音声言語情報処理 17－16
（１９９７．７．１９）

高音質 LPC 分析残差駆動合成器

赤嶺政巳　籠嶋岳彦　土谷勝美

(株)東芝　関西研究所

〒658　神戸市東灘区本山南町 8-6-26
Tel:078-435-3102
Email: akamine@krl.toshiba.co.jp

あらまし：従来の LPC 合成器は、声道パラメータの操作により声質の変更が比較的容易で合成素片辞書も比較的コンパクトにできるという利点がある反面、音質の点で問題があった。本報告では、自然音声のデータベースから代表素片辞書を自動的に学習することにより明瞭で肉声感豊かな音質を実現する残差駆動合成器を提案する。

キーワード:　音声合成、規則合成、残差駆動合成器、代表素片、学習

Residue Excited LPC Synthesizer with High Quality Speech

Masami Akamine, Takehiko Kagoshima and Katsumi Tsuchiya

Toshiba Kansai Research Laboratories

6-26,Motoyama-minami-machi,8-chome,Higashinada-ku,Kobe,658 Japan
Tel:078-435-3102
Email: akamine@krl.toshiba.co.jp

Abstract: Conventional LPC synthesizers have an advantage in flexibility of controlling speech spectrum and in memory size for synthesis units, but do not have a good speech quality. This paper proposes a new residue excited LPC synthesizer with a high speech quality and a new method for automatically generating speech synthesis units. The LPC synthesis filter and its excitation are derived from the synthesis units by LPC analysis.

Key words: Speech synthesis, Text to speech, LPC synthesizer, Synthesis units, Training

1. はじめに

PC やインターネットの普及を背景に文章を音声に変換する音声合成技術に対する需要が増大している。しかし、一般ユーザーが音声合成を違和感なく手軽に利用するには音質や自然性など解決すべき課題が少なくない。これらの課題の中でも音質の問題は自然な声で発話内容を明確に提示する上で重要である。波形編集方式は、実音声波形から切出した素片のピッチ周期や継続時間長を変更して接続することで音声を合成するもので、高音質化が比較的容易であると考えられ、方式の検討が盛んである[1][2][3]。また、高音質化のためには分析、合成などの信号処理を行わない方が良いとの立場から音韻環境及び韻律環境が一致する音声波形を自然音声のデータベースから最長単位で接続する方式が提案されている[4]。これらの方式は、従来の分析合成方式より高音質の合成音を実現している反面、素片辞書のサイズが大きくなるという問題がある。また、スペクトルパラメータが陽に表現されていないため、声質の変更などが難しいという問題もある。

本文では、波形編集方式と同等以上の高音質な音声が合成できる LPC 分析残差駆動合成器を提案する。本方式は、LPC パラメータと残差信号の形で音声素片を表現し、従来の波形編集方式と同様にピッチ同期波形重畳法[5]により韻律を制御、素片を接続するものである。音声素片の LPC 分析、残差駆動合成による音質劣化を最小限に止めることにより波形編集方式と同等の音質を実現する。同時に、予め収集した自然音声のデータベースから音声素片を自動的に学習する新しい方式を提案し、明瞭で高い音質の合成音を実現する。素片の学習は、合成音と自然音声との歪みの総和を最小化する規範でデータベースから素片を自動的に選択することで実行される[6]。この方法により、従来、音質劣化の主要因であったピッチ変更に伴う歪みが激減し、明瞭で肉声感豊かな音質を実現できるようになった。また、本方式では音声素片を LPC スペクトルパラメータと残差信号の組で表現しているため、スペクトルパラメータの操作により声質の変更が容易であると共に合成単位辞書のサイズを大幅に圧縮することができる。

2. LPC 分析残差駆動合成器

図 1 に提案する LPC 分析残差駆動合成器のブロック図を示す。本合成器は後述する方法で作成された音声素片に LPC 分析を適用して、スペクトルパラメータと残差信号を求める分析部と文解析・韻律制御部から与えられる音韻記号列と韻律情報に従って音声信号を生成する合成部から成る。以下に、各部の処理について述べる。

