TOSHIBA
Leading Innovation >>>

Iterative Estimation of Phase Using Complex Cepstrum Representation

Ranniery Maia Yannis Stylianou

Toshiba Research Europe Limited

Cambridge Research Laboratory, Cambridge, UK

ICASSP 2016, Shanghai, China

March 22nd, 2016

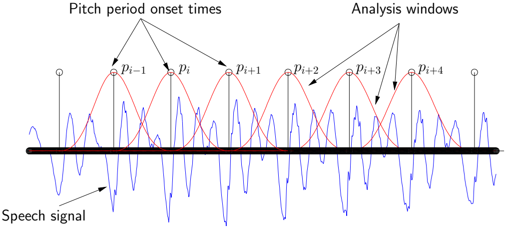
Introduction

- Goal: accurate estimation of phase from speech
- Importance of phase
  1. Speech parameterization for TTS
  2. Features for ASR
  3. Detection of speech pathologies
- Estimation of continuous phase spectrum
  1. Detection of glottal closure instants (GCI)
  2. Phase unwrapping
- Minimum MSE-based complex cepstrum analysis [Maia et al., 2013a]
  - Complex cepstrum analysis with no phase unwrapping
  - GCI are iteratively optimized in the process
- Phase estimation can be performed using the same concept!

Typical phase estimation issues

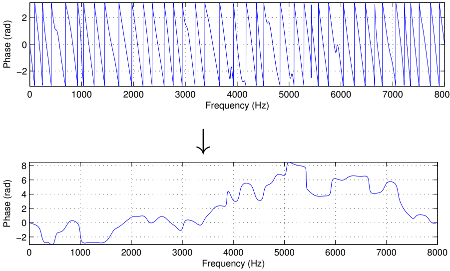
Speech segmentation

- Detection of glottal closure instants (GCIs)
- Influence of the shape and length of the windows



Phase unwrapping

- Discrete Fourier transform (DFT) gives phase modulo 2π
- Phase must be *unwrapped*



Complex cepstrum-based speech analysis and synthesis

Analysis

1. Determine pitch period onset times (or GCI): $\{p_0, \ldots, p_{Z-1}\}$
2. Complex cepstrum analysis

$$\hat{h}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left|S\left(e^{\jmath\omega}\right)\right| e^{\jmath\omega n} d\omega + \frac{\jmath}{2\pi}\int_{-\pi}^{\pi} \theta\left(\omega\right) e^{\jmath\omega n} d\omega$$

$\hat{h}(n)$ ↓ *Cepstrum at* p_z; $\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left|S\left(e^{\jmath\omega}\right)\right| e^{\jmath\omega n} d\omega$ ↓ *Amplitude response at* p_z; $\frac{\jmath}{2\pi}\int_{-\pi}^{\pi} \theta\left(\omega\right) e^{\jmath\omega n} d\omega$ ↓ *Phase response at* p_z

Synthesis

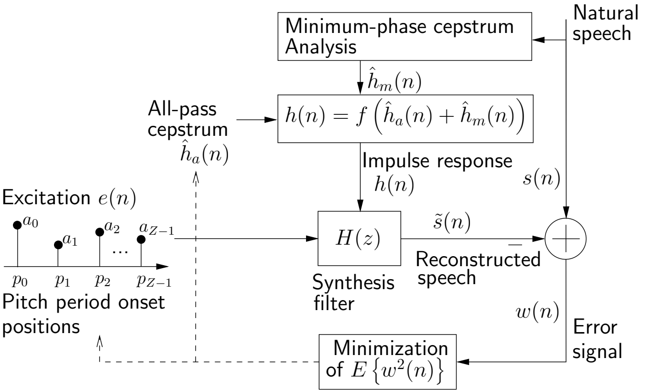
1. Derive non-causal impulse responses

$$h(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \exp\left\{\sum_{p=-C}^{C} \hat{h}(p)e^{-\jmath\omega p} + \jmath\omega n\right\} d\omega$$

$h(n)$ ↓ *Impulse response at* p_z; $\sum_{p=-C}^{C} \hat{h}(p)$ ↓ *Cepstrum at* p_z

2. Make excitation $e(n)$ with pulses located at p_z
3. Synthesize speech by making $\tilde{s}(n) = h(n) * e(n)$

Proposed phase estimation approach



- Phase iteratively estimated by minimizing the error between natural and reconstructed speech in the time domain
- Pitch period onsets jointly optimized
- No windowing: frame-based time-varying filtering
- No phase unwrapping: cepstral domain

