On the (Glottal) Inverse Filtering of Speech Signals

AN INTRODUCTION

CS578-DIGITAL SPEECH SIGNAL PROCESSING

INVITED LECTURE

On the (Glottal) Inverse Filtering of Speech Signals

- Introduction
- Inverse Filtering Techniques
- Conclusions

Introduction

On the (Glottal) Inverse Filtering of Speech Signals

- The human speech production system is a complicated system
- From an engineering point of view, it can be roughly divided into three parts [1]
  - The vocal folds, which is the source of the system
  - The vocal tract filter, which is the path from the vocal folds to the lips
  - The lip radiation, which is the final bound before system output

On the (Glottal) Inverse Filtering of Speech Signals

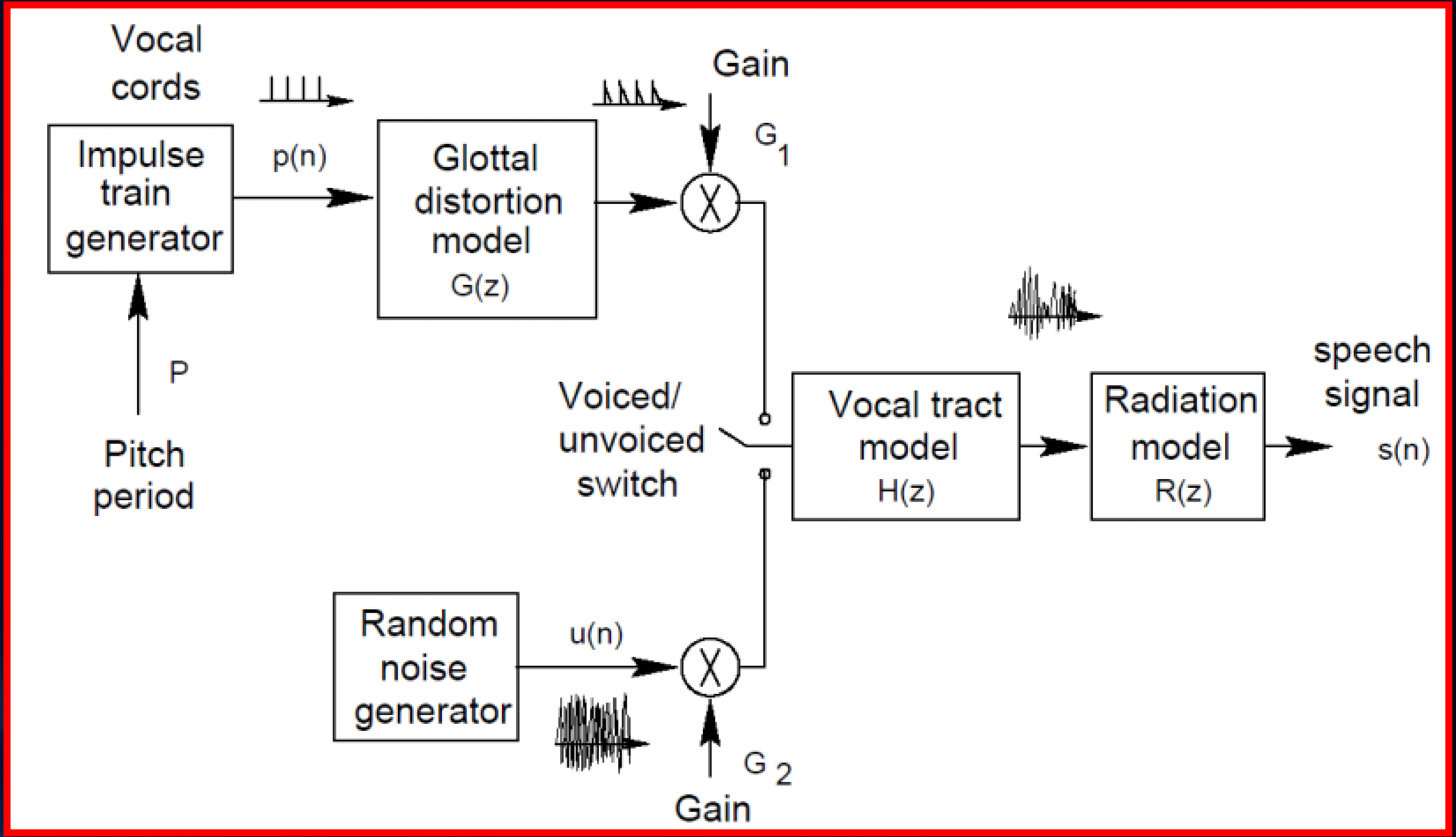
- Based on this simplification, voiced speech can be modeled as a linear filtering operation:

$$s(t) = g(t) * h(t) * r(t) \leftrightarrow S(z) = G(z)H(z)R(z)$$

where $*$ denotes convolution and

  - $g(t)$ is the glottal airflow velocity waveform
  - $h(t)$ is the vocal tract filter
  - $r(t)$ is the lip radiation filter

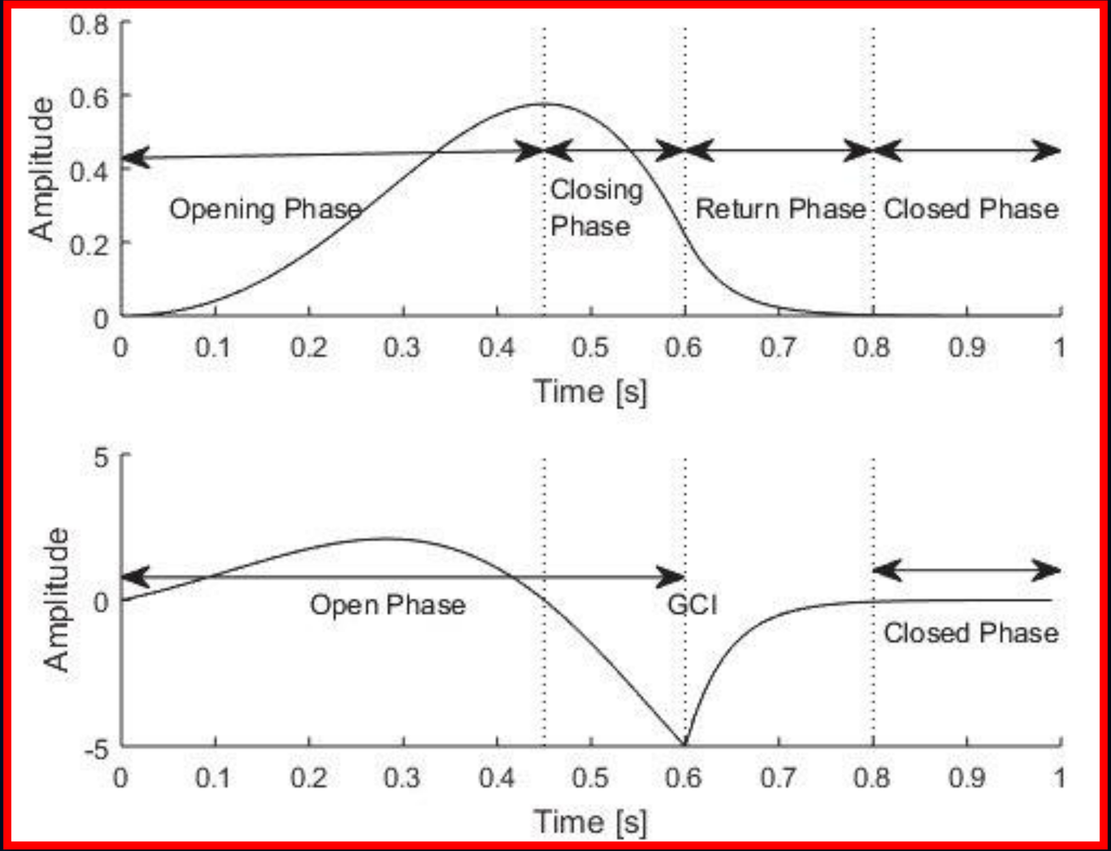
On the (Glottal) Inverse Filtering of Speech Signals