2.1 LPC 分析

従来の残差駆動合成方式は、LPC 分析と逆フィルタリングで得られる予測残差信号から 1 ピッチ分の残差を切り出し、これを駆動信号としている[7][8]。この場合、基本周波数とホルマント周波数が一致する部分でホルマント周波数幅の過少推定が発生し、韻律の変更に伴うスペクトル歪みが合成の際、問題となる。この問題を低減する方法として、ラグウインドーや適応フィルタ[8]によるスペクトル平滑化が用いられている。本文では、簡便な方法として、音声波形にピッチ周期長の窓関数を掛けてピッチ周期分の波形を切り出した後、ピッチ同期 LPC 分析を適用する方法を用いる。この方法では、窓関数により離散的なスペクトルが平滑化されるため、基本周波数の影響が低減されたスペクトル包絡を得ることができる。

2.2 位相処理

ピッチマークに従ってピッチ波形を重畳する際、各ピッチ波形の位相が揃っていないとピッチ周期の連続性が乱れて音質劣化が生じる場合がある。この問題に対しては零位相化や最小位相化

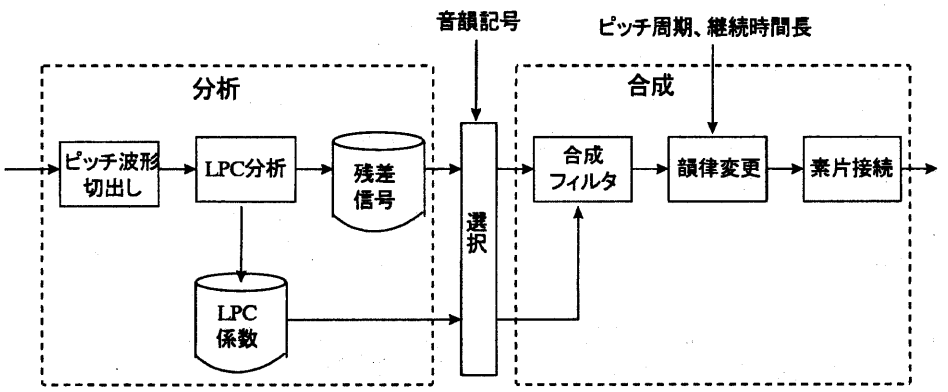

図1 LPC分析残差駆動合成器の構成

が有効であり[9]、零位相化は残差信号のエネルギーを集中させメモリ量削減の効果もある。しかし、位相処理によって合成音の明瞭性が若干低下する場合があり、音質重視の観点から本文の合成器では、位相処理を行わず、自己相関関数に基づいてピッチ波形の位置を制御している。

2.3 素片の合成

従来の残差駆動合成器は、残差信号のレベルで韻律の制御、接続を行っている。これに対し、本文の合成器は、音声素片のレベルで韻律の制御、素片の接続を行う。このため、音声素片を LPC パラメータと残差信号に分解することによる音質劣化がなければ、本合成器は波形編集方式と同等の音質を実現できる。素片の合成は、有声音の場合にはピッチ周期単位の残差信号で合成フィルタを駆動することにより行われ、無声音は白色雑音駆動により生成される。従来は、音質改善のため無声音区間も残差駆動とする場合が多く見られるが[7]、予備実験の結果から雑音駆動による音質の劣化は問題無いと判断した。