Iterative estimation of phase: requirement

- Pulse positions $\{p_0, \ldots, p_{Z-1}\}$ **must correctly indicate pitch periods** but not necessarily GCI

- Because
  1. Smooth speech spectral envelope at p_z

$$|H_z(e^{\jmath\omega})| = \left| \sum_{n=p_{z-1}}^{p_{z+1}} k(n - p_{z-1})\, s(n) e^{-\jmath\omega n} \right|, \qquad k(n): \text{ window}$$

  2. Real cepstrum

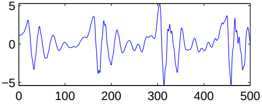
$$\hat{h}_r(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln |H(e^{\jmath\omega})| \, e^{\jmath\omega n} d\omega, \quad n = -C, \ldots, C$$

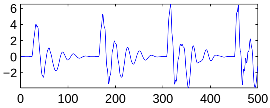
  3. Minimum-phase cepstrum

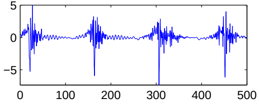
$$\hat{h}_m(n) = \begin{cases} 0 & n < 0 \\ \hat{h}_r(n) & n = 0 \\ 2\hat{h}_r(n) & n = 1, \ldots, C \end{cases}$$

All-pass/minimum-phase speech decomposition

$$s(n) \quad = \quad s_m(n) \quad * \quad s_a(n)$$







$$h(n) = h_m(n) * h_a(n) \Rightarrow \hat{h}(n) = \hat{h}_m(n) + \hat{h}_a(n)$$

Minimum-phase cepstrum	*All pass cepstrum*
$\hat{h}_m(n) = \begin{cases} 0, & -C \leq n \leq -1 \\ \hat{h}(0), & n = 0 \\ \hat{h}(n) + \hat{h}(-n), & 1 \leq n \leq C \end{cases}$	$\hat{h}_a(n) = \begin{cases} \hat{h}(n), & -C \leq n \leq -1 \\ 0, & n = 0 \\ -\hat{h}(-n), & 1 \leq n \leq C \end{cases}$

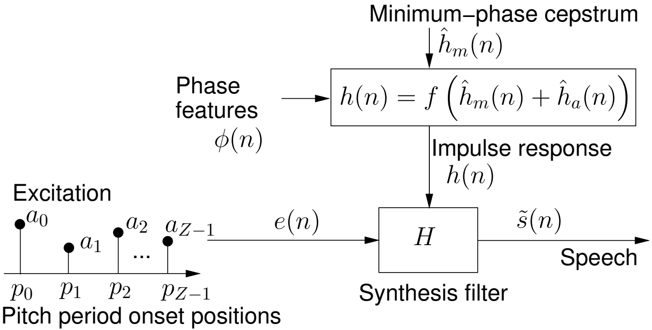
$$\hat{\boldsymbol{h}}_m = \left[\hat{h}_m(0) \quad \cdots \quad \hat{h}_m(C)\right]^\top$$

$$\boldsymbol{\phi} = \left[\hat{h}_a(1) \quad \cdots \quad \hat{h}_a(C)\right]$$

Min.-phase cepstrum: amplitude and min. phase!

Causal all pass cepstrum: residual phase!

Parameters of the model



- Excitation parameters: *hidden variables*
  - $\boldsymbol{a} = \{a_0, \ldots, a_{Z-1}\}$: pulse amplitudes
  - $\boldsymbol{p} = \{p_0, \ldots, p_{Z-1}\}$: pulse locations
- Non-causal synthesis filter parameters: variables to be determined
  - $\{\phi_0, \ldots, \phi_{T-1}\}$: phase features at every frame

A two-step optimization process at the utterance level

Step 1 Estimation of the locations and amplitudes of the excitation signal

- Keep $\hat{h}(n)$ fixed
- Optimize $\{p_0, \ldots, p_{Z-1}\}$ and $\{a_0, \ldots, a_{Z-1}\}$

