Digital speech and the Markov chain Monte Carlo method for glottal inverse filtering

Lasse Lybeck

October 27, 2015

Master of Science Thesis

University of Helsinki
Faculty of Science
Department of Mathematics and Statistics

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Mathematics and Statistics

Tekijä — Författare — Author
Lasse Lybeck

Työn nimi — Arbetets titel — Title

Digital speech and the Markov chain Monte Carlo method for glottal inverse filtering

Oppiaine — Läroämne — Subject
Applied mathematics

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's thesis	October 2015	77 p

Tiivistelmä — Referat — Abstract

Speech is the most common form of human communication. An understanding of the speech production mechanism and the perception of speech is therefore an important topic when studying human communication. This understanding is also of great importance both in medical treatment regarding a patient's voice and in human-computer interaction via speech.

In this thesis we will present a model for digital speech called the *source-filter model*. In this model speech is represented with two independent components, the *glottal excitation signal* and the *vocal tract filter*. The glottal excitation signal models the airflow created at the vocal folds, which works as the source for the created speech sound. The vocal tract filter describes how the airflow is filtered as it travels through the vocal tract, creating the sound radiated to the surrounding space from the lips, which we recognize as speech. We will also present two different parametrized models for the glottal excitation signal, the *Rosenberg-Klatt model* (RK-model) and the *Liljencrants-Fant model* (LF-model). The RK-model is quite simple, being parametrized with only one parameter in addition to the fundamental frequency of the signal, while the LF-model is more complex, taking in four parameters to define the shape of the signal. A transfer function for vocal tract filter is also derived from a simplified model of the vocal tract. Additionally, relevant parts of the theory of signal processing are presented before the presentation of the source-filter model.

A relatively new model for *glottal inverse filtering* (GIF), called the *Markov chain Monte Carlo method for glottal inverse filtering* (MCMC-GIF) is also presented in this thesis. Glottal inverse filtering is a technique for estimating the glottal excitation signal from a recorded speech sample. It is a widely used technique for example in phoniatrics, when inspecting the condition of a patient's vocal folds. In practice the aim is to separate the measured signal into the glottal excitation signal and the vocal tract filter. The first method for solving glottal inverse filtering was proposed in the 1950s and since then many different methods have been proposed, but so far none of the methods have been able to yield robust estimates for the glottal excitation signal from recordings with a high fundamental frequency, such as women's and children's voices. Recently, using synthetic vowels, MCMC-GIF has been shown to produce better estimates for these kind of signals compared to other state of the art methods.

The MCMC-GIF method requires an initial estimate for the vocal tract filter. This is obtained from the measurements with the *iterative adaptive inverse filtering* (IAIF) method. A synthetic vowel is then created with the RK-model and the vocal tract filter, and compared to the measurements. The MCMC method is then used to adjust the RK excitation parameter and the parameters for the vocal tract filter to minimize the error between the synthetic vowel and the measurements, and ultimately receive a new estimate for the vocal tract filter. The filter can then be used to calculate the glottal excitation signal from the measurements. We will explain this process in detail, and give numerical examples of the results of the MCMC-GIF method compared against the IAIF method.

Avainsanat — Nyckelord — Keywords
Signal processing, speech synthesis, glottal inverse filtering, Markov chain Monte Carlo

Säilytyspaikka — Förvaringsställe — Where deposited
Kumpula Campus Library

Muita tietoja — Övriga uppgifter — Additional information

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Matematisk-naturvetenskapliga fakulteten	Institutionen för matematik och statistik

Tekijä — Författare — Author
Lasse Lybeck

Työn nimi — Arbetets titel — Title

Digitalt tal och Markov-kedja Monte Carlo metoden för glottal inversfiltrering

Oppiaine — Läroämne — Subject		
Tillämpad matematik		
Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -avhandling	Oktober 2015	77 s.

Tiivistelmä — Referat — Abstract

Tal är den mest vanliga formen av mänsklig kommunikation. På grund av detta är det viktigt att ha en bra förståelse om hur människan producerar och uppfattar tal då man studerar mänsklig kommunikation. Denna förståelse är också högst viktig i medicinska sammanhang då man vårdar en patients röst och även i utvecklandet av talkommunikation mellan människor och maskiner.

I denna avhandling kommer vi att presentera den så kallade *källa-filter-modellen* för talproduktion. I modellen är tal representerat som två oberoende komponenter, *röstkällan* och *ansatsrörsfiltret*. Röstkällan modellerar luftflödet som uppstår vid stämbanden och fungerar som källa för det skapade talljudet. Ansatsrörfiltret modellerar hur ljudet filtreras då den rör sig genom ansatsröret till läpparna, varifrån det strålar ut till omgivningen som talljud. Vi kommer även att presentera två olika parametriserade modeller för röstkällan, *Rosenberg-Klatt modellen* (RK-modellen) och *Liljencrants-Fant modellen* (LF-modellen). Av dessa två är RK-modellen simplare och använder sig av bara en parameter tillsammans med den fundamentala frekvensen för att skapa signalen, när LF-modellen däremot använder sig av fyra parametrar för att skapa formen för signalen. Vi kommer också att härleda en överföringsfunktion för ansatsrörsfiltret från en förenklad modell för ansatsröret. Före granskningen av källa-filter-modellen kommer även relevanta delar av teorin om signalbehandling att presenteras.

En relativt ny metod för *röstkällans inversfiltrering* (eng. glottal inverse filtering, GIF), den så kallade *Markov-kedja Monte Carlo -metoden för inversfiltrering* (MCMC-GIF), presenteras också i denna avhandling. Röstkällans inversfiltrering är en teknik där man strävar efter att uppskatta röstkällan från en inspelning av tal. Tekniken används mycket i till exempel foniatri, då man granskar tillståndet av en patients stämband. I praktiken går metoden ut på att separera den inspelade talsignalen till en signal för röstkällan och ansatsrörsfiltret. Första metoden för att lösa problemet formulerades redan på 1950-talet och sen dess har många olika metoder presenterats, men tills vidare har ingen av metoderna lyckats skapa pålitliga estimat för röstkällan i sådana fall där den fundamentala frekvensen i inspelningen är hög, vilket är ofta fallet för kvinnors och barns röster. MCMC-GIF-metoden har dock under senaste tiden visats, med hjälp av syntetiska vokaler, uppnå bättre resultat än någon av de tidigare metoderna även för dessa slags mätningar.

I MCMC-GIF beräknas ett ursprungligt estimat för ansatsrörsfiltret från de inspelade mätningarna genom att använda den så kallade *iterativa adaptiva inversfiltrering* (IAIF) metoden. En syntetisk vokal skapas därefter med hjälp av RK-modellen och det beräknade ansatsrörsfiltret och jämförs med mätningarna. Därefter används MCMC-metoden för att justera RK-parametern och parametrarna för ansatsrörsfiltret för att minimera felet mellan den syntetiska vokalen och mätningarna, och till slut anhålls ett nytt estimat för anstatsrörsfiltret. Filtret kan sedan användas för att beräkna en ny uppskattning för röstkällan från mätningarna. Denna metod kommer att presenteras noggrant i avhandlingen, med numeriska exempel av en jämförelse mellan MCMC-GIF och IAIF metoderna.

Avainsanat — Nyckelord — Keywords
Signalbehandling, talsyntes, röstkällans inversfiltrering, Markov-kedja Monte Carlo

Säilytyspaikka — Förvaringsställe — Where deposited
Campusbiblioteket i Gumtäkt

Muita tietoja — Övriga uppgifter — Additional information

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen tiedekunta	Matematiikan ja tilastotieteen laitos

Tekijä — Författare — Author
Lasse Lybeck

Työn nimi — Arbetets titel — Title

Digitaalinen puhe ja Markov-ketju Monte Carlo -menetelmä äänilähteen käänteissuodatukselle

Oppiaine — Läroämne — Subject		
Sovellettu matematiikka		
Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	Lokakuu 2015	77 s.

Tiivistelmä — Referat — Abstract

Puhe on ihmisen kommunikaation yleisin muoto. Tämän vuoksi puheentuottomekanismin ja puheen käsityksen ymmärtäminen on tärkeä osa ihmisen kommunikaation ymmärtämisessä. Tämä ymmärrys on myös tärkeää lääketieteessä tutkiessa potilaan ääntä sekä ihmisen ja koneen välisessä puhekommunikaatiossa.

Tässä tutkielmassa tulemme esittämään niin kutsutun *lähde–suodin -mallin* digitaaliselle puheelle. Mallissa puhe mallinnetaan kahtena erillisenä osana, *äänilähteenä* ja *ääntöväyläsuodattimena*. Äänilähde mallintaa äänihuulilla muodostuvaa ilmavirtaa, joka toimii perustana puheäänelle. Ääntöväyläsuodatin selittää miten ääni suodattuu kulkiessaan ääntöväylän läpi, muodostaen äänen, joka säteilee huulilta ympäröivään tilaan ja jonka me miellämme puheeksi. Esitämme kaksi parametrisoitua mallia äänilähteelle, *Rosenberg–Klatt -mallin* (RK-mallin) ja *Liljencrants–Fant -mallin* (LF-mallin). Näistä kahdesta mallista RK-malli on yksinkertaisempi, sillä siinä äänilähde mallinnetaan perustaajuuden lisäksi vain yhdellä parameterilla, kun taas LF-mallissa äänilähteen muoto määritetään neljän parametrin avulla, tehden siitä huomattavasti monimutkaisemman. Johdamme lisäksi ääntöväyläsuodattimelle siirtofunktion yksinkertaistetusta mallista ääntöväylälle. Ennen lähde–suodin -mallin läpikäyntiä esitämme lisäksi tarpeelliset osat signaalikäsittelyn teoriasta.

Tutkielmassa esitämme myös melko uuden mallin *äänilähteen käänteissuodatukselle* (eng. *glottal inverse filtering*, GIF), niin kutsutun *Markov-ketju Monte Carlo -menetelmän äänilähteen käänteissuodatukselle* (MCMC-GIF). Äänilähteen käänteissuodatus on tekniikka, jossa äänilähde arvioidaan nauhoitetusta puhesignaalista. Tekniikkaa käytetään laajasti esimerkiksi foniatriikassa, kun halutaan tutkia potilaan äänihuulten kuntoa. Käytännössä menetelmissä tavoitteena on erottaa havaintosignaali äänilähteeksi ja ääntöväylän suodattimeksi. Ensimmäinen menetelmä äänilähteen käänteissuodatukselle esitettiin jo 1950-luvulla ja useita malleja ongelman ratkaisemiseksi on ehdotettu siitä lähtien, mutta vieläkään mikään tunnettu menetelmä ei ole pystynyt varmasti arvioimaan äänilähdettä tilanteissa, joissa havaintosignaalin perustaajuus on korkea, kuten naisten ja lasten äänissä tyypillisesti on. MCMC-GIF-menetelmä on kuitenkin osoittautunut synteettisiin vokaaleihin perustuvassa testauksessa toimivan muita tämän hetken parhaita menetelmiä paremmin, etenkin korkean taajuisten havaintojen tapauksessa.

MCMC-GIF-menetelmässä ääntöväylän suodattimelle tarvitaan alustava arvio, joka lasketaan niin kutsutulla IAIF-menetelmällä (eng. *iterative adaptive inverse filtering*). Tätä suodatinta ja RK-mallia käyttäen luodaan synteettinen vokaaliäänne, jota verrataan havaintosignaaliin. MCMC-menetelmää käytetään tämän jälkeen säätämään RK-mallin ja ääntöväylän suodattimen parametreja minimoimaan virhe synteettisen vokaalin ja havaintojen välillä, mistä lopulta saavutetaan uusi arvio ääntöväylän suodattimelle, jota käytetään uuden äänilähteen arvion laskemiseen. Tämä prosessi MCMC-GIF-menetelmälle esitetään tutkielmassa tarkasti, ja menetelmän tuottamia tuloksia verrataan esimerkkitapauksissa IAIF-menetelmän tuottamiin tuloksiin.

Avainsanat — Nyckelord — Keywords
Signaalinkäsittely, puhesynteesi, äänilähteen käänteissuodatus, Markov-ketju Monte Carlo

Säilytyspaikka — Förvaringsställe — Where deposited
Kumpulan tiedekirjasto

Muita tietoja — Övriga uppgifter — Additional information

Acknowledgements

This thesis was written during the spring and summer of 2015 at the Department of Mathematics and Statistics at the University of Helsinki in the Inverse Problems Research Group, which is part of the Finnish Centre of Excellence in Inverse Problems Research. First and foremost I would like to thank my supervisor Professor Samuli Siltanen for all the help and inspiration during the writing of this thesis as well as during my studies in general. I would also like to thank Dr. Ismael Rodrigo Bleyer for all the support during my time writing this thesis and for all the interesting discussions we shared regarding the subject.

Jag vill hjärtligt tacka Emilia för allt stöd du gett mig under de senaste åren och för alla de roliga stunder vi hittills haft tillsammans. Ett speciellt tack vill jag även rikta åt alla mina studiekompisar för alla de roliga och inspirerande diskussioner och trevliga stunder vi haft tillsammans under dessa år.

Tahtoisin myös kiittää äitiäni, isääni ja siskoani Lottaa kaikesta siitä tuesta, jonka olen saanut opintojeni aikana.

Contents

1 Introduction

Speech is the most common form of human communication. An understanding of the speech production mechanism and perception of speech is therefore of great interest in understanding the foundation of human communication. The importance of understanding the human speech production mechanism also rises new possibilities for diagnosing and treating anomalies and other medical conditions related to the speech production apparatus. Also, with the nowadays large number of computers and mobile devices surrounding many peoples' lives, it has become an important task for software developers to be able to create human-computer interaction based on speech.

Human speech sounds can be roughly categorized into three classes depending on the way the sound is produced. *Voiced sounds* are sounds for which the source sound signal is created by the periodic fluctuation of the vocal folds, *unvoiced sounds* use turbulent noise as the source signal, and *plosives* are created by suddenly releasing a flow of air that has been previously blocked by some part of the vocal tract [12]. Of these classes voiced sounds are the most important, and for the most part of this work we will be concentrating on these.

A simplified model of human speech production, the so called *source-filter theory* [11], divides the speech production mechanism into three parts. The source of the speech sound is located at the vocal folds, and is modelled as the *glottal excitation signal* or *glottal flow*. The flow then travels through the vocal tract, consisting of the oral and nasal cavities, where the sound is filtered with the resonance frequencies of the vocal tract, known as *formants*. The resulting flow then escapes through the lips and nostrils in a process called *lip radiation*, creating the speech pressure waveform that we hear as the speech sound.

A mathematical model for the speech production mechanism can be derived from the source-filter theory. The final part of the source-filter theory, the lip radiation, can be modelled as a first order differentiator of the volume velocity reaching the lips [12], and due to this we can model the production of a voiced speech sound as

$$m = p * v, \tag{1.1}$$

where m is the produced sound signal, p is the glottal flow derivative and v is the impulse response of the vocal tract filter. In the z-domain this model takes the form

$$S(z) = \hat{G}(z)V(z), \tag{1.2}$$

where S, $\hat{G}$, and V are the z-transforms of the measurements, the glottal flow derivative and the impulse response of the vocal tract, respectively. Here $\hat{G}(z) = G(z)L(z)$, where G is the z-transform of the glottal flow and L is the

z-transform of the lip radiation effect.

Glottal inverse filtering (GIF) is a technique for estimating the glottal flow from a recorded speech signal. Assuming the source-filter model described in equation (1.1), this can be done by first estimating the effect of the vocal tract v on the measured sound m. The measurement signal m can then be filtered with the inverse of the filter v, removing the effect of the vocal tract from the measured signal and thus revealing the glottal excitation signal p. In practice this is often done in the z-domain, as described by equation (1.2). In this setting we can acquire the glottal excitation by calculating $\hat{G}(z) = S(z)/V(z)$. Although schematically easy, GIF is a hard inverse problem to solve robustly, and many different methods have been proposed for solving the problem.

GIF is an important technique in both medicine and technology. In medicine GIF is used for studying the vocal folds, for example when there is suspicion of a medical condition with a patients voice. The vocal folds are a difficult target for making direct measurements of, as they are small and move at high speeds during voiced speech. The length of adult vocal folds vary between 1.25 cm and 1.75 cm for women and between 1.75 cm and 2.5 cm for men, and in typical speech the vocal folds vibrate with a frequency of between 165 Hz and 255 Hz for women and between 85 Hz and 180 Hz for men [31]. Thus indirect measurements with GIF are often used in inspecting the state of the vocal folds. GIF is also a non-invasive technique, meaning that no direct contact with the speech production mechanism is needed, making it a safe way of inspecting the vocal folds. In technology GIF is used for both artificial speech production and speech recognition.

