Simple Multi Frame Analysis Methods for Estimation of Amplitude Spectral Envelope in Singing Voice

Gilles Degottex, Luc Ardaillon, Axel Roebel

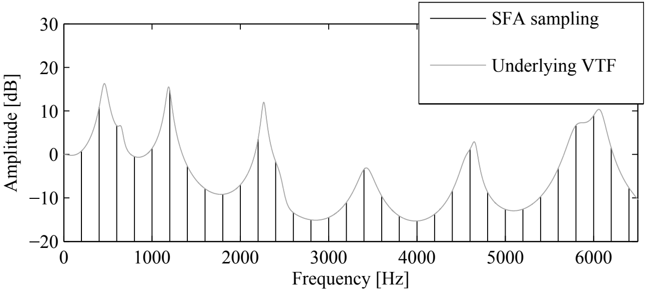
ChaNTeR National Project, France, 2014-2017
http://www.agence-nationale-recherche.fr/?Project=ANR-13-CORD-0011

Ircam - Institut de Recherche Coordination Acoustique/Musique
Paris, France

Problem

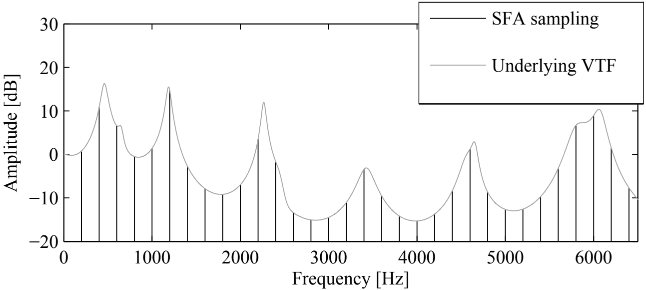
Problem

Problem



SFA sampling
Underlying VTF
Amplitude [dB]
Frequency [Hz]
30
20
10
0
−10
−20
0
1000
2000
3000
4000
5000
6000

Problem



- SFA is enough for reconstructing some shapes, not all of them.

Problem in spectral envelope estimation

Current situation

- A single frame of DFT transform is commonly used.

Problem in spectral envelope estimation

Current situation

- A single frame of DFT transform is commonly used.
- Can we exploit multiple frames to improve the accuracy?

Problem in spectral envelope estimation

Current situation

- A single frame of DFT transform is commonly used.
- Can we exploit multiple frames to improve the accuracy?
- Multiple Frame Analysis (MFA) has already been suggested [1,2].
  Consecutive neighbor frames [2], or non-consecutive among a corpus [1].
  Though, often with excessive complexity and without results used in current applications.

[1] Y. Shiga and S. King "Estimation of voice source and vocal tract characteristics based on multi-frame analysis," *Proc. EUROSPEECH*, 2003.
[2] T. Wang and T. Quatieri "High-pitch formant estimation by exploiting temporal change of pitch" *IEEE TASLP*, 2010.

Problem in spectral envelope estimation

Current situation

- A single frame of DFT transform is commonly used.
- Can we exploit multiple frames to improve the accuracy?
- Multiple Frame Analysis (MFA) has already been suggested [1,2].

  Consecutive neighbor frames [2], or non-consecutive among a corpus [1].

  Though, often with excessive complexity and without results used in current applications.

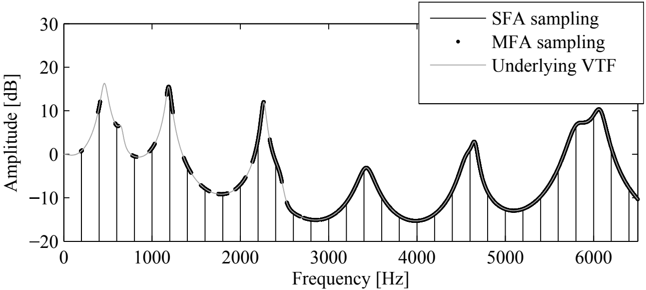
We suggest to

- Start with a simple and very controlled context:
  Sustained segments of singing voice with vibrato.

  (stationary VTF, varying f_0: ideal case for MFA!)
- Study very simple MFA-based envelope estimation methods.