On the (Glottal) Inverse Filtering of Speech Signals

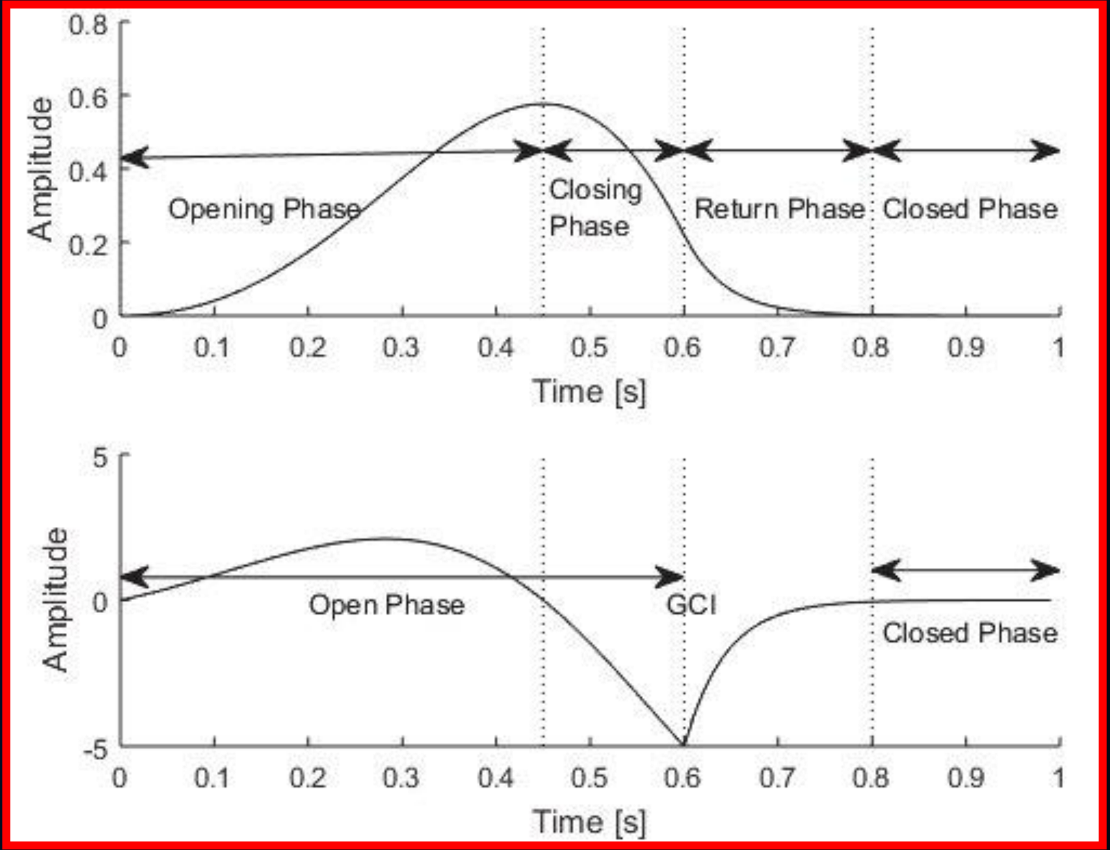
- What is Glottal Inverse Filtering (GIF)?
  - GIF refers to techniques for obtaining the source of voiced speech, the glottal airflow velocity waveform, from voiced speech itself [10]
  - How does this signal look like?

  - Open phase: air flows through the glottis
  - Return phase: vocal folds are snapping shut
  - Closed phase: glottis is shut and airflow velocity is zero



On the (Glottal) Inverse Filtering of Speech Signals

- While radiation occurs after the vocal tract filter, we often combine $G(z)$ and $R(z)$ into a single expression
  - This applies the radiation effect to the glottal source **before** it enters the vocal tract
  - Effect of differentiation on the source
- The resulting signal is the so-called **glottal flow derivative**
- Very commonly used in literature



On the (Glottal) Inverse Filtering of Speech Signals

- Why bother?
  - Basic research of speech production
  - Applications to speech analysis, synthesis, and modification
  - Environmental voice care
  - Voice pathology detection
  - Analysis of the emotional content of speech
  - Voice source modeling for TTS

On the (Glottal) Inverse Filtering of Speech Signals

- Basic idea:
  - Form a computational model for the vocal tract filter, $H(z)$
  - Cancel its effect from the speech waveform by filtering the speech signal through the inverse of the model, $\frac{1}{H(z)}$

On the (Glottal) Inverse Filtering of Speech Signals

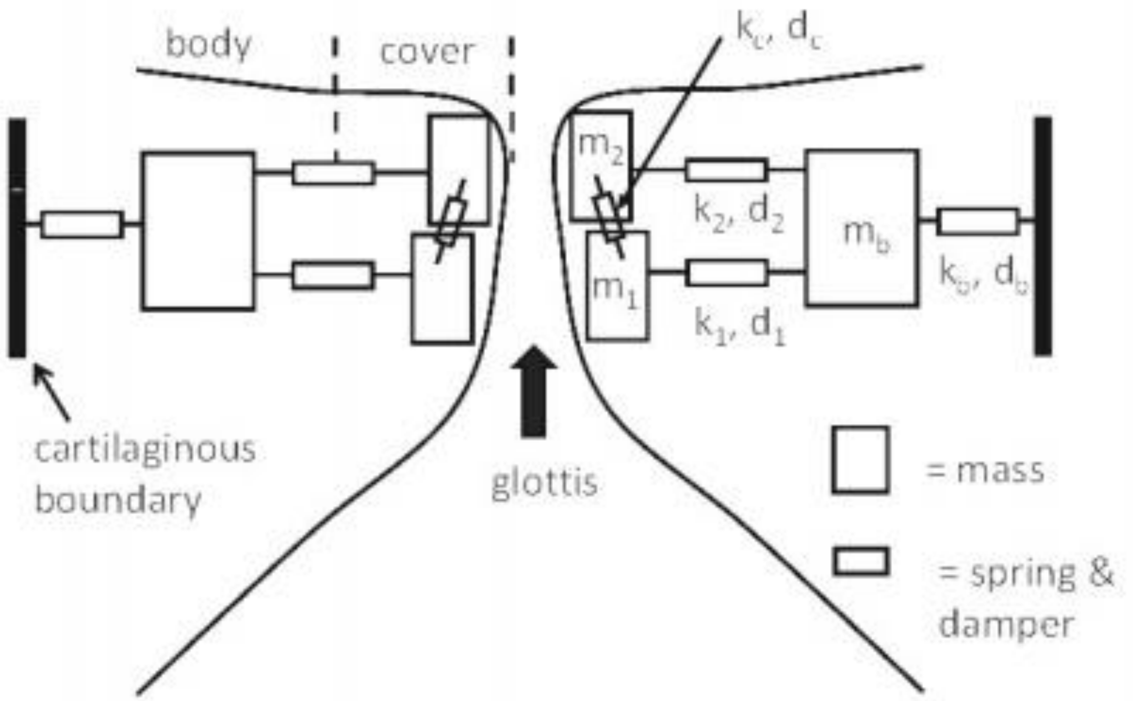
- **Problem:**
  - The actual glottal flow waveform IS NOT AVAILABLE!
  - …at least in a non-invasive manner [18]
- **Approaches:**
  - "Visual" inspection of the resulting glottal flow waveform
  - Use of synthetic speech signal produced by a known artificial excitation
  - Compare the results of different GIF algorithms
- None of the previous approaches is truly objective

On the (Glottal) Inverse Filtering of Speech Signals

- One solution is to build a physical model of the speech production mechanism
  - Generate waveforms from this model
- Time-varying waveforms are simulated
- Such waveforms are expected to provide a more firm and realistic test of GIF methods
- **Both** the speech output **and** the source are available
- A well known dataset of such signals is described in [2,6]

On the (Glottal) Inverse Filtering of Speech Signals

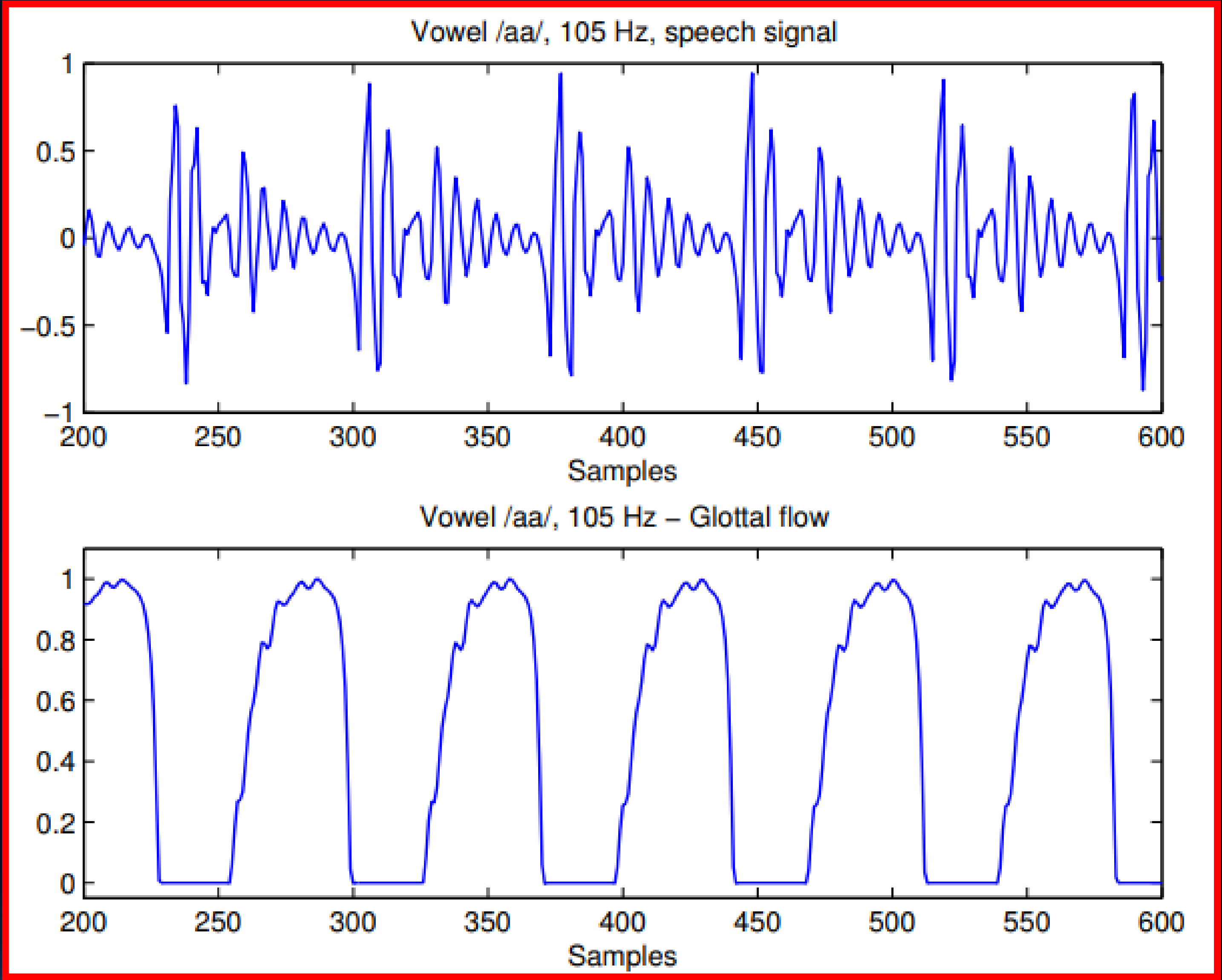
- In detail, self sustained vocal fold vibration was simulated with three masses coupled to one another through stiffness and damping elements.



