圧縮した残差を用いた規則音声合成法

6B−8

斉藤　隆
日本アイ・ビー・エム株式会社　東京基礎研究所

1．はじめに

規則音声合成において予測残差を駆動音源として用いる方式は、インパルス駆動方式に比べ、肉声に近い音質が得られるため、品質の向上が期待できる。ただ、残差駆動の規則合成への適用に際しては、異なるピッチでの合成、駆動音源のデータ量等の課題が残されている。本稿では、比較的小規模なシステムで残差駆動による規則合成を実現するため、残差データを圧縮し駆動音源として用いる方法について検討したので報告する。さらに、この方法を実際に適用したテキスト音声合成システムについても述べる。

2．残差データ圧縮の方針

規則合成においては、合成時にピッチ変更を伴うため、分析合成のように原残差をそのままの形で最大限に利用することが、必ずしも高品質化につながるとは言えない。このことを示す例として、原残差よりも位相処理を施してエネルギーを原点に集中させた残差の方が、ピッチ変更に対しては頑強であることが報告されている[1]。また、合成時に原音と異なる音韻環境で使用される場合に、合成単位の接続部ではスペクトル自体も歪を受けるため、厳密な残差スペクトルを使用しても、その効果があまり期待できないケースも考えられる。そういったことから、残差スペクトルのよりマクロな特徴を利用する方法なども検討されている[2]。ここでは、これらの知見をふまえ、クラスタリングと零位相化によって、コンパクトに表現された残差駆動音源について検討する。

3．圧縮残差を用いる規則合成方式の概要

合成単位とその接続

合成素片単位としては、LSPパラメータで表現したCV/VC/VV単位をベースとして、音韻環境の影響を特に受けやすいものに関しては、VCV/CVCの単位で保持し、総計530種類程度の単位で合成単位辞書を構成している。合成単位の接続方法は、VC単位を前後の環境に合わせてスムージングを行う適応変形法[3]を採用している。

駆動音源

基本的には、各単位素片の分析フレーム毎の残差を駆動音源として利用する。有声音に関しては、次に述べるクラスタリングと零位相化によって圧縮を行なった残差を用いる。無声音のうち、破裂音については原残差を駆動音源として用い、そのほかの摩擦音・破擦音については、白色ノイズを使用する。

4．残差データの圧縮法

圧縮処理の流れ

残差データの圧縮処理の流れを図1に示す。まず、有声音の予測残差信号について、10msの分析フレーム毎に、ピッチ区間を切り出す。次に、各ピッチ区間の残差について、256点のＦＦＴ分析を行ない残差スペクトルを計算する。こうして求められた合成単位辞書の全有声音フレーム（約6500フレーム）の残差スペクトルをクラスタリングして、代表残差スペクトルを得る。この代表残差スペクトルを用いて、各有声音フレームの残差スペクトルのベクトル量子化を行なう。また、代表残差スペクトルの位相項を零として逆ＦＦＴすることによって、駆動音源コードブックを作成する。合成時には各有声音フレームの駆動音源として、残差コードに対応する零位相化波形が駆動音源コードブックから読みだされて使用される。

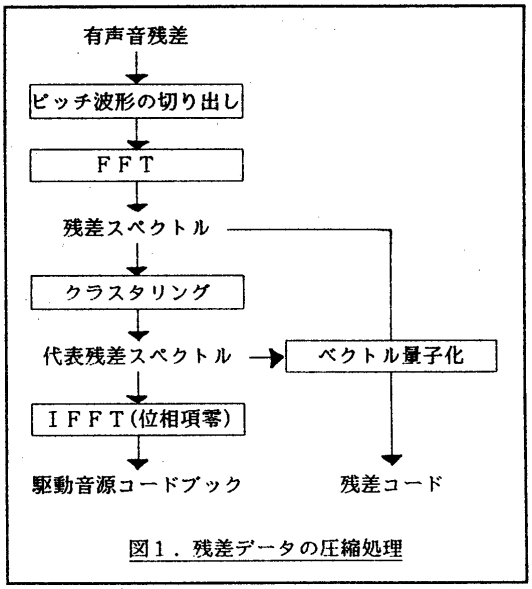

図1．残差データの圧縮処理

Speech Synthesis by Rule using Compressed Residual Signals for Excitation
Takashi Saito
IBM Research, Tokyo Research Laboratory, IBM Japan Ltd.

