音声の基本周波数パターン生成過程モデルのパラメータ自動抽出法の評価

成澤 修一[1]，　峯松 信明[1]，　広瀬 啓吉[2]，　藤崎 博也[3]

[1] 東京大学大学院 情報理工学系研究科，[2] 東京大学大学院 新領域創成科学研究科，[3] 東京大学名誉教授
Tel.: 03–5841–6393，　Fax.: 03–5841–6648
{narusawa, mine, hirose, fujisaki}@gavo.t.u-tokyo.ac.jp

あらまし　実測の F_0 パターンから藤崎モデルのパラメータを抽出することは解析的には解けない逆問題であり、初期値を与え逐次近似を行う必要がある。筆者らは、実測の F_0 パターンを至るところ連続かつ微分可能な区分的 3 次曲線で近似することにより、この逆問題を連立 1 次方程式を解く問題に帰着させ、有効な結果を既に得ている。本稿では、既報の手法の問題点をあげ、その解決策を提案するとともに、実験的見地から提案手法の有効性について検討する。

キーワード　基本周波数パターン、生成過程モデル、パラメータ自動抽出、手法の評価

Evaluation of a Method for Automatic Extraction of Parameters of the Fundamental Frequency Contour Generation Model

Shuichi Narusawa[1]，　Nobuaki Minematsu[1]，　Keikichi Hirose[2]　and　Hiroya Fujisaki[3]

[1] Graduate School of Information Science and Technology, University of Tokyo,
[2] Graduate School of Frontier Sciences, University of Tokyo,　[3]Prof. Emeritus, University of Tokyo
Tel.: 03–5841–6393，　Fax.: 03–5841–6648
{narusawa, mine, hirose, fujisaki}@gavo.t.u-tokyo.ac.jp

Abstract　The extraction of Fujisaki model parameters from an observed contour requires an iterative process starting from a set of initial parameter values. In order to guarantee a rapid convergence to an optimum solution, the values should be appropriate ones. We already have developed a method of automatically extracting these from given F_0 contours, and applied it to Japanese sentences with rather good results. The method is based on approximating an observed contour by a continuous curve differentiable everywhere. The present paper modifies the method by some tune-ups and the modification is experimentally shown to be valid.

Key words　Fundamental frequency contours, Model of the generation process of the fundamental frequency contours, Automatic extraction of the model parameters, Evaluation

1. はじめに

音声の韻律的特徴は、本来文字言語にも含まれる辞書・統語・意味・談話などの情報 (言語情報) だけでなく、文字言語には陽に含まれない話者の意図や態度などに関する情報 (パラ言語情報) や、話者の個体差、性別、年齢、感情などに関する情報 (非言語情報) をも含んでいる。高品質な音声合成を実現するためには、これらの情報を考慮した規則の獲得が不可欠であるが、まずはこれらの情報と韻律パラメータとの関係を大量の音声データの分析に基づいて明らかにすることが必要である。

音声の韻律的特徴を表現する客観的・物理的な量として、基本周波数パターン (以下、F_0 パターン) は言語の構文や意味の伝達に重要な役割を果たしている。F_0 パターンは日本語を含む多くの言語の音声の抑揚を表し、一般に単語レベルのアクセントに対応する局所的で急激な起伏と、句・節・文レベルの、より広い範囲にわたる緩やかな起伏とから成るが、この F_0 パターンが生成される過程のモデルとそのモデルを用いる分析手法は、藤崎らによって初めて考案された [1]。このモデルでは発話の言語学的情報と密接に関係のある少数のパラメータで実際に観測される F_0 パターンを極めて良く近似できることが知られている [2]。

一方、このモデルを音声言語処理に利用するためには、観測された F_0 パターンを最良近似するモデルのパラメータを抽出し、抽出されたパラメータと言語情報等との関係を調べ、把握することが重要である。しかしながら、観測された F_0 パターンからモデルのパラメータを抽出することは、いわゆる逆問題であって、この場合には解析的に解くことは出来ず、モデルのパラメータの初期値を出発点とした逐次近似を必要とする。もし、適切な初期値を自動的に求めることが可能になれば、大量の音声データの分析が容易となり、生成過程モデルに基づく大規模な韻律コーパスの構築も可能になると考えられる。

