることも多く，この場合は，音声資料はコンピュータを用いて作成される．コンピュータによる音声の合成は，これまで一般に，聴取実験の度に，必要なプログラムを書いて行われており，大変手間のかかるものであった．

そこで我々は，パーソナルコンピュータを用いて必要な人工音声が簡単に合成できるシステムを構成した．本システムでは，まず音声を音の三要素であるピッチ，パワー，スペクトル構造に分解（分析）する．次に音の三要素の時間変化パターンを，キーボードとマウスを用いて，あたかもワープロで文書を編集するように，自由自在に編集する．最後に，編集した音の三要素を用いて音声を合成する．この方法では，音の三要素を用いて音声を合成するので，聴感上，どのよう

受理：1993年5月10日

* 工学部電子・情報工学科　Dept. of Electronics and Information Eng., Fac. of Eng.

な音も作成することができる．

これまで，パーソナルコンピュータ等で動作する音声処理ソフトウエアはいくつか存在したが，これらの大半は，単に音声波形を編集するものであり，任意の音声を作成することはできない．また，音声の重要なパラメータであるスペクトル構造やホルマントを取り扱えるものもあるが，その処理は分析にとどまっており，本システムのように，合成までできるものはないようである．

本システムでは，音声の特徴（音色）に最も関係しているとされるスペクトル構造を自在に編集し，その結果を音として確認できる．そこで，これを音声合成スペクトル・エディタと名付け，以下にそのシステム構成，機能を紹介する．

2．システムの概要

本システムのハードウエア構成を図1に示す．本システムは，パーソナルコンピュータ PC-9801RA5，ディジタル信号処理プロセッサ (DSP) MSM6992 および A/D 変換器，D/A 変換器を中心に構成されている．入力装置としては，キーボード，マウス，マイク，カセットデッキ，出力装置としては，ディスプレー，プリンタ，スピーカー，カセットデッキ，外部記憶装置としては，ハードディスク (HDD)，フロッピーディスク (FDD)，光磁気ディスク (MO) が利用できる．

マイクやカセットデッキから入力された音声は，4.8kHz 低域通過フィルタを通過した後，A/D 変換器により12ビット精度で量子化される．パーソナルコンピュータや DSP で処理された信号は，外部記憶装置に出力されるほか，D/A 変換器，低域通過フィルタを通過させ，音声として聴取することができる．

本システムのソフトウエア構成を図2に示す．本システムは，入力，分析，編集，音声合成，出力，聴取

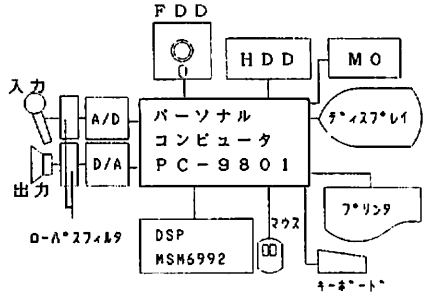

図1　システムのハードウェア構成

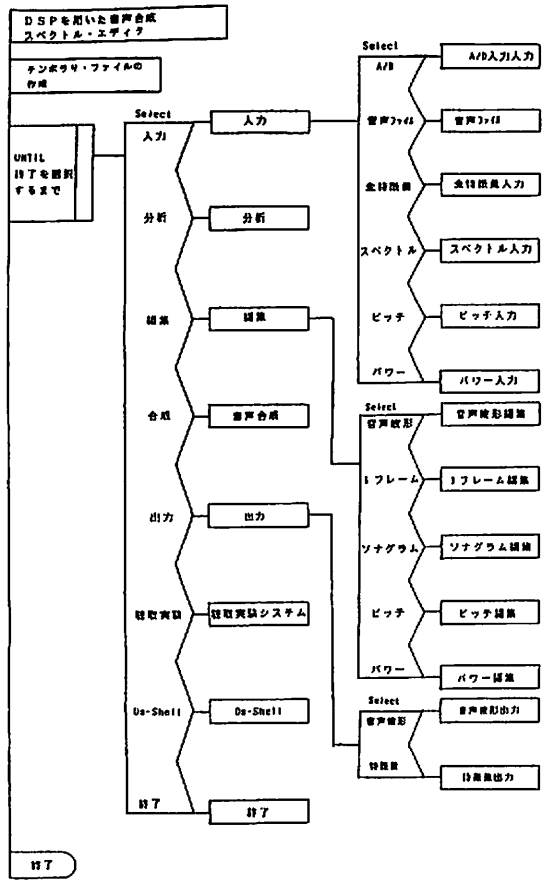

図2　システムのソフトウェア構成

実験の各サブシステムからなる．さらに，入力，編集，出力は，いくつかの下位サブシステムから構成されている．各サブシステムについては次章で説明する．

本システムのプログラムは，主として Lattice-C で記述されている．ただし，分析部は MS-FORTRAN で，A/D 変換，D/A 変換のサブルーチンは MS-DOS のアセンブラで，DSP のプログラムは DSP 専用のアセンブラで記述されている．Lattice-C のメモリ構成の制限から，本システムでは，音声データの最大サンプル数を8000とした．これは，10kHz サンプリングの時は0.8秒の音声長になる．