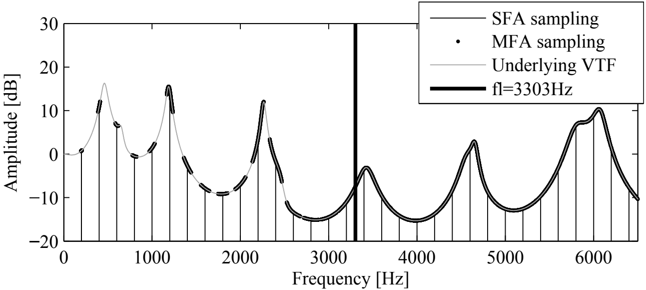
[1] Y. Shiga and S. King "Estimation of voice source and vocal tract characteristics based on multi-frame analysis," *Proc. EUROSPEECH*, 2003.
[2] T. Wang and T. Quatieri "High-pitch formant estimation by exploiting temporal change of pitch" *IEEE TASLP*, 2010.

Problem



SFA sampling
MFA sampling
Underlying VTF
Amplitude [dB]
30
20
10
0
−10
−20
0
1000
2000
3000
4000
5000
6000
Frequency [Hz]

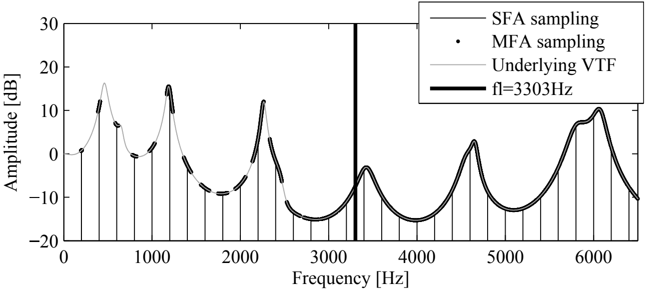
Problem



Using a time window of 2 vibrato periods ($\sim$400ms)

- The MFA coverage increases with the harmonic number
- From f_l and above the VTF is fully covered

Problem



Using a time window of 2 vibrato periods (∼400ms)

- The MFA coverage increases with the harmonic number
- From f_l and above the VTF is fully covered

Questions to answer

- How to reconstruct the remaining low frequencies of the envelope?
- Since the source is AM, how to align consecutive frames?

  (first alignment then envelope estimation? joint estimation of alignment and envelope?)

Methods

Method: Simplified Discrete Cepstral Envelope for MFA (SDCE-MFA)

The envelope model is the same as in SFA:

$$E(f) = c_0 + 2\sum_{n=1}^{P} c_n \cos(n2\pi f/f_s) \tag{1}$$

Method: Simplified Discrete Cepstral Envelope for MFA (SDCE-MFA)

The envelope model is the same as in SFA:

$$E(f) = c_0 + 2\sum_{n=1}^{P} c_n \cos(n2\pi f/f_s) \tag{1}$$

With $\mathbf{c} = [c_0 \cdots c_P]^T$ the traditional solution is [1]:

$$\mathbf{B}^T\mathbf{B}\mathbf{c} = \mathbf{B}^T\mathbf{a} \qquad (SFA) \quad (2)$$

[1] M. Campedel-Oudot, O. Cappe, E. Moulines "Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach," IEEE TSAP, 2001.

Method: Simplified Discrete Cepstral Envelope for MFA (SDCE-MFA)

The envelope model is the same as in SFA:

$$E(f) = c_0 + 2\sum_{n=1}^{P} c_n \cos(n2\pi f/f_s) \tag{1}$$

With $\mathbf{c} = [c_0 \cdots c_P]^T$ the traditional solution is [1]:

$$\mathbf{B}^T\mathbf{B}\mathbf{c} = \mathbf{B}^T\mathbf{a} \qquad (SFA) \quad (2)$$

And the MFA solution suggested by Shiga et al.[2]:

$$\Big(\sum_{k=1}^{K} \mathbf{B}_k^T\mathbf{B}_k\Big)\mathbf{c} = \sum_{k=1}^{K} \mathbf{B}_k^T(\mathbf{a}_k - d_k\mathbf{u}_k) \qquad (MFA) \quad (3)$$

[1] M. Campedel-Oudot, O. Cappe, E. Moulines "Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach," IEEE TSAP, 2001.

[2] Y. Shiga and S. King "Estimation of voice source and vocal tract characteristics based on multi-frame analysis," *Proc. EUROSPEECH*, 2003.

Method: Simplified Discrete Cepstral Envelope for MFA

In this paper, we have shown that the frame alignment d_k impacts only the overall envelope gain, but not its shape.
i.e. The estimation of the shape is independent of the frame alignment!

