北海道大学工学部研究報告
第96号（昭和54年）

Bulletin of the Faculty of Engineering,
Hokkaido University, No. 96 (1979)

発声時声道伝達特性の簡易推定法

三上直樹* 大場良次* 井戸川 徹*,†

（昭和54年6月30日受理）

A Simple Method to Estimate Transfer Characteristics of Vocal Tracts under Utterance

Naoki MIKAMI, Ryoji OHBA and Tohru IDOGAWA

(Received June 30, 1979)

Abstract

The present paper proposes a simple method to estimate the transfer characteristic of a vocal tract under utterance based on only one pitch period of vowel waveform. A vowel waveform generally contains some triangular dips. They correspond to time intervals in which the vocal tract is forced to oscillate by glottal wave. We regarded them approximately as glottal waveforms. It is ascertained by a computer simulation that the transfer characteristic can be estimated by our method. A data processing system was constructed employing a minicomputer to estimate the transfer characteristic in real-time. Estimation tests were carried out by the system on Japanese vowels from male adults. Using several features of the estimated transfer characteristics, a vowel recognition experiment was also performed on 500 vowel samples from 100 male adults. A vowel classifier was designed and trained by a half of the samples, then its performance was tested on the others. The recognition rate of 93.6% was obtained for the test samples.

1. 緒　　言

音声認識では，音声の周波数領域での特徴を用いることが多い。その際に，音声そのもののスペクトルより，個人差の大きい声帯波の影響をそれから除いた声道伝達特性を用いる方が優れているのは周知のことである。しかし，声道が生体中にあるので，その特性を測定するのは容易でない。実際の音声認識装置に声道伝達特性を利用するには，データ処理の関係上，出来るだけ短時間の音声データのみから，実時間的にそれが決定されるのが望ましい。

声道伝達特性と声帯波とを分離する方法にはケプストラム法[1]，合成による分析法[2]，線形予測法[3]等がある。しかし，いずれも比較的長い音声波データを用いており実時間処理はむずかしい。しかも，これらはスペクトルを単に平滑化するだけである。すなわち，声帯波の周期性を反映するスペクトルの細かい構造を除去するだけで，その全体像である声帯波スペクトル包絡線の音声スペクトルへの影響までは考慮されていない。

著者等は，ごく短時間の音声波から簡単に声道伝達特性を推定する方法を既に提案し，その有

* 応用物理学科応用計測学講座
† 現在，筑波大学物理工学系

効性を確かめている[4]。それは，同一話者の発声した5種類の日本語母音の各一周期分の波形から声帯波を近似的に求めるものであった。この方法は，5種の母音をピッチ周期，大きさ等をほぼ同じにして，話者にあらかじめ発声してもらう必要があった。この制約のため，実際の音声認識装置にそれを応用するのは困難である。

本論文では，2章でこの制約を除いた更に簡単な方法を提案し，3章でその妥当性をシミュレーションにより検討する。更に4章では実際に小型計算機を用いて，この方法に基づく実時間の声道伝達特性の測定システムを構成する。また5章ではこのシステムで測定される伝達特性を用いて，成年男子100名から得た500個の日本語母音に対して識別実験を行った結果について述べる。

2. 推 定 原 理

人間の発声系を線形系と見なすと，音声波スペクトル $S(j\omega)$ は

$$S(j\omega)=V(j\omega)G(j\omega) \tag{1}$$

と表わせる。ここで，$V(j\omega)$ は声道伝達特性，$G(j\omega)$ は声帯波スペクトルである。$S(j\omega)$ は音声波から決定できるから，$G(j\omega)$ が知れれば，声道伝達特性 $V(j\omega)$ を得ることができる。

著者等の先の方法では，前もって収録しておいた音声波形から声帯波を推定する。しかし，声道伝達特性を音声認識装置に利用するには，現に発声中の音声波のみから実時間的にそれらが求められることが必要である。

ここで提案する方法を次に述べる。Fig. 1 (a) は直流分を除きかつ 3 kHz に帯域制限された母音 /a/ の波形の定常な部分から2周期分を取り出したものである。この図にも見られるように，一般に母音波形の一周期中に一箇所，図中に下線を施して示すような三角波状の部分が存在する。これは，声門が開き，そこを通過する呼気流すなわち声帯波で声道が励振されている区間に対応する。この部分は声帯波形として従来から知られている非対称三角波と類似している[5]。そこで，周波数領域でその特徴を検討してみる。三角波状部分の前縁の零交差点から後縁の零交差点までを残し，他を零として，この部分を音声波から分離する。