GIF has been studied since the late 1950s, when Miller published the first study [22] on the subject. Since then many different methods for solving the problem have been proposed. Several recent comparative reviews of the different methods [3, 10, 34] have shown that both the *zeros of the z-transform* (ZZT) (or if formulated otherwise the *complex-cepstrum-based decomposition* (CCD)) [6, 7, 9, 30] and the *iterative adaptive inverse filtering* (IAIF) [2] methods of GIF perform well in most cases. However, both methods have their limitations, and no method has yet been proposed that would give robust glottal flow estimates in all circumstances.

A relatively new method, the so called *Markov chain Monte Carlo method for glottal inverse filtering* (MCMC-GIF) was proposed by Auvinen *et al.* in [5]. The authors showed that MCMC-GIF was able to yield better results than other state of the art methods for GIF in most cases.

Although GIF has been widely studied, only a few papers have been published on the subject in the mathematics and inverse problems communities. To the extent of the authors knowledge these papers are limited to [1, 15, 16, 20].

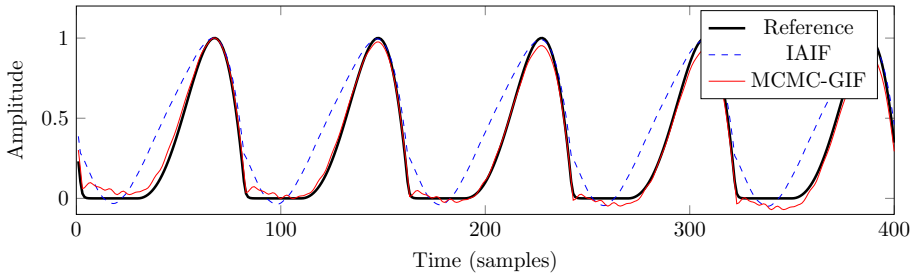


Figure 1.1: An example of the results obtained with the MCMC-GIF method compared with the IAIF method. The data used in the example was a synthetic vowel /i/ with the fundamental frequency 200 Hz.

In this thesis we will describe how the human speech production mechanism works and present a mathematical model (the source-filter model) for digital speech production. We will also describe models for solving the GIF problem and compare different methods of doing this. The emphasis of this thesis will be on the MCMC-GIF method, which we will explain in detail and give numerical examples of its performance. An example of the results obtained by MCMC-GIF compared with the IAIF method is shown in figure 1.1.

The thesis is structured as follows. In section 2 we will cover the fundamental theory of signal processing needed to describe digital speech modelling and production. In section 3 we will explain the direct problem, namely the modelling and simulation of digital speech, with an emphasis on vowel simulation. In section 4 we will explain the inverse problem of GIF and the methods we use to solve it. In section 5 we will present some numerical results obtained with different GIF methods.

2 Signal processing

To be able to give a mathematical representation of the human speech production mechanism and the resulting speech signal we will need some definitions and tools from the theory of signal processing. In this section we will give a short introduction to the necessary parts of discrete-time signals, discrete-time systems and transformations that we will need later on. For a more thorough presentation in the subject for example the book [23] by Oppenheim *et al.* is recommended.

2.1 Discrete-time signals

A discrete-time signal is a sequence of numbers (real or complex), often denoted as $x(n)$, where x denotes the signal and n is an integer variable. Thus, the signal can be thought of as a discrete function $x : \mathbb{Z} \to \mathbb{R}$ (or $x : \mathbb{Z} \to \mathbb{C}$). Even though the signals are in practical applications often real, we will present the theory using complex sequences, as it doesn't affect the formulation of the theorems or proofs presented.

The reason for us to consider discrete-time signals is the way that sound signals are represented digitally. A continuous (analog) sound signal needs to be sampled to a discrete signal for processing on for example a computer. The sampling is done with some predefined *sampling frequency* f_s (for example 44.1 kHz on typical audio CDs). A continuous signal $x_c(t)$ is then sampled into a discrete-time signal x as $x(n) = x_c(n/f_s)$.

According to the *Nyquist sampling theorem* it can be shown that an analog signal $x_c(t)$ with a bandlimit f_N (i.e., loosely speaking, the signal does not contain any frequencies higher than f_N) can be perfectly reconstructed from its (equally spaced) samples $x(n) = x_c(n/f_s)$, $n \in \mathbb{Z}$, if it holds for the sampling frequency f_s that $f_s > 2f_N$. The frequency f_N is often referred to as the *Nyquist frequency* [23].

As previously mentioned, the sampling frequency on typical audio CDs is 44.1 kHz. In practice this means that any sound signals with the frequency within the human hearing range (which ranges up to 20 kHz) can, in theory, be perfectly sampled on an audio CD (although in practice this is not quite true due to many reasons, such as measurement noise during the sound recording). In the applications we will be considering such high sampling rates are seldom used. The most important characteristics of human speech are limited to frequencies below 8 kHz [24], which is why in many speech processing applications, as in our case, the sampling rate is chosen to be 16 kHz.

An example of the sampling of a continuous signal can be seen in figure 2.1.

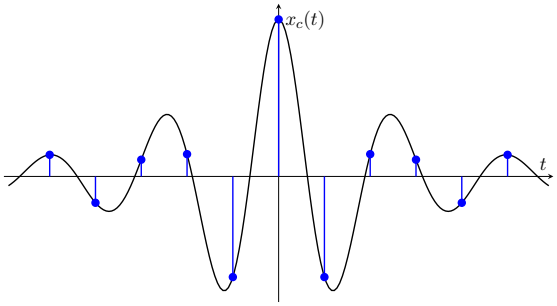


Figure 2.1: An example of the sampling of a continuous signal.

2.1.1 Some important signals

A couple of important fundamental signals, which we will are going to need in the introduction of signal processing, are the *unit sample* and the *unit step* signals.

2.1 Definition. The *unit sample* signal is defined as

$$\delta(n) = \begin{cases} 1, & n = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

2.3 Definition. The *unit step* signal is defined as

$$u(n) = \begin{cases} 1, & n \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.4}$$

One important property of the unit step signal is that it allows us to decompose any signal x as the sum

$$x(n) = \sum_{k=-\infty}^{\infty} x(k)\delta(n-k), \tag{2.5}$$

containing only scaled and shifted unit samples. This representation of signals will be a useful tool in some of the upcoming proofs.

2.2 Discrete-time systems

A *discrete-time system* (such as our model for the vocal tract will be) is an operator or mapping F, which maps a given input signal $x(n)$ to an output signal $y(n) = (F(x))(n)$. The mapping F can be expressed as an explicit mathematical function or an algorithm for transforming the input signal.

2.2.1 Properties of discrete-time systems

We will first present a couple of definitions of discrete time systems.

2.6 Definition. A discrete-time system F is called *linear* if we have for all signals $x_1(n)$ and $x_2(n)$ and constants c_1 and c_2 that

$$(F(c_1x_1 + c_2x_2))(n) = c_1(F(x_1))(n) + c_2(F(x_2))(n). \tag{2.7}$$

2.8 Definition. If a shift in the input signal of a system results in the same shift in the output signal, the system is said to be *time-invariant*.[1] In other words, if F is a time-invariant system and $y(n) = (F(x))(n)$, then for all $n_0 \in \mathbb{Z}$ we have $y(n - n_0) = (F(x))(n - n_0)$.

2.9 Definition. A system is called *causal*, if the output $y(n_0)$ of the system only depends on inputs $x(n)$ with $n \leq n_0$, for all $n_0 \in \mathbb{Z}$.

In other words, the causality of a system means that the output of the system may depend on any values of the input from the past, but on none from the future.

2.10 Definition. A system is called *stable* (in a *bounded-input, bounded-output* (BIBO) sense), if for every bounded input the system generates a bounded output. More precisely, assuming that for all $n \in \mathbb{Z}$ we have that $|x(n)| < C_1$ for some $C_1 \in \mathbb{R}$, then F is a stable system if there exists some $C_2 \in \mathbb{R}$ such that for the output $y(n) = (F(x))(n)$ we have $|y(n)| < C_2$ for all $n \in Z$.

2.2.2 Linear time-invariant systems

We will now introduce the so called *linear time-invariant* (LTI) system, which will be an important tool when defining the model for the vocal tract. An LTI system has the properties of both linear and time-invariant systems, as described earlier. The power of LTI systems is that the output signal of the system for any input can be calculated merely as the discrete convolution of the input signal and the so called *impulse response* signal. The system can therefore be completely characterized by its impulse response, which will allow us to easily analyse properties of the LTI system. We will first prove that this actually is the case and then present some theorems for the properties of LTI systems.

2.11 Definition. A *linear time-invariant* (LTI) system is a discrete-time system which is both linear and time-invariant.

[1] In a more general context this property is sometimes referred to as *shift-invariance*. However, because all the signals in this work are time domain signals we will be using the term time-invariance.

2.12 Definition. Let F be an LTI system. The *impulse response* of the system is defined as

$$h(n) = (F(\delta))(n). \tag{2.13}$$

2.14 Definition. Let x and y be discrete-time signals. The *discrete convolution* of x and y is defined as

$$(x * y)(n) = \sum_{k=-\infty}^{\infty} x(k)y(n-k). \tag{2.15}$$

2.16 Theorem. *Let h be the impulse response of an LTI system and let x be the input signal for the system. The output signal y can be calculated as*

$$y(n) = (x * h)(n)$$

Proof. Let h be the impulse response of an LTI system F, i.e.

$$h(n) = (F(\delta))(n).$$

Due to the time-invariant property of F we get that for any shift $n_0 \in \mathbb{Z}$ we have

$$h(n-n_0) = (F(\delta))(n-n_0). \tag{2.17}$$

Let now x be an input signal for the LTI system. Now we get for the output y using the linear property and equations (2.5) and (2.17) that

$$\begin{aligned} y(n) &= (F(x))(n) \\ &= \left(F\left(\sum_{k=-\infty}^{\infty} x(k)\delta(m-k)\right)\right)(n) \\ &= \sum_{k=-\infty}^{\infty} x(k)(F(\delta))(n-k) \\ &= \sum_{k=-\infty}^{\infty} x(k)h(n-k) = (x*h)(n). \end{aligned}$$

□

As the LTI systems are described by the discrete convolution between the impulse response and the input signal, we can use many well known properties of the convolution sum to get a better understanding of the systems. Next we will present some of the important properties for convolution sums.

2.18 Theorem. *The discrete convolution is commutative.*

Proof. Let $h_1(n)$ and $h_2(n)$ be sequences (e.g. discrete-time signals). Now with the substitution $k = n - m$ we get that

$$\begin{aligned}(h_1 * h_2)(n) &= \sum_{k=-\infty}^{\infty} h_1(k)h_2(n-k) \\ &= \sum_{m=-\infty}^{\infty} h_1(n-m)h_2(m) \\ &= \sum_{m=-\infty}^{\infty} h_2(m)h_1(n-m) \\ &= (h_2 * h_1)(n),\end{aligned}$$

and thus the convolution sum is commutative. □

2.19 Theorem. *The discrete convolution is distributive over addition.*

Proof. Let $x(n)$, $h_1(n)$ and $h_2(n)$ be sequences. We get that

$$\begin{aligned}(x * (h_1 + h_2))(n) &= \sum_{k=-\infty}^{\infty} x(k)(h_1(n-k) + h_2(n-k)) \\ &= \sum_{k=-\infty}^{\infty} (x(k)h_1(n-k) + x(k)h_2(n-k)) \\ &= \sum_{k=-\infty}^{\infty} x(k)h_1(n-k) + \sum_{k=-\infty}^{\infty} x(k)h_2(n-k) \\ &= (x * h_1)(n) + (x * h_2)(n),\end{aligned}$$

and thus the convolution is distributive over addition. □

2.20 Theorem. *The discrete convolution is associative, i.e. for signals h_i, $i = 1, 2, 3$ we have that*

$$h_1 * (h_2 * h_3) = (h_1 * h_2) * h_3.$$

Proof. Let h_i, $i = 1, 2, 3$ be sequences. Now we have that

$$\begin{aligned}(h_1 * (h_2 * h_3))(n) &= \sum_{k=-\infty}^{\infty} h_1(k)(h_2 * h_3)(n-k) \\ &= \sum_{k=-\infty}^{\infty} h_1(k) \left(\sum_{l=-\infty}^{\infty} h_2(l)h_3((n-k)-l) \right) \\ &= \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} h_1(k)h_2(l)h_3((n-k)-l)\end{aligned}$$

$$
\begin{aligned}
&= \sum_{l=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} h_1(k) h_2((k+l)-k) h_3(n-(k+l)) \\
&= \sum_{m=-\infty}^{\infty} \left(\sum_{k=-\infty}^{\infty} h_1(k) h_2(m-k) \right) h_3(n-m) \\
&= \sum_{m=-\infty}^{\infty} (h_1 * h_2)(m) h_3(n-m) \\
&= ((h_1 * h_2) * h_3)(n),
\end{aligned}
$$

using the substitution $m = k + l$, and we see that the discrete convolution is associative. □

Because the convolution sum, and thus LTI systems are commutative, it means that the order in which the systems are applied to a signal doesn't matter, the resulting output will be the same. Thus multiple systems can be thought of as a single system, applying the impulse response of all of the systems in question to the input signal.

The case of distributivity tells us how signals behave to several parallel LTI systems applied to an input signal. As the discrete convolution is distributive, it means that the system of several parallel LTI systems is equivalent to a single system where the impulse response is the pointwise sum of the impulse responses of the parallel systems.

For the stability of LTI systems we get an easy condition.

2.21 Theorem. *An LTI system is stable if and only if the impulse response of the system is absolutely summable.*

Proof. Let us first show that an absolutely summable impulse response defines a stable LTI system. Assume that we have a bounded input $|x(n)| < C_x$ and that

$$
\sum_{k=-\infty}^{\infty} |h(k)| = C_h,
$$

for some constants $C_x, C_h \in \mathbb{R}$. Now we can calculate for the output

$$
\begin{aligned}
|y(n)| = |(x * h)(n)| &= \left| \sum_{k=-\infty}^{\infty} h(k) x(n-k) \right| \\
&\leq \sum_{k=-\infty}^{\infty} |h(k)| \, |x(n-k)| \leq \sum_{k=-\infty}^{\infty} |h(k)| \, C_x \\
&= C_x \sum_{k=-\infty}^{\infty} |h(k)| = C_x C_h < \infty,
\end{aligned}
$$

which shows that the system is stable.

Next we want to show that every stable LTI system has an absolutely summable impulse response. This is equivalent to showing that if the impulse response is not absolutely summable, then the LTI system is not bounded (i.e. we can find some input which produces an unbounded output).

Assume now that that for the impulse response we have that

$$\sum_{k=-\infty}^{\infty} |h(k)| = \infty.$$

Let us now denote the indices of the non-zero values of the impulse response as $I := \{n \in \mathbb{Z} : h(n) \neq 0\}$, and define the input

$$x(n) = \begin{cases} \frac{\overline{h(-n)}}{|h(-n)|}, & n \in I \\ 0, & n \in \mathbb{Z} \setminus I. \end{cases}$$

Now clearly $|x(n)| \leq 1$ for all $n \in \mathbb{Z}$, but

$$\begin{aligned} y(0) = |(x * h)(0)| &= \sum_{k=-\infty}^{\infty} h(k)x(-k) = \sum_{k \in I} h(k)x(-k) \\ &= \sum_{k \in I} h(k)\frac{\overline{h(k)}}{|h(k)|} = \sum_{k \in I} \frac{|h(k)|^2}{|h(k)|} = \sum_{k \in I} |h(k)| = \sum_{k=-\infty}^{\infty} |h(k)| \\ &= \infty. \end{aligned}$$

We have now shown that if the impulse response is not absolutely summable, we can always find an input that produces an unbounded output, and hence we have shown that every stable LTI system has a bounded impulse response. This concludes our proof. □

For LTI systems we can also give an easy condition for causality.

2.22 Theorem. *An LTI system is causal if and only if it holds for the impulse response h that $h(n) = 0$ for all $n < 0$.*

Proof. Let us first prove that the mentioned condition holds for h, then the system is causal.

Let x, h be sequences, with $h(n) = 0$ for all $n < 0$. Now

$$y(n_0) = (x * h)(n_0) = \sum_{k=-\infty}^{\infty} x(k)h(n_0 - k) = \sum_{k=-\infty}^{n_0} x(k)h(n_0 - k),$$

because $h(n_0 - k) = 0$ when $k > n_0$. Now we see that the output signal $y(n_0)$

does not depend on any values $x(n)$ with $n > n_0$, which means that the system is causal.

Let us now prove that if the system is causal, then we have that $h(n) = 0$ for all $n < 0$. We will show this by showing that if $h(n) \neq 0$ for some $n < 0$, then the system cannot be causal.