On the (Glottal) Inverse Filtering of Speech Signals

- The model has a parametrized input such as
  - Lung pressure
  - Prephonatory glottal half-width (adduction)
  - Vocal fold length and thickness
  - Activation levels of the cricothyroid and thyroarytenoid muscles

On the (Glottal) Inverse Filtering of Speech Signals



GIF techniques

On the (Glottal) Inverse Filtering of Speech Signals

- Since we already know about Linear Prediction (LP), we will discuss GIF methods based only on that
- You already know two methods for estimating LP coefficients
  - Autocorrelation method: zero samples outside prediction error interval –minimize MSE everywhere
  - Covariance method: non-zero samples outside prediction error interval – minimize MSE inside prediction error interval

On the (Glottal) Inverse Filtering of Speech Signals

- LP is used to produce all-pole models of the vocal tract filter

$$H(z) = \frac{1}{\sum_{k=1}^{p} a_k z^{-k}}$$

where p is the filter order and a_k are the LP coefficients

- In general, LP minimizes the MSE over a region R

$$E = \sum_{R} e^2[n]$$

where $e[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k]$

On the (Glottal) Inverse Filtering of Speech Signals

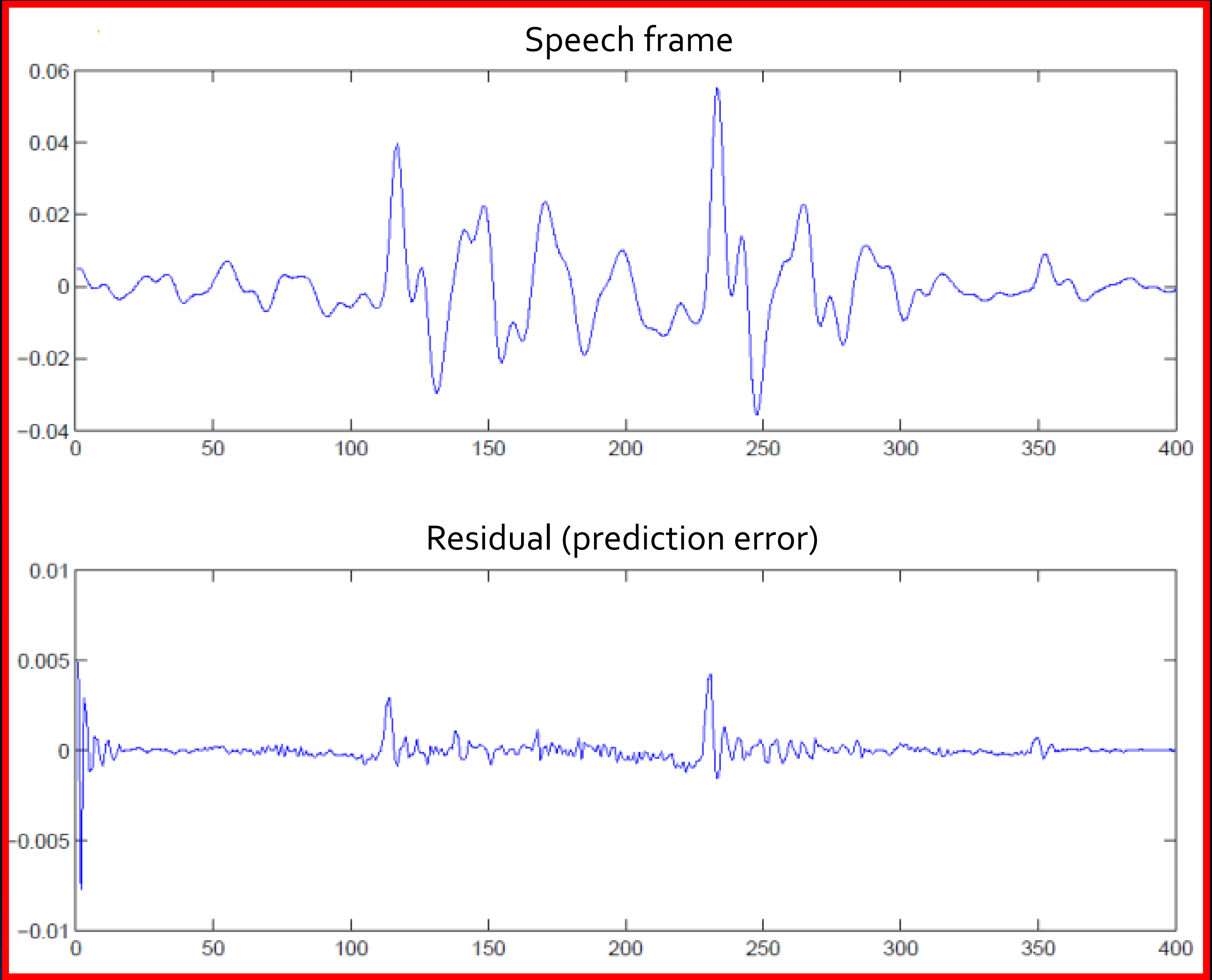
- How can we find the source excitation through LP analysis?
- If we consider speech as an AR process

$$s[n] = Ag[n] + \sum_{k=1}^{p} a_k s[n-k] \Rightarrow Ag[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k]$$

then minimization of the MSE leads to $e[n] \approx Ag[n]$

- Thus, the **prediction error** (or **residual**) can be thought of an estimation of the source excitation
- But how are glottal source and residual related?

On the (Glottal) Inverse Filtering of Speech Signals



On the (Glottal) Inverse Filtering of Speech Signals

- But how are glottal source and residual **related**?
- As you've seen, the two signals do not quite match
- The reason is that the Z transform of speech is a combined transfer function

$$Y(z) = G(z)R(z)H(z)$$

- $G(z)$ can be further decomposed as an impulse sequence passed through a glottal filter: $G(z) = I(z)U(z)$, where $I(z)$ is the impulse sequence
- Thus $Y(z)$ contains

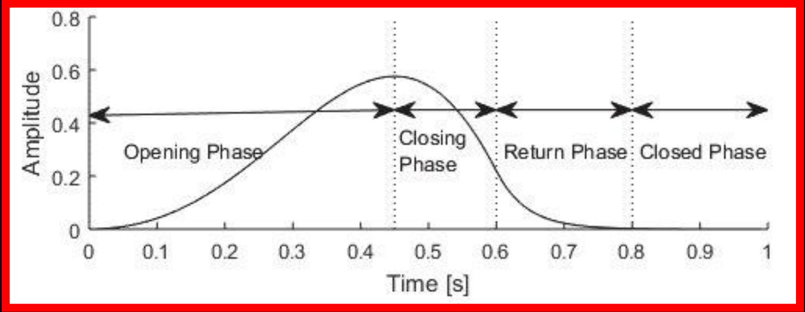
❑ Zeros from the glottal source and lip radiation

❑ Poles from the glottal filter and the vocal tract filter

On the (Glottal) Inverse Filtering of Speech Signals

- LP analysis provides an overall transfer function $\hat{H}(z)$ where all these contributions are combined!
  - …not to mention that we're using an all-pole method for a pole-zero signal…
- So what we are cancelling via simple LP-based GIF is this overall estimation
  - …resulting into something that looks like a series of impulses!
- So how can we work this out?
- Identify instants where there is no interaction between the source and the filter!