2.4 ピッチ周期の制御

ピッチ周期の制御は、従来法と同様ピッチ同期波形重畳法(PSOLA)を用いる。合成音の音質は有声音の滑らかさに負うところが大きい。そこで、ピッチ周期の変化をより滑らかにするため、ピッチ周期をサンプル単位で補間する[9]。第 j フレームと第 $j+1$ フレームの中心時刻を各々 t_1, t_2 とし、ピッチ周期を p_1, p_2 とすると、ピッチ周期が線形に変化する場合、時刻 t におけるピッチ周期 $p(t)$ は次式で表される。

$$p(t) = \frac{(t - t_1)p_2 + (t_2 - t)p_1}{t_2 - t_1} \qquad (1)$$

また、t_1 から t_2 までのピッチマークの位置を $m_k, (k=1,2,...,N)$ とすると、次式が成立する。

$$\int_{m_{k-1}}^{m_k} \frac{2\pi}{p(t)} dt = 2\pi \qquad (2)$$

式(1)(2)から次式が得られる。

$$m_k = m_{k-1} + (m_{k-1} + a)(e^b - 1) \qquad (3)$$

$$a = (t_2 p_1 - t_1 p_2)/(p_2 - p_1) \qquad (4)$$

$$b = (p_2 - p_1)/(t_2 - t_1) \qquad (5)$$

2.5 継続時間長の制御

継続時間長の制御では分析に用いた音声波形と合成波形の各々のピッチマークをどう対応付けるかが問題となる。本文では、時間的なマッピングを関数で行う。この方法によると、マッピング関数を適切に定義することによってピッチ波形の間引き・補間を接続する素片の性質に応じて自由に制御することができる。

2.6 素片の接続

提案する合成器は、CV、VC の素片を用いており、母音定常部で各素片を接続する。接続部における波形の不連続性に起因する歪みを低減するため、接続する母音のピッチ波形を全母音区間に渡って重み付き加算して接続する。

3. 素片の学習

従来、素片の作成は人手による試行錯誤的手法に頼っており、熟練した研究者が長時間かけて単音発声や無意味単語発声または連続発声された音声データから素片を切り出しては、合成音を評価するという一連の作業を繰り返し行う必要があった。一方、音声データベースから素片を自動的に生成する方法として音素環境クラスタリング(COC: Context Oriented Clustering)法[10]が知られている。この方法は、スペクトルパラメータの分散に基づいて音韻環境の拘束条件の下でデータベースから切り出された音声素片をクラスタリングし、各クラスタのセントロイドを代表素片とする方法である。COC は先見的知識に頼らずに統計的評価基準に基づいて代表素片を決定できる特徴があるが、音声合成で問題となっているピッチ変更に伴う歪みを考慮していないため、合成音の音質は必ずしも十分なものとは言えない。

本文では、ピッチ周期と継続時間長の韻律を変更することにより生じる歪みも含めて合成音の歪みを定義し、この歪みを最小化する代表素片の学習法を提案する。

3.1 閉ループ学習法

図 2 に提案する学習法のブロック図を示す。本学習法は、様々な合成器や合成単位に対して適用することができるが、ここでは、2.で述べた残差駆動合成器の CV、VC 音声素片に適用する。すなわち、学習によって音声素片を生成した後、合成フィルタの LPC パラメータと残差信号を求める。

学習に当たっては、まず事前準備として、合成単位の音声素片を音声データベースから大量に切り出し、これらを代表素片候補とする。同時に、同様な方法で学習のためのトレーニングデータを作成する。次に、トレーニングデータをターゲットに代表素片候補のピッチ周期と継続時間長を変更し、音声素片を合成する。このような方法で全ての代表素片候補と全てのトレーニングデータの組み合わせについて、音声素片を生成する。次に、生成された音声素片とトレーニングデータとの歪みを計算し、全てのトレーニングデータに対する歪みの総和を最小にする代表素片を代表素片の候補から選択する。この方法は、合成され