残差のクラスタリング

クラスタリング法としては、ＬＢＧアルゴリズム[4]を用いた。ただし、クラスター分割は、分散最大の周波数次元方向に対して行なうよう設定した。図２に、コードブックサイズと平均歪みとの関係を示す。今回はコードブックサイズとして、512 (=2^9) を採用した。この場合の圧縮率は、0.08程度となる。

駆動音源の作成

代表残差スペクトルから駆動音源を作成するに際しては、ピッチ変更によるスペクトルへの影響[1]と圧縮効率を考慮して、エネルギーを原点に集中させた零位相化波形を使用する。零位相化波形は、原点集中しているだけでなく対称波形であるため圧縮効果もさらに高くなる。

合成音のスペクトル

母音/e/について、原音のスペクトル(a)と、同じピッチで合成した合成音のスペクトル(b,c,d)を図３に示す。ベクトル量子化による圧縮を行なわない零位相化音源で合成したもの(b)と本方式で合成した合成音のスペクトル(c)との差異は小さい。また、この例からも見受けられるように、本方式で合成した合成音のスペクトル(c)は、原音スペクトルの谷の部分の再現性がパルス駆動の合成音(d)に比べ高いことが分かった。このことは、残差スペクトルが特に零点の情報を多く残していることから、当然の結果ともいえる。これは、文献[2]の結果とも同じ傾向を示している。実際に合成音を作成した結果、パルス駆動に比べて原音に近い音質を得ることができた。

５．実時間テキスト音声合成システムへの適用

テキスト音声合成処理の概要

本規則合成法を用いた実時間テキスト音声合成システムを、汎用ＤＳＰ(11.5MIPS)を搭載した音声処理ボードを用いて実現した[5]。テキスト解析と韻律制御(発話構造の組立て[6]など)は、ＰＣ上で行ない、その結果を音声処理ボードに転送し、ＤＳＰ上でピッチ生成処理と本規則合成法を用いた音声生成処理を行なう。音声生成処理に必要なデータとして、合成単位の声道情報を蓄積した合成単位辞書（130KB）と音源情報を蓄積した駆動音源データ（有声音用コードブック32KB、無声音用原残差19KB）がある。このうち合成単位辞書については、ＤＳＰメモリー(プログラム領域32Kword(96KB)データ領域32Kword(64KB))内に納めることができないため、ＰＣ側のメモリーに格納しておき、必要に応じてＤＳＰの作業メモリー領域に転送し使用する。一方、駆動音源データについては、声道情報に比べ更新レートが高いため、ＤＳＰメモリーに常駐させて使用する構成になっている。因みに、有声音用駆動音源を圧縮しない場合には、約400KBにもなり、このハードウェア構成で実現するのは極めて困難になる。

電話用諸機能・録音再生機能との同時処理も可能

このシステムは、電話を用いたアプリケーションへの対応や、定形文用の録音再生機能も重視し、1枚のボード上で、規則合成処理を、電話用諸機能と録音再生機能と並列に、動作させることができるようになっている。

６．おわりに

規則合成における残差データの圧縮方法について検討し、圧縮残差駆動を用いた実時間テキスト音声合成システムについて言及した。ここで検討したような残差データの圧縮は、高品質な規則合成システムを小規模なシステムでの実現するための有効な手段の一つと考えられる。圧縮法に関する今後の検討課題として、音韻情報や時間的相関を利用した圧縮効率の改善が挙げられる。

参考文献

[1]広川：音響講論，2-2-11(1986.3)
[2]浜田：音響講論，1-6-6(1987.10)
[3]斉藤・大嶋：音響講論，3-4-11(1986.10)
[4]Linde et al：IEEE ASSP VOL.COM-28(1980.1)
[5]大河内他：音響講論，1-P-12(1992.3)
[6]鈴木・斉藤：音響講論，1-2-22(1992.3)

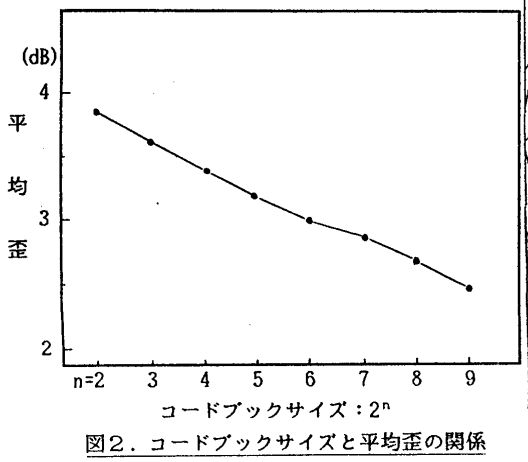

図２．コードブックサイズと平均歪の関係

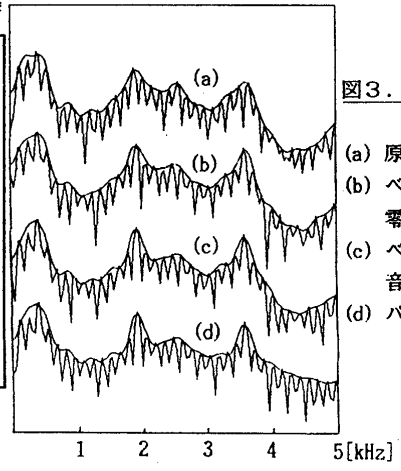

図３．合成音のスペクトルの比較

(a) 原音(母音/e/)のスペクトル
(b) ベクトル量子化を行なわない零位相化音源による合成音
(c) ベクトル量子化した零位相化音源(本手法)による合成音
(d) パルス音源による合成音