

# 音声の群遅延特性に基づく声門閉止等のイベント抽出について

河原 英紀<sup>1,2</sup>、 阿竹 義徳<sup>3</sup>

<sup>1</sup>和歌山大学システム工学部、<sup>2</sup>CREST

<sup>3</sup>奈良先端科学技術大学院大学

<sup>1</sup>〒640-8510 和歌山市栄谷930

*kawahara@sys.wakayama-u.ac.jp*

あらまし 音声波形の群遅延特性を利用して、声門閉止等の音声を駆動する主要なイベント生起時刻とイベント属性を定量的に高精度に抽出する新しい方法を発明した。Gauss型時間窓の中心時刻からその窓を用いて計算される平均時刻への写像の不動点として求められるイベントの初期推定値を、振幅スペクトルから計算される最小位相応答の群遅延特性を用いて補償することにより、声門閉止等のイベントの時刻とイベントの原因となった現象の継続時間を求めることができる。提案したアルゴリズムについて、合成音声等を用いた検証を行い、次いで、EGGと音声を同時録音したデータベースを用いて実音声の分析における定量的な評価を行った。合成音声の作成には、全極モデル、STRAIGHTにより求めた最小位相インパルス応答、パルス音源、Rosenberg-Klatt波形の様々な組合せを用いた。EGG同時収録音声のデータベースは、男女各14名がそれぞれ30文を読み上げた840文から構成されている。これらの実験結果は、本方法によれば $40\mu\text{s}$ から $200\mu\text{s}$ の誤差の標準偏差で声門の閉止時刻が推定できることを示している。本方法は、FFTを多用するものの収束計算を含まない実時間向きのアルゴリズムとして実現されているため、声帯振動の異常の診断や高品質音声合成法の音源情報抽出等に広範に応用できるものと考えられる。

キーワード 群遅延、基本周期、声門閉止、最小位相、継続時間、EGG

## Vocal fold closure and speech event detection using group delay

Hideki Kawahara<sup>1,2</sup> and Yoshinori Atake<sup>3</sup>

<sup>1</sup>Faculty of Systems Engineering, Wakayama University, <sup>2</sup>CREST

<sup>3</sup>Nara Institute of Science and Technology

<sup>1</sup>930 Sakaedani, Wakayama, Wakayama, 640-8510 Japan

*kawahara@sys.wakayama-u.ac.jp*

Abstract A new procedure for event detection and characterization was proposed based on group delay and fixed point analysis. The proposed method enables to detect precise timing and spread of speech event like a vocal fold closure. A mapping from the center of a Gaussian time window to the mean time provides event locations as its fixed points. Refining these initial estimates using minimum phase group delay functions derived from the amplitude spectra provides accurate estimates of event locations and durations of excitations of each event. The proposed algorithm was tested using synthetic speech samples and natural speech database of simultaneously recorded sound waveforms and EGG signals. These tests revealed that the proposed method provides estimates of vocal fold closure instants with timing accuracy within  $40\mu\text{s}$  to  $200\mu\text{s}$  standard deviations. The proposed method is implemented as a real-time oriented algorithm based on heavy use of FFTs without introducing any iterative procedures. The proposed method is potentially a very powerful tool for speech diagnosis and constructing very high quality speech manipulation systems.

key words group delay, fundamental period, vocal fold closure, minimum phase, electroglottograph

# 1 はじめに

音声の様々なパラメタを、音声の品質を高く保ったまま自由に変換することができれば、実用的にも、学術的にも広い適用範囲がある。著者らは、音声をスペクトル包絡と音源に分離し再合成することでパラメタ変換の高い柔軟性を有する channel VOCODER[4] の原理に基づき、最近の計算機の性能の急速な向上を生かした高品質な音声分析変換合成方法 STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTEd spectrogram) を提案してきた [5, 6, 7]。STRAIGHT は、条件によっては原音声に遜色ない自然性を有する変換音声を作成することができ、また、5ms 以下の分析フレーム周期を用いることで高い明瞭度を実現できることが示されている [13, 11]。しかし、多数の話者や様々な文章を用いた品質試験では、STRAIGHT により作成した音声は原音声に明らかに劣る品質であることも明らかになってきた。また、複数の調波成分の瞬時周波数を用いる基本周波数推定方の導入も、有声音の基本周波数推定精度の向上という本来の目的には有効であるものの、総合的な品質を必ずしも向上させるものではないことも明らかとなった [12]。問題点の所在は、これまで本格的な検討を行っていなかった音声の非周期的属性の表現にある可能性が高い。

ここでは、平均時刻と継続時間と群遅延との関係を利用して音声における非周期的属性をイベントの時刻とイベントの駆動信号の継続時間によって表現する方法を提案する。まず、最初に、平均時刻と継続時間の時間領域での表現を用いて、イベントの初期推定値が、時間窓の時刻から平均時間への写像の不動点として求められることを示し、不動点における写像の傾斜からイベントの継続時間が求められることを示す。次に、スペクトルの振幅成分と群遅延特性から平均時刻と継続時間との関係を利用してこれらの初期推定値に基づいて精密なイベントの時刻とイベントを駆動した現象の継続時間を推定する方法を導く。さらに、合成音声を用いたシミュレーションと、音声と EGG (electroglottograph) 信号を同時記録したデータベースにより、本方法を評価した結果について紹介する。