本システムは，主メモリの制約から，複数の実行プログラムに分割してあり，各プログラムをオーバーレイ方式により実行している．そのため，各プログラム間でデータの受渡しをするため，多くのテンポラリファイルを作成している．また編集部では，再編集や編集破棄をしたときのために，編集前の状態を保持す

るテンポラリファイルを作成する．これらのテンポラリファイルは，メインメニュー（図2参照）で終了を選択するまで保持されるので，例えば，OS-Shell でOSへ抜け，本システム以外のプログラムで処理できるなど，有効に利用される．

3．システム各部

システムを起動すると，図3に示すメインメニュー画面が現れる．画面の目盛りの入った四角の枠は，それぞれ上から，ソナグラム，ピッチ，パワー，音声波形の，時間変化パタンを表示する窓である．画面の最下部にメインメニューがある．これらのいずれかをマウスで選択することによりそれぞれのサブシステムへ移行する．各サブシステムを終了するとこの画面へ戻り，各窓には処理結果が表示される．

3.1 入　力　部

入力部では，A/D 変換入力，音声ファイル入力，全特徴量入力，スペクトル入力，ピッチ入力，パワー入力の各コマンドが利用できる（図2参照）．

A/D 変換では，マイクまたはカセットデッキから，-5～5 [V] のアナログ音声信号を入力し，これを-2048～2047の整数値に変換して，主メモリに取り込む．パラメータの初期値は以下のように設定されている．

A/D 変換の個数	3000[個]
サンプリング周期	100[マイクロ sec]
前読み個数	500[個]
トリガーレベル	500

これらのパラメータの値は，システムが指定する範囲で自由に変更することができる．オーバーフローまたはアンダーフローした入力データがある場合は，警告が表示され，再入力が可能である．ここでは，入力した音声データを D/A 変換して聴取・確認することができる．

その他のコマンドでは，それぞれのファイルを外部記憶装置から入力する．その場合，ファイル名の代わりに "?" をキー入力すると，OS-Shell を起動し，MS-DOS のコマンドが使用可能になるので，ファイルの一覧を見たり，ファイルをコピーしたりすることができる．スペクトル，ピッチ，パワーを独立に入力することができるので，ピッチやパワーを入れ換えた音声は簡単に合成できる．

3.2 分　析　部

分析部では，音の三要素である音の高さ，大きさ，音色にそれぞれ対応する，ピッチ，パワー，スペクト

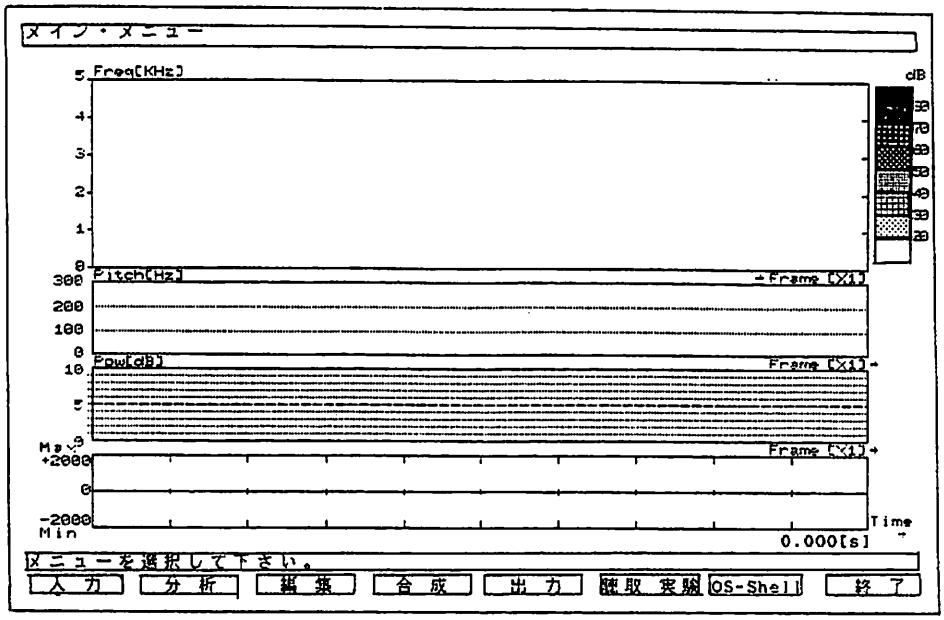

図3　メイン・メニュー画面

ル構造を音声波形から抽出する．ここでは，ケプストラム法[1]に基づいて音声分析を行っており，抽出される特徴量は以下のとおりである．

(1) スペクトル包絡の時間変化パタン
(2) ピッチ周波数の時間変化パタンと有声／無声判別結果
(3) パワーの時間変化パタン

スペクトル包絡の抽出には，改良ケプストラム法[2]を用いており，ピッチ抽出は，高ケフレンシー部のピーク位置の抽出により，また有声／無声の判別は，スペクトルの低周波数領域のパワーの大／小により行っている[2]．パワーは，本システムでは，ゼロ次のケプストラム係数の値としている．