以上のような観点から、筆者らは、実測の F_0 パターンを至るところ連続かつ微分可能な区分的 3 次曲線で近似することにより、この逆問題を連立 1 次方程式を解く問題に帰着させ、有効な結果を既に得ているが [3]、本稿では、この手法に更に改良を加えた結果について報告し、朗読調の音声資料を分析した結果および聴取実験結果から、本手法の有効性を検討する。

2. 既報の手法とその問題点

2.1 パラメータ自動抽出法の手順

(1) 実測の F_0 パターンに対する前処理

まず、抽出した F_0 パターンに含まれる gross error を、中央値平滑化と閾値処理によって修正する。次に、無声区間前後の F_0 パターンの局所的勾配を利用して microprosody を除去する。さらに、短い無音・無声区間を、無声区間前後の F_0 値と勾配により導かれた連立方程式から算出された係数を用いて 3 次曲線で補間する。最後に、最小二乗法を利用して、F_0 パターン全体を至るところ連続かつ微分可能な区分的 3 次曲線で近似する。本稿では、以下、これを平滑化 F_0 パターンと呼ぶ。

(2) アクセント指令のパラメータの初期値の推定

アクセント指令の生起位置の初期値は、平滑化 F_0 パターンの変曲点の位置から、指令の振幅の初期値は、変曲点での 1 次微係数から推定する。

(3) フレーズ指令のパラメータの初期値の推定

1 発話を幾つかの分析区間に分け、各分析区間ごとに以下の処理を発話終了時点まで left-to-right に行う。ただし、$1/\alpha$ (= 0.33ms) 以上の無音区間を休止とみなし、次の休止が現れるまでを 1 つの分析区間とする。推定したアクセント指令からアクセント成分を生成し、それを平滑化 F_0 パターンから差引いてパターン A を求める。パターン A に対する AbS 処理を行い、まず、分析区間の初頭のフレーズ指令を推定する。次に、既に推定が完了したフレーズ指令を用いてフレーズ成分を生成し、それをパターン A から差引いてパターン B を求め、その積分値を計算する。その値が閾値を超えたら、より低い閾値を超える時点まで戻り、そこに新たにフレーズ指令を挿入する。パターン B に対するフレーズ指令のみの AbS 処理を行い、その新たなフレーズ指令の生起位置・大きさの初期値を決定する。積分値をリセットして、再びフレーズ成分と残差パターンとの誤差の積分値を計算する。

(4) パラメータの最適化

推定した指令の初期値から F_0 パターンを生成し、実測の F_0 パターンとの平均二乗誤差が最小となるように、まず全てのフレーズ指令に対して、生起位置、大きさの順に AbS 処理を行ってパラメータを微小変化させ、最適値を求める。その後に全てのアクセント指令に対して、生起位置、大きさの順に AbS 処理を行う。

2.2 問題点

(1) gross error の修正が不十分である

gross error であるかどうかの判断が難しい F_0 については、アラインメント情報、音声の短時間平均パワー、自己相関関数の値などを利用して処理をしたが、gross error ではない F_0 も修正してしまい、十分な結果は得られていない。

(2) microprosody の除去が不十分である

無声区間前後の F_0 パターンの局所的勾配から microprosody であるかどうかを判断していたが、大幅に除去してしまったり、除去しきれていなかったりと、処理精度が悪い。

(3) 正解と異なる形でアクセント指令を推定する

例えば、2 単語の発話間隔が短かいときは、1 つ目のアクセント指令の立下り位置と 2 つ目のアクセント指令の立上り位置とが一致するが、これを 1 つの長いアクセント指令として推定してしまう。

(4) AbS 処理後に不正解となる

フレーズ指令の次にアクセント指令に対して AbS 処理するため、F_0 パターンが平板な部分では、フレーズ成分だけで合わせようとしてしまい、アクセント成分がほとんど無くなってしまうため、持続時間が短く大きさが小さいアクセント指令が生じてしまう。

