

声帯振動の非線形効果に基づく音声分析合成システム

Speech Analysis and Synthesis System Based on the Nonlinear Energy Damping Model

大村 浩 田中和世
H. OHMURA K. TANAKA

It is considered that the vocal folds vibration affects the vocal tract transfer characteristics through nonlinear time-varying interaction between the glottis and vocal tract. In this paper, we investigate this nonlinear effect on the experimental framework of a speech analysis and synthesis system where smooth movement of vocal folds vibration is assumed for computational feasibility of the modeling. We first analyze the relationship between the glottal cross section and the energy damping patterns of formants that appeared in individual component wave forms of formant and show its nonlinearity. Based on these experimental results, we propose a speech synthesis technique called Nonlinear Energy Damping(NED) wave function model in which formant energy damping is given by a time window function. We confirm its performance by a preference test on speech wave reconstruction. Evaluators listen two kinds of reconstructed speech sentence samples: one is synthesized by the proposed method and the other is by an ordinary LPC-based method. Results show that the proposed method is clearly superior in its quality.

§ 1 はじめに

従来の音声処理系の多くは、現実的な立場から、音声生成系を線形システムと仮定し、これを相互に独立な2つの要素、すなわち、声帯振動に基づく音源生成系とpassive-linearな声道伝送系によって構成してきた。しかし、理論的に見て、声帯の振動は声門体積流の生成に関与すると共に、声門インピーダンスの時間変化を通して声道特性にも非線形な影響を与えていることが考えられる。このような立場から、Teagerらは、分析結果に基づいて線形系という仮定の矛盾について述べている¹⁾。Rothenbergらは、声帯波抽出プロセスのなかで声帯の振動サイクルの開大区間でホルマントエネルギーの急激な減衰を指摘し²⁾、その非線形な影響について述べている³⁾。さらに、Cairnsらは、このような非線形の影響を利用して、第一ホルマント波形の包絡形状とphonationタイプとの関連を検討している⁴⁾。

合成的アプローチとしては、これを音声の自然性、個人性に深く関与する問題と捉えて、声帯の自励振機構と

声道が相互作用する音声生成システムに関する石坂⁵⁾や池田⁶⁾の研究がある。また、Maragosらは、音声波形をdamped AM-FM signalと考えてその分析法を提案しており^{7,8)}、このような声帯振動の非線形効果が注目されている。

音声合成において、パラメータによる声質(Voice Quality)のコントロールを実現するためには、声道特性-音源特性のモデルを如何に構築するかが重要な課題である。我々は、このような声帯振動による非線形効果を、音声分析-合成の枠組みをとおして十分検討する必要があると考え、実音声における声帯振動の非線形効果を観察し、これによる音声合成方式を検討した。Teager等の音声の非線形現象¹⁾やMaragos等のAM-FM信号としての音声の生成過程には声帯振動以外の要素が含まれている。我々は、問題を声帯振動に限定し、分析合成系で取り扱える複雑さという現実的な立場から、従来の線形声道システムの一つの拡張として、声帯の比較的滑らかな運動を含む音声生成系を想定した。これによって、その効果を確認し、音声規則合成システムに導入していくこ

KEY WORDS: 声帯振動, 波形包絡, 非線形効果, 音声分析合成, 窓関数モデル

とを目標としている。

本文の前半では、声門開口断面積とホルマントのエネルギー減衰について考察し、実音声におけるホルマントのエネルギー時間パターンを抽出するための分析システムの構成について述べる。ここでは、声帯振動の非線形効果が声道伝達特性にも現れていることを観察する。次に、シミュレーションによって、このような分析結果を音声合成に反映させるための非線形エネルギー減衰モデルによる音声合成方式を検討する。後半では、このモデルによる音声分析-合成実験システムを構成し、聴取による評価実験結果を示し、モデルの有効性について考察する。

§ 2 声門開口面積とホルマントのエネルギー減衰

声帯振動の影響は、音源特性と声道伝達特性に現れると考えられる。ここでは、線形システムの上で声門開口面積とホルマントのエネルギー減衰の関係を考えてみる。いま、ホルマント周波数 F_i での応答波形 $s_i(t)$ を $a_i \exp(-\sigma_i t) \sin(2\pi F_i t)$ とし、音声波形 $s(t)$ を $s(t) = \sum s_i(t)$ と表す。

声帯振動の様相は、声帯波の形状を変え、音声スペクトルの概形的な特性として現われる。これは各ホルマント波形の振幅レベル a_i を決定している。一方、声道特性への影響は、ホルマント周波数とその減衰特性に現れると考えられる。声門部のインピーダンスが声道内部に比べて十分高いと仮定すれば、ホルマントは声帯振動に対してほぼ一定であろう。これに対し、声帯振動による声門インピーダンスの変化は、主に減衰パラメータ σ_i に現れることが推測される。

Fig.1は、声道を17個の異なる断面積値を持つ直円筒で近似し、声門開口断面積 A_g とこれに隣接する声道開口断面積 A_v の比を変えて声道伝達関数を求め、その極から計算された5母音の $\{F_1, \sigma_1\}$ の変化パターンである。断面積比 A_g/A_v に対して、 F_1 が比較的一定であるのに対して、 σ_1 には大きな変化が見られる。 $\{F_2, \sigma_2\}, \{F_3, \sigma_3\}$ についても同様の傾向が見られた。したがって、声帯の動きがホルマントの共振波形に比べてゆっくり変化すると仮定すれば、ピッチ周期内のエネルギー減衰パターンを表す σ_i の変化を観察することによって、このような A_g の効果を推計することが考えられる。

§ 3 分析システム

3.1 システムの概要

ここでは、ホルマント周波数 F_i に対応する波形を、 $s_i(t) =$

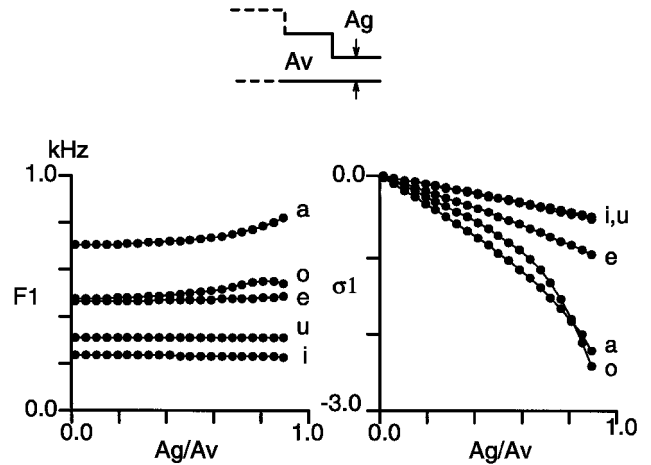


Fig.1 The first formant F_1 and attenuation parameter σ_1 as a function of glottal opening A_g , for 5 vowels.