現在のところ分析部はFORTRANで記述されており，分析時間は数十秒を要する．

3.3 編集部

編集部は，音声波形編集，1フレーム編集，ソナグラム編集，ピッチ編集，パワー編集の各サブシステムからなる．(図2参照) 本システムでは，音声波形だけでなく，上述の音声の特徴量を編集することができる．特に，音韻性に最も寄与すると考えられるスペクトル包絡に対しては，1フレーム編集およびソナグラム編集の2種の方法を提供している．

[音声波形編集]

A/D変換により，またファイルから取り込まれた音声波形には，音声データとして使用するときには不要となる部分が含まれている．音声波形編集では，音声波形から必要な部分を切り出したり，その結果を聴取して確認したりできる．図4に，音声波形編集の画面を示す．画面の下部に示されている各コマンドの機能を以下に示す．

(1) 横軸拡大：指定された音声区間を切り出す．残りの音声区間は削除される．切り出された部分は横軸が最大に拡大されて編集後波形として表示される．これは，時間軸を拡大して波形を詳しく見たいときにも使用することができる．
(2) 振幅拡大：指定された区間の音声波形の振幅を拡大または縮小する．振幅値が±2000を超えたときにはそれぞれ±2000に設定される．従って，振幅を2000倍に拡大することにより，ゼロ交差波は簡単に得ることができる．
(3) 削除：指定された区間の音声波形を削除する．
(4) 空白挿入：指定された音声区間の値を0にし，その区間を無音区間とする．
(5) D／A：波形編集をした後の波形や編集する前の波形をD/A変換し聴取・確認する．編集後の波形は繰り返し連続的に聴取することができる．
(6) 終了：編集を終了する．ただし，この時点で，“編集を継続する”，“もう一度編集し直す”，“編集結果を採用せず終了する”，“編集結果を採用して終了する”のいずれかを選択することができる．

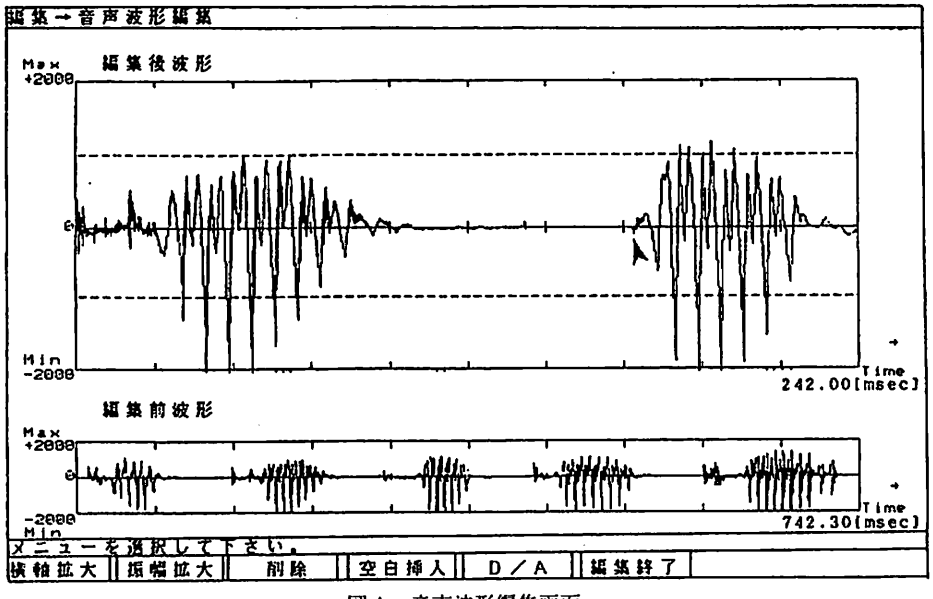

図4　音声波形編集画面

［1フレーム編集］

聴覚の神経生理学的研究および種々の音響的刺激に対する反応に関する実験心理学的研究によると，母音の音韻的情報はスペクトル包絡全体に一様に分布しているのではなく，声道の共振周波数であるホルマント周波数によって特徴づけられている．1フレーム編集では，フレーム単位でスペクトル包絡のピークや谷を強調,作成,削除したりすることによって,スペクトル包絡の形状と音韻性の関係を検討することができる．

1フレーム編集に処理が渡るとソナグラムが表示されるので，まず，編集するフレームを指定する．フレームが指定されると，図5に示す1フレーム編集画面が現れる．ただし，図5では簡単のため，指定したフレームのスペクトル包絡だけを示してある．1フレーム編集画面には，指定されたフレームとその前後のフレームのスペクトル包絡が示される．この画面上で，マウスを用いてスペクトル包絡曲線を描き変えることができる．スペクトル包絡を描き終わると，後述のFFT（高速フーリエ変換）によるスムージングが自動的に行われる．1フレーム編集におけるコマンドは以下のとおりである．

(1) 前編集：指定したフレームの1フレーム前を編集する．
(2) 後編集：指定したフレームの1フレーム後を編集する．
(3) 再編集：編集したフレームをさらに編集する．
(4) 後に複写：編集したスペクトル包絡をその後ろに複数回複写する．
(5) フレーム変更：編集するフレームを変更する．
(6) 編集終了：1フレーム編集を終了する．