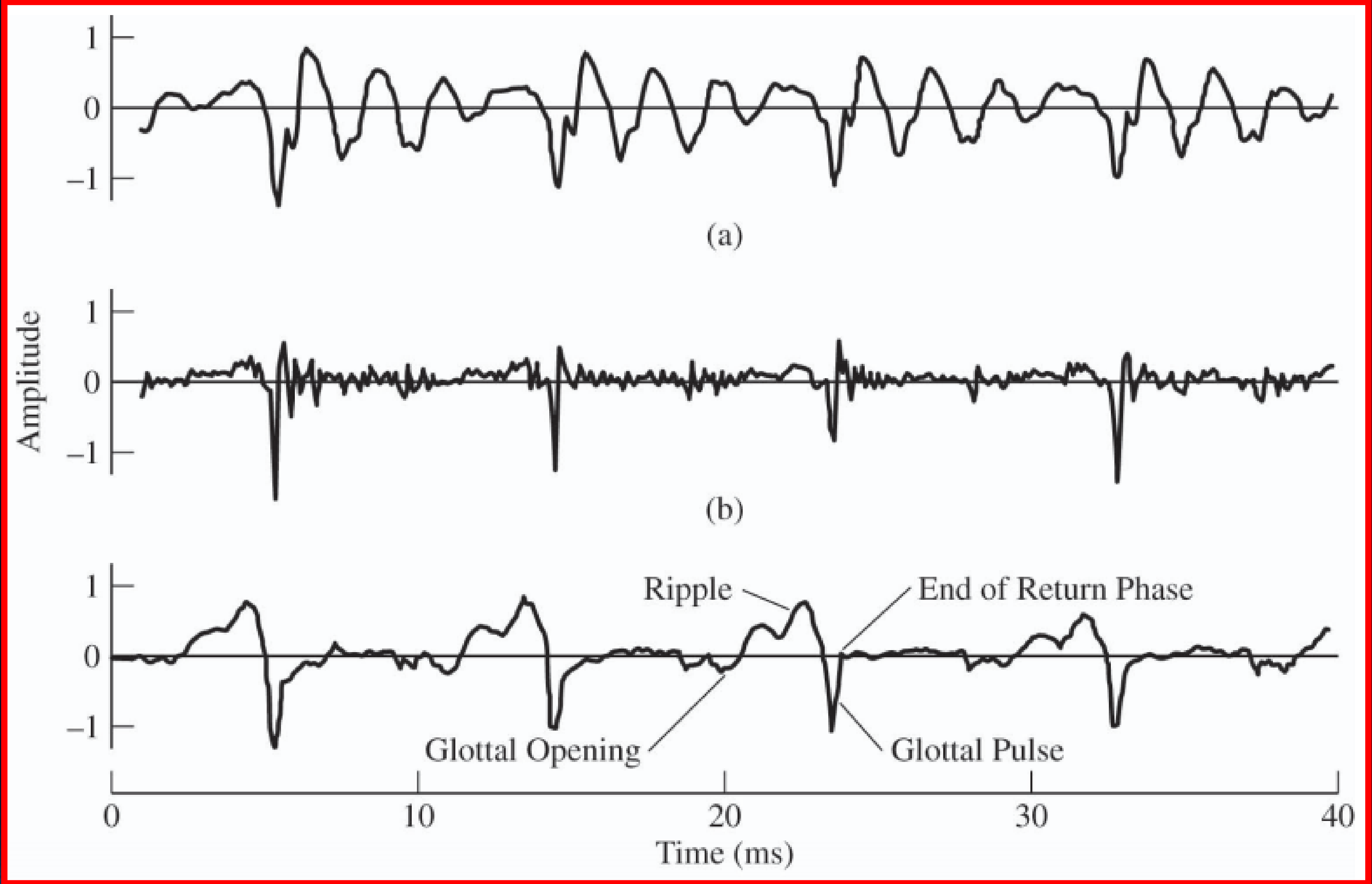
On the (Glottal) Inverse Filtering of Speech Signals



- **Closed-phase analysis**
- Identify regions where the vocal folds are closed
  - No contribution from $G(z)$, the speech signal should contain vocal tract and radiation factors $H(z)R(z)$
  - $R(z)$ can be modeled as a differentiator (single-zero FIR filter), so it can be cancelled by a simple integrator
- Vocal tract estimation in the closed phase region leads to a more precise result
- Estimation in the closed phase → cancelling vocal tract via GIF over the whole pitch period
- Use of covariance-based LP on the closed-phase [9]

On the (Glottal) Inverse Filtering of Speech Signals

- **Closed-phase analysis**



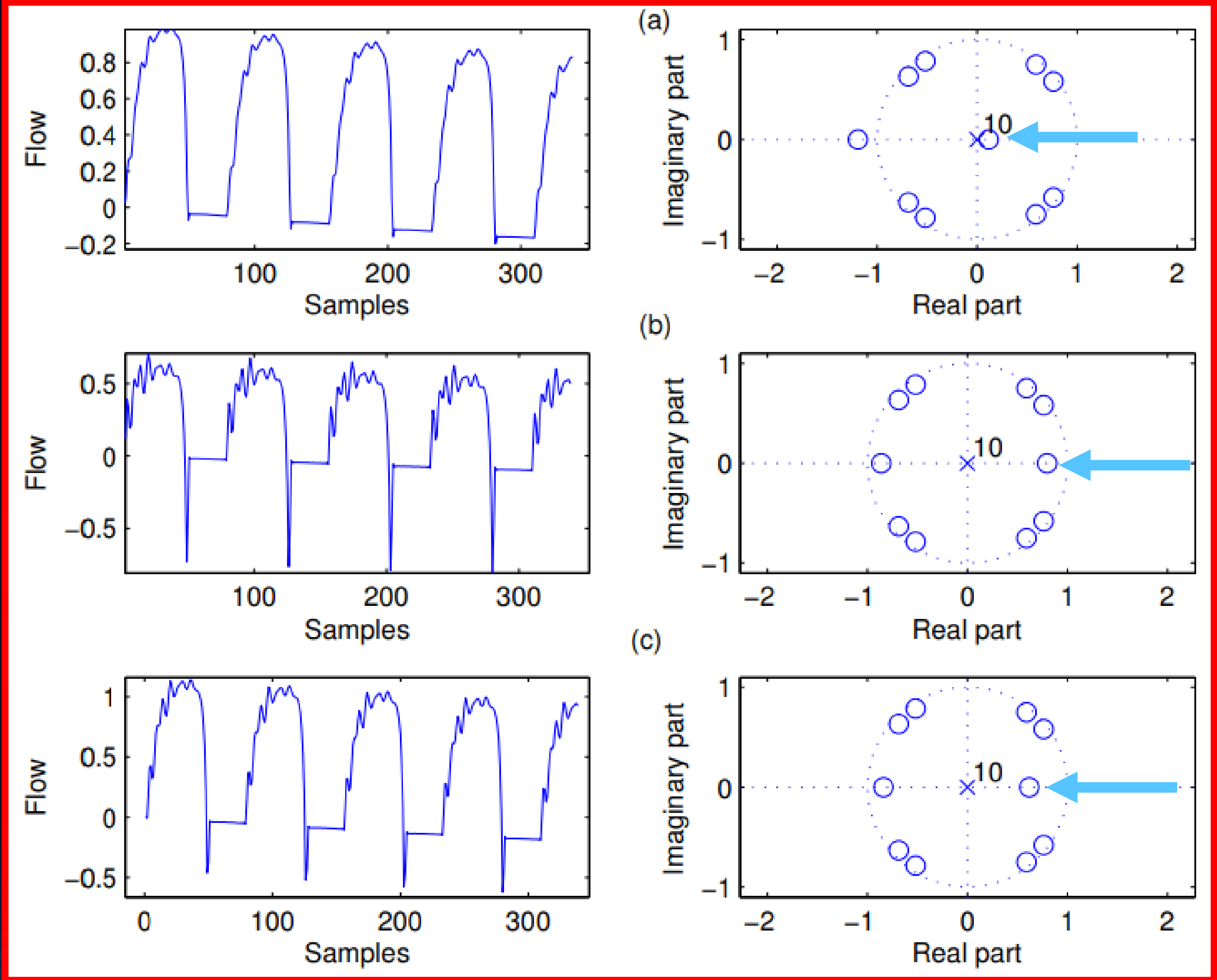
On the (Glottal) Inverse Filtering of Speech Signals

- **However,** standard closed-phase covariance LP suffers from certain shortcomings

- ❑ **Short closed phase duration (especially for high pitched speakers)**
  - **Too few samples to obtain a good estimation**
- ❑ **Sensitivity to the exact position of the covariance frame**
  - **Small variation from the exact closed phase interval produces artifacts**
- ❑ **Vocal tract filter instability**
  - **Covariance-based LP does not guarantee a stable filter**
  - **Inverse filter might not be minimum phase**