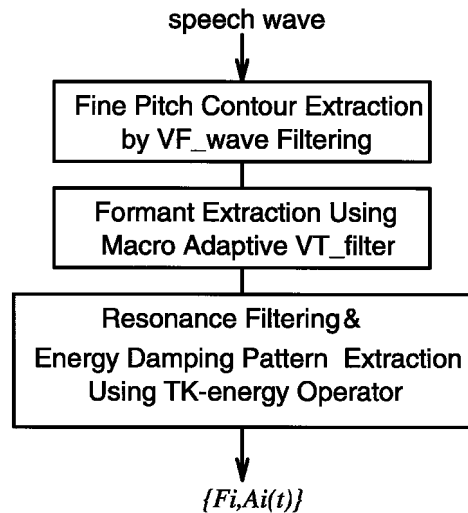


Fig.2 Block diagram for extracting formant F_i and its resonance wave envelope $A_i(t)$.

$A_i \sin(2\pi F_i t)$ と表し、実音声から F_i およびそのエネルギー減衰パターン $A_i(t)$ の抽出を行う分析システムについて述べる。

Fig.2は、音声波形を入力しピッチ同期分析によって F_i と $A_i(t)$ を出力する分析システムのブロック図である。最初に、基本波フィルタリングによって精細ピッチパターンの抽出⁹⁾を行い、次に適応化声道フィルタを用いてピッチ周期ごとにホルマント周波数 F_i の推定を行う¹⁰⁾。最後のブロックで、ホルマント周波数を中心周波数とする帯域通過型FIRフィルタ(ホルマントフィルタ)を用いて各ホルマントに対応する波形を抽出し、energy operator²³⁾によって $A_i(t)$ を計算する。その具体的な手法を次に示す。

3.2 マクロ適応化声道フィルタによるホルマント推定線形予測分析法によってホルマント抽出を行う場合、

予測モデルの次数 M をどのように決めるかが本質的な問題となってくる。標準化周波数 f_s との関係から、 $M = f_s +$ と与える方法¹¹⁾や、声道形推定の立場から、情報量基準を用いてその長さを推定する方法¹²⁾などが提案されている。ホルマントと声道長については、低次のホルマントからの声道長推定^{13,14)}やホルマント正規化¹⁵⁾等の研究がある。

ホルマントは声道形状によって変化し、また、声道長とホルマントは相互依存的であり、分析においては一方を先に（或いは同時に）決めるという構造が常に存在する。予測モデルの係数の長さ(LPCフィルタの長さ)は、音声生成モデルにおける精密な声道長というよりは、与えられた音声のスペクトルを記述するための平均的で、概ね妥当な長さとして解釈できる。そこで、新たに前述の研究では特に触れられていない基本周波数と声道長の関係を導入した。

基本周波数は、音韻性を決定する重要な因子であることが指摘されており^{16,17)}、また、話者や年齢に対するピッチのマクロ指標としての重要性が述べられている¹⁸⁾。さらに、音響モデルによる音声合成聴取実験によって、母音声質の観点から、声道長とピッチ周波数の間に一定の関係が与えられている¹⁹⁾。また、ホルマント抽出において声道長(LPCフィルタの長さ)の推定が効果的であることが報告されている²⁰⁾。このような観点から、我々は、音声生成過程でのピッチ、声道長、ホルマントの関連構造を積極的に利用したホルマント推定法を提案した²¹⁾。その手法は、入力された1音声サンプルごとに、声道フィルタの次数をマクロに推定し、その極を計算することによってホルマントを求めるものである。

始めに、声道長-平均基本周波数モデルを考える。基本周波数を f_0 とし、声道長 L から計算されるuniform-tubeの第1ホルマントを f_1 とするとき、梅田らは¹⁹⁾、声道の長さ L と基本周波数の関係を次の式で与えている。

$$f_1 = k f_0 \quad (1)$$

最も適当な合成音を得られるのは、 $k=3$ のときで、 $2.5 < k < 5.0$ で自然な声を得られること、成人男性の声では k は高めがよく、子供では低めにとるのがよいと報告されている。(1)式を声道長 L と f_0 の関係に書き換えると、

$$L = \frac{c}{4k f_0} \quad (2)$$

となる。ただし、 c は音速である。上式の f_0 は、本来その話者が最も発声し易いピッチ modal vocal frequency²²⁾ とすべきであるが、ここではその第一次近似として、一発話の平均ピッチ \bar{f}_0 に置き換えて考える。

次に、音声データから、この \bar{f}_0 に対する L の関係を実験的に求める。声道長の推定実験の手順は、一つの発話

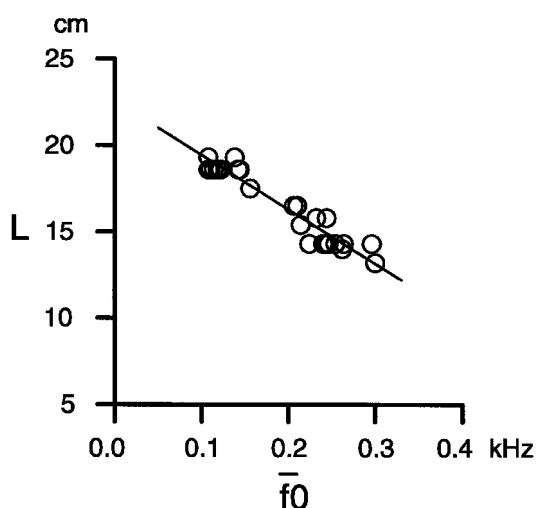


Fig.3 An experimental relation between average fundamental frequency \bar{f}_0 and vocal tract length L . The line is Eq.(4).

におけるピッチパターンを求めその平均ピッチを \bar{f}_0 とし、 $k=3$ と置いて(2)式から L を求め、 $m=2f_s L/c$ とする。伝達関数が $H(z^{-1})=1/(1-z^{-1})$ で表わされる適応一次フィルタで逆フィルタリングした音声波形をピッチ同期LPC分析し、声道フィルタの次数を $m-2 \sim m+2$ の範囲で変化させ、出力されるホルマントパターンの連続性を視察によって観察して M の値を決定する。声道長は、

$$L = \frac{cM}{2f_s} \quad (3)$$

によって計算する。

Fig.3の印は、このようにして抽出された平均基本周波数 \bar{f}_0 に対する声道長 L である。音声資料は、音響学会研究用連続音声データベースの内の1文、男10名、女10名(16kHzサンプリング)と母音連続、男2名、女3名(10kHzサンプリング)である。LPC分析は、有声音区間について行なった。この部分には鼻音、有声破裂音、有声の/h/などが含まれているがそのまま分析した。このデータに次式の関数モデルをあてはめる。

$$L(\bar{f}_0) = a_1 \bar{f}_0 + a_0 \quad (4)$$

パラメータ (a_1, a_0) は、 $(-31.3, 22.5)$ と推定された。同図の直線は(4)式によって計算したものである。

提案するホルマント抽出法は、発話の平均ピッチ \bar{f}_0 と(4)式から声道フィルタの長さ M を次式によって予測し、

$$M = \frac{2f_s L(\bar{f}_0)}{c} \quad (5)$$

M 次の声道フィルタの特性方程式を解き、求められた解を低いほうから順番にホルマントとするものである。ホルマント抽出に限らず、一般にLPC分析を行う場合、

Table 1 Average fundamental frequency \bar{F}_0 and estimated vocal tract filter length M for speech data of A-set in the Continuous Speech Corpus for Research²⁸⁾.

