ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

Ψηφιακή Επεξεργασία Φωνής

Διάλεξη: Περί Αντίστροφου Φιλτραρίσματος Σήματος Φωνής

Παρουσίαση: Γιώργος Καφεντζής

Στυλιανού Ιωάννης

Τμήμα Επιστήμης Υπολογιστών

On the Inverse Filtering of Speech

George Kafentzis

CS578-November 2012

Outline

Human Speech Production System

The human speech production system is a complicated system. However, it can be roughly divided into three parts:

- The vocal folds, which is the source of the system.

Human Speech Production System

The human speech production system is a complicated system. However, it can be roughly divided into three parts:

- The vocal folds, which is the source of the system.
- The vocal tract, which is the path from the vocal folds to the lips.

Human Speech Production System

The human speech production system is a complicated system. However, it can be roughly divided into three parts:

- The vocal folds, which is the source of the system.
- The vocal tract, which is the path from the vocal folds to the lips.
- The lips, which is the final bound before speech output.

Source-Filter model

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

$$s(t) = g(t) \star v(t) \star l(t) \leftrightarrow S(z) = G(z)V(z)L(z)$$

where $\star$ denotes convolution and

Source-Filter model

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

$$s(t) = g(t) \star v(t) \star l(t) \leftrightarrow S(z) = G(z)V(z)L(z)$$

where $\star$ denotes convolution and

  - $g(t)$ the glottal airflow velocity waveform.

Source-Filter model

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

$$s(t) = g(t) \star v(t) \star l(t) \leftrightarrow S(z) = G(z)V(z)L(z)$$

where $\star$ denotes convolution and

  - $g(t)$ the glottal airflow velocity waveform.
  - $v(t)$ the vocal tract filter.

Source-Filter model

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

$$s(t) = g(t) \star v(t) \star l(t) \leftrightarrow S(z) = G(z)V(z)L(z)$$

where $\star$ denotes convolution and

  - $g(t)$ the glottal airflow velocity waveform.
  - $v(t)$ the vocal tract filter.
  - $l(t)$ the lip radiation filter.

Source-Filter model

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

$$s(t) = g(t) \star v(t) \star l(t) \leftrightarrow S(z) = G(z)V(z)L(z)$$

where $\star$ denotes convolution and

- $g(t)$ the glottal airflow velocity waveform.
- $v(t)$ the vocal tract filter.
- $l(t)$ the lip radiation filter.
- $s(t)$ the output speech waveform.

(Glottal) Inverse Filtering - IF

What is inverse filtering?

- Inverse Filtering is a technique for obtaining the source of voiced speech: the glottal airflow velocity waveform.
- How does the glottal flow look like?

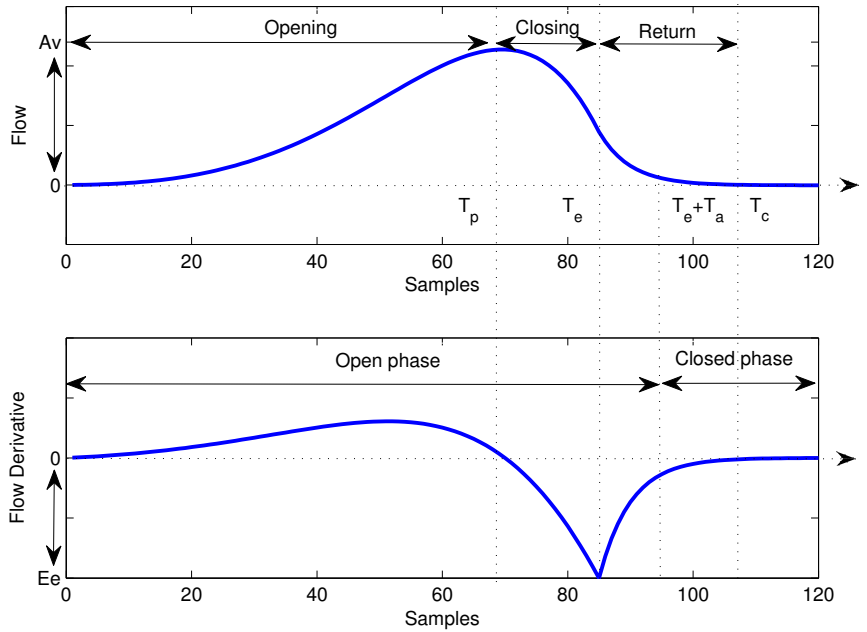


Figure: Phases of the glottal flow and its derivative.

(Glottal) Inverse Filtering - IF: Why is it important?

Inverse filtering is extensively used in:

- Basic research of speech production

(Glottal) Inverse Filtering - IF: Why is it important?

Inverse filtering is extensively used in:

- Basic research of speech production
- Applications to speech analysis, synthesis, and modification

(Glottal) Inverse Filtering - IF: Why is it important?

Inverse filtering is extensively used in:

- Basic research of speech production
- Applications to speech analysis, synthesis, and modification
- Increased interest is risen in:

(Glottal) Inverse Filtering - IF: Why is it important?

Inverse filtering is extensively used in:

- Basic research of speech production
- Applications to speech analysis, synthesis, and modification
- Increased interest is risen in:
  - Environmental voice care

(Glottal) Inverse Filtering - IF: Why is it important?

Inverse filtering is extensively used in:

- Basic research of speech production
- Applications to speech analysis, synthesis, and modification
- Increased interest is risen in:
  - Environmental voice care
  - Voice pathology detection

(Glottal) Inverse Filtering - IF: Why is it important?

Inverse filtering is extensively used in:

- Basic research of speech production
- Applications to speech analysis, synthesis, and modification
- Increased interest is risen in:
  - Environmental voice care
  - Voice pathology detection
  - Analysis of the emotional content of speech

(Glottal) Inverse Filtering - IF: How is it performed?

Basic idea:

- Form a computational model for the vocal tract filter, $\hat{V}(z)$.

(Glottal) Inverse Filtering - IF: How is it performed?

Basic idea:

- Form a computational model for the vocal tract filter, $\hat{V}(z)$.
- Cancel its effect from the speech waveform by filtering the speech signal through the inverse of the model.

(Glottal) Inverse Filtering - IF: How is it performed?

Basic idea:

- Form a computational model for the vocal tract filter, $\hat{V}(z)$.
- Cancel its effect from the speech waveform by filtering the speech signal through the inverse of the model.
- It is obvious that the heart of an IF system is the vocal tract filter estimation.

Evaluation of IF techniques

Problem: the actual glottal flow waveform is NOT available!

- ...at least in a non-invasive manner.

Evaluation of IF techniques

Problem: the actual glottal flow waveform is NOT available!

- ...at least in a non-invasive manner.
- Approaches:

Evaluation of IF techniques

Problem: the actual glottal flow waveform is NOT available!

- ...at least in a non-invasive manner.
- Approaches:
  - "Optical" inspection of the resulting glottal flow waveform,

Evaluation of IF techniques

Problem: the actual glottal flow waveform is NOT available!

