Delay compensation between the speech signal and the corresponding electroglottograph signal

Hartmut R. Pfitzinger and Uwe D. Reichel, LMU Munich

A typical close-talk microphone has a distance of approx. 20 cm from the glottis. This leads to a delay between the recorded speech signal and the electroglottograph signal of approx. 0.6 ms which is 28 samples at 48 kHz sampling rate. This delay poses a problem for glottal epoch detection reference databases because the exact delay depends on the position of the microphone capsule and on the air temperature in the recording room.

To precisely estimate and compensate the delay we use the following procedure: An autocorrelation-based LPC of the pre-emphasized speech signal yields the coefficients for inverse filtering. This is performed applying energy normalization and followed by second differentiation of the resulting signal leading to the third derivative of the quasi-glottal flow. This signal has the particular property that each glottal epoch is reflected by a small number of positive and subsequent negative samples. The same property can be found at the second derivative of the electroglottograph signal. This similarity is large enough to guide a normalized cross-correlation with a window size of 200 ms to very reliably yield a maximum value at the local delay between the two signals.

A short analysis of the delay data of several speakers shows that the delay varies within half a second by 0.1 ms which means that a global compensation of the delay leads to a typical alignment error of approx. 0.05 ms.

Bibliography

Coulton, R. H. and E. G. Conture (1990). Problems and pitfalls of electroglottography. *J. of Voice 4*(1), 10–24.

Fant, G. (1979). Glottal source and excitation analysis. Speech Transmission Lab., Quarterly Progress and Status Report 1, KTH, Stockholm, 85–107.

Höge, H., B. Kotnik, Z. Kačič, and H. R. Pfitzinger (2006). Evaluation of pitch marking algorithms. In: *ITG-Fachbericht 192: Sprachkommunikation*. Berlin, Offenbach: VDE Verlag.

Mokhtari, P., H. R. Pfitzinger, and C. T. Ishi (2003). Principal components of glottal waveforms: Towards parameterisation and manipulation of laryngeal voice-quality. In: *Proc. of the ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis (Voqual'03)*. Geneva, 133–138.

Pfitzinger, H. R. (1998). The collection of spoken language resources in car environments. In: *Proc. of the first Int. Conf. on Language Recources and Evaluation (LREC '98)*, vol. 2. Granada; Spain, 1097–1100.

Pfitzinger, H. R. (2002). Intrinsic phone durations are speaker-specific. In: *Proc. of ICSLP '02*, vol. 2. Denver, 1113–1116.

Pfitzinger, H. R. (2004). DFW-based spectral smoothing for concatenative speech synthesis. In: *Proc. of ICSLP '04*, vol. 2. Korea, 1397–1400.

Pfitzinger, H. R. (2005). Influence of differences between inverse filtering techniques on the residual signal of speech. In: *Fortschritte der Akustik (DAGA '05)*, vol. 1. Munich: Deutsche Physikalische Gesellschaft (DPG), 223–224.

Rothenberg, M. (1973). A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *J. of the Acoustical Society of America 53*(6), 1632–1645.

Strik, H. (1998). Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses. *J. of the Acoustical Society of America 103*(5), 2659–2669.

Delay Compensation between the Speech Signal and the Corresponding Electroglottograph Signal

Hartmut R. Pfitzinger & Uwe D. Reichel

Institute of Phonetics, LMU, Munich, Germany
{hpt|reichelu}@phonetik.uni-muenchen.de
http://www.phonetik.uni-muenchen.de/~hpt/