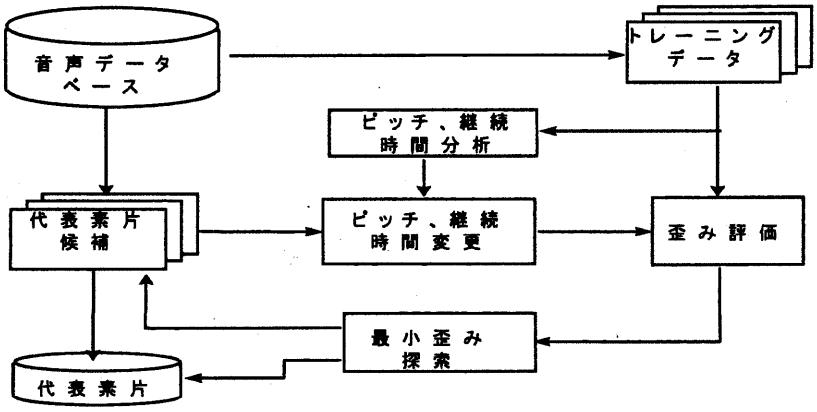

図 2 代表素片の閉ループ学習法

た音声素片の評価結果を素片の学習にフィードバックするという意味で閉ループ学習と呼ぶ。以下に、本学習法で重要になる歪み尺度と代表素片の選択法についてその具体的な方法を述べる。

3.1 歪み尺度

学習の歪み尺度は、主観評価の結果を良く反映するものである必要がある。また、合成音のパワーは合成器で制御されることから、代表素片はパワーが正規化されたレベルで評価する必要がある。このようなことを考慮して、合成素片の歪みを次式で定義する。

$$e_{ij} = \sum_n (r'_j(n) - s'_{ij}(n))^2 \qquad (6)$$

$$r'_j(n) = r_j(n) \Big/ \left(\sum_k r_j(k)^2\right)^{1/2} \qquad (7)$$

$$s'_{ij}(n) = s_{ij}(n) \Big/ \left(\sum_k s_{ij}(k)^2\right)^{1/2} \qquad (8)$$

ここで、r_jはトレーニングデータを表し、s_{ij}はr_jを目標にした代表素片候補 u_i による合成素片を表す。

3.2 代表素片の選択

合成単位当たりの代表素片数を n 、代表素片候補数を Nとすると、代表素片の選択は N個の候補からn個を選ぶ組み合わせの中から次のコスト関数を最小化する代表素片の組を一組探索する問題となる。

$$c(i_1,\ldots,i_n) = \frac{1}{M}\sum_{j=1}^{M} \min(e_{i_1 j},\ldots,e_{i_n j}) \qquad (9)$$

ここで、Mはトレーニングデータの数である。

式(9)のコスト関数を最小化する代表素片の組が求まると、全トレーニングデータを代表素片に対応するクラスタにクラスタリングすることができる。表1に4個の代表素片候補から2個の代表素片を選択する場合の例を示す。u_1～u_4の任意の二個の組み合わせの中でu_2とu_3の組み合わせのコスト関数が最小となる。この結果、u_2とu_3が代表素片として選択される。

表1 合成素片の歪みに基づく代表素片選択の例

		トレーニングデータ				
		r_1	r_2	r_3	r_4	r_5
候補代表素片	u_1	4	3	2	4	3
	u_2	2	3	3	8	1
	u_3	1	6	2	2	4
	u_4	2	1	5	5	6

4. 評価実験

CV、VC の diphone を合成単位として各合成単位に対して1個の代表素片を作成する実験を行った。視察により音韻ラベルが付けられた音声データベースからトレーニングに用いる音声素片データと代表素片候補を切り出し、3.で述べた閉ループ学習法で計 302 個の CV、VC 代表素片を作成した。学習に要した時間は Sun-Ultra2 で約 1.5 時間であった。図 3 は、合成単位(CV、VC)当たりの素片数を増加させた場合のコスト関数の値を示しており、この図から素片数の増加とともに合成音の歪みが単調に減少していることが分かる。従来からパワーやピッチにより素片を使い分けることにより合成音の音質が向上することは知られている。しかし、従来の試行錯誤による方法では、代表素片の作成に多大な労力と時間を要し、代表素片の数を増やすことは容易ではなかった。本文で提案した閉ループ学習法によれば、ラベリングされた音声データが与えられれば短