〈FFTによるスムージング〉

マウスを用いて手作業で滑らかなスペクトル包絡を描くことは容易でない．そこで1フレーム編集では，FFTを用いて自動的にスペクトル包絡のスムージングを行う．図5の水平の破線で示したように，強調したい山や谷の部分に数点の値を与えておくと，これらの点を滑らかに結ぶ補間曲線がえられる．

この処理は次のように行われる．まずスペクトル包絡をFFTにより逆フーリエ変換し，ケプストラム係数を得る．次に，ケプストラム係数の低ケフレンシー部（0～29次）をそのまま残し，高ケフレンシー部の値を0にする．このケプストラム係数をFFTによりフーリエ変換し，スムージングされたスペクトル包絡を得る．

なお，本システムの合成部においては，ケプストラム法に基づく合成が行われるので，ここで作成された滑らかなスペクトル包絡，すなわち0～29次のケプストラムと同じ特性の合成音が作成される．

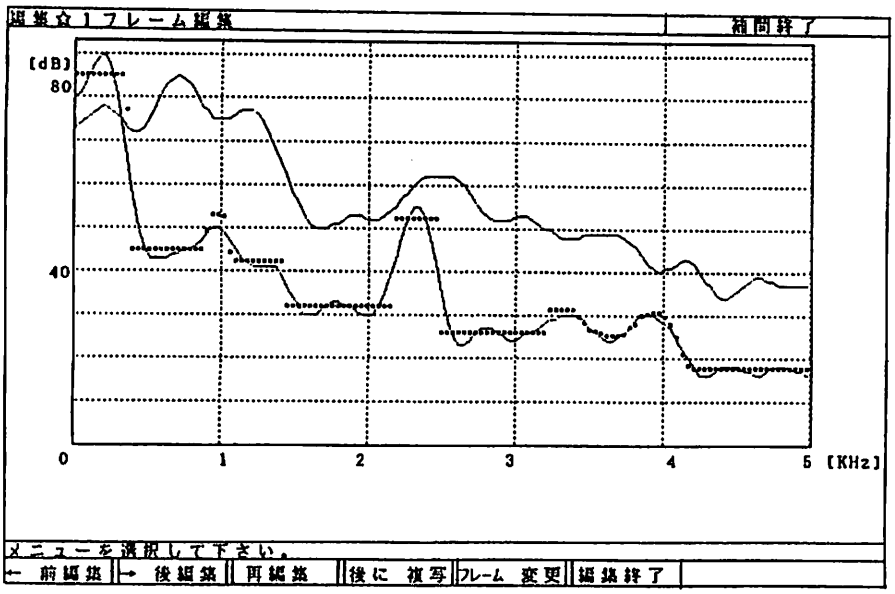

図5　1フレーム編集画面

[ソナグラム編集]

1フレーム編集は，特定の時間位置の音声信号のスペクトル包絡特性の編集を行うものであった．しかし，音声の特徴の時間変化を表すスペクトル包絡の推移，変化について検討することは，子音や単語，連続音声の音韻性を調べる上で重要である．

時間をパラメータとして取入れ，音声スペクトルの時間変化を視覚的に分かりやすい形で表示したものはスペクトログラフと呼ばれ，その代表的なものに，スペクトル強度を画像の濃度で表したソナグラムがある．

本システムのソナグラム編集では，ソナグラム上で，通常の文字エディタと同様に“削除”，“複写”，“移動”を行えるほか，山や尾根を作成する“詳細編集”，作成した山や尾根を滑らかな包絡にするための時間軸方向の“線形補間”，周波数軸方向の“FFTによるスムージング”といった操作も行うことができる．

図6にソナグラム編集画面を示す．ソナグラム編集は以下の6つの編集処理で構成されている．

(1) 削除：指定した2つのフレームの間のフレームを削除する．

(2) 値挿入：指定した2フレーム間に一定の値を挿入する．

(3) 複写：指定したフレーム間のデータを指定したフレームの後ろに指定した回数だけ複写する．ただし，システムの扱える最大フレーム数を超えるときには警告を発し，処理を行わない．ピッチ，有声／無声判別結果，パワーも複写される．

(4) 移動：指定したフレーム間のデータを指定したフレームの後ろに移動させる．従って，編集前後で全フレーム数は変化しない．なお，ピッチ，有声／無声判別結果，パワーも移動する．

(5) 詳細編集：ソナグラム上の任意の位置に，選択した値を設定する．設定値の与え方は，ソナグラムの右側に表示される階調バーによる方法と，キーボードから任意の値を設定する手入力の2方法がある．詳細編集によって，山や尾根を自由に作成することができるので，疑似的なホルマントは容易に作成することができる．しかし，手作業で作成されたスペクトル包絡は急激に変化しているので，(6)のスムージングによって滑らかな包絡にしておくことが必要である．