Step 2 Phase estimation given the new pulse positions and amplitudes

- Keep $e(n)$ fixed
- Re-estimate $\phi(n)$ using a gradient method
  - ➔ Non-linear relationship between $h(n)$ and $\phi(n)$

$$h(n) = f_1\left(\phi(n)\right)$$

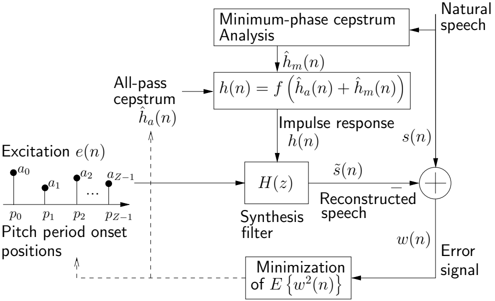
$$\downarrow$$

$$h(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp\left\{ \sum_{p=0}^{C} \hat{h}_m(p) e^{-\jmath\omega p} - 2\jmath \sum_{p=1}^{C} \phi(p-1) \sin(\omega p) + \jmath\omega n \right\} d\omega$$

TOSHIBA Leading Innovation >>>

Step 1: estimation of $\{a_0, p_0, \ldots, a_{Z-1}, p_{Z-1}\}$

$H(z)$ fixed, $e(n)$ to be optimized!



Error vector

$$\boldsymbol{w} = \boldsymbol{s} - \tilde{\boldsymbol{s}} = \boldsymbol{s} - \sum_{z=0}^{Z-1} a_z \boldsymbol{g}_{p_z}$$

where

$$\boldsymbol{g}_n = \Big[\underbrace{0 \ \cdots \ 0}_{n} \quad \boldsymbol{h}_n^\top \quad \underbrace{0 \ \cdots \ 0}_{N-n-1}\Big]^\top$$

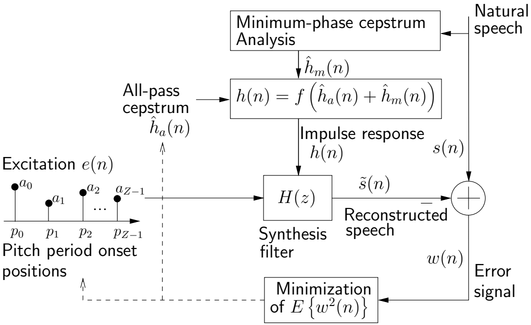
$$\boldsymbol{h}_n = \left[h_n\left(-\frac{M}{2}\right) \quad \cdots \quad h_n\left(\frac{M}{2}\right)\right]^\top$$

Cost function and new positions and amplitudes

$$\varepsilon(\boldsymbol{p}, \boldsymbol{a}) = \frac{1}{N}\left[\boldsymbol{s} - \sum_{z=0}^{Z-1} a_z \boldsymbol{g}_{p_z}\right]^\top \left[\boldsymbol{s} - \sum_{z=0}^{Z-1} a_z \boldsymbol{g}_{p_z}\right] \Rightarrow \begin{cases} \hat{p}_z = \underset{p_z-\Delta p, \ldots, p_z+\Delta p}{\arg\max} \dfrac{\left\{\boldsymbol{g}_{p_z}^\top \left[\boldsymbol{s} - \sum_{i=0, i\neq z}^{Z-1} a_i \boldsymbol{g}_{p_i}\right]\right\}^2}{\boldsymbol{g}_{p_z}^\top \boldsymbol{g}_{p_z}} \\ \hat{a}_z = \dfrac{\boldsymbol{g}_{p_z}^\top \left[\boldsymbol{s} - \sum_{i=0, i\neq z}^{Z-1} a_i \boldsymbol{g}_{p_i}\right]}{\boldsymbol{g}_{p_z}^\top \boldsymbol{g}_{p_z}} \end{cases}$$

Step 2: estimation of $\{\phi_0, \ldots, \phi_{T-1}\}$

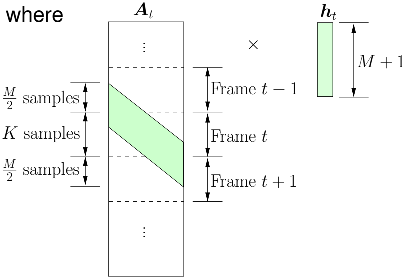
$e(n)$ fixed, $H(z)$ to be optimized!