3. パラメータ自動抽出法の改良点

3.1 gross error の修正法

F_0 と音声の短時間平均パワーは、本来相関のある物理量であるので、F_0 パターンの変化とパワーパターンの変化との相関性に着目し、gross error の判定を行う。まず、2 フレーム相当の矩形窓の短時間平均パワーを算出し、最大値を求める。その値よりも 30dB 以下のパワーを有するフレームは無声であるとみなし、F_0 を除去する。次に、それ以外のフレームで、周辺の値に比べ F_0 は高い (または低い) がパワーは同程度であったり、パワーの大きさに比べ F_0 が大きい (または小さい) など、F_0 とパワーとの相関性に反する変化をしているフレームに抽出されている F_0 を gross error の候補であるとして、既報の手法と同様の処理を行い、gross error とみなされた場合のみ、周辺の F_0 値を用いて補間または外挿する。

3.2 短い無音・無声区間の補間法

無声区間 $[t_{i+1}, t_{j-1}]$ 前後に存在する N 個の F_0 を用いて、区間 $[t_{i-N}, t_{j+N}]$ での F_0 の値を以下の 3 次曲線

$$\log_e F_0(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 \qquad (1)$$

で近似し、上式から算出される F_0 を用いて無声区間を補間する。ここで、式 (1) の係数 a_0, a_1, a_2, a_3 に関して、既報の手法では、無声区間前後の F_0 値と F_0 パターンの局所的勾配とを利用した連立方程式から算出していたが、microprosody の影響を減少するために、区間 $[t_{i-N}, t_{j+N}]$ に存在する F_0 を用いた最小二乗法により導かれる、以下の連立方程式から算出する。ただし、補間式の勾配値がある閾値を超えた場合は、無声区間を直線補間する。

$$\begin{cases} \displaystyle\sum_{n=i-N}^{i} \log_e F_0(t_n) = Na_0 + \sum_{n=i-N}^{i} a_1 t_n \\ \qquad\qquad \displaystyle + \sum_{n=i-N}^{i} a_2 t_n^2 + \sum_{n=i-N}^{i} a_3 t_n^3 \\ \displaystyle\sum_{n=i-N}^{i} \log_e F_0(t_n) t_n = \sum_{n=i-N}^{i} a_0 t_n + \sum_{n=i-N}^{i} a_1 t_n^2 \\ \qquad\qquad \displaystyle + \sum_{n=i-N}^{i} a_2 t_n^3 + \sum_{n=i-N}^{i} a_3 t_n^4 \\ \displaystyle\sum_{m=j}^{j+N} \log_e F_0(t_m) = Na_0 + \sum_{m=j}^{j+N} a_1 t_m \\ \qquad\qquad \displaystyle + \sum_{m=j}^{j+N} a_2 t_m^2 + \sum_{m=j}^{j+N} a_3 t_m^3 \\ \displaystyle\sum_{m=j}^{j+N} \log_e F_0(t_m) t_m = \sum_{m=j}^{j+N} a_0 t_m + \sum_{m=j}^{j+N} a_1 t_m^2 \\ \qquad\qquad \displaystyle + \sum_{m=j}^{j+N} a_2 t_m^3 + \sum_{m=j}^{j+N} a_3 t_m^4 \end{cases} \qquad (2)$$

3.3　アクセント指令の推定法

アクセント指令の生起位置は、平滑化 F_0 パターンの 1 次微係数が正である変曲点 (以下、第 1 種変曲点) と負である変曲点 (以下、第 2 種変曲点) のペアから推定していたが、問題点 (3) を解決するために、第 1 種 (または第 2 種) の変曲点が 2 つ連続した場合は、2 つ目の第 1 種 (または 1 つ目の第 2 種) の変曲点の位置を、1 つ目のアクセント指令の立下り位置および 2 つ目のアクセント指令の立上り位置とする。ただし、第 1 種 (または第 2 種) の変曲点が 3 つ以上連続した場合は、1 次微係数の絶対値の大きい順に 2 つ選び、同様の処理をする。

3.4　パラメータの最適化法

F_0 パターンが平板な部分において、AbS 処理後にアクセント指令の持続時間が短く、振幅が小さくなり、誤りとなってしまう場合は、直前のフレーズ指令の振幅を小さくし、再度、アクセント指令に関する AbS 処理を行う。

