A Time Regularization Technique for Discrete Spectral Envelopes Through Frequency Derivative

University of Crete, Computer Science Department
FORTH, Institute of Computer Science
FNSNF Swiss National Science Foundation

*Gilles Degottex** with follow-up works with Luc Ardaillon and Axel Roebel
University of Crete and FORTH, Heraklion, Greece — IRCAM, Paris, France
ircam Centre Pompidou

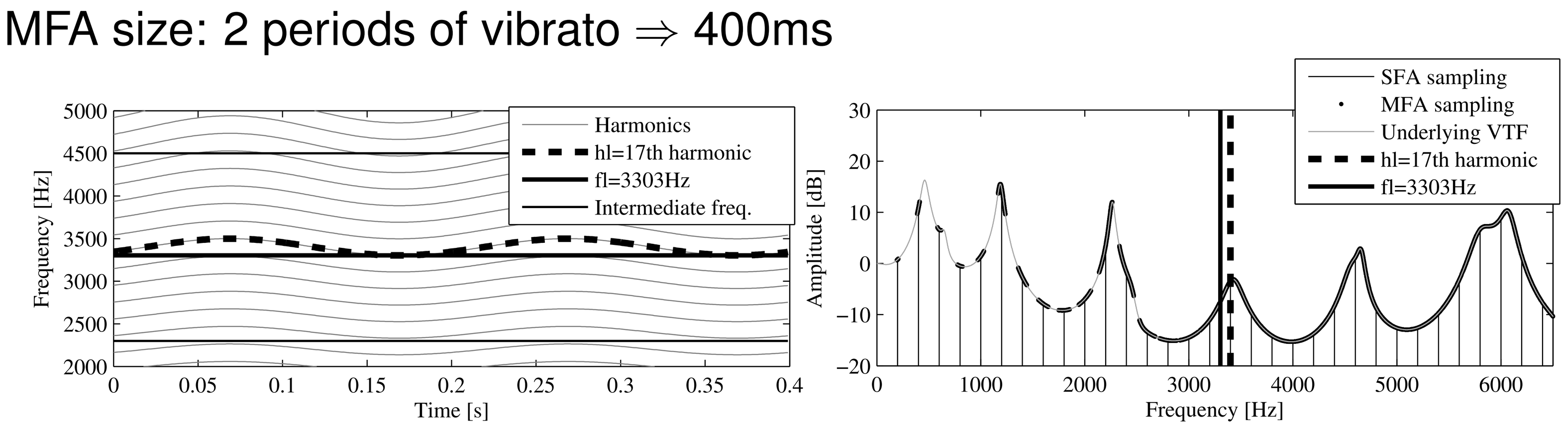
ABSTRACT

Issues in amplitude spectral envelope estimation:

- Undersampled Vocal Tract Filter (VTF)
- Shape reconstruction
- Time regularity

Traditional approach: Single Frame Analysis (SFA)
Why not using Multiple Frame Analysis (MFA) [4] ?!

Evaluation shown for singing voice [1,2]
MFA size: 2 periods of vibrato $\Rightarrow$ 400ms



MFA METHODS

The methods below use spectral peaks extracted on each frame (each 5ms).

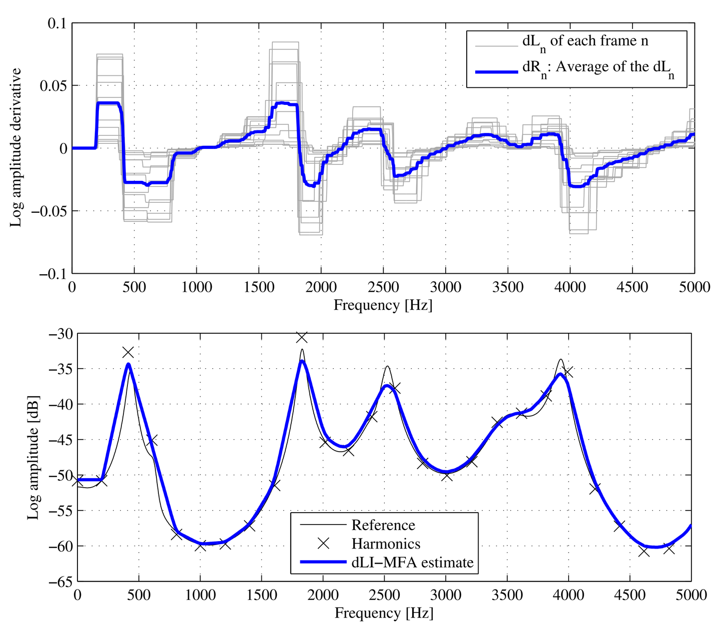
Extra issue: The frames' energy follows the amplitude modulation (AM) of the signal (e.g. tremolo in singing voice) $\Rightarrow$ **How to align the frames ?**