Error vector

$$\boldsymbol{w} = \boldsymbol{s} - \tilde{\boldsymbol{s}} = \boldsymbol{s} - \sum_{t=0}^{T-1} \boldsymbol{A}_t \boldsymbol{h}_t$$

where



- Matrix $\boldsymbol{A}_t$ contains samples of $e(n)$ at frame t
- Impulse response vector at frame t: $\boldsymbol{h}_t = \left[h_t\left(-\frac{M}{2}\right) \quad \cdots \quad h_t\left(\frac{M}{2}\right)\right]^\top$

Step 2: estimation of $\{\phi_0, \ldots, \phi_{T-1}\}$

- MSE

$$\varepsilon = \frac{1}{N}\left[\boldsymbol{s} - \sum_{t=0}^{T-1} \boldsymbol{A}_t \boldsymbol{h}_t\right]^\top \left[\boldsymbol{s} - \sum_{t=0}^{T-1} \boldsymbol{A}_t \boldsymbol{h}_t\right]$$

- Cost function

$$\varepsilon(\boldsymbol{\phi}_t) = \frac{1}{N}\left[\boldsymbol{r}_t^\top \boldsymbol{r}_t - 2\boldsymbol{r}_t \boldsymbol{A}_t f_1(\boldsymbol{\phi}_t) + \{f_1(\boldsymbol{\phi}_t)\}^\top \boldsymbol{U}_t f_1(\boldsymbol{\phi}_t)\right]$$

$$\begin{cases} \boldsymbol{r}_t = \boldsymbol{s} - \sum_{j=0, j\neq t}^{T-1} \boldsymbol{A}_j f_1(\boldsymbol{\phi}_j) \\ \boldsymbol{U}_t = \boldsymbol{A}_t^\top \boldsymbol{A}_t \end{cases}$$

- Relationship between impulse response and residual phase

$$\boldsymbol{h}_t = f_1(\boldsymbol{\phi}_t) = \frac{1}{2L}\boldsymbol{D}_2 \exp\left(\boldsymbol{D}_{m,1}\hat{\boldsymbol{h}}_{m,t} + \boldsymbol{D}_{a,1}\boldsymbol{\phi}_t\right)$$

$$\begin{cases} D_{m,1}(i,j) = e^{-\jmath\omega_i j} & -L+1 \leq i \leq L, 0 \leq j \leq C \\ D_{a,1}(i,j) = -2\jmath \sin(\omega_i j) & -L+1 \leq i \leq L, 0 \leq j \leq C \\ D_2(i,j) = e^{\jmath\omega_j i} & -\frac{M}{2} \leq i \leq \frac{M}{2}, -L+1 \leq j \leq L \end{cases}$$

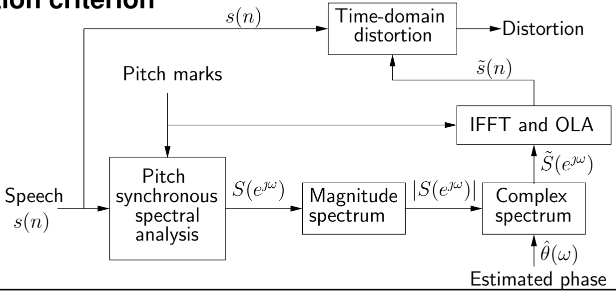
ϕ_t is determined by a gradient descent method!

Experiment

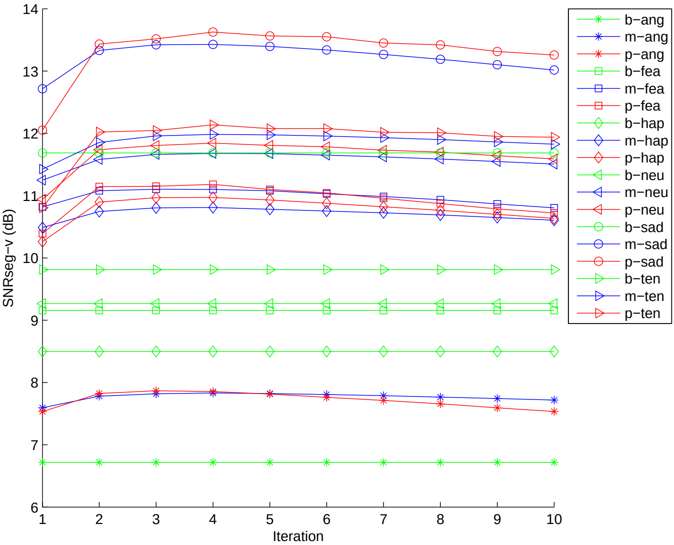
Conditions

- Female UK English speaker, 22.05 kHz, 50 sentences in each of the following styles: angry, fear, happy, neutral, sad and tender
- Methods evaluated
  1. GCI detection using DYPSA [Naylor et al., 2007] + phase unwrapping with a 8192-point DFT (g)
  2. MSE cepstrum analysis (MSE-CCEP) as in [Maia et al., 2013b] (b)
  3. Proposed (r)

