TR－H－009 23

変換聴覚フィードバックの基礎検討

非定常ピッチ変換による発声ピッチの変動について

岩谷 賢
豊橋技術科学大学 4年
河原 英紀
ATR人間情報通信研究所、第一研究室

1993. 3. 2
(1993.2.26受付)

ATR人間情報通信研究所
〒619-02 京都府相楽郡精華町光台2-2 ☎07749-5-1011

ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
Telephone: +81-7749-5-1011
Facsimile: +81-7749-5-1008

変換聴覚フィードバックの基礎検討

- 非定常ピッチ変換による発声ピッチの変動について -

岩谷　賢

豊橋技術科学大学 4年

河原英紀

人間情報通信研究所、第一研究室

あらまし

発声時における聴覚系の寄与を明らかにするための方法として、聴覚的フィードバック経路に信号処理を挿入する「変換聴覚フィードバック」を取り上げ、その基本的性質の検討を行った結果について報告する。また、これらの検討の過程で確認した新しい聴覚フィードバック現象についても併せて報告する。

これまで聴覚フィードバックの研究の多くで用いられていた「遅延」処理は、発声過程への妨害が大きく、正常な状態での発話過程の検討には適していない。本検討では、まず、これらの効果についての追試を行い、同様な傾向が観測されることを確認した。次いで、「ピッチ変換」処理を用いてかつ逐次的な追跡が不可能な速度で変換パラメタを変化させることで、より自然な状態に近い状況での自動的な過程を観測することを試みた。その結果、外から加えられたピッチ変換を打ち消す方向に、約150msの遅れで発声された音声が追従して変化する現象が、高い再現性を有して見いだされた。

なお、本報告は、1993年1月8日から2月26日にかけて、ATR人間情報通信研究所において行われた実習内容および結果の要約である。

はじめに

聴覚フィードバックとは

発声は、情報を生成する一方向だけの過程ではない。音声の生成時には、筋肉からフィードバックされる運動感覚をはじめとして触覚、聴覚などの感覚レベルのフィードバック、言い回しなどの、言語レベルのフィードバック、相手の反応を見る談話レベルのフィードバックなどの様々に異なったレベルでの調整が行なわれている双方向の複雑な過程である。聴覚フィードバックの実験は、これらの過程の中で特に聴覚に関するフィードバックループを操作することにより、音声生成過程における聴覚の役割を調べようとするものである。従来、聴覚フィードバックの研究は、顕著な効果が見られることから、主に遅延をフィードバック経路に挿入した場合について検討が進められてきた。(1)

研究の目的

しかし、これら従来行なわれてきたような遅延聴覚フィードバックの研究は、いわば破壊検査のようなものであり、正常な発声状態における聴覚フィードバックの寄与を調べるという目的には必ずしも適さないものであった。今回の研究は、正常な発声状態における聴覚フィードバック経路の寄与を調べる方法を確立するために必要となる基礎的データを蓄積するところにある。

方法

概要

正常な発声状態における聴覚系の役割を解明するという目的を達成する手段としてみたとき、従来の遅延聴覚フィードバックには以下のような問題点があることがわかる。

従来の方法（遅延聴覚フィードバック）の問題点

（1）従来の方法では正常な発声の過程を破壊。
（2）音声自体の変動が大きいので影響を分離して抽出することが困難。
（3）影響が意識されると対応に個人差が大きくなる。

これらの問題点を解決するための手段として本研究では以下のような対策を考えそれぞれの効果についての基礎データの収集を行なった。このような非破壊的なパラメタ変換を用いた聴覚フィードバック実験を変換聴覚フィードバックと呼ぶことにする。

その対策

（1）破壊的でない条件での収録。
（2）意識的な逐次的制御のかからない自動的な過程に注目。
（3）測定系の精密な時間軸を最大限に利用。
（4）音声認識で開発されたＤＴＷの利用による共通成分の除去と、それによるＳ／Ｎ比の改善。

遅延聴覚フィードバック

まず従来から研究が進められている遅延聴覚フィードバックについて、基本的な実験系を構築して追試を行ない測定への習熟を図った。

聴覚フィードバック経路に遅延を挿入した場合に、遅延の影響のみを調べるためには、

（1）通常の聴覚フィードバック経路の影響を除くこと、
（2）（1）を実現する手段により生ずる他の影響を除くこと、

が必要である。これらの影響を完全に除去することは困難であるが次のような手段を導入して、軽減を試みた。（1）のために、密閉型ヘッドフォンで耳を覆うとともに、残留する音声をマスクするためにピンクノイズをフィードバック音声に加える。（2）のために、（1）と同様の条件で遅延のみを加えない同レベルのフィードバックを行なった場合の音声を基準音声として収録し、結果の評価に当っては、この基準音声からの変化分を用いた。

これまでの研究により、遅延聴覚フィードバックでは200ms程度の遅延で最も大きな効果が得られることがわかっている。そこで 今回の研究では 200msを中心として、その二分の一と二倍にあたる100ms,400ms の遅延時間について発声資料の１と２を用いて実験した。

変換聴覚フィードバック

変換聴覚フィードバックは模式図 FIG21 に示すように聴覚フィードバック経路中に挿入した信号処理により、音声のパラメタを抽出、変換し、再び音声を合成することによる発声への影響を調べる方法一般を指している。本研究ではその中から特に音声の主要なパラメタの一つであるピッチを取り上げ、その変換による影響を調べることにする。

ピッチは日常の会話でも韻律を構成する要素として重要な役割を果たしている。また。一次元の量であるため制御、解析が容易であるという利点がある。ここでは以下の二つの種類のピッチ変換について調べた。

実験を行なう上での一般的な注意事項は遅延聴覚フィードバックと同様である。

ピッチ変換（定常）

まず時間的に一定なピッチ変換を施す。今回は +100cent,-100cent の二通りについて発声資料の１と２を用いて実験した。

ピッチ変換（非定常）

時間的に一定なピッチ変換の実験では被験者がフィードバック経路にピッチの変換が行なわれているということを意識できる[2]。ここでは意識的な逐次的制御のかからない自動的な過程に注目するために、時間的に変化するピッチ変換を施す。今回は振幅 100cent 、周波数7.3Hz,5.0Hz,3.0Hz,2.0Hz の正弦波について発声資料の３と４を用いてピッチ変換を行った。更に聴覚から音声に至る系の特性を詳しく調べるために正弦波の代りに疑似白色信号であるM系列を用いた。

実験

実験系の構成をFIG1と表1に示すこれらの系の特性は実験に先立って測定され、構成された。主要な特性を付録1に示す。

今回の実験の中心となるのは、遅延やピッチ変換の処理を行なうHARMONIZERである。この装置は本来音楽制作用に開発されたものであり、遅延、ピッチ変換など様々な処理機能を持っている。またこれらのパラメタは外部の機器からMIDIインターフェースを介して設定が可能であり時間的に変化する特性を容易に実現することができる。

なお被験者は成人男性三名（SI, HK, KO）と成人女性一名（MT）である。

また今回の実験で使用した発声資料とソフトウェアを表2と表3に示す。

実験結果

遅延聴覚フィードバック

遅延聴覚フィードバックにおいては（1）調音の誤り（言い間違い、吃りなど）、（2）発声速度の低下、（3）基本周波数の上昇、（4）音量の増加のような現象が生ずることが報告されている。今回の実験では同様な効果が認められたが、従来の報告にもある通りこれらの影響にはかなり個人差があった。このように個人差が大きいため以下の結果の分析では、個人毎に各要因の効果の統計的な有意性を調べた。

強度、継続時間長への影響

まず強度については今回の被験者のうち二名についてはほとんど影響が出なかった。残り二名のうち一人は遅延時間につれて強度が大きくなった。またうち一人は遅延時間200msまでは遅延時間につれて強度が大きくなったが400msでは若干強度が小さくなった。（FIG2）

継続時間長についてもかなり個人差が大きく、今回の被験者のうち二名については遅延時間とともに継続時間長も長くなった。残り二名については遅延を挿入しない場合に比べ挿入した場合の方が継続時間長は長くなったが、特に遅延時間が長ければ継続時間長も長くなるわけではなかった。（FIG3）

ピッチへの影響

ピッチへの影響もかなり個人差の大きなものとなった。今回の被験者のうち二名についてはほとんど影響が出なかった。残り二名のうち一人は遅延時間につれてピッチが大きくなった。またもう一人は遅延時間につれてピッチが小さくなった。（FIG4）

また今回の実験で、この方法では正常な発声を保つことは困難であるということが確認できた。（FIG5は基準音声のケプストラムと基準音声の2回目の発声のケプストラムとの距離を取ったもので、FIG6は基準音声のケプストラムと

400msの遅延をかけたときの音声のケプストラムとの距離を取ったもの。FIG6の方は正常な発声が行なわれていないことがわかる。）

変換聴覚フィードバック

ピッチ変換（定常）

強度、継続時間長への影響

ピッチ変換による強度への影響は今回の被験者についてはほとんどなかったが、継続時間長についてはピッチ変換を施した場合の方が+100cent、-100centのどちらの場合も、施さない場合に比べわずかに長くなった。（FIG7,FIG8）

ピッチへの影響

ピッチ変換によるピッチへの影響もかなり個人差の大きなものとなった。今回の被験者のうち一人についてはほとんど影響が出なかった。残りの三名のうち二名については+100centのときにピッチが下がり、-100centのときにピッチが上がった。もう一名については+100centのときにピッチが上がり、-100centのときにピッチが下がった。（FIG9）

ホルマントへ影響

ＤＴＷによって時間軸の整合を行ない、対応する位置でのホルマントの変化について調べることを試みた。予想されるピッチ変換によるホルマントへの影響よりもホルマントのデータのばらつき方が大きいことが明らかになった。したがってこの方法でホルマントへの影響を見るためには、非常に大きなピッチ変換を行なうことが必要となり、当初の意図であった正常に近い条件での効果を測定することは困難であるといえる。（FIG10）

ピッチ変換（非定常）

前項で認められたピッチ変換の影響が、意図的なものであるか、あるいは自動的な補償機構によるものであるかを調べるために、直交性のある関数を用いてピッチ変換の値を時間的に変動させる実験を行なった。

ピッチへの影響

ピッチ変換を行なったときに発声された音声と、通常の状態で発声された音声との差は、ピッチの時間的変化波形をそのまま比較しても明かではない。ここで用いたピッチ変換関数は正弦波なので、発声資料３についてピッチからその平均値を引いた波形をフーリエ変換することにより、ピッチの自発的な揺らぎの影響を分離することができる。このような分析を行ったところ、いくつかの結果においてピッチ変換の周波数にピークが見られた。被験者の男性三人については、ピッチ変換の周波数が5.0Hzのときに特に大きなピークが見られた。（FIG11は基準音声の結果、FIG12はピッチ変換周波数が5.0Hzのときのフィードバックされる音声の結果、FIG13はそのフィードバックを施したときに発声した音声の結果。）しかしこの結果だけからでは、発話者が揺らぎの速度を検出して、同様にピッチを揺らせているのか、ピッチ変換により、直接的に声帯の制御に影響が及んでいるのかを区別するのは困難である。そこでピッチ変換に用いた正弦波と、生成された音声の揺らぎの時間的な関係を調べることにした。0.1秒ごとに１秒間の基準音声と正弦波との相互相関を取ると、正弦波とピッチ揺らぎの間には安定した相関がないことがわかる。一方非定常の変換フィードバックを施したときの結果の音声と正弦波とを同様に相互相関を取ると時間的に一定の関係があることが明らかになった。

以上の実験で用いた音声は母音であり、通常の文章ではなかった。研究の狙いは、通常の文章を発声しているときの聴覚系の関与を明かにすることにある。通常の発話においてピッチは、韻律情報を伝えるため、時間的に大きく変化する。ピッチ変換の影響を調べるためには、この大きな変化の影響を除く必要がある。

そのための手段として、ここではＤＴＷにより時間軸の整合を行なって、通常の音声と、変換の影響を受けた音声に共通の傾向を除去することの効果を調べた。（FIG14,FIG15,FIG16はそれぞれ基準音声、フィードバックされる音声、フィードバックを施したときに発声した音声を前出の方法で相互相関を取った結果）

発声資料４については、変換フィードバックを施したときの結果の音声と基準音声ををＤＴＷにより時間整合して、結果のピッチから基準音声のピッチを引いたものの時間軸を結果の時間軸に戻し、フーリエ変換した。この操作によりかなりＳ／Ｎ比が向上したが発声資料３の場合のようなはっきりとしたピークは見ることができなかった。ピッチ変換による影響が持続母音の場合と同程度の大きさだと仮定すると、今回のＤＴＷによる方法での残差を三分の一以下にすることができれば、検出は可能にな

ると予想される。（FIG17はその一例で破線が基準音声２回目のピッチの揺らぎを、実線がＤＴＷにより基準音声２回目のピッチの揺らぎから１回目のものを引いたものをフーリエ変換した結果。これより今回の方法では非定常のピッチ変換の影響を見ることは困難であると言える。）

正弦波を用いた試験により、フィードバックされる音声のピッチと生成される音声のピッチとの間に相関が生じることが確認された。しかし、正弦波は周期関数であるため、どのような因果関係が両者の間に存在しているかは明らかではない。そこで、直交する系列の一つであるM系列信号をある系に入力し、その系の出力とM系列信号との相互相関を計算すると、系のインパルス応答が求められることを利用する。