Spk	\bar{F}_0^*	σ^*	M		
			min	mostly	max
1	0.112	0.0039	17	17	18
2	0.116	0.0049	17	17	18
3	0.116	0.0048	17	17	17
4	0.121	0.0037	17	17	17
5	0.131	0.0071	16	17	17
6	0.132	0.0042	16	17	17
7	0.137	0.0042	16	17	17
8	0.144	0.0047	16	16	17
9	0.146	0.0039	16	16	17
10	0.163	0.0051	16	16	16
11	0.206	0.0052	14	15	15
12	0.213	0.0080	14	14	15
13	0.225	0.0070	13	14	15
14	0.233	0.0063	14	14	14
15	0.246	0.0103	13	14	14
16	0.255	0.0095	13	13	14
17	0.259	0.0069	13	13	14
18	0.268	0.0088	12	13	14
19	0.309	0.0116	11	12	13
20	0.322	0.0109	11	11	12

* in kHz

その項数をどう決めるかが曖昧である。この手法は、ad hocに陥りやすいホルマントピーク選択法を避けて、予め当てはめるべきシステムの適応化を図りホルマントを抽出するものである。

Table 1 は、モデル $M(\bar{f}_0)$ を音響学会研究用連続音声データベースのA-set(話者20名、各50文)に適用した場合の M の値である。話者(Spk)の番号1~10は男

性、11~20は女性である。 \bar{F}_0 は、各文の平均ピッチを50文について再び平均した値、 σ はその標準偏差である。個々の話者についてみると、 M の変動は $M = \pm 1$ の範囲内に収まっており、最もよく現われる M (mostly)の範囲は、話者全体で見ると、 $11 \leq M \leq 17$ の範囲にあった。

このような M の値は、声道フィルタ長のおおよその見当であり、実際の適用では、ホルマント軌跡の連続性を高めるために、声道フィルタの終端条件として偏相関係数 (ρ) が、 $\rho < 0$ であることが望ましい。そこで、音声入力サンプルの有声音部分の平均偏相関係数 $(\bar{\rho}_M)$ が、 $M(\bar{f}_0) - 2 \leq M(\bar{f}_0) \leq M(\bar{f}_0) + 2$ の範囲で、最小となる M の値を声道フィルタ長としている。

$$M = i : \min_{M(\bar{f}_0) - 2 \leq i \leq M(\bar{f}_0) + 2} \bar{\rho}_i \quad (6)$$

Fig.4は、成人女性話者の発声した文章サンプル「あらゆる現実を全て自分の方えねじ曲げたのだ」のホルマント抽出例である。平均基本周波数は267Hzであり、 $M=12$ と推定された。

3.3 ホルマントエネルギー減衰パターンの抽出

まず、与えられたホルマント周波数 F_i を中心周波数とする帯域通過型FIRフィルタ(ホルマントフィルタ) $w(t)$ を、時間窓を $w_m(t)$ として、次式で与える。

$$w(t) = w_m(t) \cos(2\pi F_i t) \quad (7)$$

ここで、 $w_m(t)$ は、時間窓長を $2T_m$ とする、次のBlackman関数である。

$$w(t) = 0.42 + 0.50 \cos \frac{\pi t}{T_m} + 0.08 \cos \frac{2\pi t}{T_m} \quad (8)$$

時間窓長 $2T_m$ は、ピッチ周期以下の変化を抽出することを考慮して、ピッチ周期 $(1/f_0 : f_0$ はピッチ周波数)を定数とし、その倍率を表わす α を変数とする次式で与える。

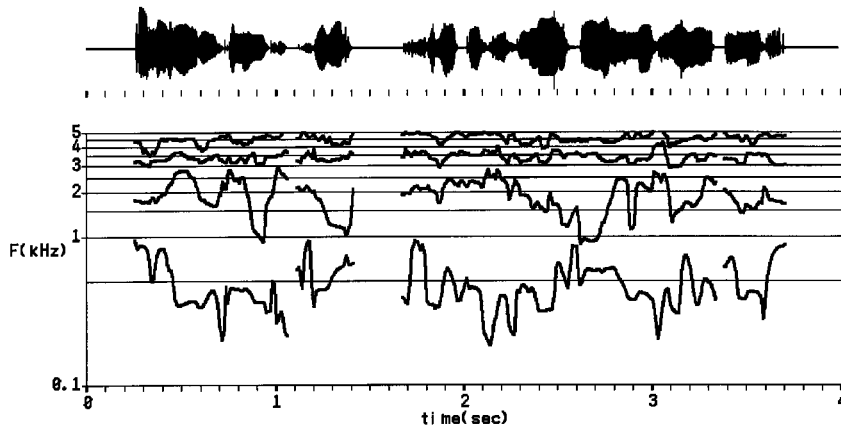


Fig.4 Examples of extracted formant contour for a speech sample, /ARAYURU GEN'JITSUO SUBETE JIBUN'NO HO-ENEJIMAGETANODA/, uttered by a female.

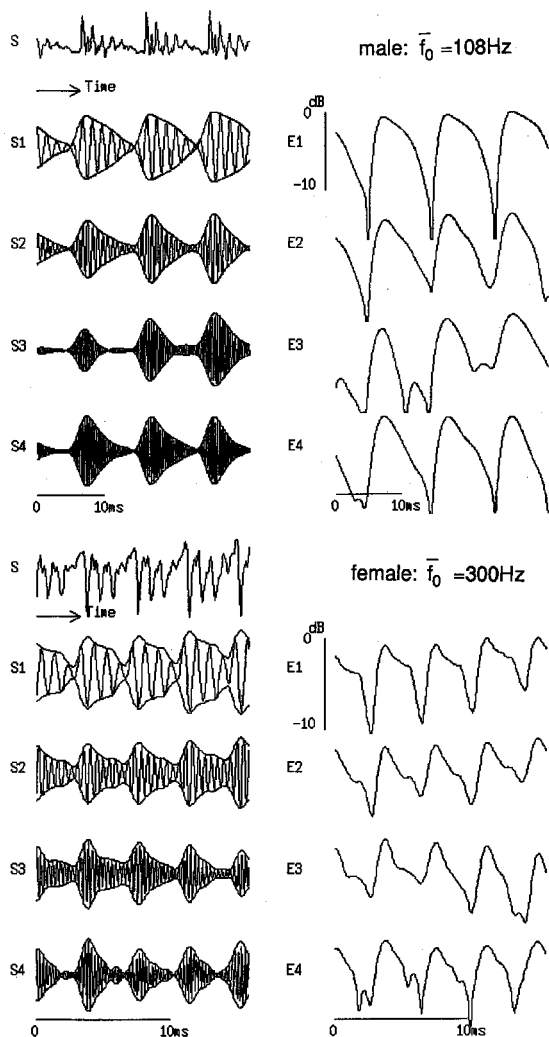


Fig.5 Examples of the resonance wave envelope extracted from a male speaker (the upper) and a female speaker (the lower) in utterance /a/. S : speech wave, S_i : resonance wave corresponding to F_i , E_i : log envelope of S_i , and f_0 : averaged fundamental frequency.