Evaluation criterion



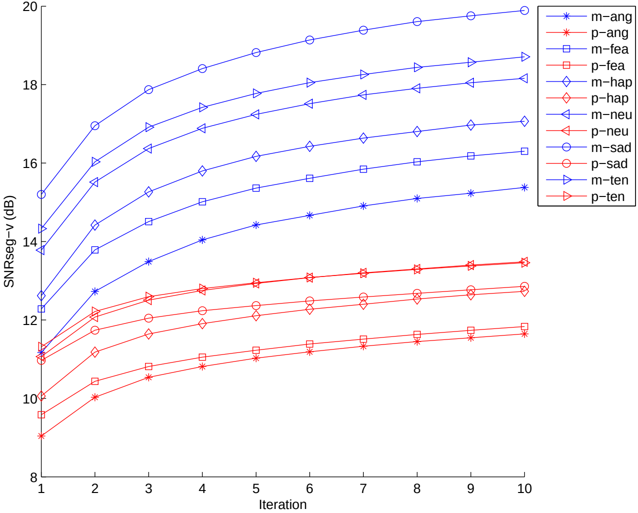
Results



- Proposed method performs better than GCI detection + unwrapping for all speech styles
- Proposed is similar to MSE-CCEP for all styles

TOSHIBA
Leading Innovation >>>

Analysis-synthesis comparison with MSE-CCEP



- MSE-CCEP performs better in terms of analysis-synthesis
- MSE-CCEP optimizes not only phase but also amplitude
- In terms of phase estimation both perform the same
- The proposed method runs in average 3 times faster

Conclusions

- Method to estimate short-term residual phase using complex cepstrum-based analysis and synthesis of speech
- Accurate markings of the pitch periods are necessary
- Better performance than GCI detection followed by multi-resolution phase unwrapping
- Similar performance to the more computationally expensive MSE complex cepstrum analysis
- An automatic way to extract phase information from pitch marks, with no need of
- Future work: application to TTS and ASR exact GCI information

Conclusions

- Method to estimate short-term residual phase using complex cepstrum-based analysis and synthesis of speech
- Accurate markings of the pitch periods are necessary
- Better performance than GCI detection followed by multi-resolution phase unwrapping
- Similar performance to the more computationally expensive MSE complex cepstrum analysis
- An automatic way to extract phase information from pitch marks, with no need of
- Future work: application to TTS and ASR exact GCI information

Conclusions

- Method to estimate short-term residual phase using complex cepstrum-based analysis and synthesis of speech
- Accurate markings of the pitch periods are necessary
- Better performance than GCI detection followed by multi-resolution phase unwrapping
- Similar performance to the more computationally expensive MSE complex cepstrum analysis
- An automatic way to extract phase information from pitch marks, with no need of
- Future work: application to TTS and ASR exact GCI information

Conclusions

- Method to estimate short-term residual phase using complex cepstrum-based analysis and synthesis of speech
- Accurate markings of the pitch periods are necessary
- Better performance than GCI detection followed by multi-resolution phase unwrapping
- Similar performance to the more computationally expensive MSE complex cepstrum analysis
- An automatic way to extract phase information from pitch marks, with no need of
- Future work: application to TTS and ASR exact GCI information

Conclusions

- Method to estimate short-term residual phase using complex cepstrum-based analysis and synthesis of speech
- Accurate markings of the pitch periods are necessary
- Better performance than GCI detection followed by multi-resolution phase unwrapping
- Similar performance to the more computationally expensive MSE complex cepstrum analysis
- An automatic way to extract phase information from pitch marks, with no need of
- Future work: application to TTS and ASR exact GCI information

References

Maia, R., Akamine, M., and Gales, M. (2013a).
Complex cepstrum analysis based on the minimum mean squared error.
In *Proc. of ICASSP*, pages 7972–7976.

Maia, R., Gales, M., Stylianou, Y., and Akamine, M. (2013b).
Minimum mean squared error based warped complex cepstrum analysis for statistical parametric speech synthesis.
In *Proc. of Interspeech*, pages 2336–2340.

Naylor, P., Kounoudes, A., Gudnason, J., and Brookes, M. (2007).
Estimation of glottal closure instants in voiced speech using the DYPSA algorithm.
IEEE Transactions on Audio, Speech, and Language Processing, 15(1):34–43.

TOSHIBA

Leading Innovation >>>