Method: Simplified Discrete Cepstral Envelope for MFA

In this paper, we have shown that the frame alignment d_k impacts only the overall envelope gain, but not its shape.
i.e. The estimation of the shape is independent of the frame alignment!

Procedure for the SDCE-MFA

1. Compute: $\mathbf{c} = \sum_{k=1}^{K} \left(\sum_{l=1}^{K} \mathbf{B}_l^T \mathbf{B}_l \right)^{-1} \cdot \left(\mathbf{B}_k^T \mathbf{a}_k \right)$
2. Align the resulting shape on the peaks of the central frame

Method: MFA Linear interp. + Cepstral liftering (Linear-MFA-LIFT)

Procedure for the Linear-MFA

1. Align the frames using the energy
2. Linear interpolation of all the peaks among all frames (used in [1])

[1] T. Wang and T. Quatieri "High-pitch formant estimation by exploiting temporal change of pitch" *IEEE TASLP*, 2010.

Method: MFA Linear interp. + Cepstral liftering (Linear-MFA-LIFT)

Procedure for the Linear-MFA-LIFT

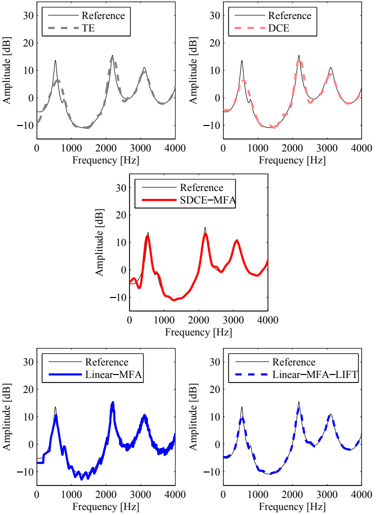
1. Align the frames using the energy
2. Linear interpolation of all the peaks among all frames (used in [1])
3. Low-pass lifter the result according to a cepstral order P

[1] T. Wang and T. Quatieri "High-pitch formant estimation by exploiting temporal change of pitch" *IEEE TASLP*, 2010.

Evaluation

Evaluation

Evaluation: Methods compared



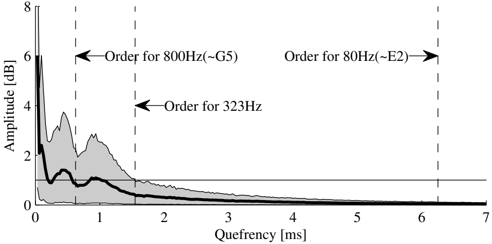
SFA

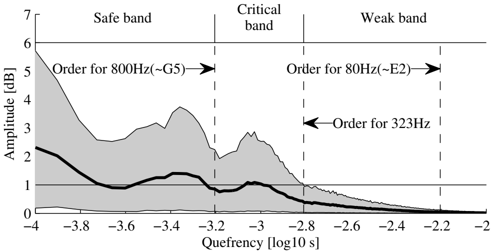
- "True-Envelope" (TE)
  Also cepstral model, iterative solution
- Discrete Cepstral Envelope (DCE)
  LS solution + regularization