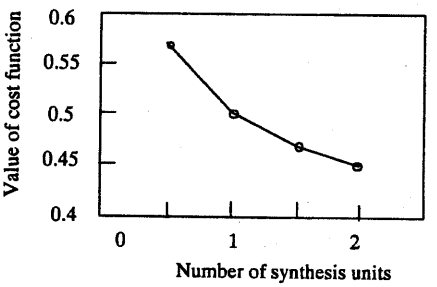

図 3 代表素片数とコスト関数の値の関係

時間で自動的に素片の作成ができ、任意の数の代表素片を生成することが容易である。しかも、パワーやピッチといった先見的な知識で素片の選択を行うのではなく、合成音の歪みの尺度で選択の規則を作成することが可能である。すなわち、トレーニングデータを選択された代表素片のクラスタにクラスタリングし、クラスタ内で共通する要因を抽出することにより素片選択の規則を生成することかできる。

次に、本文の合成器を当社のTTSシステムに適用し、音質を評価した。まず、作成した代表素片を2.で述べた手法で分析し、LPCパラメータと残差信号に分解した形で合成器に蓄積した。蓄積に当たっては、ベクトルースカラ量子化の手法を適用してLPCパラメータ及び残差信号を符号化した。この結果、データ量は一話者当たり約150kbyteと、波形編集方式と比べると1/10〜1/20の非常にコンパクトなものとなっている。大学生7名を含む計10名(男女同数)の一般の被験者による7段階(-3:非常に悪い〜+3:非常によい)の主観評価の結果、本文の合成器を用いた合成音の音質は当社従来のものと比べ、男女話者及び各種文の平均で2.5ポイント向上し、明瞭感が大幅に向上するとともにソフトでより肉声に近い音質になったとの評価が被験者から得られた。

5. おわりに

合成音の歪みを最小化する規範に基づく音声素片の学習法と学習で得られた音声素片のLPC分析で得られる残差信号を駆動信号とする分析残差駆動合成器を提案した。提案した学習法は、従来問題となっていた韻律変更に伴う歪みも含めて合成音の歪みが最小となる素片を生成しているため、明瞭で肉声感豊かな音質を実現することができる。また、本合成器は音声素片をLPCパラメータと残差信号の組で表現しているため、スペクトルパラメータの操作により声質の変更が容易であると共に合成単位辞書のサイズを大幅に圧縮することができる。ベクトル-スカラ量子化の符号化手法を適用することにより、一話者当たり約150kbyteのコンパクトなサイズを実現しており、PDA等の携帯情報端末やカーナビ等への組み込みも容易である。符号化法については別途報告したい。

参考文献

[1]広川、箱田、佐藤: "波形編集型合成方式におけるスペクトル連続性を考慮した波形選択法"、音響論2-6-10 (1990-9).

[2]岩田、他:: "パソコン向けソフトウェア日本語テキスト音声合成"、音響論2-8-13 (1993-10).

[3]小山、小泉: "VCVを基本単位とする波形規則合成方式の検討"、信学技報SP96-8 (1996-5).

[4]N. Campbell and A. W. Black: "CHATR: 自然音声波形接続型任意音声合成システム"、信学技報SP96-7 (1996-5)

[5]F. Charpentier and M. Stella: "Diphone synthesis using an overlap-add technique for speech waveforms concatenation", Proc. ICASSP86, pp.2015-2018 (1986).

[6]T. Kagoshima and M. Akamine: "Automatic generation of speech synthesis units based on closed loop training", Proc. ICASSP97, pp.963-966, (1997-4).

[7]伊藤、佐藤:"切り出し残差を用いた音声合成におけるピッチ制御法"、音響論2-7-18 (1989-3).

[8]武田、浅川、市川: "残差音源型規則合成における女声音質改善方式の検討"、信学論A, vol.J73-A, No.4, pp.700-708 (1990-4).

[9]籠嶋、赤嶺: "高音質規則音声合成器のための有声音源生成法"、音響論2-4-1 (1996-3).

[10]中嶌、浜田: "音韻環境に基づくクラスタリングによる規則合成法"、信学論D-II, vol.J72-D-II, No.8, pp.1174-1179 (1989-8).