ここではM系列として、次数５（周期３１）のものを用い、８倍のサンプリングレートに変換してピッチ変換器の制御データを生成した。このようにして生成した制御信号により測定できる上限周波数は今回の実験では８Hzであり、独立に観測できる現象の応答時間の上限は1秒である。

結果の一例をFIG18-FIG20に示す。FIG18はヘッドフォンにより発話者にフィードバックされている音声との相互相関を示しており、周期的なインパルスに近い形状であることがわかる。FIG19は発声された音声との相互相関を示す。明瞭なパルス状の反応が見える。FIG20は主要なパルスのピーク付近においてフィードバック音声に対する相関と、発声された音声に対する相関とを同一の時間軸に重ねて拡大したものである。これから遅延は約150ms、符号は逆極性であることが明らかになった。同様の傾向は他の話者についても認められた。(FIG22)

まとめ

聴覚的フィードバックパスでのピッチ変換（定常、非定常）により破壊的でない条件での音声の収録と分析とを行なうことができた。従来の遅延聴覚フィードバックの実験では正常な発声状態を保てないので行動分析が主であったのが、これによってピッチ、ホルマントなどのパラメータレベルでの比較検討が可能となった。また、このような分析が可能となったもう一つの要因として、DATやMIDI controlerなどの実験系の精密な時間軸を最大限に利用したことが挙げられよう。

このような時間軸を利用して直交性を有する信号でピッチ変換（非定常）を行うことにより、意識的な逐次的制御のかからない自動的な過程に注目することができた。持続母音を用いM系列信号を変換用の信号とした実験により、この過程が約150msの遅れを有し、補償方向の働きをするものであることが明らかになった。

さらに、より自然な状況を対象として、通常の文章を発声しているときの効果を明かにすることを狙って、音声認識で開発されたＤＴＷの利用による韻律の共通成分の除去と、それによる検出能力の改善を試みた。しかし、これまでのところ音声の自発的揺らぎや韻律成分の抑圧が不足しており、ピッチ変換（非定常）の実験において興味ある結果を検出するには至らなかった。

参考文献：

(1) Lee, Bernard S. : Effects of delayed speech feedback, Journal of the Acoustical Society of America, 22, 6, pp.824-826, 1950.

(2) Elman, Jeffrey L. : Effects of frequency-shifted feedback on the pitch of vocal productions, Journal of the Acoustical Society of America, 70, 1, pp.45-50, 1981.

表1　　実験装置

HARMONIZER　H3000S	:ピッチの変換及び遅延を挿入する
Macintosh	:ピッチを時間的に変化させて変換する際にHARMONIZERのMIDIインターフェースと接続してHARMONIZERを制御する
SINE/NOISE GENERATOR(B&K1049)	:ヘッドフォンに出力する自己音声マスキング用ノイズを発生する
MICAMP SONY MX-1000ESX	
DAT SONY DTC1000ES	
AMPLIFIER SONY TA-E901	
Head Phone Sennheizer HD-250II Linear	

表2　　発声資料

1 昔、昔、あるところにおじいさんとおばあさんが住んでいました。ある日おじいさんは山へ芝刈りに、おばあさんは川に洗濯に出かけました。

2 音声言語生成機構の研究では発声発話機構、音声パターン神経情報処理、言語生成機構などの研究項目を取り上げます。

3 長時間持続して発声した母音 /a/

4 青い葵の絵は山の上の家にある。

表3　　ソフトウェア

DSPtool	:DATに収録された音声をデータとしてUNIXに取り込む際に用いる
ESPS	:DSPtoolによって取り込まれたデータからホルマント、ピッチ、強度などを算出する際に用いる。
Matlab	:実際のデータ解析に用いる
Perform	:非定常のピッチ変換のためのMIDI信号の生成に用いる。

構成

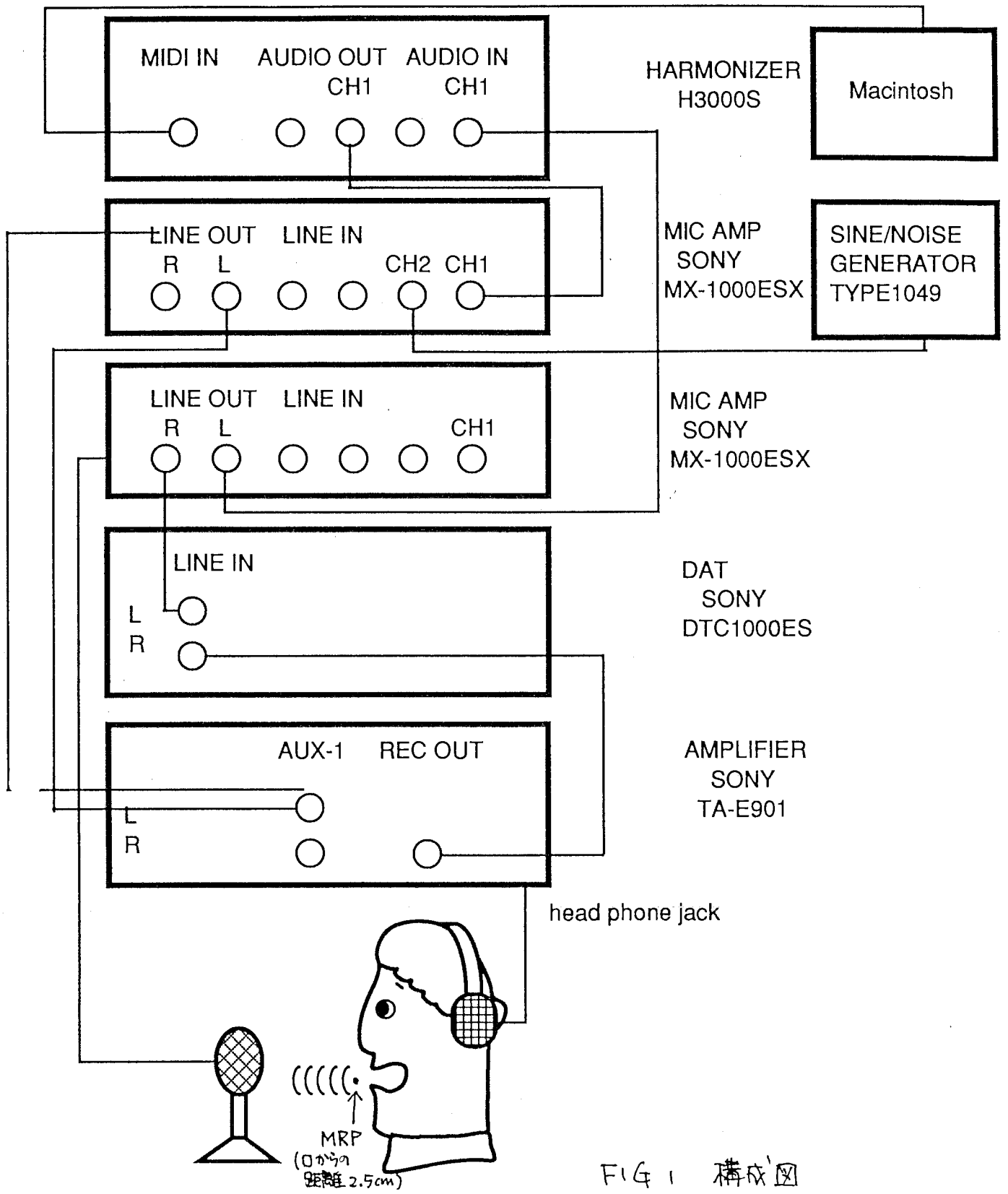
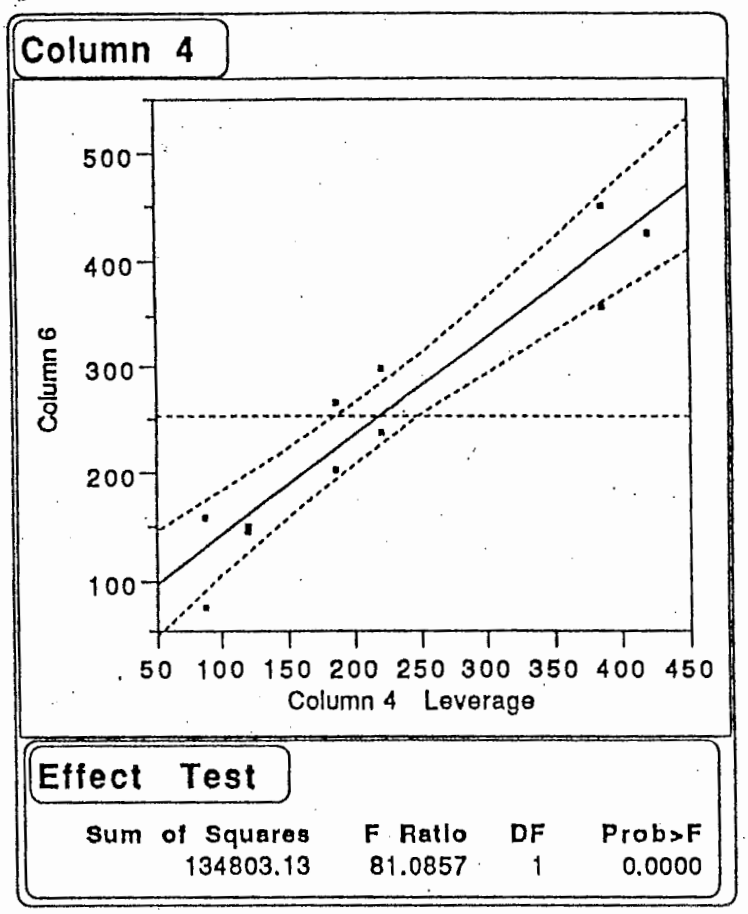

MIDI IN
AUDIO OUT CH1
AUDIO IN CH1
HARMONIZER H3000S
Macintosh
LINE OUT R L
LINE IN CH2 CH1
MIC AMP SONY MX-1000ESX
SINE/NOISE GENERATOR TYPE1049
LINE OUT R L
LINE IN CH1
MIC AMP SONY MX-1000ESX
LINE IN
L
R
DAT SONY DTC1000ES
AUX-1
REC OUT
L
R
AMPLIFIER SONY TA-E901
head phone jack
MRP
(口からの距離2.5cm)

FIG 1 構成図



Column 4 遅延時間 (ms)
Column 6 強度 (相対値)

FIG 2 遅延時間と強度の関係.