# 2 イベントの抽出と特徴付け

音声は、声門の周期的な開閉や子音部分での破裂や乱流雑音等、様々な現象に声道が駆動されて生ずる信号である。母音等の有声音の場合、高い周波数領域では、主に声門閉止における波形の不連続が声道を駆動していると見ることができる。この領域における声道の伝達特性の極に起因するピークの帯域幅が比較的大きな場合には、高域通過あるいは高域強調フィルタの出力波形  $s(t)$  の包絡は、声門閉止にやや遅れた部分で最大値を示した後、急速に減衰する。無声破裂子音では、声道の狭めが急激に解放されることによる体積流のステップ状の変化が音源となり、フィルタの出力波形は同様な包絡を示す。摩擦子音では、連続的な乱流雑音が駆動源となるため、フィルタ出力の包絡はある一定の値の周辺でランダムに変動

する。

このような現象を、「イベント」の集まりと考えることもできる。ここで言うイベントは、エネルギーが時間的に集中しているような部分を指す。声門閉止に対応する音声波形上のイベントは、高い周波数領域においてエネルギーの時間的な集中の度合いが大きいであろうし、摩擦子音の中で生ずるイベントは、エネルギーの時間的な集中の度合いは少ないであろう。有声音の場合には、類似したイベントがほぼ周期的に繰り返されるであろうし、破裂音の場合には、エネルギーの集中したイベントがあっても、単発的あるいは不規則に生ずるのみであろう。ここでは、このようなイベントの性質を記述する量として平均時刻と継続時間を用いることとする。

## 2.1 時間領域での表現

ある時間信号を  $s(t)$  としたとき、その平均時刻  $\langle t \rangle$  と継続時間  $\sigma_t$  は以下のように表される [3]。

$$\langle t \rangle = \frac{\int t |s(t)|^2 dt}{\int |s(t)|^2 dt} \quad (1)$$

$$\sigma_t^2 = \frac{\int (t - \langle t \rangle)^2 |s(t)|^2 dt}{\int |s(t)|^2 dt} \quad (2)$$

音声波形は複数のイベントを含むため、上記の量を意味のあるものとするためには、注目するイベントを時間窓等の操作によって予め分離しておくことが必要となる。

ある時間窓  $w(t)$  によって、ある一つの声門閉止の周囲を切出せば、次式によって、イベントの平均時刻  $\langle t(u) \rangle$  と継続時間  $\sigma_t(u)$  を求めることができる。

$$\langle t(u) \rangle = \frac{\int t |x(t, u)|^2 dt}{\int |x(t, u)|^2 dt} \quad (3)$$

$$\sigma_t^2(u) = \frac{\int (t - \langle t(u) \rangle)^2 |x(t, u)|^2 dt}{\int |x(t, u)|^2 dt} \quad (4)$$

$$x(t, u) = w(t - u)s(t) \quad (5)$$

ここで  $u$  は、時間窓の中心がある時刻を表し、積分の範囲は  $(-\infty, \infty)$  である。なお、 $\sigma_t(u)$  は窓が掛けられた信号の見かけの継続時間である。

### 2.1.1 窓の中心と平均時刻

ここで、以下の議論を簡単にするために次のようなガウス型の窓関数を用いる。

$$w(t) = e^{-\frac{t^2}{2\sigma_w^2}} \quad (6)$$

また、イベントの振幅包絡も次のようなガウス型を仮定する。

$$|s(t)| = e^{-\frac{(t-t_e)^2}{2\sigma_s^2}} \quad (7)$$

$t_e$  はイベントの時刻を表す。

すると、平均時刻は次のように表される。

$$\begin{aligned} \langle t(u) \rangle &= \frac{1}{p_w(u)} \int t |w(t-u)s(t)|^2 dt \\ &= \frac{1}{p_w(u)} \int t e^{-\left(\frac{(t-u)^2}{\sigma_w^2} + \frac{(t-t_e)^2}{\sigma_s^2}\right)} dt \quad (8) \end{aligned}$$

ここで  $p_w(u) = \int |x(t, u)|^2 dt$  は、切り出された波形のエネルギーを表す。式8より  $|w(t-u)s(t)|$  は中心に対して対称となるので、平均時刻は指数部の導関数が0となる位置となり、次のように求められる。  $L(t) = -\log |w(t-u)s(t)|^2$  とする。

$$\begin{aligned} \frac{dL(t)}{dt} &= -\frac{d}{dt} \left( \frac{(t-u)^2}{\sigma_w^2} + \frac{(t-t_e)^2}{\sigma_s^2} \right) \\ &= -2 \left( \frac{t-u}{\sigma_w^2} + \frac{t-t_e}{\sigma_s^2} \right) = 0 \\ \langle t(u) \rangle &= \frac{\sigma_s^2 u + \sigma_w^2 t_e}{\sigma_s^2 + \sigma_w^2} \end{aligned} \quad (9)$$

このように、平均時刻は窓の中心の時刻とイベントの時刻の加重平均となる。イベントの継続時間が短ければ短いほどイベントの時刻の重みが増す。また、平均時刻がイベントの時刻と一致するのは、窓の中心がイベント位置に重なった時であることが分かる。したがって、イベントの時刻は、窓の中心の時刻から平均時刻への写像の不動点の中で以下の条件を満たすものから求められる。

$$\{t_e\} = \{u | \langle t(u) \rangle = u, \frac{d\langle t(u) \rangle}{du} < 1\} \quad (10)$$

### 2.1.2 写像の傾斜と継続時間

ところで、式9より、上記の条件を満たす不動点における写像の傾斜  $g(t_e)$  は次式のように求められる。

$$g(t_e) = \left. \frac{d\langle t(u) \rangle}{du} \right|_{u=t_e} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_w^2} \quad (11)$$