Fig. 1 (b) は Fig. 1 (a) の波形からこのようにして分離された三角波状部分である。Fig. 2 はそれから計算したパワスペクトルであり，破線は −12 dB/Oct. の減衰線である。図から，このパワスペクトルが約 −12 dB/Oct. で減衰することがわかる。これは，従来から知られている声帯波スペクトルの特徴に一致する。多くの母音標本についても，同様の特徴が認められる。そこ

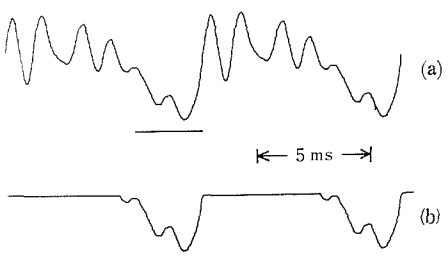


Fig. 1 Waveform of vowel /a/ (a), and estimated glottal waveform (b).

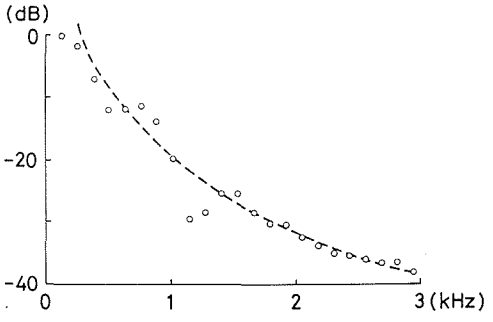


Fig. 2 Power spectrum of the estimated glottal waveform in Fig. 1 (b).

で，音声波から三角波状部分を分離して，それを近似的に声帯波と見なすことにする。すると，次式によって近似的な声道のパワ伝達特性 $|\hat{V}(j\omega)|^2$ が推定できる。

$$|\hat{V}(j\omega)|^2=\frac{|S(j\omega)|^2}{|\hat{G}(j\omega)|^2} \tag{2}$$

ここで，$|\hat{G}(j\omega)|^2$ は近似的声帯波のパワスペクトルである。なお，音声のスペクトル解析において，その位相部分は必要ない。よって，声道のパワ伝達特性をもってその伝達特性としても差支えない。

3. 推定法の検討

2章で提案した近似的声道伝達特性の推定法について検討する。声道の特性を考える場合，音源のインピーダンスが十分大きいとして，声帯音源の影響は通常無視される。しかし実際には，音源のインピーダンスは有限であるから，声道の共振点での減衰は，理想的な閉管のそれより大きくなる。減衰が大きくなることは，フォルマントの帯域幅を増加させる効果を持つ。

ところで，声帯音源のインピーダンスは，声門の開口面積にほぼ反比例する[6]。従って，呼気流が声道を励振する強制振動時には声門が開くため，音源のインピーダンスは自由振動時より減少する。つまり，強制振動時には，フォルマントの帯域幅がかなり広くなっていると考えられる。このため，励振波形である声帯波は声道の共振の影響をそれ程受けず，そのまま口から放射されることが期待できる。したがって，声道の強制振動区間に対応する部分を音声波から切り出して，それを近似的に声帯波と見なすという我々の方法は，上述のことを仮定していることになる。この仮定の当否を検討するために，次のような計算機シミュレーションを行った。

発声系のモデルとしては，声道と声帯音源の相互作用を考慮して，Fig. 3 の等価回路を用いた。ここでは簡単のため，議論を低い周波数領域の特性に限定し，第1フォルマントのみを考え，声道を単一の共振系で近似している。ここで，E_0 は電圧源で，肺の圧力に対応する一定値をとる。$R_g(t)$ は声門のインピーダンスに対応しており，強制振動区間で小さくなるように，周期的に変化させる。$U_g(t)$，$U(t)$ はそれぞれ声門，唇を通過する体積流，つまり声帯波と音声波

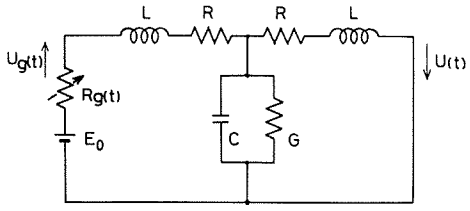


Fig. 3 Model of the vocal system.