- ...at least in a non-invasive manner.
- Approaches:
  - "Optical" inspection of the resulting glottal flow waveform,
  - Use of synthetic speech signals produced by a known artificial excitation,

Evaluation of IF techniques

Problem: the actual glottal flow waveform is NOT available!

- ...at least in a non-invasive manner.
- Approaches:
  - "Optical" inspection of the resulting glottal flow waveform,
  - Use of synthetic speech signals produced by a known artificial excitation,
  - Compare the results of different IF algorithms.

Evaluation of IF techniques

Problem: the actual glottal flow waveform is NOT available!

- ...at least in a non-invasive manner.
- Approaches:
  - "Optical" inspection of the resulting glottal flow waveform,
  - Use of synthetic speech signals produced by a known artificial excitation,
  - Compare the results of different IF algorithms.
- None of the previous approaches is truly objective.

Outline

Database of physically modeled speech

- A database of physically modeled speech signals was created by Titze and Story[3]

Database of physically modeled speech

- A database of physically modeled speech signals was created by Titze and Story[3]
- Time-varying waveforms of the glottal flow and radiated acoustic pressure are simulated in the basis of a physiological model of vocal folds and vocal tract.

Database of physically modeled speech

- A database of physically modeled speech signals was created by Titze and Story[3]
- Time-varying waveforms of the glottal flow and radiated acoustic pressure are simulated in the basis of a physiological model of vocal folds and vocal tract.
- This model generates a glottal flow waveform that is expected to provide a more firm and realistic test of IF methods than a parametric flow model.

Database of physically modeled speech

- A database of physically modeled speech signals was created by Titze and Story[3]
- Time-varying waveforms of the glottal flow and radiated acoustic pressure are simulated in the basis of a physiological model of vocal folds and vocal tract.
- This model generates a glottal flow waveform that is expected to provide a more firm and realistic test of IF methods than a parametric flow model.
- In this case, both the speech pressure waveform and the glottal flow are available.

Database of physically modeled speech

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

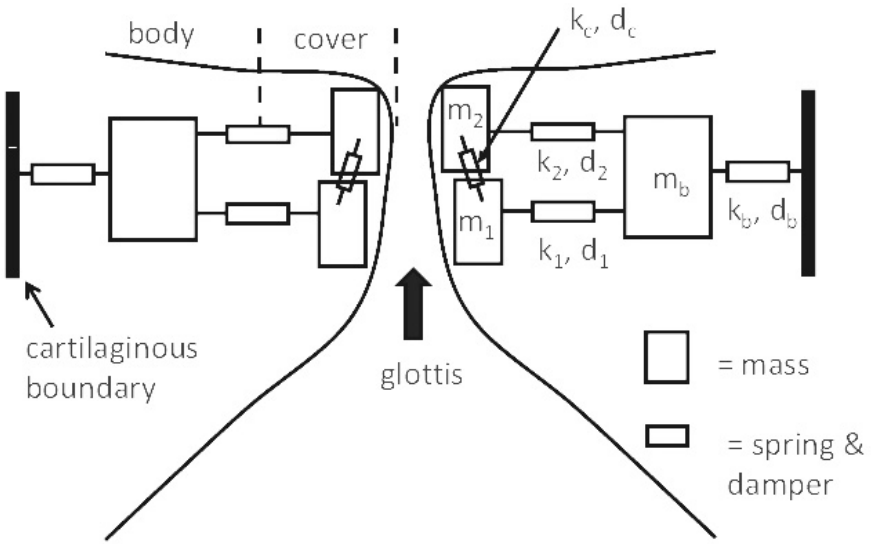


Figure: Schematic diagram of the lumped-element vocal fold model. The cover-body structure of each vocal fold is represented by three masses that are coupled to each other by spring and damping elements. Bilateral symmetry was assumed for all simulations.

Database of physically modeled speech

- The input parameters of this model are:

Database of physically modeled speech

- The input parameters of this model are:
  - the lung pressure,

Database of physically modeled speech

- The input parameters of this model are:
  - the lung pressure,
  - prephonatory glottal half-width (adduction),

Database of physically modeled speech

- The input parameters of this model are:
  - the lung pressure,
  - prephonatory glottal half-width (adduction),
  - vocal fold length and thickness,

Database of physically modeled speech

- The input parameters of this model are:
  - the lung pressure,
  - prephonatory glottal half-width (adduction),
  - vocal fold length and thickness,
  - and normalized activation levels of the cricothyroid (CT) and thyroarytenoid (TA) muscles.

Database of physically modeled speech

- The input parameters of this model are:
  - the lung pressure,
  - prephonatory glottal half-width (adduction),
  - vocal fold length and thickness,
  - and normalized activation levels of the cricothyroid (CT) and thyroarytenoid (TA) muscles.
- These parameters were transformed into mechanical parameters according to Titze and Story[3].

Database of physically modeled speech

- Four different sustained vowels (/aa/, /ae/, /eh/, /ih/) with eight different fundamental frequencies ($105, 115, 130, 145, 205, 210, 230$, and 255 Hz) were used in this work.

Database of physically modeled speech

- Four different sustained vowels (/aa/, /ae/, /eh/, /ih/) with eight different fundamental frequencies ($105, 115, 130, 145, 205, 210, 230$, and 255 Hz) were used in this work.
- In summary, the model is a simplified but physically motivated representation of a speaker.

Database of physically modeled speech

- Four different sustained vowels (/aa/, /ae/, /eh/, /ih/) with eight different fundamental frequencies (105, 115, 130, 145, 205, 210, 230, and 255 Hz) were used in this work.
- In summary, the model is a simplified but physically motivated representation of a speaker.
- It generates both the signal on which inverse filtering is typically performed (microphone signal) and the signal that is seeked to be determined (glottal flow).

Database of physically modeled speech

- Four different sustained vowels (/aa/, /ae/, /eh/, /ih/) with eight different fundamental frequencies (105, 115, 130, 145, 205, 210, 230, and 255 Hz) were used in this work.
- In summary, the model is a simplified but physically motivated representation of a speaker.
- It generates both the signal on which inverse filtering is typically performed (microphone signal) and the signal that is seeked to be determined (glottal flow).
- This provides an idealized test case for inverse filtering algorithms.