(6) SMOOTHING：全フレームについて1フレーム編集で述べたFFTによるスムージングを行う．

(7) 線形補間：指定した区間を，その区間の始点と終点の値で時間軸方向に線形補間した値で置き換える．

(8) 編集終了：ソナグラム編集を終了する．ただし，ここで，“再編集”や“編集結果の破棄”を選択することができる．

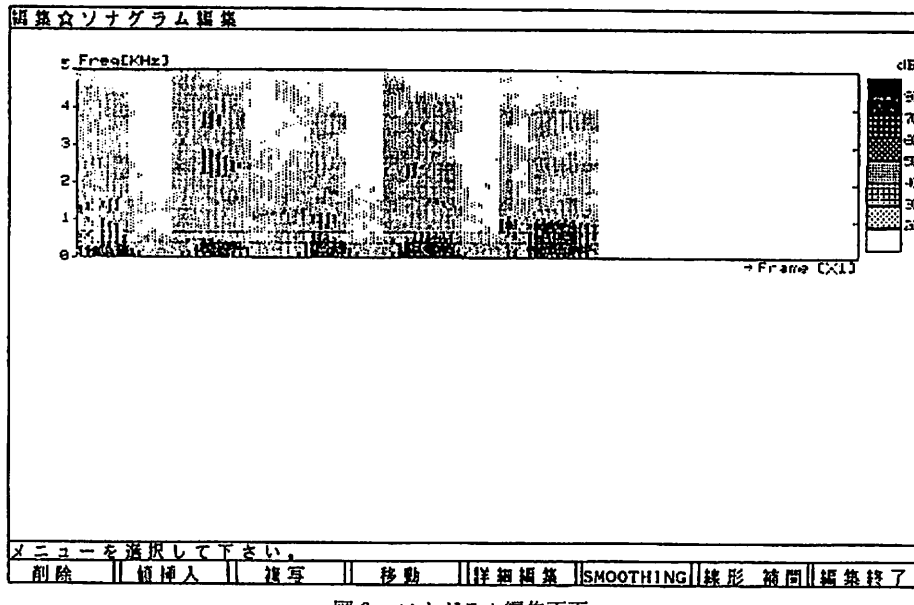

図6　ソナグラム編集画面

[ピッチ編集]

ピッチ編集では，ピッチおよび有声／無声判別結果を編集することができる．ピッチ編集画面を図7に示す．ピッチ編集におけるコマンドは以下のとおりである．

(1) ピッチ編集：ピッチ周波数の時間変化パタンを，マウスで描き直すことにより編集する．

(2) 有／無声：画面下方に有声区間が点で表示されている．マウスを左クリックすると有声に，右クリックすると無声に変化する．

(3) 戻る：ピッチ編集を終了する．

[パワー編集]

パワー編集では，パワーの時間変化パタンをマウスで描き変えることにより編集することができる．図8にパワー編集画面を示す．そのコマンドは以下のとおりである．

(1) 編集継続：編集結果を画面に残し，もう一度編集する．

(2) 再編集：編集前のパタンを画面に残し，もう一度編集する．

(3) 編集破棄：編集結果を破棄し，編集前の状態で終了する．

(4) 終了：編集結果を採用し，処理を終了する．

3.4 合　成　部

合成部では，分析，編集された特徴量を用いて音声の合成を行う．本システムで採用している音声合成方式を図9に示す．

本システムにおいて，分析され編集されるものは，スペクトル包絡の時間変化パタン，ピッチ周波数，有声／無声の判別結果，およびパワーであった．合成部では，まず，これらのパラメータのうち，スペクトル包絡とパワーを合成用フィルタに適合する形式に変換する．スペクトル包絡は逆FFTによりケプストラム係数に変換される．このとき合成に使用されるのは，ケプストラムの低ケフレンシー成分（1～30次）である．またパワーとしては，ケプストラムのゼロ次の係数を用いている．従って，編集されたパワーは，ケプストラムのゼロ次の係数として使用される．

合成フィルタの音源には，ピッチ周期 Pi のインパルス列およびM系列雑音が用意されており，有声区間では前者が，無声区間では後者が選択される．フィルタの入力における音源パワーを一定に保つために，インパルス列に対して常に $\sqrt{Pi}$ が乗算される．またフィルタの特性を与えるケプストラム係数は，無声区間では1フレーム毎に更新されるが，有声区間では，ピッチパルスの発生する時点で，その前後のフレームのケプストラム係数の線形補間によりその時点のケプストラム係数を求め，1ピッチ周期内は一定に保たれる．