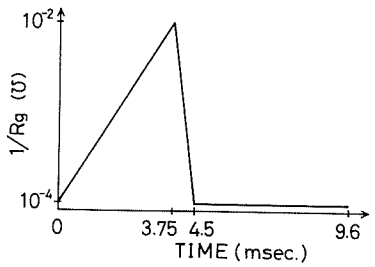


Fig. 4 Admittance of the glottis, $1/R_g(t)$.

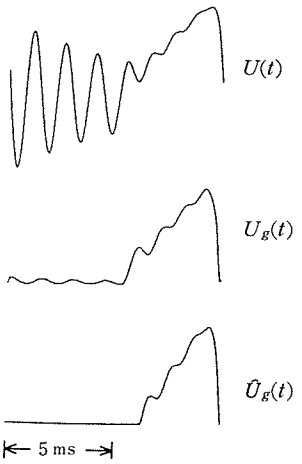


Fig. 5 Computed waveforms.

に対応している。

音声波 $U(t)$ は次の3階の変係数微分方程式を満足する。

$$\begin{aligned}&\dddot{U}+f(t)\ddot{U}+g(t)\dot{U}+h(t)U=\frac{E_0}{CL^2}\\&f(t)=\frac{2CLR+GL^2+CLR_g(t)}{CL^2}\\&g(t)=\frac{2GLR+2L+CR^2+(CR+GL)R_g(t)}{CL^2}\\&h(t)=\frac{2R+GR^2+(GR+1)R_g(t)}{CL^2}\end{aligned}\tag{3}$$

また，$U_g(t)$ は $U(t)$ を用いて (4) 式のように表わせる。

$$U_g(t)=CL\ddot{U}+(CR+GL)\dot{U}+(GR+1)U\tag{4}$$

ここで R, L, C, G の値は声道を長さ 17 cm，断面積 5 cm² の一様な太さの円管と仮定し，それを一つの T 型回路で近似したときの理論値[6]をもとに，共振周波数が /a/ の第1フォルマントに一致するように決めている。実際の計算では $C=0.5\,\mu F$, $L=0.1\,H$, $R=10\,\Omega$, $G=0\,\mho$ とした。$1/R_g(t)$ は声門の開度を参考にして，Fig. 4 に示すような非対称三角波とした。(3) 式をルンゲ・クッタ法で数値計算して $U(t)$ を，更に (4) 式より $U_g(t)$ をそれぞれ求めた。Fig. 5 は結果の一例であり，それぞれ定常状態に達した時点での音声波 $U(t)$，声帯波 $U_g(t)$ および近似的声帯波形 $\hat{U}_g(t)$ である。$U_g(t)$ と $\hat{U}_g(t)$ の波形には，かなりの類似が見られる。Fig. 6 はこれらのパワスペクトルであり，周波数領域において，両者がほぼ一致することがわかる。

Fig. 7 は $R_g(t)$ が最大 ($R_g(t)=10\,k\Omega$) つまり声門が閉じたときの声道伝達特性（実線）と，我々の方法で推定された近似的声道伝達特性（○）である。両者を比較すると，今注目している低周波域において，伝達特性が良く推定されているのがわかる。これらのことから，この近似的声帯波を用いて，(2) 式により伝達特性を推定することが許されよう。

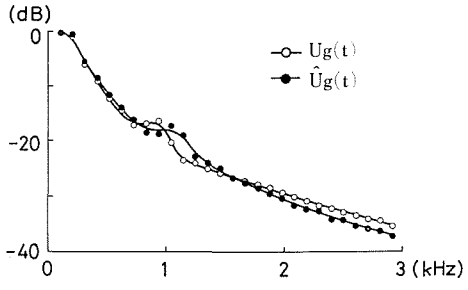


Fig. 6 Powerspectra of $U_g(t)$ and $\hat{U}_g(t)$.

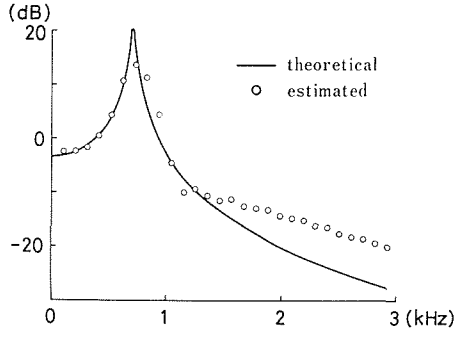


Fig. 7 Estimated transfer characteristic of the model.