$$2T_m = \frac{1}{f_0} \quad (9)$$

次に、抽出されたホルマント波形のエネルギー減衰パターン（波形包絡）の抽出法について述べる。 F_i に対応するホルマント波形のサンプル値系列を $s_{i,j}$ (j は時間軸を表す)とすると、アルゴリズムは、エネルギーオペレータ²³⁾を用いて次式によって与えられる。

$$e_{i,j} = s_{i,j}^2 - s_{i,j-1} s_{i,j+1} \quad (10)$$

$$A_{i,j} = \frac{\sqrt{e_{i,j}}}{i}$$

$e_{i,j}$ と $A_{i,j}$ は、それぞれ、同時刻の瞬時エネルギーと波形包絡のサンプル値である。パラメータ i は、サンプリング周波数 f_s を用いると $i = \sin(2\pi F_i/f_s)$ である¹⁰⁾。

3.4 分析実験

ホルマントフィルタのパラメータ α は、隣接するホルマント波形成分を除去し、また、滑らかな波形包絡抽出のためには大であることが望ましく、ピッチ周期以下の微細な変化を抽出し、また、波形近似を良くするためには小であることが望ましい。そこで次のように、平均ピッチ107Hzの男性と330Hzの女性の音声資料（各1文）を用いて実験的に α の値を設定した。視察による観察で波形包絡が比較的滑らかで、かつ原音声波形とフィルタリングされた音声波形（各ホルマントに対応する波形成分の和）との間の相互相関係数が、一つの目安として0.95以上となるように α の値を調節する。これを、ピッチ周期ごとに繰り返した結果、 α の範囲は、おおよそ0.4

0.8となった。本報告では、波形包絡の滑らかさを重視して $\alpha = 0.75$ と設定した。

Fig.5は、平均ピッチが108Hzの男性と300Hzの女性の母音/a/の波形包絡の例である。左列は、音声波形 S とホルマント F_i に対応する波形成分 S_i を、右列はその対数波形包絡パターン $E_i(t) = 20 \log A_i(t)$ を示している。特に第1ホルマントの対数波形包絡パターン E_1 では直線的減衰からかなり外れていることが読みとれる。声帯振動の影響は、声帯波のみならず声道の伝達特性にも及んでいることがわかる。

§4 声帯振動の非線形効果を考慮した音声合成モデル

非線形効果を詳細に検討するために二つのシミュレーションを行ない、非線形回路網と時間窓関数による合成方式を提案し、波形再合成実験を行った。

4.1 非線形モデルによるシミュレーション

はじめに、最も単純なエネルギー損失項を含む2次系の運動方程式を用いて、非線形システム応答に対する線形予測分析について考えてみる¹⁰⁾。

運動方程式は、

$$\ddot{x} + K\dot{x} + \frac{2}{\theta}x = 0 \quad (11)$$

ここで、 θ_0 は $K=0$ における系の共振角周波数である。いま、 $K=K(t)$ としてピッチ周期内でのエネルギー損失の時間変化を与える。これを、離散値系で扱うために、次の後退差分方程式によってシミュレーションする。

$$x_{j+1} = (2 - K_j T - \frac{2}{\theta_0} T^2)x_j - (1 - K_j T)x_{j-1} \quad (12)$$

$T=1/f_s$ で f_s はサンプリング周波数、また、 x_j と K_j はそれ

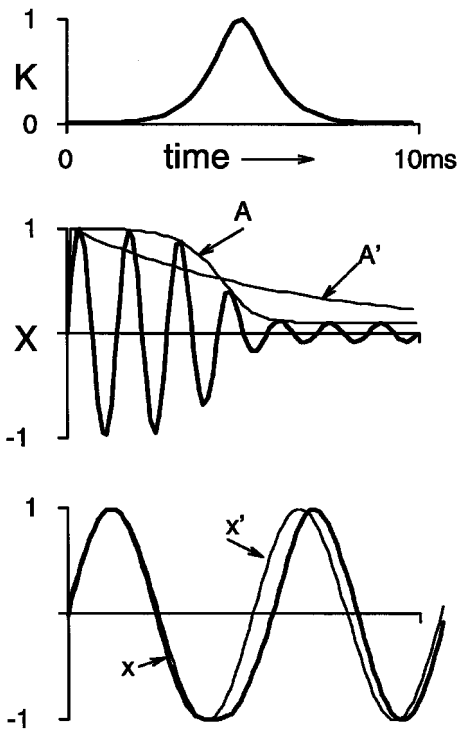


Fig.6 Responses of the second order nonlinear system indicated in Eq.(11). K is a given time pattern for $K(t)$. X is the nonlinear response at $F=700\text{Hz}$ and A is its envelope. Envelope A' is generated by the second order linear system identified by LP analysis of X . The bottom shows comparison of amplitude normalized responses: x for the nonlinear system and x' for the linear system at $F=200\text{Hz}$.

それ(11)式の自由振動波形 $x(t)$ と系のエネルギー減衰特性を決定する時変パラメータ $K(t)$ の時刻 $t=jT$ における標本値である。

Fig.6は、 K に時間変化を与え、 $\omega_0=2\pi F$, $F=700\text{Hz}$ とし、(12)式で合成した波形 $x(t)$ を、2次の線形予測フィルタで分析した例である。原波形の包絡 A に対してLPC-filterの応答 A' は K を一定として与えられる。一方、非線形システムの共振周波数は、 K の時間変化に伴って瞬時変動(下降)し、また K がすべての共振周波数について同じ値とすれば、高域ホルマントに比べて低域ホルマントに対する影響が大きい。しかし、同図3段目の低域ホルマント($F=200\text{Hz}$)における、振幅正規化した非線形応答 x とLPC-filterの応答 x' の例では、ホルマントの瞬時変動が大きくなる時間帯は、同時に非線形応答波形 X の振幅 A が急激に減衰する部分でもある。 x と x' を観察すると、線形予測分析によって推定されたホルマントの周期は、原波形振幅(X)の高い区間のホルマント周期に対応していることがわかる。波形エネルギーの高い部分での正確な記述が重要であるという観点にたてば、LPC分析によるホルマント推定は、ほぼ妥当なものであろう。これは、

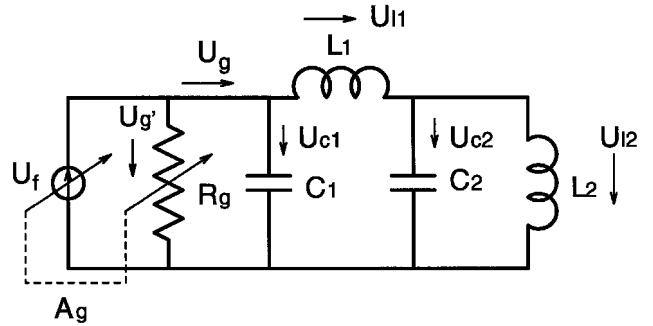


Fig.7 The fourth order nonlinear circuit model for the speech synthesis simulating the vocal folds vibration effect.

