ACCURATE VOCAL EVENT DETECTION METHOD BASED ON A FIXED-POINT ANALYSIS OF MAPPING FROM TIME TO WEIGHTED AVERAGE GROUP DELAY

Hideki Kawahara[1], *Yoshinori Atake*[2] *and Parham Zolfaghari*[3]

[1]Wakayama University/ATR/CREST, Wakayama, 640-8510 Japan
[2]NAIST, Ikoma, Nara, Japan, [3]CIAIR/Nagoya University, Nagoya, Aichi, Japan

ABSTRACT

A new procedure for event detection and characterization is proposed based on group delay and fixed point analysis. This method enables the detection of precise timing and spread of speech events such as 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 60 μs to 210 μs standard deviations. This algorithm is implemented to be suitable for real-time operation by making extensive use of FFTs without introducing any iterative procedures. It is potentially a very powerful tool for speech diagnosis and construction of very high quality speech manipulation systems.

1. EVENT: DEFINITION AND ATTRIBUTES

Speech is a signal excited by periodic air flow modulations due to vocal fold vibration, explosions of vocal tract constriction, random noise due to turbulence, and various other stimulations. In a high frequency region, the main driving source for vowel sounds is a discontinuity of the air flow at a vocal fold closure. Vocal tract response to this discontinuity shows its maximum magnitude somewhat later to the driving instance. For example, the envelope of a high-pass filtered signal $s(t)$ has its maximum shortly after the vocal fold closure which then roughly decays exponentially. A voice less stop also has a similar envelope as the driving waveform is a step function. Fricatives have a randomly varying envelope since the driving signal is approximately a continuous stable noise.

One way to represent these aspects of speech is to interpret speech as a collection of events. Here, an event is defined as an energy concentration in time. For example, an event corresponding to vocal fold closure may display a highly concentrated energy distribution, especially in a high frequency region. In a fricative sound an event may display low energy concentration. Events may occur periodically in vowels, however, in stops events may not show such periodic behavior. In this paper these aspects of an event are represented by two attributes; mean time (event location) and duration.

2. TEMPORAL VIEW

As a speech signal consists of multiple events, it is necessary to isolate each event by using a time windowing function $w(t)$. The mean time of an isolated event $\langle t(u)\rangle$ and the windowed duration $\sigma_t(u)$ are defined using the following equation [2]

$$\langle t(u)\rangle = \int t|x(t,u)|^2 dt \qquad (1)$$

$$\sigma_t^2(u) = \int (t-\langle t\rangle)^2 |x(t,u)|^2 dt \qquad (2)$$

$$x(t,u) = w(t-u)s(t)$$

where u represents the time of the window center location. Note that the windowed signal $x(t,u)$ is normalized. Integrals that do not have bounds represent integrations in $(-\infty,\infty)$ throughout this report.

2.1. Window Center to Mean Time Mapping

Assume that the window $w(t)$ has a Gaussian shape as

$$w(t) = e^{-\frac{t^2}{2\sigma_w^2}}. \qquad (3)$$

Also assume that an event has a Gaussian shape envelope given by

$$|s(t)| = e^{-\frac{(t-t_e)^2}{2\sigma_s^2}}, \qquad (4)$$

where t_e represents the location of the event.

Then, the windowed mean time is represented as

$$\langle t(u)\rangle = \frac{\sigma_s^2 u + \sigma_w^2 t_e}{\sigma_s^2+\sigma_w^2}. \qquad (5)$$

This equation indicates that the windowed mean time is a weighted average of the window center and the event location. When the event duration is relatively shorter than the window length, the event location is weighted more. It is also indicated that the windowed mean time is equal to the location of the event when the window center location equals the location of the event. Therefore, a set of event locations $\{t_e\}$ is defined as a set of fixed points of the mapping that satisfies the following condition

$$\{t_e\} = \{u|\langle t(u)\rangle = u, \frac{d\langle t(u)\rangle}{du} < 1\}. \qquad (6)$$

2.2. Duration and mapping gradient

Based on equation 5, the temporal gradient $g(t_e)$ of the mapping at a fixed point yields the duration of the event. The standard deviation parameter $\sigma_s(t_e)$ of this event can be represented using the following equation

$$\sigma_s(t_e) = \sigma_w \sqrt{\frac{g(t_e)}{1-g(t_e)}}. \qquad (7)$$

Figure 1 represents the mapping from the time window center to the mean time using the onset of a vowel segment spoken by a Japanese male speaker. The circle symbols represent extracted fixed points. Comparison with the speech signal at the top of the figure illustrates that extracted fixed points are found to be located shortly after the vocal fold closure instants. This discrepancy is not desirable and should be compensated using a spectral domain method described in the following section.