この状態での波形の包絡は、不動点における写像の傾斜  $g(t_e)$  を用いて次のように表されることに注意する。

$$\begin{aligned} |w(t-t_e)s(t-t_e)| &= e^{-\frac{\sigma_s^2 + \sigma_w^2}{2\sigma_s^2 \sigma_w^2} (t-t_e)^2} \\ &= e^{-\frac{1}{2g(t_e)\sigma_w^2} (t-t_e)^2} \end{aligned} \quad (12)$$

この式12を用いて継続時間を求めると、以下が得られる。

$$\sigma_t(t_e) = \sigma_w \sqrt{\frac{g(t_e)}{2}} \quad (13)$$

すなわち、不動点における写像の傾斜と分析に用いた時間窓の標準偏差  $\sigma_w$  を用いることにより、窓で切り出された信号の見かけの継続時間を表すことができることが分かる。また、次式によってイベントのパラメタである  $\sigma_s(t_e)$  を求めることができる。

$$\sigma_s(t_e) = \sigma_w \sqrt{\frac{g(t_e)}{1-g(t_e)}} \quad (14)$$

## 2.2 周波数領域での表現

ここで、平均時刻と継続時間の周波数領域での表現 [3] を利用し、窓を掛けた信号の平均時刻  $\langle t(u) \rangle$  と継続時間  $\sigma_t(u)$  を群遅延  $t_g(\omega, u) = -\psi'(\omega, u)$  を用いて表現する。

ここで  $'$  は  $\omega$  に関する微分を表す。

$$\langle t(u) \rangle = -\int \psi'(\omega, u) |S(\omega, u)|^2 d\omega \quad (15)$$

$$\begin{aligned} \sigma_t^2(u) &= \int \left( \frac{B'(\omega, u)}{B(\omega, u)} \right)^2 B^2(\omega, u) d\omega \\ &+ \int (\psi'(\omega, u) + \langle t(u) \rangle)^2 B^2(\omega, u) d\omega \end{aligned} \quad (16)$$

$$\begin{aligned} S(\omega, u) &= \frac{1}{\sqrt{2\pi}} \int x(t, u) e^{-j\omega t} dt \\ &= |S(\omega, u)| e^{j\psi(\omega, u)} = B(\omega, u) e^{j\psi(\omega, u)} \end{aligned} \quad (17)$$

式16の第一項は、スペクトルの振幅変動による継続時間への寄与分、第二項は、位相変動による寄与分を表す。  $B(\omega, u)$  は、スペクトルの振幅成分を表す。

### 2.2.1 最小位相応答の補償

式15は、平均時刻が群遅延の加重平均であることを示している。すなわち、声門閉止に対応する不動点は、声道のインパルス応答の群遅延分だけ実際の声門の閉止時刻から遅れた位置に生ずることが分かる。声道のインパルス応答が因果律を満たしているのであれば、振幅スペクトルから例えば次のように複素ケプストラム  $C(q, u)$  を介して最小位相インパルス応答 [9] に対応する群遅延  $\tau_m(\omega, u)$  を計算することができる。

$$\tau_m(\omega, u) = -\frac{d}{d\omega} \left( \text{imag} \left[ \frac{1}{\sqrt{2\pi}} \int C(q, u) e^{j\omega q} dq \right] \right) \quad (18)$$

$$\begin{aligned} C(q, u) &= \begin{cases} 2c(q, u) & q > 0 \\ c(q, u) & q = 0 \\ 0 & \text{otherwise} \end{cases} \\ c(q, u) &= \frac{1}{\sqrt{2\pi}} \int \log B(\omega, u) e^{-j\omega q} d\omega \end{aligned} \quad (19)$$

ここで  $q$  はケフレンシーである。この最小位相成分を用いて群遅延  $-\psi'(\omega, u)$  を補償すれば、声道の影響を受ける前の駆動信号の群遅延特性を求めることができる。

声道による群遅延を補償した平均時刻  $\langle \tilde{t}(u) \rangle$  と位相変動による寄与分  $\tilde{\sigma}_P^2(u)$  は次のように求められる。

$$\langle \tilde{t}(u) \rangle = -\int (\psi'(\omega, u) + \tau_m(\omega, u)) |S(\omega, u)|^2 d\omega \quad (20)$$

$$\tilde{\sigma}_P^2(u) = \int (\psi'(\omega, u) + \langle \tilde{t}(u) \rangle + \tau_m(\omega, u))^2 B^2(\omega, u) d\omega \quad (21)$$

声門閉止による駆動が高域ではインパルスで近似できるのであれば、声道の応答による群遅延を補償した位相変動による寄与分は、ほぼ0となる。なお、振幅成分を補償して平坦なスペクトルとすると第一項は0となる。これは逆フィルタ処理に外ならない。<sup>1</sup>

<sup>1</sup>この処理は、例えばLPCによる逆フィルタを用いて残差波形を推定し、その残差波形についてイベント位置やイベントの継続時間を求めることと類似している。しかし、LPC等による逆フィルタ処理は、スペクトルの平坦化のためにSNの悪いレベルの低い部分を増幅し、SNの良いレベルの高い部分を減衰させるため、イベントの推定値の誤差を増幅する。ここで提案する処理は、レベルの高い部分の情報を主に利用するため、そのような問題による劣化を避けることができる。

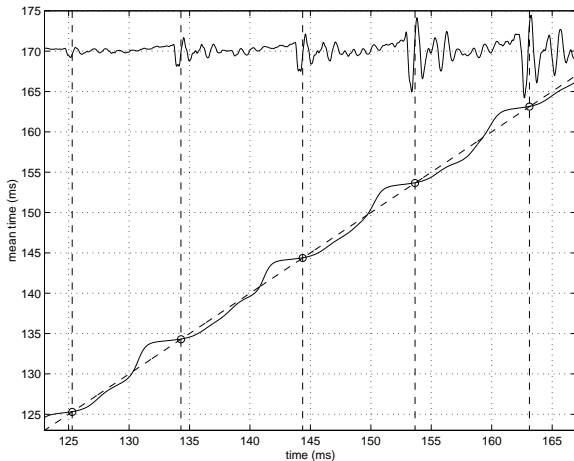


図 1: Time domain event extraction. The original speech waveform is plotted at the top of the figure. The diagonal solid line represents the mapping from the window center location to the mean time. Small circles represent the extracted fixed points.