4.　パラメータの自動抽出実験

4.1　音声資料とその処理

音声資料として、NHK ラジオ第 1 放送の朗読番組「私の本棚」で男性アナウンサ 1 名による朗読 (1 回分、15 分間、全 85 文) を録音したもの、および ATR503 文 (話者 MHT) の朗読音声を使用した。ただし、NHK85 文は試行錯誤用のデータとして、ATR503 文は評価用のデータとして使用した。

これらの音声を 10kHz、16bit で A/D 変換し、遅れ時間比例窓長の自己相関関数を用いた手法 [4] により、10 ms 間隔で F_0 を抽出した。さらに、F_0 パターンに対して前記の前処理を施したのち、得られた平滑化 F_0 パターンからフレーズ指令とアクセント指令のパラメータの初期値を抽出し、AbS 処理を行いパラメータを最適化した。

4.2　実験結果とその評価

4.2.1　前処理およびパラメータ自動抽出

提案した手法を用いて分析した結果例を図 1 に示す。分析に用いた音声資料は、“いくつかの音だけそこから借用して、ライブの雰囲気を損なわずに編集することができる。” という文である。

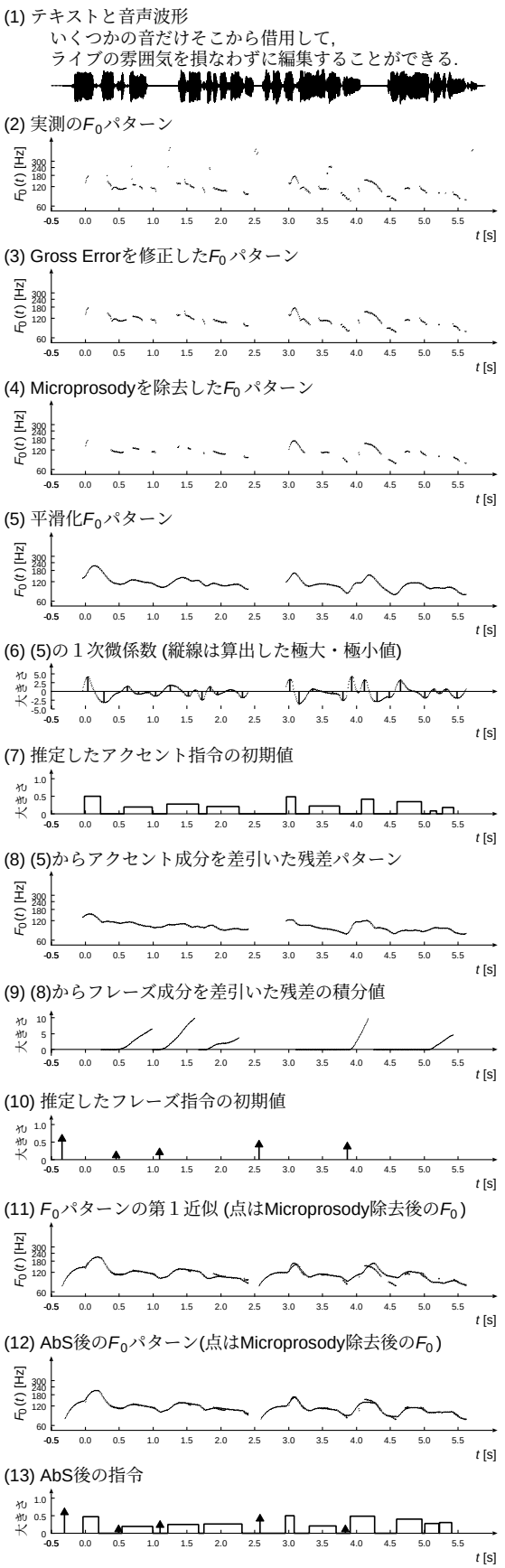

図 1.　前処理およびパラメータ自動抽出の結果例

(1) の音声データから (2) の F_0 パターンを抽出した後、前処理として、gross error の修正を行った結果が (3) であり、それに加えて microprosody の除去を行った結果が (4) であり、更に無声区間の補間および全体の平滑化を行った結果、(5) の平滑化 F_0 パターンを得る。次に (5) を微分して得られる (6) から (7) に示すアクセント指令の初期値を得る。更に (7) のアクセント指令を用いて生成されるアクセント成分を (5) から除去した残差パターンが (8) であり、これまでに推定されたフレーズ成分を更に (8) から差引いたのちの残差の積分値 (9) から (10) に示すフレーズ指令の初期値を得る。この (9) と (10) の処理は反復的に行われるが、ここでは、便宜上、最終結果を 1 つの図として示した。また、推定した (7) と (10) の指令から (11) に示す F_0 パターンを得るが、(7) と (10) の指令を初期値として (5) に対する AbS 処理を行うことにより、(12) の F_0 パターンとそれを生成するフレーズ指令、アクセント指令 (13) を得る。ただし、(11) と (12) 中の点が示すのは (4) の F_0 パターンであり、生成した F_0 パターンがどれほど実測の F_0 パターン (変動を除去したもの) に類似した値であるかを判断するのに役立つものである。(11) の推定された F_0 パターンでは実測値からかなり外れる部分も見られるが、(12) の AbS 処理後の F_0 パターンは、(4) の実測の F_0 パターンから gross error を修正し、microprosody を除去したパターンに極めて類似した形状である。これを定量的に評価するために、この音声資料の有声フレームに対して、(4) と (12) との間の、また (4) と手動推定によるパラメータから生成した F_0 パターンとの間の 1 標本値あたりの平均二乗誤差を計算すると、それぞれ 0.0016、0.0011 であった。これは、百分率誤差の実効値としては、それぞれ約 4.0%、約 3.3%の差に相当し、小さい量である。つまり、本手法を用いて生成した F_0 値は、全体にわたって実測値に極めて近い値が得られているといえる。

