Speech Analysis/Synthesis by Non-parametric Separation of Vocal Source and Tract Responses

Kanru Hua *(khua5@illinois.edu)*

University of Illinois at Urbana-Champaign

Introduction

- Recent speech analysis/synthesis methods based on explicit separation of glottal flow and vocal tract responses are vulnerable to parameter estimation errors. We tackle this problem by using a non-parametric representation of glottal flow signal.

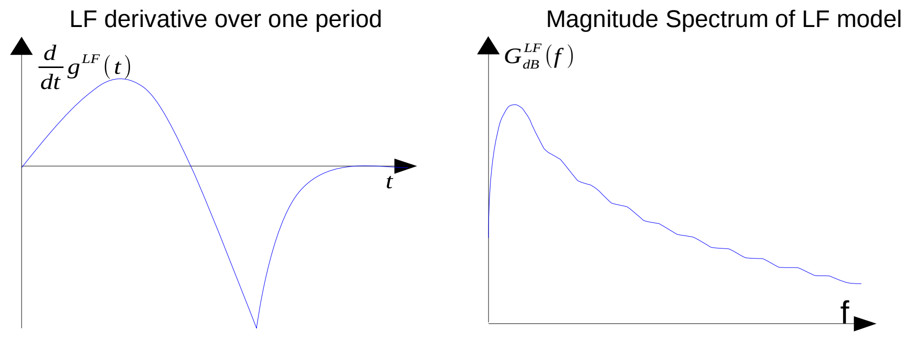
Motivation

- Classical source-filter model assumes that vocal-tract and source characteristics are independent from pitch:

$$S(\omega)=H^{f_0}(\omega)C(\omega)$$

- The SVLN method (Degottex, et al., 2013) introduces a glottal pulse parametrized by Liljencrants-Fant model:

$$S(\omega)=H^{f_0}(\omega)G^{LF}(\omega)C(\omega)L(\omega)$$

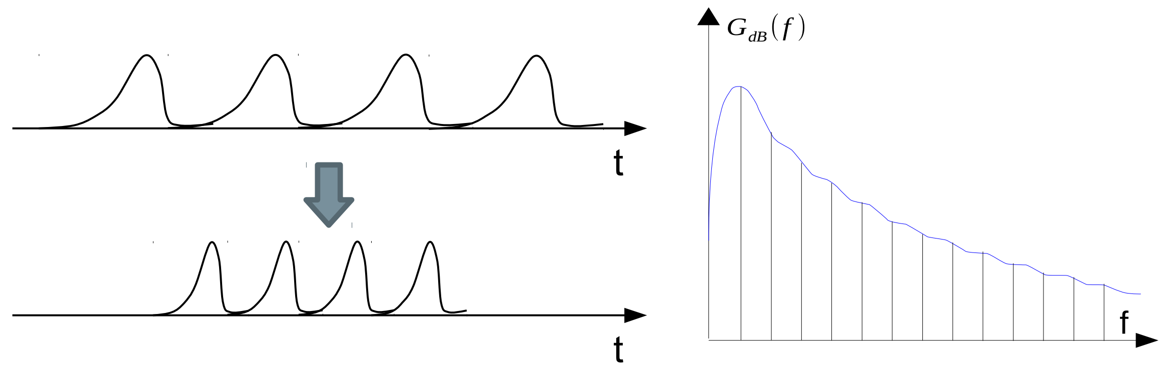


(fig. 1 waveform and spectrum of a single period of LF model)

- However robust estimation of LF parameters is a non-trivial problem.

Our Approach

- Assumption: the effect of pitch on glottal flow signal is mainly time scaling.