ピッチ周期内で、ホルマント周波数は一定とみなすことであり、従って、このような非線形応答 $x(t)$ に対する近似解として、ホルマント周波数は一定、波形包絡に対してはピッチ周期内での自由度を与えるモデル、 $x(t) = A(t)\sin(\omega t)$ が考えられる。

次に、声帯振動 $A_g(t)$ の音声波形への効果をより詳細に見るために、声門部におけるNorton equivalent model³⁾を含む非線形声道システムによるシミュレーションを行った²⁴⁾。Fig.7は、Norton equivalent modelにおけるインダクタンス L_g を省略し、二つのホルマントを持つ声帯音源/声道回路モデルである。電流源 $U_f(t)$ は、声門開口断面積 $A_g(t)$ に比例し、声門抵抗 $R_g(t)$ は、反比例するように制御される。同図の L_2 を流れる体積流 U_{l2} を x とすると、その微分方程式は、

$$\ddot{x} + \frac{1}{T_1} \dot{x} + (\omega_1^2 + \frac{1}{2}\omega_2^2)x + \frac{1}{2}\omega_2^2 x = \frac{\omega_1^2 \omega_2^2}{R_g} P_g \quad (13)$$

ここで、 $T_1 = C_1 R_g(t)$, $\omega_2 = 1 / (\omega_1^2 + \frac{1}{2}\omega_2^2 - 1/2)$, $i=2, F_i, P_g/R_g(t) = U_f(t)$ である。 R_g に2種類の時間パターンを与えて音声合成実験を行った。Fig.8にその結果を示す。左上の A_g は、振幅を1に正規化した声門断面積波である。波形 W_v は、 A_g によって R_g を制御した場合の応答(放射特性に対応して微分した値)、 S_{pv} はそのスペクトルである。一方、 A_g の平均値(同図左上の横線)を与えて、 R_g 一定(線形系)とした場合の波形とスペクトルが、 W_c と S_{pc} である。ホルマント F_1, F_2 は、それぞれ $0.75\text{kHz}, 1.25\text{kHz}$ である。両者とも、声帯音源 U_f はピッチ周期で変化する同じ A_g パターンで制御している。

この図からも分かるように、両者の波形では、エネルギー減衰の違いが観察される。また、周波数スペクトルにはホルマントの強さと帯域幅に大きな差があり、平均的減衰が同一であっても声帯が振動している場合、ホルマント情報がよりよく保たれていることがわかる。

4.2 音声合成方式

このような声帯振動の影響を考慮した 幾つかの合成方

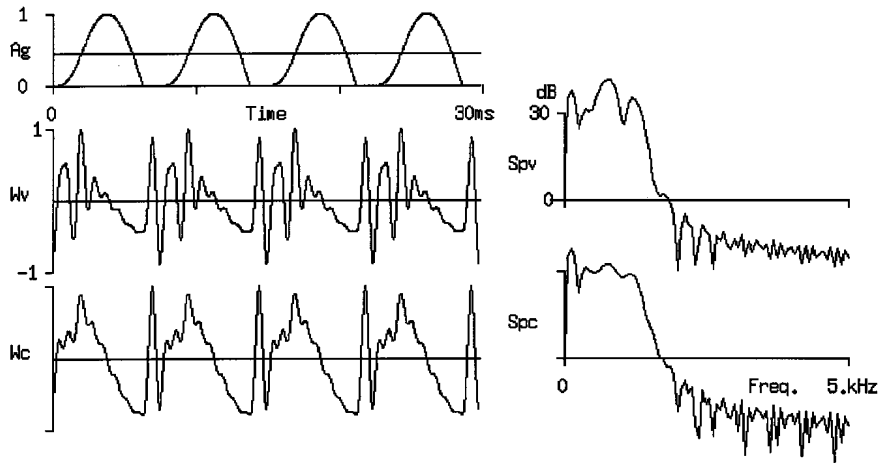


Fig.8 Time and frequency responses of the nonlinear circuit shown in Fig.7. A_g is a given glottal area wave, W_{vis} the time response and W_c is the one for the glottal area wave kept in a constant shown by a horizontal line in the top of the figure, and S_{pv} and S_{pc} are their spectra.

式が考えられる。最も精密なシミュレーションとしては、声帯の自励振機構そのものを模擬する石坂の Two-mass model⁵⁾を含む合成系や声帯形状をより精密に模擬した池田の流体力学に基づく音声生成モデル⁶⁾がある。また、声門部を簡略化して、声道線形システムの単純な拡張として声門を瞬時コントロールしながら、声門下部との結合を考えた、Fig.9に示す音声生成モデル¹⁰⁾も考えられる。

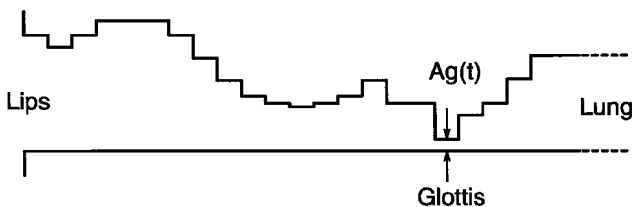


Fig.9 A simplified vocal tract system with time varying glottal area function.

分析-合成系を通して考えた場合、このような高次非線形モデルは極めて魅力的ではあるが、これに供するパラメータの抽出が困難であり、そのモデル自身が研究対象となっている。そこで現実的なアプローチとして、声帯振動に対応する時間変化素子を含み、一組のホルマントと、一定の仮定の基に計算される回路モデル (Fig.7) が挙げられる。これを実現するためには、基本的にスペクトルの概形とホルマント波形の波形包絡の概形を同時に満足させる声門開口断面積波の同定が必要である。そこで、これらの要素をホルマントごとに比較的自由に扱える2次系モデル (11) 式の並列接続で実現する方法が考えられる。さらに分析方式に直結させたアプローチとして、非線形応答に対する近似的記述モデルがある。この場合、ホルマント周波数 F_i はピッチ周期内で一定で、その波形包絡 $A_i(t)$ のみを制御するホルマント関数型合成方

式である。その形式は、

$$x_i(t) = a_i \exp(K_i(t)) \sin(2 F_i t + \phi_i) \quad (14)$$

である。ここで、 a_i はホルマント波形の振幅の大きさ、 $\exp(K_i(t))$ は波形包絡の形状を与える。 ϕ_i は相対的な位相差である。従って、このモデルのホルマントのエネルギー減衰パターン (波形包絡) は、時間窓関数 $\exp(K_i(t))$ によって制御されることになる。音声波形は、ホルマントの数を N として、これらの和 $x(t) = \sum_{i=0}^N x_i(t)$ で与えられる。ここで、周波数 F_0 は、 f_0 F_0 F_1 の範囲にあり、女性などの高ピッチでは、 $F_0 = f_0$ である。これは、声帯波の基本波成分⁹⁾と考えられ、音質を考える上では重要な要素である。

4.3 波形再合成実験

Fig.10 は、分析再合成実験システムである。左図の分析ブロックは図2に対応し、右図は音声合成ブロックである。抽出された基本周波数 f_0 、声帯基本波とホルマントの周波数 F_i ($i=0,1,2,3,4$)、各周波数成分の振幅レベル a_i および対数エネルギー減衰パターン $K_i(t)$ を用いて1ピッチの波形再合成を行った。音声波形は、(14)式の各

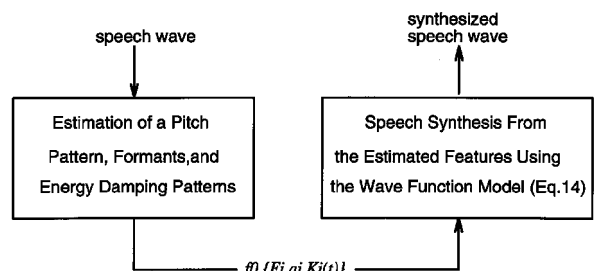


Fig.10 The speech analysis and synthesis system using the wave function model.

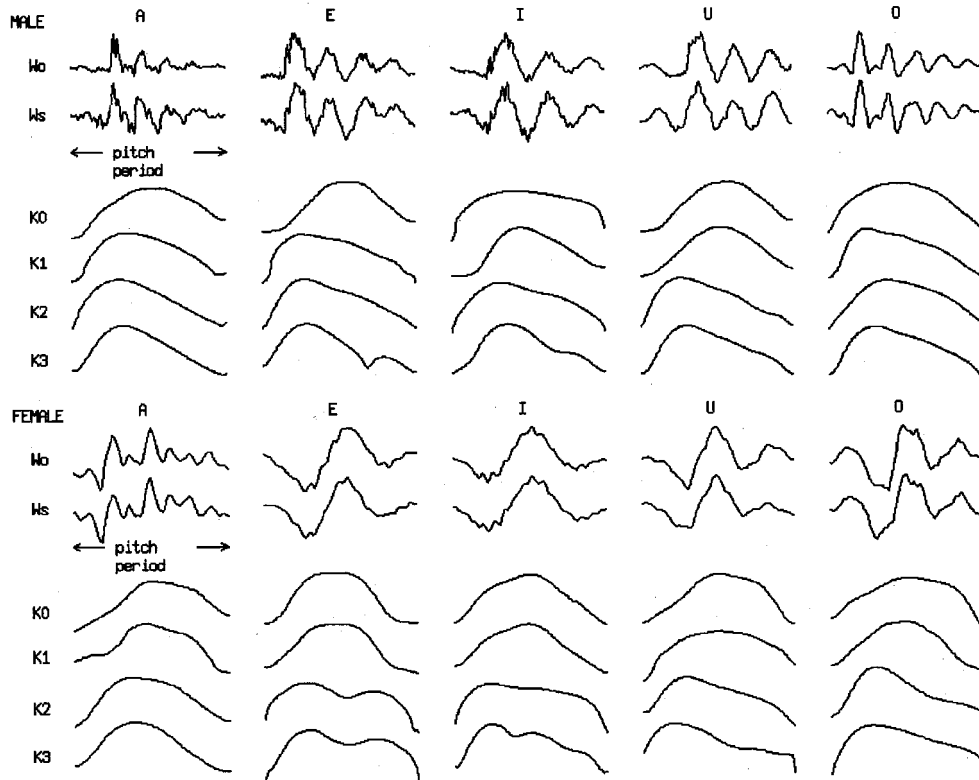


Fig.11 Original and synthesized wave forms, W_o and W_s respectively, for 5 vowels uttered by a male and a female speakers. K_i is the logarithmic energy damping pattern of resonance F_i .

ホルマント波形の和で与えられる。分析によれば、各ホルマントの K_i パターンは、励起-保持-減衰のサイクルがあり、ホルマント間でほぼ同期するが若干のずれも観察される。特に、 K_0 は、第1ホルマントが比較的高い/a/などの場合、 K_1 に比べて、およそ半ピッチ程度シフトしている。 F_1 の低下する母音や有声子音では声帯波が重畳し、見かけ上 K_1 パターンがフラット（減衰なし）となることもある。この場合でも K_2, K_3 では、声帯の振動に対応した変化が観察される。 K_4 以上では一般に変化は不規則になり易く、摩擦性成分の存在や信号レベルの低下などによる分析精度の問題が考えられる。

Fig.11の上段は、音響学会研究用データベースのA-set中の男性話者(平均ピッチ111Hz)、下段は女性話者(平均ピッチ213Hz)の分析再合成の例である。 W_o は原波形、 W_s は(14)式の和による再合成波形、 K_0, K_1, K_2, K_3 は位相はそろえ、振幅は正規化して表示している。いま、2つの波形 W_s, W_o の時系列を、それぞれ、 $\{o_0, o_1, \dots, o_{J-1}\}, \{s_0, s_1, \dots, s_{J-1}\}$ (J は1ピッチ周期における波形サンプル点の数)として相互相関関数を $r_{o,s}(k) = \frac{1}{J} \sum_{j=0}^{J-1-k} (o_j - \bar{o})(s_{j+k} - \bar{s})$ と定義する。ただし、 \bar{o}, \bar{s} はそれぞれの平均値、 σ_o, σ_s は標準偏差である。このとき、 W_o と W_s との相互相関係数、 $r_{o,s}(0)$ は概ね0.8以上となり、関数モデルは非線形波形応答の近似として十分機能していると考えられる。

4.4 ホルマントエネルギー減衰の窓関数モデル

より柔軟で自由な音声処理系を考えた場合、音声のパラメータ化は重要であり、上で述べたようなホルマントのエネルギー減衰パターンに対するモデルが必要となる。理論的には、声門開口断面積波との関連を基礎に考えるべきであるが、ホルマントエネルギー減衰の非線形効果を試すための第一段階として、次のようなモデルを考えた。これは、線形なエネルギー減衰を与える項と非線形な減衰を与える項を持つモデルであり、次式で与えられる。

$$K_i(t) = k_{0,i} + k_{1,i}(t) \quad (15)$$

波形包絡は、 $\exp(K_i(t))$ によって計算される。ここで $k_{0,i}$ は、Dunnのホルマント-帯域幅測定データ²⁵⁾に基づき、与えられたホルマントから帯域幅を与える関数モデル (Fig.12)を用いて設定した。具体的には、ホルマント F_i としてFig.12の関数から、その帯域幅 $B(F_i)$ を計算し、 $k_{0,i} = B(F_i)/f_s$ と与える。非線形項関数 $\exp(k_{1,i}(t))$ は、全ホルマント共通のパターンを使用し、Fig.13の上段に示すように立ち上がりと下がり二つの窓関数 (Hanning型)で表わすモデルを設定した。同図 T_0 はピッチ周期、 T_2, T_1 は、それぞれ立ち上がり終了時刻、下がり開始時刻である。それぞれ T_0 に対する割合で設定される。同図

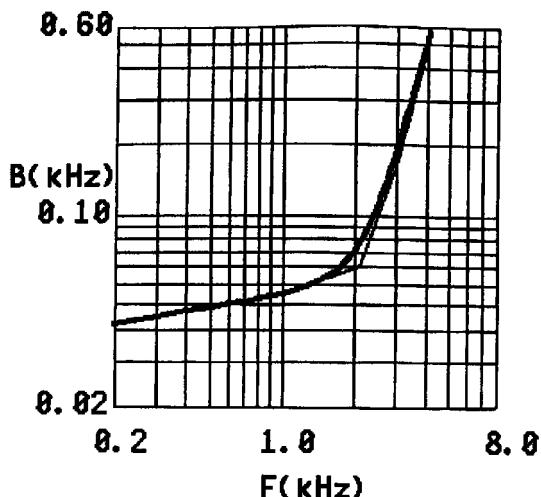


Fig.12 Relationship between formant F and its bandwidth B derived from Dunn's data²⁵⁾.

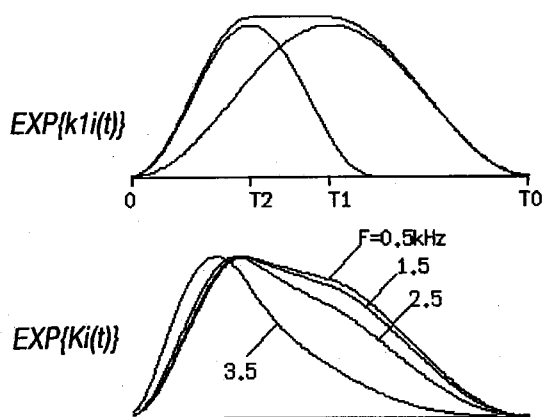


Fig.13 The window function model for nonlinear energy damping of resonance (see Eq.(15)).

下は、窓関数モデル(15)式による波形包絡 $\exp(K_i(t))$ のホルマント周波数 0.5, 1.5, 2.5, 3.5kHz に対する形状を示す。周波数が低いほど非線形性を表わす。

§ 5 聴取実験による評価

5.1 実験方法

このようなホルマント波形の非線形エネルギー減衰モデルの効果を確かめるために、線形システムとして最も一般的な線形予測法による合成音声との比較評価実験を行った²⁶⁾。まず、Table 2 にそれらの合成に用いたパラメータを示す。ここで、NED (Nonlinear Energy Damping Model) 法が提案法、LPF (Linear Predictive Filter) 法が線形予測フィルターによる方法である。

NED法では、Fig.10に示したように、原音声波形から、ピッチ周波数、ホルマント、その振幅が推定され、これらのパラメータとエネルギー減衰の窓関数モデルによっ

Table 2 Parameters of the analysis-synthesis experiments.

Information	NED	LPF
Pitch	f_0	f_0
Intensity	a_i	p
Spectrum	F_i	α_i
Energy Damping	$K_i(t)$	-
Length	$i = 0, 1, \dots, N$	$i = 1, 2, \dots, 20$

て、4.2節に述べた方法で再合成される。一方、LPF法では、線形予測分析によって求められた予測係数を用いて再合成される。両者とも同じピッチパターンによってピッチ同期分析し、有声音区間ではインパルス音源を用いて、ピッチ同期で再合成する。無声子音区間の音源はノイズを用いている。

スペクトル情報を表わすパラメータは、NEDでは F_0, F_1, \dots, F_N である。ここで、 N はホルマントの数であり、 $N=M/2$ と与えられる。また、 M は、3.2節で述べた方法で設定された。実験資料に対するホルマント数は $5 \sim 9$ の範囲で与えられた。

これに対して、線形予測法の場合、予測係数の数を決定する一般的な手法はない。そこで予測モデルを、音声生成系を近似する全極モデルと考えれば、声道特性と音源特性を合わせて十分に表せる項数が必要であると考られる。本実験で推定した話者(4名)のホルマントの数は6から9であり、音源特性に対して2~3個程度の予測係数を仮定すれば、ほぼ十分な項数として20が得られる。また、Markel¹¹⁾の提唱している項数も f_s+4 であり、これらを参考にすると、予測係数の数として20が妥当であろう。これにより、LPFのパラメータは $1, 2, \dots, 20$ とした。したがって、両者の合成法での基本的な違いは、ホルマントに対応する波形の減衰特性が非線形(NED法)か線形(LP法)かである。なお、NEDの $\exp(K_i(t))$ のパラメータ $T_1/T_0, T_2/T_0$ は、男性話者に対しては{0.2,0.5}、女性話者には{0.5,0.5}と設定した。また、(14)式の位相項 $\theta_0, \theta_1, \dots, \theta_N$ はすべて0とした。

音声資料は、音響学会研究用連続音声データベース²⁷⁾中のATR音素バランス文から選択した15文であり、男性2名、女性2名の発声したもの(全60サンプル)である。文の長さは、約15~30モーラである。評価試験は、Fig.14のランダム刺激系列による対比較法である。予め編集録音された合成音のDATテープを静かな居室でスピーカから再生して音を提示し、NED法とLPF法の内どちらか一方を選ぶものである。被験者(評定者)に対しては、「自然音声に近い、聞きやすい」を評価基準とするよう指示した。

被験者は、音声研究の専門家でない6名(女性)で、

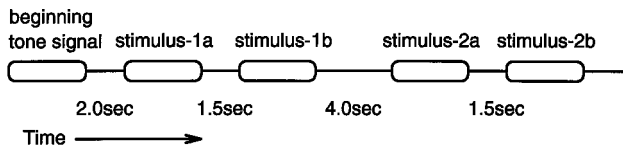


Fig.14 An example of the stimulus sequences.

全員がこの類の実験は経験したことがない。ただし、別の被験者への予備試験で、訓練の必要性が認められなかったため、被験者への訓練は行っていない。

5.2 実験結果

まず、Table 3に、4名の発声者（M1, M2, W1, W2）についての結果を示す。全体としては、ほぼ3：2の割合でNED法が優位である。ただし、話者W1（女性）については、後述するような理由から著しく逆の結果となっており、これを除くと、NED法の優位はより明確（7：3）になる。また、6名の被験者についての結果をTable 4に示す。一部の被験者（CD03, CD04）で評価が拮抗ないし逆転しているが、これは、上記の話者W1についての結果が影響しているためである。

Table 3 Preference scores for individual speakers.