### 3 駆動情報の抽出手順

以上をまとめると、次の手続きによって駆動点の情報を抽出することができる。

ステップ1 式10を用いて不動点として駆動点の候補を抽出する。同時に、式13により継続時間を求める。

ステップ2 それぞれの駆動点において、式18を用いて最小位相インパルス応答に対応する群遅延を求める。

ステップ3 ステップ1で求めた駆動点のそれぞれの候補における平均時刻とステップ2で求めた最小位相インパルス応答に対応する群遅延から、式20を用いて駆動点の位置を求め、式21を用いて駆動源の継続時間を求める。

次の節では、実際の音声を例として、各ステップの具体的な動作について説明する。

#### 3.1 実音声の分析例

男性の発声した日本語母音の連鎖「アイウエオ」を例にとって、分析の各ステップを説明する。音声の収録には圧力型マイク (Sony EMC-77S) を用い、22050 Hz 16 bit で標本化した。

時間領域での不動点の抽出 図1に、時間窓の中心位置から平均時刻への写像を示す。図中の印は抽出された不動点を示す。図の最上部に示した音声波形と比較すると、不動点は声門閉止の位置から少し遅れたところにあることが分かる。本方法で用いている写像は窓内のエネルギーで正規化された無次元の量であり、図中の 125 ms 附近のようにレベルが低い場合も、163 ms 附近のようにレベルが高い場合も、同様に安定に求めることができる。

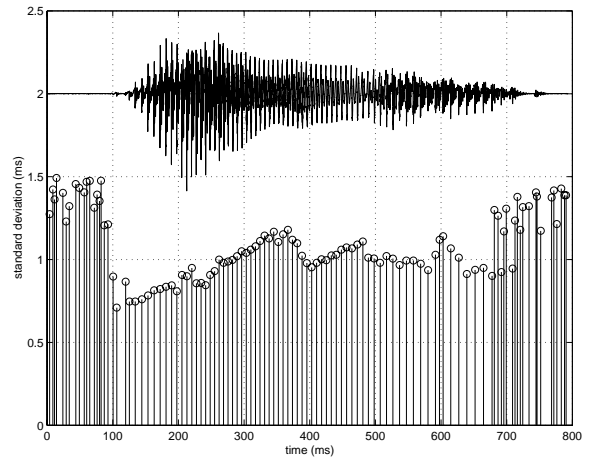


図 2: Estimated durations of events and the original waveform using the time domain procedure.

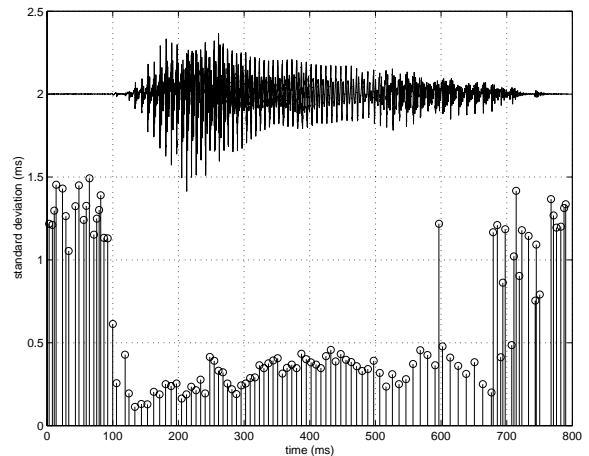


図 3: Estimated durations of driving signals of events and the original waveform using the spectral domain compensation.

通常のイベント検出に必要な閾値の設定は、本方法では不要である。

波形から求めたイベントの継続時間 図2に、それぞれ不動点について求めた継続時間を音声波形とともに示す。無声部分ではイベントの継続時間はほぼ窓長に一致し、有声部分では短くなっていることが分かる。

最小位相成分の補償による駆動情報の抽出 図3に、スペクトルの振幅情報を用いて補正したイベント時刻と、イベントの駆動信号の継続時間を示す。ここでも無声部分のイベントの継続時間は窓長のあたりに存在することが分かる。有声部分については、継続時間が明らかに小さな値を示しており、駆動源が非常に短い時間に集中していることが分かる。

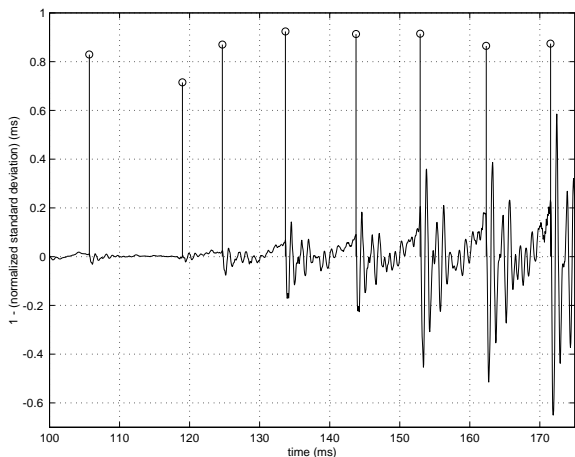


図 4: Updated event locations and the original waveform at the beginning of the utterance. Event energy concentration is represented by the vertical stems.

