Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Tolg: A classical and neural hybrid LF glottal vocoder for generative source-filter speech synthesis and modification

Xiao Zhang

September 25, 2023

Thesis for the M.Phil. Speech and Language Processing
Phonetics and Speech Laboratory
School of Linguistic, Speech and Communication Sciences
Trinity College Dublin

Declaration

I declare that this dissertation has not been submitted as an exercise for a degree at this or any other university and that it is entirely my own work.

I agree that the Library may lend or copy this dissertation on request.

Signed: _____ Date: August 30, 2023

Total words: 13205

Abstract

Can we utilize neural networks with parametric glottal models to generate the excitation and speech signal? In light of current generative speech trends with the popular LF (Liljencrants-Fant) parametric glottal model, this can be a critical issue. The high-quality and versatile excitation generation synthesizer with an easily controllable LF speech vocoder certainly can bring a new era for the voice quality modification, which can be used in speech-related applications and scientific research.

The aim of this research is to address this problem by proposing a novel hybrid LF vocoder that is compatible with both classic and deep learning methods, called "Tolg," which bridges the gap between parametric vocoders and neural network frameworks. Tolg's development, based on voice quality research, includes glottal acoustics parameters R_d that can be used to generate neural glottal excitation signals. LF excitation signals generated by the classical Tolg glottal vocoder can be linear and highly controllable, as well as the neurally predicted approach from the neural Tolg. With this work, we will be able to develop future copy synthesis works, especially to dynamically adjust the glottal parameters to ensure seamless speech synthesis and precise voice quality modification tasks.

Keywords: Source filter vocoding, speech modification, speech synthesize, LF speech modeling, glottal vocoder, neural glottal vocoder, generative speech

Acknowledgements

It is amazing to see how life unfolds, and speech processing has brought me to a country where I never expected to complete this master's degree.

I would like to express my deepest gratitude to my main supervisor, Dr. Christer Gobl, who introduced me to the source filter theory, and Dr. Andy Murphy, and Zihan Wang for their help. Their assistance and feedback have been essential throughout the whole process. My co-supervisor was Dr. Andy Murphy, whose technical support and advice throughout the development of the Tolg vocoder were much appreciated.

I also would like to express my gratitude to all the academic staff who have taught me over the past year. Without them, this work would be challenging.

Also, I would like to thank Dr. John Kane and Dr. Lauri Juvela, whose open works in voice glottal analysis and synthesize provided me with many inspirations. A special thanks to the GrabVoice team. We are extremely fortunate to have the opportunity to integrate innovative speech ideas with funding that can be applied to real-life applications.

Contents

List of Figures

List of Tables

1 | Introduction

1.1 Motivation

Vocoder, derived from merging "voice" and "coding," represents a technique employed by early researchers in the field of speech analysis and synthesis. However, since then, it has evolved into a broad category that encompasses a range of algorithms that extend beyond voice coding and speech synthesizer applications. Vocoder-based systems can produce and modify speech waveforms using various acoustic features, such as spectrum envelopes or prosodic information, such as voice quality parameters. These are used for speech synthesis and modification tasks, predicting or generating the required waveform samples with the designed effects.

From its long history, the vocoder is a crucial interval component to focusing on extracting acoustic features from waveforms to synthesize speech, which can be an essential part inside of speech synthesis. Speech synthesis, also known as text-to-speech (TTS) technology, constitutes one of the subfields of artificial intelligence. TTS takes user text as input, and it involves converting text into natural and fluent speech using specific algorithms. The objective of TTS systems is to map the textual input to the corresponding waveforms. To achieve these goals, the process typically involves two main components: an acoustic model responsible for predicting spectrograms, which are sequences of Mel spectral values over time from textual input, and a subsequent vocoder, which generates the speech waveform from the speech acoustic representation. To solve the problem of vocoding for TTS, speech production models have adopted source-filter configurations and are widely used within vocoders; In accordance

with the speech production model of Fant (Fant, 1970), vocalized speech signals go through a convolutional process that involves the modulation of the glottal source by the vocal tract's transfer function. Two types of excitation signals featuring pulse train and white noise are prepared and mixed according to voiced and unvoiced information, and then filters defined by the spectrum envelope are applied to generate synthetic speech. Followed by these motivations, over the years, research in source-filter and glottal vocoders has evolved from rule-based synthesizers to waveform concatenation and statistical parameter-based approaches, as shown in Figure 1.1, leading to breakthrough improvements in the naturalness and quality of synthesized speech. Furthermore, with the development of the latest machine learning as shown in Figure 1.2, recent advances in generative neural network models, exemplified by the WaveNet neural vocoder (Oord et al., 2017) which is incorporated in Tacotron (Y. Wang et al., 2017), have significantly improved the state of the art in TTS synthesis. These models also hold promise as statistical vocoders to generate machine learning-based speech waveforms. With the need for higher-quality speech synthesizers, more demands for vocoders have constantly been proposed, which makes this field change dynamically, especially during the developments of machine learning.

The latest trend followed by the end-to-end (E2E) strategy, which is to map directly from text to speech without the vocoder in a fully E2E manner, brings to the challenge of the development of vocoders for speech synthesis. The fully E2E method was proposed and it efficiently generates high-quality speech waveform samples without the needs of the vocoders, which in some ways seems to declare the "ending" of the vocoders in the TTS task. At the same time, the newly proposed universal speech generation model in 2023 based on a large-scale generative approach such as Voicebox (Le et al., 2023) also accelerates this trend.

However, there is an increasing demand to reduce the complexity of the model while preserving the quality of the synthesis in combination with the latest and traditional vocoding strategies. Significantly, computing the digital waveform for glottal representations involves less complexity compared to acoustic waveforms, which is a compelling motivation to adopt neural models. For example, the novel approach is based on deep neural networks (DNN) for the generation of glottal excitation (Juvela et al., 2016), involving the decomposition of speech

signals into glottal source signals and vocal tract filters through the glottal inverse filtering technique (GIF). Additionally, in various end-user applications, advanced speech and vocoder technologies enable users to discern and calibrate speech quality and prosodic effects. An example of the latest system, the GlórCáil system (Murphy et al., 2021) was developed to allow users to fine-tune prosodic parameters and the transformation of speaker characteristics in synthesis. It enables synthetic utterances with various voice qualities (Murphy, 2021) using the transformed LF glottal parametic model (Fant, 1995; Fant, Liljencrants, & Lin, 1995), which characterizes differentiated glottal flow with parameters R_d, which has been found to have covariate voice source parameters based on source-filter theory. The resultant waveform can be effectively represented by parametric models, allowing straightforward speech modifications. These technologies find applications in various domains, such as speech anti-spoofing, speaker verification, speech pathology, and language education, that E2E speech synthesis cannot directly give to low-resource languages. Speech synthesizers based on vocoders, especially those integrated with LF glottal modeling that can be parametrized and controlled, and neural networks for speech quality modification, will continue to play an indispensable role for speech synthesis and modification in the long future.

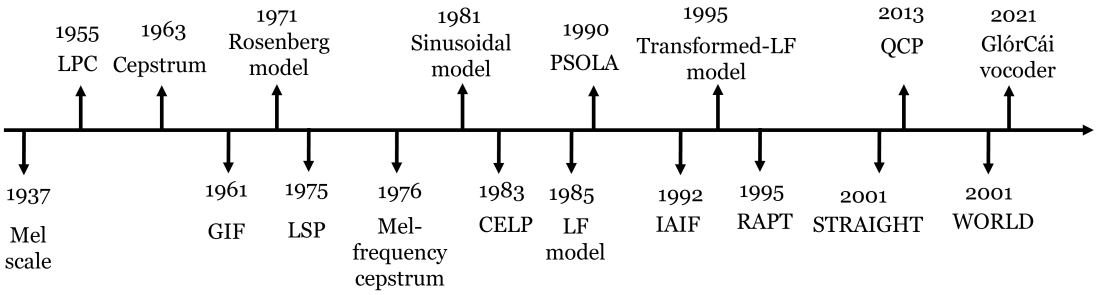

Figure 1.1: A timeline of some of the major developments in parametric vocoding.

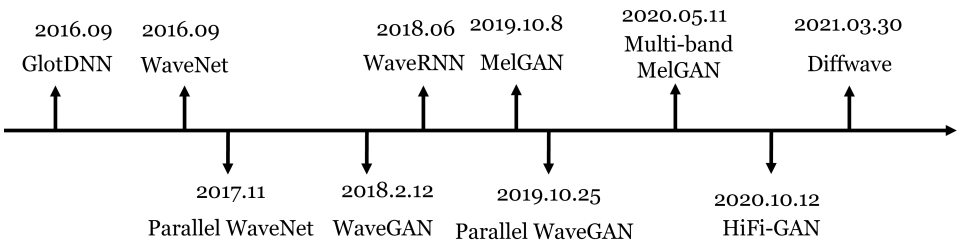

Figure 1.2: A timeline of some of the major developments in neural vocoding.

1.2 Aims and objectives

The focus of this research is to tackle the challenging task of proposing an LF vocoder with both classic and deep learning model called "Tolg" (which is the inverse word of "Glot"), which can lead to the primary objective of this dissertation being to address the following question: Can DNNs be used effectively to generate the LF glottal source signal?

The primary contributions of this study are as follows.

(1) The first establishment of an LF-DNN vocoder named "Tolg" was designed to construct a linear and DNN E2E architecture for LF parameter-to-glottal pulse conversion, with a specific emphasis on voice quality modification tasks.

(2) Implementation of instant LF pulse and generation modification features with a global scaling factor R_d that replaces the previous complicated program setups gives the potential for highly efficient real-time speech processing and transformation applications in the future.

1.3 Structure of the document

There are two types of non-neural vocoders including source-filter and glottal vocoder which will be discussed here; they differ mainly in the addition of the glottal inverse filtering processing; in other ways, the glottal vocoder gives a deeper investigation of the source signals using the parametric method.

This thesis is structured as follows.

Chapter 2 provides an overview of non-neural vocoders, particularly at the source-filter-based vocoding strategy. This chapter establishes the foundations for the analysis and synthesis methods discussed for the vocoding task.

Chapter 3 focuses on glottal models, specifically the LF glottal vocoder. It is with a particular emphasis on acoustic glottal modeling, and glottal vocoding, that inheritable works are

obtained from Chapter 2.

Chapter 4 provides an exhaustive examination of the methodologies implemented in neural vocoders, serving as a foundation for the in-depth exploration of key deep learning paradigms that unfolds subsequently in Chapter 5.

Chapter 5 elaborates to propose a neural network vocoder, named "Tolg," which integrates LF modeling to the neural glottal vocoder. It also illustrates the effectiveness of the proposed Tolg vocoder in the context of modification of R_d; prediction of neural speech synthesis, respectively.

Chapter 6 presents and analyzes the conclusions drawn from the research and outlines potential avenues for future work.

2 | Source-filter vocoders

Source-filter vocoders are traditional vocoders based on speech modeling theory that do not involve neural networks to generate speech waveforms from the input of phonetic or acoustic features.

Throughout the advancement of speech synthesis technology, initial attempts largely adopted expert knowledge-based approaches for vocoding tasks and mainly based on the source-filter theory. Prior laboratory phonetics research, such as the source-filter model, played a critical role in early attempts. In rule-based speech synthesis, phonetic units were manually crafted according to specific rules to improve intelligibility, as an example called Pink Trombone [1] is shown in Figure 2.1. During this time, Klatt introduced a serial/parallel formant synthesizer (Klatt, 1980), with the aim of addressing intelligibility problems. Based on these works, an additional approach to speech synthesis emerged in the 1990s, known as corpus-based concatenative synthesis. This technique involved concatenating speech units from a database, either in acoustic feature or waveform. There were two primary synthesis methods used: diphone synthesis with a single inventory and unit selection synthesis with multiple inventories. However, that approach had significant limitations. By the late 1990s, corpus-based statistical synthesis techniques had exploded. That approach combined the source-filter model with a statistical acoustic model, employing Hidden Markov Models (HMMs) starting in 1995 and DNNs from 2013. Until now, source-filter theory still remains fundamental for processing and feature extraction, provides the materials and a deeper view of speech research, and is capable of continuing to reduce the complexity and controllability of speech.

[1] https://dood.al/pinktrombone/

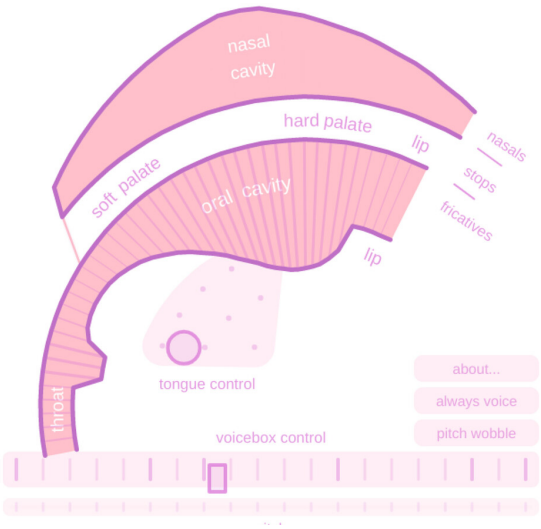

Figure 2.1: In Pink Trombone, users can manually synthesize speech by modifying vocal tract and glottis parameters (from Thapen, 2023).

The Source-Filter paradigm serves as an ideal model in which speech articulation is conceptualized as the consequence of transmitting a glottal excitation signal, often referred to as the "source," through a temporally dynamic linear filter responsible for encapsulating the vocal tract's inherent resonant attributes. Vocoder implementations based on source-filter-based architectures have gained considerable traction and are widely acknowledged as some of the most effective methodologies for the parameterization, modification, and subsequent reconstruction of speech waveforms. During speech production, the main excitation comes from variations in acoustic pressure caused by glottal vibrations. To generate speech, the vocal tract system filters the excitation signal. Numerous research endeavors have been undertaken to enhance the characteristics of the excitation signal with the objective of attaining superior harmonic resolutions, thereby contributing to an enriched auditory experience. This includes, for example, STRAIGHT (Kawahara, Estill, & Fujimura, 2001; Kawahara et al., 2008), and code-excited linear prediction (CELP) in mel-cepstrum vocoding. Similarly, a family of speech representations called sinusoidal models also proposes an alternative representation of the excitation signal. In all, the source-filter model has played a fundamental role in the development of glottal vocoders, especially in terms of modeling and controllability.

2.1 Linear predictive coding vocoder

The most naive source-filter system is based on linear prediction (LP) modeling. In that case, the spectral characteristics are frequently modeled through a time-variant AR filter, which is excited either by quasi-periodic pulses in the case of voiced segments or by a stochastic noise source during unvoiced speech segments. Classical source-filter vocoding employs linear predictive coding (LPC) analysis to derive linear predictive coefficients together with source-filter modeling. This strategy utilizes pitch extraction to obtain pitch-related information from pulse train signals, and an all-pole digital filter is applied to the input to produce the filter output. For the future autoregressive (AR) neural glottal vocoder, the AR strategy still provides the fundamental concept. As depicted in Figure 2.2, the Linear Predictive Coding (LPC) model encapsulates the spectral attributes of the source spectrum, the vocal tract transfer function, and the radiation characteristics by employing a time-variant digital filter. A complete speech waveform is generated by concatenating the processed frames using LPC coefficients extracted from previous frames and then predicting future speech waveforms based on these coefficients. To create a windowed data sequence, we convert the extracted frame (10 to 30 milliseconds) into a data sequence, then window the data sequence. To determine the transfer function of the time domain filter, we compute optimal LPC coefficients or autoregressive coefficients for each segment.

The nearby speech samples are found to exhibit a high degree of correlation shown in Figure 2.3. During the early history of LPC, it was a direct or naive way of producing speech using source-filters. In 1955, Norbert Wiener introduced a seminal framework for predictive coding of signals, grounded in the mathematical theories he had previously developed in the 1940s (Atal, 2006). These theories were aimed at optimizing filters and predictors for the explicit purpose of identifying signals obscured by stochastic noise (Elias, 1955a, 1955b). Subsequently, in 1966, Fumitada Itakura and Shuzo Saito made a notable advancement by formulating a statistical methodology to estimate the spectral density of speech signals. Intriguingly, their approach, based on maximum likelihood estimation (MLE), was conceived independently of the research efforts conducted at Bell Labs on predictive coding (Saito &

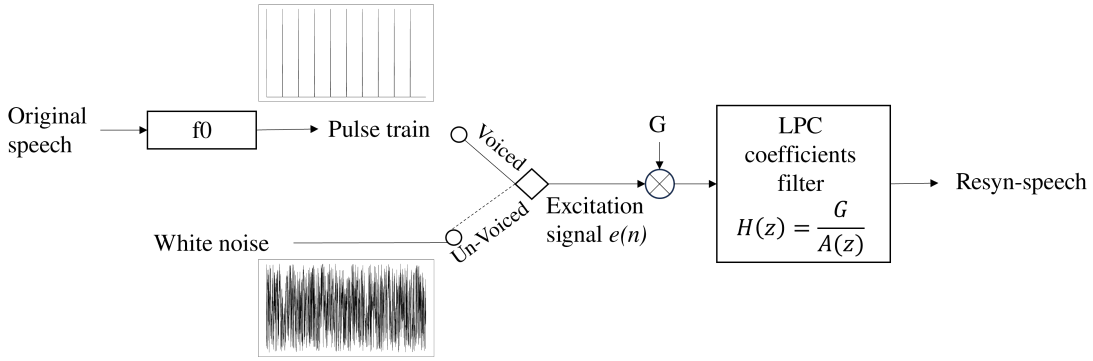

Figure 2.2: Block diagram of a LPC vocoder.

Itakura, 1967; Itakura & Saito, 1970).

The mathematical representations are as follows in an AR approach shown in equations (1-6). The basic principle is that it can predict the waveform sample at time n by taking the weighted sum of past waveform samples plus the error noise signal or the LPC residual e.

$$s(n) = \underbrace{\sum_{k=1}^{p} \alpha_k s(n-k)}_{\text{Linear predictor}} + \underbrace{e(n)}_{\text{Prediction error}} \tag{1}$$

$$s(n) = a_1 s(n-1) + \cdots + a_p s(n-p) + e(n) \tag{2}$$

Based on the autoregressive approach, the LPC vocoder utilizes the linear weighted sum of past k speech waveform samples. a_k is called predictor, auto regressive or predictor coefficient computed by minimizing the mean squared prediction error (MSE) $e[n]$ or energy parameter E_n with mainly two different approaches - auto-correlation or covariance methods (Rabiner & Schafer, 1978).

Using i as a frame index, we can use those LPC coefficients to perform waveform conversions independently for each frame. To estimate the LPC coefficients for the future, we calculate and minimize the MSEs of the autoregressive coefficients.

$$\widehat{a}_{1:P} = (\widehat{a}_1, \cdots, \widehat{a}_P) = \underset{a_{1:P}}{\operatorname{argmin}} e_n^2 = \underset{a_{1:P}}{\operatorname{argmin}} \left(o_n - \sum_{k=1}^{P} a_k o_{n-k} \right)^2 \tag{3}$$

The target here is to minimize the energy of the residual or loss function, where $e[n]$ represents

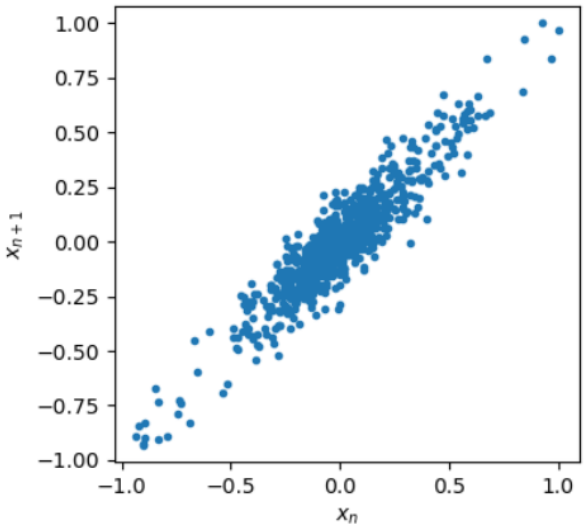

Figure 2.3: Scatter plot of subequent samples x_n and x_{n+1} (from Bäckström et al., 2022).

the prediction error, which is also known as the LPC residual. Depending on the modeling of the formant analysis task, the order of the LPC synthesis filter p may vary. In other words, the difference between the waveform granted at time n and the waveform predicted at time n, which is the sum of p weighted by some assumptions.

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \tag{4}$$

The sequence representing the prediction error emanates from a system characterized by the following transfer function:

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{5}$$

Consequently, the prediction error filter, denoted as $A(z)$, serves as an inverse filter with respect to the system function $H(z)$.

$$H(z) = \frac{G}{A(z)} \tag{6}$$

Generally, Figure 2.2 illustrates a block diagram for a LPC vocoder, the parameters governing the LPC model can be systematically categorized into four primary elements: the voiced/unvoiced speech classification, the pitch period specifically pertaining to voiced speech,

the gain parameter denoted by G, and the set of LPC coefficients represented as (a_k). The pitch parameter conveys information about the glottis, the gain value pertains to the overall vocal amplitude, and the LSPs (Line spectral pairs) provide information about the vocal tract spectral characteristics. The subsequent sections will facilitate a comprehensive introduction to these constituent elements.

The LPC concept is remarkably straightforward and uses prediction to achieve efficient signal coding, thereby reducing speech latency. Additionally, this linear predictive framework serves as a foundational construct for elucidating the fundamental principles embedded within MLE, which are then deployed massively on neural networks. The latest, but similar, neural network-based LPC vocoder replaces the AR strategy with a probabilistic AR neural network. Although it shares extreme simplicity, it lacks the resolution of speech details, especially the harmonic structure. The reason for this is that when the vocal folds are closed, the airflow ceases, but when they are opened, the airflow resumes. However, it also opens gradually, differing from the naive stimulus pulse suggested by the LPC in the pulse generation section. Accordingly, it is linked to the Mel-cepstral vocoder in in Section 2.2.

2.1.1 Fundamental frequency

The estimation of the fundamental frequency, denoted as F_0, and commonly referred to as pitch estimation or pitch tracking, constitutes a critical challenge in speech signal processing. This problem is of paramount importance, given its far-reaching applications in various subdomains.

Traditional approaches to pitch estimation predominantly leverage audio signal processing techniques to ascertain the periodicity embedded within the individual frames of the speech signal. It is viable to analyze F_0 in segmented speech waveforms in three different domains - the frequency, time, and cepstrum domains - using autocorrelation or linear prediction techniques. Traditional estimation strategies use autocorrelation in the time domain to determine F_0. In contemporary speech synthesis paradigms, some of the prevailing methodologies also employ various autocorrelation technique, such as Robust Algorithm for

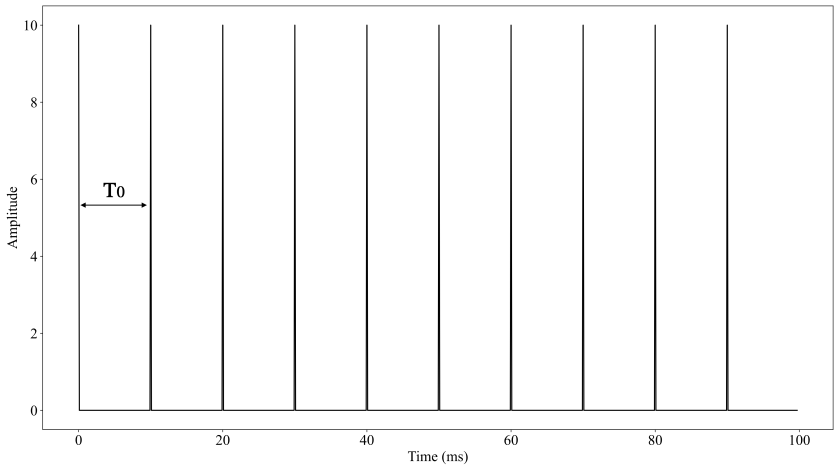

Figure 2.4: An impulse train for f_0 at 100 Hz.

Pitch Tracking (RAPT) (Talkin, 1995), YIN (de Cheveigné & Kawahara, 2002), and the Sawtooth Waveform Inspired Pitch Estimator (SWIPE) which was introduced in 2008 (Camacho & Harris, 2008). Approaches operating in the frequency domain frequently capitalize on the energy of residual harmonics resulting from linear predictive analysis; an illustrative example being the Spectral Root Harmonic (SRH) technique (Drugman & Alwan, 2011). In addition, extensive academic attention has been devoted to the estimation of the fundamental frequency (F_0) within the cepstrum domain (Noll, 1967). Methods based on cepstral analysis perform a separation of the excitation and vocal tract information by means of a homomorphic transformation; within this framework, the fundamental period is ascertained as the interval corresponding to the first salient peak in the cepstrum. The fundamental frequency constitutes an essential acoustic feature in the landscape of speech analysis.

In this example as shown in Figure 2.4, pulse segments are interpolated at a constant sampling rate F_s of 16 kHz for a framming duration of 25 ms (400 samples). The period (T) is the duration of a complete pulse vibration, and the frequency is inversely proportional to it.

$$T_0 = \frac{1}{f_0} \tag{7}$$

Fundamental frequency serves as an indispensable component across a broad spectrum of domains, encompassing pitch perception associated with intonational paradigms, phonetic functions—with particular prominence in tonal languages—vocal fold vibrational mechanics,

and models of periodic signal representations.

2.1.2 Line spectral pairs

Line Spectrum Pairs (LSPs) also referred to as Line Spectral Frequencies (LSFs) are an alternative LPC spectral representation of speech frames that are perceptually meaningful in coding systems. Historically, Itakura introduced the LSP as an alternative to the LPC spectral representation (Itakura, 1975). To address the issues of filter stability and representational efficiency with less space, linear prediction-based coders typically use line spectral pairs (LSPs) to represent LPCs (McLoughlin, 2008; Kim, Kim, & Hahn, 2006). Originating from the LPC domain, these LSPs offer a robust representation of the vocal tract resonances within the context of analyzed speech signals. They have quickly become popular in speech research and are now widely used for speech-coding applications.

Speech parameters include the LP filter of the vocal tract, which is converted to LSFs.

$$A_p(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_p z^{-p} \tag{8}$$

It can be stated as follows:

$$A_p(z) = \frac{P(z) + Q(z)}{2} \tag{9}$$

A symmetric polynomial $Q(z)$ and an antisymmetric polynomial $P(z)$ related to $A_p(z)$ are formed by adding and subtracting the system function in time.

$$P(z) = A_p(z) - z^{-(p+1)} A_p\left(z^{-1}\right) \tag{10}$$

$$Q(z) = A_p(z) + z^{-(p+1)} A_p\left(z^{-1}\right) \tag{11}$$

The roots of the two auxiliary polynomials serve as the determinants for LSFs, thereby constituting the mathematical basis for spectral analysis in the given context.

$$H(z) = \frac{1}{A_p(z)} \tag{12}$$

Based on the transfer function of the all-pole LPC filter, here, which transmits pseudo-stationary vocal tract spectral information using a set of LSPs as in Equation (11).

2.1.3 Glottal closure instants

Glottal Closure Instants (GCIs), also referred to as epochs, are defined as quasiperiodic temporal instances in which significant excitation occurs in the glottal source, formally recognized as the instant of closure of the glottal source (Drugman et al., 2012). Within the context of concatenative speech synthesis methodologies, it has been acknowledged since the nascent stages of TTS technology that a comprehensive understanding of a reference instant is imperative for the mitigation of concatenative discontinuities. Thus, GCIs have been incorporated into the widely-acclaimed Time-Domain Pitch-Synchronous Overlap-Add (TD-PSOLA) algorithm (Moulines & Charpentier, 1990), as well as a mechanism for removing phase mismatches (Stylianou, 2001). The location of the GCI is of particular significance during the analysis stage, so robust GCI algorithms are of crucial importance.

In speech signal processing, GCIs are defined as temporally precise instances characterized by substantial excitation within the vocal tract. These discrete temporal events occur with periods of elevated energy in the glottal waveform, specifically during segments of voiced speech. Several innovative approaches have been proposed to detect GCI automatically, using a range of signal processing techniques. As an example, using a simple peak pick algorithm, GCIs are detected from the differentiated glottal flow signal. Another notable finding is the use of a residual excitation and mean-based signal (SEDREAMS) algorithm to detect speech events. This method uses the mean-based signal to identify short intervals where GCIs are likely to occur, and then assigns a more accurate estimation of the location of the GCI based on the LP residual (Drugman & Dutoit, 2009; Drugman, 2011). More importantly, with SEDREAMS, GCIs are now capable of being extracted from signals produced by different types of

phonation, which includes a variety of voice qualities, such as creaky, breathy, tense, harsh, and falsetto voices (Kane & Gobl, 2013b).

2.1.4 Pitch-synchronous overlap-add

The Pitch-Synchronous Overlap-Add (PSOLA) technique, originally delineated in seminal works (Moulines & Charpentier, 1990; Moulines & Laroche, 1995), serves as an effective methodology for the implementation of overlap addition, pitch modification, and temporal scaling in speech signal processing.

Historically, the waveform concatenation-based synthesis approach, which performs the straightforward task of concatenating pre-existing waveforms in a designated database to produce a contiguous speech stream. In concatenative synthesis, two predominant paradigms exist: one LPC as we discussed earlier and the other, PSOLA. In the LPC-based paradigm, the speech signal undergoes LPC coding to achieve data compression, thereby minimizing storage requirements. The ensuing synthesis is executed via a rudimentary decoding and concatenation process. While this mechanism yields an audibly natural synthesis for isolated lexical items, its limitations become conspicuously evident in continuous speech. Unlike isolated words, natural speech embodies a more complex temporal and spectral structure, rendering the naive concatenation of individual speech units suboptimal, particularly when the continuity between successive frames is considered. This phenomenon is exemplified in Figure 2.5, where a typical overlap-add procedure is schematically illustrated.

In contrast, PSOLA improves these deficiencies by placing a greater focus on the control and modification of prosody. This is achieved by adaptively modulating the prosodic characteristics of each concatenative unit in alignment with the target contextual environment. Such an approach distinguishes PSOLA from its LPC-based counterpart, facilitating the generation of a synthesized waveform that not only preserves the innate quality of the original phonetic elements, but also adapts the prosodic attributes of the concatenated units to suit the targeted context.

The advent of PSOLA has had a profound impact on the landscape of voice source-vocoder

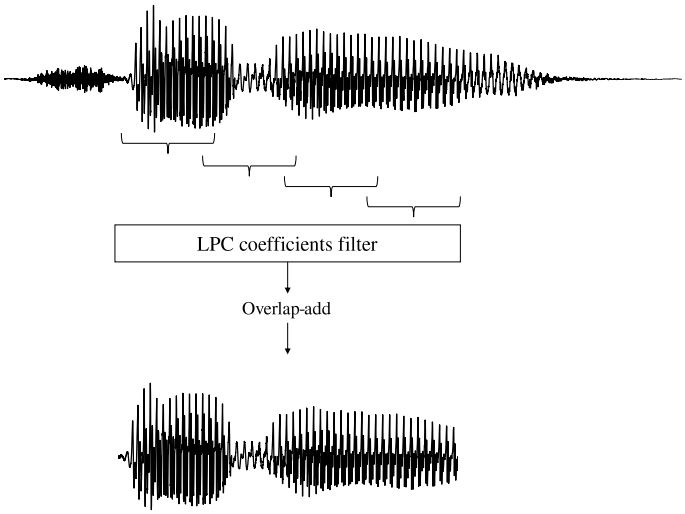

Figure 2.5: Block diagram of over-lap add.

research, particularly at the synthetic phase. Originating in 1990, PSOLA (Moulines & Charpentier, 1990) significantly improved the fidelity and naturalness of speech waveforms generated by time-domain concatenative synthesis techniques. The algorithm serves as a foundational framework for pitch-synchronous analysis, as well as for the modeling and synthesis of speech waveforms.

The latest work such as (Drugman, Wilfart, & Dutoit, 2009; Drugman & Dutoit, 2012) has extended the application of Principal Component Analysis (PCA) to pitch-synchronous glottal flow signals, offering a more nuanced representation of the excitation waveform. Subsequently, the construction of a pitch-synchronous residual and glottal flow codebook has been reported in studies such as (Drugman et al., 2009) and (Raitio et al., 2011), facilitating the selection of optimal glottal pulses.

Contemporary research efforts have ventured into the extraction, interpolation, and storage of pitch-synchronous glottal flow time-domain waveforms within a specialized codebook. Employing machine learning algorithms, these waveforms are predicted directly in the time domain (Raitio et al., 2014). Spanning from its inception to the most recent advancements, PSOLA has demonstrated its efficacy as a robust algorithmic approach, particularly when implemented within frame-level synthesis paradigms.

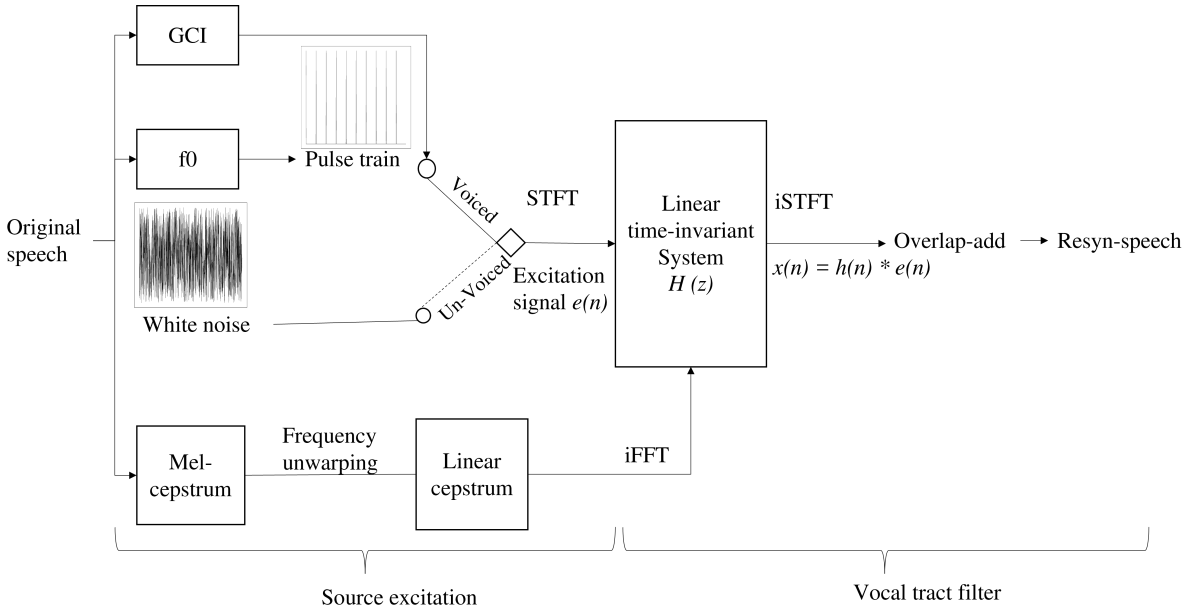

Figure 2.6: Block diagram of a Mel-cepstral vocoder.

2.2 Mel-cepstral vocoder

The Mel-cepstral vocoder is a source-filter model based vocoder that includes mixed excitations i.e. not only white noise but also pulse trains with harmonic structures that approximate the glottal spectrum. Using source-filter decomposition, we approximate the vocal fold to a pulse and the turbulence to random noise. In general, cepstral analysis-synthesis involves convolution of the vocal tract filter response with a complex impulse excitation signal, which is equivalent to a homomorphic decomposition, as shown in Figure 2.6.

2.2.1 Mel scale

The Mel scale is a psychoacoustic scale to describe pitch perception, the unit of measurement is "Mel" (Stevens, Volkmann, & Newman, 1937).

The pitch concept differs from the frequency value; it requires auditory scaling. Figure 2.7 shows a relationship between linear frequency and two Mel scale approximations. One Mel scale approximation is the Technical Mel scale (Fant, 1968).

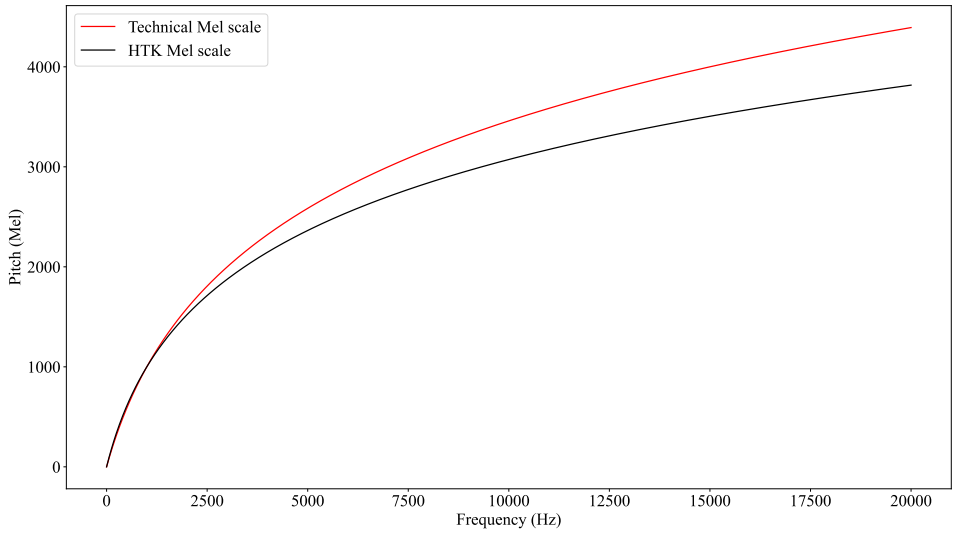

Figure 2.7: Mel scale approximations.

$$m_{Tech}(f) = \frac{1000}{\log 2} \log\left(1 + \frac{f}{1000}\right) \tag{13}$$

There is also a popular linear-to-Mel frequency formula in (O'Shaughnessy, 1987) for the HTK speech recognition toolkit [2]:

$$m_{HTK}(f) = 2595 \log\left(1 + \frac{f}{700}\right) \tag{14}$$

In terms of frequency, Mel scales are roughly linear at low frequencies and logarithmic at higher frequencies.

2.2.2 Cepstrum

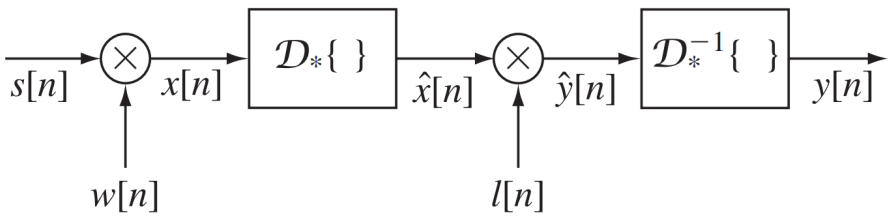

Figure 2.8: Flowchart of the cepstrum analysis (from Rabiner & Schafer, 2010).

[2] HTK toolkit website: https://htk.eng.cam.ac.uk/

The cepstrum is defined as the inverse Fourier Transform of the logarithmic magnitude spectrum, as shown in equation (15). The discrete ceptrum is obtained from the inverse Discrete Fourier Transform (iDFT) of the logarithmic magnitude spectrum obtained from the Discrete Fourier Transform (DFT). The term cepstrum was derived by reversing the first four letters of spectrum. The domain is called quefrency, the term derived by reversing the order of fre and que in frequency. Filtering in the quefrency domain is sometimes referred to as liftering. The spectrum envelope is basically created after computing the DFT, which converts time-domain signals into frequency domain. It has been posited that the utilization of logarithmic scaling in conjunction with the Mel frequency scale furnishes a more effective representation of the spectral envelope of speech signals. This claim is motivated by the logarithmic relationship between auditory perception and acoustic production mechanisms within the human auditory system, which requires the incorporation of a specialized logarithmic function for optimal signal representation (Taylor, 2009). It is recalled that the aim of the vocoding process will be to invert the log-Mel spectrum representation into a time-domain waveform representation since the output spectrum displays both features of the glottal source and the vocal tract filter.

$$C(m) = \mathcal{F}^{-1}\left[\log\left|S\left(e^{j\omega}\right)\right|^2\right] \tag{15}$$

In this way, we may compute the spectrum envelope like the cepstrum and then convert the cepstrum to LPCs. At the same time, the pitch parameter can also be obtained using a peak selection algorithm for the higher-quefrency cepstrum. Overall, the cepstrum transform, in conjunction with the upcoming Mel scale filtering, offers low processing pressure while improving the quantization of the coefficient, ultimately reducing the complexity of speech analysis.

2.2.3 Mel-frequency cepstral coefficients

Mel Frequency Cepstral Coefficients (MFCCs) constitute an integral component of the Mel frequency cepstrum, which is instrumental in the transformation of cepstral information into

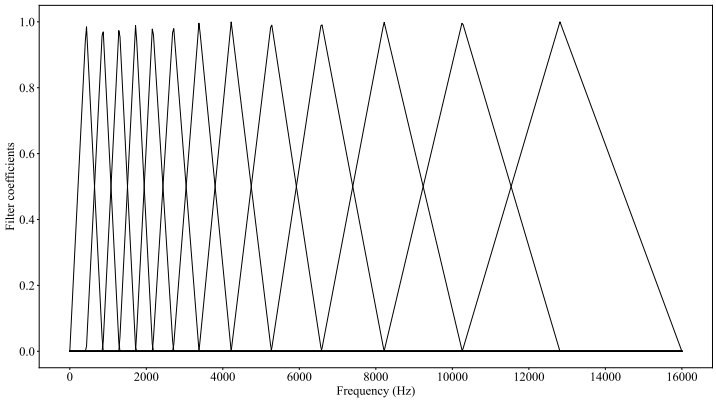

Figure 2.9: Mel filterbank.

the Mel scale via a Mel filterbank, as shown in Equation (16). Within auditory signal processing, the Mel-frequency cepstral transform serves as a sophisticated mathematical framework for the computation of the short-term power spectrum of auditory signals through the extraction of cepstral coefficients. The computational process of this transformation takes place in a sequence of discretized steps. Advanced adaptations of this technique encompass procedures such as normalization of triangular-shaped Finite Impulse Response (FIR) filters, as depicted in Figure 2.9, to a unitary area, as well as the application of the Mel filterbank to the amplitude of the power spectrum. Such modifications put a heightened emphasis on the lower end of the frequency spectrum. Consequently, frequencies proximal to the center of the spectrum confer greater weightage compared to those that are more peripherally situated.

$$\mathrm{MFCC}(x) = \mathbf{DCT}\{\log(\mathbf{Mel}|\mathrm{FFT}(x)|)\} \tag{16}$$

In further applications, Mel-log spectrum approximation (MLSA) synthesis filters combine impulse and noise components, which are controlled by Mel-cepstral coefficients. Utilizing a Mel-cepstral Maximum Likelihood Spectral Approximation (MLSA) filter results in a notable decrease in computational burden, which is critical for real-time applications, when placed with the traditional Fast Fourier Transform (FFT)-based processing integral to the conventional STRAIGHT methodology (Kawahara et al., 2001). Subsequently, these filter frames are combined via the Pitch-Synchronous Overlap-Add (PSOLA) technique, thereby producing the final synthetic speech waveform.

2.3 Sinusoidal vocoder

Sinusoidal modeling can be considered as an evolutionary development of the Short-Time Fourier Transform (STFT) and phase vocoder methodologies (McAulay & Quatieri, 1986). This evolution is primarily driven by the inherent limitations and insufficiencies observed in the STFT approach, which necessitates the formulation of a more generalized and parametric sinusoidal model. Such a model effectively serves as an extension of the STFT and increases its capabilities. The inspiration for the incorporation of a sinusoidal source module comes from earlier seminal works that focus on the generation of sine-wave-based excitation signals with pre-specified fundamental frequencies (Hedelin, 1981). Subsequent empirical studies and evaluations have determined that sinusoidal models exhibit unparalleled efficiency in the accurate representation of voiced speech phenomena (McAulay & Quatieri, 1986).

Sinusoidal modeling represents voiced and unvoiced speech frames as sums of sinusoids, so that the speech signal can also be represented as a sum of K sinusoids:

$$s(n) = \sum_{k=1}^{K} A_k \exp\left(j\left(2\pi k f_0 n + \theta_k\right)\right) \tag{17}$$

As shown in Equation (17), where A_k, f_0, and θ_k denote the amplitude, frequency, and phase of the k th sinusoids. Speech signals are particularly well modeled by harmonically related sinusoids, which makes it easy to apply various modifications to the signal while still maintaining a low mean-squared error.

The challenge lies in the precise determination of time-varying parameters, namely amplitude, frequency, and phase, for a series of sinusoids, such that their algebraic summation faithfully approximates the temporal behavior of the original signal under study. As a result of the periodic nature of glottal excitation, the speech spectrum exhibits a harmonic structure. Therefore, when analyzing voiced speech signals in a short time window (roughly three periods), the DFT will reveal integer multiples of the fundamental frequency in the amplitude spectrum. In voiced speech signals, the spectral peaks serve as repositories for the most perceptually salient components, thus playing a crucial role in human auditory perception of

the spoken utterance.

2.4 Summary

Until now, we have discussed recent advances in source-filter models based vocoders. Considering the development of such models, it is quite reasonable to conduct research on the source signals. Can we parameterize the glottal signals into a mathematical model in order to better understand the source signals? The question here leads to how we use the glottal signals after they have been parametricated. This will be the core question that leads from the source-filter vocoder to the next phase of the study, the LF glottal model.

3 | LF glottal model

The glottal model fundamentally integrates AR models with specialized parametric glottal representations, aligning itself with the source-filter paradigm of speech production. Using the technique of glottal inverse filtering, one can disentangle the original speech signals into constituent elements: the glottal source signals and the corresponding vocal tract filters. We computationally estimate glottal source signals by negating the inherent contributions of the vocal tract transfer function.

Glottal vocoders differ from source-filter vocoders in that they are able to extract and parameterize the excitation signal so that they can control the voice signals further. A physiologically plausible glottal waveform is extracted by glottal inverse filtering (GIF) (Alku, 1992) and parameterized to reflect glottal characteristics (Raitio et al., 2011).

A variety of glottal models have been proposed, including Rosenberg (Rosenberg, 1971), Liljencrants-Fant (LF) (Fant, Liljencrants, & Lin, 1985), Fujisaki-Ljungqvist (FL) (Fujisaki & Ljungqvist, 1986) and Rosenberg++ (R++) (Veldhuis, 1998). In speech prosody, we are particularly interested in controlling voice qualities such as creaky and tense voice. To control excitation signals or excitation functions, we use popular LF models (Fant, Liljencrants, & Lin, 1985).

3.1 Glottal inverse filtering

Glottal Inverse Filtering (GIF) is a technique specifically engineered for the acoustic analysis of the human vocal production apparatus. This methodological approach involves the

construction of a computational representation of the acoustic filtering characteristics imposed by both the vocal tract and lip radiation. Subsequently, the influence of these physiological components is computationally nullified through the application of inverse filtering operations predicated on the aforementioned model. GIF is generally incorporated in a synergistic fashion with the parameterization of the approximated glottal excitation, serving as a descriptive metric for the intricacies involved in the human voice production mechanism.

In the late 1950s, Miller pioneered an approach for estimating glottal excitation by processing the recorded speech signal through specialized hardware, effectively nullifying the vocal tract characteristics. Importantly, his work concentrated solely on the first formant (F_1) as a representative feature of the vocal tract (Miller, 1959). Subsequently, in the early 1960s, Fant and Lindqvist-Gauffin advanced the field by spearheading a comprehensive series of studies focusing on GIF (Fant, 1961; Fant & Sonesson, 1962).

Progress was further augmented in the mid to late 1970s with the advent of the Closed Phase (CP) covariance technique (Strube, 1974; Wong, Markel, & Gray, 1979). This innovative method used all-pole digital models derived from linear prediction and based on the covariance method to model the vocal tract (Rabiner & Schafer, 1978). Remarkably, thus enhancing the precision of the inverse vocal tract model.

In 1992, Alku introduced the Iterative Adaptive Inverse Filtering (IAIF) methodology, an automated but straightforward approach to GIF (Alku, 1992). IAIF employs an iterative algorithm designed to segregate the glottal flow and the spectral envelopes of the vocal tract, an innovation that subsequently evolved into the more advanced GFM-IAIF approach (Perrotin & McLoughlin, 2019).

The most recent advancement in this domain is the Quasi-Closed Phase (QCP) analysis, an algorithm that extrapolates the principles of CP analysis to identify the characteristics of the vocal tract during the closed phase of the glottal flow (Airaksinen, Story, & Alku, 2013). Unlike traditional methods relying on linear regression, QCP adopts Weighted Linear Prediction (WLP) to conduct multi-pitch period analysis, with a particular focus on the samples acquired during the closed phase. This innovation significantly enhances the

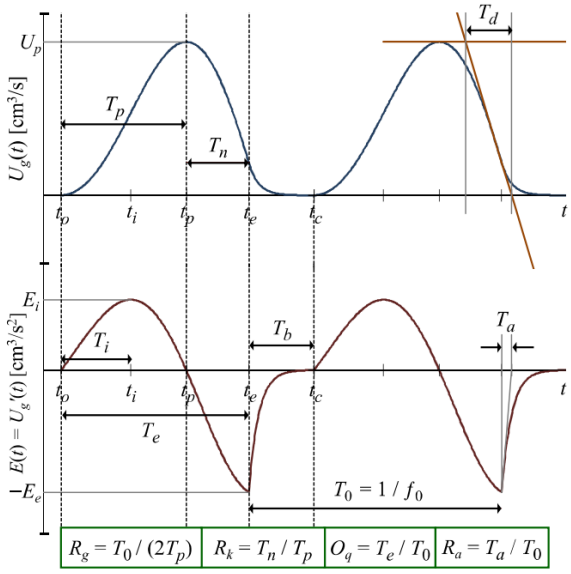

Figure 3.1: LF Glottal model pulses (top), flow derivative (bottom) (from Gobl, 2017) .

robustness of closed phase identification. Moreover, the Quasi-Closed Phase-Deep Neural Net (QCP-DNN) technique integrates a novel neural-based GIF methodology, thus optimizing the estimation of glottal flow, particularly for high-pitched speech (Juvela et al., 2016). Empirical evidence, obtained through a series of subjective listening tests, substantiates the claim that the QCP-DNN method appreciably enhances the naturalness of the resulting speech signal.

3.2 LF model

The LF model (Fant, Liljencrants, & Lin, 1985) uses a five-parameter model of differentiated glottal flow. The Equation (18) is used to determine the time derivative of the glottal pulses, which consists of one exponential part modulated by a sinusoid and an exponential part.

The parameters E_e, R_g, R_k, and R_a as shown in Figure 3.1 are referred to as the "LF parameters" and are typically used to describe model pulses.

$$U_{g'}(t) = e_{LF}(t) = \begin{cases} E_0 e^{\alpha t} \sin(w_g t), & t_o \leq t \leq t_e \text{ open-phase} \\ -\frac{EE}{\epsilon T_a}\left[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}\right], & t_e < t \leq t_c \text{ return-phase} \\ 0, & t_c < t \leq T_0 \end{cases} \tag{18}$$

The LF model is parameterized by a set of temporal landmarks, namely t_p, t_e, t_c, T_a, and T_0, together with an amplitude-based excitation strength denoted by EE. The LF model consists of two distinct phases: the open phase and the glottal return phase. The glottal opening phase is defined as the time interval during which the vocal folds remain in an open state. Conversely, the glottal returning phase refers to the temporal segment during which the vocal folds return to their closed state. The normalised glottal frequency, termed R_g, is obtained by the quotient of T_0 and twice the duration of the pulse (t_p). R_k serves as an indicator of glottal pulse symmetry and is calculated as $\frac{T_n}{T_p}$ (see Figure 3.1). Lastly, the parameter R_a quantifies the normalized duration of the return phase and is calculated as the ratio of t_a to T_0.

Despite the provision of a comprehensive set of parameters to regulate glottal excitation, ongoing scholarly discussions persist on the topic of simplifying the control mechanisms for the voice source in TTS systems and speech modification protocols, particularly when considering the transformation capabilities of a single-parameter-driven LF vocoder which will be discussed in the next section.

3.3 Transformed LF model

The transformed-LF model is a particular parameterization of the LF model proposed by Fant (Fant, 1995). The analytical framework employed here is aligned with that of the LF model, albeit with a reduced shape space governed by a singular global shape parameter, denoted as R_d. Previous work by Fant has empirically validated that this singular parameter serves as an optimal quantitative descriptor for characterizing voice quality (Fant et al., 1995). The parameter R_d serves not only as a critical determinant of voice quality but also provides a mechanism for modification of paralinguistic attributes in synthetic speech generation. Furthermore, the efficacy of R_d in modulating both linguistic prominence and prosodic variations in synthetic speech has been substantiated through rigorous experimental evaluations (Murphy et al., 2019).

The parameter R_d serves as a key parameter in the transformed LF model, encapsulating a significant portion of the intrinsic covariation exhibited among the glottal parameters. Based

on the relationship between R parameters measured on various speakers, we describe the parameter R_d as a function of a transition from tense to breathy voices. The most recent methodology for automated extraction of R_d employs the exhaustive search algorithms, dynamic programming techniques, and optimization strategies (Kane & Gobl, 2013a).

The R parameters of the glottal pulse, as we discussed before, such as R_a, R_k and R_g, can be determined from R_d. It is necessary to obtain the full set of parameters for the LF model to generate the LF glottal waveform. Using the equations presented below, we can calculate R_d:

$$R_d = (1/0.11)\,(0.5 + 1.2R_k)\,(R_k/4R_g + R_a) \tag{19}$$

From the parameter R_d, we can inversely predict the ratio parameters:

$$\begin{aligned} R_{ap} &= (-1 + 4.8Rd)/100 \\ R_{kp} &= (22.4 + 11.8Rd)/100 \\ R_{gp} &= 1/\left(4\left((0.11Rd/\left(1/2 + 1.2R_{kp}\right)\right) - R_{ap}\right)/R_{kp}\right) \end{aligned} \tag{20}$$

In the study presented by (Fant, 1995), the primary domain of R_d is explicitly defined as $[0.3; 2.7]$. However, the interval $[2.7; 5]$ is postulated to emulate scenarios related to the abduction of the vocal folds.

Using the parameter R_d, one can generate glottal pulse shapes that can range from tense to lax in terms of voice quality. Given the importance of the tense-lax dimension in prosodic modulation, the parameter R_d seemingly presents an optimal choice for reducing the requisite control parameters essential for voice quality transformation.

One of the most notable proposed transformed LF vocoders based on transformed LF modeling is GlórCáil (Murphy et al., 2021). The GlórCáil vocoder is designed for the intricate analysis and subsequent re-synthesis of speech signals. Developed within the MATLAB environment, the system emphasizes manipulation of the voice source parameters. Initially, both the characteristics of the vocal tract and the parameters of the voice source are estimated

during a dedicated analysis phase, as illustrated in Figure 3.3. Subsequent to this analytical stage, the speech waveform is re-synthesized, opting for an acoustically based glottal source model in lieu of the original glottal source. The vocoder incorporates a graphical user interface (GUI) as shown in Figure 3.2, affording end-users the ability to modify the perceptual quality of a given speech utterance.

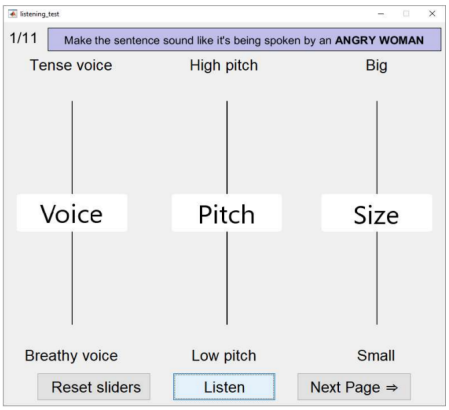

Figure 3.2: User interface for voice transformation task in GlórCáil (from Murphy et al., 2021).

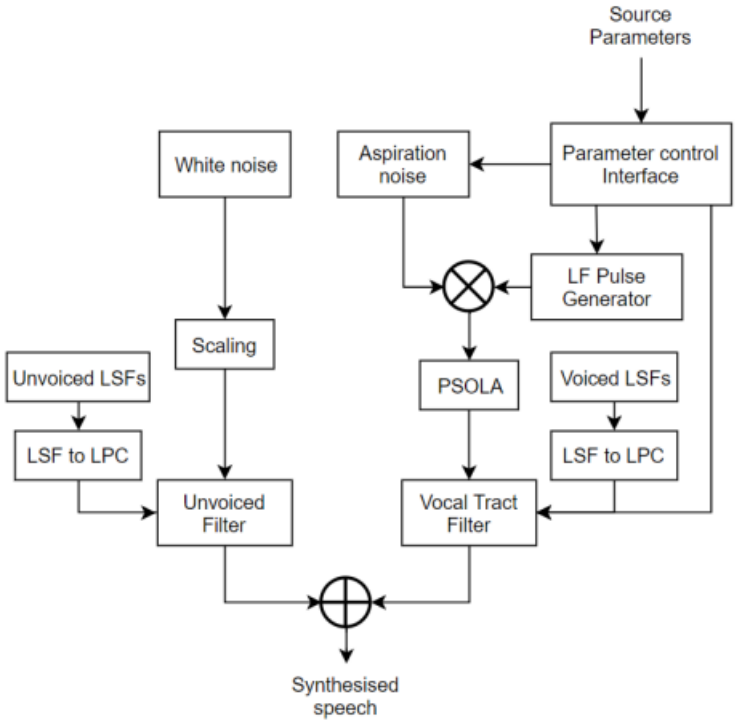

Figure 3.3: An overview of the GlórCáil analysis stage (from Murphy et al., 2021).

In general, the transformed LF vocoder provides an efficient method for modeling the glottal waveforms with the goal of reducing parameter redundancy.

3.4 Summary

In this chapter, an overview of recent advances is presented. From LF modeling in speech production theory to using LF vocoders in real-life applications. The LF modeling inherits but surpasses previous source-filter models, showing high parametric and controllability. Nevertheless, with the development of neural networks, glottal vocoders are constantly being developed, which are of higher quality than the naive source-filter vocoders described in Chapter 2. In the following chapter, we will primarily discuss neural vocoders and show how deep learning works, which provides a compelling reason for the work in Chapter 5.

4 | Neural network and vocoder

In recent years, advances in deep learning and neural network architectures have catalyzed transformative innovations in the domain of speech synthesis. These contemporary methodologies have demonstrably outperformed their traditional counterparts in terms of both synthesis quality and computational efficiency. In particular, in the context of neural vocoding, deep learning paradigms have been instrumental in enhancing the capabilities of TTS systems based on neural networks (Tan et al., 2021) as shown in Table 4.1. This has led to a marked shift in preference from traditional parametric vocoders to neural vocoding techniques, which are rapidly gaining traction within the community.

In the domain of deep learning, which refers to the training of multi-layered neural networks to approximate complex functions (LeCun, Bengio, & Hinton, 2015), a significant advancement has been made in the vocoding technologies that facilitate the transformation of log-Mel spectra or phonetic representations into time-domain waveforms. A paradigmatic example is the introduction of WaveNet (Oord et al., 2017) in 2016. This innovative technology employs an AR neural network model to directly represent speech waveforms, taking spectrograms as input and generating audio sequences represented in compressed linear Pulse-Code Modulation (PCM) values (Tokuda & Zen, 2015, 2016). In particular, WaveNet has demonstrated superior performance compared to other vocoding techniques within the architecture of TTS systems (Wang et al., 2018).

Subsequently, WaveNet has been integrated into the Tacotron framework (Wang et al., 2017) , a comprehensive neural speech synthesis toolkit designed to transform textual input into linear-scale spectrograms. Within the Tacotron system, WaveNet assumes the responsibility of

converting these linear-scale spectrograms into speech output. This integration underscores the symbiotic relationship between classical and neural vocoders, a topic that will be further elaborated upon in subsequent sections of this chapter.

However, it is crucial to delineate the marked distinctions between these two vocoding paradigms. Unlike classical approaches that are primarily rooted in mathematical derivations, neural vocoding technologies leverage probabilistic models grounded in statistical inference. Furthermore, while the Hidden Markov Model (HMM) confines itself to non-autoregressive (NAR) generation, neural networks, being inherently more flexible, are capable of accommodating both AR and NAR models. This imbues them with the capability to achieve a generative modeling approach that has recently become more nuanced and computationally efficient.

Vocoder	Input	AR/NAR	Modeling	Architecture
WaveNet	Phonetic Feature	AR	/	CNN
WaveRNN	Phonetic Feature	AR	/	RNN
Par. WaveNet	Phonetic Feature	NAR	Flow	CNN
WaveGlow	Mel-Spectrogram	NAR	Flow	Hybrid/CNN
WaveFlow	Mel-Spectrogram	AR	Flow	Hybrid/CNN
WaveGAN	/	NAR	GAN	CNN
Par. WaveGAN	Mel-Spectrogram	NAR	GAN	CNN
MelGAN	Mel-Spectrogram	NAR	GAN	CNN
HiFi-GAN	Mel-Spectrogram	NAR	GAN	Hybrid/CNN
Wave-VAE	Mel-Spectrogram	NAR	VAE	CNN
WaveGrad	Mel-Spectrogram	NAR	Diffusion	Hybrid/CNN

Table 4.1: A catalog of neural vocoders along with their associated features (adapted from Tan et al., 2021).

Neural networks, one observes a structured interconnection of artificial neurons, colloquially referred to as nodes. Each of these nodes is responsible for generating a sequence of

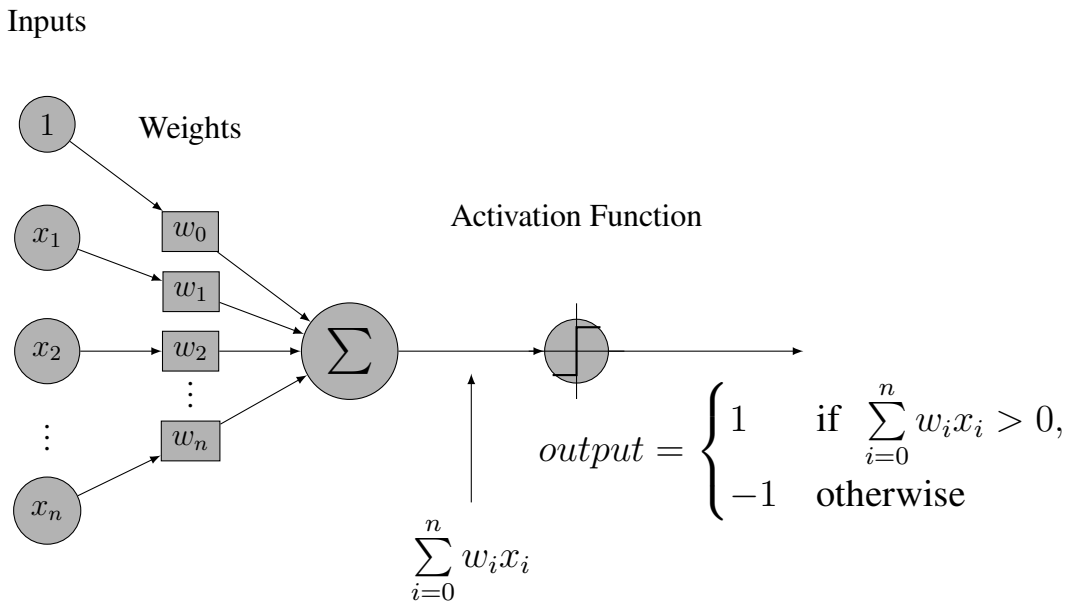

Figure 4.1: Block diagram of the perceptron model.

activations aimed at a specific target. Excluding the nodes in the input layer, every node can essentially be regarded as a distinct perceptron characterized by a unique set of weights as shown in Figure 4.1. Mathematically, a perceptron is conceptualized as a processing unit incorporating a multitude of components: weights (w), biases (b), summation functions $(\sum)$, activation function (f) and output signal (y).

Building upon this foundational understanding, there exist four cardinal constituents that collectively define a machine learning system:

(1) Data: inputs x and outputs y.

(2) Model: the function $f(x;\theta)$ to be fitted to the predicted data $\hat{y}$.

(3) Loss function: $\ell(y,\hat{y})$, which measures the discrepancy between the output of the fitted model $\hat{y}$ and the real value y of the dataset.

(4) Optimizer: $\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$, which seeks to minimize the loss to obtain the optimized θ^* parameter.

In the following sections, these core components will be elaborated individually to provide a comprehensive understanding of the neural network and their relevance to the specialized field of neural vocoding.

4.1 Data

In neural networks, the objective involves the manipulation of vectorial and matrix-valued functions, a procedure colloquially referred to as "learning" within the context of machine learning paradigms. The pivotal inquiry centers on the systematic identification of optimal parameter vectors, denoted $\boldsymbol{\theta}$, predicated on a given dataset
$X = (x_1, x_2, \ldots, x_N), Y = (y_1, y_2, \ldots, y_N)$. It is imperative to acknowledge the notational distinction where x is in vector notation, while y is in scalar notation.
In machine learning methodologies, two predominant paradigms can be delineated: supervised and unsupervised learning techniques. In supervised learning, the model is trained on a dataset that comprises both input variables and their corresponding target outputs, thereby facilitating the optimization of the predictive accuracy with respect to the annotated labels. Conversely, in unsupervised learning, the algorithm is presented solely with input variables, devoid of any accompanying target labels, thus compelling the model to discern inherent structures within the data autonomously.

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{pmatrix} = (x_1, x_2, \ldots, x_D)^\top \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_D \end{pmatrix}$$

In the data matrix shown above, each entry in the designated column corresponds to an integer, serving as an index to a one-hot encoded vector, thus facilitating the categorical representation. For those well-versed in linear algebraic operations, the scalar projection of two vectors can be conveniently described as the inner product between these vectors.

4.2 Model

In the proposed framework, the architecture of the model is predicated on a given speech input, denoted as x. A hypothesis function, herein referred to as predictor f produces a predicted value $\hat{y}$ based on the given predicted parameters θ.

$$\hat{y} = f(x; \theta) \tag{21}$$

In the context of substituting the antecedent AR model from LPC, as previously elucidated, with a machine learning-based approach, it can reformulate the original functional representation.

For the vector of regression coefficients, here:

$$\boldsymbol{a} \equiv \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} \qquad \boldsymbol{s}_p[n] \equiv \begin{pmatrix} s[n-1] \\ s[n-2] \\ \vdots \\ s[n-p] \end{pmatrix}$$

So, the original AR model can be expressed equally as follows:

$$\hat{s}[n] = \sum_{k=1}^{p} a_k s[n-k] = \langle \boldsymbol{a}, \boldsymbol{s}_p[n] \rangle \tag{22}$$

In machine learning developments, two predominant categories emerge: generative and discriminative modeling. Discriminative models strive to learn a predictor predicated on empirical observations, whereas generative models seek to determine a joint distribution that encompasses all pertinent variables. In the following sections, we shall delve into each of these paradigms independently.

4.2.1 Discriminative modeling

In deep learning, there are several discriminative learning techniques, such as artificial neural networks (ANNs), DNNs, convolutional neural networks (CNNs), and recurrent neural networks (RNNs). The discussion will follow the historical development of these models.

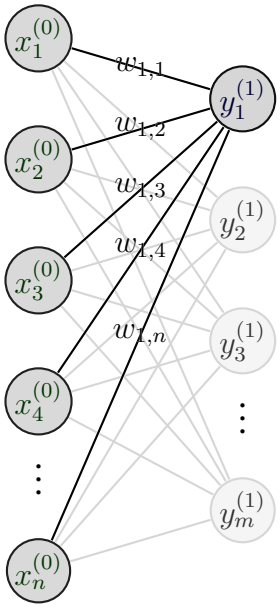

Figure 4.2: The block diagram of the artificial neural networks.

$$\hat{y}_m = \sigma\left(\sum_n W_{mn} x_n + b_m\right) \tag{23}$$

In ANNs, as shown in Figure 4.2, vertically organized structural components are commonly designated as "layers," each comprising an arbitrary number of interconnected perceptrons or neurons. Intermediate layers within this architecture are frequently categorized under the term "hidden layers," distinguishing them from the input and output layers that interact directly with the external data environment. An ANN that incorporates two or more such hidden layers between its input and output strata is formally called a DNN, as shown in Figure 4.3 (Hinton et al., 2012) and Equation (24).

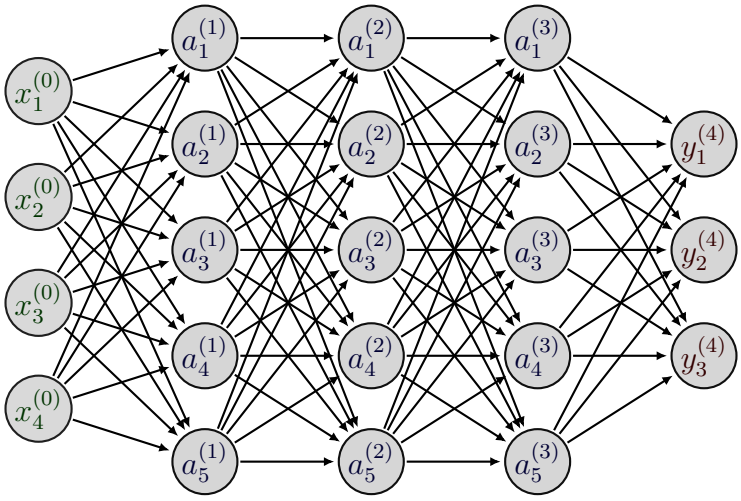

Figure 4.3: The block diagram of the deep neural network.

$$\begin{pmatrix} y_1^{(1)} \\ y_2^{(1)} \\ \vdots \\ y_m^{(1)} \end{pmatrix} = \sigma \left[\begin{pmatrix} w_{1,0} & w_{1,1} & \dots & w_{1,n} \\ w_{2,0} & w_{2,1} & \dots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,0} & w_{m,1} & \dots & w_{m,n} \end{pmatrix} \begin{pmatrix} a_1^{(0)} \\ a_2^{(0)} \\ \vdots \\ a_n^{(0)} \end{pmatrix} + \begin{pmatrix} b_1^{(0)} \\ b_2^{(0)} \\ \vdots \\ b_m^{(0)} \end{pmatrix} \right] \tag{24}$$

In the realm of deep learning architectures, Convolutional Neural Networks (CNNs) have garnered considerable attention and widespread adoption due to their distinctive advantages over traditional DNNs (Yao et al., 2019). In a CNN, the neurons in the hidden layers are intricately arranged to form sparse connections only with a subset of neurons from the preceding layer. This design choice facilitates implicit learning of important features. The element of a CNN architecture is the convolutional layer, which comprises an array of convolutional filters, colloquially referred to as kernels. The architecture of CNNs offers three main advantages: translational invariance, sparse connectivity, and parameter sharing, as elaborated in (Goodfellow et al., 2016).

Expanding on this architectural foundation, WaveNet (Oord et al., 2017) serves as a paradigmatic example of applying CNN-based architectures, or alternatively Recurrent Neural Networks (RNNs) (Oord, Kalchbrenner, & Kavukcuoglu, 2016), to the domain of image generation. WaveNet is a fully probabilistic AR model designed to forecast the likelihood

distribution of an impending audio sample conditioned on its antecedent samples. The model uses dilated causal convolutions to compute the t th sampling point only using sampling points from 0 to $t - 1$.

For pre-processing of raw audio data $y_1 \ldots y_t \ldots y_T$, the μ law is used (Lyon, 2008), where μ = 255, resulting in a transformed series $x_1 \ldots x_t \ldots x_T$. The constraint $-1 < x_t < 1$ is imposed for each t. Subsequently, each x_t is quantized into 256 values and then encoded into one-hot vector representations as shown in Figure 4.4.

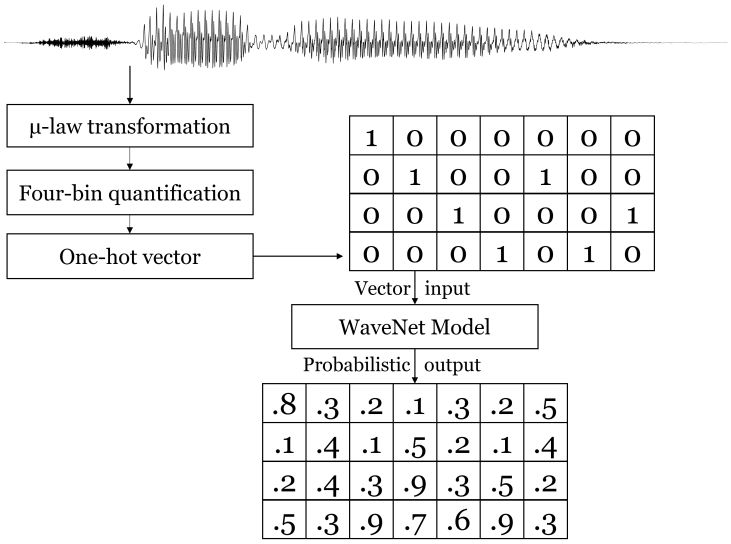

Figure 4.4: WaveNet's speech analysis stage.

$$x_t = \text{sign}\,(y_t)\,\frac{\ln\left(1 + \mu\,|y_t|\right)}{\ln(1+\mu)} \tag{25}$$

For a waveform X = {x_0, x_1, ..., x_{T-1}}, WaveNet models the conditional probability with a deep neural network,

$$P\left(x_t \mid x_0 \ldots x_{t-1}\right), \tag{26}$$

Utilizing AR connections, the WaveNet architecture sequentially synthesizes individual waveform samples, where the generation of each sample is conditional upon its immediate predecessor. The cumulative likelihood of the entire waveform can thus be mathematically formalized as follows:

$$p(X) = P\left(x_0\right) \prod_{t=1}^{T-1} p\left(x_t \mid x_0, x_1, \ldots, x_{t-1}\right) \tag{27}$$

Through probabilistic computation, a time-domain waveform can be synthesized from an intermediate representation, specifically the Mel-frequency spectrogram denoted by h.

$$p(Y) = \prod_{t=1}^{t} P\left(y_t \mid y_1, \ldots, y_{t-1}, h_1, \ldots, h_t\right) \tag{28}$$

Dilated convolutions differ from standard convolutions in that they apply the convolutional kernel across an extended input span, achieved by the strategic omission of certain input values at each progressive layer. When the dilation factor is set to 1, a convolutional filter of length 2 will sample the input variables x_t and x_{t-1} at time t. On the contrary, with a dilation factor of 2, the filter will intersperse the sampling, effectively producing a receptive field that still comprises the input values x_t and x_{t-1}. Figure 4.5 elucidates the mechanism by which the output is obtained at a given time t, employing a sequence of dilated convolution layers with dilation factors, respectively. Using dilated convolutions increases the receptive field of the network, thereby facilitating the capture of long-range dependencies without necessitating an increase in either the network depth or the convolutional filter size.

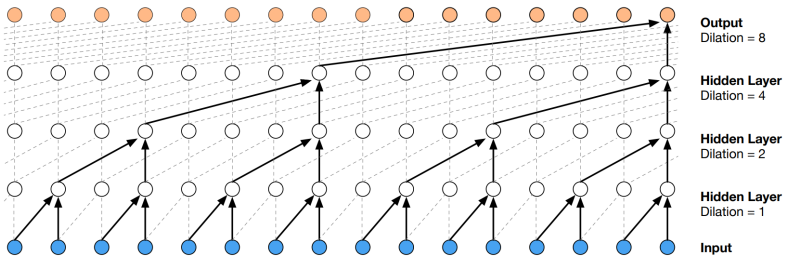

Figure 4.5: The figure shows a series of dilated causal convolutional layers with a dilation cycle size of 4, including dilation values of 1, 2, 4, and 8 (from Oord et al., 2017).

The WaveNet AR architecture, renowned for its high-fidelity audio synthesis, faces computational inefficiencies in real-time deployment due to its inherent sequential nature. Specifically, the generation of audio waveforms occurs in a temporally linear fashion, where each sample point is predicated upon its antecedents. Although recent advances, such as the introduction of parallel WaveNet inference methodologies (van den Oord et al., 2018), have

been made to these latency issues, the computational burden remains substantial for real-time systems.

Consequently, there exists a compelling rationale for investigating more streamlined, computationally efficient architectures that maintain high-quality audio output, thereby enabling large-scale deployment of WaveNet-derived models. One such avenue of exploration lies in the domain of neural glottal vocoders. Transitioning from the conventional speech waveform domain to the glottal excitation domain has been shown to permit the utilization of more compact neural architectures without compromising acoustic fidelity (Tamamori et al., 2017).

Given the rapid evolution of generative neural network frameworks, it is plausible to consider the design of glottal excitation models that leverage the computational efficiencies inherent in WaveRNN-based architectures, rather than being solely reliant on WaveNet. These endeavors could potentially culminate in the realization of real-time, high-quality, and computationally efficient voice synthesis systems.

In the advanced domain of RNNs, particular subclasses such as Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks have been extensively investigated. Subsequently, WaveRNN emerges as a distinct NAR neural architecture. Remarkably, WaveRNN is engineered to condense its architecture to a singular layer while maintaining a high degree of efficacy in capturing contextual information. This is achieved via the incorporation of a specialized recurrent layer capable of encoding the entire historical context within its hidden states.

4.2.2 Generative learning modeling

In recent developments, deep-generative models employing intricate neural network architectures have exhibited a transformative impact on improving voice synthesis quality (Tan et al., 2017). These models operate on the principle of direct probabilistic modeling of the distribution of training samples, obviating the need for supervised guidance. Such an approach facilitates an unsupervised learning paradigm, where the model ingests copious amounts of training data to self-organize and generalize underlying patterns.

Over the last few years, there has been a discernible elevation in the fidelity of the synthesized samples emanating from an array of deep generative models. Specifically, these advances have been manifested in various domains, including, but not limited to, Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models. Although the ensuing discussion will provide an overview elucidating the foundational logic governing these models, it will not cover the mathematical intricacies that underpin them.

The core idea is that for any observed x, the model will provide a probability distribution function that is close to the true data distribution, denoted by $p^*(X)$.

$$p_{\boldsymbol{\theta}}(X) \approx p^*(X) \tag{29}$$

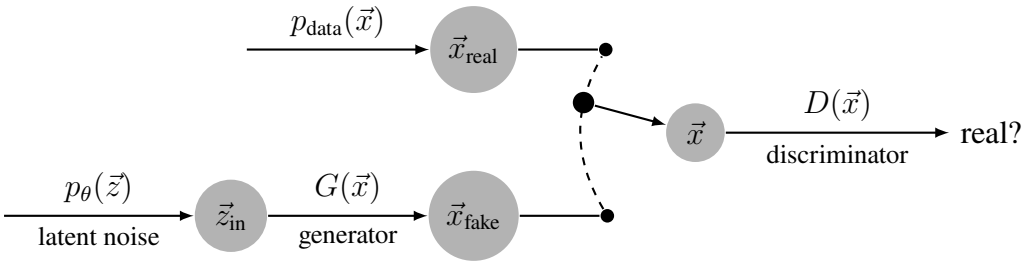

Figure 4.6: Block diagram of the diffusion model (modified from Velikovi, 2023).

The architecture of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), inherently comprises a bidirectional mapping paradigm that involves low-dimensional latent vectors and high-dimensional data spaces as shown in Figure 4.6. Within this framework, two distinct, yet interdependent models are concomitantly trained: a generative model denoted G, responsible for capturing the underlying distribution of the data, and a discriminative model, denoted D, tasked with estimating the probability that a given sample originates from the empirical training data as opposed to the generative model.

In the specific application to vocoding, the GAN model comprises two critical sub-components. The first, the generator, is engineered to synthesize speech data. The second, the discriminator, functions as an evaluator of the similarity of the generated data, serving as a criterion against which the fidelity of the generated speech data can be assessed. This dualistic interplay forms the core of GAN-based vocoding methodologies.

In contrast, Variational Autoencoders (VAEs), as explained in (Rezende et al., 2014), are made

up of two interdependent but distinct submodels: the encoder and the decoder. Within this architecture, the computational prowess of DNNs is complemented by the statistical rigor of approximate Bayesian inference, thereby engendering scalable, deep, and directed generative models. In the specialized context of WaveVAE, the encoder $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{c})$ and the decoder $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{c})$ are optimized together, where $\boldsymbol{z}$ are latent variables and c are Mel spectrogram conditioners (Peng et al., 2020). This dual-optimization schema facilitates a more cohesive and effective generative model, particularly in applications requiring complex conditional dependencies.

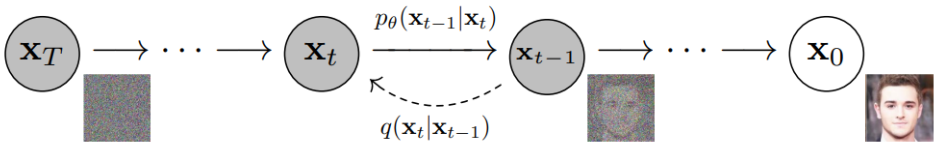

Figure 4.7: The visualization of the diffusion process (from Ho, Jain, & Abbeel, 2020).

The diffusion model, as delineated in (Sohl-Dickstein et al., 2015), employs a parameterized Markov chain to initiate a stochastic process that incrementally perturbs the data, operating in a manner antithetical to conventional sampling schemes, as shown in Figure 4.7. This methodology uses variational inference techniques to generate data samples that progressively approximate the underlying data distribution. The generative paradigm underlying this model is articulated through a bi-directional diffusion mechanism. Specifically, the forward or inference phase of diffusion transforms intricate data distributions into more tractable Gaussian distributions, thereby facilitating a simplified computational framework. Subsequently, the model undergoes a finite-time reverse diffusion process aimed at recovering the original complex data distribution. This dual mechanism achieves a parametric mapping between the data and latent variable spaces. Within this framework, diffusion processes involve a controlled infusion of stochastic noise into waveform data samples, ultimately rendering the Gaussian noise. On the other hand, the reverse processes involve the systematic removal of Gaussian noise to recover the waveform data samples. Despite its ability to generate high-fidelity speech output via diffusion-based vocoders, the model suffers from latency issues in the inference stage, attributable to the inherently iterative nature of diffusion processes (Ho, Jain, & Abbeel, 2020).

4.3 Loss/Error function

Loss or error functions serve as quantifiable metrics that elucidate the discrepancy between predicted and ground-truth outcomes. The selection of an appropriate loss function is crucial, depending on the specific nature of the computational task at hand. It is imperative that these functions are computationally tractable, facilitating straightforward optimization. Prominent examples of such objective functions include Mean Squared Error (MSE), Cross-Entropy Loss, and Logarithmic Loss, each offering distinct advantages depending on the underlying mathematical properties of the model and the domain-specific constraints of the problem being addressed.

$$E = \sum_{j} (\boldsymbol{y}_m - \hat{\boldsymbol{y}}_m)^2 \tag{30}$$

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=0}^{N-1} (s[n] - \hat{s}[n])^2 = \frac{1}{N} \sum_{n=0}^{N-1} \left(s[n] + \sum_{k=1}^{p} \theta_k s[n-k] \right)^2 \tag{31}$$

In the context of our computational model, we introduce a carefully formulated loss function, denoted by $\mathcal{L}(\boldsymbol{\theta})$ as shown above, which is specifically designed to quantify the discrepancies between the predicted and actual results. Our primary objective is to identify an optimal set of θ, that minimizes this loss function. To achieve this minimization, the use of an optimization algorithm is necessary.

4.4 Optimizer

In both conventional vocoding methodologies and neural vocoding frameworks, iterative optimization techniques are indispensable for model parameter refinement. Using a designated optimizer, the objective is to iteratively adjust the loss function in a manner that facilitates its minimization. This iterative process is imperative to determine the optimal model parameters, culminating in the identification of the global or local minimum.

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\theta}) \tag{32}$$

In the process of optimization, two predominant methodologies are mainly used, namely gradient descent techniques and the Newton-Raphson method. Both approaches manifest indispensable attributes, each uniquely contributing to the broader landscape of optimization problems. A comprehensive discussion delineating their respective characteristics is presented herein.

4.4.1 Gradient descent

In the domain of optimization, the gradient descent algorithm serves as a method to locate a local minimum of a given objective function. The method operates iteratively, beginning with an initial estimate and progressing in the direction opposite to the gradient of the objective function. Formally, given a function $f : R^D \rightarrow R$, the aim is to identify the global minimum of $\mathcal{L}(\boldsymbol{\theta})$ as denoted by $\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$. The algorithm proceeds by updating the parameter vector $\boldsymbol{\theta}$ in the direction antiparallel to the gradient, symbolized by $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$.
It should be noted that obtaining an analytical solution to the optimization problem is often infeasible. Hence, the gradient descent algorithm is used, which involves iterative updates parameterized by η, commonly referred to as the learning rate in the context of machine learning. For each iteration t, the parameter vector $\boldsymbol{\theta_t}$ is updated as follows:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta_t}) \tag{33}$$

The magnitude of the learning rate η dictates the rate of convergence, with larger values inducing faster but potentially more unstable updates to the parameter vector $\boldsymbol{\theta}$. After iterating the algorithm until convergence is achieved, employing infinitesimal increments in the antipodal direction of the gradient vector.

Algorithm 1 Gradient descent

Initialize $\theta = \theta_0$ randomly and choose learning rate parameter $\eta > 0$
for $t = 1, 2, 3, \ldots$ **do**
    $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta \nabla \mathcal{L}\left(\boldsymbol{\theta}_t\right)$
end for

4.4.2 Newton-Raphson method

The Newton-Raphson method, initially conceptualized by Sir Isaac Newton and later refined by Joseph Raphson, was originally devised to determine the roots of polynomial equations. Furthermore, the method was extended by Thomas Simpson in 1740 to address the minimization of nonlinear equations.

In the context of Newton's optimization method, each iteration comprises an update that incorporates the Hessian matrix $\mathbf{H}$ of the objective function. This is accomplished by taking into account the second-order derivatives of the loss function evaluated at each iterative step.

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial \theta_1^2} & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_n} \\ \frac{\partial^2 f}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 f}{\partial \theta_2^2} & \cdots & \frac{\partial^2 f}{\partial \theta_2 \partial \theta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial \theta_n \partial \theta_1} & \frac{\partial^2 f}{\partial \theta_n \partial \theta_2} & \cdots & \frac{\partial^2 f}{\partial \theta_n^2} \end{bmatrix} \tag{34}$$

Subsequently, the iterative procedure governed by Newton's method is articulated as:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - H^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta_t}) \tag{35}$$

If we add a step size η:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - \eta H^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta_t}) \tag{36}$$

The distinction arises from its intrinsic capability to account for the curvature of the function's surface, thereby refining the convergence characteristics. As an example , the Newton-Raphson method is used to find the parameters of the LF model, where the variables ε and α are computed through the Newton-Raphson approach (Gobl, 2017). However, its deployment in neural networks is challenged by the computational intensity associated with

the derivation and evaluation of the Hessian matrix, the second-order derivative of the objective function. This computational burden makes it less feasible for large-scale applications compared to gradient descent algorithms.

Algorithm 2 Newton method

for $t = 1, 2, 3, \ldots$ **do**
  $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta \mathrm{H}^{-1} \nabla \mathcal{L}\left(\boldsymbol{\theta}_t\right)$
end for

4.5 Summary

In this chapter, we introduce the foundational underpinnings of neural networks within deep learning. An analysis of contemporary models is presented, demonstrating a paradigmatic shift in performance in contrast to the literature surveyed in Chapter 2. Subsequently, we delineate the architecture of neural networks, thereby bridging the theoretical knowledge between classical and neural speech modeling. Our discourse then turns to the application of generative models in the emergent domain of neural vocoders. In the next section, we expand our exploration by investigating the work of neural network architectures within LF vocoders.

5 | Neural LF glottal vocoder

Based on the above works, this chapter proposes a hybrid LF glottal vocoder called Tolg [1]. The name comes from the inversion of the word "Glot". To begin exploring hybrid systems, it is useful to consider the HTS-LF synthesis system as a significant exemplar of machine learning. The HTS-LF system exhibits the ability to separate and synthesize LF glottal source signals (Cabral, 2011). The latest GlottDNN vocoder [2] (Airaksinen et al., 2016; Juvela et al., 2016; Juvela, 2020) as we discussed before represents remarkable work on the source-filter model and DNN-based glottal signal synthesis. Despite the fact that GlottDNN does not follow the LF model, it has shown improved TTS quality in subjective listening tests. At Trinity College Dublin, the recently introduced GlórCáil speech synthesis system is also of considerable importance. By taking a limited set of parameters from the transformed LF model, it can adjust the emotion and voice quality of speech (Murphy, 2021). We have been inspired by the transformed LF model and the GlórCáil system, and therefore we have included the global wave shape parameter R_d as a key component of our training model, as it can accurately capture the subtleties of voice quality (Murphy et al., 2022; Yanushevskaya et al., 2022). We propose to incorporate the R_d LF parameter into the existing neural glottal vocoder in the present study.

5.1 Neural source-filter vocoder

The neural source-filter waveform model proposed in (Wang, Takaki, & Yamagishi, 2019) is a NAR model that converts phonetic characteristics into speech waveforms. It is possible to

[1] Tolg work community at: https://github.com/tolg-voice/tolg
[2] https://github.com/ljuvela/GlottDNN

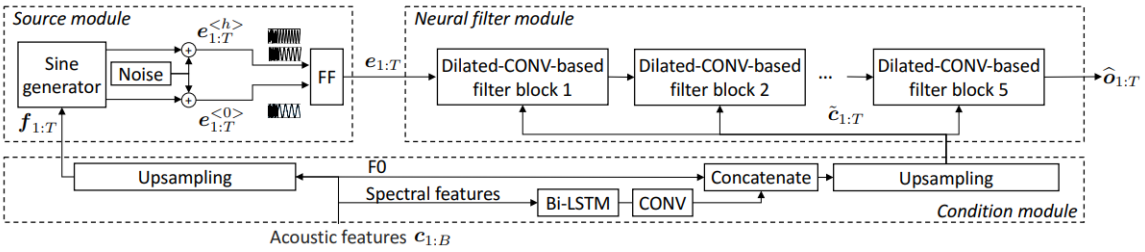

Figure 5.1: A block diagram of the neural source-filter vocoder (from Wang et al., 2019).

minimize the spectral amplitude and phase distances in training by using DFTs. Using dilated convolution, the source module generates a sine wave excitation signal with a specified fundamental frequency. The source module generates an excitation signal from f_0, which carries the source information. Using a simple equation, we can generate a sine wave that includes the sampling rate, the initial phase, and the active noise. In the accompanying Figure 5.1, this framework is illustrated. Here, a feedforward layer, harmonics and fundamental components are merged into the final excitation signal. As a result, each neural filter block generates the same 1D signal output, with the input excitation transformed into an output waveform by the neural filter module. The basic framework for all neural sinusoidal filter models consists of three modules: one module deals with the source, one module deals with the neural network itself, and another module deals with conditional features.

5.2 Neural glottal vocoder

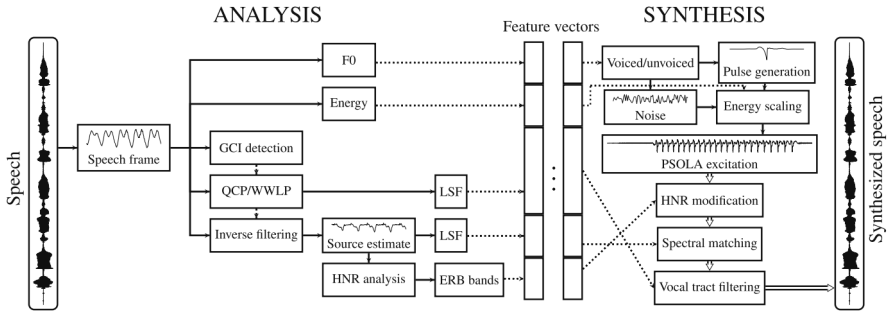

Figure 5.2: Overview of the GlottDNN vocoder analysis and synthesis processes (from Juvela, 2020).

Neural glottal vocoders are a type of hybrid neural vocoder that replaces AR glottal functions with neural ones to represent excitation signals. GlottDNN is one of the most suitable neural

glottal vocoders, since it uses a quasi-closed phase analysis (QCP) GIF technique (Airaksinen et al., 2014) and a feedforward neural network to generate pitch-synchronous glottal excitation pulses in segments of two pitch periods (Juvela et al., 2016). GlottDNN is often rated similarly to the WaveNet type of neural vocoder (Juvela, 2020). Based on this idea, the later GlotNet (Juvela et al., 2019) is usually ranked the highest, and a nine-layer GlotNet has performance comparable to a 30-layer WaveNet.

5.3 Proposed hybrid system

In the proposed hybrid system, two versions are offered, a classical version and a neural version. For classical speech analysis, we will use well-established R_d extraction scripts in Matlab from Voice_Analysis_Toolkit[3] and Covarep (Kane, 2012; Degottex et al., 2014), which we have used in our previous research at Trinity College Dublin's Phonetics and Speech Laboratory. We redeveloped the initial code in C++ to speed up the computation and make it more accessible. There are many similarities between the neural approach here and GlottDNN, especially in terms of the network training pipeline and some pre-processing analysis steps. Generally, in this study, we are interested in the structural aspects of the analysis and training of models. In all, we use a GlottDNN-like glottal excitation model for statistical parametric speech synthesis in the DNN training and decoding part.
In Tolg, the LF excitation signal is mathematically represented in equation (37). E_{LF} is the excitation signal and c is the conditional feature, E_{LF} is the classical or neural LF glottal model that we proposed, a_k are the AR coefficients.

$$s(n) \approx \sum_{k=1}^{P} a_k s_{n-k} + E_{LF}\left(n, c_{1:M}\right) \tag{37}$$

The CMU ARCTIC database (Kominek & Black, 2003) was used as the speech corpus sampled at 16 kHz. The subject, a native speaker of American English, serves as a prevalent standard in the domains of speech processing and synthesis.
The experimental setups are shown in Figure 5.3:

[3]https://github.com/jckane/Voice_Analysis_Toolkit

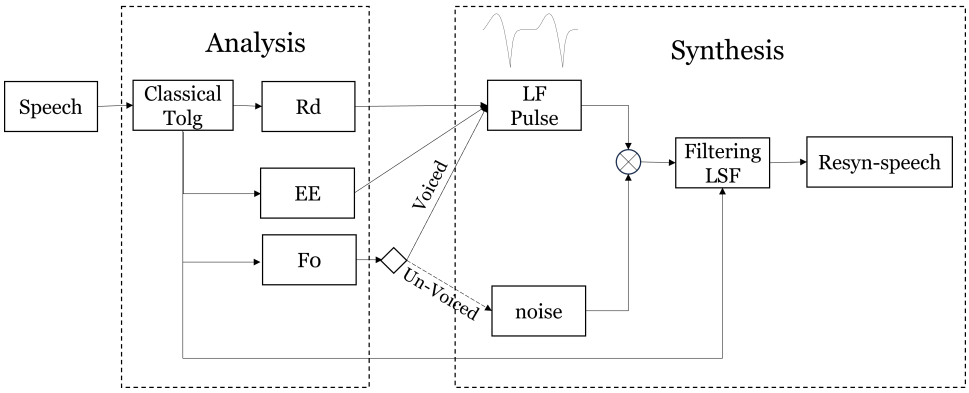

Figure 5.3: Decoding and re-synthesizing block diagram of classical Tolg vocoder.

(1) The classical Tolg model is utilized to estimate the R_d parameter, serving as a foundational step for subsequent analyses.

(2) LF pulse signals are synthesized by employing the classical Tolg model as the generative mechanism.

(3) The LF pulses are transformed into a structured feature matrix, which acts as the input for the neural network embedded within the neural Tolg framework.

(4) The previously computed Rd parameters data are leveraged as input features, which are mapped to the corresponding LF pulses derived from the classical Tolg model, for the training of the neural network model.

As delineated in the aforementioned workflow, we commence by using the classical vocoder as an analytical script to facilitate the preparation of training materials. Following this, LF pulses generated from the classical vocoder are integrated into the neural Tolg framework for subsequent pulse predictions. Hence, the procedure manifests itself in a twofold approach: initial deployment of the classical Tolg model followed by the incorporation of its neural counterpart. In essence, LF pulses are extracted via the classical Tolg model and then fed into the neural Tolg framework through a DNN model for predictive analysis.

5.3.1 Classical Tolg

The classical Tolg is divided into analysis and synthesis stages, followed by a linear generation process. We implemented the GIF method for the analysis stage using the quasi-closed phase (QCP) (Airaksinen et al., 2013), which has been proven effective for high-pitched speech. We

used 25 ms analysis frames with a frameshift of 5 ms during the analysis. We determined the F_0 and GCIs using the REAPER pitch estimator (Talkin, 2023). In the proposed glottal vocoder framework, the LSFs were partitioned into two distinct sets: LSF_{glott} and LSF_{VT}. Specifically, LSF_{glott} consisted of a 10-dimensional feature vector, while LSF_{VT} was characterized by a 30-dimensional feature vector. Additionally, the Harmonic-to-Noise Ratio (HNR) feature was extracted and represented across five distinct frequency channels. As we fit the pulses of the LF model to one of our estimated glottal flow waveforms, we used mainly the dyProg Algorithm (Kane, 2012). To extract R_d, a comprehensive search mechanism is used. That mechanism uses a dynamic programming algorithm specifically tailored to select the optimal set of parameters through R_d. Subsequent to the overlap-add procedure, the focus shifts towards the synthesis of pitch-synchronous LF glottal flow waveforms in the time domain.

The classical Tolg vocoder provides a mechanism for manipulating R_d, instrumental in controlling the degree of tension or breathiness in the voice. Within the configuration file `dnn_demo/config_dnn_demo.cfg`, an adjustable parameter denoted as Rd_Ratio is specified. This parameter facilitates the alteration of the resultant waveform files by either multiplication or division by a user-defined scalar factor.

$$R_d^* = R_d \times Rd\_Ratio \tag{38}$$

The methodology introduced effectively accomplishes the synthesis of the LF glottal flow waveform. Furthermore, this method facilitates straightforward manipulation of the generated waveform by tuning the input parameters associated with the classical LF vocoding functions, as specified in Tolg's framework. Upon entering all relevant parameters into the classical Tolg LF model, we will obtain re-synthesized speech that includes the expected R_d effects.

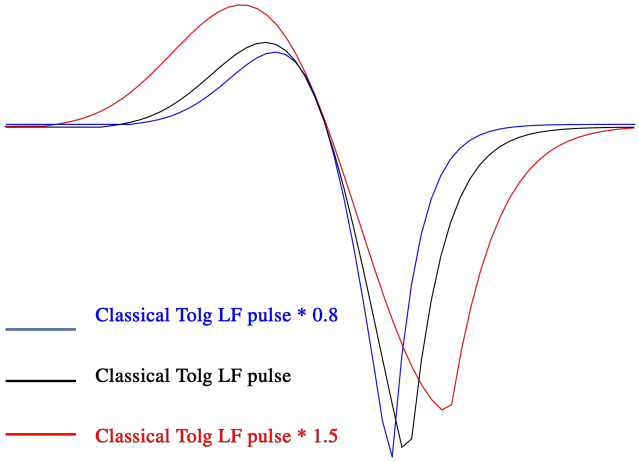

Figure 5.4: Modification effects on the classical Tolg vocoder using Rd_Ratio on the time domain.

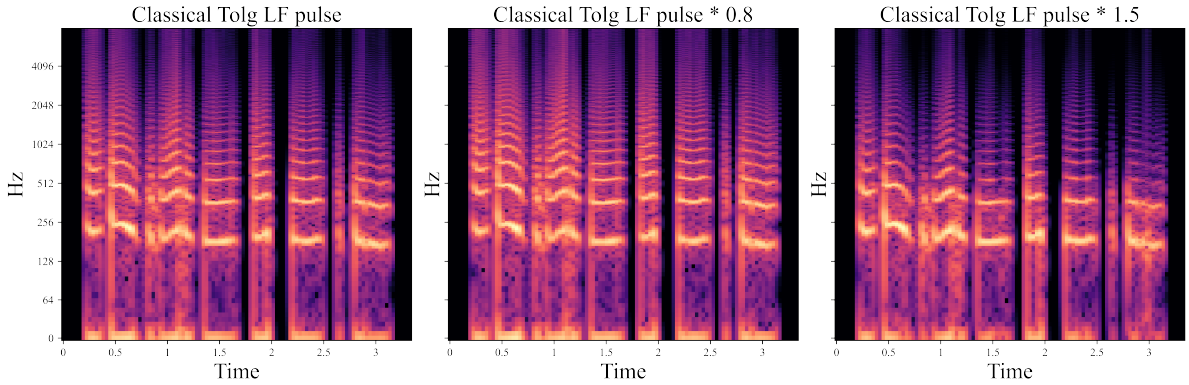

Figure 5.5: Modification effects on the classical Tolg vocoder using Rd_Ratio on the frequency domain.

The speech sample from "arctic_a0001" was chosen for the purpose of visualizing the results; it features a transcription that reads "Author of the Danger Trail, Philip Steels, etc.". As depicted in Figure 5.4, it can be observed that in the time domain and showing a single pulse randomly selected from a vowel /ɔː/, an increase in the global value of R_d results in pulses generated with higher peak flow and a longer return phase. Note that the three pulses in Figure 5.4 are slightly time-shifted relative to each other to enhance visualization. The corresponding spectral representations in Figure 5.5 further elucidate a consequential attenuation of the higher frequency components. Such transformations engender a breathier voice quality. Conversely, a reduction in the R_d values not only attenuates the peak amplitude in the time domain, but also amplifies higher frequencies in the frequency domain, resulting in the voice

quality characterized by increased tension.

The tree diagram 5.6 above and Table 5.1 list the main operations folder options, including descriptions of the general settings and how to understand them. For convenience, the structure of the output folders is depicted visually.

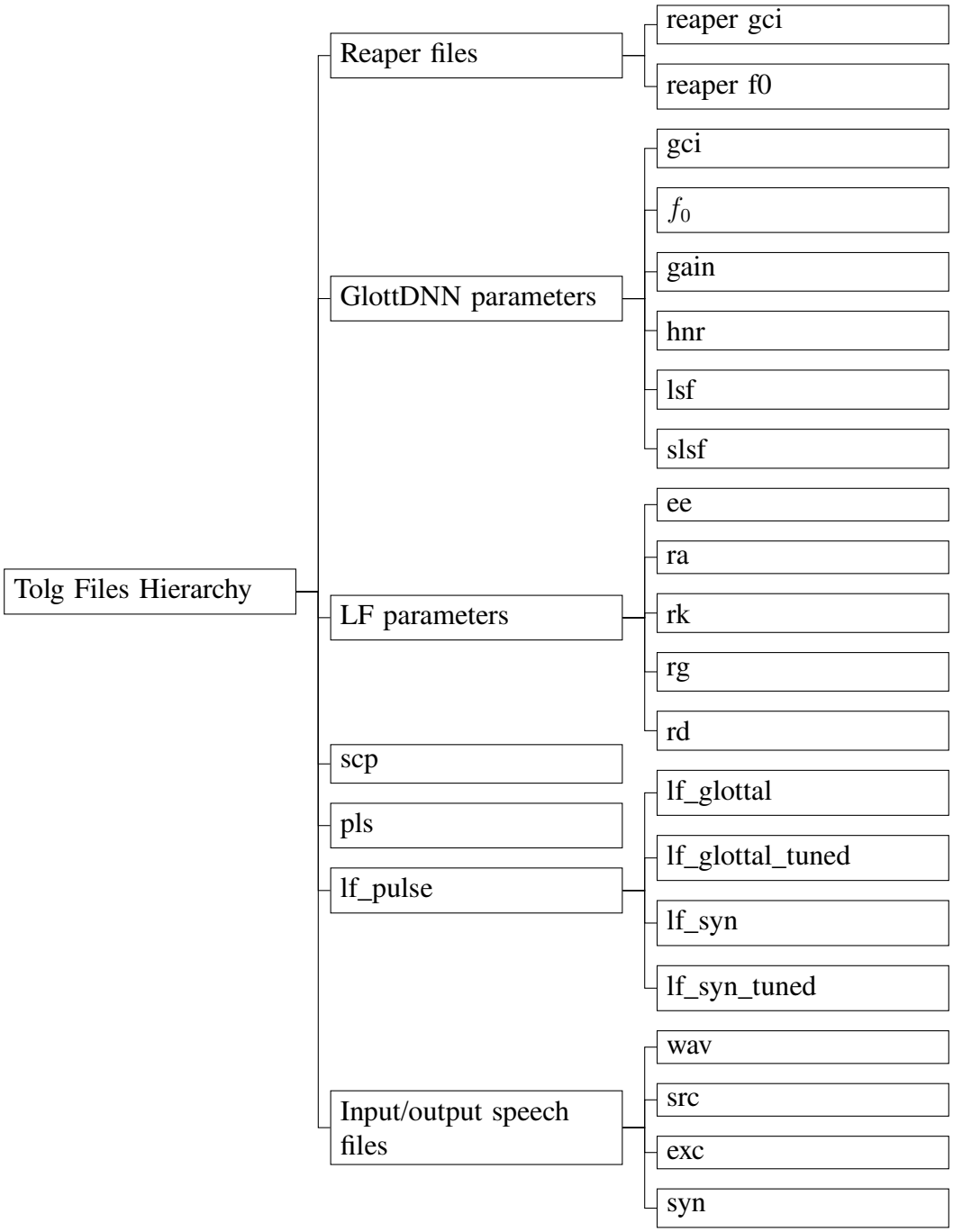

Figure 5.6: The architecture of files and folders within the Tolg system.

Folder name	Extension	Description
$reapergci$	.reaper_gci	GCI values from Reaper
$reaper f_0$	.reaper_f0	f_0 values from Reaper
gci	.gci	gci values from GlottDNN using SEDREAMS
f_0	.f0	f_0 extracted from GlottDNN
$gain$	.gain	$gain$ extracted from GlottDNN
hnr	.hnr	hamnonic noise ratio values from GlottDNN
lsf	.lsf	lsf extracted from GlottDNN
$slsf$	.slsf	slsf extracted GlottDNN
E_e	.ee	E_e from classical Tolg
R_a	.ra	R_a from classical Tolg
R_k	.rk	R_k from classical Tolg
R_g	.rg	R_g from classical Tolg
R_d	.rd	R_d from classical Tolg
scp	.scp	scp for DNN trainning from classical Tolg
pls	.pls	glottal signal feature matrixs from classical Tolg
$lf\_glottal$	.lf_pulse.wav	linear LF glottal signal from classical Tolg
$lf\_glottal\_tuned*$	.lf_pulse_tuned.wav	modified glottal signal through classical Tolg
lf_syn	.lf_syn.wav	synthesized speech through classical Tolg
lf_syn_tuned*	.lf_syn_tuned.wav	modified syn-speech through classical Tolg
src	.src.wav	source signals from the QCP GIF analysis
exc	.wav	predicted LF glottal signal from the neural Tolg
syn	.wav	synthesized speech from the neural Tolg
wav	.wav	original speech files for training and re-syn

Table 5.1: In the configuration directory, files originating from the analytical phase of the Tolg framework are explicitly with a red hue, those emanating from the synthesis phase are in blue. Parameters that are conditionally based on user-specific values are marked with *.

5.3.2 Neural Tolg

Figure 5.7 illustrates a feedforward neural network (FNN) trained to generate the LF glottal flow waveform using the proposed neural LF excitation vocoding method. Using the R_d synthetic glottal flow of the classical Tolg, the DNN is trained to predict these glottal pulse vectors; the network architecture consists of three hidden layers comprised of 150, 250, and 300 neurons, respectively. To train the model, we use Adam Optimizer with a learning rate of 1e-4. To ensure efficient training, we used a batch size of 100 and set the maximum number of epochs to 100 and will input the training data 100 times to compute the loss parameter at each time. To avoid overfitting, we implemented an early stop criterion with patience of 5. As a result of these settings, we were able to achieve optimal performance and fine-tune well for the given task. As a result, a DNN is trained to map the LF parameter R_d to normalized LF glottal waveforms of the same duration. Synthesized neural Tolg sections are based on pitch-synchronous pulse frames. In the stage of decoding and synthesis, as illustrated in Figure 5.8, the R_d representation is used as an input to the pre-trained DNN to predict the LF glottal flow waveform. Subsequently, framewise estimations of the LF glottal flow signals are made, using each individual frame extracted from the comprehensive speech database. Therefore, the neural LF pipeline maps acoustic characteristics to the glottal flow waveform in the time domain using an FNN and the output is shown in Figure 5.9.

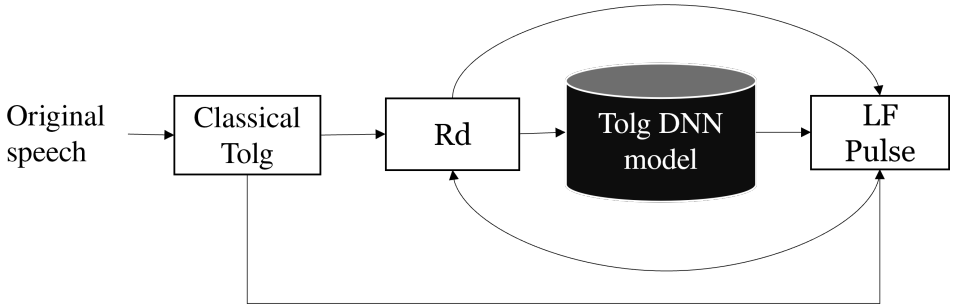

Figure 5.7: Model training block diagram of neural Tolg vocoder.

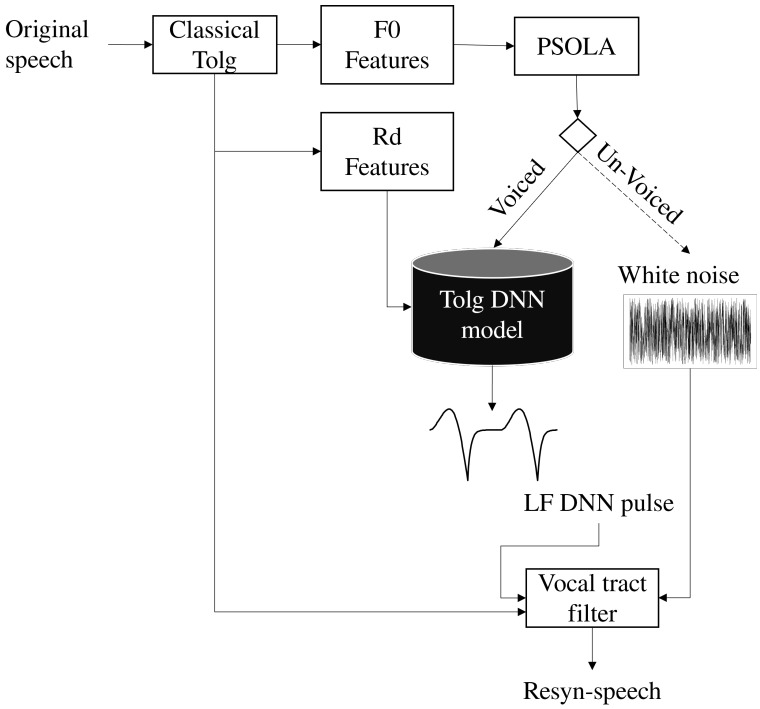

Figure 5.8: Decoding and re-synthesizing block diagram of neural Tolg vocoder.

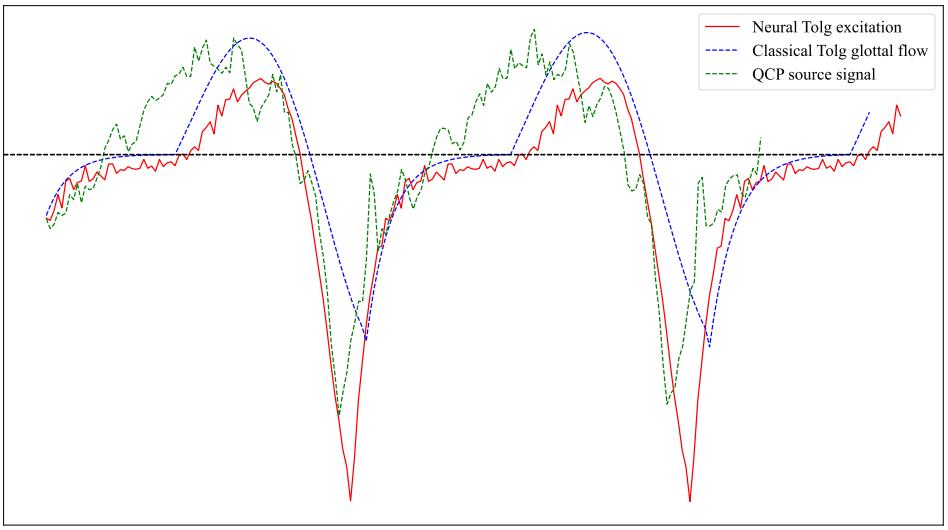

Figure 5.9: The proposed neural-based Tolg model facilitates the generation of the LF glottal excitation signal.

5.4 Experimental evaluation

To assess performance and draw comparative insight between the proposed classical and neural Tolg, a preliminary perception test was performed via an online interface. The perceptual quality of these two methods was quantitatively analyzed using a simple A/B preference test,

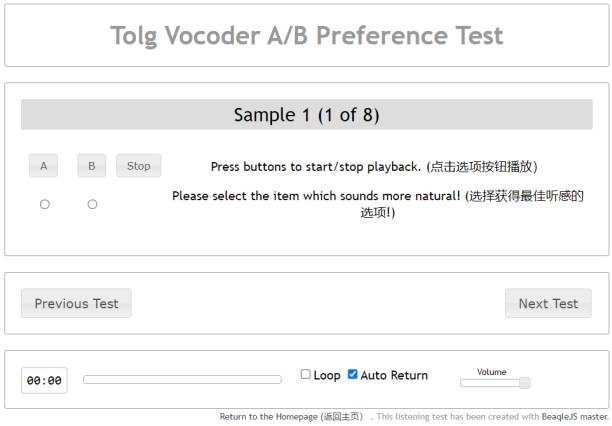

Figure 5.10: Screenshot of the Tolg vocoder A/B preference test: testing page.

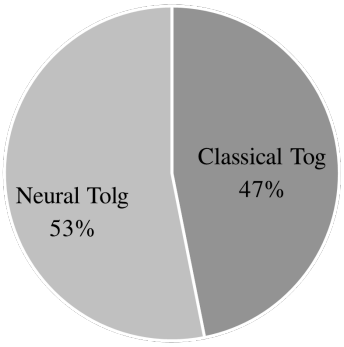

Figure 5.11: Results of A/B preference test (%).

the results of which are elucidated in Figure 5.10. This allowed for a side-by-side comparison with the classical and neural implementations of the Tolg vocoder.

In the A/B preference test that used the BeaqleJS framework for online deployment (Kraft, 2014), we selected 10 utterances from each system and 12 subjects, mainly students at Trinity College Dublin. With the consent of the prior data collection policy, participants in the listening test were asked to provide quality judgments and preferences for the naturalness of the given synthesised utterances.

Figure 5.11 presents the outcomes of the conducted A/B preference test, which relies on the data submitted by the participants. On average, the proposed neural-based glottal vocoder is slightly preferred compared to its traditional parametric glottal vocoder counterpart. However,

since the number of participants was small, the difference is not statistically significant, and further evaluation will be required to establish whether neural architecture contributes to better naturalness.

5.5 Summary

In this chapter, we create a hybrid LF vocoder that supports both classical and neural vocoding functions. Through a small-scale auditory experiment, it suggest a small difference between the classical and neural vocoders, which is worth further investigation. This chapter contains works that can serve as a starting point for building neural LF vocoders.

6 | Conclusion

The work here proposes a novel glottal vocoding system, Tolg. Three main characteristics of the vocoder are described below:

(1) it generates LF glottal excitation using a hybrid method that integrates neural networks with classical linear generation.

(2) it utilizes more high-performance scripts that replace previous MATLAB environment scripts for classical LF speech synthesis, providing the potential for a production environment for real-time speech modification.

(3) it deploys a simple global parameter to better perform the R_d modification and re-synthesize tasks.

The results of a preliminary perception test show a small difference between the LF classical and neural glottal vocoders. This may be due to the limited testing samples and listeners, and therefore, it would be necessary to carry out further auditory evaluations and a more comprehensive experimental methodology to investigate the findings.

A critical avenue for future research lies in the exploration of the mapping capacity of the neural vocoder modification with respect to the parameter R_d. Specifically, two pertinent queries are: 1) the feasibility of using the neural LF vocoder to re-synthesize speech according to a predetermined R_d parameter, and 2) the potential to map textual information in conjunction with the R_d parameter. Although the current study corroborates the efficacy of classical LF vocoding methods, it also illuminates the need for a more comprehensive examination of controllability aspects within neural vocoding techniques with larger training datasets.

Further improvements can be found in the following directions.

- Analysis module: For the analysis stage, specifically, in terms of pitch extraction and GCI detection work.
- Synthesis module: For classical synthesis, there is a need to enhance the classical LF source generation program to improve training materials. At the same time, the deployment of more advanced models on the neural vocoding approach urgently needs to be deployed, such as the GAN/diffusion model.
- Test evaluation: More broad evaluation and validation of the proposed vocoder through subjective listening tests, to gauge its effectiveness.
- User experience: Better interface experience instead of global range controlling; more acoustic features can be added inside to reveal the full control of the speech modification.
- Real-life application deployment: Integration of the features developed in this study with the Abair Irish TTS system (Ní Chasaide et al., 2017), to implement online TTS tasks with voice quality tuning features.

References

Airaksinen, M., Bollepalli, B., Juvela, L., Wu, Z., King, S., & Alku, P. (2016). GlottDNN–A Full-Band Glottal Vocoder for Statistical Parametric Speech Synthesis. In *Proc. interspeech 2016* (pp. 2473–2477).

Airaksinen, M., Raitio, T., Story, B., & Alku, P. (2014, March). Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(3), 596–607.

Airaksinen, M., Story, B., & Alku, P. (2013). Quasi closed phase analysis for glottal inverse filtering. In *Proc. interspeech 2013* (pp. 143–147).

Alku, P. (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Commun.*, *11*(23), 109–118.

Atal, B. (2006, 04). The history of linear prediction. *Signal Processing Magazine, IEEE*, *23*, 154-161.

Bäckström, T., Räsänen, O., Zewoudie, A., Zarazaga, P. P., Koivusalo, L., Das, S., . . . Alku, P. (2022). *Introduction to speech processing* (2nd ed.).

Cabral, J. (2011). *Hmm-based speech synthesis using an acoustic glottal source model* (Doctoral thesis). University of Edinburgh.

Camacho, A., & Harris, P. R. (2008). A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, *124*(3), 1638–1652.

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, *111*(4), 1917–1930.

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). Covarep a collaborative voice analysis repository for speech technologies. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 960-964.

Dhillon, A., & Verma, G. (2019, 12). Convolutional neural network: a review of models, methodologies and applications to object detection. *Progress in Artificial Intelligence*, *9*. doi: 10.1007/s13748-019-00203-0

Drugman, T. (2011). *Advances in glottal analysis and its applications* (Unpublished doctoral

dissertation). University of Mons.

Drugman, T., & Alwan, A. (2011). Joint robust voicing detection and pitch estimation based on residual harmonics. In *Proc. interspeech* (pp. 1973–1976).

Drugman, T., & Dutoit, T. (2009). Glottal closure and opening instant detection from speech signals. In *Proc. interspeech* (pp. 2891–2894).

Drugman, T., & Dutoit, T. (2012). The deterministic plus stochastic model of the residual signal and its applications. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(3), 968–981.

Drugman, T., Thomas, M., Gudnason, J., Naylor, P., & Dutoit, T. (2012). Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(3), 994–1006.

Drugman, T., Wilfart, N., & Dutoit, T. (2009). A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In *Proc. interspeech* (pp. 1779–1782).

Elias, P. (1955a, Mar.). Predictive coding i. *IRE Trans. Inform. Theory*, *IT-1*(1), 16–24.

Elias, P. (1955b, Mar.). Predictive coding ii. *IRE Trans. Inform. Theory*, *IT-1*(1), 24–33.

Fant, G. (1961). A new anti-resonance circuit for inverse filtering. *Speech Transmission Laboratory Quarterly Progress and Status Report*, *2*(1), 1–6.

Fant, G. (1968). Analysis and synthesis of speech processes. In B. Malmberg (Ed.), *Manual of phonetics* (pp. 173–177). Amsterdam.

Fant, G. (1970). *Acoustic theory of speech production* (2nd ed.). The Hague, Paris: Mouton.

Fant, G. (1995). The lf-model revisited . transformations and frequency domain analysis. *STL-QPSR*.

Fant, G., Liljencrants, J., & Lin, Q. (1985). A four-parameter model of glottal flow. *Speech Transmission Laboratory Quarterly Progress and Status Report*, *26*, 1–13.

Fant, G., Liljencrants, J., & Lin, Q. (1995, 01). The lf-model revisited. transformations and frequency domain analysis. *KTH, Speech Transmission Laboratory, Quarterly Report*, *2-3*, 119-156.

Fant, G., & Sonesson, B. (1962). Indirect studies of glottal cycles by synchronous inverse

filtering and photo electrical glottography. *Speech Transmission Laboratory Quarterly Progress and Status Report*, *3*(4), 1–3.

Fujisaki, H., & Ljungqvist, M. (1986). Proposal and evaluation of models for the glottal source waveform. In *Icassp* (pp. 1605–1608).

Gobl, C. (2017). Reshaping the transformed lf model: Generating the glottal source from the waveshape parameter rd. In *Interspeech 2017* (pp. 3008–3012). Stockholm, Sweden.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. (`http://www.deeplearningbook.org`)

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27). Curran Associates, Inc. Retrieved from `https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf`

Hedelin, P. (1981). A tone oriented voice excited vocoder. In *Proc. icassp* (Vol. 6, pp. 205–208).

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., . . . Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82-97. doi: 10.1109/MSP.2012.2205597

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *CoRR*, *abs/2006.11239*. Retrieved from `https://arxiv.org/abs/2006.11239`

Itakura, F. (1975). Line spectrum representation of linear predictive coefficients of speech signals. *J. Acoust. Soc. Am.*, *57*(535(A)).

Itakura, F., & Saito, S. (1970). A statistical method for estimation of speech spectral density and formant frequencies. *Electron. Commun. Japan*, *53-A*, 36–43.

Juvela, L. (2020). *Neural waveform generation for source-filter vocoding in speech synthesis* (Doctoral thesis, School of Electrical Engineering). Retrieved from `http://urn.fi/URN:ISBN:978-952-60-3910-7`

Juvela, L., Bollepalli, B., Airaksinen, M., & Alku, P. (2016). High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network. In *Proc. int. conf. acoust., speech signal process.* (pp. 5120–5124).

Juvela, L., Bollepalli, B., Tsiaras, V., & Alku, P. (2019). Glotneta raw waveform model for the glottal excitation in statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(6), 1019-1030. doi: 10.1109/TASLP.2019.2906484

Kane, J. (2012). *Tools for analysing the voice: Developments in glottal source and quality analysis* (Unpublished doctoral dissertation). Trinity College (Dublin, Ireland), Centre for Language and Communication Studies.

Kane, J., & Gobl, C. (2013a). Automating manual user strategies for precise voice source analysis. *Speech Communication*, *55*(3), 397–414. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0167639312001422` doi: 10.1016/j.specom.2012.12.004

Kane, J., & Gobl, C. (2013b). Evaluation of glottal closure instant detection in a range of voice qualities. *Speech Communication*, *55*(2), 295–314.

Kawahara, H., Estill, J., & Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high-quality speech analysis, modification, and synthesis system straight. In *Second international workshop on models and analysis of vocal emissions for biomedical applications* (p. N/A).

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. In *Acoustics, speech and signal processing, 2008. icassp 2008. ieee international conference on* (pp. 3933–3936).

Kim, S.-j., Kim, J.-j., & Hahn, M. (2006). Hmm-based korean speech synthesis system for hand-held devices. *IEEE Transactions on Consumer Electronics*, *52*(4), 1384-1390. doi: 10.1109/TCE.2006.273160

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the*

Acoustical Society of America, *67*(3), 971–995.

Kominek, J., & Black, A. W. (2003). *Cmu arctic databases for speech synthesis* (Tech. Rep.). Language Technologies Institute.

Kraft, S. (2014). Beaqlejs : Html 5 and javascript based framework for the subjective evaluation of audio quality.. Retrieved from `https://api.semanticscholar.org/CorpusID:49529466`

Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., . . . Hsu, W.-N. (2023). *Voicebox: Text-guided multilingual universal speech generation at scale.*

LeCun, Y., Bengio, Y., & Hinton, G. (2015, 05). Deep learning. *Nature*, *521*, 436-44. doi: 10.1038/nature14539

Lyon, D. (2008, 11). The -law codec. *The Journal of Object Technology*, *7*, 21. doi: 10.5381/jot.2008.7.8.c2

McAulay, R., & Quatieri, T. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, *34*(4), 744–754.

McLoughlin, I. V. (2008). Line spectral pairs. *Signal Processing*, *88*, 448–467.

Miller, R. (1959). Nature of the vocal cord wave. *J. Acoust. Soc. Am.*, *31*(6), 667–677.

Moinet, A., Drugman, T., Dutoit, T., & Wilfart, G. (2009, 04). Using a pitch-synchronous residual codebook for hybrid hmm/frame selection speech synthesis. In *Acoustics, speech, and signal processing, ieee international conference* (p. 3793-3796). Los Alamitos, CA, USA: IEEE Computer Society.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, *9*, 453–467.

Moulines, E., & Laroche, J. (1995). Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, *16*(2), 175–205.

Murphy, A. (2021). *Controlling the voice quality dimension of prosody in synthetic speech using an acoustic glottal model* (Unpublished doctoral dissertation). Trinity College Dublin, School of Linguistic, Speech & Communication Sciences.

Murphy, A., Yanushevskaya, I., Chasaide, A. N., & Gobl, C. (2019). The Role of Voice

Quality in the Perception of Prominence in Synthetic Speech. In *Proc. interspeech 2019* (pp. 2543–2547). doi: 10.21437/Interspeech.2019-2761

Murphy, A., Yanushevskaya, I., Chasaide, A. N., & Gobl, C. (2021). Integrating a voice analysis-synthesis system with a tts framework for controlling affect and speaker identity. In *2021 32nd irish signals and systems conference (issc)* (p. 1-6). doi: 10.1109/ISSC52156.2021.9467853

Murphy, A., Yanushevskaya, I., Chasaide, A. N., & Gobl, C. (2022). Affect Expression: Global and Local Control of Voice Source Parameters. In *Proc. speech prosody 2022* (pp. 525–529).

Noll, A. M. (1967). Cepstrum pitch determination. *The Journal of the Acoustical Society of America*, *41*(2), 293–309.

Ní Chasaide, A., Ní Chiaráin, N., Wendler, C., Berthelsen, H., Murphy, A., & Gobl, C. (2017, 08). The abair initiative: Bringing spoken irish into the digital space. In *Proc. interspeech 2017* (p. 2017-1407).

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., & Kavukcuoglu, K. (2017). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Oord, A. v. d., Kalchbrenner, N., & Kavukcuoglu, K. (2016). Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*.

O'Shaughnessy, D. (1987). *Speech communication: human and machine*. United States: Addison-Wesley.

Peng, K., Ping, W., Song, Z., & Zhao, K. (2020, 13–18 Jul). Non-autoregressive neural text-to-speech. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 7586–7598). PMLR. Retrieved from `https://proceedings.mlr.press/v119/peng20a.html`

Perrotin, O., & McLoughlin, I. (2019). Gfm-voc: A real-time voice quality modification system. In *Proc. interspeech 2019* (pp. 3685–3686).

Rabiner, L., & Schafer, R. (1978). *Digital processing of speech signals*. Prentice-Hall.

Raitio, T., Lu, H., Kane, J., Suni, A., Vainio, M., King, S., & Alku, P. (2014). Voice source

modelling using deep neural networks for statistical parametric speech synthesis. In *Proc. eusipco.*

Raitio, T., Suni, A., Pulakka, H., Vainio, M., & Alku, P. (2011). Utilizing glottal source pulse library for generating improved excitation signal for hmm-based speech synthesis. In *Proceedings of the ieee international conference on acoustics, speech, and signal processing* (pp. 4564–4567).

Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., & Alku, P. (2011, 02). Hmm-based speech synthesis utilizing glottal inverse filtering. *Audio, Speech, and Language Processing, IEEE Transactions on*, *19*, 153 - 165. doi: 10.1109/TASL.2010.2045239

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). *Stochastic backpropagation and approximate inference in deep generative models.*

Rosenberg, A. (1971). Effects of the glottal pulse shape on the quality of natural vowels. *J. Acoust. Soc. Am.*, *49*, 583–590.

Saito, S., & Itakura, F. (1967). Theoretical consideration of the statistical optimum recognition of the spectral density of speech. *J. Acoust. Soc. Japan.*

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *CoRR*, *abs/1503.03585*. Retrieved from `http://arxiv.org/abs/1503.03585`

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, *8*(3), 185–190.

Strube, H. (1974). Determination of the instant of glottal closure from the speech wave. *J. Acoust. Soc. Am.*, *56*(5), 1625–1629.

Stylianou, Y. (2001). Removing linear phase mismatches in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, *9*(3), 232–239.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn & K. K. Palatal (Eds.), *Speech coding and synthesis* (pp. 497–518). Elsevier Science B.V.

Talkin, D. (2023). *REAPER: Robust Epoch And Pitch Estimator.*

https://github.com/google/REAPER.

Tamamori, A., Hayashi, T., Kobayashi, K., Takeda, K., & Toda, T. (2017). Speaker-dependent wavenet vocoder. In *Interspeech 2017* (pp. 1118–1122).

Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). *A survey on neural speech synthesis.*

Taylor, P. A. (2009). Text-to-speech synthesis. In (p. 361).

Thapen, N. (2023). *Pink trombone.* https://dood.al/pinktrombone/. ((visited time 2023))

Tokuda, K., & Zen, H. (2015). Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis. In *Proceedings of the ieee international conference on acoustics, speech, and signal processing (icassp)* (pp. 4215–4219).

Tokuda, K., & Zen, H. (2016). Directly modeling voiced and unvoiced components in speech waveforms by neural networks. In *2016 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5640–5644).

van den Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., ... Hassabis, D. (2018, 10–15 Jul). Parallel WaveNet: Fast high-fidelity speech synthesis. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 3918–3926). PMLR. Retrieved from https://proceedings.mlr.press/v80/oord18a.html

Veldhuis, R. (1998). A computationally efficient alternative for the liljencrants–fant model and its perceptual evaluation. *Journal of the Acoustical Society of America*, *103*, 566–571.

Veličković, P. (2023). *Generative adversarial network examples in tikz.*

Wang, X., Lorenzo-Trueba, J., Takaki, S., Juvela, L., & Yamagishi, J. (2018). A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis. In *Proc. icassp* (pp. 4804–4808).

Wang, X., Takaki, S., & Yamagishi, J. (2019). Neural source-filter-based waveform model for statistical parametric speech synthesis. In *Icassp 2019 - 2019 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 5916-5920). doi: 10.1109/ICASSP.2019.8682298

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R., Jaitly, N., & Le, Q. (2017).

Tacotron: Towards end-to-end speech synthesis. *arXiv*. Retrieved from `http://arxiv.org/abs/1703.10135`

Wong, D., Markel, J., & Gray, A. (1979). Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Trans. Acoust. Speech Signal Process.*, *27*, 350–355.

Yanushevskaya, I., Murphy, A., Gobl, C., & Ní Chasaide, A. (2022). Global waveshape parameter *Rd* in signaling focal prominence: Perceptual salience in the absence of *f0* variation. *Frontiers in Communication, 7*, 1-23.

Yao, G., Lei, T., & Zhong, J. (2018, 05). A review of convolutional-neural-network-based action recognition. *Pattern Recognition Letters*, *118*. doi: 10.1016/j.patrec.2018.05.018