合成フィルタとしては，対数振幅特性を2乗平均誤差が最小となるように近似できる対数振幅近似フィル

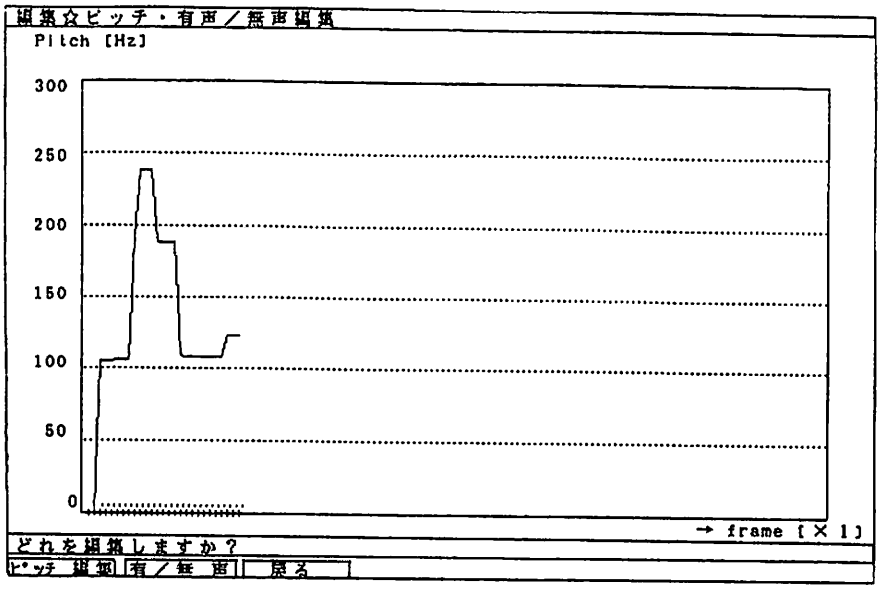

図7　ピッチ編集画面

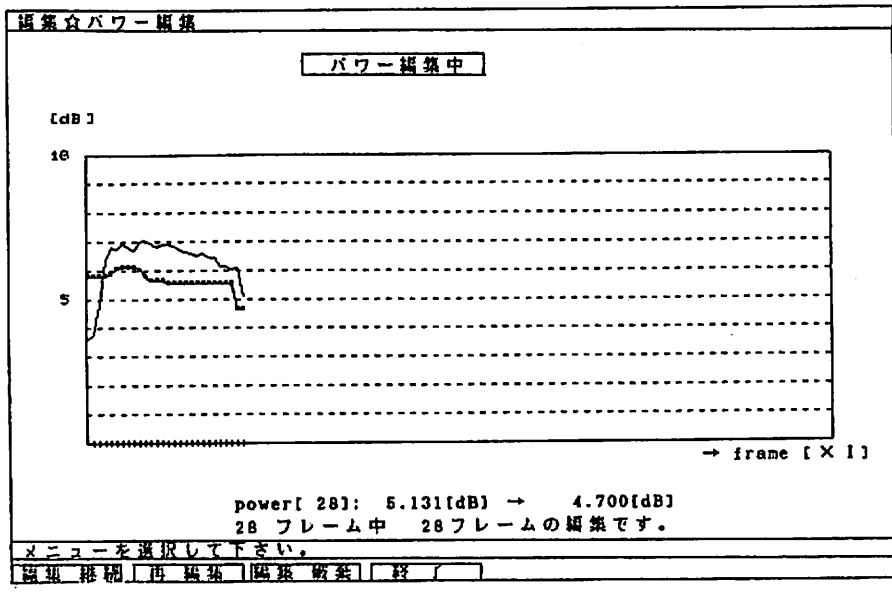

図8　パワー編集画面

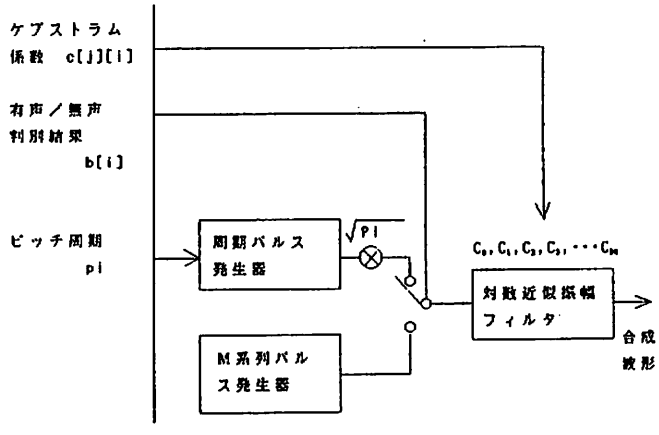

図9　合成システムの構成

タ[3]を用いた．これにより，スペクトル包絡に対するフィルタの特性の誤差が，聴覚の特性にあった対数目盛り上において一様に小さくなり，スペクトル包絡の時間推移で決まる音声の音韻的特徴を精度よく表現できる．またフィルタは，沖電気社製のディジタル信号処理プロセッサ MSM6992 を用いて構成し，処理速度の高速化を図った．

波形が合成されると，直ちに D/A 変換を行い，合成波形を表示する．

3.5 出　力　部

出力部は，音声波形出力と特徴量出力とからなる．(図2参照)

[音声波形出力]

音声波形出力画面を図10に示す．ここで使用できるコマンドは以下のとおりである．

(1) 印刷：4つの波形をプリンタで印刷する．

(2) 拡大表示：4つの波形から任意の1波形を選択し，その振幅を拡大して画面に出力する．拡大した波形は，印刷，D/A変換することができる．

(3) 保存：4つの波形から任意の1波形を外部記憶装置に保存する．ファイル名は，MS-DOS形式にし

たがう．

(4) ④へ出力：①ACTIVE，②原音声，③合成音声のいずれかを④バッファーへコピーする．

(5) ④→①：④バッファーの波形を① ACTIVE へコピーする．

(6) D／A：4つの波形から任意の1波形を選択し，D/A 変換する．

(7) 戻る：音声波形出力を終了する．

[特徴量出力]