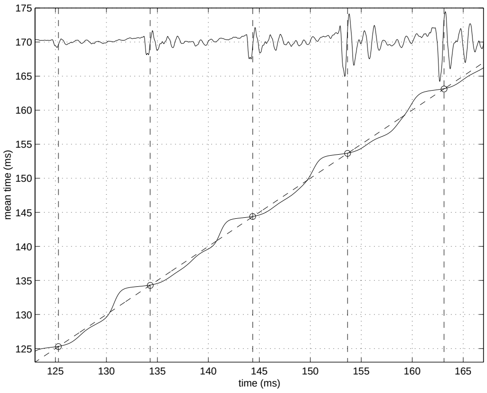

Figure 1. Time domain event extraction. The original speech waveform is plotted at the top of the figure. The diagonal solid line represents the mapping from the window center location to the mean time. Small circles represent the extracted fixed points.

3. SPECTRAL VIEW

The spectral domain representations [2] of the mean time $\langle t(u)\rangle$ and the duration $\sigma_t(u)$ yield the following equation using group delay $\tau_g(\omega,u) = -\psi'(\omega,u)$, where $'$ represents the derivative in terms of the angular frequency ω

$$\langle t(u)\rangle = -\int \psi'(\omega,u)|S(\omega,u)|^2 d\omega \qquad (8)$$

$$\sigma_t^2(u) = \int \left(\frac{B'(\omega,u)}{B(\omega,u)}\right)^2 B^2(\omega,u)d\omega + \int (\psi'(\omega,u)+\langle t(u)\rangle)^2 B^2(\omega,u)d\omega \qquad (9)$$

$$S(\omega,u) = \frac{1}{\sqrt{2\pi}}\int x(t,u)e^{-j\omega t}dt = |S(\omega,u)|e^{j\psi(\omega,u)} = B(\omega,u)e^{j\psi(\omega,u)}. \qquad (10)$$

Here $B(\omega,u)$ represents the amplitude spectrum. The first term of Equation 9 represents the contribution of the spectral deviations and the second term represents the contribution of the phase deviations.

3.1. Compensation of the minimum phase component

Equation 8 indicates that the mean time is a weighted average in the frequency domain where the weight is the power spectrum of the windowed signal. This suggests that, if the system is a passive system, the mean time is located somewhat later to the actual excitation due to the group delay. This is the case for speech signals, where a vocal tract is the passive system. If the system obeys the causality, then the phase spectrum can be calculated from the amplitude spectrum. The group delay $\tau_\phi(\omega,u)$ due to the minimum phase component [5] is calculated through the complex cepstrum $C(q,u)$.

$$\tau_\phi(\omega,u) = -\frac{d}{d\omega}\left(\text{imag}\left[\frac{1}{\sqrt{2\pi}}\int C(q,u)e^{j\omega q}dq\right]\right) \qquad (11)$$

$$C(q,u) = \begin{cases} 2c(q,u) & q>0 \\ c(q,u) & q=0 \\ 0 & \text{otherwise} \end{cases}$$

$$c(q,u) = \frac{1}{\sqrt{2\pi}}\int \log B(\omega,u)e^{-j\omega q}d\omega \qquad (12)$$

where q represents the quefrency. The group delay of the original driving signal that drove the vocal tract can be estimated by compensating the signal group delay $-\psi'(\omega,u)$ using the minimum phase group delay $\tau_\phi(\omega,u)$.

Finally, the compensated mean time $\langle\tilde{t}(u)\rangle$ and the compensated phase component's contribution $\tilde{\sigma}_P^2(u)$ are calculated using the following equation.

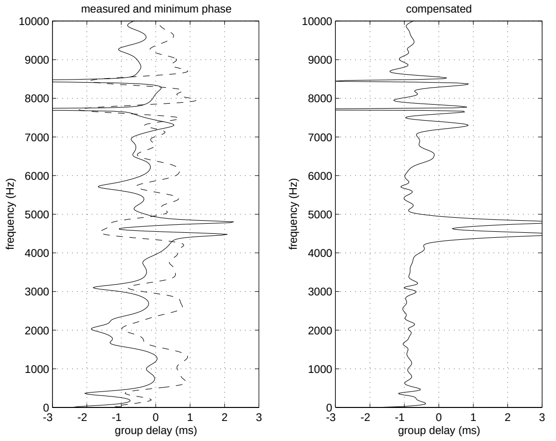

Figure 2. Measured group delay and compensated group delay. Left plot shows the measured group delay (solid line) and the minimum phase group delay (dashed line). Right plot shows the compensated group delay.