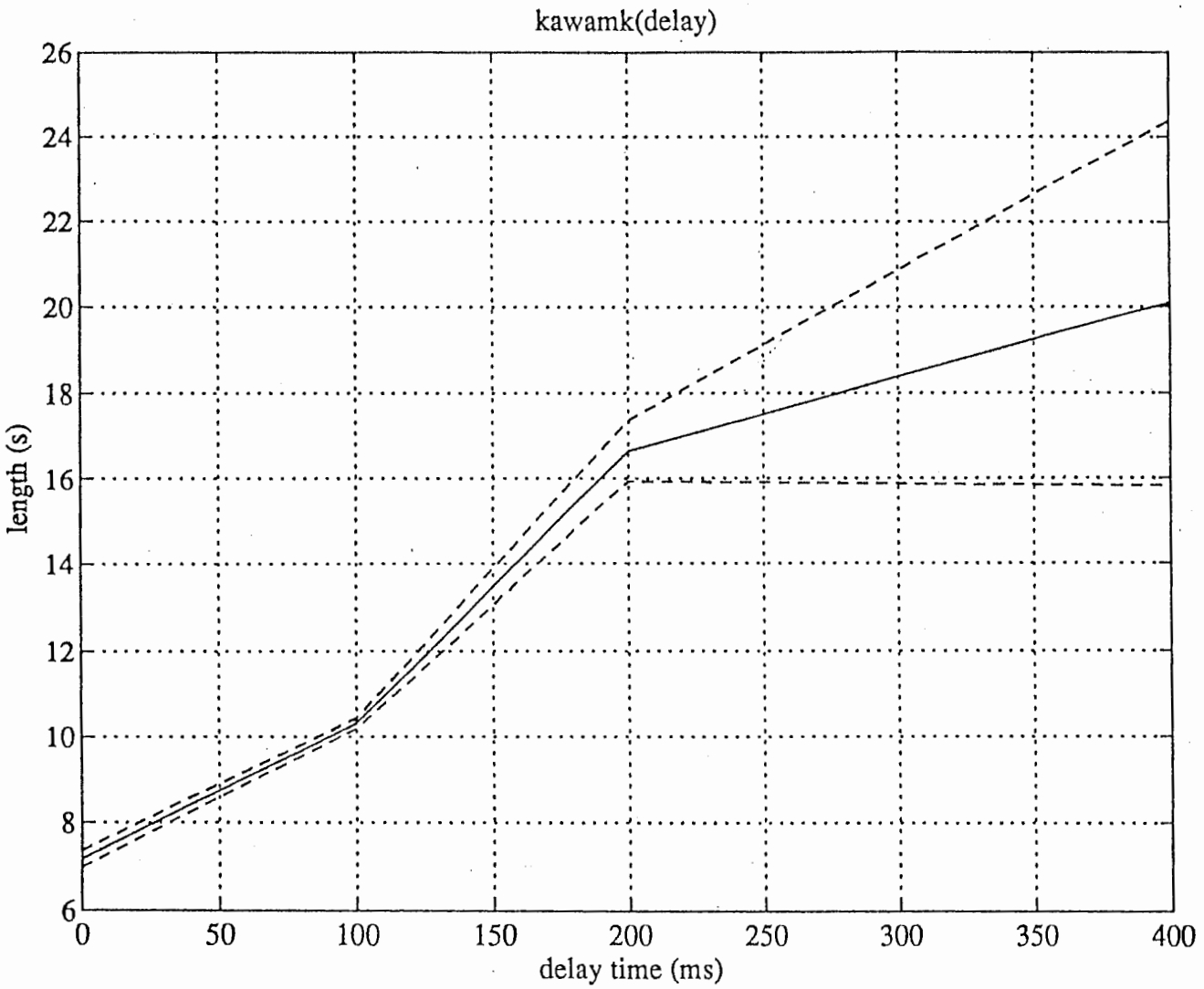
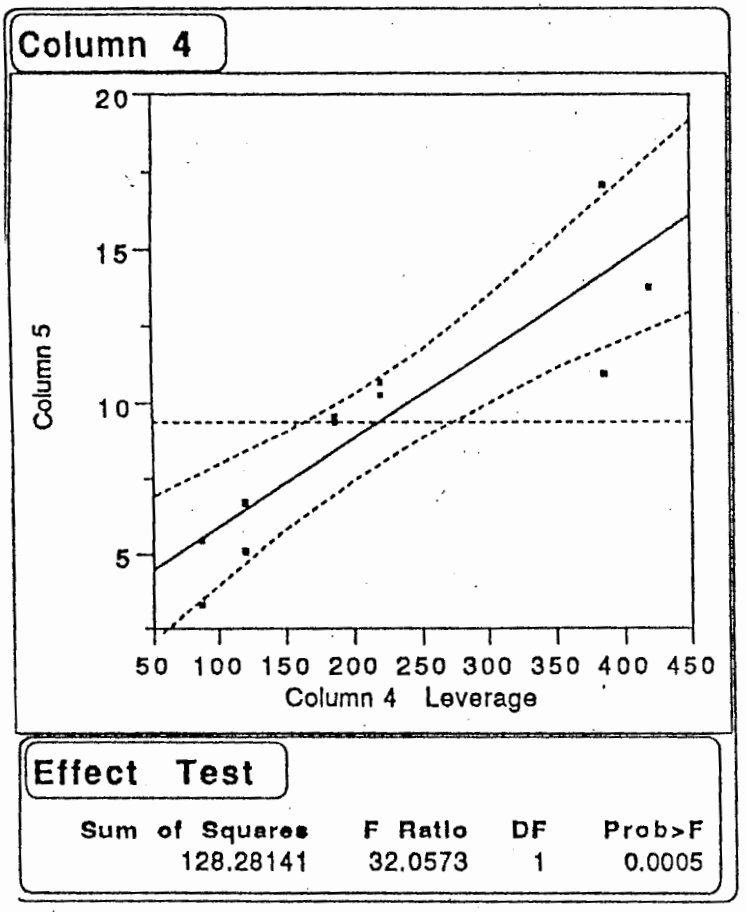


FIG3 遅延時間と維続時間長との関係



Column 4 遅延時間 (ms)

Column 5 ピッチ (相対値) (Hz)

FIG4 遅延時間とピッチとの関係

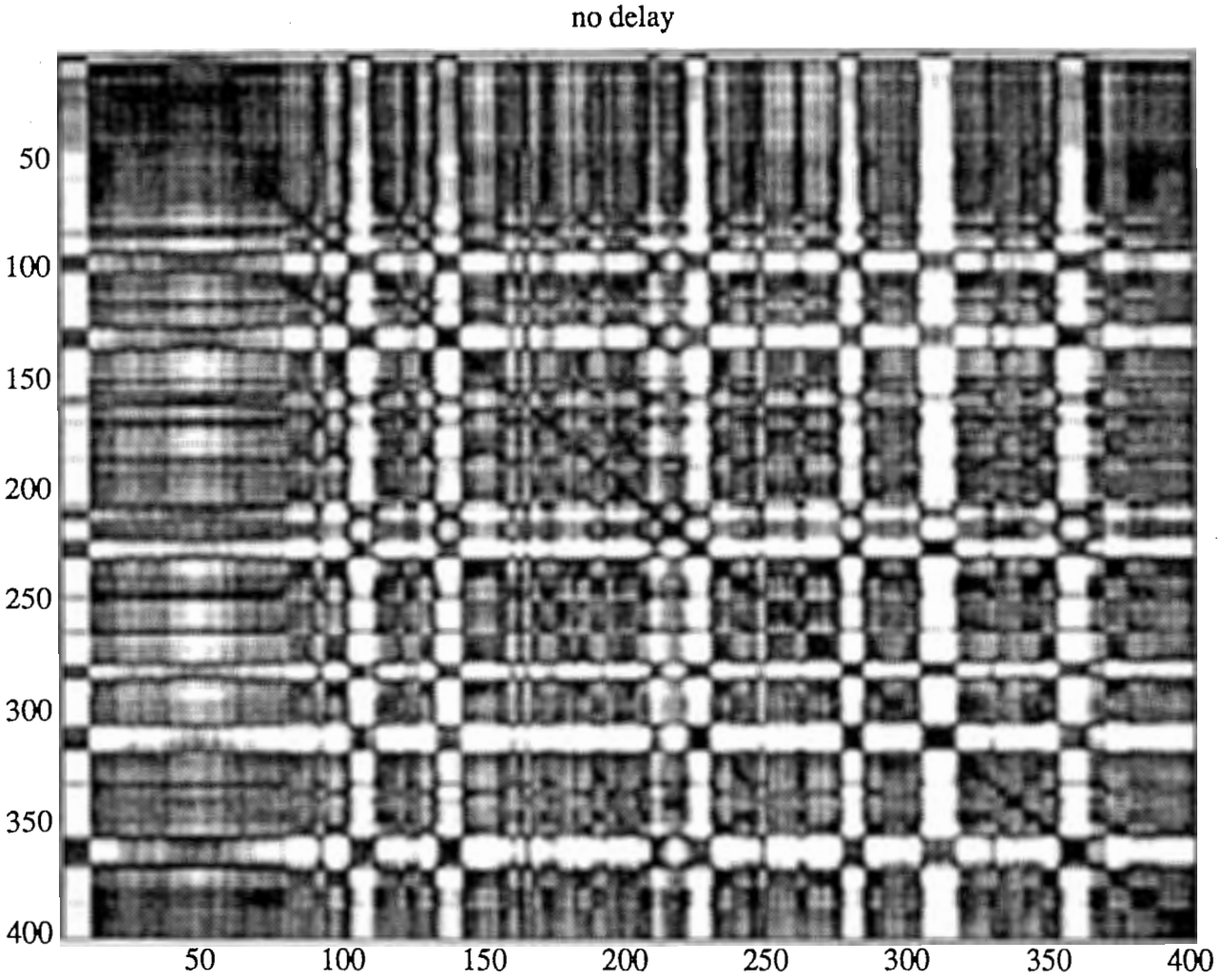


FIG 5 基準音の1回目と2回目のケプストラムの距離

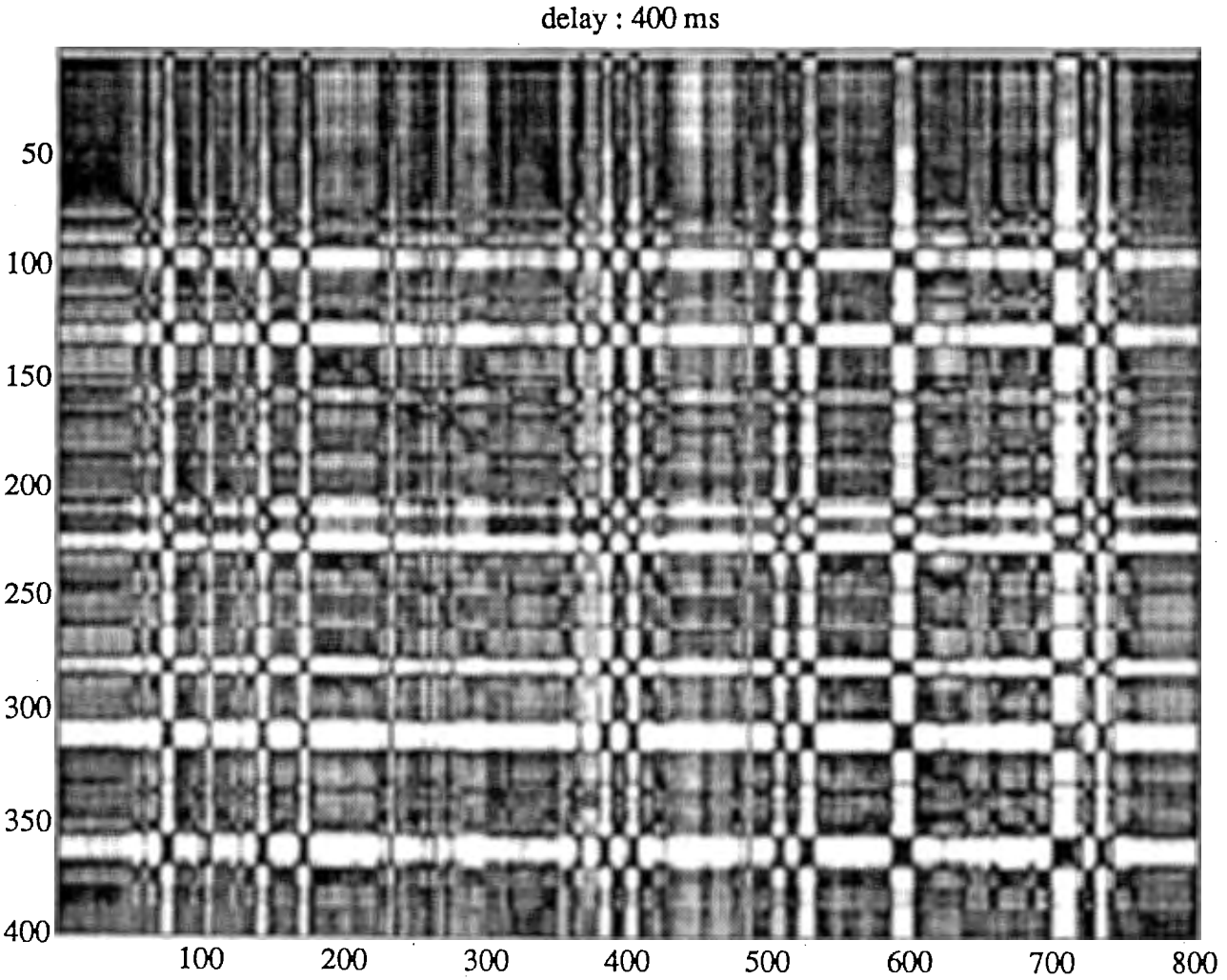
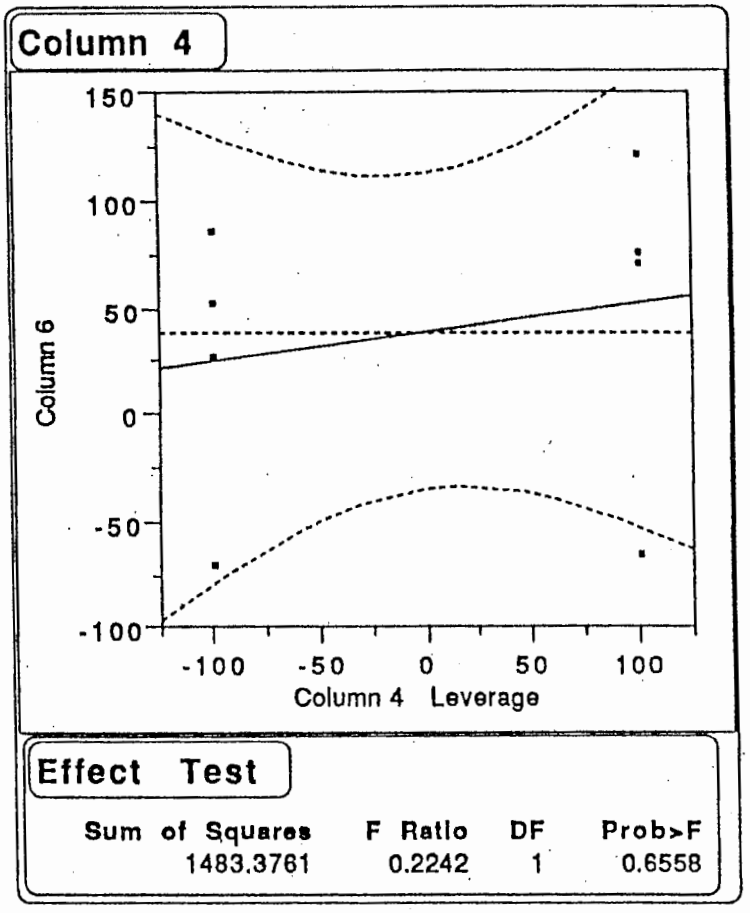


FIG6 基準音声と遅延時間400(ms)のときの音声とのケプストラムの距離.