特徴量出力画面を図11に示す．ここで使用できるコマンドは以下のとおりである．

(1) 保存：表示されている特徴量を外部記憶装置に保存する．ファイル名はMS-DOS形式にしたがう．

(2) 印刷：表示されている特徴量をプリンタで印刷する．

(3) 3D表示：スペクトル包絡の時間系列を立体表示する．表示結果はプリンタで印刷することができる．

(4) メインへ：メインメニュー画面に戻る．

3.6 聴取実験システム

聴取実験システムでは，音声波形のバッファーを3つ用意し，作成した音声を相互に聴取して比較できるようになっている．聴取実験システム画面を図12に示す．ここで使用できるコマンドは以下のとおりである．

(1) 印刷：4つの波形をプリンタで印刷する．

(2) 拡大表示：4つの波形から任意の1波形を選択し，拡大して画面に出力する．拡大した波形は，印刷，D/A変換することができる．

(3) 保存：4つの波形から任意の1波形を選択し，外部記憶装置に保存する．ファイル名は，MS-DOS形式にしたがう．

(4) ACTIVEへ：バッファー①～③のいずれかの波形をACTIVEへコピーする．

(5) BUFFERへ：ACTIVEの波形をバッファー①～③のいずれかへコピーする．

(6) D／A：4つの波形から任意の1波形を選択し，D/A変換する．

(7) 戻る：メインメニュー画面に戻る．

3.7 OS-Shell

OSへぬけ，MS-DOSのコマンドや実行ファイルを実行することができる．これにより，MS-DOSのコマンドラインから実行できるソフトウエアは，すべて本システムのコマンドと見なすことができ，システムの汎用性が大幅に増す．OS-Shell から本システムに戻るためには“exit”と入力する．