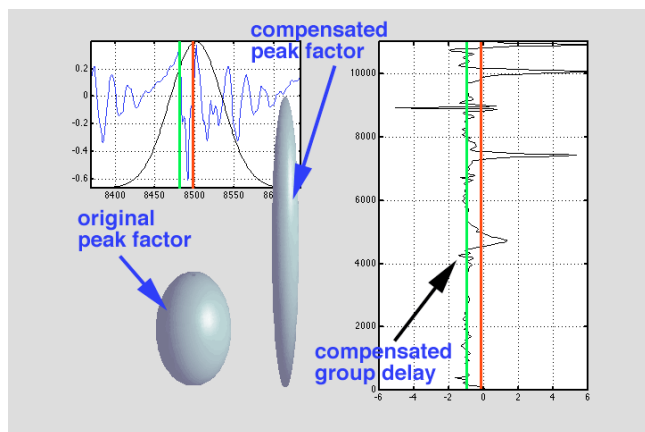


図 5: Snapshot of a demonstration video to illustrate operating principles of the proposed event extraction method. Top left panel shows speech waveform and the time window. Dark solid thick vertical lines represent the initial mean time estimates. Light solid thick vertical lines represent the compensated mean time estimates.

修正されたイベント位置と駆動情報 図 4 に、修正されたイベント位置とイベントの鋭さを示す。イベントの鋭さ  $\eta(u)$  を表す指標として以下の式で表されるものを用いた。

$$\eta(u) = \frac{\sigma_w - \tilde{\sigma}_P}{\sigma_w} \quad (22)$$

この指標が 1 となるのは、駆動源がインパルスの場合で最小位相の補償が完全に行われた場合である。また、駆動源が定常的なランダム雑音の場合には、この指標は 0.3 よりもやや少ない値の周辺に分布する。

アニメーションによる紹介 図 5 に、提案した方法の動作を紹介するビデオの一画面を示す。ビデオ機器の利用が

可能であれば、研究会場でも紹介する予定である。

図 5 の左上の部分には、分析対称の音声波形と、分析に用いた Gauss 窓を示す。図中の黒い太い実線は、式 3 で求められるイベントの平均時刻  $\langle t(u) \rangle$  を表す。図中の灰色の太い実線は、式 20 で求められる最小位相成分の影響を補償された平均時刻  $\langle \tilde{t}(u) \rangle$  を表す。音声は、男性の発声した持続母音「ア」である。

図 5 の右側には、補償された群遅延特性を示す。ここで縦軸は周波数であり、横軸は遅延時間である。図中の太線は前の段落と同様に、式 3 で求められるイベントの平均時刻  $\langle t(u) \rangle$  と最小位相成分の影響を補償された平均時刻  $\langle \tilde{t}(u) \rangle$  を表す。

左下の二つの楕円体は、窓の継続時間をイベントの継続時間で割った値を長径とし、窓の継続時間を短径として画面に納まるようにスケーリングしたものである。この楕円体は、縦長であるほどエネルギーが短い区間に集中していることを表す。左の楕円体は、波形からそのまま求めた値を利用したもの、右の楕円体は、最小位相成分を補償した群遅延特性から求めた値を利用したものである。

この画面は、分析窓の中央付近に声門閉止による応答の大部分が単独で選択されている場合をとらえたものである。このように主要なイベントが単離された場合には、最小位相応答による影響を補正された群遅延特性は、4 kHz 付近まで非常に小さな変動に納まる。逆の見方をすると、この周波数領域における一見すると複雑な群遅延特性 [1] は、振幅スペクトルから決まる最小位相成分によるものであることが分かる。

## 4 合成音声等による検証と評価

前節で紹介したように、本方法の実現可能性は示された。ここでは次に、試験用の単純な信号や合成音声を用いて本方法による分析の精度を調べた。

試験信号による検証  $\sigma_s = 1\text{ms}$  のガウス型信号の周期系列を分析した例を図 6 に示す。標準化周波数は 22050 Hz であり、図の縦棒の位置が平均時刻を表し上の印が継続時間から推定した  $\sigma_s$  を表す。分析の時間窓は  $\sigma_w = 1.5\text{ms}$  とした。この場合、 $\sigma_s$  の推定値は有効数字 4 桁で真の値と一致した。

### 4.1 合成音声による検証と評価

ここでは、実音声から求めたパラメタを用いて様々な合成音声を作成し、提案した方法の検証と評価を行った。ここでも、標準化周波数は 22050 Hz とした。

合成用音源 合成音声の音源には、周期的なパルス列から平均値を除去して積分したものと、Rosenberg-Klatt [8] の波形 (以下 RK 波形と略) を微分したものの 2 種類を用いた。RK 波形のパラメタとしては、OQ を 0.66 とし、基本周波数を 110.25 Hz とした。

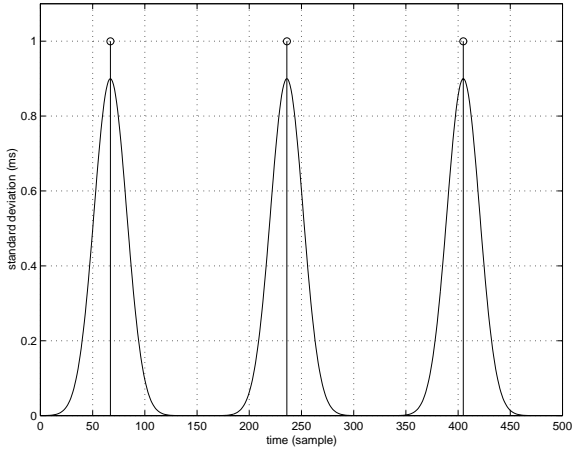


図 6: Extracted excitation position and estimated standard deviation. The test signal is a synthetic signal with 1 ms for the standard deviation.