Column4　変換ピッチ（cent）
Column6　強度　（相対値）

FIG7 定常的なピッチ変換における
変換ピッチと強度の関係

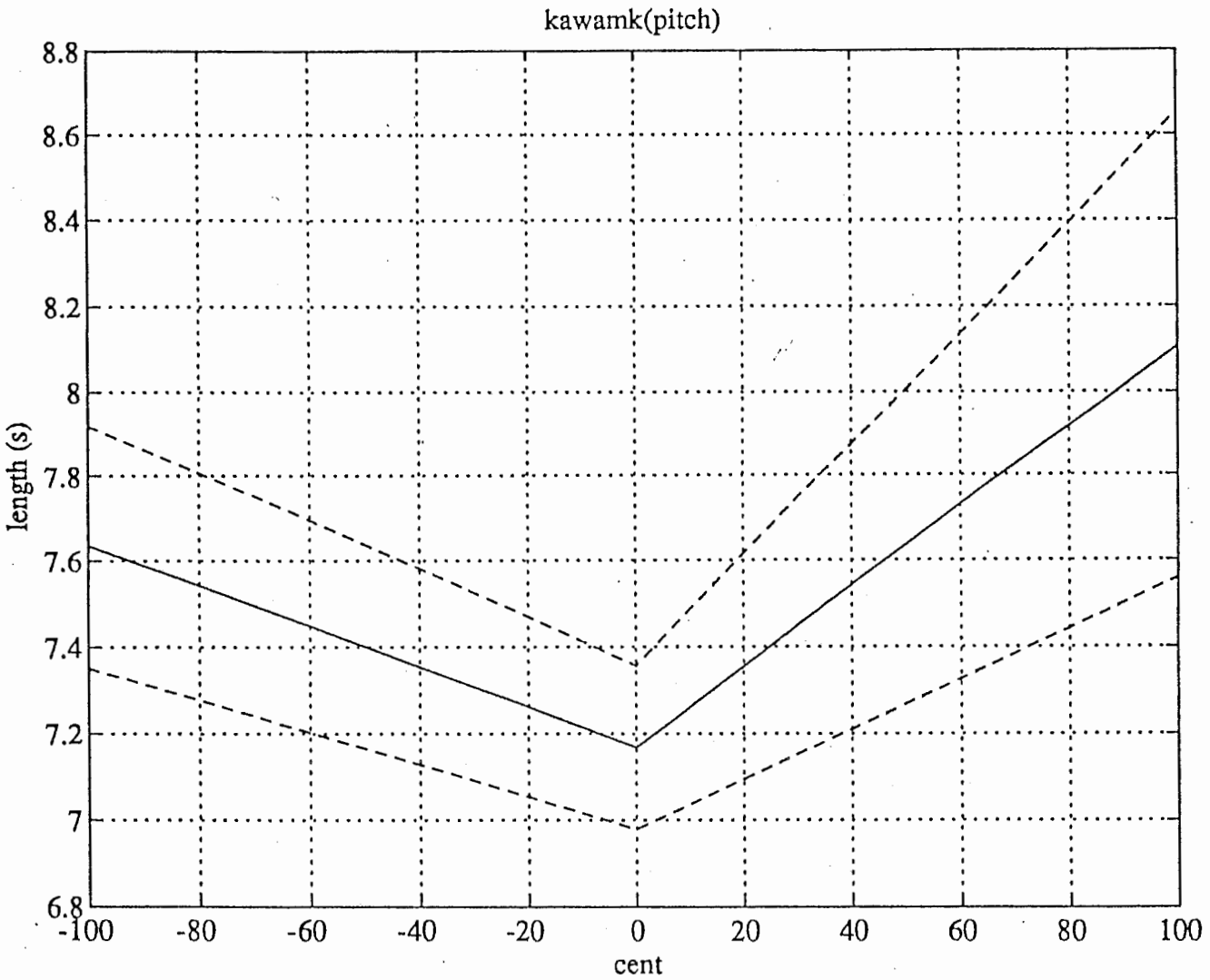
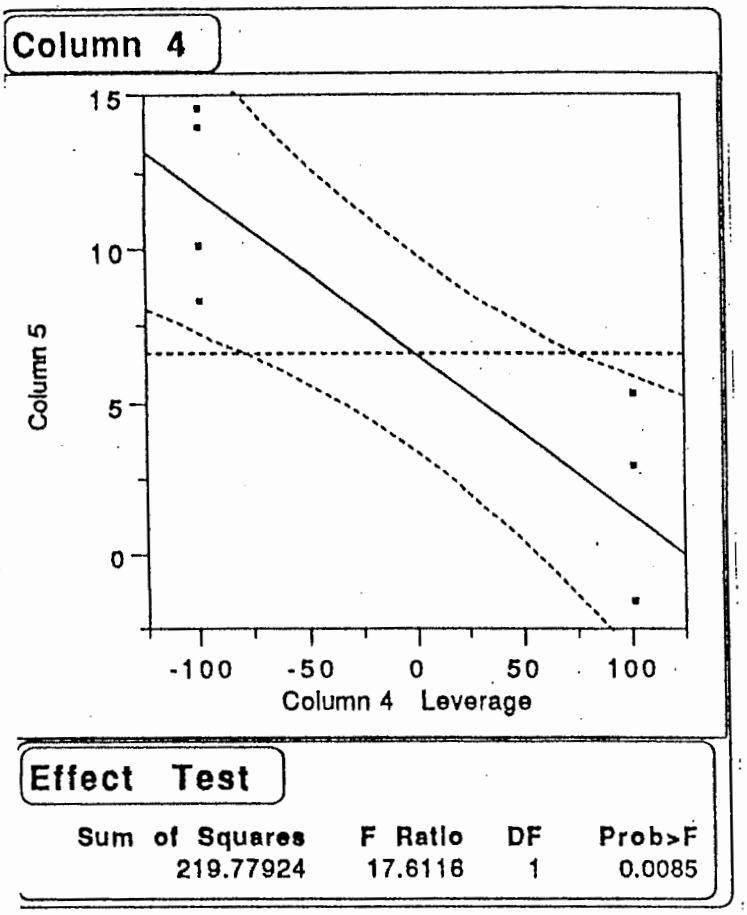
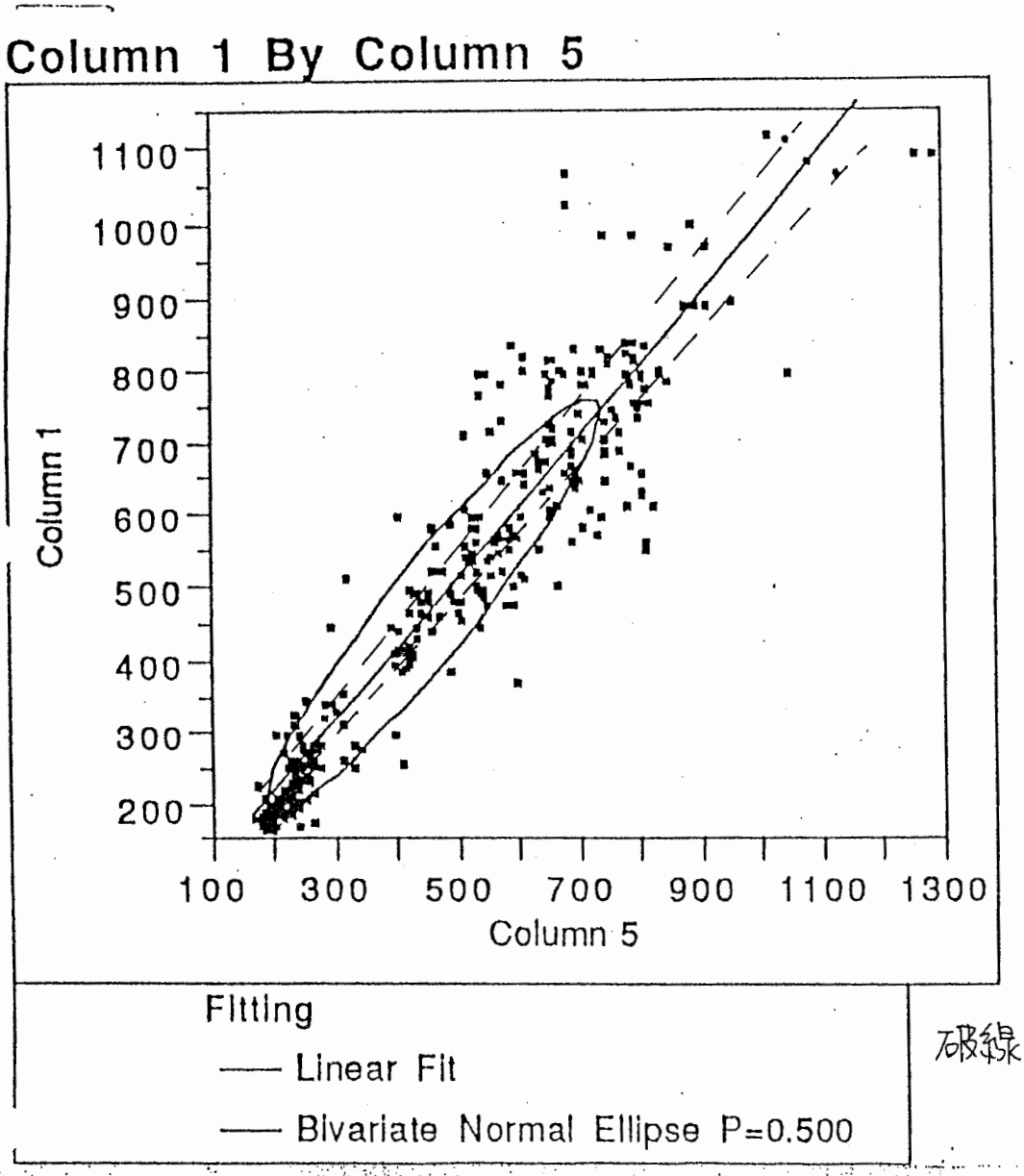


FIG 8 定常的なピッチ変換における
変換ピッチと継続時間長との関係



Column 4 変換ピッチ (cent)
Column 5 ピッチ (相対値) (Hz)

Fig9 定常的なピッチ変換における
変換ピッチとピッチとの関係



破線：予想される
ピッチ変換における
影響.

Column 1　基準音声2回目の第一ホルマント
Column 5　基準音声1回目の第一ホルマント

FIG 10　DTWで時間整合したときの基準音声
1回目と2回目のホルマント.