Let x, h be sequences, with $h(-m) \neq 0$ for some fixed $m > 0$. Now

$$\begin{aligned} y(n_0) &= (x * h)(n_0) \\ &= \sum_{k=-\infty}^{\infty} x(k)h(n_0 - k) \\ &= \sum_{k=-\infty}^{n_0+m-1} x(k)h(n_0 - k) + x(n_0 + m)h(-m) \\ &\quad + \sum_{k=n_0+m+1}^{\infty} x(k)h(n_0 - k), \end{aligned}$$

and as we know the $h(-m) \neq 0$, we know that the output $y(n_0)$ always depends on a future value $x(n_0 + m)$, and thus the system cannot be causal. This concludes our proof. □

One important class of LTI systems are those systems, whose input $x(n)$ and output $y(n)$ satisfy the *linear constant coefficient difference equation*

$$\sum_{k=0}^{N} a_k y(n - k) = \sum_{k=0}^{M} b_k x(n - k). \tag{2.23}$$

We will see later on that the vocal tract can be described as such a system.

2.3 Frequency domain representations

We will now introduce some ways to represent discrete-time signals in the frequency domain. First, we will make a quick reminder about definition of eigenfunctions and eigenvalues, and then inspect the eigenfunctions and eigenvalues of LTI systems.

2.24 Definition. The function f is an *eigenfunction* of the operator H, if it holds that

$$Hf = \lambda f,$$

for some constant λ. Here, λ is called the *eigenvalue* corresponding to the eigenfunction f of the operator H.

2.25 Theorem. *Let h be the impulse response of an LTI system. Now the complex exponential $x(n) = e^{i\omega n}$, $\omega \in \mathbb{R}$, is an eigenfunction for the LTI system, with the eigenvalue*

$$H(\omega) = \sum_{k=-\infty}^{\infty} h(k)e^{-i\omega k}.$$

Proof. Let h be the impulse response of an LTI system, and let $x(n) = e^{i\omega n}$, for some fixed $\omega \in \mathbb{R}$. We have now that

$$\begin{aligned} y(n) = (h * x)(n) &= \sum_{k=-\infty}^{\infty} h(k)x(n-k) \\ &= \sum_{k=-\infty}^{\infty} h(k)e^{i\omega(n-k)} = e^{i\omega n} \sum_{k=-\infty}^{\infty} h(k)e^{-i\omega k}. \end{aligned}$$

If we denote

$$H(\omega) = \sum_{k=-\infty}^{\infty} h(k)e^{-i\omega k},$$

we can write $y(n)$ as

$$y(n) = (h * x)(n) = x(n)H(\omega),$$

and we see that x is an eigenfunction of the LTI system with the eigenvalue $H(\omega)$. □

The function $H(\omega)$ defined in theorem 2.25 is called the *frequency response* of the LTI system. The frequency response defines both the change in magnitude and phase of an input signal by the impulse response, represented by $|H(\omega)|$ and $\arg(H(\omega))$, respectively.

The frequency response has a couple of important properties, which we will present next.

2.26 Theorem. *The frequency response of LTI systems is periodic with the period 2π.*

Proof. Let $H(\omega)$ be the frequency response of an LTI system with the corresponding impulse response $h(n)$, and let $n \in \mathbb{Z}$ be fixed. Now

$$\begin{aligned} H(\omega + 2\pi n) &= \sum_{k=-\infty}^{\infty} h(k)e^{-i(\omega+2\pi n)k} = \sum_{k=-\infty}^{\infty} h(k)e^{-i\omega k}e^{-2i\pi nk} \\ &= \sum_{k=-\infty}^{\infty} h(k)e^{-i\omega k} = H(\omega), \end{aligned}$$

which proves the theorem. □

2.27 Theorem. *Let $H(\omega)$ be the frequency response of an LTI system with the impulse response $h(n)$. Now the frequency response $\overline{H(-\omega)}$ describes an LTI system with the impulse response $\overline{h(n)}$.*

Proof. Let $H(\omega)$ be the frequency response corresponding to an LTI system with the impulse response $h(n)$. Now

$$\begin{aligned}\overline{H(-\omega)} &= \overline{\sum_{k=-\infty}^{\infty} h(k)e^{-i(-\omega)k}} = \sum_{k=-\infty}^{\infty} \overline{h(k)e^{i\omega k}} \\ &= \sum_{k=-\infty}^{\infty} \overline{h(k)}e^{-i\omega k},\end{aligned}$$

as claimed. □

2.28 Corollary. *If the impulse response $h(n)$ of an LTI system is a real sequence, then we have for the frequency response that $H(-\omega) = \overline{H(\omega)}$.*

Proof. According to theorem 2.27 and the fact that $h(n)$ is real for all $n \in \mathbb{Z}$ we get that

$$H(-\omega) = \overline{\sum_{k=-\infty}^{\infty} \overline{h(k)}e^{-i\omega k}} = \overline{\sum_{k=-\infty}^{\infty} h(k)e^{-i\omega k}} = \overline{H(\omega)}.$$

□

The above theorem also implies, that the magnitude of the frequency response for real valued impulse responses is an even function, $|H(\omega)| = |H(-\omega)|$.

Because the frequency response of LTI systems is periodic with the period 2π and conjugate symmetric around zero, it is usual to restrict the values of the variable ω to the range $[-\pi, \pi]$ or $[0, \pi]$. Here the values near zero represent the low frequencies, and the values near $\pm\pi$ represent the high frequencies.

2.3.1 The discrete-time Fourier transform

We will now explain how arbitrary signals can be expressed in the frequency domain, much like we did in the case of the frequency response for the LTI systems. The *discrete-time Fourier transform* (DTFT) is an important tool in such frequency domain analysis of signals.

2.29 Definition. The discrete-time Fourier transform of a sequence $x(n)$, denoted by $\mathscr{F}\{x\}$, is defined as

$$X(\omega) = \sum_{k=-\infty}^{\infty} x(k)e^{-i\omega k}. \tag{2.30}$$

As one can easily see, the DTFT is defined in the same way as the frequency response of an LTI system. In other words, the frequency response of an LTI system is just the DTFT of the system's impulse response.

However, the sum in equation (2.30) does not always converge, meaning that not all sequences have a DTFT. Before showing a condition for the convergence, let us review a couple of examples.

2.31 Example. Let us calculate the DTFT $\mathscr{F}\{\delta\}$ for the unit sample. We get the DTFT

$$\Delta(\omega) = \sum_{k=-\infty}^{\infty} \delta(k)e^{-i\omega k} = \delta(0) = 1.$$

So the DTFT of the unit sample is a constant function 1.

2.32 Example. For the DTFT of the unit step u we get

$$U(\omega) = \sum_{k=-\infty}^{\infty} u(k)e^{-i\omega k} = \sum_{k=0}^{\infty} e^{-i\omega k} = \frac{e^{i\omega}}{e^{i\omega} - 1}.$$

This is however not defined at $\omega = 0$, as $|U(\omega)| \to \infty$, as $\omega \to 0$. This means that the unit step does not have a DTFT.

Let us now give a condition for the convergence of the DTFT.

2.33 Theorem. *The DTFT converges uniformly to a continuous function for absolutely summable sequences.*

Proof. Let us assume that $x(n)$ is an absolutely summable sequence, i.e.

$$\sum_{k=-\infty}^{\infty} |x(k)| = S < \infty,$$

for some $S \in \mathbb{R}$.

First, we want to show that the DTFT described in equation (2.30) actually converges to a continuous function for the sequence $x(n)$. After this we want to prove that the convergence is uniform.

Let us begin with the first step of the proof. We get for the DTFT that

$$\begin{aligned} |X(\omega)| &= \left| \sum_{k=-\infty}^{\infty} x(n)e^{-i\omega k} \right| \leq \sum_{k=-\infty}^{\infty} |x(n)| \left| e^{-i\omega k} \right| \\ &= \sum_{k=-\infty}^{\infty} |x(n)| = S < \infty, \end{aligned}$$

so the series converges for all $\omega \in \mathbb{R}$. The continuity of X comes directly from the continuity of the exponential function $\omega \mapsto e^{\omega}$.

We now need to show that the partial sum

$$X_n(\omega) = \sum_{k=-n}^{n} x(n)e^{-i\omega k}$$

convergences uniformly toward the function X.

Let $\varepsilon > 0$, and let us denote the upper and lower partial sums as

$$S_n := \sum_{k=-n}^{\infty} |x(k)|$$

and

$$S^n := \sum_{k=-\infty}^{n} |x(k)| .$$

Because $x(n)$ is absolutely summable, we know for the partial sum that there exists some $N_1 \in \mathbb{N}$, such that

$$|S - S_{N_1}| = S - S_{N_1} = \sum_{k=-\infty}^{-N_1-1} |x(k)| < \frac{\varepsilon}{2}.$$

Similarly, we know that there exists some $N_2 \in \mathbb{N}$, such that

$$\left|S - S^{N_2}\right| = S - S^{N_2} = \sum_{k=N_2+1}^{\infty} |x(k)| < \frac{\varepsilon}{2}.$$

Let now $N = \max\{N_1, N_2\}$ and $\omega \in \mathbb{R}$. We get

$$\begin{aligned}
|X(\omega) - X_N(\omega)| &= \left| \sum_{k=-\infty}^{\infty} x(k)e^{-i\omega k} - \sum_{k=-N}^{N} x(k)e^{-i\omega k} \right| \\
&= \left| \sum_{k=-\infty}^{-N-1} x(k)e^{-i\omega k} + \sum_{k=N+1}^{\infty} x(k)e^{-i\omega k} \right| \\
&\leq \left| \sum_{k=-\infty}^{-N-1} x(k)e^{-i\omega k} \right| + \left| \sum_{k=N+1}^{\infty} x(k)e^{-i\omega k} \right| \\
&\leq \sum_{k=-\infty}^{-N-1} |x(k)| \left|e^{-i\omega k}\right| + \sum_{k=N+1}^{\infty} |x(k)| \left|e^{-i\omega k}\right| \\
&= \sum_{k=-\infty}^{-N-1} |x(k)| + \sum_{k=N+1}^{\infty} |x(k)| \\
&< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

□

2.34 Corollary. *Every stable LTI system has a continuous frequency response.*

Proof. Follows directly from theorems 2.21 and 2.33. □

We can now introduce the inverse DTFT.

2.35 Theorem. *The inverse mapping the DTFT (the inverse discrete-time Fourier transform, IDTFT) is*

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\omega) e^{i\omega n} \,\mathrm{d}\omega. \tag{2.36}$$

We denote the inverse mapping as $x = \mathscr{F}^{-1}\{X\}$.

Proof. Let $x(n)$ be a absolutely summable sequence, and $X(\omega)$ its DTFT. Let us now denote

$$y(n) := \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\sum_{k=-\infty}^{\infty} x(k) e^{-i\omega k} \right) e^{i\omega n} \,\mathrm{d}\omega.$$

We would now like to show that $y(n) = x(n)$. We know from theorem 2.33 that the infinite sum in our expression converges uniformly. This allows us to change the order of the sum and the integral, yielding

$$\begin{aligned}
y(n) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\sum_{k=-\infty}^{\infty} x(k) e^{-i\omega k} \right) e^{i\omega n} \,\mathrm{d}\omega \\
&= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} x(k) \left(\int_{-\pi}^{\pi} e^{-i\omega k} e^{i\omega n} \,\mathrm{d}\omega \right) \\
&= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} x(k) \left(\int_{-\pi}^{\pi} e^{i\omega(n-k)} \,\mathrm{d}\omega \right) \\
&= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} x(k) \frac{2\sin(\pi(n-k))}{n-k} \\
&= \sum_{k=-\infty}^{\infty} x(k) \frac{\sin(\pi(n-k))}{\pi(n-k)} \\
&= \sum_{k=-\infty}^{\infty} x(k) \delta(n-k) \\
&= x(n),
\end{aligned}$$

which proves our claim. □

Some of the central properties of the DTFT are listed in table 2.1. The proofs for the identities can be found in appendix A. In the table $x(n)$ and $y(n)$ are sequences, $X(\omega)$ and $Y(\omega)$ their respective DTFTs, $c_1, c_2, \omega_0 \in \mathbb{R}$ and $n_0 \in \mathbb{Z}$.

Property	Transform		
Linearity	$\mathscr{F}\{c_1x(n)+c_2y(n)\}$	$=$	$c_1X(\omega)+c_2Y(\omega)$
Delay	$\mathscr{F}\{x(n-n_0)\}$	$=$	$e^{-i\omega n_0}X(\omega)$
Modulation	$\mathscr{F}\left\{e^{i\omega_0 n}x(n)\right\}$	$=$	$X(\omega-\omega_0)$
Conjugation	$\mathscr{F}\left\{\overline{x(n)}\right\}$	$=$	$\overline{X(-\omega)}$
Time reversal	$\mathscr{F}\{x(-n)\}$	$=$	$X(-\omega)$
Convolution	$\mathscr{F}\{(x*y)(n)\}$	$=$	$X(\omega)Y(\omega)$

Table 2.1: Properties of the DTFT.

One of the most important properties for the DTFT is the convolution property (also know as the convolution theorem). The property states that the DTFT of the convolution of two sequences is just the product of the sequences' DTFTs. What makes this particularly valuable is, that the convolution is computationally relatively expensive (the discrete convolution has the time complexity of $\mathcal{O}\left(n^2\right)$), whereas the DTFT using fast Fourier transform (FFT) algorithms is computationally very efficient (having the time complexity of $\mathcal{O}\left(n\log(n)\right)$). This way the convolution can be done efficiently by first taking the DTFT of the sequences, then multiplying them pointwise, and then applying the IDTFT.

Next we are going to introduce a generalization of the DTFT, the so called z-transform. This will be an important transformation in the modelling of the vocal tract.

2.3.2 The z-transform

The z-transform is a generalization of the DTFT. As the DTFT is defined on the unit circle in the complex plane (or more specifically, on the range $\omega \in [-\pi, \pi]$, which is the mapped to the point $e^{i\omega}$ on the unit circle), the z-transform generalizes this to the whole complex plane.

2.37 Definition. Let $x(n)$ be a sequence. The *z-transform* of $x(n)$ is defined as

$$\mathcal{Z}\{x(n)\} = \sum_{k=-\infty}^{\infty} x(k)z^{-k}. \tag{2.38}$$

2.39 Theorem. *The z-transform is equivalent to the DTFT, when* $|z| = 1$. *More specifically*

$$X_{\mathscr{F}}(\omega) = X_{\mathcal{Z}}\left(e^{i\omega}\right), \tag{2.40}$$

where $X_{\mathscr{F}}$ *is the DTFT and* $X_{\mathcal{Z}}$ *is the z-transform of the sequence* $x(n)$.

Property	Transform		
Linearity	$\mathcal{Z}\{c_1x(n)+c_2y(n)\}$	$=$	$c_1X(z)+c_2Y(z)$
Delay	$\mathcal{Z}\{x(n-n_0)\}$	$=$	$z^{-n_0}X(z)$
Multiplication by z_0^n	$\mathcal{Z}\{z_0^n x(n)\}$	$=$	$X(z/z_0)$
Conjugation	$\mathcal{Z}\left\{\overline{x(n)}\right\}$	$=$	$\overline{X(\overline{z})}$
Time reversal	$\mathcal{Z}\{x(-n)\}$	$=$	$X(z^{-1})$
Convolution	$\mathcal{Z}\{(x*y)(n)\}$	$=$	$X(z)Y(z)$

Table 2.2: Properties of the z-transform.

Proof. Let $X_{\mathscr{F}}$ be the DTFT and $X_{\mathcal{Z}}$ the z-transform of the sequence $x(n)$, and $\omega \in \mathbb{R}$. Now

$$X_{\mathscr{F}}(\omega) = \sum_{k=-\infty}^{\infty} x(k)e^{-i\omega k} = \sum_{k=-\infty}^{\infty} x(k)\left(e^{i\omega}\right)^{-k} = X_{\mathcal{Z}}\left(e^{i\omega}\right).$$

□

Some important properties of the z-transform are listed in table 2.2. The proofs of the properties can be found in appendix A.

Series of the type defined in equation (2.38) are so called *Laurent series* familiar from complex analysis. We will not go through the theory of Laurent series in detail here, but we will use some results from the theory, which we will present when needed. For a good reference work on the subject, see for example [29] by Stewart and Tall.

The Laurent series of an analytic function f in an annulus around a point $z_0 \in \mathbb{C}$ is defined as

$$f(z) = \sum_{n=-\infty}^{\infty} a_n (z-z_0)^n, \tag{2.41}$$

where $a_n \in \mathbb{C}$ are constants defined in a specific manner by the function f. The series in equation (2.38) falls within this definition, which we can easily see with the substitution $n = -k$.

We know, that every Laurent series has a *region of convergence* (ROC), where the series converges to an analytic function (i.e. a function which is infinitely differentiable in a neighbourhood around every point). The convergence is also uniform in all compact subsets of the ROC. More specifically, for every Laurent series defined as in equation (2.41) there exists two radii R_1 and R_2, such that the series converges in the open annulus $\{z \in \mathbb{C} : R_1 < |z-z_0| < R_2\}$. Here we might have $R_1 = 0$ or $R_2 = \infty$.