MFA

- SDCE-MFA
  LS solution
- Linear-MFA-LIFT
  Liftered linear interpolation

Evaluation: Quefrency bands

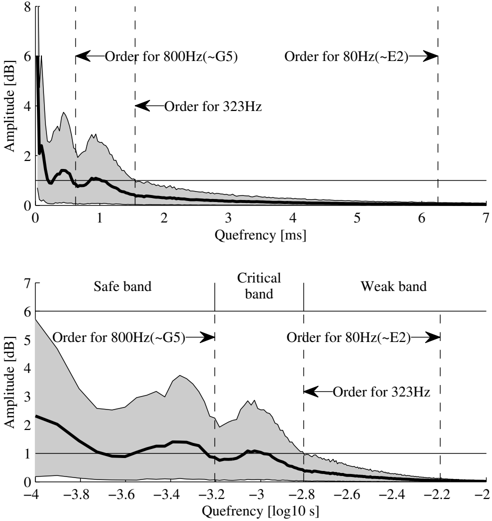




- 5%, 50% and 95% of cepstral magnitude distribution

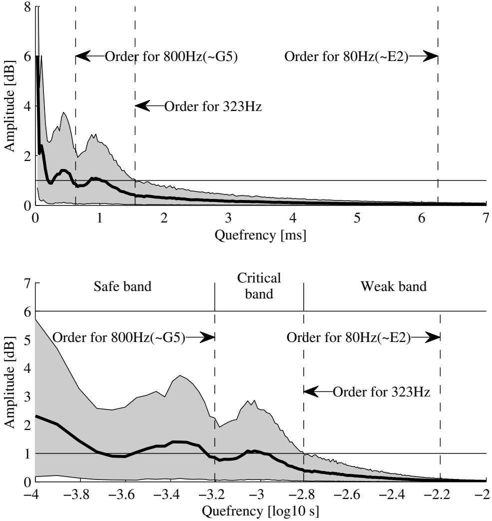
  (Maeda digital acoustic synthesizer)

Evaluation: Quefrency bands



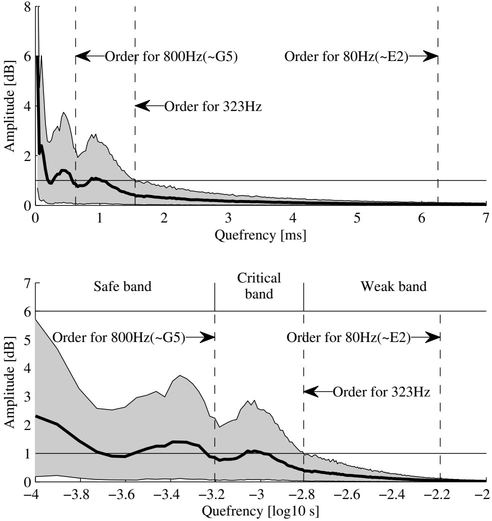
- 5%, 50% and 95% of cepstral magnitude distribution
  (Maeda digital acoustic synthesizer)
- Assuming $f_0 \in [80, 800]$Hz
- Usual order for cepstral models: $\mathrm{P} = \frac{0.5 f_s}{f_0}$

Evaluation: Quefrency bands



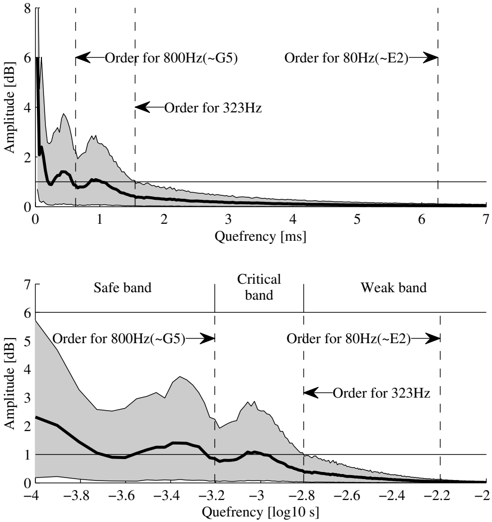
- 5%, 50% and 95% of cepstral magnitude distribution
  (Maeda digital acoustic synthesizer)
- Assuming $f_0 \in [80, 800]$Hz
- Usual order for cepstral models: $\mathrm{P} = \frac{0.5 f_s}{f_0}$
- **Weak band**:
  From auditory threshold to inf

Evaluation: Quefrency bands



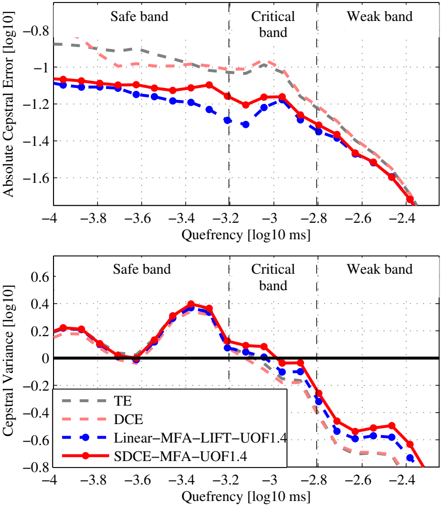
- 5%, 50% and 95% of cepstral magnitude distribution
  (Maeda digital acoustic synthesizer)
- Assuming $f_0 \in [80, 800]$Hz
- Usual order for cepstral models: $\mathrm{P} = \frac{0.5 f_s}{f_0}$
- **Weak band**:
  From auditory threshold to inf
- **Safe cepstral band**:
  Zero to minimum order