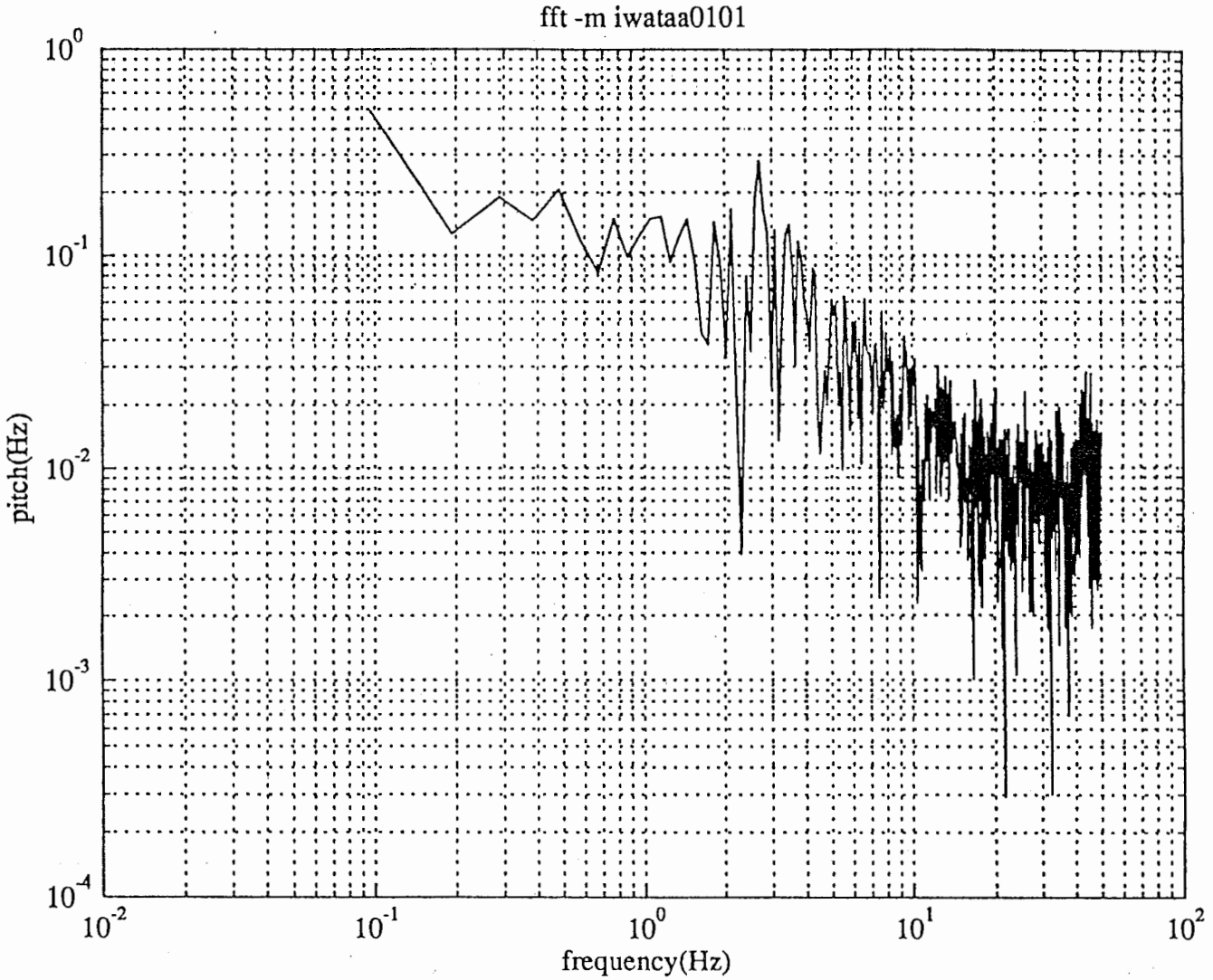


FIG11 基準音声のピッチのフーリエ変換結果

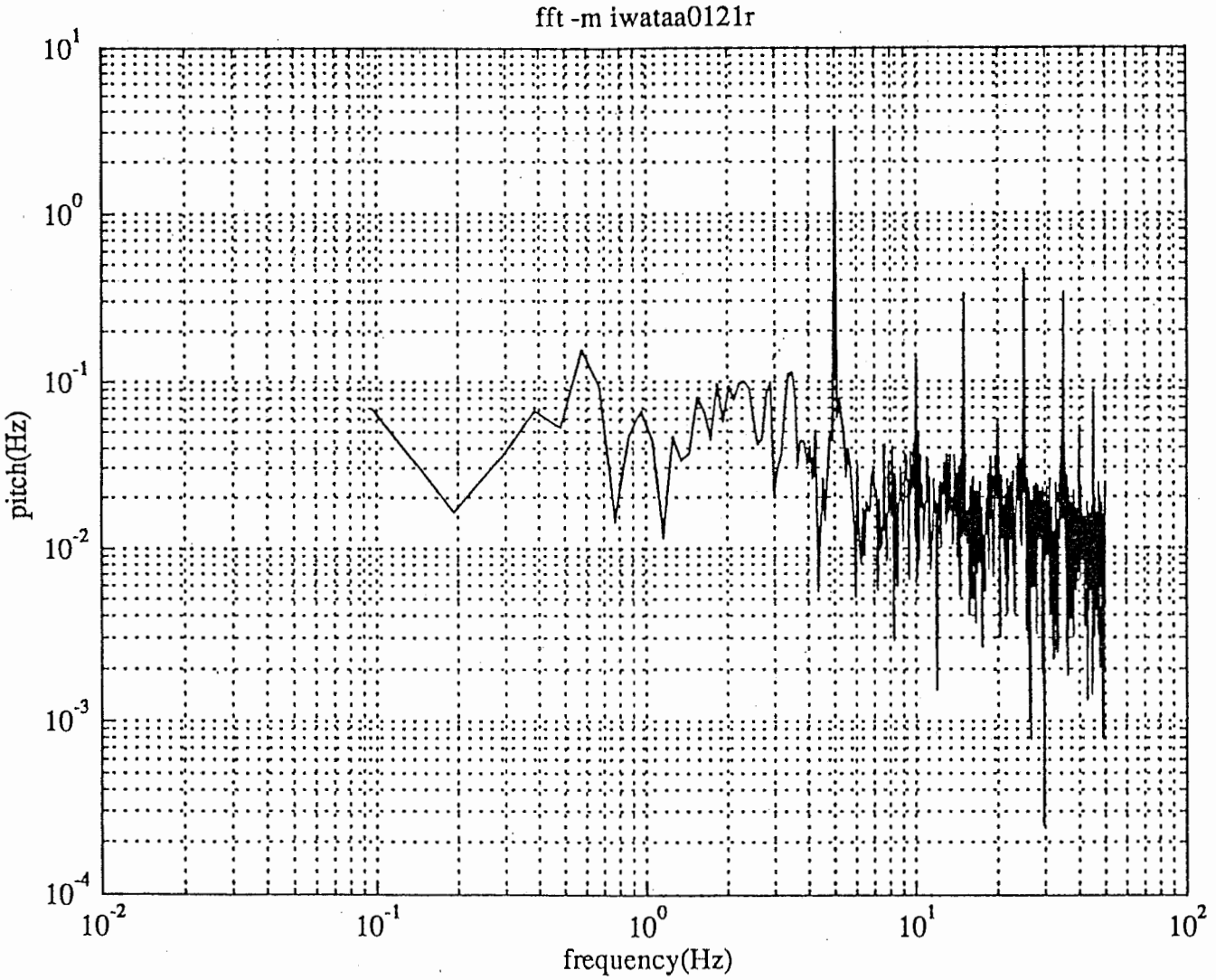


FIG12 ピッチ変換周波数5Hzのときの
フィードバックされる音声のフーリエ変換結果.

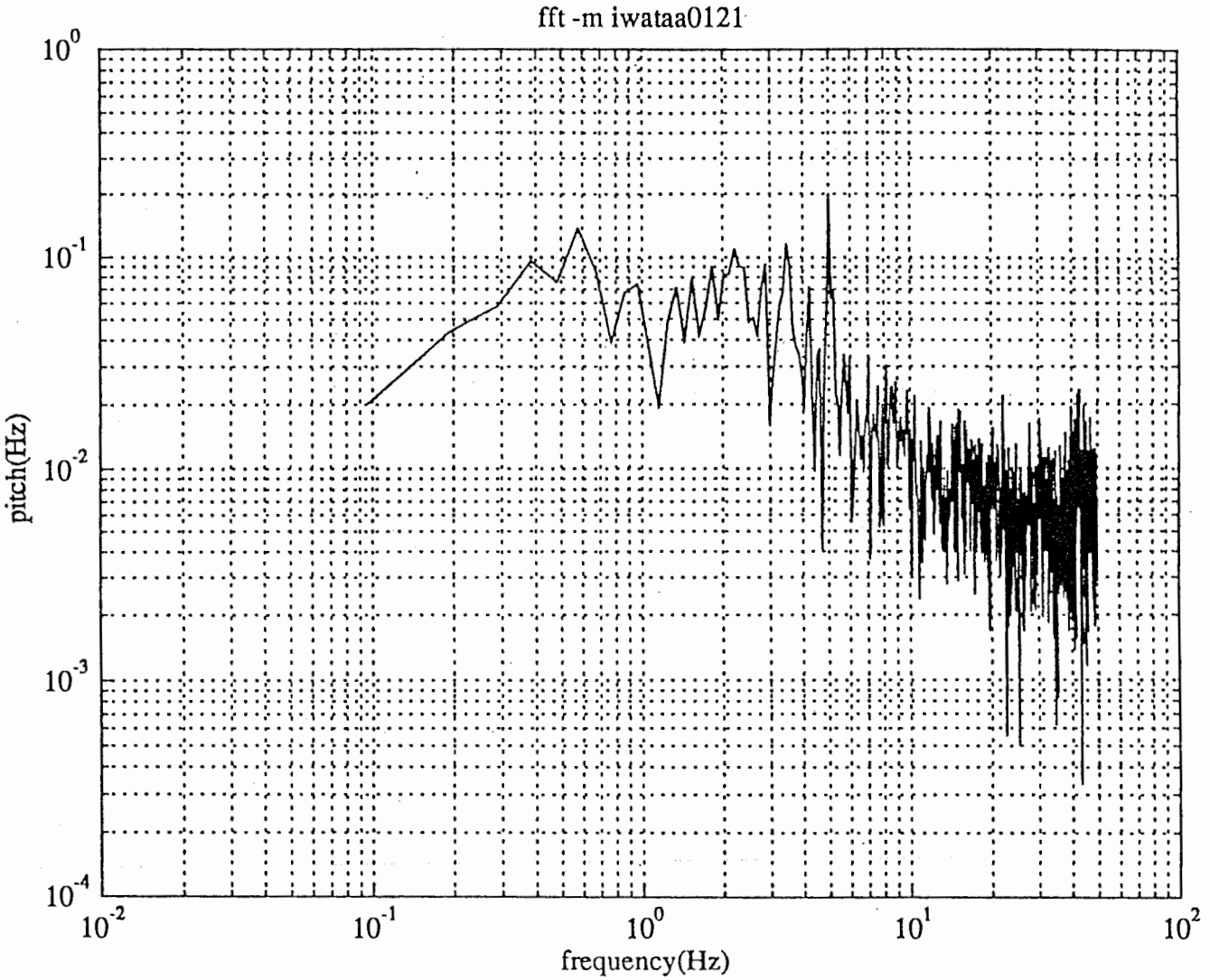


FIG 13. ピッチ変換周波数5Hzのときの
発声した音声のフーリエ変換の結果.