4. 声道伝達特性の測定

4.1 測 定 装 置

データの処理を実時間で行えるように，小型計算機 PDP-8/E (1語 12 ビット，サイクルタイム 1.2 μs，コア 12 k 語) を中心とした Fig. 8 のような測定システムを開発した。

ここで，LPF は遮断周波数 3 kHz，減衰特性 −12 dB/Oct. の低域フィルタ。ADC は 8 ビッ

トの A/D 変換器である。標本化用の同期信号を発生するクロックはコンピュータからの指令で，パルス間隔を 1～127 μs の範囲で 1 μs 刻みで設定できる。MT-6 は標本化された音声信号を格納するディジタル式カセット磁気テープである。出力装置として CRT ディスプレーと X-Y レコーダを備えている。リモートスイッチはデータ処理の種類，手順等をコンピュータに指示するものである。

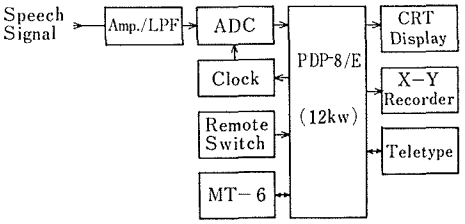


Fig. 8 Data processing system.

4.2 測 定 手 順

2 章で述べた方法で，実際に近似的声道伝達特性を測定する。Fig. 9 は測定手順の流れ図である。被験者に通常の会話時の強さ，高さで 5 種の母音を発声させ，それを入力する。コンピュータは初めの一周期分の波形を十分小さな間隔で標本化して，それからピッチ周期を抽出する。このピッチ周期の 1/128 を標本化周期として，次の一周期分の波形を 128 点で標本化し，コアに格納する。このようにして標本化された一ピッチ分の音声データの直流分を除いてから，まず先に述べた方法で近似声帯波を求める。次に音声波，近似声帯波の分散を規格化し，両者のスペクトルを FFT を用いて計算し，ハニング窓で平滑化する。最後に，(2) 式により近似的声道伝達特性を計算し，CRT ディスプレーあるいは X-Y レコーダに出力する。

ピッチの抽出は次のようにして行われる。先に述べたように，連続して入力される音声波の一つの三角波状部分の後縁の零交差点から，次の三角波状部分の同様の点までの時間間隔を測定し

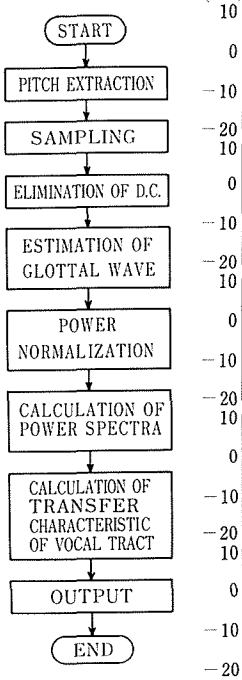


Fig. 9 Flowchart of the data processing.

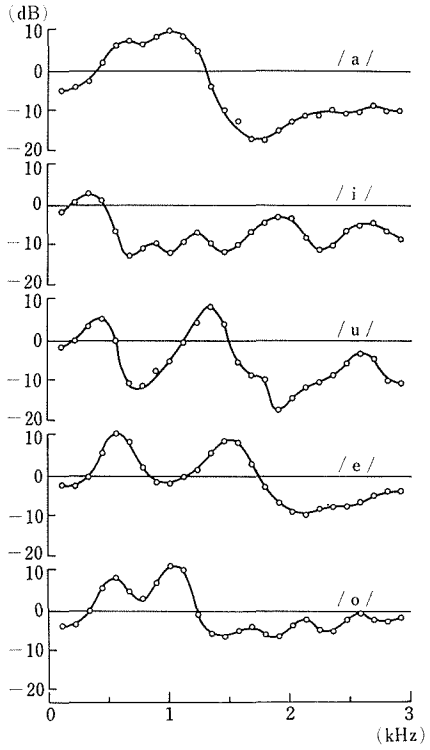


Fig. 10 Estimated transfer characteristics of vocal tracts.