ID	conditions	
	envelope	source
SYN1	STRAIGHT	Impulse
SYN2	STRAIGHT	Rosenberg-Klatt
SYN3	LPC	Impulse
SYN4	LPC	Rosenberg-Klatt

表 1: Synthetic speech samples and their conditions

合成音声の作成用パラメタ STRAIGHT による実音声 /a/ (男声) の分析から求めた振幅スペクトル包絡ならびにその振幅スペクトル包絡の自乗から自己相関関数を求めて推定した全極型モデル (LPC) のパラメタを用いた。極の個数は 30 とした。振幅スペクトル包絡からは最小位相インパルス応答を求め、合成用のパラメタとした。なお、前処理としてスペクトル包絡の +6dB/oct のプリエンファシスを行った。音源とこれらのパラメタを組み合わせ、表 1 に示す 4 種類の合成音声を用意した。

合成音声の分析結果 こうして作成した合成音声を、提案の方法により分析して幾つかの SN について、駆動点の位置と位置推定誤差の標準偏差を求めた。図 7 に、抽出された駆動点の例を波形と重ねて示す。分析の前処理として、波形の差分を行った。また、分析のための時間窓は、 $\sigma = 1.0\text{ms}$  のガウス窓とした。

表 2 に様々な条件で合成した母音「ア」について、声門閉止時点の推定精度と SN 比の関係を示す。雑音は正規白色雑音を用い、それぞれの条件で 500 個の声門閉止時点を推定した。20 dB 以上の SN であれば、標準偏差は 1 サンプル ( $45\mu\text{s}$ ) の半分以下であることが分かる。ただし、1 サンプル程度の系統誤差は存在する。

この様子を詳しく見るため、図 8 に SN を 6 dB から 70 dB まで 2 dB ステップで変えながら求めた声門閉止時点の誤差と推定された継続時間の散布図を示す。継続時間

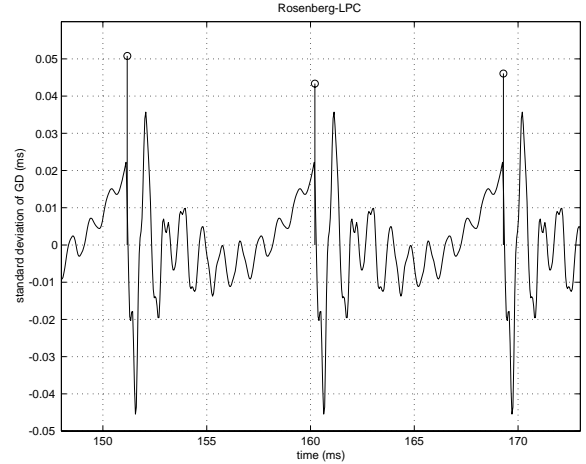


図 7: Synthesized waveform and extracted events.

measure	location ( $\mu\text{s}$ )			SD ( $\mu\text{s}$ )		
	SN (dB)	60	40	20	60	40
SYN1	33.2	37.7	91.6	0.8	4.1	21.4
SYN2	9.0	13.2	66.4	1.1	4.0	21.7
SYN3	4.2	5.3	61.3	0.5	3.3	21.4
SYN4	-15.5	-14.6	39.0	0.9	3.4	19.5

表 2: Event accuracy for synthetic vowel /a/.

は、ここでは時間窓の  $\sigma_w$  を継続時間で割った値 (ピーク率) として表示している。この図から、ピーク率が 5 以上であれば、声門閉止時点の推定値の誤差は 1 サンプルの時間長以下となることが分かる。このように、本方法はスペクトル上でのエネルギーの大きな部分の情報に基づいているため、比較的雑音に強い。

## 5 EGG データベースの分析

試料 EGG と音声を同時収録したデータベースを用い、提案した方法の実音声分析での精度を調べた。データベースには、男女それぞれ 14 名が各 30 文章を発声した計 840 文の発話のデータが収録されている。データベースの詳細を付録に示す。今回の分析では、標準化周波数を 16kHz にダウンサンプリングしたものを試料として用いた。

EGG による時刻との比較 本方法によって求めた EGG 波形の駆動点の時刻を基準として、音声波形から求められた駆動点の時刻の分布を調べた。図 9 に男性話者 (M04) の分析例を示す。 $\sigma_w = 1\text{ms}$  を用いた。上の図は、駆動点の時刻とその継続時間の散布図である。縦軸は時間窓の継続時間を駆動点の継続時間で割った値 (ピーク率) である。下の図は、駆動点の時刻のヒストグラムである。集中度が 5 以上の駆動点の時刻の標準偏差は  $60\mu\text{s}$  であった。