Evaluation: Quefrency bands



- 5%, 50% and 95% of cepstral magnitude distribution
  (Maeda digital acoustic synthesizer)
- Assuming $f_0 \in [80, 800]$Hz
- Usual order for cepstral models: $P = \frac{0.5 f_s}{f_0}$
- **Weak band**: From auditory threshold to inf
- **Safe cepstral band**: Zero to minimum order
- **Critical band**: From minimum order to auditory threshold of 1dB

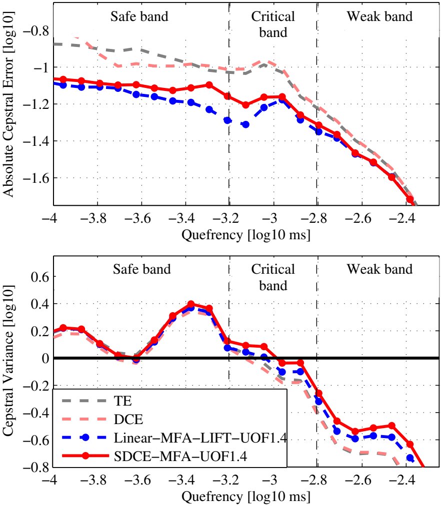
Evaluation: Numerical evaluation using synthetic signals



Experimental data

- 1000 samples of 2s
- $f_0 \in [80, 800]$Hz
- Vibrato extent $\in [0, 150]$cents
- Vibrato freq $\in [4, 6]$Hz
- Source Dirac impulse
- Random AM of std$= 0.5$dB
- Convolved by a random VTF

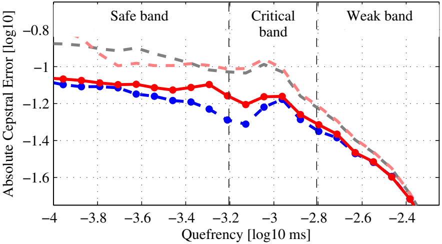
Evaluation: Numerical evaluation using synthetic signals

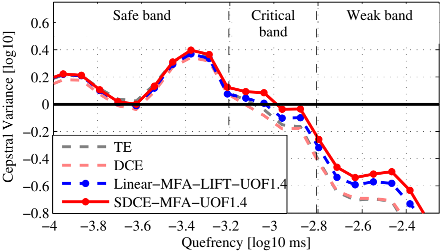


Estimation setup

- Window length for MFA:
  2 periods of vibrato $\Rightarrow \sim$400ms
- We take advantage of the MFA by *boosting* the cepstral order:
  $P = 1.4 \cdot \frac{0.5 f_s}{f_0}$

Evaluation: Numerical evaluation using synthetic signals





Absolute cepstral error:

$$\epsilon_{n,i} = \frac{1}{M} \sum_{m=1}^{M} \left| c^*_{n,i} - c_{m,n,i} \right| \qquad (4)$$

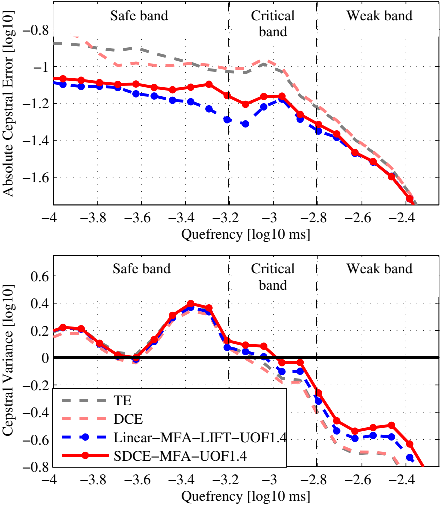
$c^*_{n,i}$ the reference sample i; M the number of frames in i

Cepstral Variance:

$$\bar{\sigma}_n = \frac{\mathrm{std}_i(\bar{c}_{n,i})}{\mathrm{std}_i(\bar{c}^*_{n,i})} \quad \bar{c}_{n,i} = \frac{1}{M} \sum_{m=1}^{M} c_{m,n,i} \qquad (5)$$

$\bar{c}_{n,i}$ the average cepstrum over M; $\mathrm{std}_i(.)$ the standard-deviation over i