Example

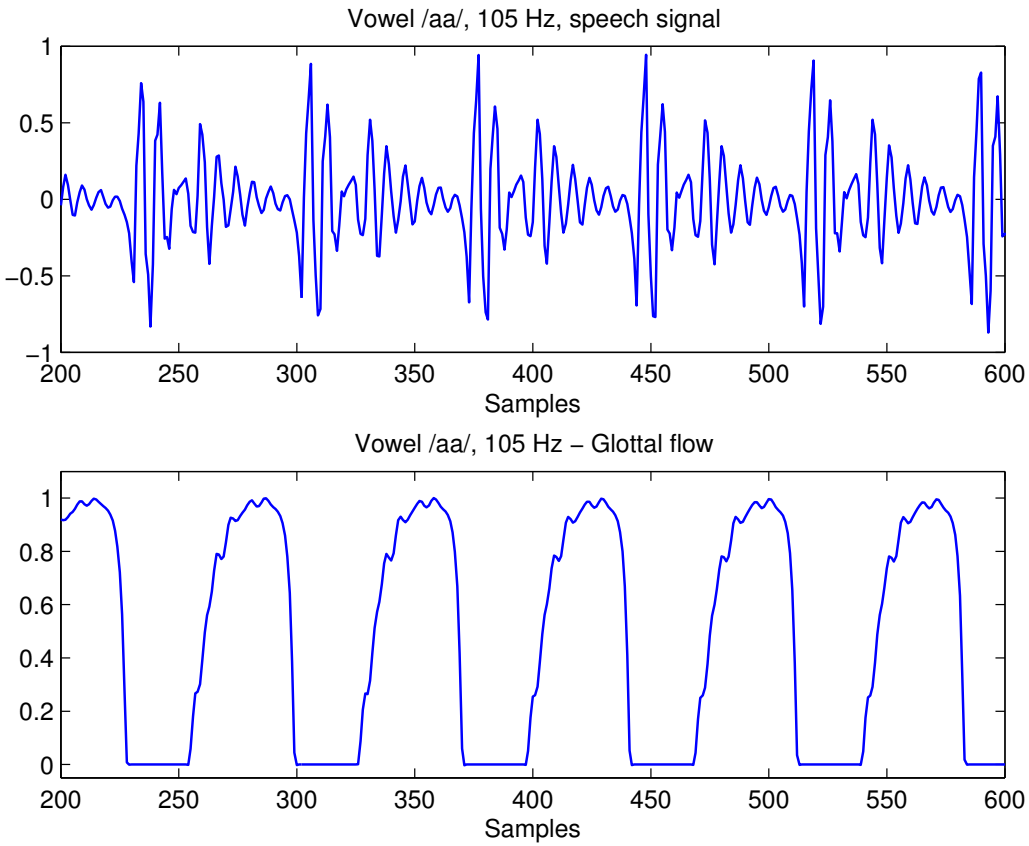


Figure: Upper panel: Speech waveform.
Lower panel: Glottal flow waveform.

Outline

Vocal tract estimation techniques

A number of techniques have been developed to robustly estimate the vocal tract filter.

- Most of them are based on *Linear Prediction* (LP).
- LP is used to produce an all-pole model of the system filter, $H(z)$, which turns out to be a model of the vocal tract and its resonances or formants.

$$V(z) = \frac{1}{\sum_{k=1}^{p}(\alpha_k z^{-k})},$$

- In general, LP minimizes the mean squared error over a region R

$$E = \sum_{R} e^2[n], \text{ where } e[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k],$$

where p is the prediction order and a_k are the prediction coefficients.

- The selection of this region leads to different approaches.

Linear Prediction variants

- In this work, we will discuss the performance of four vocal tract estimation techniques, all based on LP:

Linear Prediction variants

- In this work, we will discuss the performance of four vocal tract estimation techniques, all based on LP:
  - Autocorrelation-based Linear Prediction,

Linear Prediction variants

- In this work, we will discuss the performance of four vocal tract estimation techniques, all based on LP:
  - Autocorrelation-based Linear Prediction,
  - Closed Phase Covariance-based Linear Prediction,

Linear Prediction variants

- In this work, we will discuss the performance of four vocal tract estimation techniques, all based on LP:
  - Autocorrelation-based Linear Prediction,
  - Closed Phase Covariance-based Linear Prediction,
  - Stabilised Weighted Linear Prediction,

Linear Prediction variants

- In this work, we will discuss the performance of four vocal tract estimation techniques, all based on LP:
  - Autocorrelation-based Linear Prediction,
  - Closed Phase Covariance-based Linear Prediction,
  - Stabilised Weighted Linear Prediction,
  - Closed Phase (CP) Covariance-based Linear Prediction with Mathematical Constraints.

Linear Prediction - Autocorrelation method

- Assuming that the speech signal $s[n]$ is zero outside an interval $0 \leq n \leq N-1$
- the total error minimization leads to the matrix equation

$$\mathbf{\Phi}\vec{a} = \vec{r}, \tag{1}$$

- where the matrix Φ is called the *autocorrelation matrix* and its elements are given by $\Phi_{i,j} = R(|i-j|) = \sum_{n=i}^{N-1} s[n]s[n-|i-j|], \quad 0 \leq i \leq p$, and the other two vectors are given by

$$\vec{a} = [a_1, a_2, a_3, ..., a_p]^T \quad , \vec{r} = [R(1), R(2), R(3), ..., R(p)]^T. \tag{2}$$

- A tapered window (e.g. Hanning) is often used to eliminate beggining and end effects.

Linear Prediction - CP Covariance method

- Assuming that the speech signal $s[n]$ is zero outside an interval $-p \leq n \leq N-1$
- the total error minimization leads to the matrix equation

$$\mathbf{\Phi}\vec{a} = \vec{\psi}, \tag{3}$$

- where the matrix Φ has the properties of a *covariance matrix* and its elements are given by

$$\phi_{i,j} = \sum_{n=0}^{N-1} s[n-i]s[n-j], \tag{4}$$

where $1 \leq i, j \leq p$, and the other two vectors are given by

$$\vec{a} = [a_1, a_2, a_3, ..., a_p]^T \quad , \vec{\psi} = [\phi_{0,1}, \phi_{0,2}, \phi_{0,3}, ..., \phi_{0,p}]^T. \tag{5}$$

- However, in CP analysis, the total error is minimized over a region where the glottis is closed.

Stabilised Weighted Linear Prediction

Stabilized Weighted Linear Prediction (SWLP)[1], is an all-pole modeling method based on Weighted Linear Prediction (WLP).

- SWLP uses time domain weighting of the square of the prediction error signal.
- Short Time Energy (STE) Weighting function: $w_n = \sum_{i=0}^{M-1} x^2[n-i-1]$, where $x[n]$ is the signal and M is the duration of the STE window.
- The prediction error energy E in the SWLP method is

$$E = \sum_{n=1}^{N+p} (e_n)^2 w_n = \mathbf{a}^T \Big(\sum_{n=1}^{N+p} w_n \mathbf{x}[n] \mathbf{x}^T[n] \Big) \mathbf{a} = \mathbf{a}^T \mathbf{R} \mathbf{a}, \tag{6}$$

where w_n is the weight imposed on sample n, N is the length of the signal $x[n]$, and

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

Stabilised Weighted Linear Prediction

- Constrained minimization problem:

$$\text{minimize } E \text{ subject to } \mathbf{a}^T\mathbf{u} = 1,$$

where $\mathbf{u}$ is the vector defined as $\mathbf{u} = [1\,0\,\ldots\,0]^T$.

Stabilised Weighted Linear Prediction

- Constrained minimization problem:

$$\text{minimize } E \text{ subject to } \mathbf{a}^T\mathbf{u} = 1,$$

where $\mathbf{u}$ is the vector defined as $\mathbf{u} = [1\,0\,\ldots\,0]^T$.

- It can be shown that $\mathbf{a}$ satisfies the linear equation

$$\mathbf{R}\mathbf{a} = \sigma^2\mathbf{u}, \tag{7}$$

where $\sigma^2 = \mathbf{a}^T\mathbf{R}\mathbf{a}$ is the error energy.