$$\langle\tilde{t}(u)\rangle = -\int (\psi'(\omega,u)+\tau_\phi(\omega,u))|S(\omega,u)|^2 d\omega \qquad (13)$$

$$\tilde{\sigma}_P^2(u) = \int (\psi'(\omega,u)+\langle t(u)\rangle+\tau_\phi(\omega,u))^2 B^2(\omega,u)d\omega \qquad (14)$$

Note that the compensated phase component's contribution approaches to zero if the excitation due to the vocal fold closure can be approximated (with an appropriate equalization) by an impulse train. It also should be noted that Equations 13 and 14 can be extended to include an additional frequency weighting. Such weighting is useful for eliminating errors due to observation noise and the estimation bias due to the glottal source waveform.

Figure 2 shows how this minimum phase compensatin works. The group delay was caluculated at one of the fixed points ($t_e = 153.8$ (ms)) in the previous figure. Note that the minimum phase group delay that was calculated from the amplitude spectrum has the similar shape with the measured group delay. The compensated group delay shown in the right plot approximately has a constant delay about -850 μs under 4 kHz region. This constant delay is used to compensate the event location.

3.2. Duration as a P/N indicator

When the duration of an event is very short in relation to the window length, another factor determines the observed compensated duration. Assume that a background noise has a time independent variance ε^2. The excitation energy of the event is independent of the window shape, because the energy is assumed to be concentrated at the center of the window. Then the total energy $p_w^2(t_e)$ measured at the event location is given by

$$p_w^2(t_e) = \varepsilon^2\sigma_w\sqrt{2\pi}+\eta^2. \qquad (15)$$

Also, the observed duration at the event location $\sigma_P(t_e)$ is represented as follows.

$$\sigma_P^2(t_e) = \langle t^2(t_e)\rangle = \frac{\varepsilon^2\sigma_w^3\sqrt{2\pi}}{2\varepsilon^2\sigma_w\sqrt{2\pi}+2\eta^2} \qquad (16)$$

Using these equations, the apparent P/N (Pulse power to Noise power) ratio $R(t_e)$, defined by the ratio of the event energy η^2 divided by the windowed noise enery $\varepsilon^2\sigma_w\sqrt{2\pi}$, yield the following relation

$$R(t_e) = \frac{\sigma_w^2 - 2\sigma_P^2(t_e)}{2\sigma_P^2(t_e)}. \qquad (17)$$

Figure 3 shows the original speech waveform and the refined estimates of event locations and P/N ratios. Note that the detected event marks are precisely corresponds to the vocal fold closure instants. The P/N ratio levels around 20 dB also indicates that the energy concentration around an event is very high.

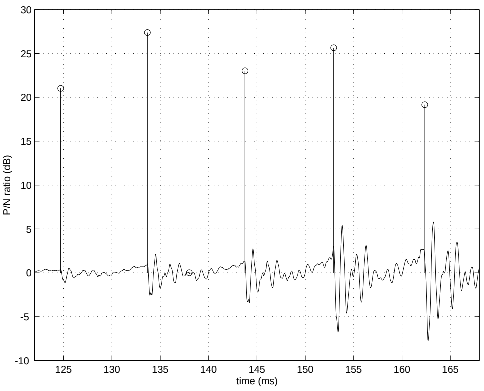

Figure 3. Compensated event locations and the original waveform at the beginning of the utterance. Estimated P/N values are represented by the vertical stems.

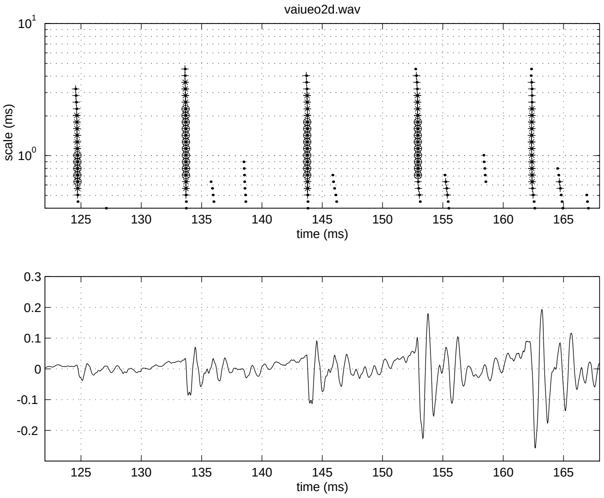

Figure 4. Scale dependency of the detected event. The upper plot shows extracted event locations for different scale paremeter σ_w. The event marks are designed to produce dense symbol when P/N is high. The lower plot shows the corresponding waveform.