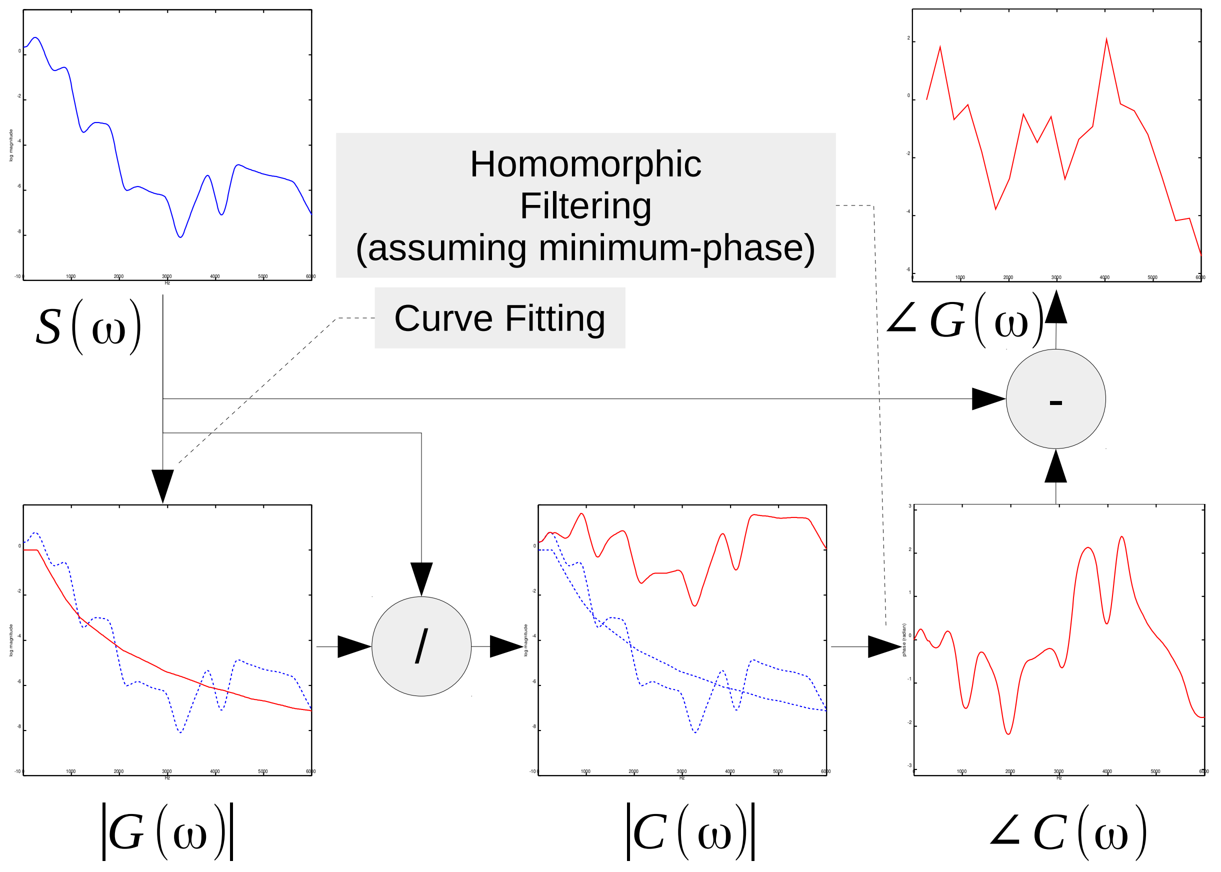


(fig.2 Left: example of time scaling a glottal flow signal; right: harmonics taken from the spectrum of a glottal flow signal)

- Then we can explicitly store the magnitude and phase spectrum of glottal pulse, sampled at each harmonic.