On the (Glottal) Inverse Filtering of Speech Signals

- **Frame position sensitivity**

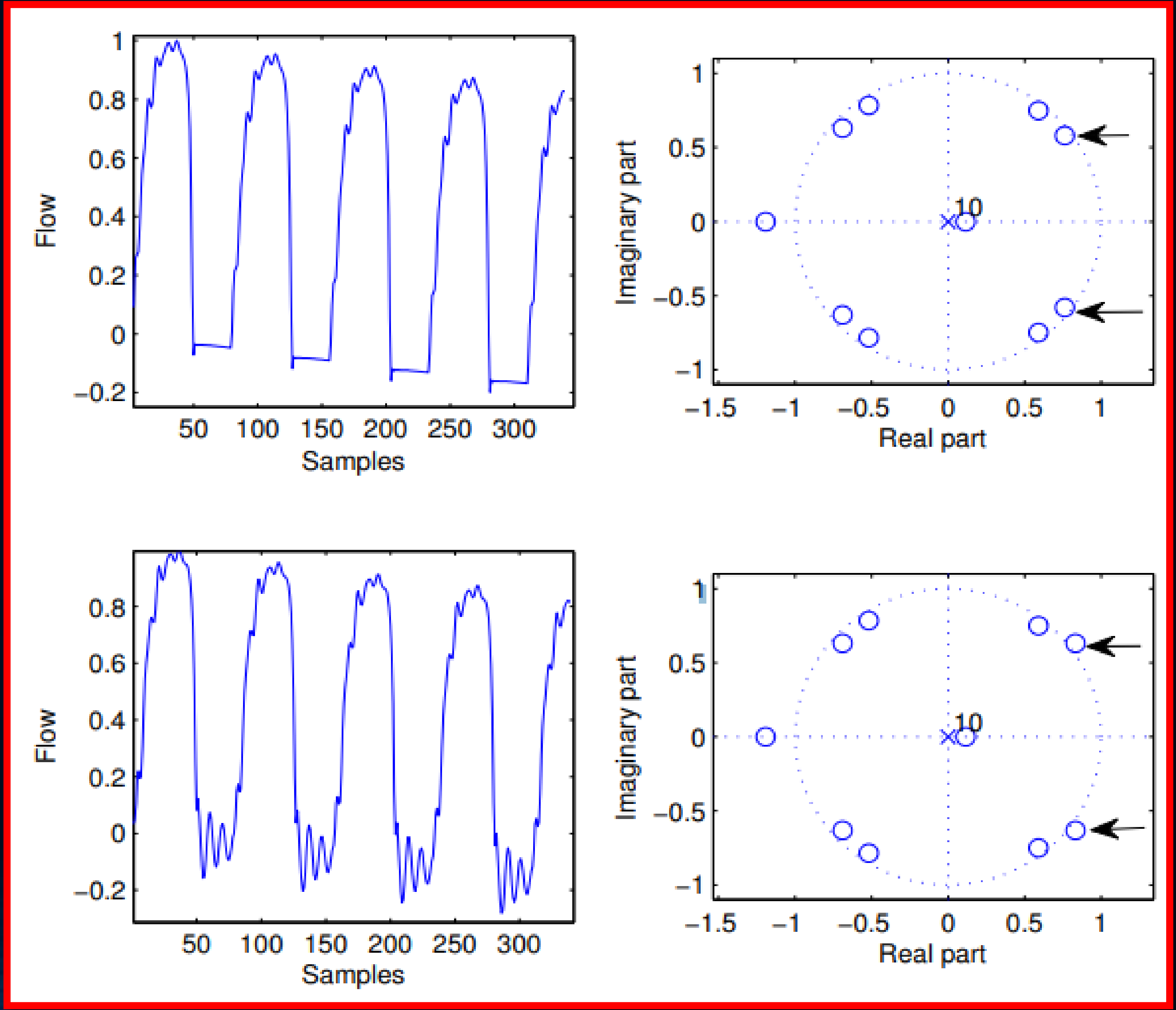


On the (Glottal) Inverse Filtering of Speech Signals

- The effect of an inverse filter root which is located on the positive real axis has the properties of a first order differentiator, when the root approaches the unit circle
- A similar effect is also produced by a pair of complex conjugate roots at low frequencies
- This distortion is more apparent at the time instants where the glottal flow changes more rapidly, that is, near glottal closure
- The presence of such roots are in contrast to the source-filter suggested theory
- The removal of such roots results in less dependency on the covariance frame location

On the (Glottal) Inverse Filtering of Speech Signals

- **Non-minimum phase inverse filter**



On the (Glottal) Inverse Filtering of Speech Signals

- The inverse filter $1/H(z)$ might **not** be minimum phase
- As we know from basic DSP, it can become minimum phase by replacing each zero by its mirror image partner
- That leaves the magnitude spectrum unchanged
- The phase characteristics change, though

On the (Glottal) Inverse Filtering of Speech Signals

- **Constrained Covariance-based Closed-Phase LP [3]**
- Idea: modification of the conventional CP covariance analysis in order to provide more realistic root locations, in the acoustic sense
- How?
  - Not allow mean square error to locate the roots freely on the z-plane
  - Impose mathematical restrictions in a form of concise mathematical equations
  - DC-constraint

On the (Glottal) Inverse Filtering of Speech Signals

- **Constrained Covariance-based Closed-Phase LP [3]**
- DC-constraint:

$$H\left(e^{j0}\right) = \sum_{k=0}^{p} a_k e^{-j0n} = \sum_{k=0}^{p} a_k = l_{DC}$$

- Why?
  - Magnitude response of voiced sounds approaches unity at zero frequency [1]
  - A short and misplaced covariance frame might lead to a response with higher gain at DC than at formants
- With such a constraint, one might expect a better match of the magnitude response to the source-filter theory

On the (Glottal) Inverse Filtering of Speech Signals

- **Constrained Covariance-based Closed-Phase LP [3]**
- Constrained convex minimization problem
- Minimize $a^T \Phi a$ subject to $\Gamma^T a = b$

$$a = \left[1, a_1, \ldots, a_p\right]^T \qquad \Phi = \left[\Phi_{ij}\right], \Phi_{ij} = \sum_{n=0}^{N-1} s[n-i]s[n-j], \qquad 1 \leq i, j \leq p$$

$$b = [1, l_{DC}]^T \qquad \Gamma = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix}^T$$

Solution:
$a = \Phi^{-1}\Gamma(\Gamma^T\Phi^{-1}\Gamma)^{-1}b$

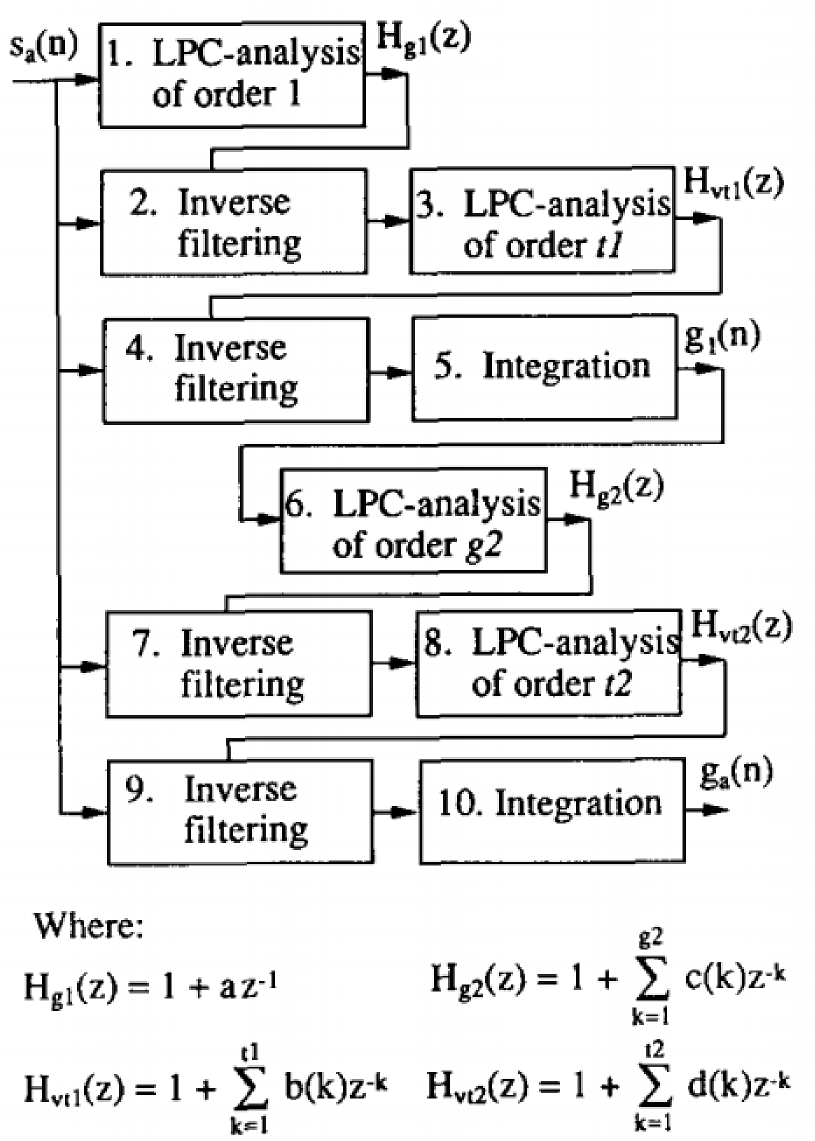
On the (Glottal) Inverse Filtering of Speech Signals

- **Still,** the computational load of covariance-based LP along with its shortcomings (cases of very small CP, frame dependent, CP identification) might make the method not appropriate

- Idea: use autocorrelation method with "enhancements"
  - Fast & stable
  - Not optimal but good enough
  - Try to introduce "enhancements"
  - Try to approach performance of CP analysis without detecting CP