In many cases we might have a z-transform which consists of a sum of many

Laurent series (due to the linear property of the z-transform). This is still a Laurent series, and its ROC will be the intersection of the sums' ROCs.

Let us now formulate a condition for the convergence of the z-transform on a specific circle around the origin. The proof of the theorem comes directly from the convergence of the DTFT of an exponentially weighted sequence, as we will see.

2.42 Theorem. *The z-transform of the sequence $x(n)$ converges uniformly for $|z| = r$, $r > 0$, if the sequence $x(n)r^{-n}$ is absolutely summable, i.e.*

$$\sum_{k=-\infty}^{\infty} \left|x(k)r^{-k}\right| < \infty.$$

Proof. Let $x(n)$ be a sequence, $r > 0$ and $x(n)r^{-n}$ absolutely summable. Now, according to theorem 2.33, the DTFT

$$\sum_{k=-\infty}^{\infty} x(k)r^{-k}e^{-i\omega k}$$

converges uniformly. This, however, is precisely the z-transform $X(z)$ of $x(n)$ at $|z| = r$, because

$$\sum_{k=-\infty}^{\infty} x(k)r^{-k}e^{-i\omega k} = \sum_{k=-\infty}^{\infty} x(k)\left(re^{i\omega}\right)^{-k} = X\left(re^{i\omega}\right).$$

Thus the sequence $x(n)$ converges uniformly for $|z| = r$. □

The previous theorem states, that the convergence of the z-transform at the point $z \in \mathbb{C}$ is only dependent on the magnitude $|z|$. This is consistent with the previously mentioned fact that Laurent series have an annular region of convergence around the point about which they are defined (the origin in the case of z-transforms). In practice this means, that if the z-transform converges at a point $z_0 \in \mathbb{C}$, it will also converge for all other values with the magnitude $|z_0|$ (and actually also for all magnitudes in some neighbourhood of $|z_0|$, as Laurent series always converge in an *open* annulus, but we will not go into further detail regarding this).

Let us now take a look at a couple of examples regarding z-transforms and their ROCs.

2.43 Example. Let $x(n) = a^n u(n)$, for some $a \in \mathbb{R}$, where $u(n)$ is the unit step signal. Now we get the z-transform

$$X(z) = \sum_{k=-\infty}^{\infty} a^k u(k) z^{-k} = \sum_{k=0}^{\infty} a^k z^{-k} = \sum_{k=0}^{\infty} \left(\frac{a}{z}\right)^k = \frac{1}{1 - az^{-1}},$$

when $|z| > |a|$. Here the resulting function gets a pole at $z \in \mathbb{C}$ where $1 - az^{-1} = 0$, i.e. at $z = a$.

2.44 Example. Let us look at the z-transform of a combination of sequences like the one in theorem 2.43. Let $x(n) = a^n u(n) + b^n u(n)$ for $a, b \in \mathbb{R}$. Now according to theorem 2.43, when $|z| > |a|$ and $|z| > |b|$ we get the z-transform

$$\begin{aligned} X(z) &= \frac{1}{1 - az^{-1}} + \frac{1}{1 - bz^{-1}} = \frac{1 - bz^{-1} + 1 - az^{-1}}{(1 - az^{-1})(1 - bz^{-1})} \\ &= \frac{2 - (a+b)z^{-1}}{(1 - az^{-1})(1 - bz^{-1})} = \frac{z^{-2}(2z^2 - (a+b)z)}{z^{-2}(z-a)(z-b)} \\ &= \frac{2z^2 - (a+b)z}{(z-a)(z-b)}. \end{aligned}$$

Here the resulting function has zeros at $z = 0$ and $z = \frac{a+b}{2}$ and (first order) poles at $z = a$ and $z = b$, and the series converges when $|z| > \max\{|a|, |b|\}$.

2.45 Example. Let us look at one more example of a sum of two sequences, namely $x(n) = a^n u(n) - b^n u(-n-1)$. We know the z-transform of the first term from the previous examples, but we need to calculate it for the second term. We get

$$\begin{aligned} \mathcal{Z}\left\{-b^n u(-n-1)\right\} &= -\sum_{k=-\infty}^{\infty} b^k u(-k-1) z^{-k} = -\sum_{k=-\infty}^{-1} b^k z^{-k} \\ &= -\sum_{k=1}^{\infty} b^{-k} z^k = 1 - \sum_{k=0}^{\infty} \left(b^{-1} z\right)^k = 1 - \frac{1}{1 - b^{-1} z} \\ &= \frac{-b^{-1} z}{1 - b^{-1} z} = \frac{-b^{-1} z}{-b^{-1}(z - b)} = \frac{z}{z(1 - bz^{-1}} = \frac{1}{1 - bz^{-1}}, \end{aligned}$$

when $\left|b^{-1} z\right| < 1$, or equivalently $|z| < |b|$. Now we get for the z-transform of $x(n)$ with the linear property that

$$\begin{aligned} X(z) &= \frac{1}{1 - az^{-1}} + \frac{1}{1 - bz^{-1}} = \frac{1 - bz^{-1} + 1 - az^{-1}}{(1 - az^{-1})(1 - bz^{-1})} \\ &= \frac{2 - (a+b)z^{-1}}{(1 - az^{-1})(1 - bz^{-1})} = \frac{z^{-2}(2z^2 - (a+b)z)}{z^{-2}(z-a)(z-b)} \\ &= \frac{2z^2 - (a+b)z}{(z-a)(z-b)}, \end{aligned}$$

and the z-transform converges when $|z| > |a|$ and $|z| < |b|$, or $|a| < |z| < |b|$. We can note that the algebraic expression (and thus also the zeros and poles) of the z-transform are precisely the same as in example 2.44, even though the signal is different. The ROC for the z-transform is however different.

As we saw in the previous examples that in the case where the z-transform can be written as a rational function, it is enough to specify the ROC and the zeros and poles of the z-transform to specify the complete expression for the transform. This is because we can uniquely find the polynomials for the numerator and the denominator based on the poles and zeros. We will formalize this in a theorem a little bit later on.

Let us now take a look at some properties the z-transforms related to LTI systems, and then continue to a special case of LTI systems, namely those which can be defined by a rational system function.

2.46 Definition. Let $h(n)$ be the impulse response of an LTI system. Its z-transform, denoted as $H(z)$ is called the *system function* of the system.

2.47 Theorem. *Let $H(z)$ be the system function of an LTI system and $X(z)$ the z-transform of the input signal. Now the z-transform of the output signal is*

$$Y(z) = X(z)H(z).$$

Proof. Let $h(n)$ be the impulse response of the LTI system, $H(z)$ its z-transform (the system function), $x(n)$ the input signal and $X(z)$ its z-transform. Now the output of the system is $y(n) = (h * x)(n)$. Directly from the convolution property of the z-transform we now get that

$$\mathcal{Z}\{y(n)\} = \mathcal{Z}\{(h * x)(n)\} = \mathcal{Z}\{h(n)\}\,\mathcal{Z}\{x(n)\} = H(z)X(z).$$

□

2.3.3 Systems with rational system functions

We will now take a closer look at a specific set of LTI systems, namely those which can be described by a rational system function.

2.48 Theorem. *Let $H(z)$ be the system function of an LTI system, which satisfies the linear constant coefficient difference equation defined in equation* (2.23)*. Now the system function can be written as*

$$H(z) = \frac{\sum_{k=0}^{M} b_k z^{-k}}{\sum_{k=0}^{N} a_k z^{-k}}. \tag{2.49}$$

We call such a system function a rational system function*.*

Proof. Let $H(z)$ be the system function of an LTI system satisfying equation (2.23). Let $x(n)$ be the input signal and $y(n)$ the resulting output signal,

and $X(z)$ and $Y(z)$ their respective z-transforms. Now taking the z-transform on both sides of equation (2.23) we get

$$\mathcal{Z}\left\{\sum_{k=-0}^{N} a_k y(n-k)\right\} = \sum_{k=0}^{N} a_k \mathcal{Z}\left\{y(n-k)\right\} = \sum_{k=0}^{N} a_k z^{-k} Y(z)$$
$$= Y(z)\sum_{k=0}^{N} a_k z^{-k},$$

and similarly for the other side

$$\mathcal{Z}\left\{\sum_{k=0}^{M} b_k x(n-k)\right\} = \sum_{k=0}^{M} b_k \mathcal{Z}\left\{x(n-k)\right\} = \sum_{k=0}^{M} b_k z^{-k} X(z)$$
$$= X(z)\sum_{k=0}^{M} b_k z^{-k},$$

and we now get the equation

$$Y(z)\sum_{k=0}^{N} a_k z^{-k} = X(z)\sum_{k=0}^{M} b_k z^{-k}.$$

Recalling from theorem 2.47 that for an LTI system we have that $Y(z) = H(z)X(z)$, we can write this as

$$\frac{\sum_{k=0}^{M} b_k z^{-k}}{\sum_{k=0}^{N} a_k z^{-k}} = \frac{Y(z)}{X(z)} = H(z),$$

which is what we set out to prove. □

2.50 Corollary. *Every LTI system with a rational system function can be characterized by its poles and zeros, within a linear scaling factor.*

Proof. Both the numerator and the denominator in equation (2.49) are polynomials of the variable z^{-1}, with the degrees M and N, respectively. This means that we can factorize the polynomials with their respective roots, call them $c_k \in \mathbb{C}$, $k = 1, \ldots, M$, for the numerator and $p_k \in \mathbb{C}$, $k = 1, \ldots, M$, for the denominator. We can now write the system function in equation (2.49) as

$$H(z) = \frac{b_0}{a_0}\frac{\prod_{k=1}^{M}\left(1 - c_k z^{-1}\right)}{\prod_{k=1}^{N}\left(1 - p_k z^{-1}\right)}. \tag{2.51}$$

Now, as defined, c_k are the zeros and p_k are poles of the system function. □

As the the systems with a rational system function can be described only with their zeros and poles, we can also describe the frequency response of such a system with the help of these (if it exists). In fact, the magnitude of the frequency response gets a nice expression in this way.

2.52 Theorem. *The magnitude of the frequency response of an LTI system with a rational system function can (if it exists) be written as*

$$|H(\omega)| = |B| \frac{\prod_{k=1}^{M} \left|c_k - e^{i\omega}\right|}{\prod_{k=1}^{N} |p_k - e^{i\omega}|}, \tag{2.53}$$

where c_k are the zeros and p_k the poles of the system function, and $B = b_0/a_0$ a scaling factor. The frequency response in a point on the unit circle is thus just the ratio between the product of the zeros' distances and the poles' distances to that point.

Proof. Let $H_{\mathcal{Z}}$ be the z-transform of an LTI system, with

$$H_{\mathcal{Z}}(z) = B \frac{\prod_{k=1}^{M} \left(1 - c_k z^{-1}\right)}{\prod_{k=1}^{N} \left(1 - p_k z^{-1}\right)},$$

and assume that the frequency response exists (i.e. the DTFT converges, or the z-transform converges on the unit circle). As mentioned earlier in theorem 2.39, we can now write the frequency response as

$$H(\omega) = H_{\mathcal{Z}}\left(e^{i\omega}\right) = B \frac{\prod_{k=1}^{M} \left(1 - c_k e^{-i\omega}\right)}{\prod_{k=1}^{N} \left(1 - p_k e^{-i\omega}\right)}.$$

We can now write the magnitude of the frequency response as

$$\begin{aligned} |H(\omega)| &= \left| B \frac{\prod_{k=1}^{M} \left(1 - c_k e^{-i\omega}\right)}{\prod_{k=1}^{N} \left(1 - p_k e^{-i\omega}\right)} \right| = |B| \frac{\prod_{k=1}^{M} \left|1 - c_k e^{-i\omega}\right|}{\prod_{k=1}^{N} |1 - p_k e^{-i\omega}|} \\ &= |B| \frac{\prod_{k=1}^{M} \left|e^{-i\omega}\right| \left|e^{i\omega} - c_k\right|}{\prod_{k=1}^{N} \left|e^{-i\omega}\right| |e^{i\omega} - p_k|} = |B| \frac{\prod_{k=1}^{M} \left|e^{i\omega} - c_k\right|}{\prod_{k=1}^{N} |e^{i\omega} - p_k|}, \end{aligned}$$

as we wanted. □

2.54 Note. The frequency response is often presented in logarithmic form in decibels (dB). We get the so called *log magnitude* as

$$20\log_{10}(|H(\omega)|) = 20\log_{10}(|B|) + \sum_{k=1}^{M} 20\log_{10}\left(\left|e^{i\omega} - c_k\right|\right) - \sum_{k=1}^{N} 20\log_{10}\left(\left|e^{i\omega} - p_k\right|\right).$$

One thing we would still be interested in is the location of the poles for the system function. For certain types of systems this can be shown quite specifically. We will need to present a lemma first.

2.55 Lemma. *If $x(n)$ is a bounded right-sided sequence (i.e. there exists a number $N \in \mathbb{Z}$ such that $x(n) = 0$ for all $n < N$), then the ROC of the sequence's z-transform is $|z| > R$ for some $R > 0$.*

Proof. Let $x(n)$ be a sequence with $x(n) = 0$ when $n < N$ for some $N \in \mathbb{Z}$. Also, let $C > 0$ such that $|x(n)| < C$ for all $n \in \mathbb{Z}$. Now we get that

$$\sum_{k=-\infty}^{\infty} \left|x(k)r^{-k}\right| = \sum_{k=N}^{\infty} |x(k)|\, r^{-k} = \sum_{k=N}^{\infty} Cr^{-k} = C\sum_{k=N}^{\infty} \left(r^{-1}\right)^k = C\frac{-\left(r^{-1}\right)^{-N}}{r^{-1}-1} = C\frac{r^{1-N}}{r-1},$$

which converges when $\left|r^{-1}\right| < 1$, or $r > 1$. According to theorem 2.42 also the z-transform of $x(n)$ converges for $r > 1$. □

Note. What theorem 2.55 actually says, is that if a bounded right-sided sequence converges for some $z_0 \in \mathbb{C}$, it will also converge for all $|z| > |z_0|$.

2.56 Theorem. *All the poles of the system function of a stable and causal LTI system are located inside the unit disc $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$.*

Proof. Recall from theorem 2.34, that every stable LTI system has a continuous frequency response. This means that the DTFT of the system converges, which again means that the z-transform converges for $|z| = 1$. From theorem 2.22 we know that a causal LTI system is right-sided, and thus by theorem 2.55 we now know that the z-transform converges at least for $|z| \geq 1$. As the ROC cannot contain any poles (as the z-transform by definition diverges at poles), we can conclude that all the poles of the system are inside the unit disc. □

3 The direct problem – digital speech

The aim of this section is to give an overview of the human speech apparatus by first describing how the speech production mechanism works and then describing the whole system mathematically. The system will be described in terms of a simplified so called *source-filter theory*, which is a widely used model for digital speech production. The source-filter theory will be described in sections 3.3 and 3.4. We will first describe conceptually the different parts of the speech production mechanism and then give a mathematical description for them one by one, using the tools described in section 2.

3.1 The speech production mechanism

Human speech can be said to consist of three major parts: the respiratory system (i.e. the lungs), the larynx and the vocal tract. Speech is produced by pressing out air from the lungs, which then travels through the larynx and the vocal tract both of which modify the airflow in a specific manner. In voiced speech (which we will be concentrating on) the larynx modifies the constant airflow to characteristic pulses, known as the *glottal flow*. The vocal tract, which consists of the pharynx and the oral and nasal cavities, then modifies frequency spectrum of the flow with resonances and anti-resonances depending on the tract's shape. The flow is then finally radiated through the lips and nostrils to the surrounding air, creating the speech signal.

Speech is created by combining signals of the type described above. Only a part of the different sounds in speech are voiced, the rest are different unvoiced sounds where the larynx lets through the airflow from the lungs unaltered. Although we will be concentrating on voiced sounds, we will also give a short review of the different kinds of unvoiced sounds in section 3.1.3, where we will also describe the other different kinds of speech sounds.

3.1.1 Glottal excitation

As described earlier, in voiced speech the flow from the lungs is modified at the larynx to a periodic signal. This is done by two elastic flaps know as the *vocal folds*. The opening between the vocal folds is known as the *glottis*. In voiced speech the muscles of the larynx tighten the vocal folds which then start to vibrate due to the pressure from the lungs; first the glottis is closed and pressure builds up behind the vocal folds, then the pressure forces the glottis open and releases a pulse of air, releasing the pressure and causing the glottis to close again. This mechanism then repeats creating a periodic pulse know as the *glottal excitation* or *glottal flow*.

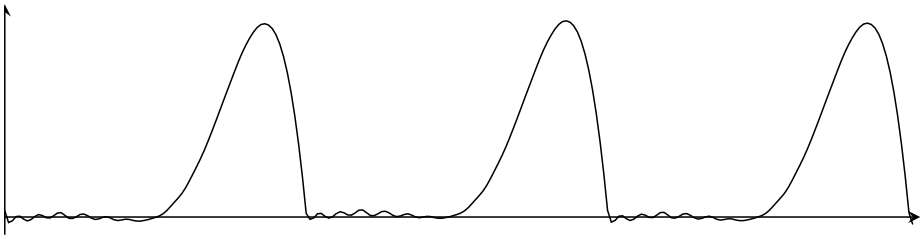


Figure 3.1: A glottal flow signal acquired with glottal inverse filtering.