AST Workshop, Univ. of Maribor, Slovenia

6th of July, 2006

Introduction

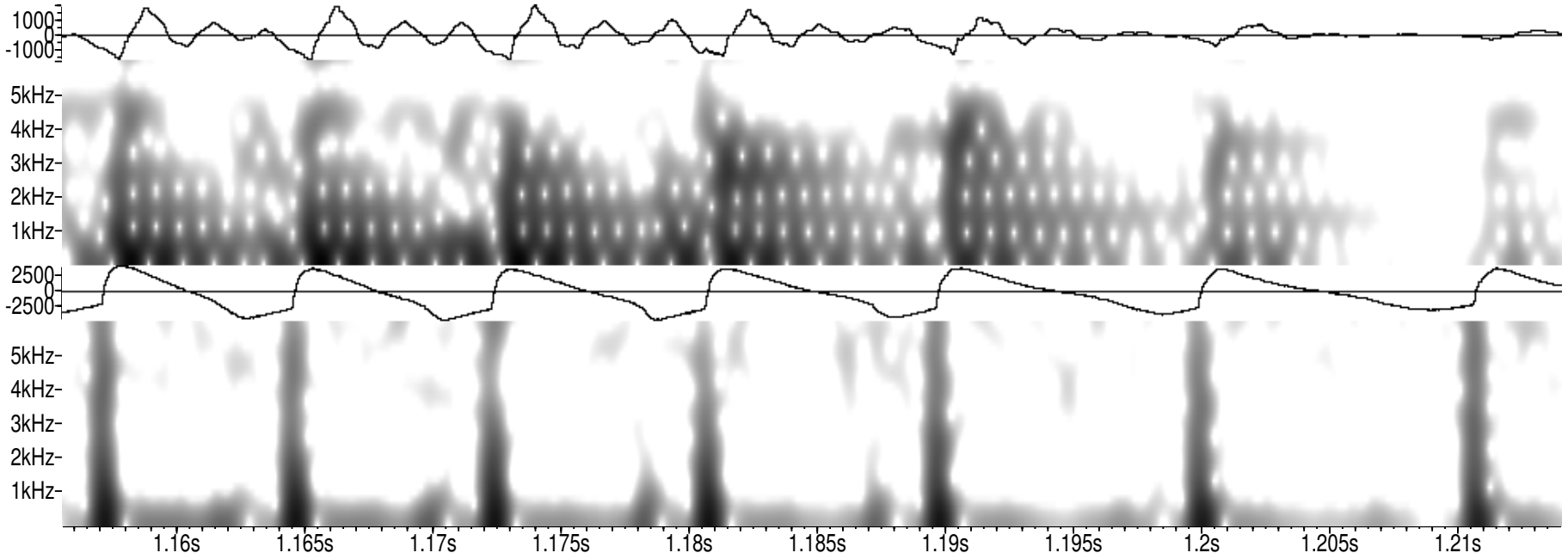
- Multi-channel speech recordings have **channel-individual delays**.
- A speech database providing **glottal excitation positions** needs **compensation for the delays**.
- The delay of the **electroglottograph (EGG) signal** is **difficult** to estimate.
- **This paper presents a solution for the delay estimation.**

Details of the Problem

- The **speed of sound** in air is quite slow: 340 m/s.
- **Voiced sounds are produced** by the glottis.
- The sound travels through the **vocal tract** (length: approx. 17 cm).
- A microphone **cannot be positioned closer** than 20–25 cm to the glottis.
- Speech recordings with an additional electroglottograph (EGG) show a **minimum delay** of $\frac{0.20 \text{ m} - 0.25 \text{ m}}{340 \text{ m/s}} = \mathbf{0.59 \text{ ms} - 0.74 \text{ ms}}$.
- Two microphones produce **sufficiently similar signals for cross-correlation** to estimate the delay.
- But the **EGG signal is too different** and cross-correlation fails.
- A new method to **reliably estimate the delay** is needed.

Two Possible Solutions

I. Use a **two-dimensional cross-correlation** between the sonagrams of the speech signal and the EGG signal:



II. Use inverse filtering of the speech signal to estimate the quasi-glottal flow and cross-correlate it with the EGG signal, which is similar to the glottal flow.