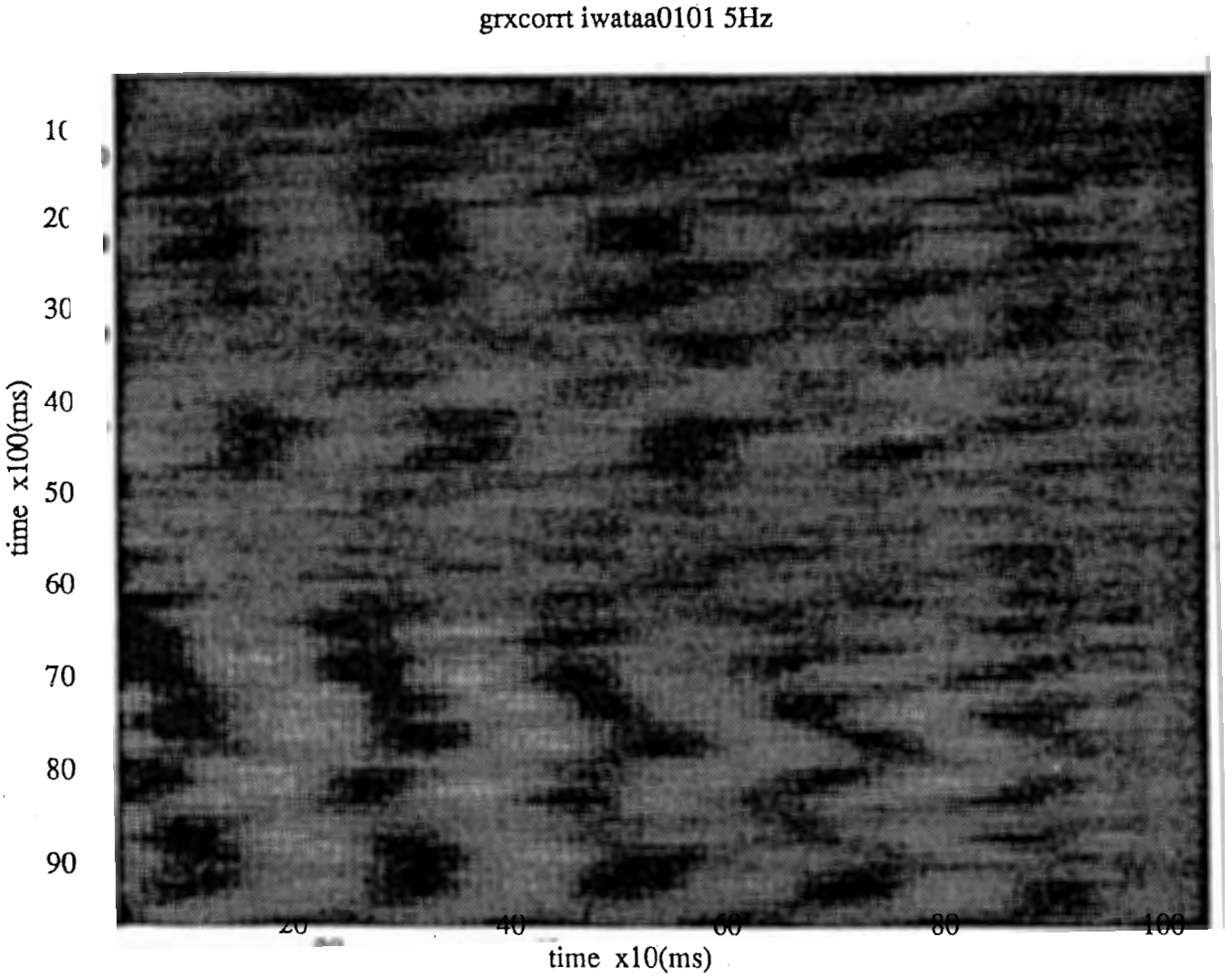


FIG 14 基準音声と5Hzの正弦波との相互相関

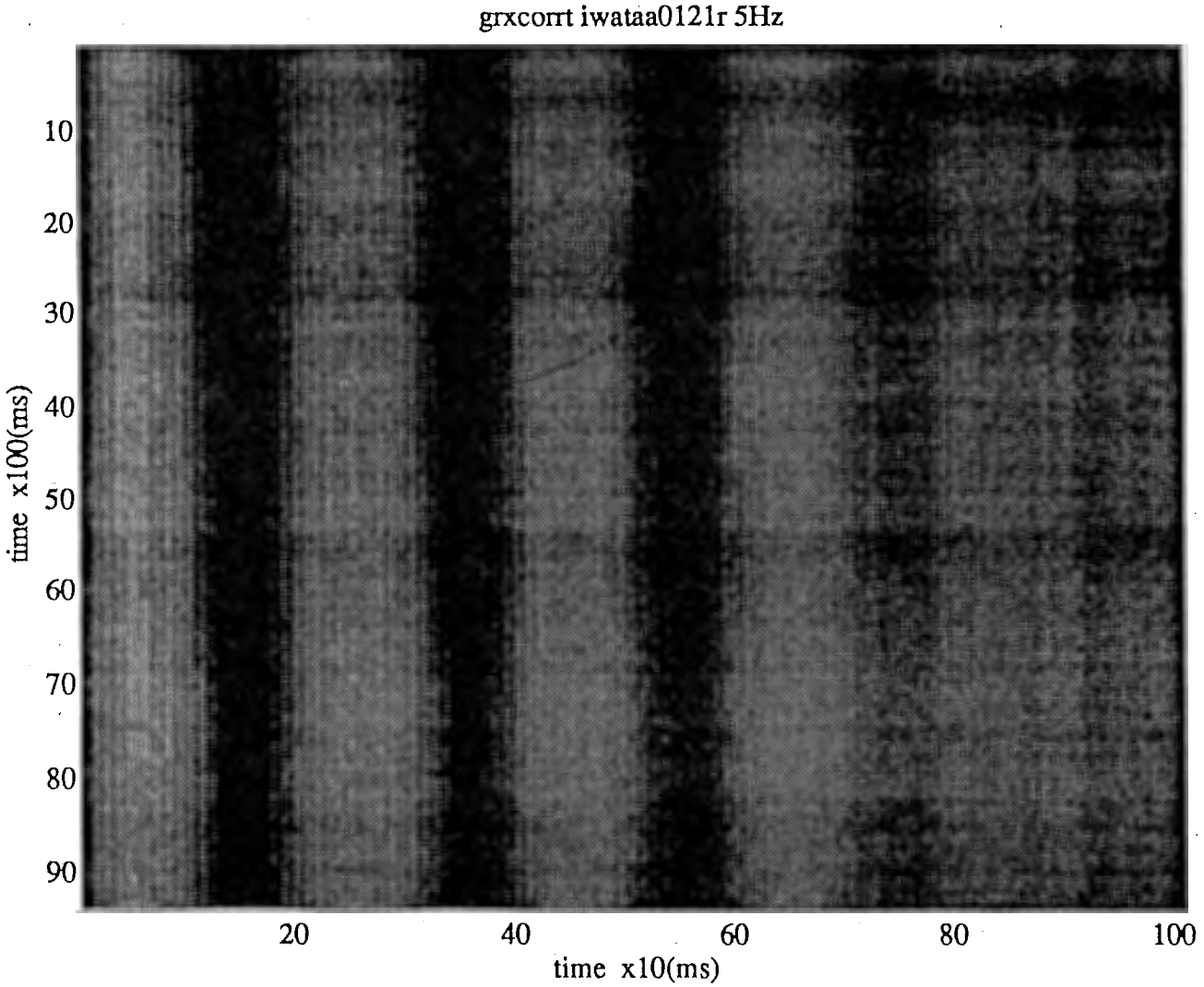


FIG15. ピッチ変換周波数 5Hz のときの
フィードバックされる音声と 5Hz の正弦波との相互相関

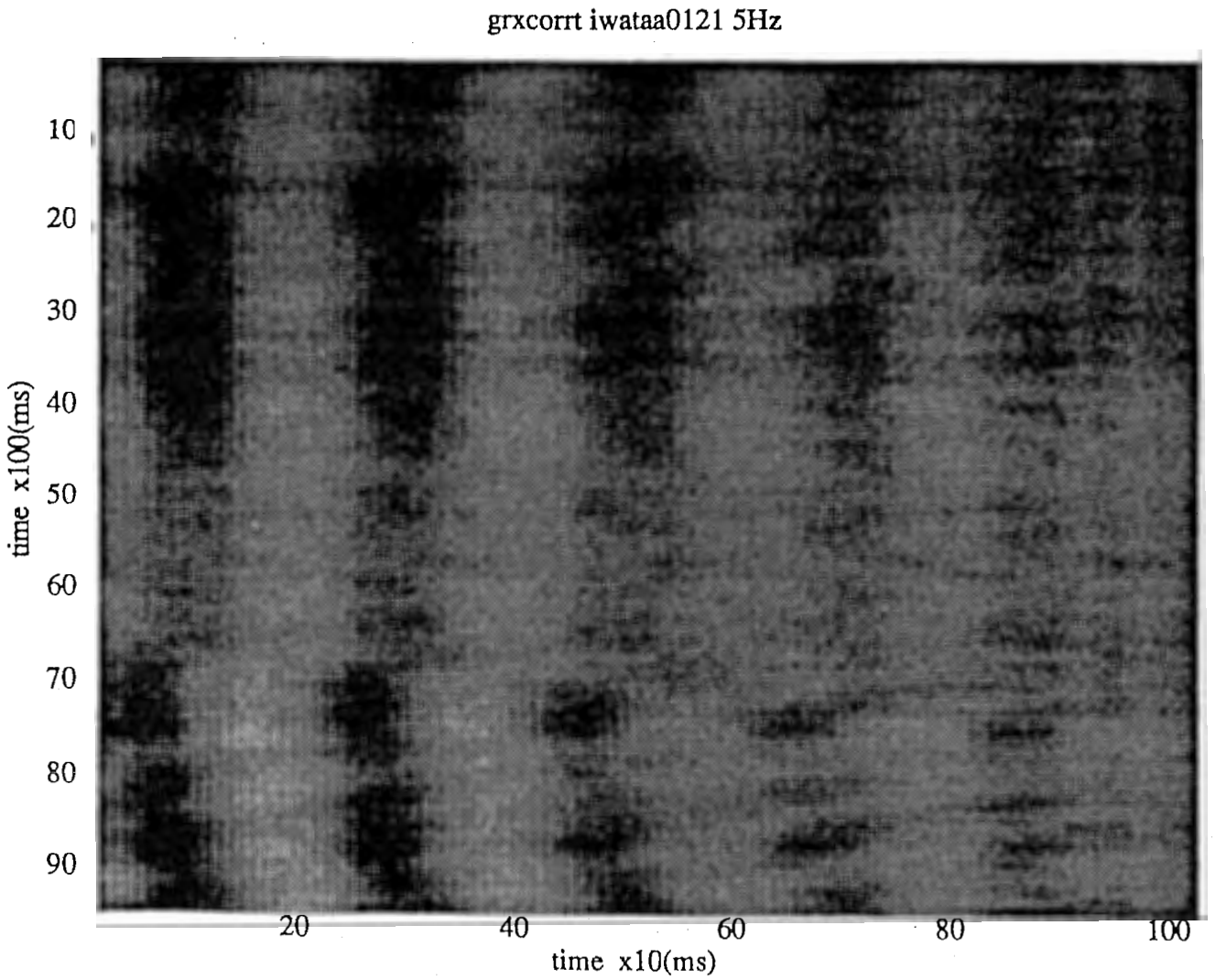
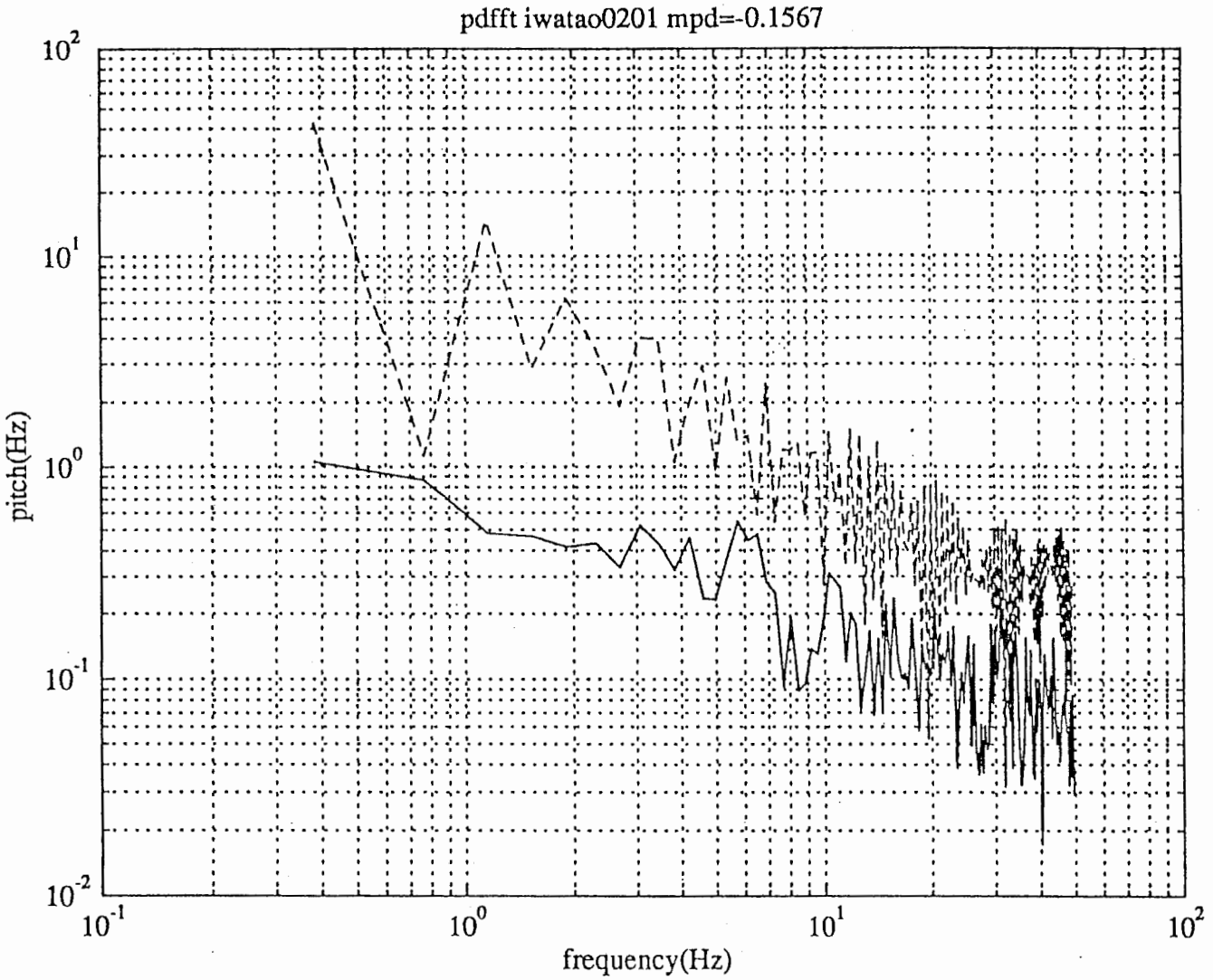


FIG 16 ピッチ変換周波数5Hzのときの
発声した音声と5Hzの正弦波との相互相関



破線：基準音声の2回目のピッチのフーリエ変換結果、
実線：基準音声の2回目からDTWにより1回目のピッチを引いたもののフーリエ変換結果.