4.2.2 NHK85 文と ATR503 文の分析結果

NHK85 文と ATR503 文の分析結果を表 1 に示す。表中の数字の単位は個数であるが、正解率、欠落誤り率および挿入誤り率は、韻律研究に深く携わっているエキスパートが手動で抽出したフレーズ指令数・アクセント指令数を基準として算出した、自動抽出結果の百分率である。

表 1. NHK85 文と ATR503 文の分析結果の比較

	NHK		ATR
	既報手法	改良手法	
フレーズ指令			
手動抽出数	393	393	1733
自動抽出数	418	408	2558
正解数	332	342	1635
(率)	(84.5%)	(87.0%)	(94.4%)
欠落誤り数	64	51	98
(率)	(16.3%)	(13.0%)	(5.7%)
挿入誤り数	86	66	923
(率)	(21.9%)	(16.8%)	(53.3%)
アクセント指令			
手動抽出数	608	608	3372
自動抽出数	603	605	3119
正解数	492	506	2296
(率)	(80.9%)	(83.2%)	(68.1%)
欠落誤り数	129	115	1076
(率)	(21.2%)	(18.9%)	(31.9%)
挿入誤り数	111	99	823
(率)	(18.3%)	(16.3%)	(24.4%)

ここで、アクセント指令に関しては、音声資料を朗読した話者の話速における平均モーラ長を 1 モーラとして、推定したアクセント指令の立上り/立下り位置が、手動分析結果のアクセント指令の立上り/立下り位置を中心とした前後 1 モーラの範囲以内に存在するものだけを正解とした。なお、手動分析において振幅の値がわずかに異なる 2 個のアクセント指令が連続して抽出されたものを、自動分析では 1 個のアクセント指令として抽出した場合にも、それを正解とみなした。また、フレーズ指令に関しては、手動分析結果のフレーズ指令の直前に存在するアクセント指令の立下り位置から直後に存在するアクセント指令の立上り位置までにおいて推定されたものだけを正解とした。

NHK85 文に関しては、改良した手法では、既報の手法と比べて、フレーズ指令に関しては、正解率が 2.5% 増加し、欠落誤り率・挿入誤り率がそれぞれ 3.3%・5.1% 減少した。アクセント指令に関しては、正解率が 2.3% 増加し、欠落誤り率・挿入誤

り率がそれぞれ 2.3%・2.0% 減少した。僅かではあるが、改良を加えることにより、パラメータ自動抽出の精度が向上した。

一方、評価用のデータである ATR503 文に関しては、フレーズ指令の正解率が 94%と非常に高いが、挿入誤り率も 53%と高く、フレーズを必要としない部分にも多くの指令を推定してしまった。アクセント指令は、正解率 68%、欠落誤り率 31%、挿入誤り率 24%と平凡な結果であった。