Method Step I: Transform the Speech Signal

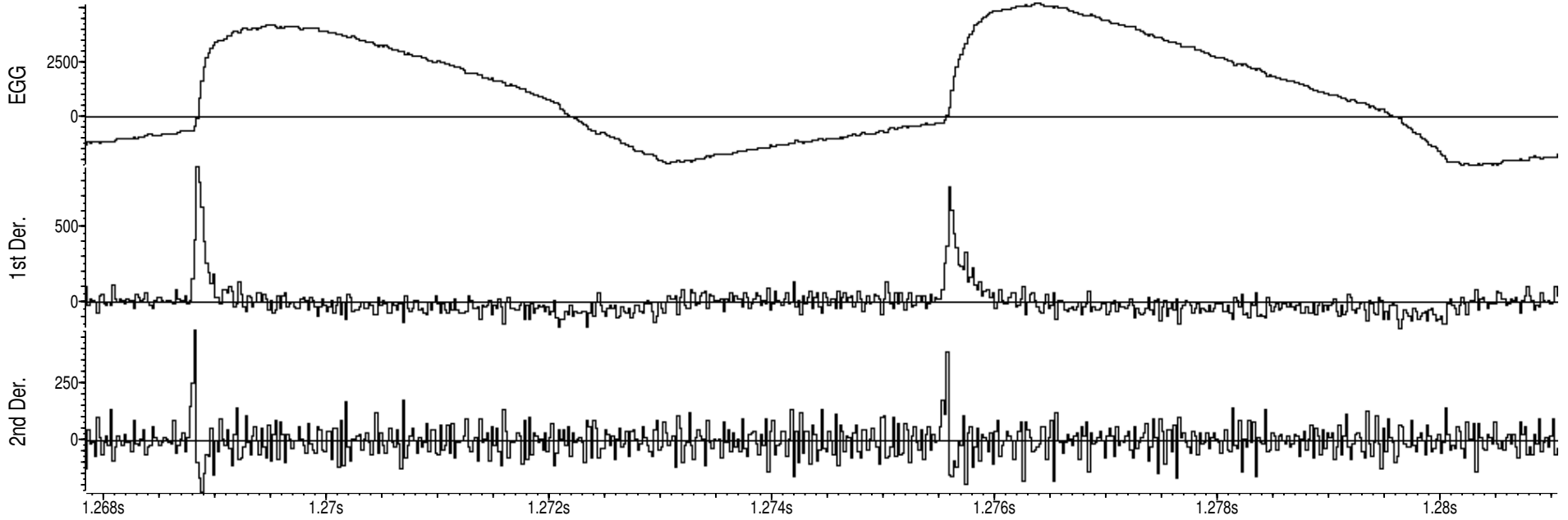
- Pre-emphasis of the speech signal (+6dB/Oct.)
- Autocorrelation-based LPC (every 10 ms, 50 coefficients for 48 kHz)
- Inverse filtering with amplitude normalization
- Differentiation

The result is the 3rd derivative of the quasi glottal flow

Method Step II: Transform the EGG Signal

- Estimate 2nd derivative of the EGG signal

The result is a signal similar to the 3rd derivative of the quasi glottal flow



Method Step III: Cross-Correlation

- Apply **normalized cross-correlation**:

$$\text{ncc}_l = \frac{\sum_n x_n y_{n+l}}{\sum_n x_n^2 + \sum y_{n+l}^2}$$

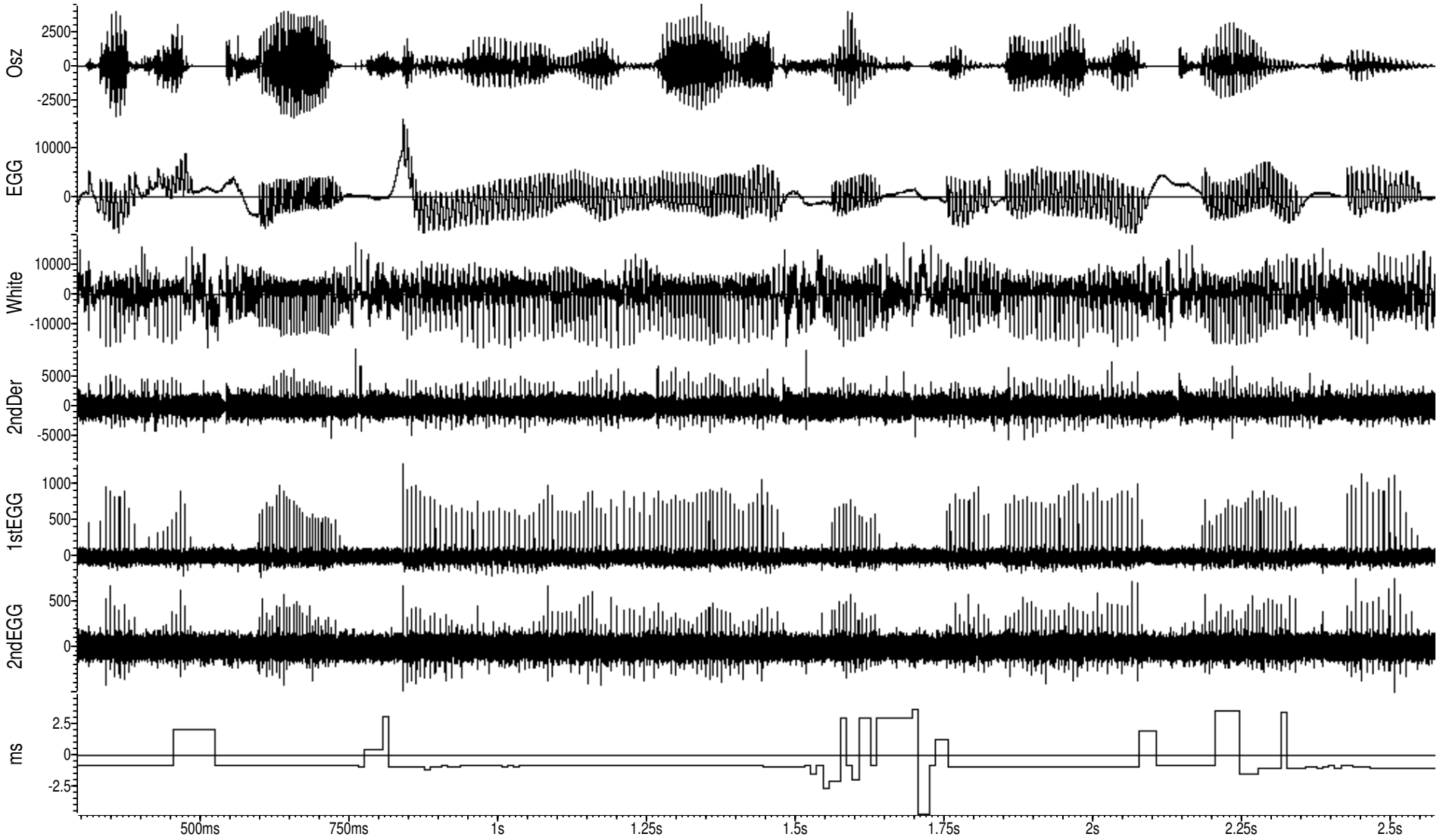
- Practical parameters are:
  step size of 10 ms,
  rectangular window
  with 200 ms duration

- Pick the **maximum** of the normalized cross-correlation function (its the delay!).