FIG17 DTWによるS/N比の向上

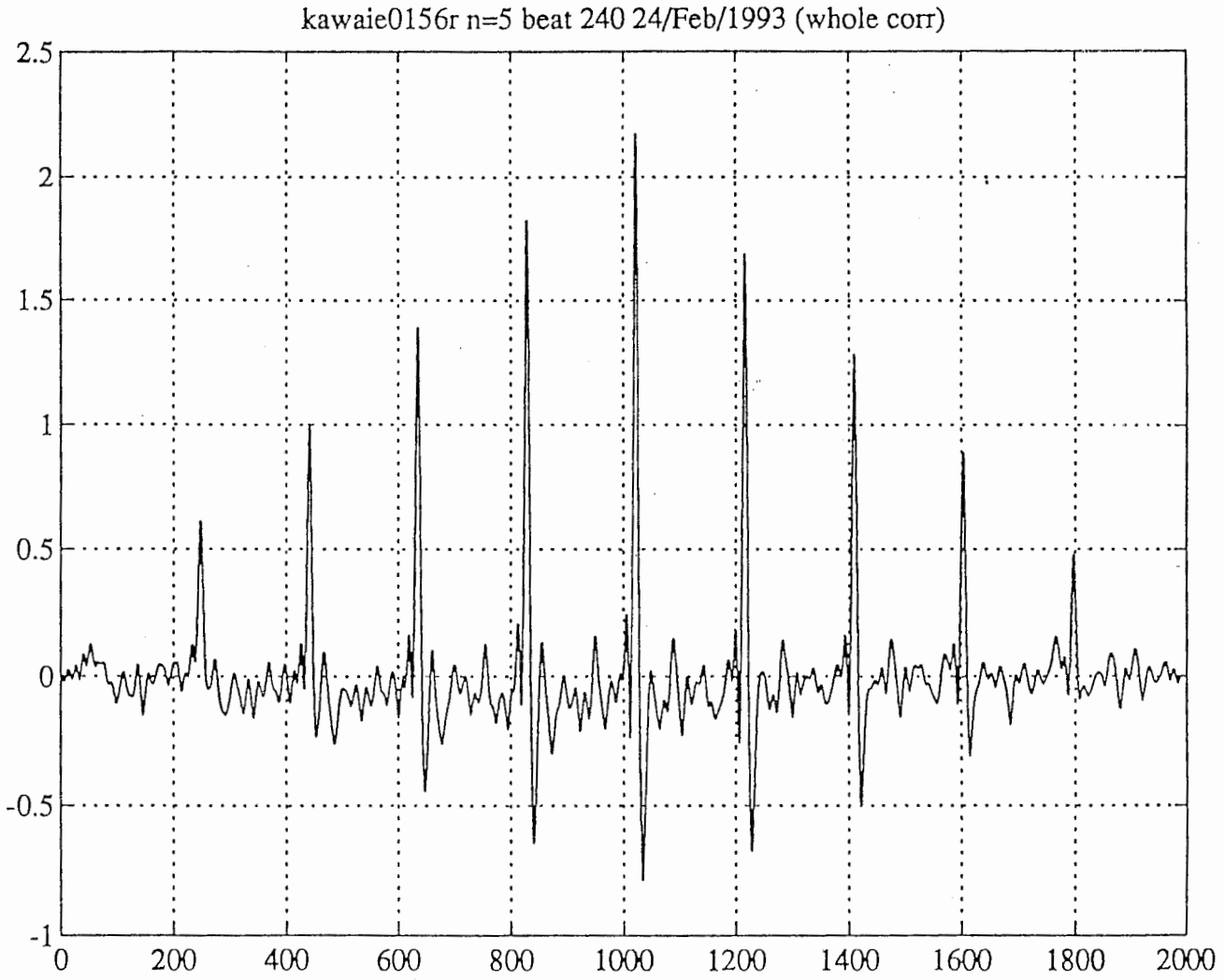


FIG 18 フィードバック音声とM系列との
相互相関

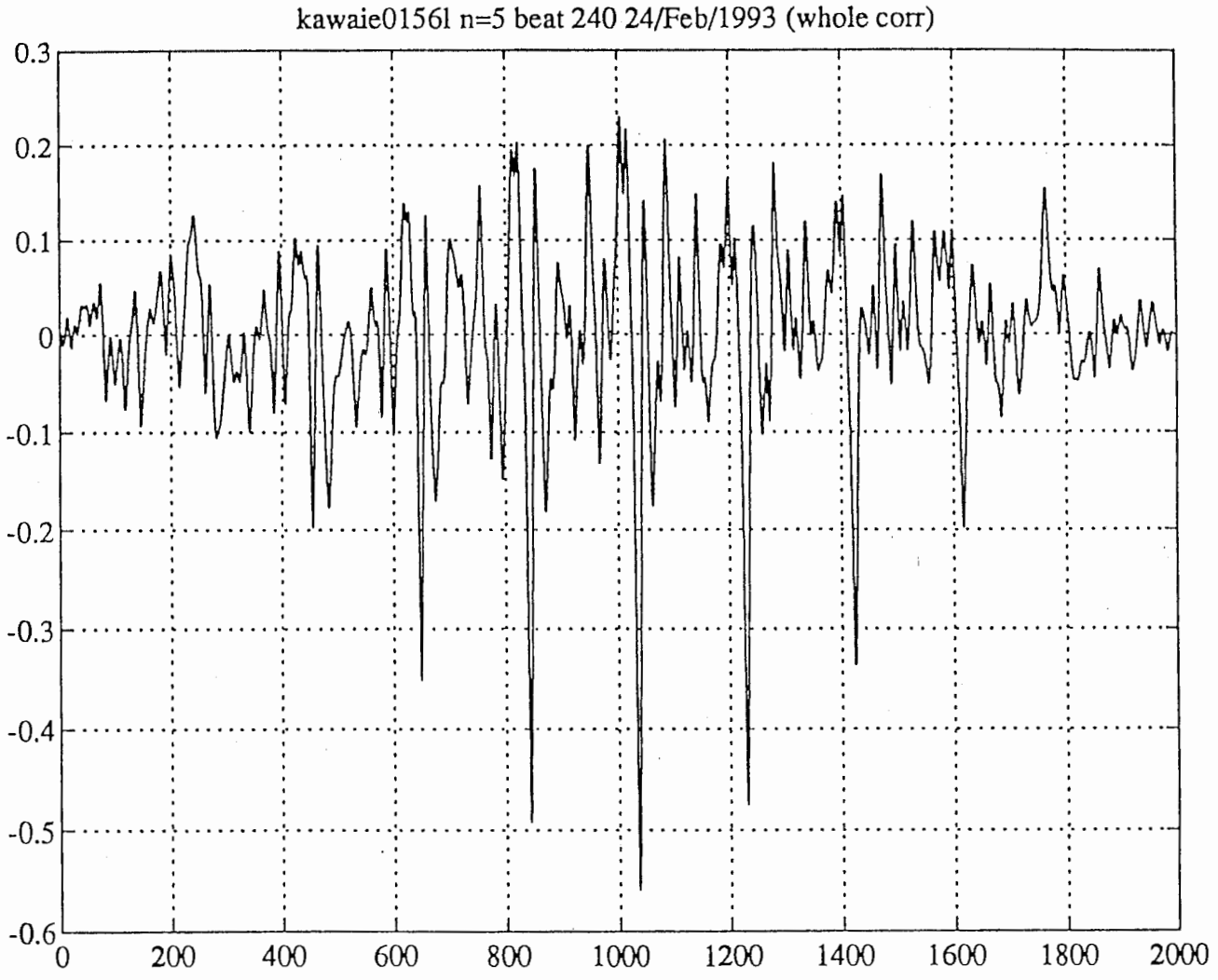
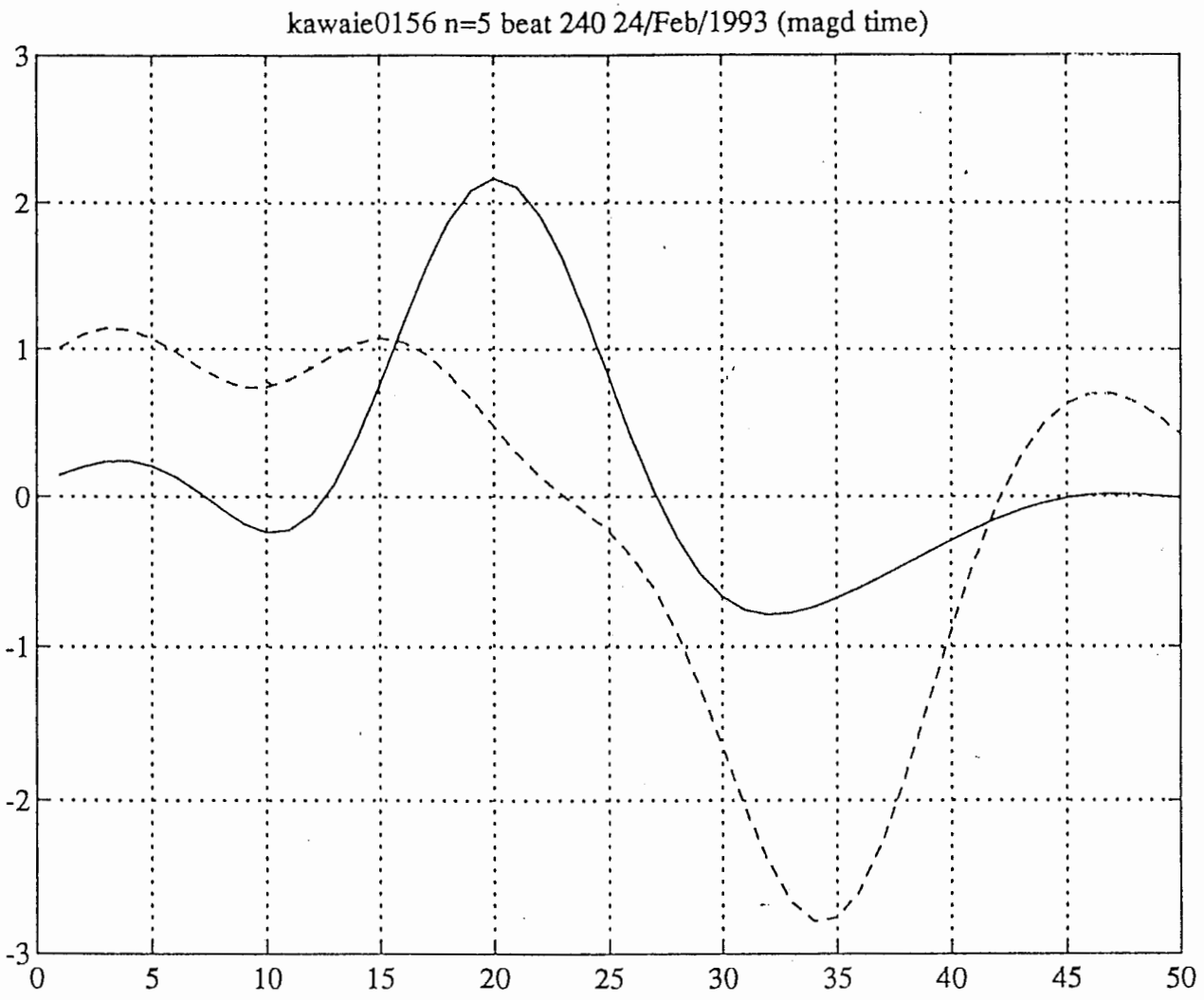


FIG 9 発声している音声とH系列との相互相関



実線：フィードバック音声とM系列との相互相関

破線：発声している音声とM系列との相互相関

FIG 20 パルスのピーク付近の拡大図.

変換聴覚フィードバック

(TAF: Transformed Auditory Feedback)

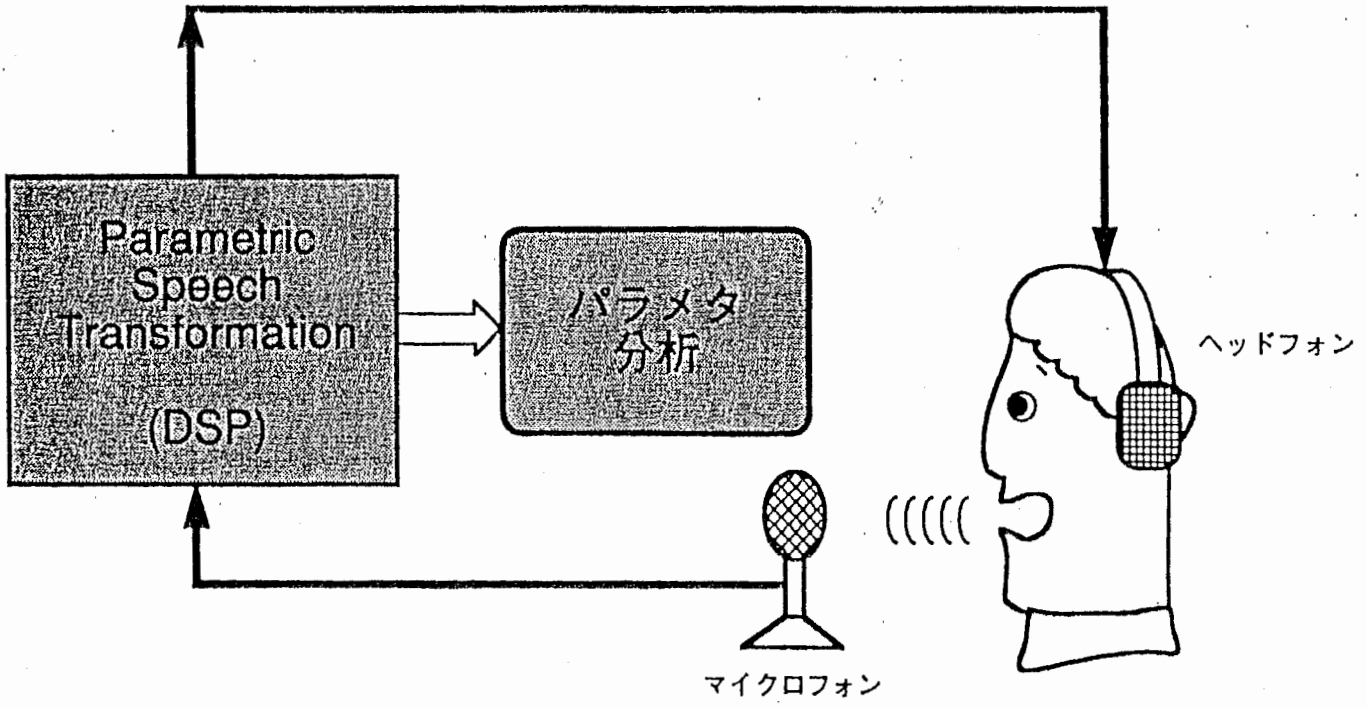
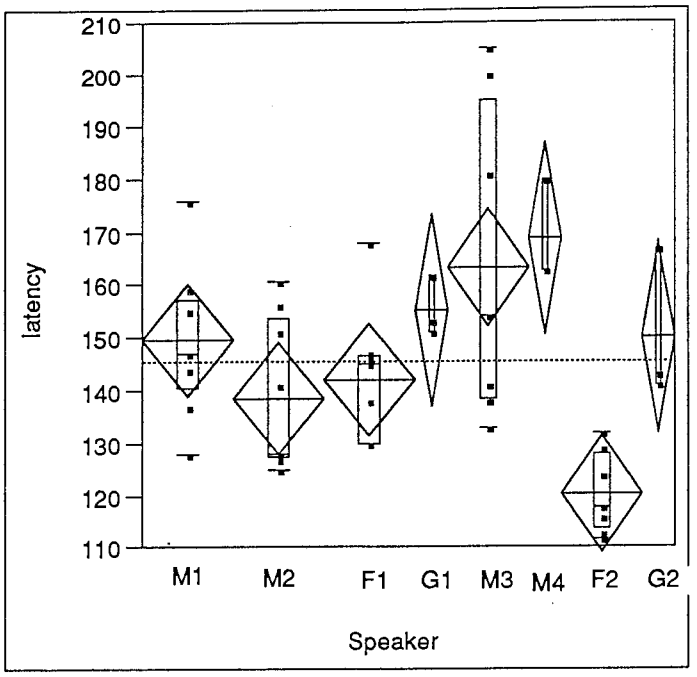


FIG 21　　変換聴覚フィードバックの模式図.