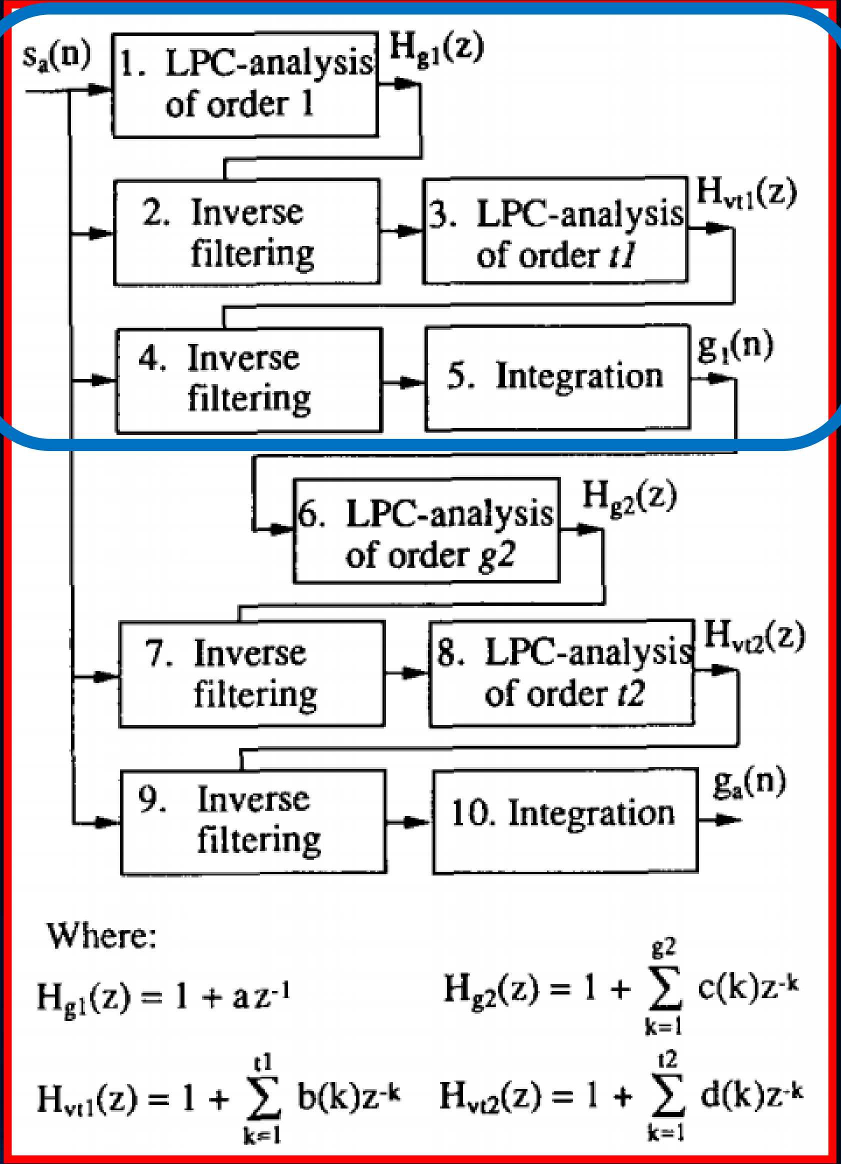
On the (Glottal) Inverse Filtering of Speech Signals

- **Iterative Adaptive Inverse Filtering [4]**
- An iterative method for obtaining the glottal source
- Motivation:
    - A priori knowledge of the overall shape of the vocal tract
    - Cancel the tilting effect of the glottal source
    - Estimate vocal tract filter



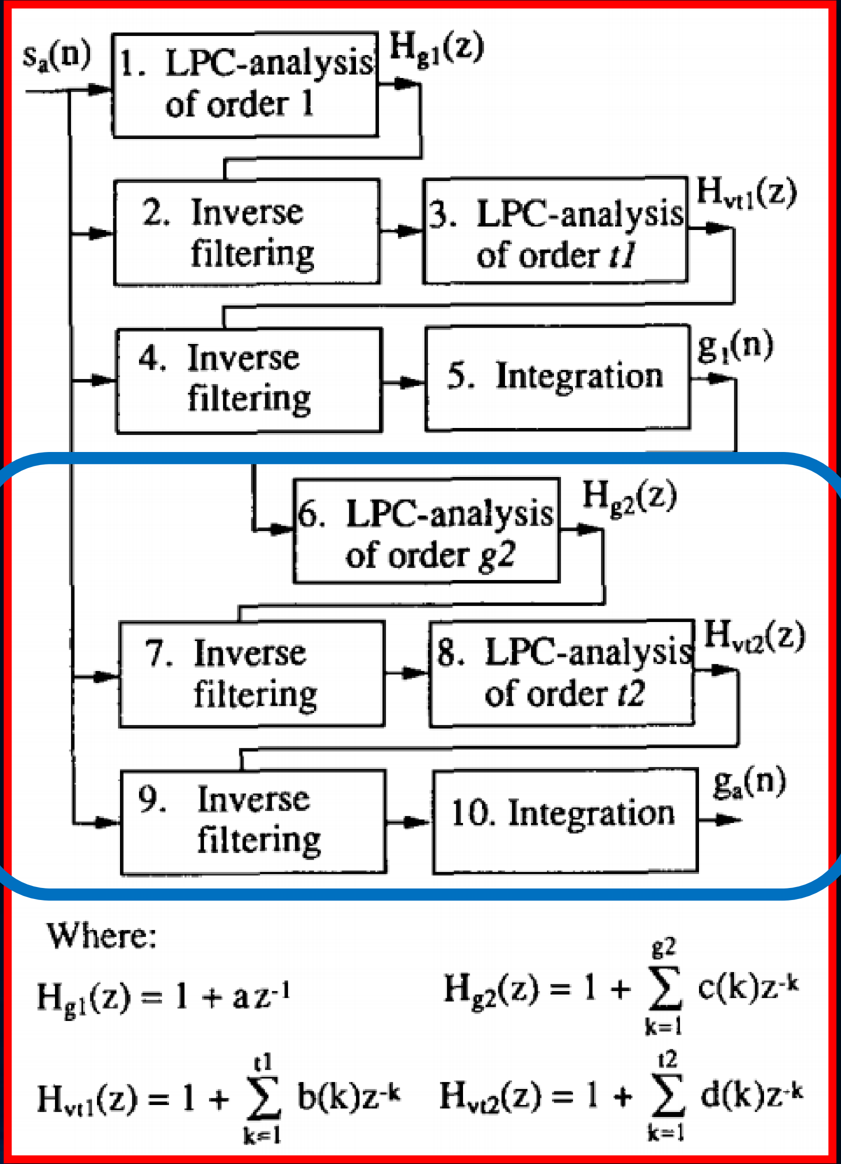
On the (Glottal) Inverse Filtering of Speech Signals

- **Iterative Adaptive Inverse Filtering [4]**
- **First iteration**
- 1. LPC of order 1 to model the effect of the glottal source on the speech spectrum
- 2. Cancel 1.
- 3.LPC of high order to model the vocal tract
- 4&5. Cancel vocal tract and lip radiation



On the (Glottal) Inverse Filtering of Speech Signals

- **Iterative Adaptive Inverse Filtering [4]**
- **Second iteration**
- 6. LPC of order 2-4 to more accurately model the effect of the glottal source on the speech spectrum
- 7. Cancel 6.
- 8.LPC of high order to model the vocal tract
- 9&10. Cancel vocal tract and lip radiation