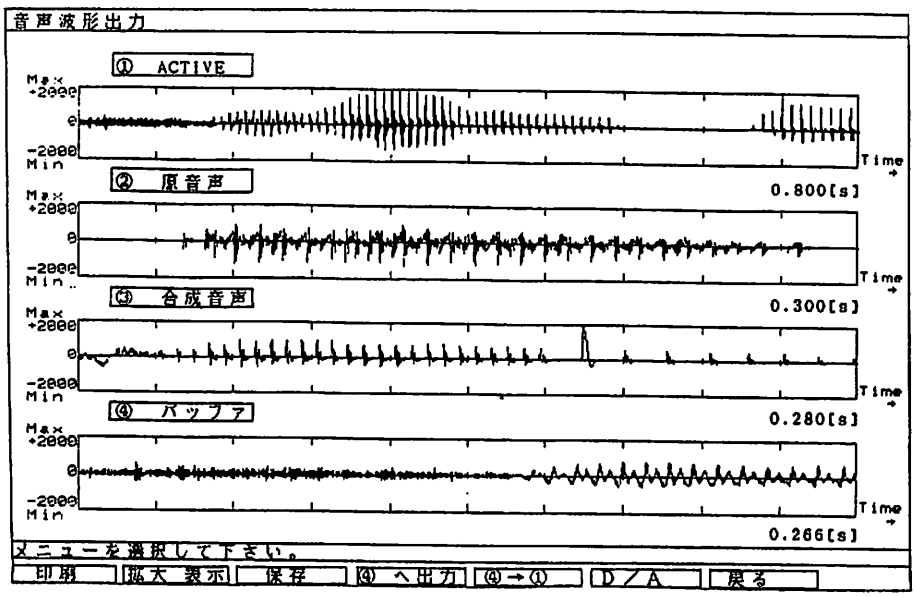

図10　音声波形出力画面

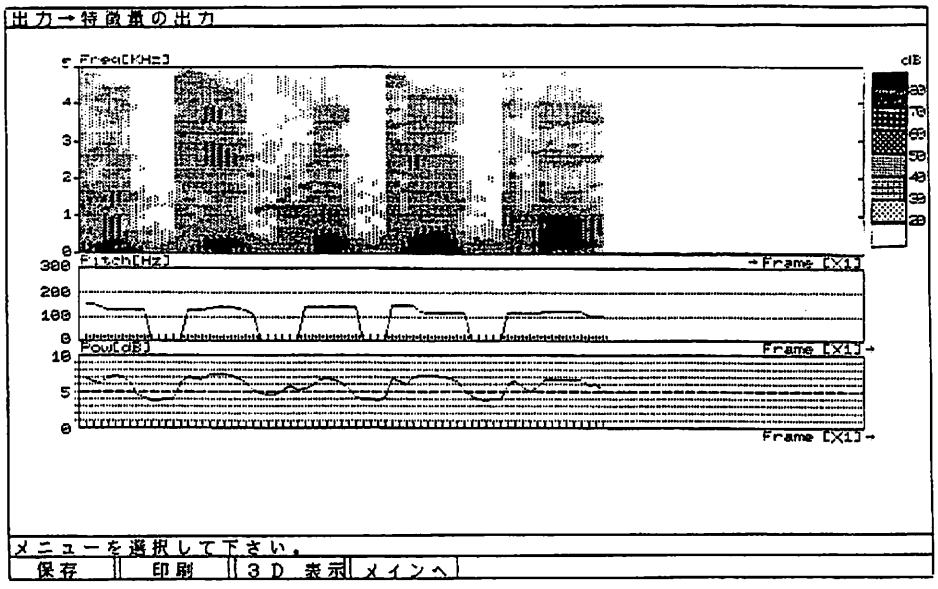

図11　特徴量出力画面

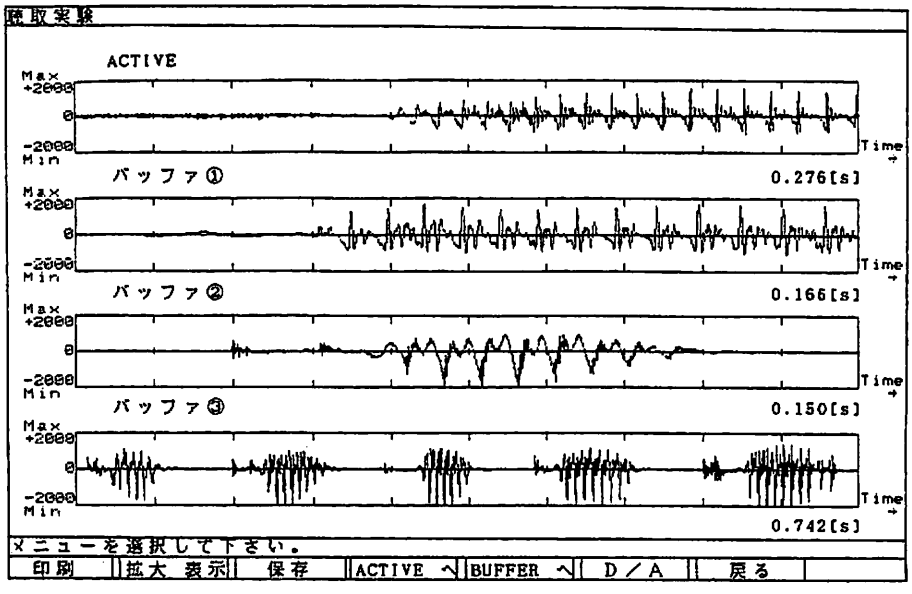

図12　聴取実験システム画面

5．評価・検討

作成したシステムを３人が使用し，“対話型コンピュータ・システムのユーザー評価シート”[4]を用いて本システムの評価をした．評価シートは22項目の質問事項からなり，各項目について11段階で評価するものである．まず，最初に作成したシステム[5]について評価を行い，この結果に基づき改良を加えたシステムで再び評価を行った．

主な改良点は以下のとおりである．

(1) 改良前のメニュー選択の深さは，各サブシステムによってまちまちで，深さ３以上のものもあり，サブシステム相互の移行が面倒であった．これを，最も深いもので深さ２とするようにした．
(2) 大きなメニュー項目画面を使用したため最下層のメニューを選択するまでデータが見えなかったが，メニュー項目を画面の最下段に表示し，常にデータが画面上に表示されるようにした．
(3) 聴取実験システムおよびOS-Shell を追加した．
(4) キー入力ミスに対してシステムが必ず警告を発するようにした．
(5) マウスのクリックにおけるチャタリングを極力抑えるため，時定数を設定しなおした．

これにより，評価シートの大半の項目について評価値が高くなった．特に，“画面のレイアウトは常に作業をしやすくしていますか”と，“画面の表示順序は明らかになっていますか”の２項目については，３人とも評点が大幅に向上した．しかし，“動作速度”，“作業と操作のイメージを近くする”等の項目はまだ評点が低い．

6．むすび

音声の音韻的特徴を調べるため，音の三要素であるピッチ，パワー，スペクトル構造の時間変化パタンを編集し，聴取実験のための任意の音声が簡単に合成することのできる，音声合成スペクトル・エディタを作成した．本システムでは，DSPを使用することにより音声合成を高速化した．さらに，作成したシステムの評価試験を行い，ユーザー・インターフェースについても考慮した．このシステムはこれまで，人工内耳シミュレーションのためのホルマント合成器として，また琉球方言の音声分析機として有効に利用されている．

文　献

[1] L．R．Rabiner & R．W．Shafer著　鈴木久喜訳：“音声のディジタル信号処理（下）”，pp.135-140，コロナ社（昭58-04）．
[2] 阿部，今井：“ＣＶ音節のケプストラムパラメータからの音声合成”，電子通信学会論文誌(D)，J64-D，9，pp.861-868（昭56-09）．
[3] 今井　聖：“対数振幅近似（LMA）フィルタ”，電子通信学会論文誌（A），J63-A，12，pp.886-893（昭55-12）．
[4] B．Shneiderman著　東　基衛，井関　治　監訳：“ユーザー・インターフェースの設計”，日経マグロウヒル社（1987-12）．
[5] 山城辰也：“DSPを用いた音声合成スペクトル・エディタ”，琉球大学工学部電子・情報工学科卒業研究中間発表会予稿集(B)（1989-11）．