DLINEAR-MFA[THIS LETTER]

Steps:

1. Compute derivative of the linear envelope (on log scale).
2. Averaging of K frames across time.
3. Compute integral of the averaged derivated envelope.
4. Final alignment on the central frame $(K-1)/2$.

(No need of frame pre-alignment!)



SDCE-MFA[1,2]

Model for the Discrete Cepstral Envelope (DCE):

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

c_n: the cepstral coefficients, P: the cepstral order.

For MFA, Shiga et al. [4] suggested to minimize:

$$\epsilon = \sum_{k=1}^{K} \|\boldsymbol{a}_k - d_k \boldsymbol{u}_k - \boldsymbol{B}_k \boldsymbol{c}\| \qquad (2)$$

k: the frame, $\boldsymbol{a}_k$: the log amplitudes, d_k: correction term for the AM, $\boldsymbol{u}_k = [1,\ldots,1]^T$, $\boldsymbol{c}$: the cepstral coefficients, $\boldsymbol{B}$: the Fourier basis.

Steps:

1. Compute:

$$\boldsymbol{c} = \sum_{k=1}^{K} \left(\sum_{l=1}^{K} \boldsymbol{B}_l^T \boldsymbol{B}_l\right)^{-1} \cdot \left(\boldsymbol{B}_k^T \boldsymbol{a}_k\right) \qquad (3)$$

   (No need of frame pre-alignment! Shown in [1,2])

2. Final alignment on the central frame.

We can increase the order ! (e.g. x1.4 in the following experiments)

LINEAR-MFA-LIFT[1,2]

MFA version of the basic linear interpolation, which has already been used for comparison [5]. We only add a low-pass liftering.

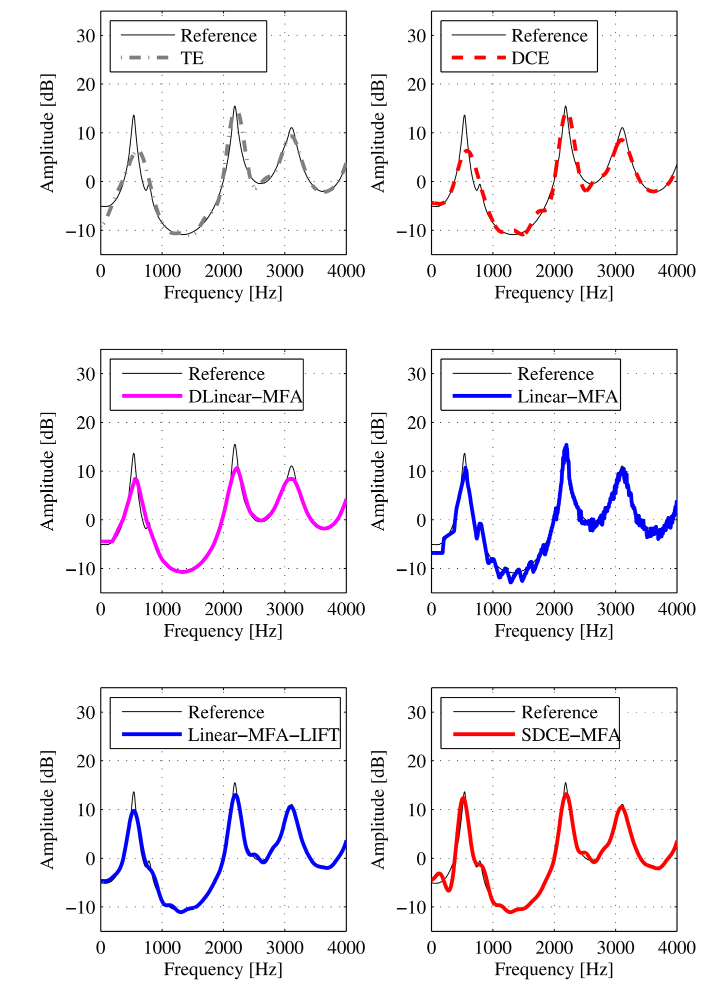
Steps:

1. Pre-alignement of K successive frames using energy in [0-4]kHz.
2. Linear interpolation of all peaks of the K frames [5] $\Rightarrow$ Linear-MFA.
3. Low-pass lifter of the Linear-MFA to alleviate erratic shapes.
4. Final alignment on the central frame.

EVALUATION

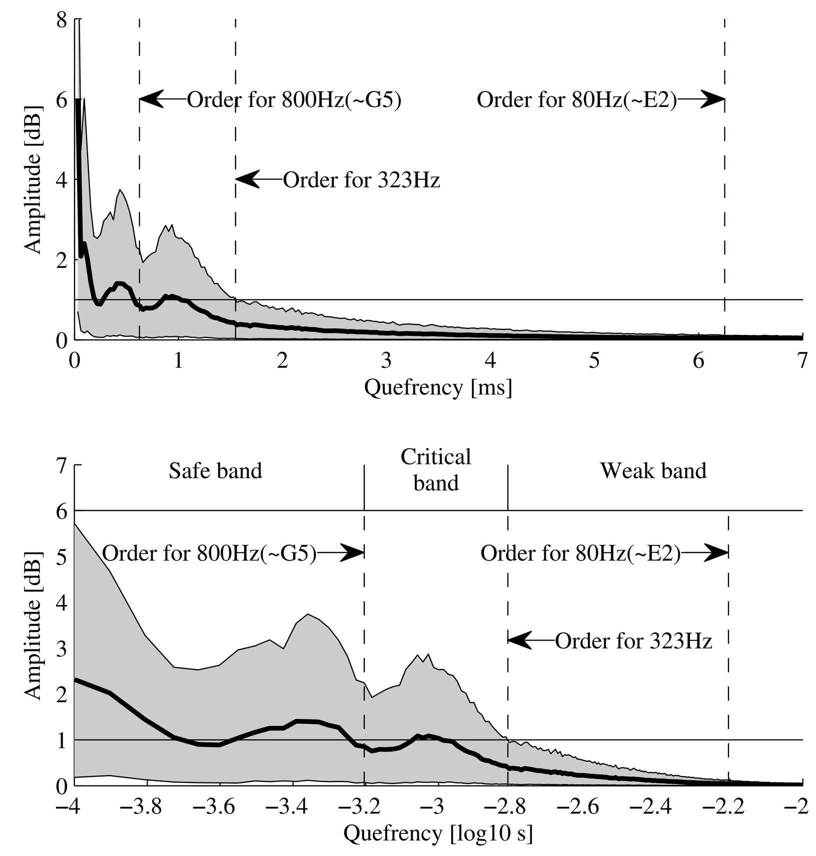
Compared methods

TE=True Envelope (SFA)
DCE=Discrete Cepstral Envelope (SFA)



Material

1000 synthetic signals of 2s; f_0: vibrato with central freq. in $[80, 800]$Hz; AM component: low-passed filtered Gaussian noise; VTF from digital acoustic synthesizer.



MEASUREMENTS

Absolute cepstral error (AC Error)

$$\epsilon_n = \frac{1}{M}\sum_{m=1}^{M} |c_n^* - c_{m,n}|$$

M: the number of frames
c_n^*: The known VTF.

Relative cepstral error (RC Error)

$$\rho_n = \frac{1}{M}\sum_{m=1}^{M} \left|\frac{c_n^* - c_{m,n}}{c_n^*}\right|$$

Cepstral Variance: The capacity to reproduce the *global* variance:

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

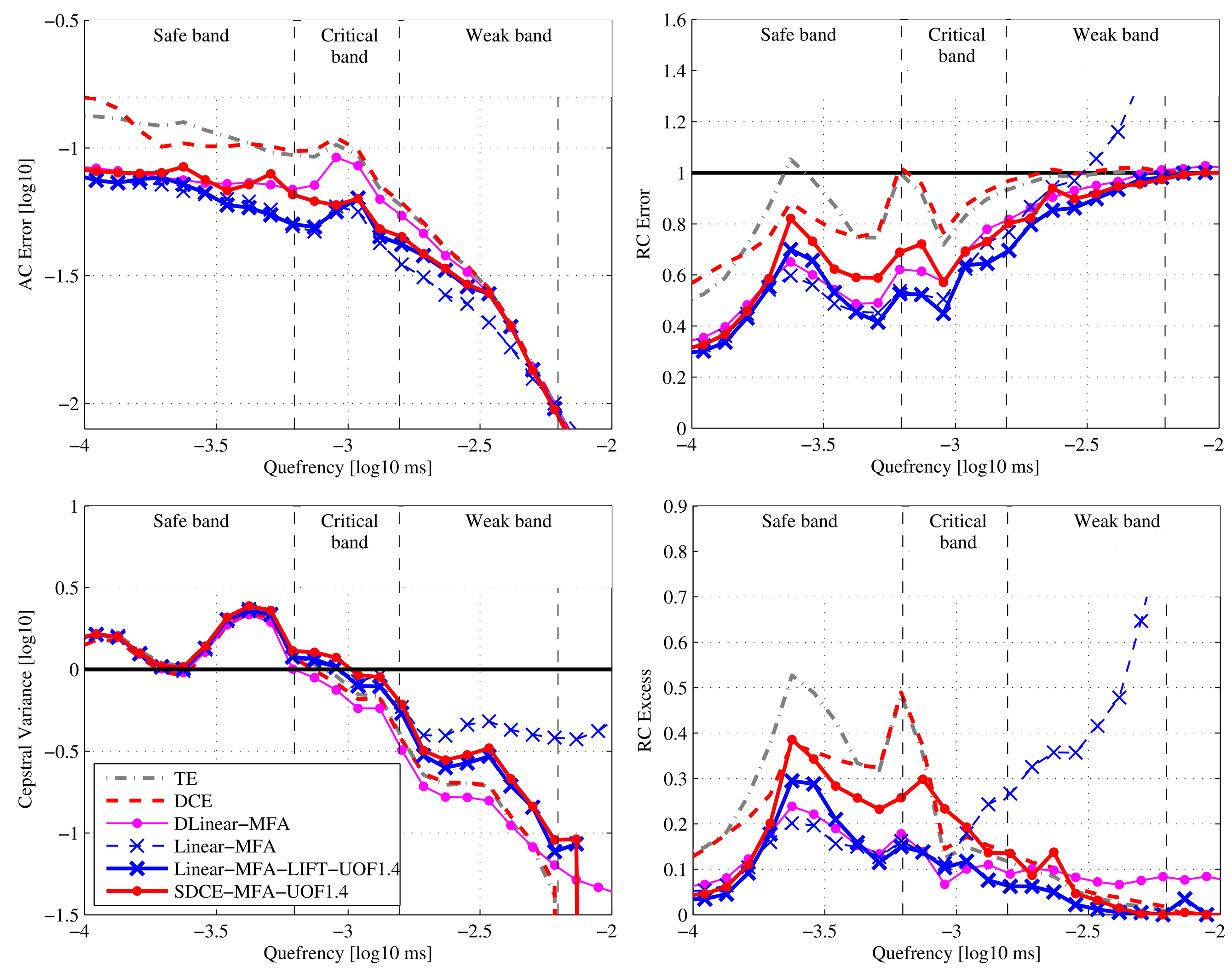
$\bar{c}_{n,i}$: the average cepstrum of sample i.

Relative Cepstral Excess (RC Excess): Risk of degeneration or incoherent resonances.

$$\chi_n = \frac{1}{M}\sum_{m=1}^{M} \max(\{\rho_{m,n}, 1\}) - 1$$

$$\rho_{m,n} = \left|\frac{c_n^* - c_{m,n}}{c_n^*}\right|$$

RESULTS

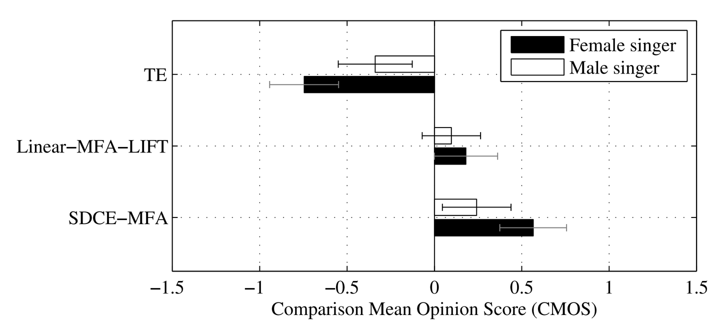


- RC Error, safe band: MFA divides the error by ~2 compared to SFA.
- Critical and weak bands, cepstral Variance < 1 $\Rightarrow$ averaging of the envelopes (without any statistical modeling!). The DLinear-MFA suffers the most. SDCE-MFA is the best.
- RC Excess: SFA produces substantial erratic shapes, even in the safe band. Much less problems for the MFA.

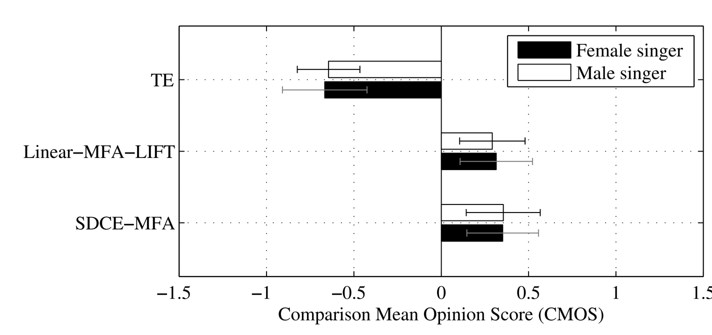
LISTENING TESTS

Material: 1-3s of sustained vibrato; 15 French vowels; 2 voices (!)
Method: Harmonic synthesis[6]

Pitch scaling: x1.25 and x0.75



Timbre conversion: from *mf* to *fff*



- MFA clearly prefered to TE.
- Linear-MFA-LIFT: good results and efficient !

[1] G. Degottex, L. Ardaillon, A. Roebel, "Simple Multi Frame Analysis methods for estimation of Amplitude Spectral Envelope estimation in Singing Voice", in ICASSP, 2016.
[2] G. Degottex, L. Ardaillon, A. Roebel, "Multi-Frame Amplitude Envelope Estimation for Modification of Singing Voice", IEEE Transactions on Audio, Speech, and Language Processing, Accepted 2016.
[3] M. Campedel-Oudot, O. Cappe, E. Moulines "Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach", IEEE Transactions on Speech and Audio Processing, 2001.
[4] Y. Shiga, S. King, "Estimating the spectral envelope of voiced speech using multi-frame analysis", EUROSPEECH, 2003.
[5] T. Wang, T. Quatieri, "High-pitch formant estimation by exploiting temporal change of pitch", IEEE Transactions on Audio, Speech, and Language Processing, 2010.
[6] G. Kafentzis, G. Degottex, O. Rosec, Y. Stylianou, "Pitch Modifications of speech based on an Adaptive Harmonic Model", in ICASSP, 2014.

* Now supported by Horizon2020 Marie Sklodowska-Curie Actions (MSCA) - Research Fellowship - 655764 - HQSTS