	M1	M2	W1	W2
NED(%)	57	79	32	73
LPF(%)	43	21	68	27

Table 4 Preference scores for individual evaluators.

	CD01	CD02	CD03	CD04	CD05	CD06	in all
NED(%)	58	75	48	43	65	72	60
LPF(%)	42	25	52	57	35	28	40

話者W1のサンプルについては、聞こえの滑らかさに問題があるとの評価があった。これは、この話者のピッチが高く、抽出されたホルマントパターンの滑らかさに起因すると考えられる。したがって、その改善は分析側の問題に帰着される。ほとんどの被験者は、手法の違いによる音質の差異はかなりはっきりと知覚できると回答している。

§ 6 むすび

ホルマントのエネルギー減衰パターンを抽出し、理論的に推測される声帯振動の非線形な効果を確認した。声帯振動に対応した非線形回路網によるシミュレーションによって、線形システムに比べてホルマントの周波数スペクトル情報がよりよく保たれることが分かった。この

結果に基づき、これを実現するための音声合成法として、低次ホルマントによって構成される非線形回路網と非線形応答の近似表現である時間窓関数合成法を提案した。音声分析再合成システムによって、時間窓関数方式による波形近似の性能を確認した。この方式による再合成音声と線形予測合成法による音声をを用いた聴取評価実験の結果、提案法の有効性が示された。

これによりエネルギー減衰パターンを声帯振動に対応して非線形制御することによって合成音声の声質の改善を図ることができる。また、線形予測合成法や波形編集方式などと比べて、音声をホルマント周波数とその振幅レベルおよびエネルギー減衰パターン（波形包絡）によって記述しているために、話者特徴などの声質に関するより柔軟な処理や操作が可能であると考えられる。

今後、ホルマントのエネルギー減衰パターンの収集とモデルによるそれらの組織化、規則音声合成システムにおける話者性（声質）の制御、非線形音声生成モデルの検討を行う計画である。

謝 辞

本研究をご支援いただく当所知能情報部大津展之部長、並びに音声研究グループの皆様様に深謝します。

参 考 文 献

- 1) H.M.Teager and S.M. Teager, "Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract," *Speech Production and Speech Modeling* (Kluwer Academic, Netherland, 1990) 241-261.
- 2) M. Rothenberg, "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," *J. Acoust. Soc. Am.* 53 (1973) 1632-1645.
- 3) M. Rothenberg and S. Zahorian, "Nonlinear inverse filtering technique for estimation the glottal area waveform," *J. Acoust. Soc. Am.* 61, (1977) 1063-1071.
- 4) D. A. Cairns and J. H. L. Hansen, "Nonlinear Speech Analysis Using the Teager Energy Operator with Application to Speech Classification under Stress," *Proc. ICSLP 94*, s19-1.1 (1994).
- 5) 石坂謙三, ジェームズ L. フラナガン, "声帯音源の自励振動モデル," *日本音響学会誌* 34 (1978) 122-131.
- 6) 池田忠繁, 松崎雄嗣, "一次元非定常声門流れによる音声の生成," *日本機械学会論文集* 60-572B (1994) 1226-1233.

- 7) P. Maragos, T. F. Quatieri, F. Kaiser, "Speech Nonlinearities, Modulations, and Energy Operators," IEEE Proc. ICASSP91 S7.2 (1991) .
- 8) H.M. Hanson, P. Maragos, A. Potamianos, "Finding Speech Formant and Modulations via Energy Separation: With Application to a Vocoder," IEEE Proc. ICASSP93, (1993) 716-718.
- 9) 大村浩, 田中和世, "基本波フィルタリング法による精細ピッチパターンの抽出," 日本音響学会誌 51 (1995) 509-518.
- 10) 大村浩, "声帯振動による非線形性を考慮した振幅制御型音声合成方式," 信学技報 SP95-78 (1995) 39-46.
- 11) J.D. Markel, "Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation," IEEE Trans., Vol. AU-20, No.2 (1972) 129-137.
- 12) 石崎 俊, 中島隆之, "情報量を用いた声道長の推定," 音講論集 1-4-8 (1975.10).
- 13) A. Paige, V.W. Zue, "Calculation of Vocal Tract Length," IEEE Trans. Vol. AU-18, No.3 (1970) 268-270.
- 14) R.L. Kirlin, "A Posteriori Estimation of Vocal Tract Length," IEEE Trans. Vol. ASSP-26, No.6 (1978) 571-574, .
- 15) 田中和世, "音韻性を表わす特徴空間の構成とそのダイナミックモデル," 音響学会音声研究会資料S74-20 (1974-11).
- 16) R.K. Potter, J.C. Steinberg, "Toward the Specification of Speech," J.A.S.A., 22 (1950) 807-820.
- 17) 藤崎博也, 川島崇子, "母音の韻質に対するピッチおよび高次ホルマントの影響," 音講論集1-2-6 (1996, 11).
- 18) 粕谷英樹, 鈴木久喜, 城戸健一, "年令, 性別による日本語5母音のピッチ周波数とホルマント周波数の変化," 日本音響学会誌, Vol. 24, No.6 (1968) 355-364.
- 19) 梅田規子, 寺西立年, "声の韻質と声質," 日本音響学会誌22巻4号 (1965) 195-203.
- 20) 田中和世, 中島隆之, "適応的エンハンスフィルタによる声道長の推定とホルマント抽出," 音講論集1-2-24 (1974.6).
- 21) 大村浩, "基本周波数と声道長に関する分析的検討," 音講論集1-7-14 (1993,10).
- 22) R. Danikoff, G. Schuckers, L. Feth, The Physiology of Speech and Hearing, Prentice-Hall, Englewood Cliffs, (1980) 203.
- 23) J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," IEEE Proc. ICASSP90, s7.3 (1990).
- 24) H. Ohmura, K. Tanaka, "Speech Synthesis Using a Nonlinear Energy Damping Model for the Vocal Folds Vibration Effect," Proc. ICSLP96, (1996) 1241-1244.
- 25) J.L. Flanagan, Speech Analysis Synthesis and Perception (Springer-Verlag, New York, 1972) 182.
- 26) H. Ohmura, K. Tanaka, "Speech Synthesis Using a Nonlinear Modeling of Vocal Folds Vibration Effect," Acoust. Soc. Japan and Acoust. Soc. America 3rd Joint Meeting Proc., (1996) 1047-1050 .
- 27) 小林哲則, 板橋秀一, 速水悟, 竹沢寿幸, "日本音響学会研究用連続音声データベース," 日本音響学会誌 48 (1992) 888-893.

(1998.11.30受付)

著者紹介



大村 浩
Hiroshi OHMURA
知能情報部 音声信号処理ラボ
ohmura@etl.go.jp
音声情報処理, 特に音声生成システムに基づく分析, 合成の研究に従事。



田中和世
Kazuyo TANAKA
知能情報部 音声信号処理ラボ
ktanaka@etl.go.jp
音声分析, 音声認識・合成, 音声対話処理の研究に従事。