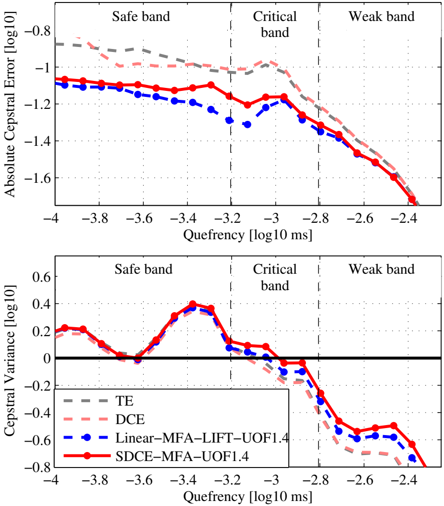
Evaluation: Numerical evaluation using synthetic signals



Results

- In safe band: MFA almost divides the error by 2

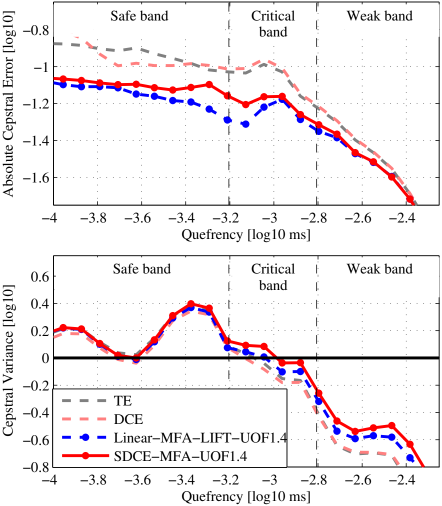
Evaluation: Numerical evaluation using synthetic signals



Results

- In safe band: MFA almost divides the error by 2
- In safe band: High absolute error explains the higher variance

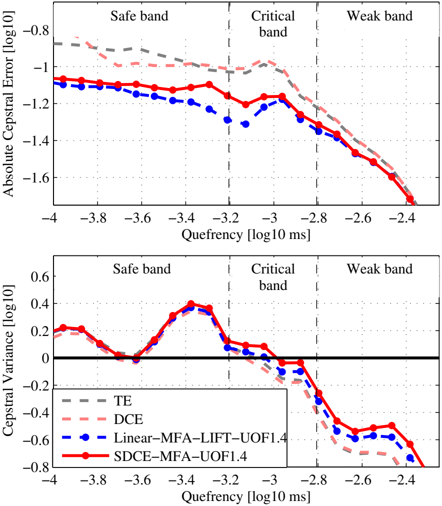
Evaluation: Numerical evaluation using synthetic signals



Results

- In safe band: MFA almost divides the error by 2
- In safe band: High absolute error explains the higher variance
- In the critical band: The variance drops quickly

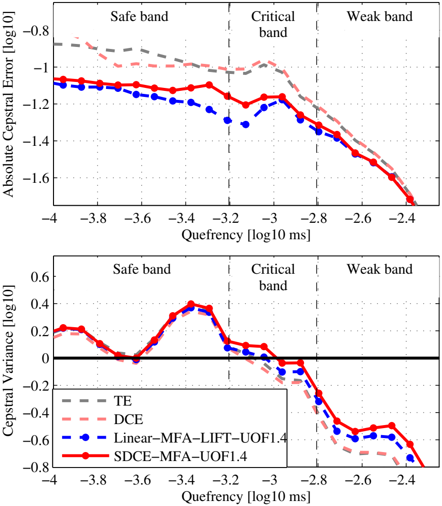
Evaluation: Numerical evaluation using synthetic signals



Results

- In safe band: MFA almost divides the error by 2
- In safe band: High absolute error explains the higher variance
- In the critical band: The variance drops quickly
  Averaging effect!
  (without any statistical modeling!)

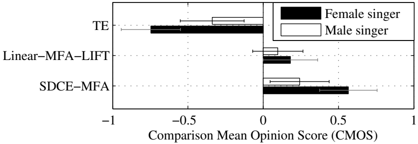
Evaluation: Numerical evaluation using synthetic signals



Results

- In safe band: MFA almost divides the error by 2
- In safe band: High absolute error explains the higher variance
- In the critical band: The variance drops quickly
  Averaging effect!
  (without any statistical modeling!)
- MFA methods better recover the variance