図 10 に女性話者 (F06) の分析例を示す。 $\sigma_w = 0.6\text{ms}$  を用いた。女性の場合にはピーク率は男性よりも一般に小

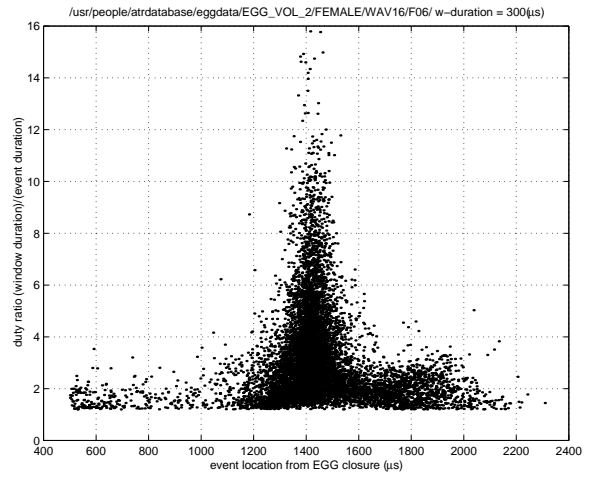
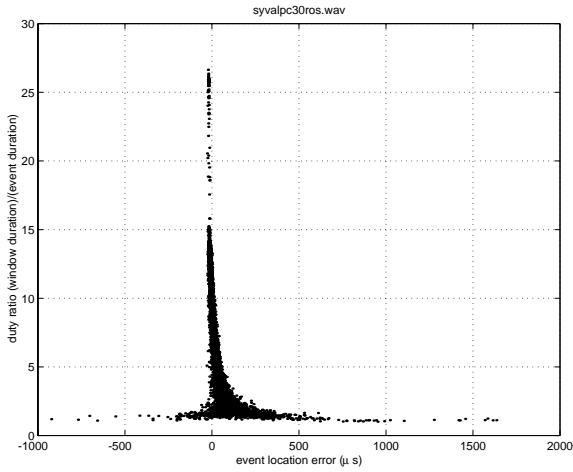


図 8: Distribution of event location and peak factor for SYN4.

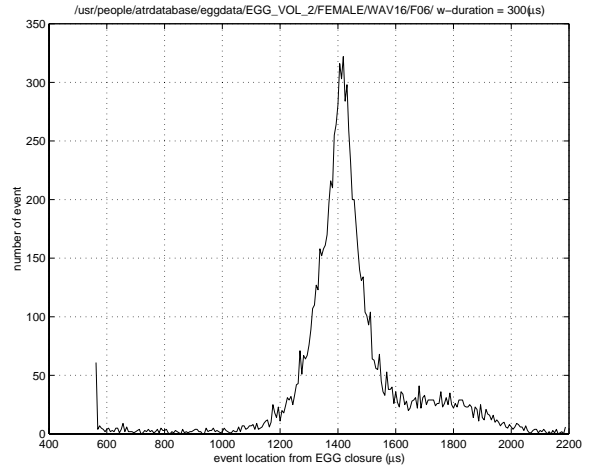
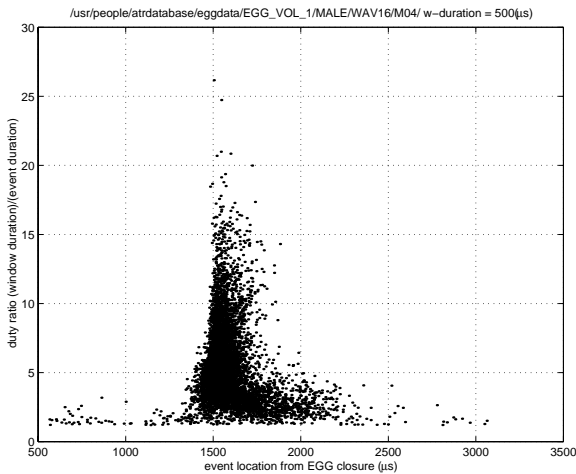


図 10: Distribution of event location and peak factor for a female speaker (F06).

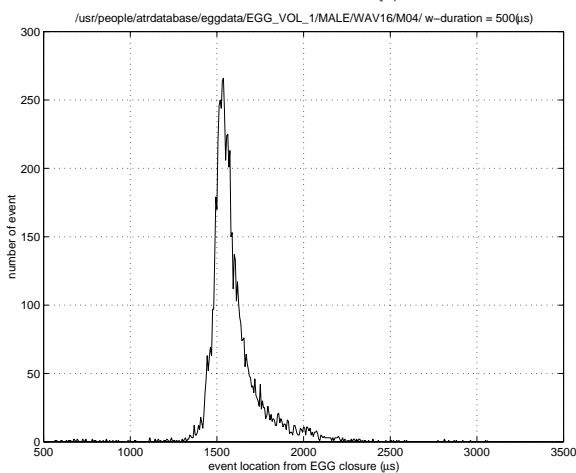


図 9: Distribution of event location and peak factor for a male speaker (M04).

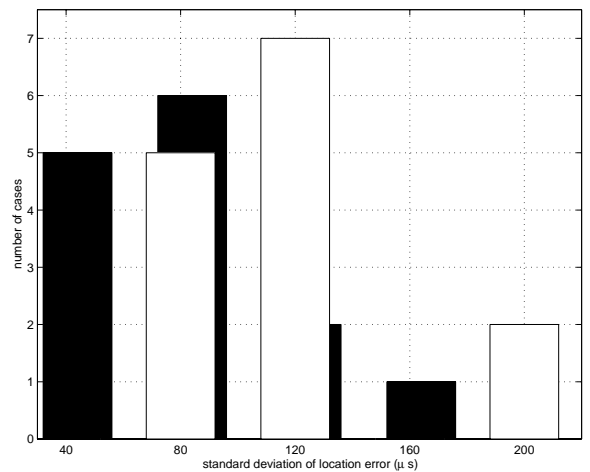


図 11: Histogram of standard deviation of location estimation errors. Filled bars represent female results. Open bars represent male results.