3.3. Scale dependency

The proposed method does not require ad-hoc thresholds. The only parameter of the proposed method is the size of the time window determined by σ_w. In other words, the proposed method provides a unique solution once the scale of the phenomena under study is deteminind.

Figure 4 shows the scale dependency of the proposed method for the same speech sample. Event locations were calculated for the range $0.4 < \sigma_w < 5$ (ms) in $2^{1/6}$ steps. Note that the prominent event locations extracted are insensitive to the scale change from 0.5 ms to 3 ms.

4. EVALUATION USING SYNTHETIC SIGNALS

In this section we demonstrate the relevance of the method and evaluate the accuracy of the estimates of attributes using synthetic signals. Various combinations of synthesis architecture were tested in the verification and evaluation of the proposed method. The sampling rate was 22050 Hz throughout the tests.

Excitation source Two types of excitation source signals were used to synthesize speech sounds. First signal is an integrated pulse train with mean value subtraction. The second signal is a differentiated Rosenberg-Klatt (RK) [4] waveform. Open quotient of RK wave was set to 0.66.

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

Table 1. Synthetic speech signals and their conditions

measure	location (μs)			SD (μs)		
SN (dB)	60	40	20	60	40	20
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

Table 2. Event accuracy for synthetic vowel /a/.

Other synthesis parameters Two types of spectral models were used. First model is a minimum phase impulse response calculated from the amplitude spectrum of Japanese vowel /a/ reconstructed using STRAIGHT smoothing. The second model is LPC coefficients calculated from the amplitude spectrum of Japanese vowel /a/ reconstructed using STRAIGHT smoothing. The order of LPC analysis was 30. Both models were pre-processed to have a +6dB/oct frequency weighting for equalization. Table 1 shows a list of synthetic speech samples with their synthesis conditions.

4.1. Estimation accuracy

The described synthetic speech signals were analyzed by the proposed method under several SN conditions. The mean biases and standard deviations of the estimates were evaluated. Differentiation of the signal for analysis pre-processing was also introduced. The standard deviation of the Gaussian time window was set to $\sigma = 1.0$ ms.

Table 2 shows a list of estimation biases and standard deviations for each SN and synthetic signal condition. Gaussian white noise was used as an additional noise source. For each condition, 500 events of vocal fold closures were evaluated. These results indicate that the standard deviation is about one half of one sample period (45 μs) when SN is better than 20 dB. However, in all conditions there is a systematic bias of about one sampling period.

5. EVALUATION BY EGG DATABASE

EGG database A database that consists of simultaneously recorded EGG (electroglottograph) and speech signals was analyzed to evaluate the estimation accuracy of vocal fold closure locations. The EGG database consists of 840 utterances of Japanese sentences spoken by 14 male subjects and 14 female subjects. This database was created using a sampling rate of 16 kHz.

EGG events and speech events Using events extracted from EGG as reference points, the distribution of relative position of events extracted from speech waveforms was investigated. Figure 5 shows the scatter plot and histogram for a male speaker (M04). A window with $\sigma_w = 1$ ms was used. The upper plot shows the scatter plot in the location - duration plane. The duration is represented in terms of P/N ratio. The bottom plot is the histogram of event locations. The standard deviation of locations was less than 60 μs, when P/N ratio is larger than 7 dB. Figure 6 shows results for a female speaker (F04). Here a window with $\sigma_w = 0.7$ ms was used. P/N ratios are generally smaller than the male data. The standard deviations for 14 male speakers was found to be distributed from 60 μs to 170 μs. The standard deviations for the female speakers was found to be distributed from 80 μs to 210 μs. Events with P/N ratios larger than 7 dB were taken into account in this statistics.