Evaluation: Listening tests about pitch scaling



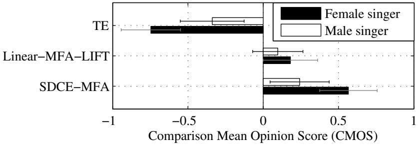
Experimental setup

- 2 voices (female and male) (proof of concept!)
- 15 sustained French vowels + natural vibrato
- Up and down pitch scaling (x0.75, x1.25)
- Dropped DCE-SFA to keep test duration low
- Each listener assessed 4 random vowels
- Web-based sent to mailing-lists

Samples accessible at:

`http://gillesdegottex.eu/Demos/DegottexG2016mfaenvsing/`

Evaluation: Listening tests about pitch scaling



Results (31 listeners)

- MFA methods clearly preferred
- Linear-MFA-LIFT very good improvement

  (and very efficient computationally!)
- SDCE-MFA shows better improvement for the female voice

  (might be due to the better variance reconstruction)

Samples accessible at:

`http://gillesdegottex.eu/Demos/DegottexG2016mfaenvsing/`

Conclusions: Points worth remembering

- With MFA, the filter response is fully covered above some freq.
  $\Rightarrow$Need to reconstruct only the low frequencies.

Conclusions: Points worth remembering

- With MFA, the filter response is fully covered above some freq.
  $\Rightarrow$Need to reconstruct only the low frequencies.
- Numerical evaluation shows error reduction by factor $\sim$2

Conclusions: Points worth remembering

- With MFA, the filter response is fully covered above some freq.
  $\Rightarrow$Need to reconstruct only the low frequencies.
- Numerical evaluation shows error reduction by factor $\sim$2
- Cepstral variance drops quickly above the lowest cepstral order
  (averaging effect at features' estimation level!)

Conclusions: Points worth remembering

- With MFA, the filter response is fully covered above some freq.
  $\Rightarrow$Need to reconstruct only the low frequencies.
- Numerical evaluation shows error reduction by factor $\sim$2
- Cepstral variance drops quickly above the lowest cepstral order
  (averaging effect at features' estimation level!)
- Linear-MFA-LIFT is very simple and efficient

Conclusions: Points worth remembering

- With MFA, the filter response is fully covered above some freq.
  $\Rightarrow$Need to reconstruct only the low frequencies.
- Numerical evaluation shows error reduction by factor $\sim$2
- Cepstral variance drops quickly above the lowest cepstral order
  (averaging effect at features' estimation level!)
- Linear-MFA-LIFT is very simple and efficient
- MFA-based methods seem to improve pitch scaling

Conclusions: Points worth remembering

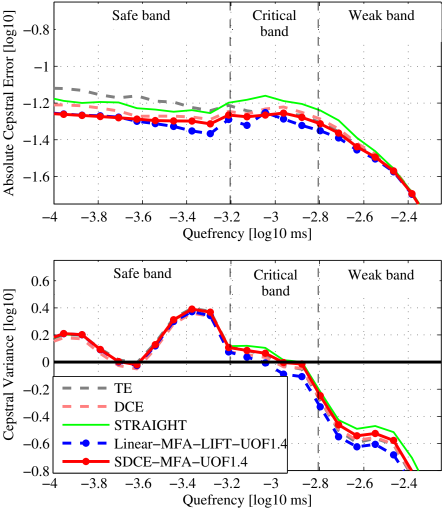
- With MFA, the filter response is fully covered above some freq.
  $\Rightarrow$Need to reconstruct only the low frequencies.
- Numerical evaluation shows error reduction by factor $\sim$2
- Cepstral variance drops quickly above the lowest cepstral order
  (averaging effect at features' estimation level!)
- Linear-MFA-LIFT is very simple and efficient
- MFA-based methods seem to improve pitch scaling

Journal article just accepted!

G. Degottex, L. Ardaillon, A. Roebel "Multi-Frame Amplitude Envelope Estimation for Modification of Singing Voice", *IEEE TASLP*, accepted 2016.