On the (Glottal) Inverse Filtering of Speech Signals

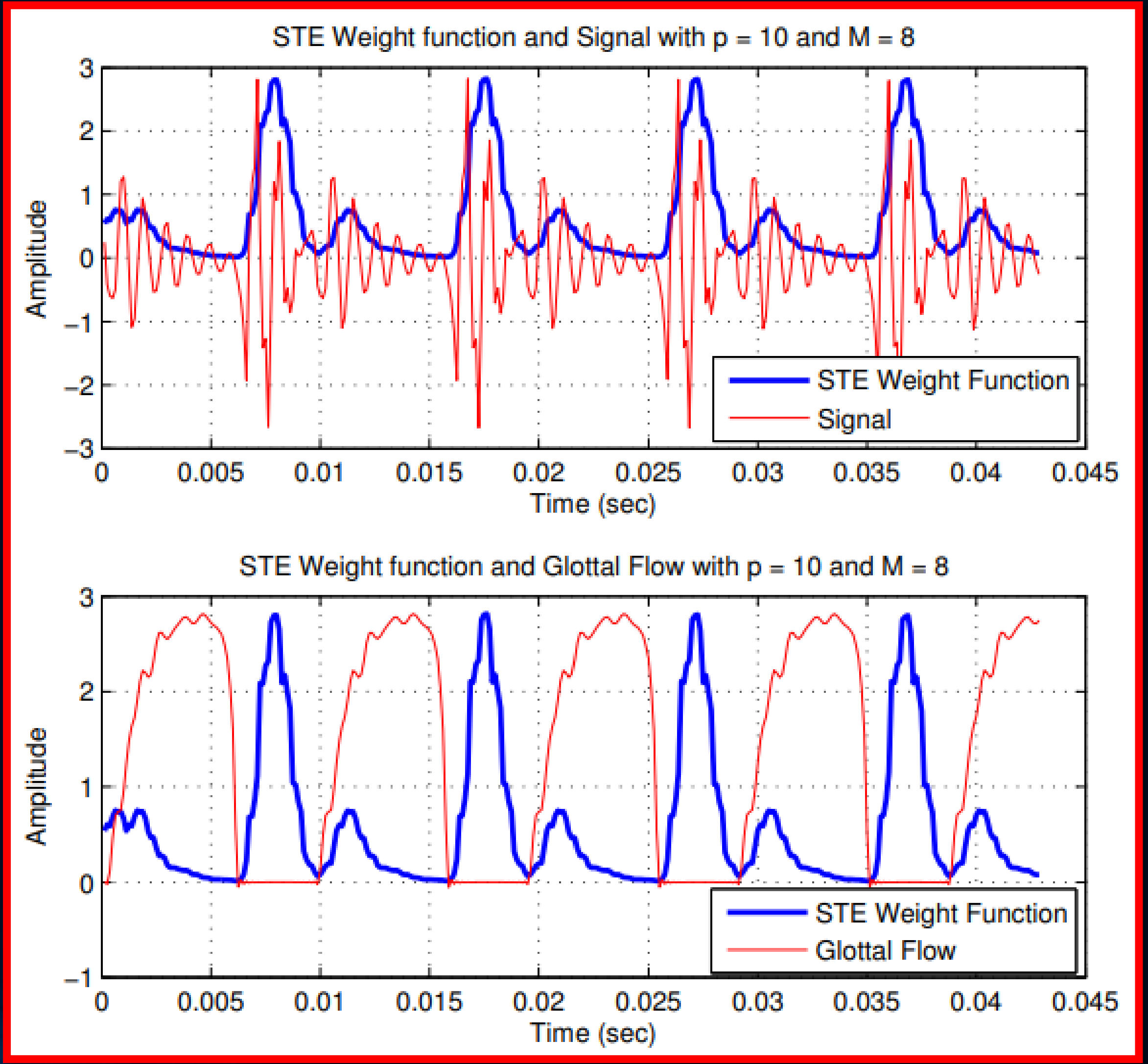
- **Stabilized Weighted Linear Prediction [5,7]**
- An all-pole method based on Weighted Linear Prediction (WLP)
- Idea: use standard autocorrelation method but give more weight to some samples of the autocorrelation matrix compared to others
- How to give more weight?

On the (Glottal) Inverse Filtering of Speech Signals

- **Stabilized Weighted Linear Prediction**

- Compute *the short time energy (STE)* of the signal

- High energy samples fall in the closed phase region!



On the (Glottal) Inverse Filtering of Speech Signals

- **Stabilized Weighted Linear Prediction**
- STE function emphasizes the speech samples of large amplitude, which typically occur during the closed phase interval
- By emphasizing on these samples that occur during the glottal closed phase, it is likely to yield more robust acoustical cues for the formants
- The method depends on a parameter M, the energy window length
- A high value of M increases the sharpness of the resonances of the spectrum, whereas a low value of M increases the smoothness of the spectrum

On the (Glottal) Inverse Filtering of Speech Signals

- **Stabilized Weighted Linear Prediction**
- STE:

$$w_n = \sum_{i=0}^{M-1} x^2[n-i-1]$$

- Prediction error energy:

$$E = \sum_{n=1}^{N+p} e^2[n]\, w_n = a^T \left(\sum_{n=1}^{N+p} w_n x[n] x^T[n] \right) a = a^T R a$$

where

$$R = \sum_{n=1}^{N+p} w_n x[n] x^T[n]$$

On the (Glottal) Inverse Filtering of Speech Signals

- **Stabilized Weighted Linear Prediction**
- Constrained minimization problem (again ☺)
- Minimize E subject to $a^T u = 1$, where $u = [1,0,0, \dots, 0]^T$
- It can be shown that a satisfies the linear equation