4.2.3 聴取実験結果

抽出したパラメータから生成した F_0 パターンを用いて、音声資料の再合成音を作成し、12 名の被験者に対して聴取実験を行った。ただし、上述の分析結果において、誤りであると判定されたパラメータが聴覚的にどのような意味を持つのかを検証するために、欠落誤りまたは挿入誤りを含むものを、NHK、ATR の音声資料からそれぞれランダムに 10 文 (NHK: N1∼N10、ATR: A1∼A10) だけ選定し、再合成音の評価点 (5 段階、1: 悪い ∼5: 良い) および韻律的に違和感を感じる箇所を記述させた。評価点の平均値を表 2 に示す。

表 2. ランダムに抽出した 10 文の聴取実験結果

NHK		ATR	
再合成音	評価点の平均	再合成音	評価点の平均
N1	3.3	A1	2.2
N2	2.1	A2	1.9
N3	4.2	A3	2.7
N4	4.4	A4	4.1
N5	3.1	A5	2.0
N6	3.8	A6	3.1
N7	2.1	A7	1.2
N8	3.5	A8	3.8
N9	1.8	A9	4.4
N10	4.2	A10	2.2

評価点の平均が 2 程度の再合成音 (N2、N7、N9、A1、A2、A5、A7、A10) は、フレーズ指令の挿入誤りが存在するものであり、評価点の平均が 4 を超えている再合成音 (N3、N4、N10、A4、A9) は、振幅の平均が 0.18 と小さいアクセント指令が欠落していたり、振幅の平均が 0.16 と小さいアクセント指令が挿入されているものであった。また、韻律的に違和感を感じると記述された箇所において、挿入されたフレーズ指令の振幅 A_p とアクセント指令の振幅 A_a がほぼ同じ再合成音 (N1: $A_a = 0.42$ と A3: $A_p = 0.43$、N2: $A_p = 0.51$ と A6: $A_a = 0.52$、N5: $A_a = 0.38$ と A10: $A_p = 0.38$、N9: $A_p = 0.58$ と A1: $A_a = 0.57$) であっても、フレーズ指令が挿入されたときの方が評価点は低いという結果が得られた。振幅の大きな指令が欠落している再合成音 (N6: $A_a = 0.42$、N8: $A_p = 0.33$、A8: $A_a = 0.39$) に関しては、評価点の平均が 3.7 と評価は高かった。

5. まとめ

以上、著者らが既に発表している F_0 パターン生成過程モデルのパラメータ自動抽出法に関して、その問題点を挙げ、解決策を幾つか提案した。実際に音声資料を用いた分析から、提案した手法の有効性を示す結果が得られた。また、音声資料の再合成音を用いた聴取実験の結果から、聴覚的に、フレーズ指令の挿入誤りは影響が大きく、振幅値の小さなアクセント指令の欠落誤りと挿入誤りは影響が少ないことが示された。指令の欠落誤りと挿入誤りとでは、挿入誤りの方が韻律的に違和感を感じると判断された。これらの観点から、今後の方針として、さらに多くの音声資料に対する分析を行い、指令の挿入誤りの減少に着目した改良を行うと共に、パラメータと言語情報との関係を調べ、韻律データベースの構築を行う予定である。

参考文献

[1] H. Fujisaki and S. Nagashima : "A model for synthesis of pitch contours of connected speech," *Annual Report of Engineering Research Institute, University of Tokyo*, vol.28, pp.53–60 (1969).

[2] H. Fujisaki and K. Hirose : "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Jpn* (*E*), vol.5, no.4, pp.233–242 (1984).

[3] 成澤 修一, 峯松 信明, 広瀬 啓吉, 藤崎 博也 : "音声の基本周波数パターン生成過程モデルのパラメータ自動抽出法," 情報処理学会論文誌, vol. 43, no. 7, pp. 2155–2168 (2002).

[4] K. Hirose, H. Fujisaki and S. Seto : "A scheme for pitch extraction of speech using autocorrelation function with frame length proportional to the time lag," *Proc. ICASSP '92*, vol.1, pp.149–152 (1992).