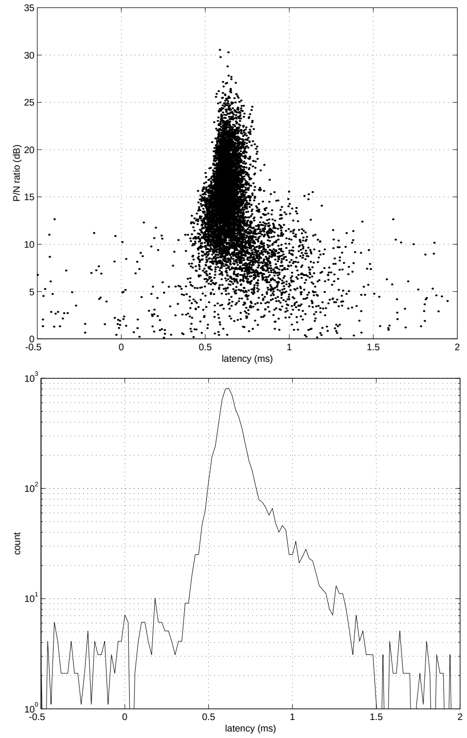

Figure 5. Distribution of event location and peak factor for a male speaker (M04).

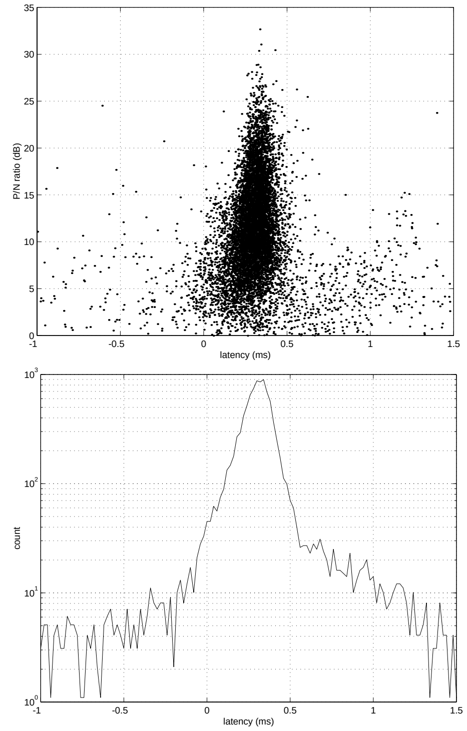

Figure 6. Distribution of event location and peak factor for a female speaker (F04).

Figure 7 shows the standard deviation of the estimated location as a function of the P/N threshold to extract events. Each line represents a single male speaker.

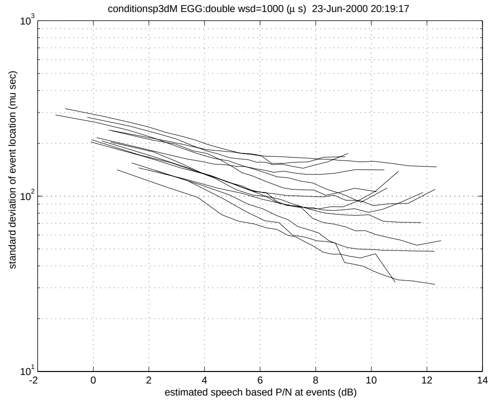

Figure 7. Standard deviations of location estimation errors in terms of minimum P/N for event detection. Each line represents the result of a single male speaker.

6. DISCUSSION

The EGG event and the corresponding glottal airflow discontinuity does not necessarily coincide with each other, suggested by schematic analyses of the relation between the EGG signal and the glottal airflow [1, 6]. However, the proposed method still provides useful information about events in speech as it does not rely on a vowel production model and detects events as points where energy is concentrated. It is a scale dependent non parametric event detector and can be used for various types of sounds.

The simple shape of the compensated group delay suggests that the proposed method provides a powerful tool to investigate group delay characteristics of sounds. It is likely that the proposed method will provide important clues to enhance the high-quality speech analysis/modificatin/synthesis method STRAIGHT [3, 7]

7. CONCLUSION

A new procedure to extract events in a time series was proposed based on a mapping from the location of the time window to the mean time. It was demonstrated that the proposed method can reliably extract events like a vocal fold closure. It also enables the characterization of extracted events by the estimated duration of the events. Furthermore, refinements to the location and duration, by compensating effects of minimum phase component enabled highly precise estimates of the attributes. The proposed method yields a powerful research tool for speech.

REFERENCES

[1] 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.

[2] L. Cohen. *Time-frequency analysis.* Prentice Hall, Englewood Cliffs, NJ, 1995.

[3] 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.

[4] 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.

[5] A. Oppenheim and R. Schafer. *Discrete-Time Signal Processing.* Prentice Hall, Englewood Cliffs, NJ, 1989.

[6] I. Titze. *Principles of voice production.* Prentice Hall, 1994.

[7] Parham Zolfaghari and Hideki Kawahara. Investigation of analysis and synthesis parameters of STRAIGHT by subjective evaluations. In *Proc. ICSLP'2000*, Beijin, 2000. [to appear].