INTERSPEECH 2016: Show & Tell Contribution
September 8–12, 2016, San Francisco, USA

SparkNG: Interactive MATLAB tools for introduction to speech production, perception and processing fundamentals and application of the aliasing-free L-F model component

Hideki Kawahara[1]

[1]Wakayama University, Wakayama, 640-8510 Japan
kawahara@sys.wakayama-u.ac.jp

Abstract

This article introduces a set of interactive tools for studying fundamentals of speech production, perception and processing. In addition to this voice production simulator, it consists of interactive time-frequency representation, auditory representation visualizer and a vocal tract shape visualizer for introductory materials. It consists of compiled executables for Windows and Mac environment, which do not require MATLAB license. The MATLAB sources of the tools and their constituent functions are publicly accessible under open source license.
Index Terms: speech production, speech perception, vocal tract, glottal source, linear prediction

1. Introduction

Combination of the source-filter model of speech production [1] and the one dimensional acoustic tube model of vocal tract [2] still serve as a relevant introduction to speech science. Recent advances in computer science make it possible to provide an environment for students to explore relations between constituent components of speech production and their physical instantiation interactively in real-time. This article introduces an example of such environment based on a closed form representation [3] of the L-F model of glottal excitation source [4] and additional interactive tools for understanding fundamentals of auditory signal processing. The tools and MATLAB souce codes are publicly accessible [5] under open source license. The following sections mainly describe the speech production simulator and briefly describe additional tools for introducing fundamentals of auditory signal processing. It also introduces an application example of the aliasing-free L-F model function.

2. Speech production simulator

The speech production simulator has two GUIs. The main GUI is for designing vocal tract shape and transfer function and synthesizes speech sounds using the designed vocal tract and the source signal. The source model GUI is for designing glottal source model based on the L-F model. These GUIs call elementary functions. They consist of conversion functions of LPC parameter family [6], anti-aliased L-F model signal generator [3], lattice filter based on PARCOR and so on. The tools page [5] has links to tutorial movies of the simulator.

2.1. Main GUI

Figure 1 shows the main GUI. The upper left plot shows the vocal tract area function. The horizontal axis of the plot represents the distance from the lip opening. The vertical axis represents the area using logarithmic scaling. The vertical cyan line is a handle to manipulate the vocal tract length.

The lower left plot shows the vocal tract shape modifier. The left-side vertical line is a slider to control the magnitude of modification to the vocal tract shape. The black line in the plot shows the modification shape to be magnified and copied to the upper left plot. Many color lines are basis functions to shape the modification shape, the black line.

The three dimensional shape in the lower right part is prepared to help grasping the vocal tract shape more intuitively.

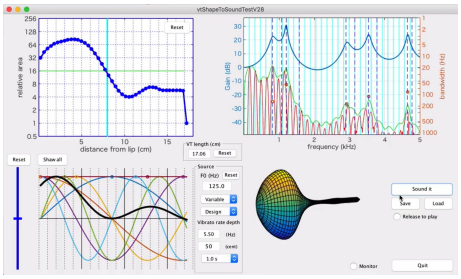

Figure 1: Main GUI of the speech production simulator.

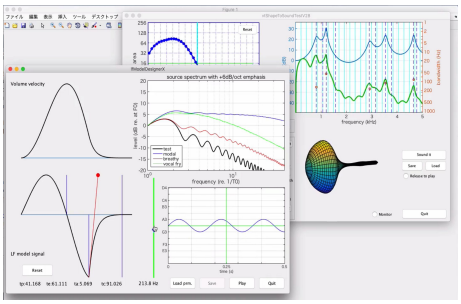

Figure 2: Source design GUI and the main GUI. The manipulation of the source parameter modifies the plots of the main GUI simultaneously.

The shape keeps slowly panning back and forth to enhance three dimensional perception.

The top right plot shows the frequency domain representations. The blue thick line represent the vocal tract transfer function of the designed vocal tract shape. The green solid thin line represents the composite spectral envelope of the vocal tract transfer function and the glottal source with radiation from the opening. The thin cyan vertical lines represent harmonic frequency locations. The dashed black lines represent the line spectrum frequencies. The red circles in the plot show poles calculated from the vocal tract transfer function. These poles can be interactively manipulated and corresponding changes of the transfer function and the vocal tract shape are displayed simultaneously. The red line simultaneously displays the power spectrum of the played back sound.

2.2. Source design GUI

Figure 2 shows a screen shot of the source signal design process. The foreground panel shows the source design GUI. This GUI uses the aliasing-free implementation of the L-F model [3] to generate the actual signal.

The upper left plot represents the volume velocity at the glottis. The lower left plot represents the L-F model signal, the differentiated volume velocity. Knobs on this lower left plot enable interactive manipulatio of the L-F model parameters.

The upper right plot shows the equalized spectral envelope.

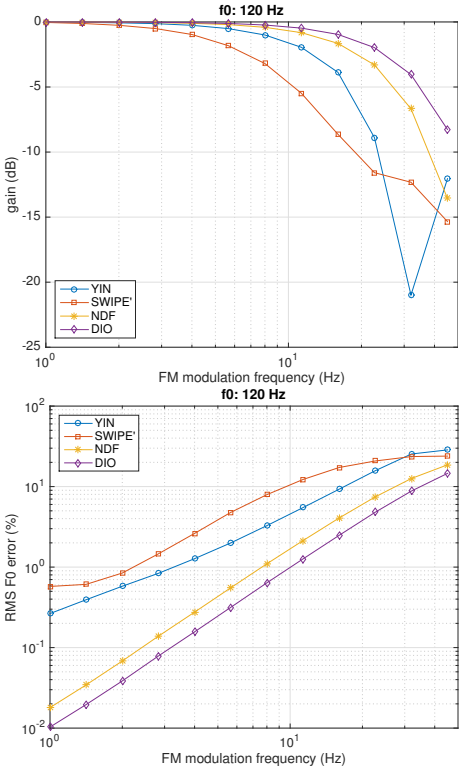

Figure 3: Application of the aliasing-free L-F model for F0 extractor evaluation.

(Nominal -6 dB/oct slope of the L-F model spectrum is equalized.) The black line represents the envelope of the designed signal. The other color lines represent characteristics of typical voice quality (modal, fly and breathy [7]). Manipulation of the L-F model parameters is simultaneously updates this spectral envelope and the composite spectral envelope in the main GUI, which is shown in the background.

The lower right panel is for the control of F0 trajectory. The vertical green slider at the left side of the plot controls the mean value of F0. The vertical thick green line in the plot is the control knob of the signal duration. Similarly to the L-F model parameter control, manipulation results simultaneously update frequency characteristics display in the background main GUI.

3. Other tools and constituent functions

In addition to this speech production simulator, realtime FFT analyser and interactive spectrogram with time-frequency region playback, realtime auditory spectrogram display and realtime vocal tract shape visualizer are prepared. The MATLAB source code of constituent functions called from these applications are also publicly available under open source license. They provide a set of useful building blocks for implementing speech processing applications. The aliasing-free L-F model [3] is the most useful and original contribution.

3.1. Example application of the aliasing-free L-F model

The aliasing-free L-F model is defined on the continuous time axis and represented in a closed form equation. The parameters of the L-F model can be updated each cycle and the F0 trajectory can be defined also on the continuous time axis. The spurious level other than harmonic component is attenuated more than 120 dB around the fundamental component. These make this model output as an ideal signal for testing tracking ability of F0 extractors. This model is useful to extend a new framework for profiling F0 extractors [8].

Figure 3 shows example results using frequency modulated F0 trajectories. In this figure, YIN [9], SWIPE' [10], NDF [11] and DIO [12] were tested. The frequency modulation depth was set 100 cent peak to peak. The upper plot shows magnitude response to the FM modulation frequency. The lower plot shows relative RMS error of the true F0 trajectory and the extracted F0 trajectories. YIN and SWIPE' introduced strong nonlinear distortion and magnitude attenuation in the extracted trajectories. A new F0 extraction framework [13] used this model to evaluate its tracking ability. Using a temporally variable lattice filter in the constituent functions also makes quantitative analysis of the time varying group delay effects.

4. Conclusions

A set of interactive tools for studying fundamentals of speech production, perception and processing are introduced. The constituent functions are mainly classical ones and the aliasing-free L-F model is our original contribution, in terms of signal processing. All these are open sourced hoping for them to be useful for beginners as well as experienced tutors and researchers.

5. Acknowledgements

This work was partly supported by Kakenhi (Grant in Aids for Scientific Research) B 15H02726 and 16K12464 of JSPS.

6. References

[1] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960.

[2] J. L. Kelly and C. C. Lochbaum, "Speech synthesis," *4th International Congress on Acoustics*, p. G42, 1962.

[3] H. Kawahara, K.-I. Sakakibara, H. Banno, M. Morise, T. Toda, and T. Irino, "Aliasing-free implementation of discrete-time glottal source models and their applications to speech synthesis and F0 extractor evaluation," in *APSIPA 2015*, Hong Kong, 2015, pp. 520–529.

[4] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech.*, vol. 4, pp. 1–13, 1985.

[5] H. Kawahara, "Matlab realtime speech tools and voice production tools." [Online]. Available: http://www.wakayama-u.ac.jp/%7ekawahara/MatlabRealtimeSpeechTools/

[6] S. Sagayama and F. Itakura, "Symmetry between linear predictive coding and composite sinusoidal modeling," *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, vol. 85, no. 6, pp. 42–54, 2002.

[7] D. G. Childers and C. Ahn, "Modeling the glottal volumevelocity waveform for three voice types," *JASA*, vol. 97, no. 1, pp. 505–519, 1995.

[8] M. Morise and H. Kawahara, "TUSK: A framework for overviewing the performance of F0 estimators," in *Interspeech 2016*, San Francisco, 2016, [Accepted].

[9] A. de Chevengné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *JASA*, vol. 111, no. 4, pp. 1917–1930, 2002.

[10] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *JASA*, vol. 124, no. 3, pp. 1638–1652, 2008.

[11] H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," in *Interspeech 2005*, Lisbon, 2005, pp. 537–540.

[12] M. Morise, H. Kawahara, and N. Nshiura, "Rapid F0 estimation for high-SNR speech based on fundamental component extraction," *Trans. IEICEJ*, vol. J93-d, no. 2, pp. 109–117, 2010, [in Japanese].

[13] H. Kawahara, Y. Agiomyrgiannakis, and H. Zen, "Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis," in *ISCA SSW9*, San Francisco, 2016, [Submitted]. [Online]. Available: http://arxiv.org/abs/1605.07809