さい。14名の男性の標準偏差は、同様な条件の下で60から200 $\mu$ s、女性の標準偏差は40から180 $\mu$ sの範囲に分布していた。この様子を図11に示す。

## 6 討論

本方法で求めたEGG上のイベントと音声波形上のイベントの時間差は1.5msの周辺に分布している。音声波

形にはマイク位置に基づく補正を行ってあるので、この時間差は唇位置における時間差であることになる。平均的な声道長を17cmとして声道長の長さの影響を補正すると、声門位置での時間差は1msとなる。図式的な検討[2, 10]を参考にすれば、本方法がEGG波形から抽出するのは、声門の一端での接触の開始時刻であり、実際に体積流が切断される時刻とは異なると考えられる。しかし、接触の開始から閉止までに1msもの時間が必要であるとは考えにくい。この食い違いの解明は、今後の検討課題である。

遅延量の絶対値に上記のような問題があるにせよ、補償後の継続時間が200 $\mu$ s以下と推定されるようなイベントについては、本方法により高い再現性のある測定が可能であることが分かる。

## 7 まとめ

群遅延を用いた音声の駆動情報の抽出法を提案し、声門閉止等の音声を駆動するイベントとの関係を調べた。提案した方法は、音声波形から声門閉止のタイミングを高い精度で求めるだけでなく、イベントを駆動する刺激の継続時間についての定量的な情報を同時に抽出することができる。シミュレーションならびに実音声を用いた実験結果は、本方法が声門閉止のタイミングを精密に(40 $\mu$ sから200 $\mu$ sの誤差の標準偏差で)推定できることを示した。今後は、本方法を高品質音声分析変換合成システム[7]に組み込むと共に、音声・聴覚研究への応用を検討して行きたい。本方法は、これらの他にも、音声生成過程の研究・診断等、様々な方面への応用が可能であろうと考えられる。

## 謝辞

EGGデータベースの作成では、器機の使用と測定法についてATR人間情報通信研究所 正木信夫 博士、データベースの内容の決定や収録において奈良先端大の 鹿野清宏 教授に支援頂いたことに深謝します。また、技術的な内容に関して日ごろ議論頂くCRESTの聴覚脳グループのメンバーに感謝します。最後に、本検討は助手の 山本英里 博士との群遅延に関する議論から始まったことを特記して感謝します。本研究は、CRESTの「聴覚脳プロジェクト」の一環であり、また、一部に科学研究費(基盤C:11650425)の支援を受けた。

## 参考文献

- [1] Hideki Banno, Jinlin Lu, Satoshi Nakamura, Kiyohiro Shikano, and Hideki Kawahara. Efficient representation of short-time phase based on group delay. In *Proc. ICASSP'98*, pp. 861–864, Seattle, 1998.
- [2] D. G. Childers, D. M. Hicks, G. P. Moore, and Y. A. Alsaka. A model for vocal fold vibratory motion, contact area, and the electroglottogram. *J. Acoust. Soc. Am.*, Vol. 80, No. 5, pp. 1309–1320, 1986.
- [3] L. Cohen. *Time-frequency analysis*. Prentice Hall, Englewood Cliffs, NJ, 1995.
- [4] H. Dudley. Remaking speech. *J. Acoust. Soc. Am.*, Vol. 11, No. 2, pp. 169–177, 1939.

- [5] 河原英紀, 増田郁代. 時間周波数領域での補間を用いた音声の変換について. 信学技報, Vol. EA96-28, , August 1996.8.
- [6] Hideki Kawahara. Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited. In *Proceedings of IEEE int. Conf. Acoust., Speech and Signal Processing*, Vol. 2, pp. 1303–1306, Muenich, 1997.
- [7] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction. *Speech Communication*, Vol. 27, No. 3-4, pp. 187–207, 1999.
- [8] D. Klatt and L. Klatt. Analysis synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.*, Vol. 87, pp. 820–857, 1990.
- [9] A. Oppenheim and R. Schaffer. *Discrete-Time Signal Processing*. Prentice Hall, Englewood Cliffs, NJ, 1989.
- [10] I. Titze. *Principles of voice production*. Prentice Hall, 1994.
- [11] Parham Zolfaghari, 河原英紀. Subjective evaluation of STRAIGHT. 音響学会秋季講演論文集, pp. 193–194, 1999.
- [12] 阿竹義徳, 陸金林, 中村哲, 鹿野清宏, 河原英紀. STRAIGHTの分析合成方式パラメタの主観評価による検討. 音響学会春季講演論文集, pp. 205–206, 2000.
- [13] 河原英紀, 山田玲子, 久保理恵子. STRAIGHTを用いた音声パラメタの操作による印象の変化について. 聴覚研究会資料, Vol. H-97-63, , 1997.9.

## A EGG同時記録データベース

本データベースは、著者の一人(阿竹)が修士論文用の作成過程において資料として収集したものである。

話者: 平均24.5才の男女各14名

発話内容: 30個の日本語文章。

収録に使用した機材:

マイク: SONY ECM-23F3

DATデッキ: SONY DTC-2000ES

EGGアンプ: LARYNGOGRAPH BS5724

オシロスコープ: KENWOOD CS-8010

収録期間: 1999年11月25日 - 12月9日

収録場所: 奈良先端科学技術大学院大学 情報科学棟

B117 音響実験室

室温: 約24

実験状況: マイクは、椅子に着席した発声者の口の前約20cmの位置に、発声者に正対して設置した。EGGの一組の電極は、発声者の喉を挟むようにバンドを用いて装着した。発声時のEGG波形をオシロスコープを用いて監視しながら、音声とEGGを同時にDATに収録した。

記録条件: 標本化周波数: 48kHz, 16bit, stereoでDATに収録。右チャンネルにEGG波形、左チャンネルにマイクからの音声を収録。デジタルオーディオインタフェースであるDAT-Link+を用いてワークステーションに転送。