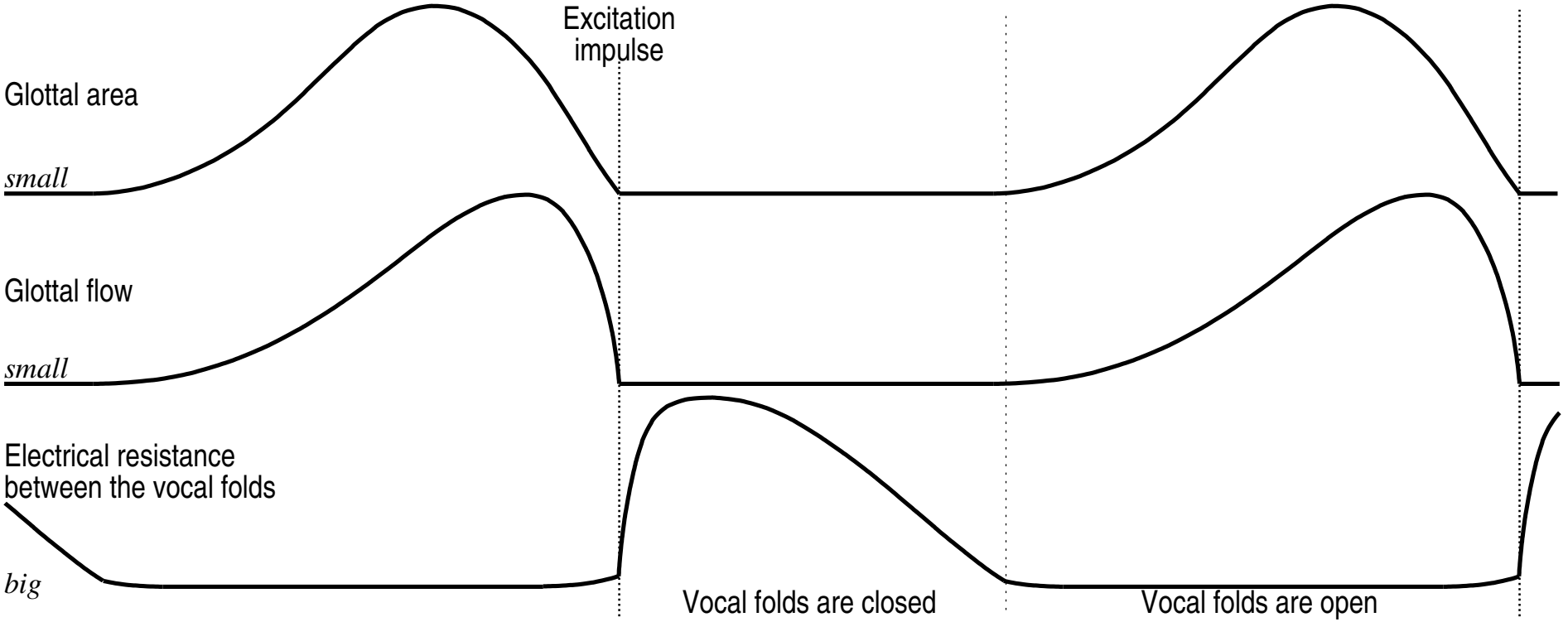
- Estimate the **modal value** of those frames of one utterance, which are **voiced**.

This leads to the global delay value!

Example


Osz
2500
0
-2500
EGG
10000
0
White
10000
0
-10000
2ndDer
5000
0
-5000
1stEGG
1000
500
0
2ndEGG
500
0
ms
2.5
0
-2.5
500ms
750ms
1s
1.25s
1.5s
1.75s
2s
2.25s
2.5s

Explanation: Relation between Glottal Flow and EGG Signal



(After Colton&Conture 1990)

The glottal flow and the EGG signal have **one crucial common feature**:
The **discontinuity** at the excitation impulse
(which is emphasized by differentiation)

Conclusion

- The proposed procedure is **fast and reliable**
- The best method to find the **correct global delay value** is a **continuity criterion** in the time-domain (and not an extensive analysis of minima and maxima in the correlation functions).
- The data of several speakers reveal that: **the local delay varies by ±0.05 ms within half a second of speech.**

Thus global compensation of the delay applied individually to each utterance leads to a typical alignment error of approx. ±0.05 ms.