The glottal flow plays an important role in the quality and characteristics of the produced speech sound, in particular with vowel sounds. The glottal flow cannot be directly measured due to the larynx's position deep in the throat, which creates a challenge for the creation of digital speech; the modelling of the signal is a hard task when no complete reference signal can be presented. One of the main aims of this thesis is, apart from giving a description of the modelling of digital speech, to describe means to recreate the signal for the glottal flow from a recorded sound signal. This task is called *glottal inverse filtering* (GIF), and it will be reviewed in detail in section 4.

A couple of different models for the modelling of the glottal flow will be presented in section 3.3. An example of the vocal flow, acquired by glottal inverse filtering, is shown in figure 3.1.

3.1.2 Vocal tract

The vocal tract include the pharynx and the oral and nasal cavities and is the primary source shaping the frequency spectrum of the speech signals. The vocal tract can be thought of as a tube from the larynx to the lips, with a branch to the nasal cavity. In sounds where the glottal airflow is not obstructed the cross-sectional area of the vocal tract is what mostly defines the resonant and anti-resonant frequencies of the tract. The resonant frequencies of the vocal tract are called *formants*, and can be seen in the frequency spectrum of the vocal tract system as peaks. The airflow can also be modified by obstructing the airflow at some point, creating different turbulent effects resulting most often in different consonants. Also the amount of flow directed to the nasal cavity can be adjusted.

The vocal tract varies from person to person, giving each speaker a characteristic sound and allowing people to be recognized by the sound of their speech. The personal differences in the vocal tract are however smaller than the similarities when uttering the same sound, so different sound (for example vowels) can be characterized in a general context for all speakers.

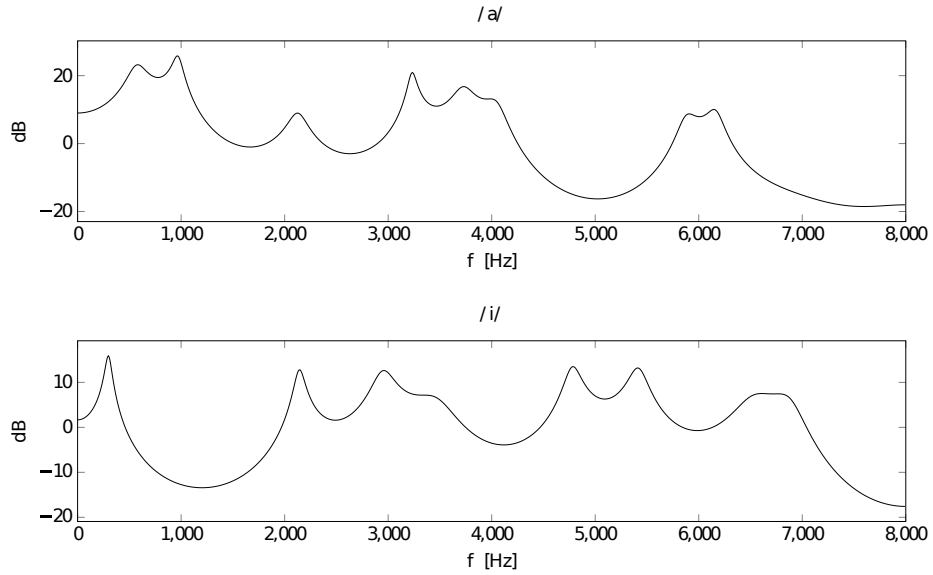


Figure 3.2: The frequency spectrum of the vowels /a/ and /i/.

3.1.3 Categorization of speech sounds

A typical way of categorizing speech sounds is to divide them into *vowels* and *consonants*. Vowels are said to be the voiced sounds where the airflow from the larynx is not obstructed in any way on its way to the lips. The rest of the speech sounds are regarded as consonants. However, many consonants are very different from each other, which is why they need to be further categorized to subcategories.

As previously stated, vowels can be characterized by their distinct formants, which can be seen as peaks in the frequency spectrum of the vocal tract system. It has been noted, that the greatest impact on the created vowel comes from the first two or three formants, while the rest correspond to personal differences between the speakers [26]. The vocal tract can be regarded as if it would be stationary when looking at the characteristics of vowels, because the vocal tract moves quite slowly and it is almost stationary when looking at short periods of time. An example of the frequency spectrum of a couple of different vowels is shown in figure 3.2.

The consonants can be roughly categorized as follows. A more complete categorization of the pulmonic consonants can be found in table 3.1.

Nasals Nasals are consonants where the whole airflow is directed to the nasal cavity, and no air escapes through the lips. Nasals have characteristic anti-resonant frequencies in addition to resonant frequencies, due to the closed oral

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Retroflex	Alveolo-palatal	Palatal	Velar	Uvular	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ		ɲ	ŋ	ɴ		
Stop	p b			t d		ʈ ɖ		c ɟ	k g	q ɢ		ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ɕ ʑ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ								
Approximant		ʋ		ɹ		ɻ		j	ɰ			
Lateral approximant				l		ɭ		ʎ	ʟ			
Flap or tap				ɾ		ɽ						
Trill	ʙ			r		ɽr				ʀ	ʜ ʢ	

Table 3.1: Pulmonic consonants according to the International Phonetic Association, 2015.

cavity trapping certain frequencies. Examples of nasals include [m] and [n].

Stops Stops (or plosives) are created by stopping the airflow for a short time at some point in the vocal tract and then releasing a short burst of air. Stops include [t], [d], [k], [g], [p] and [b].

Fricatives Fricatives are created by narrowing the vocal tract at some point so much that a turbulent noise is created. Fricatives include [s], [z], [f] and [v].

Approximants Approximants are created by partially obstructing the airflow, but only to an extent that it does not create much turbulent noise. Approximants include [l], [ɹ] and [j].

Flaps or taps Tap or flaps are created by hitting one part of the vocal tract against some other. Examples include [ɾ] (as in the (American) English word "latter") and [ɺ] (as in the Japanese word "ラーメン", "ramen").

Trills Trills are produced when an active articulator vibrates against a passive articulator. Trills include [r] and [ʀ].

Other consonants include for example clicks, but are very uncommon in western languages.

3.2 The source-filter theory

The source-filter theory is a simplified model for the human speech production mechanism. As speech production consists of different parts, the source-filter theory also (as the name states) consists of a model for both the sound source (the glottal flow) and the filter (the vocal tract). We will give a more detailed

review of the model in sections 3.3 and 3.4, but we will first demonstrate the general idea of the model.

As previously stated, the source-filter model consists of different parts. The source signal is created at the glottis by the vibrating vocal folds, after which the created airflow travels through the vocal tract getting its spectral structure modified, and finally radiates to the surrounding air from the lips and nostrils. In the z-domain, the produced speech signal can be written as

$$S(z) = G(z)V(z)L(z) \tag{3.1}$$

where S is the speech signal, G the glottal flow, V the vocal tract filter and L the lip radiation [11]. The lip radiation can however be expressed as a first order differentiator

$$L(z) = 1 - \alpha z^{-1}, \tag{3.2}$$

where usually $0.96 \leq \alpha < 1$ [12]. For this reason equation (3.1) is often expressed as

$$S(z) = \hat{G}(z)V(z), \tag{3.3}$$

where $\hat{G}(z) = G(z)L(z)$ is the glottal flow derivative or the glottal pressure.

We will now take a closer look at how the glottal flow and the vocal tract are modelled. We will discuss glottal flow models in section 3.3 and the vocal tract filter in section 3.4.

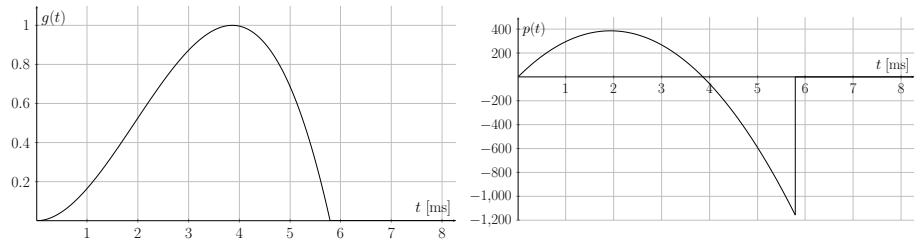
3.3 Glottal flow models

We will present two different models for the glottal flow: the *Rosenberg-Klatt model* (*RK-model*) and the *Liljencrants-Fant model* (*LF-model*). Of these two the RK-model is more simple, using only a single parameter in addition to the fundamental frequency of the sound, whereas the LF-model uses four parameters. The RK-model is still useful even though it gives less freedom in creating the sound signal than the LF-model, because it is easier to use in more complicated situations, such as in Markov chain Monte Carlo simulations.

The two different models described next will give closed form expressions for the glottal flow and the glottal flow derivative. The models will be given for a continuous time case, but the discrete-time signals may be acquired by sampling the expressions, as described in section 2.

3.3.1 Rosenberg-Klatt model

The RK-model was first proposed by Rosenberg in 1971 [28] and later used in creating the synthesisers KLSYN [17] by Klatt, D, and KLSYN88 [18] by



(a) (b)

Figure 3.3: The (a) airflow and (b) pressure generated by the RK-model.

Klatt, D and Klatt, L.

The airflow g for the RK-model is defined as

$$g(t) = \begin{cases} at^2 + bt^3 & \text{if } 0 \leq t \leq QT \\ 0 & \text{if } QT < t \leq T, \end{cases} \tag{3.4}$$

where t denotes the time, $T = 1/f_0$ is the period of the pitch, f_0 is the fundamental frequency and $Q \in [0, 1]$ is the Klatt-parameter. The air pressure p can be calculated as the derivative of the airflow function, namely

$$p(t) = g'(t) = \begin{cases} 2at + 3bt^2 & \text{if } 0 \leq t \leq QT \\ 0 & \text{if } QT < t \leq T. \end{cases} \tag{3.5}$$

An example of the airflow and pressure generated by the RK-model can be seen in figure 3.3. The parameters used in the example for the fundamental frequency and the Klatt-parameter are $f_0 = 120$ Hz and $Q = 0.7$, respectively.

Let us now calculate the values for the variables a and b. We can assume that $Q > 0$, because else we would have $g(t) = 0$ for all $0 \leq t \leq T$. We can also assume that $f_0 > 0$ and $T > 0$. Now define $T_0 := QT$ as the closing instant for the vocal folds. Now we only need to inspect the situation for $t \in [0, T_0]$. We will need two assumptions. Firstly, we know from the definition of the airflow that

$$g(0) = g(T_0) = 0. \tag{3.6}$$

Secondly, we can choose to normalize the airflow, as

$$\max_{t \in [0, T_0]} g(t) = 1. \tag{3.7}$$

From equation (3.6) we get that

$$0 = g(T_0) = a{T_0}^2 + b{T_0}^3 = {T_0}^2(a + bT_0). \tag{3.8}$$

Because we assumed $Q > 0$ and $T > 0$ we know that $T_0 > 0$, and thus we get from equation (3.6) that

$$b = -\frac{a}{T_0}. \tag{3.9}$$

Inserting this into equation (3.4) for the glottal airflow we get for $0 < t < T_0$ that

$$g(t) = a\left(t^2 - \frac{t^3}{T_0}\right). \tag{3.10}$$

Now our assumption in equation (3.7) states that the maximum value of the airflow should be 1. As the function for the airflow is a polynomial (when $0 < t < T_0$) we know that it is continuous. Because we know that all continuous functions on closed intervals have a maximum value, we know that such a maximum exists for the airflow function. We can find the point for the maximum, call it t_0, from the derivative of the function as

$$g'(t_0) = 0. \tag{3.11}$$

Now we get

$$0 = g'(t_0) = a\left(2t_0 - \frac{3{t_0}^2}{T_0}\right) = at_0\left(2 - \frac{3t_0}{T_0}\right), \tag{3.12}$$

which means that either $t_0 = 0$ or $t_0 = 2T_0/3$. However, we assumed that $g(0) = 0$, which means that we get

$$g(t_0) = g\left(\frac{2T_0}{3}\right) = 1. \tag{3.13}$$

We can now calculate

$$1 = a\left(t_0^2 - \frac{t_0^3}{T_0}\right) = a\left(\left(\frac{2}{3}T_0\right)^2 - \frac{1}{T_0}\left(\frac{2}{3}T_0\right)^3\right) = a \cdot \frac{4}{27}T_0^2, \tag{3.14}$$

and finally

$$a = \frac{27}{4T_0^2}, \quad b = -\frac{27}{4T_0^3}. \tag{3.15}$$

We have now acquired numerical values for the variables a and b, which makes the RK-model easy to implement.

3.3.2 Liljencrants-Fant model

The LF-model is a more complicated model for the glottal excitation than the RK-model. The LF-model uses four different parameters to define the shape of the signal, and thus gives more freedom in shaping the signal as wanted and allows to create more realistic excitation signals. The LF-model takes the

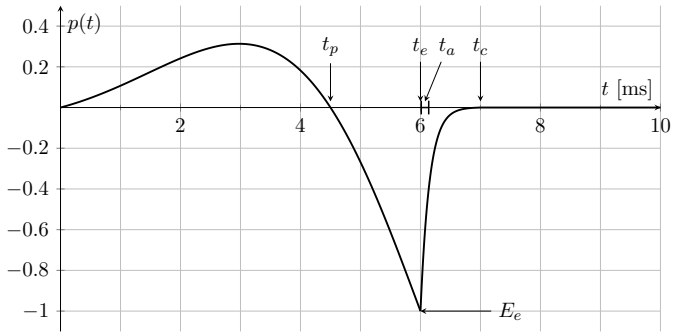


Figure 3.4: The pressure created by the LF-model.

parameters f_0 (the fundamental frequency), t_p, t_e, t_a and t_c, and optionally a scaling parameter E_e.

As presented in [33] by Touda, the air pressure according to the LF-model is defined as

$$p(t) = \begin{cases} E_0 e^{\alpha t} \sin(\omega t), & 0 \leq t \leq t_e \\ -\frac{E_e}{\varepsilon t_a} \left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)} \right], & t_e \leq t \leq t_c \\ 0, & t_c \leq t \leq T, \end{cases} \tag{3.16}$$

and the airflow as the integral

$$g(t) = \begin{cases} \int_0^t p(t')\,\mathrm{d}t', & 0 \leq t \leq t_c \\ 0, & t_c \leq t \leq T, \end{cases} \tag{3.17}$$

where

$$E_0 = -\frac{E_e}{e^{\alpha t_e} \sin(\omega t_e)}$$

and

$$\omega = \frac{\pi}{t_p}.$$

The variables ε and α can be retrieved by solving the system

$$\begin{cases} \eta(\varepsilon) &= 0 \\ \xi(\alpha) &= 0, \end{cases} \tag{3.18}$$

where

$$\eta(\varepsilon) = \varepsilon t_a + e^{-\varepsilon(t_c-t_e)} - 1$$

and

$$\xi(\alpha) = \int_0^T p(t)\,\mathrm{d}t. \tag{3.19}$$

The system (3.18) can then be solved with Newton's iteration as

$$\varepsilon_{n+1} = \varepsilon_n - \frac{\eta(\varepsilon_n)}{\eta'(\varepsilon_n)} \quad \text{and} \quad \alpha_{n+1} = \alpha_n - \frac{\xi(\alpha_n)}{\xi'(\alpha_n)}. \tag{3.20}$$

In order to easily generate similar signals of different fundamental frequencies we will be using the following notations. Let us denote the period of the signal as $T = 1/f_0$. Then, instead of defining the parameters t_p, t_e, t_a and t_c directly, we define them as relative parameters to the period T. We can choose the parameters q_p, q_e, q_a and q_c, where

$$0 < q_p, q_e, q_a, q_c < 1$$

and

$$t_p = Tq_p, \quad t_e = Tq_e, \quad t_a = Tq_a, \quad t_c = Tq_c.$$

Now we do not need to adjust the parameters when changing the fundamental frequency of the signal to match the length of the fundamental period of the signal.

An example of the pressure generated with the LF-model can be found in figure 3.4. The fundamental frequency used in the example is $f_0 = 100$ Hz and the parameters used are $q_p = 0.45$, $q_e = 0.6$, $q_a = 0.015$ and $q_c = 0.7$.

Now we still want to give explicit formulations for the equations of the LF-model. These can then be used as a reference when making an implementation of the model. Let us first calculate the integral in equation (3.17). Firstly, when $0 \leq t \leq t_e$ we get

$$\begin{aligned}\int_0^t p(t')\,\mathrm{d}t' &= \int_0^t E_0 e^{\alpha t'} \sin(\omega t')\,\mathrm{d}t' \\ &= \frac{E_0}{\alpha^2 + \omega^2}\left[e^{\alpha t}\left(-\omega\cos(\omega t) + \alpha\sin(\omega t)\right) + \omega\right].\end{aligned}$$

Secondly, when $t_e < t \leq t_c$ we have the following situation. Let us denote the pressure function when $0 \leq t \leq t_e$ as p_1 and the pressure function when $t_e < t \leq t_c$ as p_2. We know from equations (3.18) and (3.19) that

$$\begin{aligned}0 &= \int_0^T p(t')\,\mathrm{d}t' = \int_0^{t_c} p(t')\,\mathrm{d}t' \\ &= \int_0^{t_e} p_1(t')\,\mathrm{d}t' + \int_{t_e}^{t_c} p_2(t')\,\mathrm{d}t' \\ &= \int_0^{t_e} p_1(t')\,\mathrm{d}t' + \int_{t_e}^{t} p_2(t')\,\mathrm{d}t' + \int_{t}^{t_c} p_2(t')\,\mathrm{d}t' \\ &= \int_0^{t} p(t')\,\mathrm{d}t' + \int_{t}^{t_c} p_2(t')\,\mathrm{d}t',\end{aligned}$$

and thus

$$\begin{aligned}\int_0^t p(t')\,\mathrm{d}t' &= -\int_t^{t_c} p_2(t')\,\mathrm{d}t' \\ &= \int_t^{t_c} \frac{E_e}{\varepsilon t_a}\left[e^{-\varepsilon(t'-t_e)} - e^{-\varepsilon(t_c-t_e)}\right]\mathrm{d}t' \\ &= \frac{E_e}{t_a\varepsilon^2}\left[e^{-\varepsilon(t-t_e)} + (\varepsilon t_e - \varepsilon t_c - 1)e^{-\varepsilon(t_c-t_e)}\right].\end{aligned}$$

We can now express the equation (3.17) of the glottal flow in a more explicit form as

$$g(t) = \begin{cases} \frac{E_0}{\alpha^2+\omega^2}\left[e^{\alpha t}\left(-\omega\cos(\omega t) + \alpha\sin(\omega t)\right) + \omega\right], & 0 \le t \le t_e \\ \frac{E_e}{t_a\varepsilon^2}\left[e^{-\varepsilon(t-t_e)} + (\varepsilon t_e - \varepsilon t_c - 1)e^{-\varepsilon(t_c-t_e)}\right], & t_e \le t \le t_c \\ 0, & t_c \le t \le T, \end{cases} \tag{3.21}$$

We can now also write the function ξ as

$$\begin{aligned}\xi(\alpha) &= \int_0^T p(t')\,\mathrm{d}t' \\ &= \int_0^{t_e} p(t')\,\mathrm{d}t' + \int_{t_e}^{t_c} p(t')\,\mathrm{d}t' \\ &= \frac{E_0}{\alpha^2+\omega^2}\left[e^{\alpha t_e}\left(-\omega\cos(\omega t_e) + \alpha\sin(\omega t_e)\right) + \omega\right] \\ &\quad + \frac{E_e}{t_a\varepsilon^2}\left[1 + (\varepsilon t_e - \varepsilon t_c - 1)e^{-\varepsilon(t_c-t_e)}\right].\end{aligned}$$

In order to calculate the Newton's iterations defined in equation (3.20) we need the derivatives $\eta'(\varepsilon)$ and $\xi'(\alpha)$. We get with a simple calculation that

$$\eta'(\varepsilon) = t_a - (t_c - t_e)e^{-\varepsilon(t_c-t_e)}$$

and

$$\begin{aligned}\xi'(\alpha) = \frac{E_0}{\left(\alpha^2+\omega^2\right)^2}\big[e^{\alpha t_e}\big(&2\alpha\omega\cos(\omega t_e) - \left(\alpha^2-\omega^2\right)\sin(\omega t_e)\big) \\ &-\omega\left(\alpha^2 t_e + 2\alpha + \omega^2 t_e\right)\big].\end{aligned}$$

3.4 The vocal tract filter

As the source signal created in the glottis travels through the vocal tract, the sound is filtered due to the resonant and anti-resonant frequencies of the tract. This process is very similar to an equalizer, where some frequencies are boosted and others are dampened.

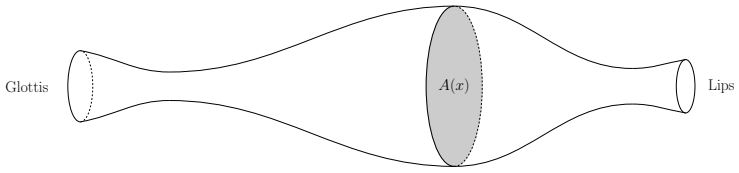


Figure 3.5: An example of the cross-sectional area of the simplified vocal tract.

As we described earlier in section 3.2, we will assume that the vocal tract is unaltered with time. Also, we will only be concentrating on vowel sounds, where the airflow can move unobstructed through the whole vocal tract without any turbulent behaviour.

The most important factor in the creation of different vowel sounds is the cross-sectional area of the vocal tract; we will assume that the vocal tract is straight and rotationally symmetric. This is of course not quite the case in reality, but it has been noted that it is a plausible simplification [26]. The vocal tract can then be characterized by a so called cross-sectional *area function*, $A(x)$, where x denotes the distance travelled along the rotational axis from the glottis toward the lips. An example of this situation can be seen in figure 3.5.

3.4.1 The uniform lossless tube model

One way to look more closely at a vocal tract with a variable cross-sectional area is to use the so called *uniform lossless tube model* [26] (or just *tube model*). The idea of the tube model is to discretize the area function of the vocal tract, and then see how the tract affects the airflow, when assuming there are no losses in the tract. The vocal tract will thus be regarded as a concatenation of cylinders with different radii. The area function of a vocal tract with the length L will then become a simple function $A : [0, L] \to \mathbb{R}_+$, where

$$A(x) = A_k, \quad x \in I_k,$$

where $I_k = [(k-1)\Delta x, k\Delta x[$, $k \in \{1, \ldots, N\}$, and $\Delta x = L/N$. An example of the discretized vocal tract can be seen in figure 3.6.

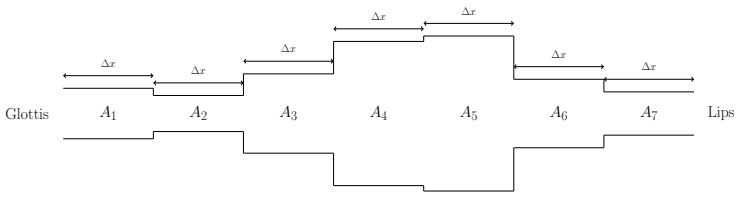


Figure 3.6: An example of the discretization of the vocal tract in the tube model.

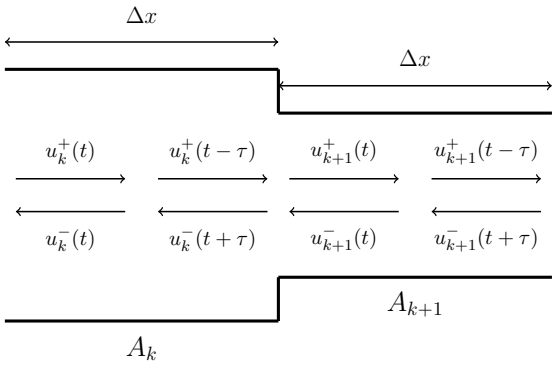


Figure 3.7: The airflow at a tube junction.

The idea with the tube model is to be able to calculate the output signal of an input signal to a tube, given the cross-sectional areas A_k. As it turns out, the lossless tube will work as an LTI system and given the cross-sectional areas we can calculate an explicit transfer function for the system, which will allow us to calculate the output of any input signal easily. Next we will explain briefly how the transfer function is derived.

In the inspection of the tube model we will be using the following notation. The airflow in the k:th tube at the distance x from the start of the k:th tube at the time t will be denoted with $u_k(x,t)$. The pressure will be denoted similarly as $p_k(x,t)$. The speed of sound in the whole system will be assumed to be constant and will be denoted with c. The pressure inside the tube will be denoted with ρ, and will also be assumed to be constant.

For the airflow inside the k:th cylinder we have the identities

$$u_k(x,t) = u_k^+(t - x/c) - u_k^-(t + x/c) \tag{3.22}$$

and

$$p_k(x,t) = \frac{\rho c}{A_k}\left(u_k^+(t - x/c) + u_k^-(t + x/c)\right), \tag{3.23}$$

where u_k^+ and u_k^- are the positive and negative direction volume flows at the beginning of the k:th tube, respectively [26].

Let us look at the situation at the junction between two tubes, as shown in figure 3.7. Because of the physical principle that volume flow and pressure are continuous everywhere in both space and time, we get a constraint for the junction between the k:th and $(k+1)$:st tube, namely

$$\begin{cases} u_k(\Delta x, t) = u_{k+1}(0,t) \\ p_k(\Delta x, t) = p_{k+1}(0,t). \end{cases} \tag{3.24}$$

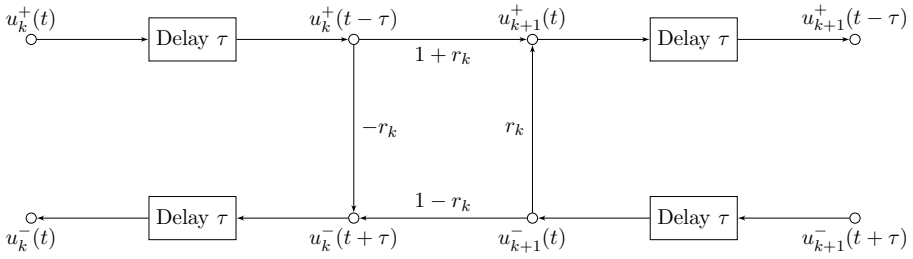


Figure 3.8: A schematic image of the signals at the k:th junction.

Now, by substituting equations (3.22) and (3.23) into the system (3.24), and denoting $\tau := \Delta x/c$, we acquire the system

$$\begin{cases} u_k^+(t-\tau) - u_k^-(t+\tau) = u_{k+1}^+(t) - u_{k+1}^-(t) \\ \frac{A_{k+1}}{A_k}\left(u_k^+(t-\tau) + u_k^-(t+\tau)\right) = u_{k+1}^+(t) + u_{k+1}^-(t). \end{cases} \tag{3.25}$$

Solving the system (3.25) for u_{k+1}^+ and u_k^- we get

$$\begin{cases} u_{k+1}^+(t) = r_k u_{k+1}^-(t) + (1+r_k)u_k^+(t-\tau) \\ u_k^-(t+\tau) = (1-r_k)u_{k+1}^-(t) - r_k u_k^+(t-\tau), \end{cases} \tag{3.26}$$

where

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k}. \tag{3.27}$$

The coefficients r_k, $k = 1, \ldots, N$, are the so called *reflection coefficients* for the k:th junction. Loosely speaking, the reflection coefficient r_k describes how much of the flow u_{k+1}^- is reflected back to the positive direction at the k:th junction, and similarly $-r_k$ tells how much of u_k^+ is reflected back in the negative direction. Because we assumed the tube to be lossless, the rest of the volume flow continues in its original direction, adding up to a no change in the total volume flow but only redirection of the flow.

The process of signal propagation and reflection at the k:th junction is shown schematically in figure 3.8. Each delay corresponds to the time it takes for the sound to traverse through one tube. As we can clearly see, the shortest possible time for a signal to arrive at the lips from the glottis is $N\tau$. If we now assume that there is an impulse starting at the lips at time zero, the impulses from the system will arrive at the lips at times $N\tau + 2n\tau$, $n = 0, 1, \ldots$, i.e. always when the impulses have had time to traverse one cylinder back and forth.

Before we start examining the system in more detail, we will define exactly what kind of system the tube model describes, in order to know what kind of tools we have to work with.

3.28 Theorem. *The uniform lossless tube model describes a causal LTI system.*

Proof. We can easily see that the tube model describes an LTI system; the system is time-invariant because we defined the cross-sectional areas to be constant, and we can see from the system (3.26) that the volume flow behaves linearly at the tube junctions. The system is also clearly causal; as we previously mentioned, the least time in which a signal from the glottis can reach the lips is $N\tau > 0$ and thus the output at time t cannot depend on any input values after the time $t - N\tau$. □

We now know how the volume flow behaves at the tube junctions, but our ultimate goal with the tube model is to describe the output of the last tube as compared to the input of the first tube. Thus we will now start working our way toward a transfer function for the vocal tract. For this we will need boundary conditions for the glottis and lips ends of the tube.

The boundary condition for the glottis can be done in a multitude of ways, but one common way is to assume the glottis to be completely lossless, reflecting all of the incoming signals back toward the lips. This can be described as

$$u_1^+(t) = u_g^+(t) + u_1^-(t), \tag{3.29}$$

where u_g^+ is the source flow from the glottis.

To get expressions for the flow u_l at the lips, we will think of a fictional $(N+1)$:st tube, which is thought to be infinitely long, resulting in no flow back in the negative direction, as

$$u_l^-(t) = 0. \tag{3.30}$$

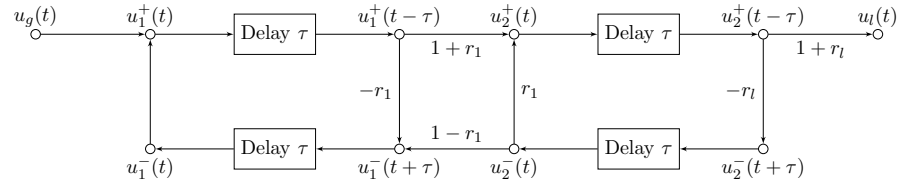
A reflection coefficient for the lips is still needed to account for the reflection at the junction to the fictional $(N+1)$:st tube. We will call this coefficient r_l, resulting in the output

$$u_l^+(t) = (1 + r_l)u_N^+(t). \tag{3.31}$$

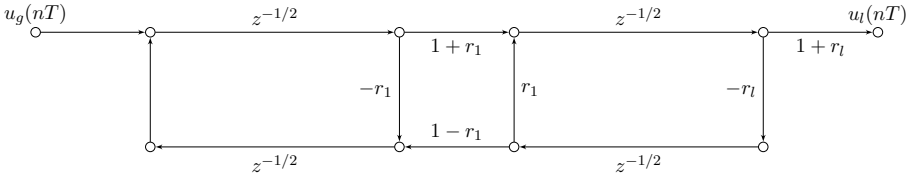
The reflection at the lips is the only source of loss in the system. The value of r_l represents the amount of loss at the lips, and also determines the strength of the resonances of the system. A value of $r_l = 1$ results in an acoustic short circuit with no loss, but usually the value is chosen as $r_l < 1$ to give reasonable bandwidths for the resonances of the system [26].

An example of a complete tube model with two tubes is seen in figure 3.9a.

Let us now consider the discrete-time case where we sample the signals with the sample rate $T = 2\tau$. Now each delay in figure 3.8 will become a delay of half a sample. Recall from section 2.3.2, that a delay of n_0 samples corresponds to a factor z^{-n_0} in the z-plane. Thus, the delays of τ time correspond to a factor



(a) Signal flow graph of the system.



(b) The discrete time system.

Figure 3.9: An example of a two tube model including the boundary conditions for the glottis and the lips.

$z^{-1/2}$. An example of a two tube model with its respective discrete-time system can be seen in figure 3.9.

As previously mentioned, an impulse at the glottis will arrive at the lips at times $t = N\tau + 2k\tau$, for $k = 0, 1, \ldots$. In the discrete time case this will correspond to the samples at the indices $n = NT/2 + kT$, $k = 0, 1, \ldots$, which we can see is an integer as long as N is even. If it is not, the problem can be solved with interpolation between samples. For simplicity, we will assume N to be even in the following discussion.

With the information so far, we can now give an expression for the transfer function of the vocal tract described by the uniform lossless tube model.

3.32 Theorem. *The transfer function of the system described by the uniform lossless tube model can be written as*

$$V(z) = \frac{G}{D(z)}, \tag{3.33}$$

where

$$D(z) = 1 - \sum_{k=1}^{N} \alpha_k z^{-k} \tag{3.34}$$

and G and α_k, $k = 1, \ldots, N$ are constants depending on the reflection coefficients r_k, $k = 1, \ldots, N$ of the tube.

Proof. Let r_k, $k = 1, \ldots, N$, be the reflection coefficients of a uniform lossless tube, with the reflection coefficient at the lips denoted as $r_N = r_l$. As we know by theorem 3.28 that the tube model describes an LTI system, we know that there exists an impulse response $v(n)$ such that the output signal $u_l(n)$ at the

lips can be described by the convolution

$$u_l(n) = (u_g * v)(n), \tag{3.35}$$

where u_g is the input signal at the glottis. Taking the z-transform of equation (3.35), we obtain the expression

$$U_L(z) = U_G(z)V(z), \tag{3.36}$$

where U_L, U_G and V are the z-transforms of the signal at the lips, the signal at the glottis and the impulse response, respectively. Solving equation (3.36) for $V(z)$ we can now note that the transfer function we are looking for can be expressed as

$$V(z) = \frac{U_L(z)}{U_G(z)}. \tag{3.37}$$

Let us now recall the system (3.26), which describes the change of flow at the k:th tube junction. Sampling this with the sampling frequency $T = 2\tau$ we get the discretized version of the system as

$$\begin{cases} u_{k+1}^+(n) = r_k u_{k+1}^-(n) + (1 + r_k)u_k^+\left(n - \frac{1}{2}\right) \\ u_k^-\left(n + \frac{1}{2}\right) = (1 - r_k)u_{k+1}^-(n) - r_k u_k^+\left(n - \frac{1}{2}\right). \end{cases} \tag{3.38}$$

Taking the z-transforms of the two equations in the system (3.38) we obtain the z-domain system

$$\begin{cases} U_{k+1}^+(z) = r_k U_{k+1}^-(z) + (1 + r_k)z^{-\frac{1}{2}}U_k^+(z) \\ z^{\frac{1}{2}}U_k^-(z) = (1 - r_k)U_{k+1}^-(z) - r_k z^{-\frac{1}{2}}U_k^+(z), \end{cases} \tag{3.39}$$

where U_k^+ and U_k^- are the z-transforms of the volume flows in the positive and negative directions, respectively. Solving the first equation in the system (3.39) for U_k^+ we obtain

$$U_k^+(z) = \frac{z^{\frac{1}{2}}}{1 + r_k}U_{k+1}^+(z) - \frac{r_k z^{\frac{1}{2}}}{1 + r_k}U_{k+1}^-(z). \tag{3.40}$$

Substituting equation (3.40) into the second equation in the system (3.39) and solving for U_k^- we get

$$\begin{aligned} U_k^-(z) &= (1 - r_k)z^{-\frac{1}{2}}U_{k+1}^-(z) - r_k z^{-1}\left(\frac{z^{\frac{1}{2}}}{1 + r_k}U_{k+1}^+(z) - \frac{r_k z^{\frac{1}{2}}}{1 + r_k}U_{k+1}^-(z)\right) \\ &= -\frac{r_k z^{-\frac{1}{2}}}{1 + r_k}U_{k+1}^+(z) + \left((1 - r_k)z^{-\frac{1}{2}} + \frac{r_k^2 z^{-\frac{1}{2}}}{1 + r_k}\right)U_{k+1}^-(z) \end{aligned}$$

$$= -\frac{r_k z^{-\frac{1}{2}}}{1+r_k} U_{k+1}^{+}(z) + \left(\frac{z^{-\frac{1}{2}} \left((1+r_k)(1-r_k) + r_k^2 \right)}{1+r_k} \right) U_{k+1}^{-}(z)$$

$$= -\frac{r_k z^{-\frac{1}{2}}}{1+r_k} U_{k+1}^{+}(z) + \frac{z^{-\frac{1}{2}}}{1+r_k} U_{k+1}^{-}(z),$$

i.e.

$$U_k^{-}(z) = -\frac{r_k z^{-\frac{1}{2}}}{1+r_k} U_{k+1}^{+}(z) + \frac{z^{-\frac{1}{2}}}{1+r_k} U_{k+1}^{-}(z). \tag{3.41}$$

Now let us define

$$\mathbf{U_k}(z) = \begin{pmatrix} U_k^{+}(z) \\ U_k^{-}(z) \end{pmatrix}$$

and

$$\mathbf{Q_k}(z) = \frac{z^{\frac{1}{2}}}{1+r_k} \hat{\mathbf{Q}}_\mathbf{k}(z),$$

where

$$\hat{\mathbf{Q}}_\mathbf{k}(z) = \begin{pmatrix} 1 & -r_k \\ -r_k z^{-1} & z^{-1} \end{pmatrix}.$$

Now we can write equations (3.40) and (3.41) in matrix form as

$$\mathbf{U_k}(z) = \mathbf{Q_k}(z)\mathbf{U_{k+1}}(z). \tag{3.42}$$

Let us denote the z-transform of the flow at the lips as $\mathbf{U_L} := \mathbf{U_{N+1}}$. As described in equations (3.30) and (3.31), we can describe the z-transform at the lips as

$$\mathbf{U_L}(z) = \begin{pmatrix} U_L(z) \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} U_L(z). \tag{3.43}$$

Applying equation (3.42) recursively through the whole tube we get

$$\mathbf{U_1}(z) = \prod_{k=1}^{N} \mathbf{Q_k}(z)\mathbf{U_L}(z). \tag{3.44}$$

From equation (3.29) we get the boundary condition for the z-transform of the flow at the glottis as the matrix product

$$U_G(z) = \begin{pmatrix} 1, & -1 \end{pmatrix} \mathbf{U_1}(z). \tag{3.45}$$

Now, combining equations (3.43)–(3.45) we get

$$U_G(z) = \begin{pmatrix} 1, & -1 \end{pmatrix} \mathbf{U_1}(z)$$

$$= \begin{pmatrix} 1, & -1 \end{pmatrix} \prod_{k=1}^{N} \mathbf{Q_k}(z)\mathbf{U_L}(z)$$

$$= \begin{pmatrix} 1, & -1 \end{pmatrix} \prod_{k=1}^{N} \mathbf{Q_k}(z) \begin{pmatrix} 1 \\ 0 \end{pmatrix} U_L(z),$$

and according to equation (3.37) we can write

$$V(z) = \frac{U_L(z)}{U_G(z)} = \frac{1}{\begin{pmatrix} 1, & -1 \end{pmatrix} \prod_{k=1}^{N} \mathbf{Q_k}(z) \begin{pmatrix} 1 \\ 0 \end{pmatrix}},$$

and further

$$V(z) = \frac{\prod_{k=1}^{N}(1+r_k)z^{-N/2}}{\begin{pmatrix} 1, & -1 \end{pmatrix} \prod_{k=1}^{N} \hat{\mathbf{Q}}_\mathbf{k}(z) \begin{pmatrix} 1 \\ 0 \end{pmatrix}}. \tag{3.46}$$

The constant delay $z^{-N/2}$ in the numerator can be dropped, as it does not account to anything else than a shift of $N/2$ samples in the output. This can be compensated by advancing the input with the respective amount of samples. Thus, the transfer function gets the expression

$$V(z) = \frac{G}{D(z)}, \tag{3.47}$$

where G is a constant,

$$G = \prod_{k=1}^{N}(1+r_k). \tag{3.48}$$

Let us now take a closer look at the denominator. By writing out $\hat{\mathbf{Q}}_\mathbf{k}$, we obtain

$$D(z) = \begin{pmatrix} 1, & -1 \end{pmatrix} \prod_{k=1}^{N} \begin{pmatrix} 1 & -r_k \\ -r_k z^{-1} & z^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{3.49}$$

It can easily be seen that this yields a polynomial of the variable z^{-1}, with the degree N. Thus, the denominator can be written as

$$D(z) = 1 - \sum_{k=1}^{N} \alpha_k z^{-k}, \tag{3.50}$$

and we have proven our claim. □

We would still be interested in finding the numerical values of the polynomial coefficients α_k in equation (3.33). We can calculate them easily from the reflection coefficients r_k, as we will soon see. However, we will first need to formulate a lemma.

3.51 Lemma. *Let r_k, $k = 1, \ldots, N$, be the reflection coefficients of a uniform lossless tube. The denominator $D(z)$ of the transfer function in equation* (3.33) *can be calculated recursively by the formula*

$$\begin{cases} D_0(z) = 1 \\ D_m(z) = D_{m-1}(z) + r_m z^{-m} D_{m-1}\left(z^{-1}\right), \quad m = 1, \ldots, N \\ D(z) = D_N(z). \end{cases} \tag{3.52}$$

Proof. Let r_k, $k = 1, \ldots, N$ be the reflection coefficients of a uniform lossless tube. From equation (3.49) we know that

$$D(z) = \begin{pmatrix} 1, & -1 \end{pmatrix} \prod_{k=1}^{N} \hat{\mathbf{Q}}_{\mathbf{k}}(z) \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{3.53}$$

where

$$\hat{\mathbf{Q}}_{\mathbf{k}}(z) = \begin{pmatrix} 1 & -r_k \\ -r_k z^{-1} & z^{-1} \end{pmatrix}.$$

Let us now define the vectors

$$\mathbf{P_0}(z) = \begin{pmatrix} 1, & -1 \end{pmatrix}$$

and

$$\mathbf{P_m}(z) = \mathbf{P_{m-1}}(z)\hat{\mathbf{Q}}_{\mathbf{m}}(z),$$

for $m = 1, \ldots, N$. Now we can write equation (3.53) as

$$D(z) = \mathbf{P_N}(z) \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

because by definition

$$\mathbf{P_N}(z) = \mathbf{P_{N-1}}(z)\hat{\mathbf{Q}}_{\mathbf{N}}(z) = \ldots = \mathbf{P_0}(z) \prod_{k=1}^{N} \hat{\mathbf{Q}}_{\mathbf{k}}(z).$$

Now, let us define

$$D_0(z) = 1$$

and

$$D_m(z) = D_{m-1}(z) + r_m z^{-m} D_{m-1}\left(z^{-1}\right),$$

for $m = 1, \ldots, N$. We want to show that

$$\mathbf{P_m}(z) = \begin{pmatrix} D_m(z), & -z^{-m} D_m\left(z^{-1}\right) \end{pmatrix}, \tag{3.54}$$

for $m = 0, \ldots, N$. The proof is done by induction.

For $m = 0$ we get

$$\mathbf{P_0}(z) = \begin{pmatrix} 1, & -1 \end{pmatrix} = \begin{pmatrix} D_0(z) & -z^{-0} D_0\left(z^{-1}\right) \end{pmatrix}.$$

Now assume that equation (3.54) holds for $m = n - 1$, $n = 1, \ldots, N$, i.e.

$$\mathbf{P_{n-1}}(z) = \begin{pmatrix} D_{n-1}(z), & -z^{-n+1} D_{n-1}\left(z^{-1}\right) \end{pmatrix}.$$

Now we get

$$\begin{aligned}
\mathbf{P_n}(z) &= \mathbf{P_{n-1}}(z)\hat{\mathbf{Q}}_\mathbf{n}(z) \\
&= \begin{pmatrix} D_{n-1}(z), & -z^{-n+1} D_{n-1}\left(z^{-1}\right) \end{pmatrix} \begin{pmatrix} 1 & -r_k \\ -r_k z^{-1} & z^{-1} \end{pmatrix} \\
&= \begin{pmatrix} D_{n-1}(z) + r_n z^{-n} D_{n-1}\left(z^{-1}\right), & -z^{-n}\left(D_{n-1}\left(z^{-1}\right) + r_n z^n D(z)\right) \end{pmatrix} \\
&= \begin{pmatrix} D_n(z), & -z^{-n} D_n\left(z^{-1}\right) \end{pmatrix}.
\end{aligned}$$

Thus, equation (3.54) holds for $m = 0, \ldots, N$. We see now, that

$$D(z) = \mathbf{P_N}(z) \begin{pmatrix} 1 \\ 0 \end{pmatrix} = D_N(z),$$

and thus we have proven our claim. □

3.55 Theorem. *Let r_k, $k = 1, \ldots, N$ be the reflection coefficients of a uniform lossless tube. The coefficients for the transfer function*

$$V(z) = \frac{G}{1 - \sum_{k=1}^{N} \alpha_k z^{-k}}$$

can be calculated as

$$G = \prod_{k=1}^{N} (1 + r_k) \tag{3.56}$$

and by the recursion formula

$$\begin{cases}
\alpha_0^{(0)} = 1 & \\
\alpha_m^{(n)} = 0, & m = 1, \ldots, N \text{ and } n < m \\
\alpha_m^{(n)} = \alpha_m^{(n-1)} + r_n \alpha_{n-m}^{(n-1)}, & m = 1, \ldots, n \text{ and } n = 1, \ldots, N \\
\alpha_m = \alpha_m^{(N)}, & m = 1, \ldots, N.
\end{cases} \tag{3.57}$$

Proof. Let r_k, $k = 1, \ldots, N$ be the reflection coefficients of a uniform lossless tube. We already know from equation (3.48), that equation (3.56) holds.

Now we can prove equation (3.57) by induction, using the recursion formula in equation (3.52). We see that for $m = 1$ we get

$$D_0(z) = 1 = \sum_{k=0}^{0} \alpha_k^{(0)}.$$

Now, assume that the values hold for $m = p - 1$, $p = 1, \ldots, N$, i.e.

$$D_{p-1}(z) = 1 + \sum_{k=1}^{p-1} \alpha_k^{(p-1)} z^{-k}.$$

Now we get for $m = p$, that

$$\begin{aligned}
D_p(z) &= D_{p-1}(z) + r_p z^{-p} D_{p-1}\left(z^{-1}\right) \\
&= 1 + \sum_{k=1}^{p-1} \alpha_k^{(p-1)} z^{-k} + r_p z^{-p} \left(1 + \sum_{k=1}^{p-1} \alpha_k^{(p-1)} z^{k}\right) \\
&= 1 + \sum_{k=1}^{p-1} \alpha_k^{(p-1)} z^{-k} + \sum_{k=1}^{p-1} r_p \alpha_k^{(p-1)} z^{k-p} + r_p z^{-p} \\
&= 1 + \sum_{k=1}^{p-1} \alpha_k^{(p-1)} z^{-k} + \sum_{k=1}^{p-1} r_p \alpha_{p-k}^{(p-1)} z^{-k} + r_p z^{-p} \\
&= 1 + \sum_{k=1}^{p-1} \underbrace{\left(\alpha_k^{(p-1)} + r_p \alpha_{p-k}^{(p-1)}\right)}_{=\alpha_k^{(p)}} z^{-k} + \underbrace{\left(0 + r_p \alpha_0^{(p-1)}\right)}_{=\alpha_p^{(p)}} z^{-p} \\
&= 1 + \sum_{k=1}^{p} \alpha_k^{(p)} z^{-k}.
\end{aligned}$$

Now we know that $\alpha_k^{(m)}$ are the polynomial coefficients of the m:th polynomial $D_m(z)$. As $D(z) = D_N(z)$ we can conclude that $\alpha_k = \alpha_k^{(N)}$ are the coefficients of the denominator $D(z)$. We have now proven our claim. □

3.58 Note. An algorithm to obtain the values in equations (3.56) and (3.57) from the reflection coefficients is shown in algorithm 3.1.

The final thing we want to show regarding the uniform lossless tube model is that all the poles of the transfer function are located strictly inside the unit circle. We are going to need a powerful theorem from complex analysis for this proof, namely the so called *Rouché's theorem* [8]. We will not prove the theorem here, as it would require a much deeper discussion in the theory of complex analysis than possible for the scope of this work.

Algorithm 3.1 Coefficients for the transfer function

```
1: function REFLECTION_COEFF2TRANSFER_COEFF(r)
2:     let G = 1
3:     let α_0^(0) = 1
4:     for n = 1, ..., N do
5:         let G = G × (1 + r_n)
6:         for m = 1 to n do
7:             let α_m^(n) = α_m^(n−1) + r_n α_{n−m}^(n−1)
8:         end for
9:     end for
10:    return G, α^(N)
11: end function
```

3.59 Theorem (Rouché's theorem)**.** *Let γ be a simple closed contour in $\mathbb{C}$ and let f and g be analytic within and on γ. Assume that $f(z) \neq 0$ and $|g(z)| < |f(z)|$ for all z on γ. Then f and $f + g$ have the same number of zeros inside γ.*

Proof. See theorem 5.3.2 in [8]. □

Using Rouché's theorem we have the tools to present a result about the position of the transfer function's poles. We will first present a lemma regarding the values of the reflection coefficients for a uniform lossless tube and then move to present the actual result regarding the poles of the transfer function.

3.60 Lemma. *Let r_k, $k = 1, \ldots, N$ be the reflection coefficients of a uniform lossless tube. We have that $|r_k| < 1$, for all $k = 1, \ldots, N$.*

Proof. Let A_k, $k = 1, \ldots, N$ be the cross-sectional areas of a uniform lossless tube. We can assume that $A_k > 0$ for all $k = 1, \ldots, N$. Now, according to equation (3.27), we get the upper bound for the reflection coefficients as

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} = \frac{A_{k+1}}{A_{k+1} + A_k} < \frac{A_{k+1}}{A_{k+1}} = 1,$$

and the lower bound as

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} > \frac{-A_{k+1} - A_k}{A_{k+1} + A_k} = -1.$$

Thus we see that $|r_k| < 1$, and we have proven our claim. □

3.61 Theorem. *All the poles of the transfer function of a system defined by the uniform lossless tube model are strictly inside the unit circle.*

Proof. Let r_k, $k = 1, \ldots, N$ be the reflection coefficients of a uniform lossless tube. The poles of the transfer function are the zeros of the denominator $D(z)$ defined in equation (3.52). As D_k, $k = 1, \ldots, N$, are polynomials of the degree k

we know that they have exactly k roots (as each root is calculated as many times as its multiplicity). This means that we have to show that for each $k = 1, \ldots, N$ the function D_k has all its k roots inside the unit circle, i.e. for each root z_l, $l = 1, \ldots, k$, of D_k we have that $|z_l| < 1$. This proves that $D_N = D$ has all its N roots inside the unit circle, which is exactly what we want to prove. We will prove this by induction.