Stabilised Weighted Linear Prediction

- Constrained minimization problem:

$$\text{minimize } E \text{ subject to } \mathbf{a}^T\mathbf{u} = 1,$$

  where $\mathbf{u}$ is the vector defined as $\mathbf{u} = [1\,0\,\ldots\,0]^T$.
- It can be shown that $\mathbf{a}$ satisfies the linear equation

$$\mathbf{R}\mathbf{a} = \sigma^2\mathbf{u}, \tag{7}$$

  where $\sigma^2 = \mathbf{a}^T\mathbf{R}\mathbf{a}$ is the error energy.
- Finally, the SWLP all-pole model is obtained as $H(z) = 1/A(z)$, where $A(z)$ is the z-transform of vector $\mathbf{a}$.

Stabilised Weighted Linear Prediction

- STE function emphasizes the speech samples of large amplitude, which typically occur during the closed phase interval.

Stabilised Weighted Linear Prediction

- STE function emphasizes the speech samples of large amplitude, which typically occur during the closed phase interval.
- It is well-known that applying LP analysis on speech samples that belong to the glottal closed phase interval will generally result in a more robust spectral representation of the vocal tract.

Stabilised Weighted Linear Prediction

- STE function emphasizes the speech samples of large amplitude, which typically occur during the closed phase interval.
- It is well-known that applying LP analysis on speech samples that belong to the glottal closed phase interval will generally result in a more robust spectral representation of the vocal tract.
- By emphasizing on these samples that occur during the glottal closed phase, it is likely to yield more robust acoustical cues for the formants.

Stabilised Weighted Linear Prediction

- STE function emphasizes the speech samples of large amplitude, which typically occur during the closed phase interval.
- It is well-known that applying LP analysis on speech samples that belong to the glottal closed phase interval will generally result in a more robust spectral representation of the vocal tract.
- By emphasizing on these samples that occur during the glottal closed phase, it is likely to yield more robust acoustical cues for the formants.
- 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.

Stabilised Weighted Linear Prediction

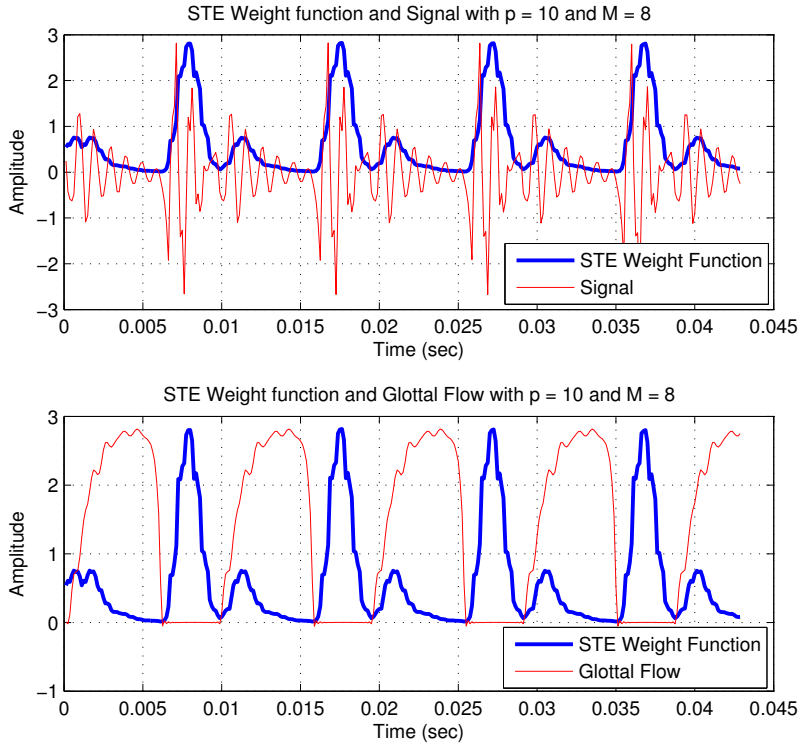


Figure: Upper panel: time domain waveforms of speech (vowel /a/ produced by male speaker) and short-time energy (STE) weight function (M=8).
Lower panel: Glottal flow waveform of the vowel /a/ together with the STE weight function (M=8).

Stabilised Weighted Linear Prediction

- Stability is ensured by the following formula:
- $$\mathbf{R} = \mathbf{Y}^T \mathbf{Y}, \tag{8}$$

  where

  $$\mathbf{Y} = [\mathbf{y_0}\ \mathbf{y_1} \cdots \mathbf{y_p}] \in \Re^{(N+p)x(p+1)}$$

  and

  $$\mathbf{y}_0 = [\sqrt{w_1}x[1] \cdots \sqrt{w_N}x[N]\ \ 0 \cdots 0]^T.$$

- The column vectors are given by

  $$\mathbf{y}_{k+1} = \mathbf{B}\mathbf{y}_k,\ \ k = 0, 1, \cdots, p-1, \tag{9}$$

  where

  $$\mathbf{B} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ \sqrt{w_2/w_1} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{w_3/w_2} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \sqrt{w_{N+p}/w_{N+p-1}} & 0 \end{bmatrix}$$

Stabilised Weighted Linear Prediction

- Before forming the matrix $\mathbf{Y}$, the elements of the secondary diagonal of the matrix $\mathbf{B}$ are defined for all $i = 1, \cdots, N + p - 1$ as

$$\mathbf{B}_{i+1,i} = \begin{cases} \sqrt{w_{i+1}/w_i}, & \text{if } w_i \leq w_{i+1} \\ 1, & \text{if } w_i > w_{i+1} \end{cases}$$

Stabilised Weighted Linear Prediction

- Before forming the matrix $\mathbf{Y}$, the elements of the secondary diagonal of the matrix $\mathbf{B}$ are defined for all $i = 1, \cdots, N + p - 1$ as

$$\mathbf{B}_{i+1,i} = \begin{cases} \sqrt{w_{i+1}/w_i}, & \text{if } w_i \leq w_{i+1} \\ 1, & \text{if } w_i > w_{i+1} \end{cases}$$

- This method of computing matrix R is called the *Stabilized Weighted Linear Prediction* model, and the stability of the all-pole filter is ensured.

Sources of Distortion in conventional CP Covariance LP

- Conventional CP Covariance LP suffers from certain shortcomings.

Sources of Distortion in conventional CP Covariance LP

- Conventional CP Covariance LP suffers from certain shortcomings.
  - Short CP length (especially for high pitched speakers).

Sources of Distortion in conventional CP Covariance LP

- Conventional CP Covariance LP suffers from certain shortcomings.
  - Short CP length (especially for high pitched speakers).
  - Sensitivity to the position of the covariance frame.

Sources of Distortion in conventional CP Covariance LP

- Conventional CP Covariance LP suffers from certain shortcomings.
  - Short CP length (especially for high pitched speakers).
  - Sensitivity to the position of the covariance frame.
  - Vocal tract filter may not be stable.