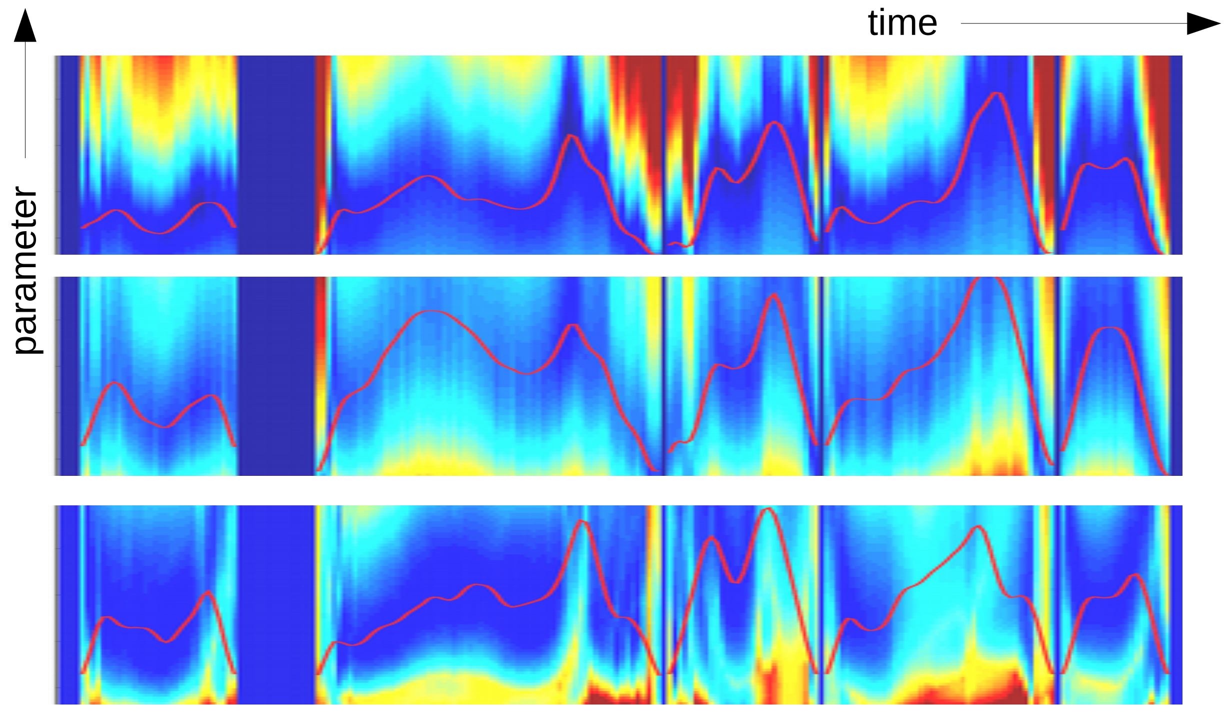
Analysis/Synthesis Method

- First recover magnitude responses ($|G(\omega)|$ and $|C(\omega)|$), then recover glottal phase $\angle G(\omega)$ by subtracting $\angle C(\omega)$ from the input spectrum.



(fig. 3 schematic diagram of the analysis method)

- $|G(\omega)|$ is directly estimated from $|S(\omega)|$ by minimizing distance between the two spectra with respect to LF model parameter(s).

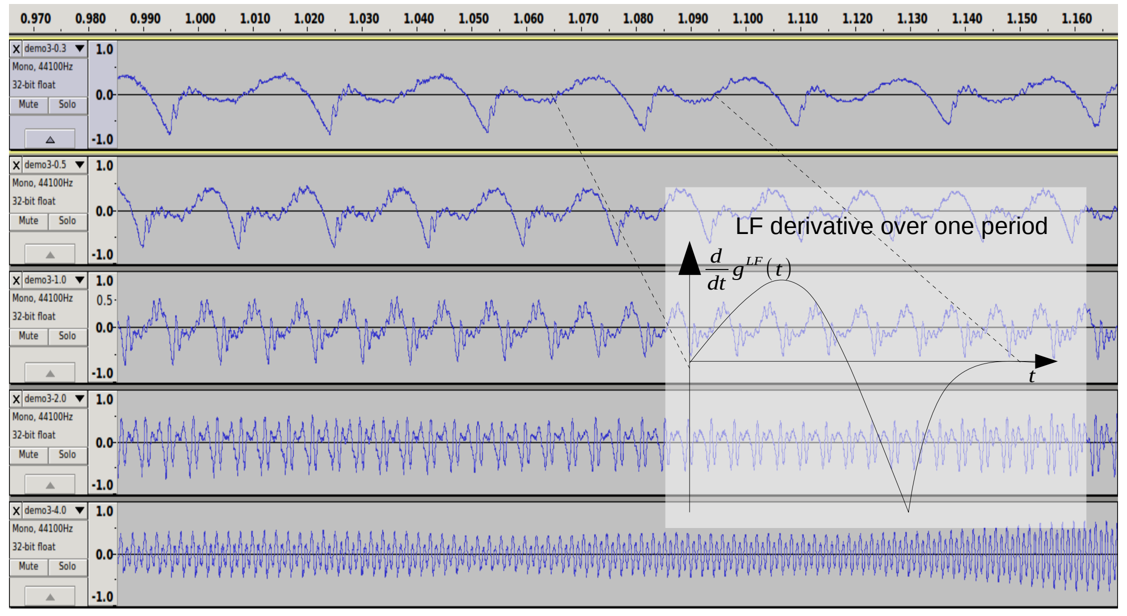


(fig. 4 From top to bottom: error surfaces of LF parameter fitting, based on Itakura-Saito distance, Euclidean distance between log magnitude spectra, and MSP method, respectively.)

- Magnitude-only fitting gives smoother error surface than the phase-based method; Itakura-Saito distance performs better than Euclidean distance.
- Synthesis is basically the reverse of analysis procedures.

Evaluation

- Apply the proposed method to scale the f0 of speech samples.



(fig. 5 waveforms of pitch scaled (by 0.3, 0.5, 1.0, 2.0 and 4.0 times from top to bottom) speech)

- As pitch becomes lower, the waveform converges to the shape of glottal flow derivative. This implies that the model *implicitly* captures phase characteristics of vocal source.
- Informal test shows the result sounds more natural than methods based on parametric glottal models.

Conclusion

- The proposed method describes glottal flow signal by the amplitudes and phases of harmonics, assuming the signal is shape-invariant under pitch scaling. Magnitude-only fitting of LF model turned out to be sufficient for robust inverse analysis of speech.
- Next step: investigate different models and distance measures for estimation of $|G(\omega)|$; investigate how to meaningfully modify recovered $\angle G(\omega)$.

References

- Degottex, Gilles, et al. "Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis." *Speech Communication* 55.2 (2013): 278-294.
- Degottex, Gilles, Axel Roebel, and Xavier Rodet. "Phase minimization for glottal model estimation." *Audio, Speech, and Language Processing, IEEE Transactions on* 19.5 (2011): 1080-1090.
- Fant, Gunnar. "The LF-model revisited. Transformations and frequency domain analysis." *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm* 2.3 (1995): 40.

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN