Phonation modes synthesis using voice source-filter models

MUMT 618 Final Project

Ziyue Piao

Music Technology, McGill University
Ziyue.piao@mail.mcgill.ca

Abstract

This project aims to explore and compare different measurements for singing tension in voice synthesis. By implementing the Liljencrants-Fant model for synthesizing the glottal source, we can explore tension-related parameters. We then add a vocal tract filter to the LF model and synthesize voice samples. By comparing the synthesized results at two levels of tension parameters, we find that the high-level shape parameter R_d can be used as a tension measurement.

1. Introduction

Singing is a popular and enjoyable activity for many people, but it is challenging to learn because it relies on subtle muscle movements. One common problem that singers of all levels may face is incorrect phonation, particularly hyper-functional voice. This project uses a glottal flow model to synthesize voice with different tension levels.

1.1. Singing Phonation Modes

Besides pitch and vocal loudness, the phonation mode is a third important dimension to measure the voice source from the extreme of hyperfunction to hypofunction phonation. According to Sundberg (Sundberg, 1995)'s definition, the four phonation modes are pressed, neutral(modal), flow, and breathy. The four phonation modes are defined by subglottal pressure and glottal airflow: Neutral and breathy phonations involve less subglottal pressure than pressed and flow phonations, and neutral and pressed phonations to have lower glottal airflow than breathy and flow phonations.

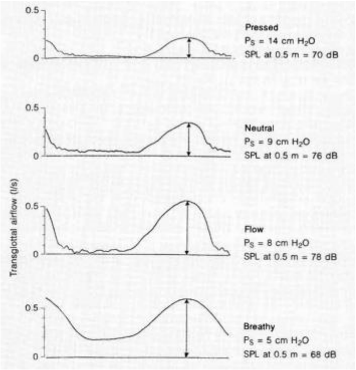

Figure 1: *Effects of phonation types on voice source waveform (Sundberg, 1995).*

1.2. Glottal Flow Model (GFM)

The basic acoustic theory of speech production is that voice involves a vocal fold vibration source and a vocal tract filtering process, known as the source-filter theory. Fant (Fant, 1970) assumes independence and linearity between the airflow modulated in the glottis by vocal folds vibration, called glottal flow, and the resonance effect of the vocal tract that shapes the glottal flow into a speech signal.

Many aspects of voice source, such as breathiness, tenseness, and vocal force, are directly related to voice quality perception. As a result, glottal flow modeling is widely used in research on expressive speech, including speech analysis, synthesis, and perception. Glottal flow waveforms can be obtained using inverse filtering or indirect measurements such as electroglottography. To match glottal filter impulse responses with glottal flow waveforms, many glottal flow models are proposed which are defined in the time domain by analytic and parametric formulations of the glottal flow waveform and its derivative (Perrotin et al., 2021). Liljencrants–Fant (LF) model

(Fant et al., 1985) is one of the most widely used glottal flow models for the analysis and synthesis of speech signals.

Two phases are considered in GFMs: first is the open phase which means lung pressure forces the vocal folds to spread, an increasing airflow passes through the glottis, and the elasticity of the vocal folds takes over, closing the air passage; the second is the closed phase which means the airflow is blocked (Perrotin et al., 2021).

2. Methodology

2.1. Liljencrants–Fant (LF) Model

The project focuses on using the LF model to generate voice source with various tension, so the first step is to implement the LF model with different levels of parameters.

The LF model is initially a four-parameters time domain model that aims to provide a reasonable approximation to actual flow conditions. The model separately measures the open phase and closing phase. The following parameters decide on the final shape of the glottal waveform:

- the length of one glottal pulse t_c, depends on the fundamental frequency f_0
- the maximum airflow is reached at t_p with a flow derivative of 0
- at t_e the vocal folds collide with the maximum energy excitation and the flow derivative is E_e
- the return phase time interval t_a

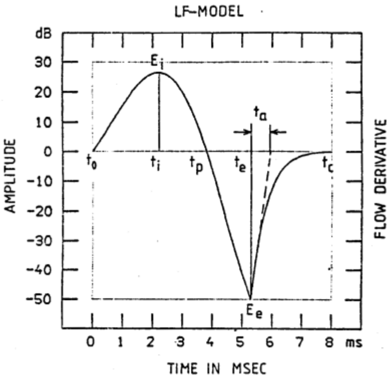

Figure 2: *Typical representation of the glottal pulse (Fant et al., 1985)*

Thus, in the open phase and close phase, the glottal pulse can be expressed as:

$$E(t) = \begin{cases} E_0 e^{\alpha t} sin\omega_g t & \text{for } 0 < \mathrm{t} \leq \mathrm{t_e} \\ \dfrac{-E_e}{\varepsilon t_a}\left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}\right] & \text{for } \mathrm{t_e} < \mathrm{t} < \mathrm{t_c} \end{cases}$$

The determining mathematical relationship is that the minimal airflow gain in a glottis period is zero, which means there is no air leak, thus

$$\int_0^{t_c} E(t) = 0$$

And the ε can be determined by t_a:

$$\begin{cases} \text{when } t_a \text{ is small,} & \varepsilon t_a = 1 \\ \text{other } t_a, & \varepsilon t_a = 1 - e^{-\varepsilon(t_c-t_e)} \end{cases}$$

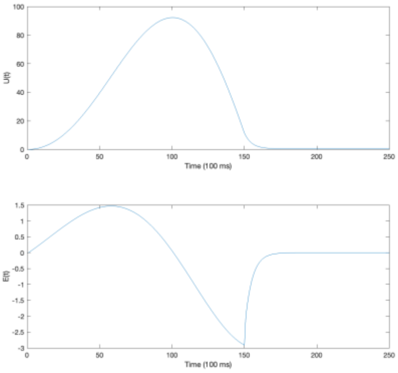

Figure 3: *The implementation version of LF model. The initial waveform audio is generated by using a wavetable. The audio sample named LF.wav.*

To use higher level parameters to separately represent different glottal phases, four new intermediate shape parameters are specified (Fant, 1995). The return phase accounts for the degree of spectral tilt cut-off frequency through $F_a = \frac{1}{2\pi T_a}$ and the $F_g = \frac{1}{2T_p}$ is the glottal formant. The initial four shape parameters can be represented by the shape parameters:

- Return phase: $R_a = \frac{T_a}{T_c}$
- Close phase: $R_k = \frac{T_e}{T_p} - 1$
- Open phase: $R_g = \frac{T_c}{2T_p} = \frac{F_g}{F_0}$

- The open $quotient$ is $O_q = \frac{T_e}{T_c} = \frac{1+R_k}{2R_g}$

Intermediate shape parameters (Fant, 1995) R_a, R_k and R_g can be qualified to a single shape parameter $R_d = T_d \frac{F_0}{110}$, which is close related to the pulse declination time $T_d = \frac{U_0}{E_e}$ of the close phase:

- $R_a = \frac{-1+4.8R_d}{100}$
- $R_k = \frac{22.4+11.8R_d}{100}$
- $R_g = \frac{R_k(0.5+1.2R_k)}{0.44R_d - 4R_a(0.5+1.2R_k)}$

2.2. Pressed-related Characteristics

After implementing the LF model, we design three LF functions which can be controlled by low level shape parameters, intermediate parameters and the R_d -parameter.

When increasing the LF model parameters to higher level, the parameters can be connected more with real perceptual variables. Intermediate shape parameter R_a is to set the spectral tilt frequency F_a, and an increase of R_a can also occur in breathy phonation. And an increase of R_g will lead to the shortening of the glottal open interval and result in a pressed voice (Fant, 1995).

In theory, the R_d parameter can be more close to a tension parameter of glottal source(Perrotin et al., 2021). Low values of R_d lead to higher center frequency and bandwidth of the glottal formant and a higher spectral tilt frequency. The combined effect is typical for tensed and loud voice when the vocal folds open and close abruptly. Inversely, high values of Rd lower the center frequency and bandwidth of the glottal formant as well as the spectral tilt cut-off frequency. This is soft voice when the vocal folds oscillate more symmetrically. So, R_d may directly control the tension level in phonation.

2.3. Vocal Tract Synthesizer

To make the synthesized sound closer to human voice, we add a vocal tract filter after the glottal source (Gold & Rabiner, 1968). A digital formant is a resonant network based on the dynamics of a second-order linear difference equation and a serial chain of digital formants can approximate the vocal tract during vowel production. The vocal tract filter can specify 3 formants in its all-pole transfer function.

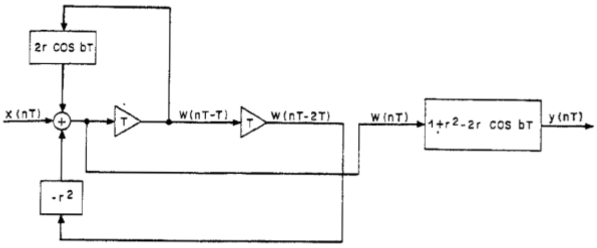

Figure 4: *Second digital network of a single formant (Gold & Rabiner, 1968)*

2.4. The System Implementation

The source-filter voice synthesizer can generate voice by controlling the:

- Generating vowels
- Fundamental frequency
- Using different levels of LF model parameters and can manually input parameters or use the preset phonation mode parameters of breathy, modal, and pressed

3. Result and Discussion

We generate three phonation modes of different pressed level as table 1 shows and the breathy, modal, and pressed modes are used to represent hypofunction, appropriate and hyperfunction phonations. The breathy and pressed modes are two relatively extreme setting of the two levels of LF parameters. The modal mode is a middle status between breathy and pressed. Finally, we aim at comparing the two parameter-tuning methods and its perception result in generating different phonation modes.

In general, the three phonation modes can result in similar glottal shapes and next we will introduce the detailed differences between phonation modes. The code and audio examples can be found in https://github.com/piaoziyue/LF_voice_synthesizer.git. The codes partly refers to the repository generation codes (Alku et al., 2019). All samples are generated below the fundamental frequency of 100 Hz and the vowel of ‘ae’.

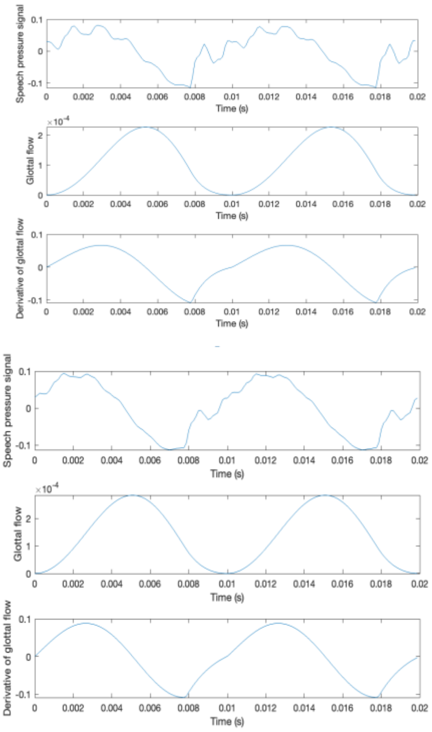

Figure 5, 6: *Breathy mode. The first figure is tuning by mid-level LF parameters, and the second figure is tuning by high-level LF parameters.*

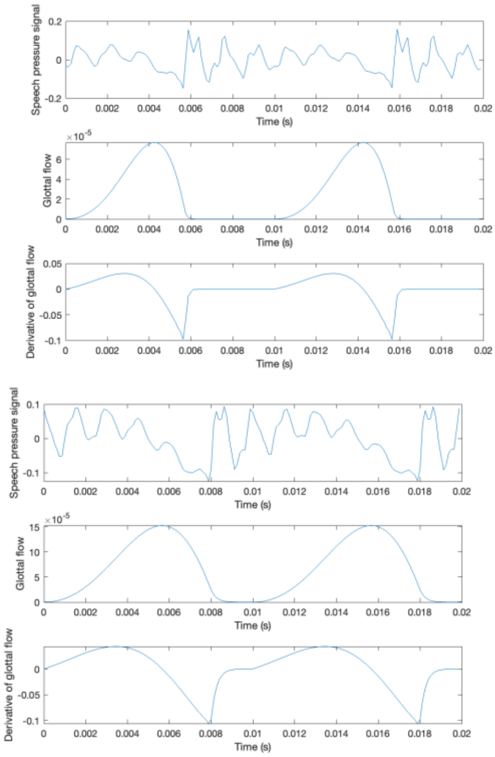

Figure 7, 8: *Modal mode. The first figure is tuning by mid-level LF parameters, and the second figure is tuning by high-level LF parameters.*

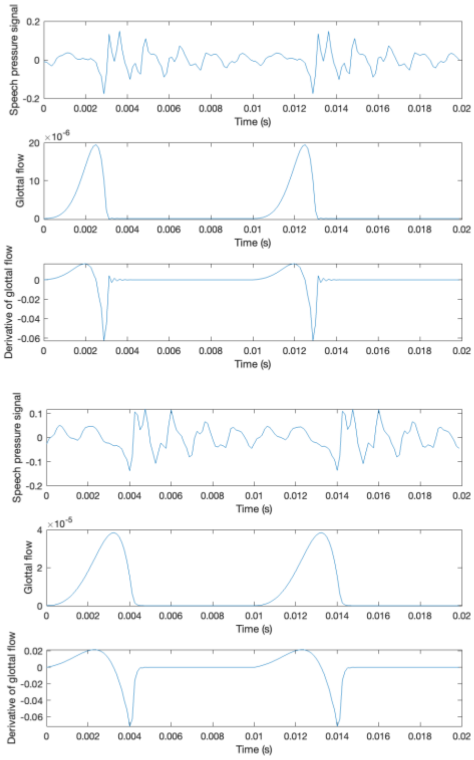

Figure 9, 10: *Pressed mode. The first figure is tuning by mid-level LF parameters, and the second figure is tuning by high-level LF parameters.*

From the audio samples, we can easily percept the different tension levels in breathy, modal, and pressed samples. Although the high-level method only uses one parameter to tune the shape of the glottal shape, the result of two levels of parameters can be close in three phonation modes. As the figures show, the longer the open phase is and the more symmetry the glottal flow is, the more relaxed the final synthesis voice will be. The modal mode is a reference sound to be compared with breathy and pressed voice, so the result of the two levels of parameters is not as close as the other two modes. The mid-level parameters can reach more extreme breathy and pressed sound, but this method needs to adjust two or three of the variables simultaneously, which is more complex to interact with. Thus, R_d can be used as a good controllable measurement of the phonation tension.

The range of the R_d is about 0.4 to 2.7 and the larger the R_d is, the more pressed the synthesis voice will get.

4. Conclusion and Future Work

The system produces a method to generate voice with different phonation modes. After comparing the synthesized effect of varying levels of parameters, we found the high-level parameter R_d can represent the tension level in voice synthesis. The next step is how to use the model in real-time voice detection which can enable to use of the model in more scenarios. The future goal of the project is to allow singers to monitor the source of their voices without invasive devices and improve their singing skills.

There are three limitations of the existing LF model. First, the complexity of the computation of the LF model restricts to use in real-time synthesizer and analysis. Second, the LF model has great potential in phonation mode detection and can be used in phonation monitor and singing learning. To achieve this goal and apply the above models, we need to evaluate an effective glottal inverse filtering method. Third, voice tension is a relatively roughness concept without a clear definition, and there may need more acoustic experiments about human phonation so that we can know how to specify the tension parameters with glottis synthesis models.

5. References

Alku, P., Murtola, T., Malinen, J., Kuortti, J., Story, B., Airaksinen, M., Salmi, M., Vilkman, E., & Geneid, A. (2019). OPENGLOT–An open environment for the evaluation of glottal inverse filtering. *Speech Communication, 107*, 38–47.

Fant, G. (1970). *Acoustic theory of speech production* (Issue 2). Walter de Gruyter.

Fant, G. (1995). *The LF-model revisited. Transformations and frequency domain analysis*. 40.

Fant, G., Liljencrants, J., & Lin, Q. (1985). *A four-parameter model of glottal flow*. 15.

Gold, B., & Rabiner, L. (1968). Analysis of digital and analog formant synthesizers. *IEEE Transactions on Audio and Electroacoustics*, *16*(1), 81–94. https://doi.org/10.1109/TAU.1968.1161954

Perrotin, O., Feugère, L., & d'Alessandro, C. (2021). Perceptual equivalence of the Liljencrants–Fant and linear-filter glottal flow models. *The Journal of the Acoustical Society of America*, *150*(2), 1273–1285. https://doi.org/10.1121/10.0005879

Sundberg, J. (1995). Vocal Fold Vibration Patterns and Modes of Phonation. *Folia Phoniatrica et Logopaedica*, *47*(4), 218–228. https://doi.org/10.1159/000266353