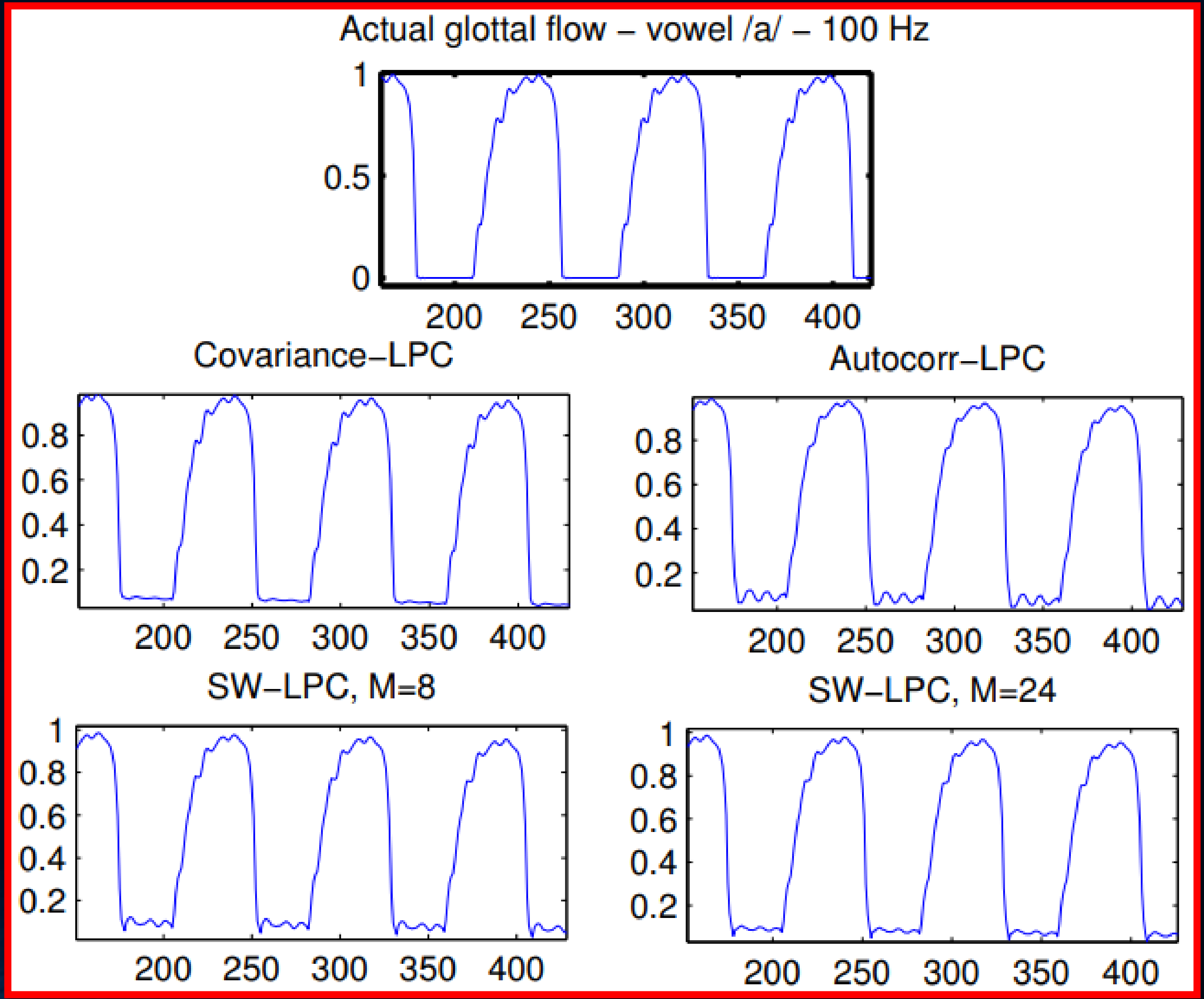
$$Ra = \sigma^2 u$$

where σ^2 is the error energy

- Stability is ensured by a specific algorithm

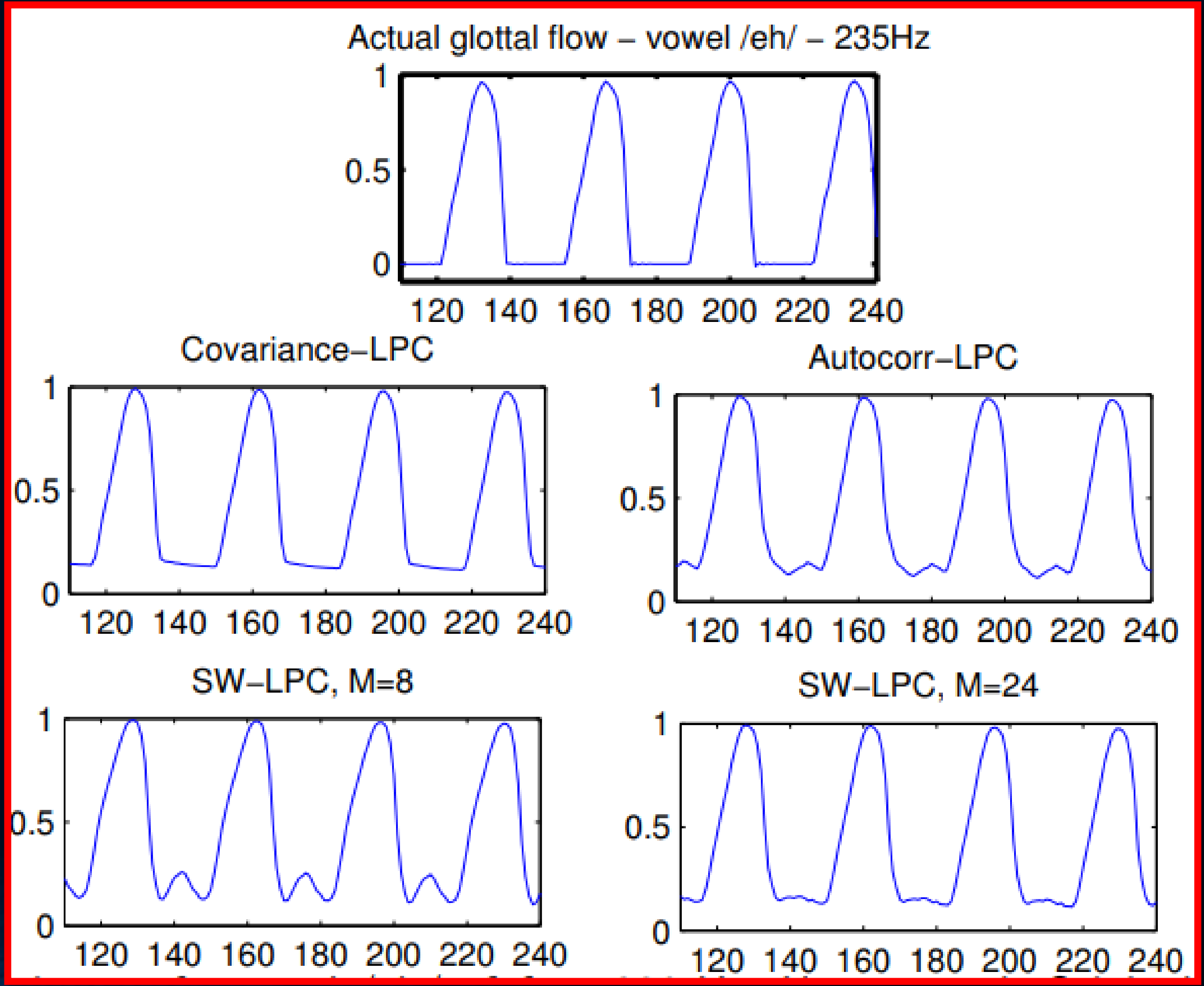
On the (Glottal) Inverse Filtering of Speech Signals

- **Results**



On the (Glottal) Inverse Filtering of Speech Signals

- **Results**



On the (Glottal) Inverse Filtering of Speech Signals

- **Performance metric**

$$SRER = 20\log_{10}\left(\frac{\sigma_s[n]}{\sigma_e[n]}\right)$$

The bigger the better!

Signal to Reconstruction Error Ratio

- $s[n]$: original glottal source
- $e[n]$: error between original and synthetic
- $\sigma_x[n]$: standard dev. of $x[n]$

SRER				
Vowel	$SWLP_8$	$SWLP_{24}$	LPC	CovLPC
$/aa/$	33.5 (±2.0)	39.7 (±4.5)	36.2 (±5.7)	41.9 (±6.3)
$/ae/$	32.7 (±4.4)	35.2 (±2.9)	37.8 (±3.0)	40.4 (±6.4)
$/eh/$	34.0 (±1.9)	38.4 (±4.2)	33.9 (±4.0)	40.5 (±5.2)
$/ih/$	32.3 (±1.5)	37.6 (±3.1)	35.3 (±4.6)	39.2 (±5.6)

On the (Glottal) Inverse Filtering of Speech Signals

The smaller the better!

- **Performance metric**

$$ER(H_1 - H_2) = |Ref(H_1 - H_2) - Synth(H_1 - H_2)|$$

- Ref: original glottal source spectrum
- $Synth$: synthetic glottal source spectrum
- $H_1 - H_2$ is the difference between the first two glottal source harmonics
- This metric is an indication of the spectral tilt

	ER_{H1H2}			
Vowel	$SWLP_8$	$SWLP_{24}$	LPC	CovLPC
$/aa/$	0.68 (±0.10)	0.23 (±0.09)	0.75 (±0.09)	0.20 (±0.20)
$/ae/$	0.15 (±0.12)	0.15 (±0.05)	0.55 (±0.05)	0.18 (±0.13)
$/eh/$	0.34 (±0.09)	0.30 (±0.07)	0.54 (±0.08)	0.38 (±0.17)
$/ih/$	0.72 (±0.14)	0.39 (±0.11)	0.85 (±0.12)	0.35 (±0.24)

Conclusions

On the (Glottal) Inverse Filtering of Speech Signals

- GIF has been around for more than five decades

- Attractive analysis method
  - Non-invasive
  - Using only speech signal
  - Mostly automatic
  - Applications in many speech technologies
  - Still improving! (QCP Analysis [16])

- Software: OPENGlot, Aparat, etc

On the (Glottal) Inverse Filtering of Speech Signals

- GIF has been around for more than five decades

- Shortcomings
  - Recording should be made with caution
    - Introducing non-linearities that distort GIF result
  - "Ground truth" is very rarely available
    - Synthetic speech or physiologically modeled data is used
  - Unreliable analysis of certain voice types [11,12]
    - High-pitch speech, low F1, vulnerability of best method (closed-phase CP)
  - Based on all-pole methods → speech is pole-zero (nasal sounds)
  - Fixed filter coefficients over successive periods

References

- [1] Fant G., Acoustic Theory of Speech Production (Mouton, The Hague).
- [2] Story, B. and Titze, I. Voice simulation with a body-cover model of the vocal folds," J. Acoust. Soc. Am., 97, 1249–1260, 1995.
- [3] Alku, P., Magi, C., Santeri, Y., Backstrom, T., Story, B., Closed phase covariance analysis on constrained linear prediction for glottal inverse filtering, Journal of Acoustical Society of America, 125(5):3289-3305, 2009.
- [4] Alku, P., Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. Speech Communication, *11(2-3), 109–118, 1992.*
- [5] Magi, C. , Pohjalainen, J., Backstrom, T., Alku, P., Stabilised Weighted Linear Prediction, Speech Communication, 51:401-411, 2009.
- [6] Alku P., Story B., Airas M., Estimation of the voice source from speech pressure signals: Evaluation of an inverse filtering technique using physical modelling of voice production, Folia Phoniatr. Logo. 58(1): 102–113, 2006.

References

- [7] Kafentzis, G., Stylianou, Y. and Alku, P.,International Conference on Acoustics, Speech, and Signal Processing, Prague, Czech Republic, May 22-27, 2011. p. 5408-5411, 2011.
- [8] Airaksinen, M., Raitio, T., Story, B., & Alku, P., Quasi Closed Phase Glottal Inverse Filtering Analysis With Weighted Linear Prediction. IEEE/ACM Transactions on Audio, Speech, and Language Processing, *22(3), 596-607, 2014.*
- [9] M. Plumpe, T. Quatieri, and D. Reynolds, Modeling of the glottal flow derivative waveform with application to speaker identification, IEEE Transactions on Speech and Audio Processing, 7:569-586, 1999.
- [10] Wong D., Markel J., Gray A., Least squares glottal inverse filtering from the acoustic speech waveform, IEEE Trans. Acoust. Speech Signal Process. 27: 350–355, 1979.

Other references

- [11] Childers D., Ahn C., Modeling the glottal volume-velocity waveform for three voice types, J. Acoust. Soc. Am. 97(1): 505–519, 1995.
- [12] Childers D., Lee C., Vocal quality factors: Analysis, synthesis, and perception. J. Acoust. Soc. Am. 90(5): 2394–2410, 1991.
- [13] Milenkovic P., Glottal inverse filtering by joint estimation of an AR system with a linear input model, IEEE Trans. Acoust. Speech Signal Process. 34(1): 28–42, 1986.
- [14] Raitio T., Suni A., Yamagishi Y., Pulakka H., Nurminen J., Vainio M., Alku P., HMM-based speech synthesis utilizing glottal inverse filtering, IEEE Trans. Audio Speech Lang. Process. 19(1): 153– 165, 2011.
- [15] Shiga Y., King S., Estimation of voice source and vocal tract characteristics based on multi-frame analysis. CD Proc. Eurospeech, 1749–1752, 2003.

Other references

- [16] J. Pohjalainen, R. Saeidi, T. Kinnunen, and P. Alku, Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions, in Proc. Interspeech, 2010, pp. 1477–1480.
- [17] T. Drugman, B. Bozkurt, and T. Dutoit, A comparative study of glottal source estimation techniques, Comp. Speech & Lang., vol. 26, no. 1, pp. 20–34, 2012.
- [18] Kitzing, P., & Löfqvist, A., Subglottal and oral air pressures during phonation—preliminary investigation using a miniature transducer system. Medical & Biological Engineering, 13(5), 644–648, 1975.