Means with confid. interval
Quantiles: 90%,75%,50%,25%,10%

Means

Summary of Fit

Rsquare	0.482884
Root Mean Square Error	15.83014
Mean of Response	145.3846
Observations (or Sum Wgts)	52

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	10296.197	1470.89	5.8696
Error	44	11026.111	250.59	Prob>F
C Total	51	21322.308		0.0001

Mean Estimates

Level	number	Mean	Std Error
M1	9	149.444	5.2767
M2	9	138.333	5.2767
F1	9	142.111	5.2767
G1	3	155.333	9.1395
M3	8	163.250	5.5968
M4	3	168.667	9.1395
F2	8	120.250	5.5968
G2	3	150.333	9.1395

図２２　発声者による応答時間の変化
（M：男性、F：女性、G：女児）

付録1.

実験系の特性

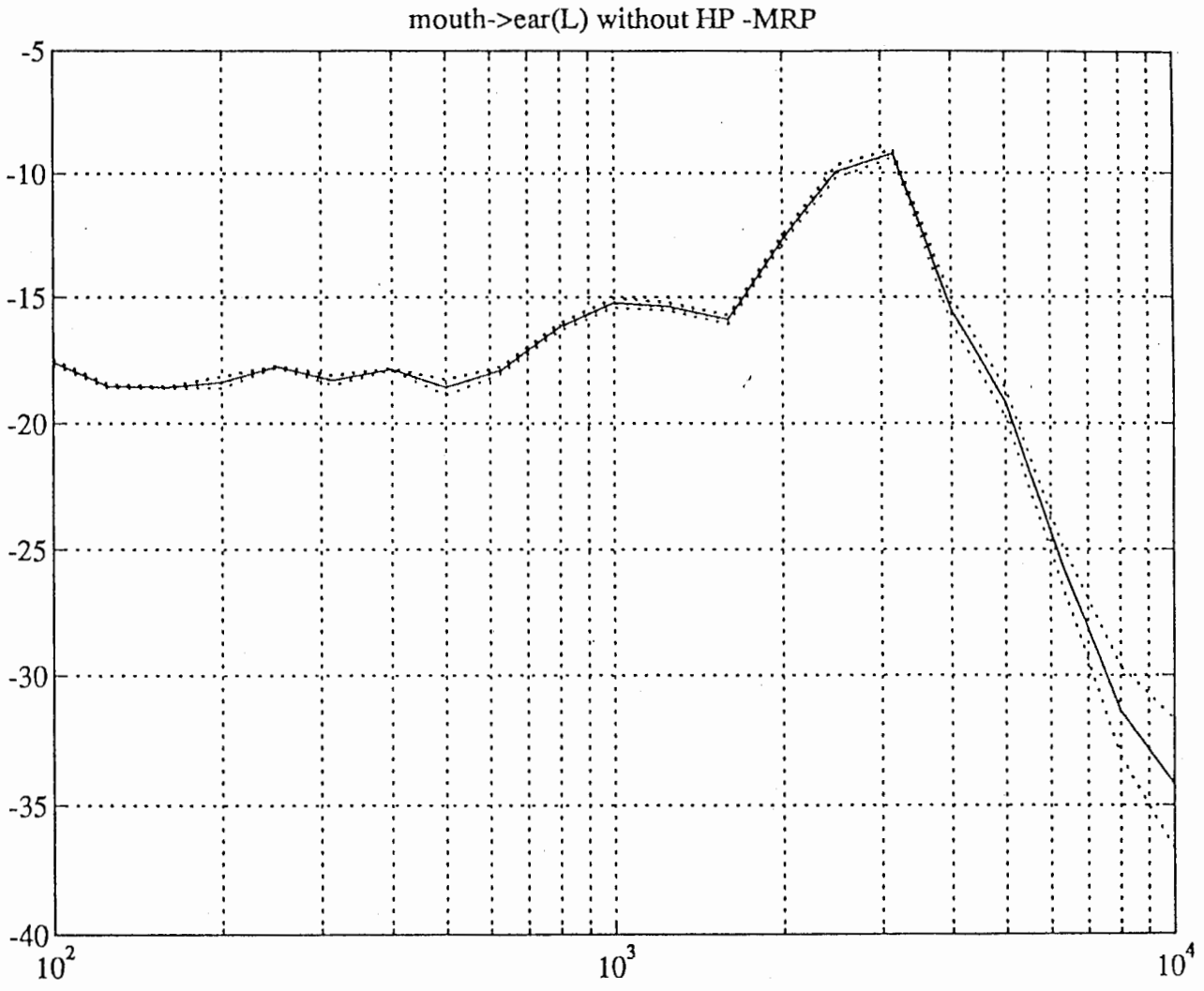


FIG 22 ヘッドフォン非装着時の口から耳への伝達特性.

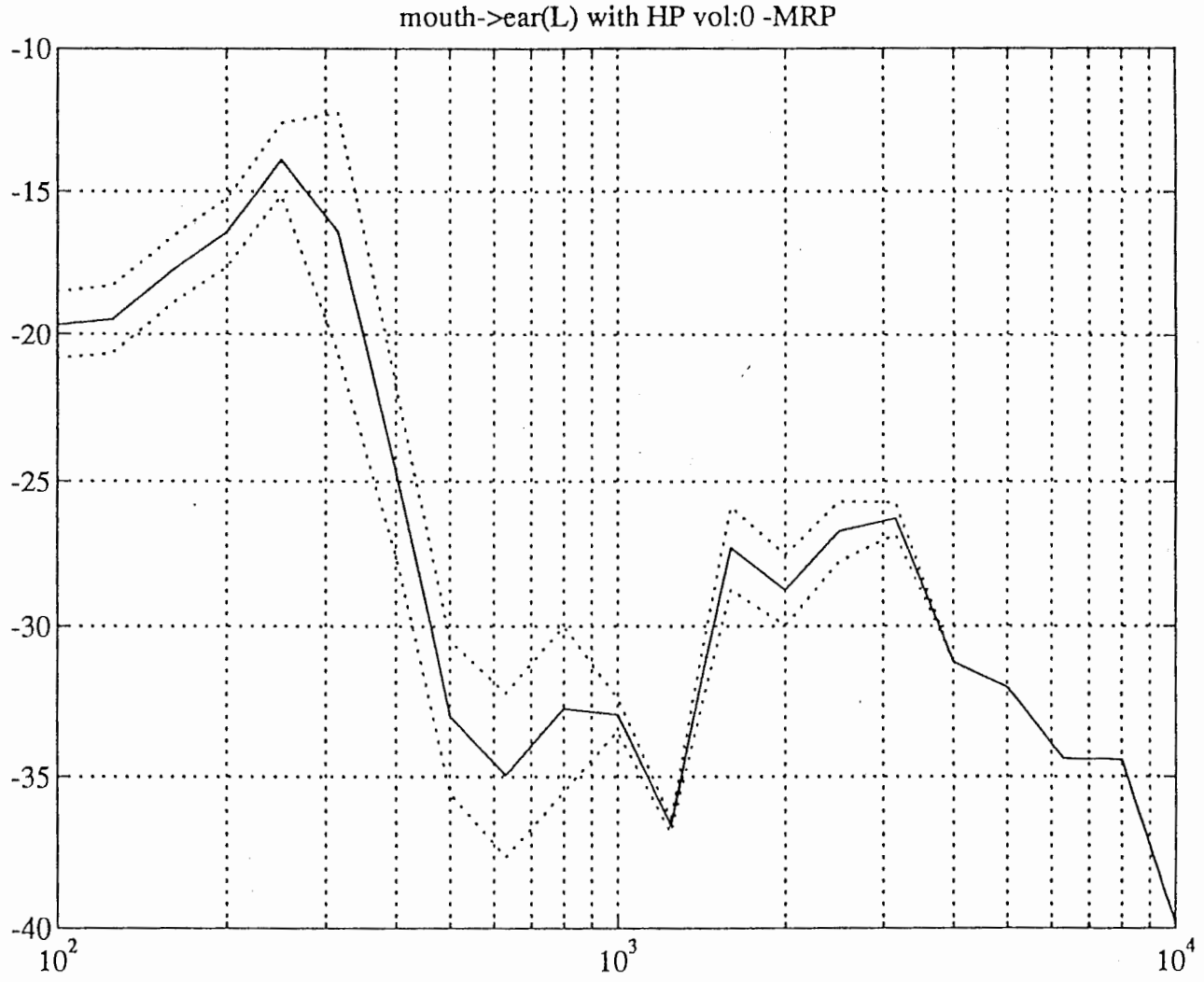


FIG 23 ヘッドフォン装着時の口から耳への伝達特性.

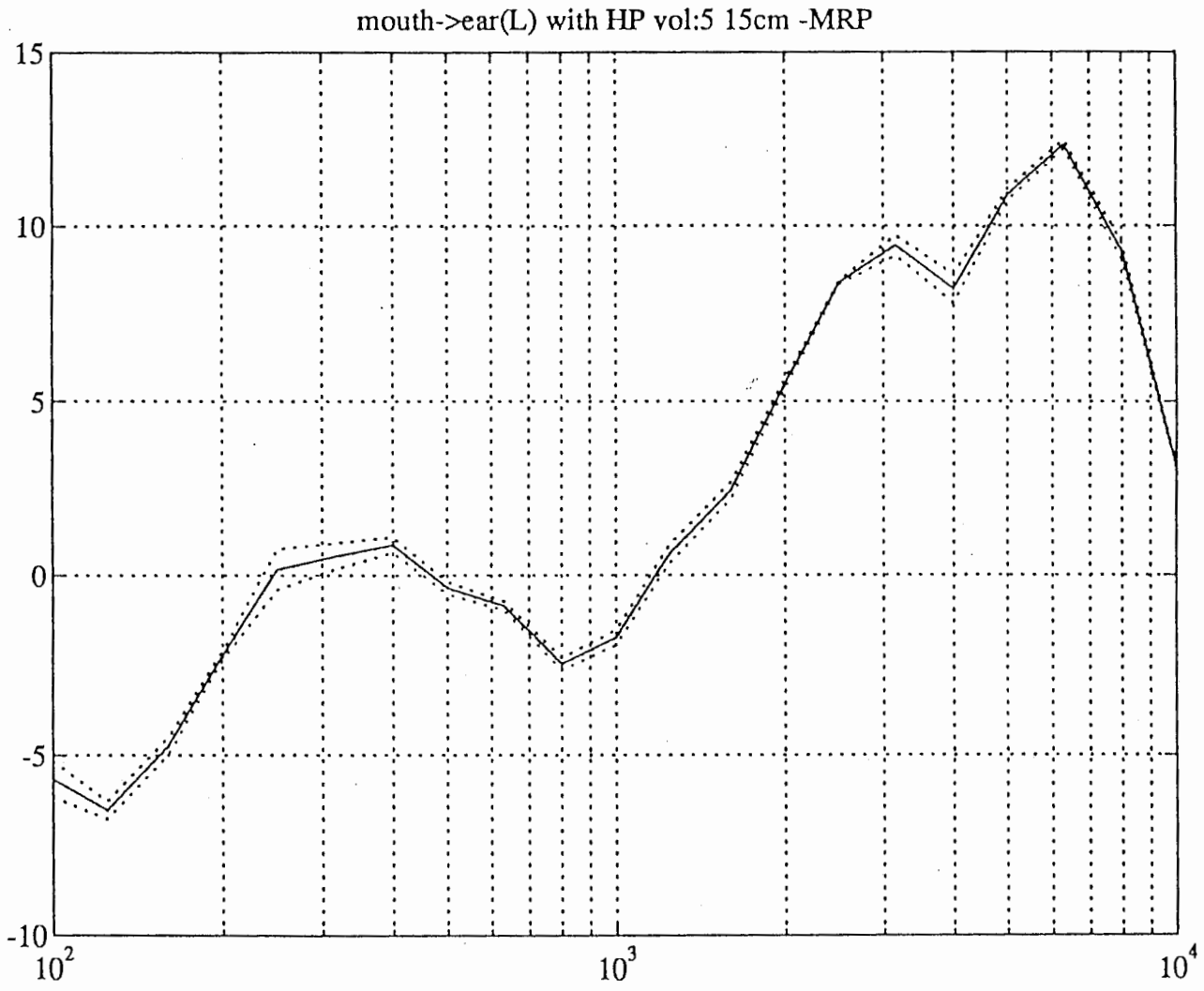


FIG24. 聴覚フィードバック系における
口から耳への伝達特性.