Sources of Distortion in conventional CP Covariance LP

Sensitivity of the covariance frame position:

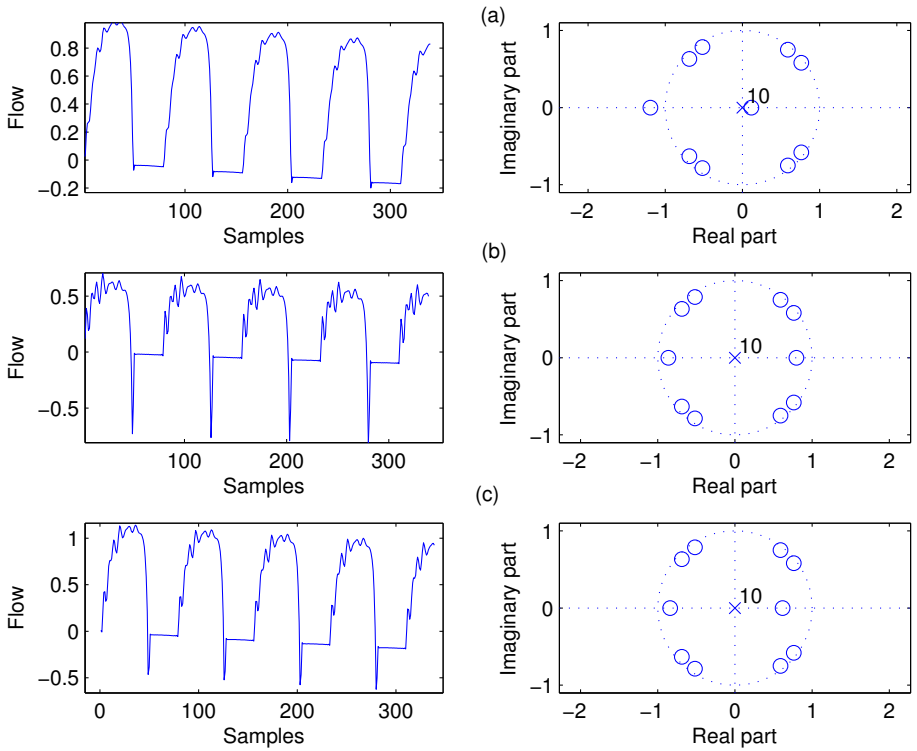


Figure: Covariance Frame Misalignment and Glottal Flow Distortion.

Sources of Distortion in conventional CP Covariance LP

- 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.

Sources of Distortion in conventional CP Covariance LP

- 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.

Sources of Distortion in conventional CP Covariance LP

- 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.

Sources of Distortion in conventional CP Covariance LP

- 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 from Fant[5].

Sources of Distortion in conventional CP Covariance LP

- 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 from Fant[5].
- The removal of such roots results in less dependency on the covariance frame location.

Sources of Distortion in conventional CP Covariance LP

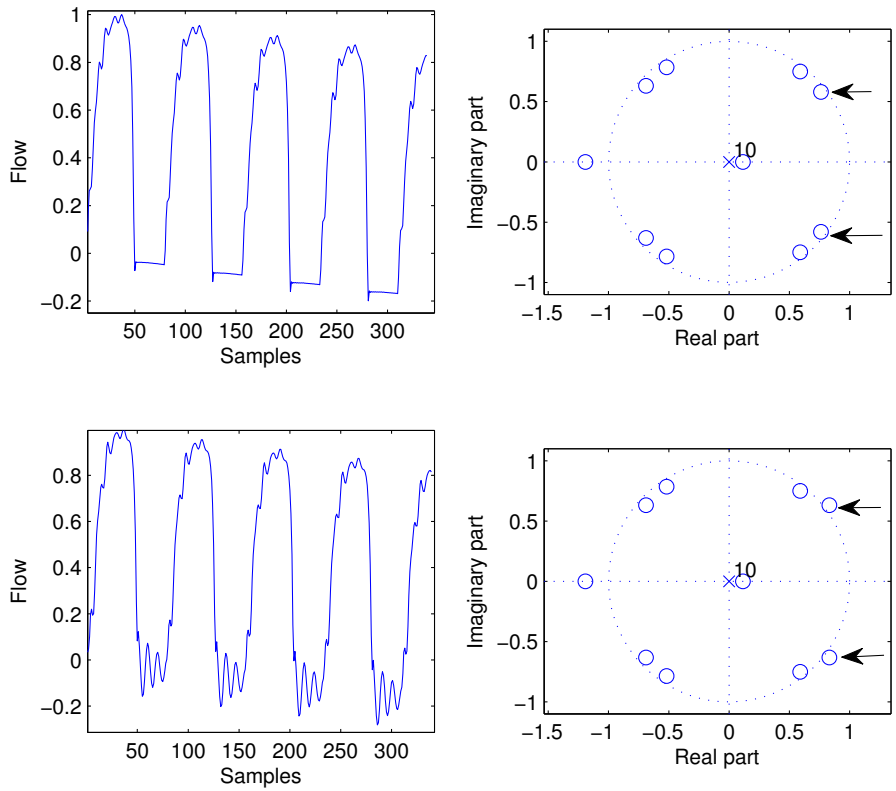


Figure: Distortion caused by non-minimum phase filter.

Sources of Distortion in conventional CP Covariance LP

- The inverse filter $1/V(z)$ might not be minimum phase.

Sources of Distortion in conventional CP Covariance LP

- The inverse filter $1/V(z)$ might not be minimum phase.
- It can become minimum phase by replacing each zero by its mirror image partner.

Sources of Distortion in conventional CP Covariance LP

- The inverse filter $1/V(z)$ might not be minimum phase.
- It can become minimum phase by replacing each zero by its mirror image partner.
- That leaves the amplitude spectrum unchanged.

Sources of Distortion in conventional CP Covariance LP

- The inverse filter $1/V(z)$ might not be minimum phase.
- It can become minimum phase by replacing each zero by its mirror image partner.
- That leaves the amplitude spectrum unchanged.
- The phase characteristics change, though.

Constrained CP Covariance Linear Prediction

Concept:

- Modification of the conventional CP covariance analysis in order to provide more realistic root locations, in the acoustic sense.

Constrained CP Covariance Linear Prediction

Concept:

- Modification of the conventional CP covariance analysis in order to provide more realistic root locations, in the acoustic sense.
- How?

Constrained CP Covariance Linear Prediction

Concept:

- 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.

Constrained CP Covariance Linear Prediction

Concept:

- 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.

Constrained CP Covariance Linear Prediction

Concept:

- 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.
  - Two suggestions: DC-constraint and/or π-constraint.

DC-Constrained CP Covariance Linear Prediction

- DC-constraint:

$$V(e^{j0}) = \sum_{k=0}^{p} \alpha_k e^{-j0n} = \sum_{k=0}^{p} \alpha_k = I_{DC}. \qquad (10)$$

DC-Constrained CP Covariance Linear Prediction

- DC-constraint:

$$V(e^{j0}) = \sum_{k=0}^{p} \alpha_k e^{-j0n} = \sum_{k=0}^{p} \alpha_k = I_{DC}. \qquad (10)$$

- Why DC-constraint?

DC-Constrained CP Covariance Linear Prediction

- DC-constraint:

$$V(e^{j0}) = \sum_{k=0}^{p} \alpha_k e^{-j0n} = \sum_{k=0}^{p} \alpha_k = I_{DC}. \qquad (10)$$

- Why DC-constraint?
  - Amplitude response of voiced sounds approaches unity at zero frequency[5]

DC-Constrained CP Covariance Linear Prediction

- DC-constraint:

$$V(e^{j0}) = \sum_{k=0}^{p} \alpha_k e^{-j0n} = \sum_{k=0}^{p} \alpha_k = I_{DC}. \qquad (10)$$

- Why DC-constraint?
  - Amplitude response of voiced sounds approaches unity at zero frequency[5]
  - A short and misplaced covariance frame might lead to a response with higher gain at DC than at formants.

DC-Constrained CP Covariance Linear Prediction

- DC-constraint:

$$V(e^{j0}) = \sum_{k=0}^{p} \alpha_k e^{-j0n} = \sum_{k=0}^{p} \alpha_k = I_{DC}. \qquad (10)$$

- Why DC-constraint?
  - Amplitude response of voiced sounds approaches unity at zero frequency[5]
  - 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 amplitude response to Fant's source-filter theory.

DC-Constrained CP Covariance Linear Prediction

Formulation:

- minimize $\mathbf{a}^T \mathbf{\Phi} \mathbf{a}$ subject to $\mathbf{\Gamma}^T \mathbf{a} = \mathbf{b}$,
- where $\mathbf{a} = [a_0, \cdots, a_p]^T$ is the filter coefficient vector with $a_0 = 1$, $\mathbf{\Phi}$ is the covariance matrix, $\mathbf{b} = [1,\ l_{DC}]^T$, and $\mathbf{\Gamma}$ is a $(p+1)$ by 2 constraint matrix defined as

$$\mathbf{\Gamma} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \tag{11}$$

- Using Lagrange multipliers (convex problem), we have the filter coefficients:
-

$$\mathbf{a} = \mathbf{\Phi}^{-1}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Phi}^{-1}\mathbf{\Gamma})^{-1}\mathbf{b} \tag{12}$$

DC-π-Constrained CP Covariance Linear Prediction

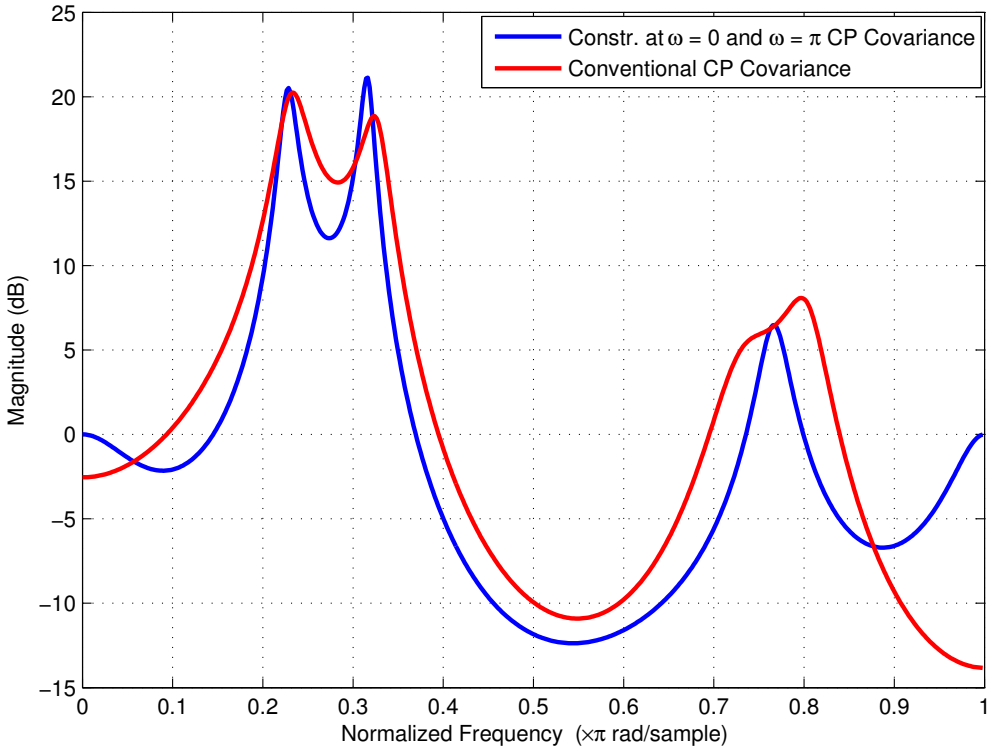
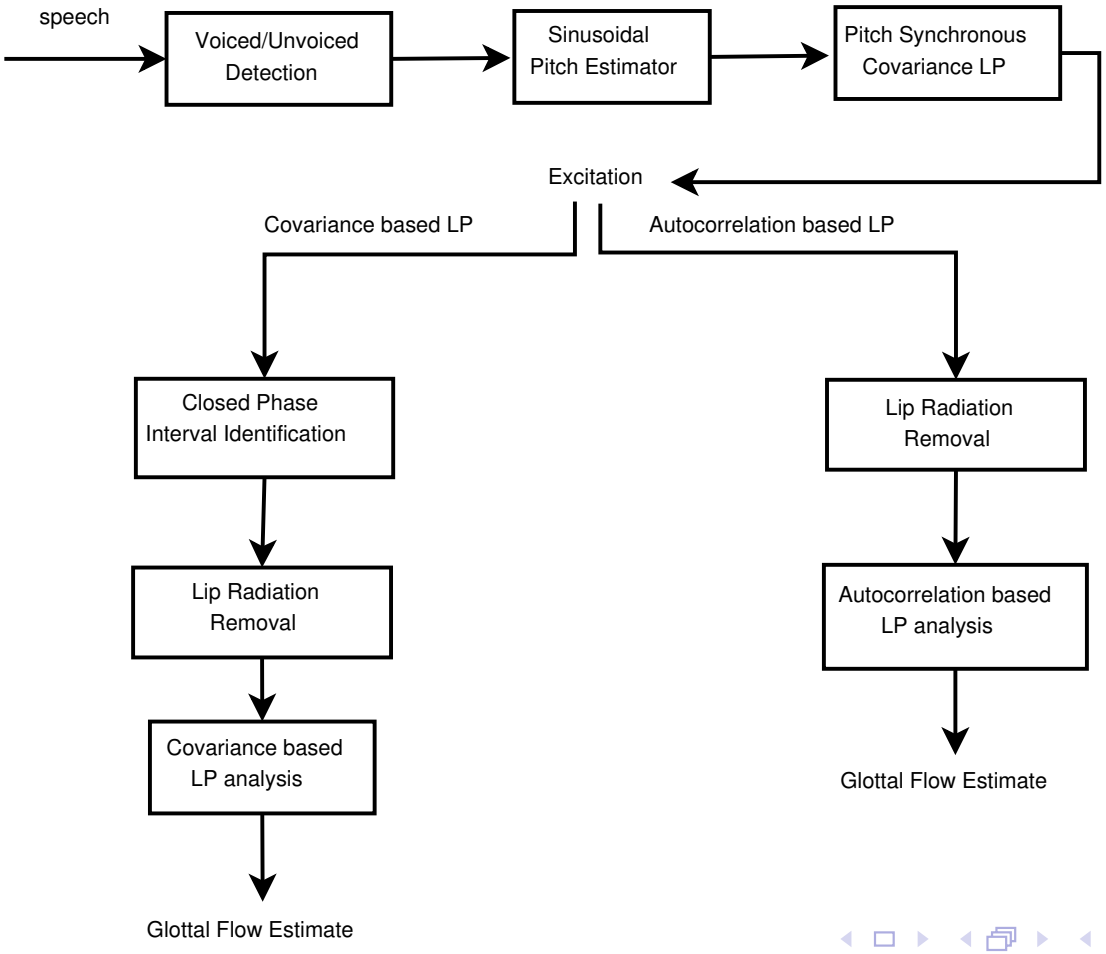


Figure: Examples of all-pole spectra computed in the closed phase covariance analysis by the conventional LP and by the DC-π constrained LP.

Outline

Inverse Filtering Procedure - Flowchart



Inverse Filtering Procedure - Details

- Sampling Frequency: $f_s = 8$kHz.
- Order of LP Analysis: $p = 10$.
- The lip radiation effect was canceled by a first order all-pole filter with its pole at $z = 0.999$.
- Autocorrelation approaches:
  - Analysis window: hanning type, 250 ms duration.
  - M-parameter for SWLP: $M = 8$, $M = 24$.
- Covariance approaches:
  - Analysis window: rectangular type,
  - Duration determined by the detected closed phase interval[6].
  - Analysis frame rate: one pitch period
- Inverse filtered speech signals were computed in a frame by frame basis.
- Two local pitch periods and a frame rate of one pitch period was applied.
- The overall glottal flow was synthesized using the Overlap-Add (OLA) method.

Outline

Metrics for evaluating the IF techniques

- Signal to Reconstruction Error Ratio - SRER
  - SRER is a standard index for measuring the effectiveness of modeling a waveform and is defined as:

$$SRER = 20log10\Big(\frac{\sigma_{s[n]}}{\sigma_{e[n]}}\Big) \tag{13}$$

    where $s[n]$ is the original (or true) glottal flow signal in our case, $e[n]$ is the modeling (or reconstruction) error, $e[n] = s[n] - \hat{s}[n]$, and σ denotes the corresponding standard deviation.
  - SRER was computed from the overall glottal flow waveforms.
- Difference between the first two harmonics - H1-H2
  - H1-H2 is an index of the spectral decay (or spectral tilt) of the glottal spectrum.
  -

$$ER_{H1H2} = \Big|Ref_{H1H2} - Est_{H1H2}\Big| \tag{14}$$

    where Ref_{H1H2} and Est_{H1H2} denote the H1-H2 metric for the true (or reference) and the estimated glottal airflow, respectively.
  - For a good estimation ER_{H1H2} should be close to zero.

Examples

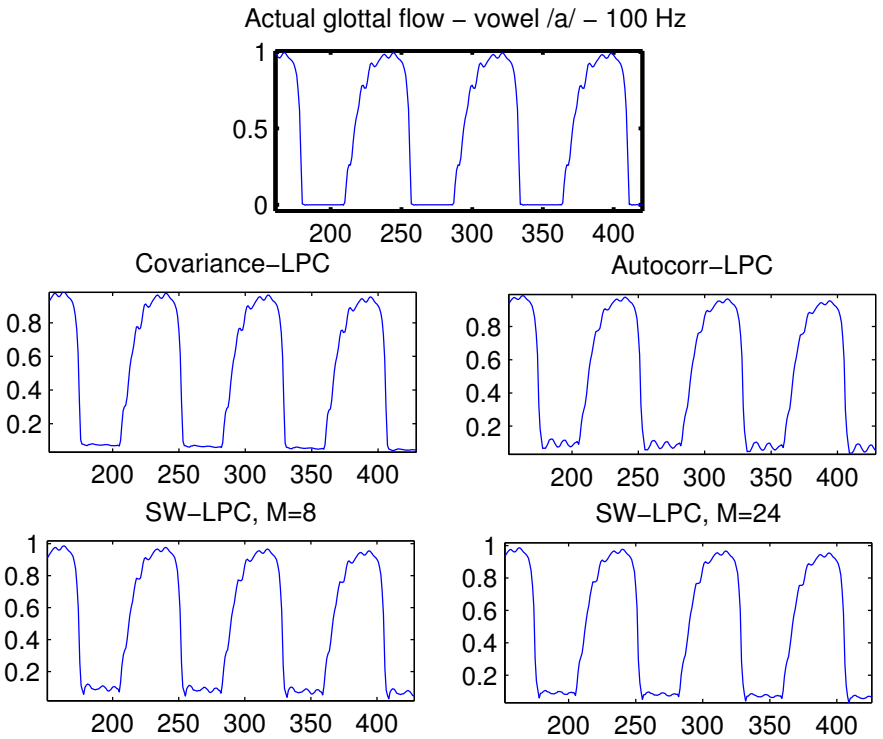


Figure: Glottal flow estimates for vowel /aa/ of $f_0 = 105$ Hz. Upper panel: Original glottal flow. Middle panel: Covariance (left) and Autocorrelation (right) based glottal flow estimates. Lower panel: SWLP with $M = 8$ and $M = 24$ glottal flow estimates. In all panels, time is indicated in samples.

Examples

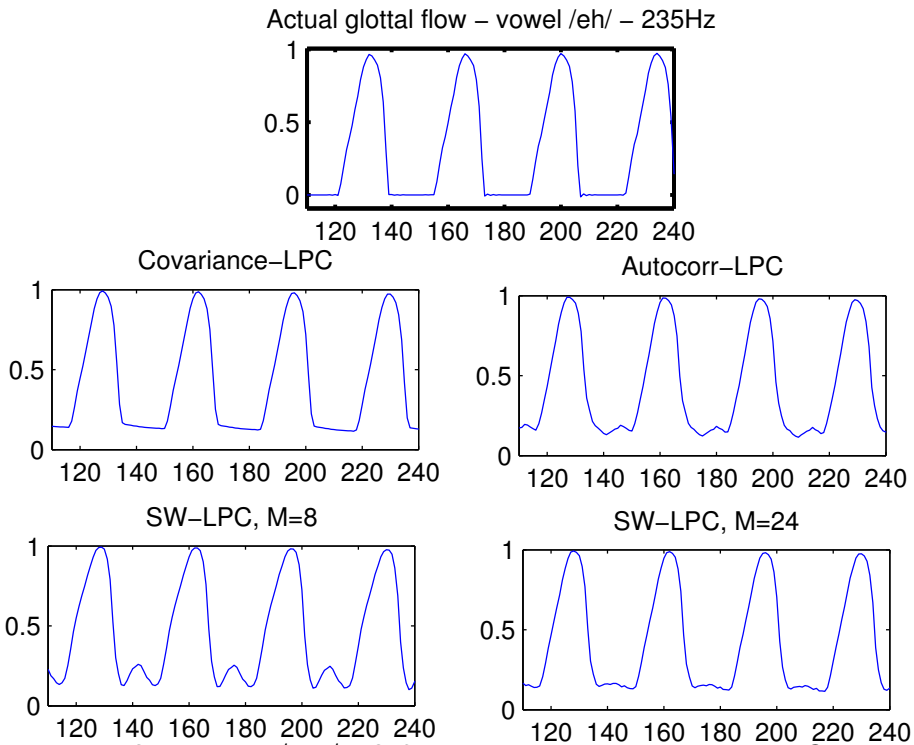


Figure: Glottal flow estimates for vowel /eh/ of $f_0 = 230$ Hz. Upper panel: Original glottal flow. Middle panel: Covariance (left) and Autocorrelation (right) based glottal flow estimates. Lower panel: SWLP with $M = 8$ and $M = 24$ glottal flow estimates. In all panels, time is indicated in samples.