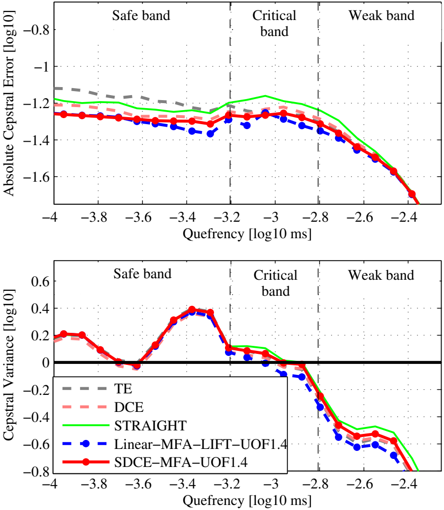
感謝您的關注
Gracias por su atención
Thank you for your attention
आप अपना ध्यान के लिए धन्यवाद
شكرا لكم على اهتمامكم
Obrigado pela sua atenção
Спасибо за ваше внимание
ご清聴ありがとう
మీ శ్రద్ధకు ధన్యవాదాలు
Je vous remercie de votre attention
Σας ευχαριστώ για την προσοχή σας
Dankon pro via atento
please no question about the window

Evaluation: Numerical evaluation using synthetic signals



What about speech?

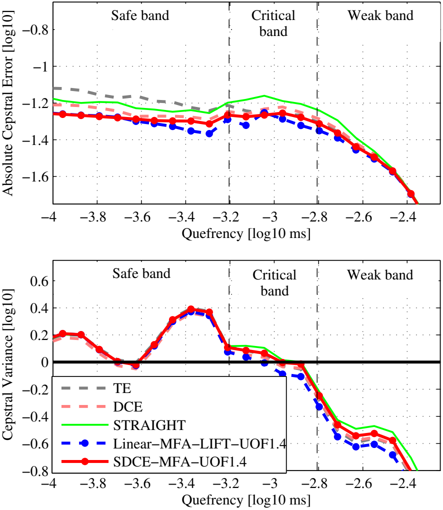
Evaluation: Numerical evaluation using synthetic signals



What about speech?

- Same experiment using TIMIT database to build f_0 and AM

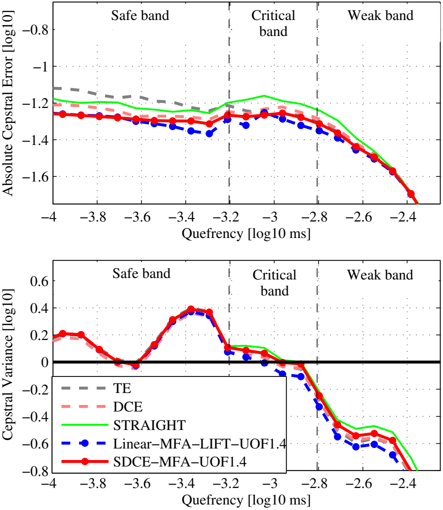
Evaluation: Numerical evaluation using synthetic signals



What about speech?

- Same experiment using TIMIT database to build f_0 and AM
- MFA window length: 30ms
  $\Rightarrow$ Less disparities of errors

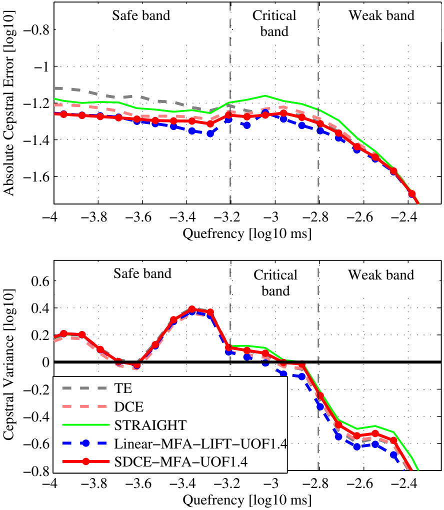
Evaluation: Numerical evaluation using synthetic signals



What about speech?

- Same experiment using TIMIT database to build f_0 and AM
- MFA window length: 30ms
  $\Rightarrow$ Less disparities of errors
- $f_0 \in [80, 500]$Hz
  $\Rightarrow$ Less disparities of variance
  $\Rightarrow$ Variance of SFA similar to MFA

Evaluation: Numerical evaluation using synthetic signals



What about speech?

- Same experiment using TIMIT database to build f_0 and AM
- MFA window length: 30ms
  $\Rightarrow$ Less disparities of errors
- $f_0 \in [80, 500]$Hz
  $\Rightarrow$ Less disparities of variance
  $\Rightarrow$ Variance of SFA similar to MFA
- For MFA we need:

  $$\frac{\mathrm{var}(\mathrm{VTF})}{\mathrm{var}(f_0)} \text{ as small as possible}$$

  Small enough for speech?
  Doesn't seem to be the case :(