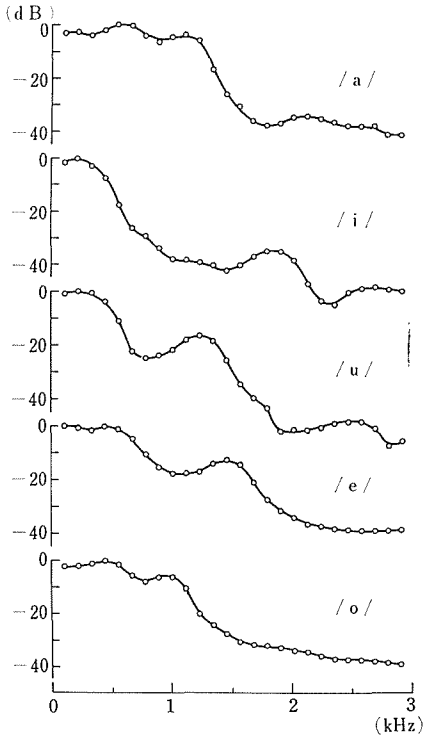


Fig. 11 Powerspectra calculated from the same sample vowels as used in Fig. 10.

て，それをピッチ周期とする。

4.3 測 定 結 果

実際に測定された声道伝達特性の例を Fig. 10 に示す。この図からわかるように，従来知られているフォルマント周波数と，ここで求めた伝達特性のピークの周波数が良く一致している。Fig. 11 は Fig. 10 の伝達特性を求めるのに用いたのと同一の波形標本から求めた音声波のスペクトルである。声帯波の影響のため，この図では特に低い周波数領域にある第1フォルマントのピークが判別しがたい。一方，声道伝達特性は声帯波の影響が除かれているため，低い周波数にあるピークもはっきり現れている。したがってフォルマント周波数を容易に求めることが出来る。Table 1 に，5名の話者の声道伝達特性からピークピッキング法[7]で求めたフォルマント周波数を示す。ここで(−)の記号は，フォルマントに対応するピークが見出せなかったものである。

5. 母音識別への応用

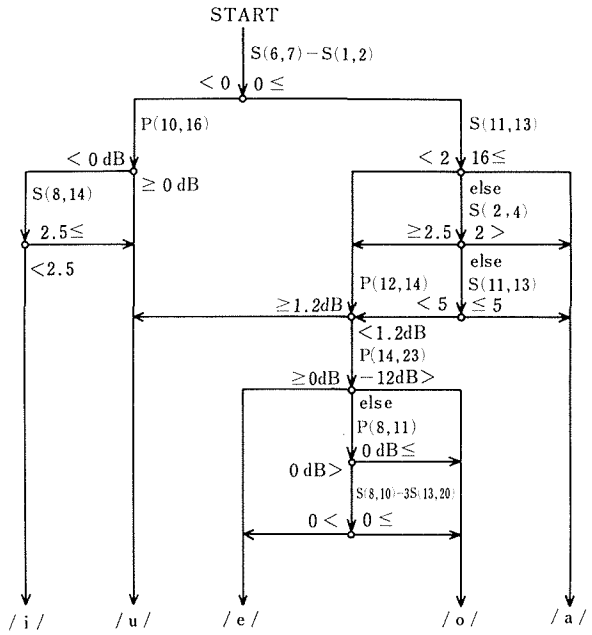
このような簡単な方法によってもフォルマント周波数が十分求められることが確認できたの

Table 1 Formant frequencies obtained from transfer characteristics of the vocal tract.
F_1, F_2, F_3: Formant frequency (Hz)
F_0: Pitch frequency (Hz)

vowel	speaker	F_0	F_1	F_2	F_3
/a/	1	132	792	1056	2904
	2	182	728	1274	—
	3	112	672	1120	3136
	4	113	678	1130	3277
	5	159	954	—	2385
/i/	1	137	411	2466	3151
	2	186	372	2976	3720
	3	112	336	—	2912
	4	112	336	2240	—
	5	163	326	1956	2445
/u/	1	135	540	1350	2295
	2	182	364	—	3094
	3	110	440	1320	2640
	4	112	448	1344	2352
	5	159	318	1113	2067
/e/	1	132	660	1848	3168
	2	186	558	1488	3720
	3	109	545	1744	2725
	4	112	560	1456	3024
	5	156	624	1560	3276
/o/	1	132	560	—	2772
	2	178	534	1068	3560
	3	109	545	872	3052
	4	113	565	1017	3277
	5	156	624	936	2340