Results - SRER

SRER					
Vowel	$SWLP_8$	$SWLP_{24}$	LPC	CovLPC	CLPC
/aa/	33.5 ($\pm$2.0)	39.7 ($\pm$4.5)	36.2 ($\pm$5.7)	41.9 ($\pm$6.3)	41.5 ($\pm$6.6)
/ae/	32.7 ($\pm$4.4)	35.2 ($\pm$2.9)	37.8 ($\pm$3.0)	40.4 ($\pm$6.4)	40.9 ($\pm$6.9)
/eh/	34.0 ($\pm$1.9)	38.4 ($\pm$4.2)	33.9 ($\pm$4.0)	40.5 ($\pm$5.2)	41.1 ($\pm$7.4)
/ih/	32.3 ($\pm$1.5)	37.6 ($\pm$3.1)	35.3 ($\pm$4.6)	39.2 ($\pm$5.6)	40.8 ($\pm$8.7)

Table: Mean and standard deviation of the SRER value for each vowel (all 8 frequencies) and method is illustrated. LPC stands for autocorrelation LP, CovLPC for Closed Phase covariance LP, CLPC for Constrained Closed Phase covariance LP, $SWLP_8$ for SWLP with $M = 8$, and $SWLP_{24}$ for SWLP with $M = 24$

Results - H1-H2

ER_{H1H2}					
Vowel	$SWLP_8$	$SWLP_{24}$	LPC	CovLPC	CLPC
/aa/	0.68 ($\pm$0.10)	0.23 ($\pm$0.09)	0.75 ($\pm$0.09)	0.20 ($\pm$0.20)	0.08($\pm$0.28)
/ae/	0.15 ($\pm$0.12)	0.15 ($\pm$0.05)	0.55 ($\pm$0.05)	0.18 ($\pm$0.13)	0.07($\pm$0.30)
/eh/	0.34 ($\pm$0.09)	0.30 ($\pm$0.07)	0.54 ($\pm$0.08)	0.38 ($\pm$0.17)	0.16($\pm$0.30)
/ih/	0.72 ($\pm$0.14)	0.39 ($\pm$0.11)	0.85 ($\pm$0.12)	0.35 ($\pm$0.24)	0.34($\pm$0.50)

Table: Mean and the standard deviation of ER_{H1H2} for each vowel (all 8 frequencies) and each method is illustrated. LPC stands for autocorrelation LP, CovLPC for CP covariance LP, CLPC for Constrained Closed Phase covariance LP, $SWLP_8$ for SWLP with $M = 8$, and $SWLP_{24}$ for SWLP with $M = 24$

Outline

Conclusions

- For both metrics, it is obvious that $SWLP_{24}$ outperforms conventional autocorrelation LP.

Conclusions

- For both metrics, it is obvious that $SWLP_{24}$ outperforms conventional autocorrelation LP.
- As expected, covariance methods prevails for both metrics.

Future Work

- Furtherly investigate the role of M parameter of SWLP.

Future Work

- Furtherly investigate the role of M parameter of SWLP.
- Evaluation of other IF methods or closed phase detection algorithms on the discussed database.

Future Work

- Furtherly investigate the role of M parameter of SWLP.
- Evaluation of other IF methods or closed phase detection algorithms on the discussed database.
- Apply IF in databases produced by a more sophisticated model of human production system.

Future Work

- Furtherly investigate the role of M parameter of SWLP.
- Evaluation of other IF methods or closed phase detection algorithms on the discussed database.
- Apply IF in databases produced by a more sophisticated model of human production system.
- Real time system approach to IF.

References

Carlo Magi, Jouni Pohjalainen, Tom Backstrom, Paavo Alku

Stabilised Weighted Linear Prediction
Speech Communication, 51:401-411, 2009.

Paavo Alku, Carlo Magi, Santeri Yrttiaho, Tom Backstrom, Brad Story

Closed phase covariance analysis on constrained linear prediction for glottal inverse filtering
Journal of Acoustical Society of America, 125(5):3289-3305, 2009.

Ingo Titze and Brad Story

Rules for controlling low-dimensional vocal fold models with muscle activites
Journal of Acoustical Society of America, 112:1064-1076, 2002.

Ingo Titze

Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model
Journal of Acoustical Society of America, 111:367-376, 2002.

Gunnar Fant

Acoustic Theory of Speech Production
(Mouton, The Hague).

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.

Acknowledgments (say thanks to everyone!)

Time for Questions!

Any questions? (Hope not...)

Τέλος Ενότητας

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Κρήτης**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Σημειώματα

Σημείωμα αδειοδότησης

- Το παρόν υλικό διατίθεται με τους όρους της άδειας χρήσης Creative Commons Αναφορά, Μη Εμπορική Χρήση, Όχι Παράγωγο Έργο 4.0 [1] ή μεταγενέστερη, Διεθνής Έκδοση. Εξαιρούνται τα αυτοτελή έργα τρίτων π.χ. φωτογραφίες, διαγράμματα κ.λ.π., τα οποία εμπεριέχονται σε αυτό και τα οποία αναφέρονται μαζί με τους όρους χρήσης τους στο «Σημείωμα Χρήσης Έργων Τρίτων».

[1] http://creativecommons.org/licenses/by-nc-nd/4.0/

- Ως **Μη Εμπορική** ορίζεται η χρήση:
  - που δεν περιλαμβάνει άμεσο ή έμμεσο οικονομικό όφελος από την χρήση του έργου, για το διανομέα του έργου και αδειοδόχο
  - που δεν περιλαμβάνει οικονομική συναλλαγή ως προϋπόθεση για τη χρήση ή πρόσβαση στο έργο
  - που δεν προσπορίζει στο διανομέα του έργου και αδειοδόχο έμμεσο οικονομικό όφελος (π.χ. διαφημίσεις) από την προβολή του έργου σε διαδικτυακό τόπο

- Ο δικαιούχος μπορεί να παρέχει στον αδειοδόχο ξεχωριστή άδεια να χρησιμοποιεί το έργο για εμπορική χρήση, εφόσον αυτό του ζητηθεί.

.

Σημείωμα Αναφοράς