で，音声識別に応用してみた。音声データには成年男子100名から各5母音，計500個の日本語母音を採集して用いた。通常の実験室で自然に発声された音声を，口から約10 cm離して設置したマイクロフォンで受け，データレコーダに一旦記録する。次にこれを再生しながら，各母音の定常部分から一周期分を128点で標本化して，ディジタル磁気テープに格納する。このようにして得られた標本をもとに $|\hat{V}(j\omega)|^2$ を推定し，更に一次補間により 100 Hz 毎の値 $A_n=|\hat{V}(j2\pi\cdot 100n)|^2$ $(n=0, 1, 2, ..., 30)$ を求めて用いた。

まず，各母音について初めの20名のサンプルから A_n を求め，ピークの位置，パワの集中度等の共通の特徴を見出す。この特徴を用いて母音判別の流れ図を作る。次にこの20名を含む50名のサンプルを利用して，誤り率が最も小さくなるように閾値を調整する。Fig. 12 はこのようにして決定された母音識別の流れ図である。ここで $P(p, q)$ は $p\leq n\leq q$ における A_n のピーク値，$S(p, q)$ は $p\leq n\leq q$ の A_n の和 $(\sum_{n=p}^{q} A_n)$ を表わす。調整終了時の識別正解率は96.0%



$P(p, q)$: Peak value of A_n, $p\leq n\leq q$

$S(p, q)$: $\sum_{n=p}^{q} A_n$

Fig. 12 Flowchart of vowel recognition.

Table 2 Recognition result of training samples.

Input	Recognition Output				
	/a/	/i/	/u/	/e/	/o/
/a/	50	—	—	—	—
/i/	—	49	1	—	—
/u/	—	2	47	1	—
/e/	1	1	1	47	—
/o/	2	—	—	1	47

correct-recognition rate 96.0%

Table 3 Recognition result of test samples.

Input	Recognition Output				
	/a/	/i/	/u/	/e/	/o/
/a/	49	—	—	—	1
/i/	—	47	2	1	—
/u/	—	5	45	—	—
/e/	—	—	1	46	3
/o/	—	—	—	3	47

correct-recognition rate 93.6%

であった。

残りの50名の未知のサンプルを用いて識別実験を行った結果，識別正解率は93.6%となった。Table 2，3はこの2つの結果をまとめたコンフュージョンマトリックスである。

なお，一つのサンプルの識別に要する時間は約6秒である。その大部分は2つのスペクトルの計算に費されているので，専用のFFT装置を用いれば，実時間処理も可能であろうと思われる。

6. むすび

本論文では，音声認識にとって重要な情報を与える声道伝達特性の簡単な推定法を提案した。それは，従来知られている声道と声門の相互作用についての知識と，著者等が見出した事実とに基づいて考案されたもので，発声時の声道伝達特性を一ピッチ分の母音波形のみから極めて簡単に推定するものである。

この方法の妥当性を計算機シミュレーションで検討した。また，発声時の声道伝達特性を求める信号処理システムを開発し，実際の音声波から得られた伝達特性の構造が従来のフォルマントパターンと良く一致することを確かめた。

更に，声道伝達特性を用いて100名の成年男子から得た500個の母音波形標本について識別実験を行った。半数の標本で識別システムを決定し，残り250個の未知標本に対して識別実験を行い93.6%の識別率を得た。

日頃，種々の点でご助言いただく応用光学講座村田和美教授に感謝致します。また，推定方法を検討した際の計算機シミュレーションには，北大大型計算機センターを利用した。

参考文献

1) A. V. Oppnheim: A speech analysis-synthesis system based on homomorphic filtering, J. Acoust. Soc. Am., **45,** (1969), 2, pp. 458-465.
2) C. G. Bell. et al.: Reduction of speech spectra by analysis-by-synthesis techniques, J. Acoust. Soc. Am., **33,** (1961), 12, pp. 1725-1736.
3) B. S. Atal and S. L. Hanauer: Speech analysis and synthesis by linear prediction of speech wave, J. Acoust. Soc. Am., **50,** (1971), 2, pp. 637-655.
4) 大場，土肥，井戸川：発声時における声道伝達特性の推定，北大工学部研究報告，74，(昭50)，pp. 41-51.
5) R. L. Milles: Nature of the vocal cord wave, J. Acoust. Soc. Am., **31,** (1959), 6, pp. 667-677.
6) J. L. Flandgan: Speech analysis, synthesis and perception, 2nd ed., (1972), chapter 3, Springer-Verlag.
7) J. L. Flanagan: Automatic extraction of formant frequencies from Continuous speech, J. Acoust, Soc. Am., **28,** (1956), 1, pp. 110-118.