The case $k = 0$ is easy. We know from the definition of D_k in equation (3.52) that $D_0(z) = 1$ for all z. Thus it has no roots, meaning that all its roots are inside the unit circle.

Let us now assume that for $k = n$, with $0 \leq n < N$, we have that D_n has all its n roots inside the unit circle. From the proof of theorem 3.55 we know that D_k can be written as

$$D_k(z) = \sum_{m=0}^{k} \alpha_m^{(k)} z^{-m},$$

for all $k = 0, \ldots, N$. We also know from the recursive definition of D_{n+1} in equation (3.52) that

$$D_{n+1}(z) = D_n(z) + r_{n+1} z^{-(n+1)} D_n\left(z^{-1}\right).$$

Multiplying this with z^{n+1} ($z \neq 0$) and reorganizing the terms yields the equation

$$z^{n+1} D_{n+1}(z) - z^{n+1} D_n(z) = r_{n+1} D_n\left(z^{-1}\right). \tag{3.62}$$

Let now $|z| = 1$. Because $\alpha_m^{(k)} \in \mathbb{R}$ for all m and k, we get that

$$\begin{aligned}
\overline{D_k\left(z^{-1}\right)} &= \overline{\sum_{m=0}^{k} \alpha_m^{(k)} z^m} = \sum_{m=0}^{k} \overline{\alpha_m^{(k)} z^m} = \sum_{m=0}^{k} \overline{\alpha_m^{(k)}}\, \overline{z}^m \\
&= \sum_{m=0}^{k} \alpha_m^{(k)} \left(z^{-1}\right)^m = \sum_{m=0}^{k} \alpha_m^{(k)} z^{-m} = D_k(z),
\end{aligned}$$

and thus

$$\left|D_k\left(z^{-1}\right)\right| = \left|D_k(z)\right|. \tag{3.63}$$

Using lemma 3.60, equation (3.63) and the fact that $\left|z^{n+1}\right| = 1$ we can rewrite equation (3.62) for $|z| = 1$ as

$$\begin{aligned}
\left|z^{n+1} D_{n+1}(z) - z^{n+1} D_n(z)\right| &= \left|r_{n+1} D_n\left(z^{-1}\right)\right| = |r_{n+1}| \left|D_n\left(z^{-1}\right)\right| \\
&= |r_{n+1}| \left|D_n(z)\right| = |r_{n+1}| \left|z^{n+1}\right| \left|D_n(z)\right| \\
&= |r_{n+1}| \left|z^{n+1} D_n(z)\right| \\
&< \left|z^{n+1} D_n(z)\right|,
\end{aligned}$$

yielding

$$\left|z^{n+1}D_{n+1}(z) - z^{n+1}D_n(z)\right| < \left|z^{n+1}D_n(z)\right|. \tag{3.64}$$

Both sides of equation (3.64) are now analytic[2] and thus by theorem 3.59 we can conclude that $z^{n+1}D_{n+1}(z)$ and $z^{n+1}D_n(z)$ have the same number of zeros inside the unit circle.

Writing out $z^{n+1}D_n(z)$ as

$$z^{n+1}D_n(z) = z^{n+1}\sum_{m=0}^{n}\alpha_m^{(n)}z^{-m} = \sum_{m=0}^{n}\alpha_m^{(n)}z^{n+1-m},$$

we see that $z^{n+1}D_n(z)$ has a total of $n+1$ zeros. Setting $z^{n+1}D_n(z) = 0$ we see that either $z^{n+1} = 0$, yielding $z = 0$, or $D_n(z) = 0$, which we assumed to be n zeros inside the unit circle. Thus $z^{n+1}D_n(z)$, and therefore also $z^{n+1}D_{n+1}(z)$, has $n+1$ zeros inside the unit circle.

Writing out $z^{n+1}D_{n+1}(z)$ as

$$z^{n+1}D_{n+1}(z) = z^{n+1}\sum_{m=0}^{n+1}\alpha_m^{(n+1)}z^{-m} = \sum_{m=0}^{n+1}\alpha_m^{(n+1)}z^{n+1-m},$$

we see that it has a total of $n+1$ zeros. These zeros are already proven to be inside the unit circle. Noting that $D_{n+1}(z)$ already has $n+1$ zeros in total, we can conclude that the zeros of $z^{n+1}D_{n+1}(z)$ are actually the zeros of $D_{n+1}(z)$.

We have now shown by induction that D_N, and thus also D, has all its N zeros inside the unit circle, which is what we set out to prove. □

[2] This is strictly speaking not completely true, as equation (3.62) is not valid for $z = 0$, as we mentioned earlier. However, the functions on both sides can be continued to analytic functions over $z = 0$, as multiplying $D_n(z)$ and $D_{n+1}(z)$ makes the singularities at $z = 0$ *removable singularities*. Due to this reasoning it is reasonable to call the functions analytic.

4 The inverse problem – glottal inverse filtering

Glottal inverse filtering (GIF) is a technique for recovering the glottal excitation signal from a recorded speech signal. The basic idea of the technique is to cancel the vocal tract filter and lip radiation from the speech signal, revealing the original source signal. However, this is a challenging task, as it turns out that GIF is an ill-posed inverse problem.

As GIF has been studied for decades, since around the 1950s, many different methods for solving the problem have been proposed. Several methods are also robust and work within reasonable error margins for low enough fundamental frequencies [3]. One of the challenges with GIF is still to construct a robust method for estimating the glottal flow from recorded speech signals with a high fundamental frequency, such as children's or women's voices.

In this section we will present a GIF method based on Markov chain Monte Carlo proposed by Auvinen *et al.* in [5]. In section 4.2 we will also briefly present the so called IAIF (iterative adaptive inverse filtering) method, proposed by Alku in [2]. The IAIF method has been shown to give very good estimates of the glottal flow, but is prone to error with high-pitched voices. The IAIF is an important method to the MCMC based GIF method, as the initial guess of the solution of GIF in the MCMC method is based on the IAIF result.

We will now move on to describe the problem in more detail, and then describe different methods for solving the problem.

4.1 Glottal inverse filtering

As previously stated, GIF is a technique where we attempt to recover the glottal flow from a recorded speech signal. As we know from equation (3.1) in section 3, the speech signal can be view in the z-domain as $S(z) = G(z)V(z)L(z)$, where S, G, V and L represent the z-transforms of the recorded speech signal, the glottal flow, the vocal tract and lip radiation, respectively. This means that the glottal flow can be recovered as $G(z) = S(z)/(V(z)L(z))$. Recalling from equation (3.2), the lip radiation can be modelled with a fixed first-order differentiator, resulting in the fact that only the vocal tract needs to be estimated.

As we described in section 3.4, we will only need to estimate the parameters G and α_k in equations (3.33) and (3.34) in order to recover a good estimation for the vocal tract filter. This is however not that straightforward, due to a number of factors.

GIF is a typical inverse problem, which we can be write as

$$m = A(g) + \varepsilon, \tag{4.1}$$

where $m \in \mathbb{R}^n$ represents the measured data, $A : \mathbb{R}^k \to \mathbb{R}^n$ is an operator, $g \in \mathbb{R}^k$ is the original signal we want to recover and $\varepsilon \in R^n$ is random measurement noise. For simplicity the noise is assumed to be Gaussian additive noise, with $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]^T$ satisfying $\varepsilon_m \sim \mathcal{N}\left(0, \sigma^2\right)$ for all $m = 1, \ldots, n$, with a standard deviation $\sigma > 0$.

The task related to equation (4.1) is "given the measurements m, reconstruct g". In practice, the problem with solving these kinds of inverse problems is often the ill-posedness of the problem. Typically this means that two different solutions g and g' will result in almost the same measurements, $A(g) \approx A(g')$. With the addition of the measurement noise, this makes recovering g from the measurement m a difficult task.

To be able to give robust solutions to the inverse problem in equation (4.1) we need to use some a-priori knowledge of the problem at hand. As an example, in our case we have defined a detailed mathematical model of the speech production mechanism in section 3 that we can use to our advantage.

4.2 The IAIF method

We will now briefly present the *iterative adaptive inverse filtering* (IAIF) method for solving the glottal inverse filtering problem. The idea in IAIF is to estimate both the glottal excitation and the vocal tract using *linear predictive coding* (LPC) [25]. We will go into further detail about the algorithm in section 4.2.2, but we will first need to take a closer look at LPC analysis.

4.2.1 Linear predictive coding and analysis

In LPC analysis we would like to find such *linear prediction coefficients* α_k, $k = 1, \ldots, p$, such that given the signal x the signal

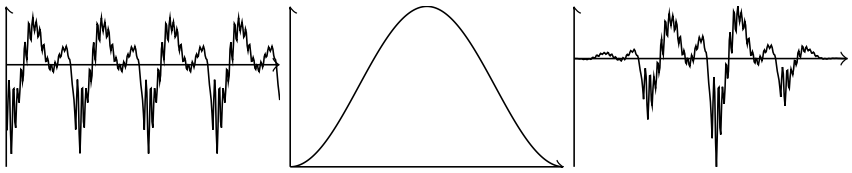
$$y(n) = \sum_{k=1}^{p} \alpha_k x(n-k) \tag{4.2}$$

is as good an approximation of x as possible. More specifically, we want to minimize the sum of squares of the error

$$d(n) = x(n) - y(n) = x(n) - \sum_{k=1}^{p} \alpha_k x(n-k).$$

If we assume $\alpha_0 = -1$ we can write the error simply as

$$d(n) = -\sum_{k=0}^{p} \alpha_k x(n-k). \tag{4.3}$$



(a) (b) (c)

Figure 4.1: Example of the weighting of a signal with the Hann window. The graphs show (a) the original signal, (b) the Hann window and (c) the weighted signal.

Here p is the order of the LPC analysis.

Assume now that we have a frame of K samples as our given signal x, i.e.

$$\begin{cases} x(n) \in \mathbb{R}, & 1 \leq n \leq K \\ x(n) = 0, & \text{otherwise.} \end{cases}$$

Writing equation (4.3) in matrix form we acquire

$$X\alpha = y', \tag{4.4}$$

where

$$X = \begin{pmatrix} x(1) & 0 & \cdots & 0 \\ x(2) & x(1) & \ddots & \vdots \\ \vdots & x(2) & \ddots & 0 \\ \vdots & \vdots & \ddots & x(1) \\ \vdots & \vdots & \vdots & x(2) \\ x(K) & \vdots & \vdots & \vdots \\ 0 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & x(K) \end{pmatrix}, \quad \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix} \quad \text{and} \quad y' = \begin{pmatrix} x(2) \\ x(3) \\ \vdots \\ x(K) \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Now equation (4.4) can easily be solved in the least squares sense for the coefficients α_k.

In practice the LPC analysis is often done to a somewhat modified signal $\hat{x}$ instead of the original signal x. The reason for this is that the errors at the boundaries near $x(1)$ and $x(K)$ tend to become large due to the fact that the frame ends and the remaining samples are padded with zeros. A common modification is to use the weighted signal using the Hann (or Hanning) window,

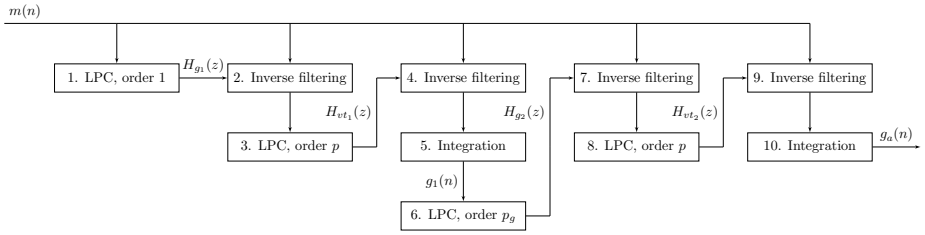


Figure 4.2: A diagram showing the structure of the IAIF algorithm.

where the samples near the center are given high weights and samples near the boundaries close or equal to zero weight. An example of the weighting of a signal using the Hann window is shown in figure 4.1.

4.2.2 The IAIF algorithm

The IAIF method relies on the assumption that the combined effect of the glottal excitation and lip radiation can be estimated from the speech measurement frame using low-order LPC analysis. The algorithm proceeds as follows.

Let $m(n)$ be the measurement speech frame.[3] The algorithm begins by first estimating the effect of glottal excitation and lip radiation from $m(n)$ with an LPC analysis of order 1. We receive the first estimate for the effect of the glottal excitation and the lip radiation as an all pole filter $H_{g_1}(z)$. The measurement signal $m(n)$ is then filtered with the inverse of the filter $H_{g_1}(z)$ to receive a new signal, with the effect of $H_{g_1}(z)$ removed. An LPC analysis of order p (usually 20) is then performed on this signal, giving the first estimate of the vocal tract filter, $H_{vt_1}(z)$. The measurement signal $m(n)$ is then filtered with the inverse of this filter and integrated, leaving us with the first estimate for the glottal flow, $g_1(n)$. This process is then repeated to receive the estimates $H_{g_2}(z)$ and $H_{vt_2}(z)$, but this time the estimate for the effect of the glottal excitation and the vocal tract is done with an LPC analysis of order p_g (usually 8) instead of 1. Finally the estimate $g_a(n)$ for the glottal flow is received with filtering $m(n)$ with the inverse of $H_{g_2}(z)$ and integrating the resulting signal.

A diagram of the specifics of the algorithm is shown in figure 4.2.

4.3 The MCMC-GIF method

As we already noted, we described in section 3 that the we can model the speech production mechanism in the z-domain as $S(z) = G(z)V(z)L(z)$, where S, G, V and L represent the z-transforms of the recorded speech signal, the

[3] The measurements can be high-pass filtered before using the algorithm, but we will not consider that part of the algorithm itself.

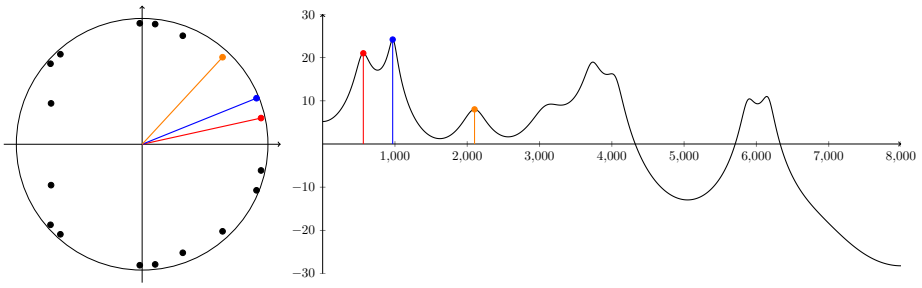


Figure 4.3: An example of the connection between the formant frequencies and poles of the transfer function.

glottal flow, the vocal tract and lip radiation, respectively. In theorem 3.32 we showed that the vocal tract can be modelled as an all-pole filter

$$V(z) = \frac{G}{1 - \sum_{k=1}^{N} \alpha_k z^{-k}}, \tag{4.5}$$

allowing us to estimate the vocal tract filter simply by estimating the parameters for the filter.

As we noted in theorem 2.50, all LTI systems with a rational transfer function can be characterized by their poles and zeros. In the case of the vocal tract filter, which is an all-poles filter having no zeros, we can characterize it by only its poles. Having a denominator written as a real polynomial of degree N, the transfer function V has exactly N poles, all of which are either purely real or appear in conjugate pairs. Let us assume for simplicity that we have no real poles, but all poles are complex having a conjugate pair.[4]

Let us now call the poles of the transfer function with a positive argument $p_1, \ldots, p_{N/2}$ and their conjugates $\tilde{p}_1, \ldots, \tilde{p}_{N/2}$. Recalling theorem 2.52 from section 2, and assuming that the arguments of the poles p_k are somewhat evenly divided in the range $[0, \pi]$, we can see that each pole corresponds to a formant in the frequency spectrum.[5] The closer the pole is to the unit circle, the higher the formant. An example of the connection between poles and formants is shown in figure 4.3.

[4] This is actually a feasible assumption. Recall from section 2, that the argument of a pole of a transfer function corresponds to the frequency, with an argument close to 0 corresponding to a low frequency and an argument close to $\pm\pi$ corresponding to a high frequency. This means that a real pole corresponds to a frequency of either 0 Hz or the Nyquist frequency $f_s/2$, where f_s is the sampling frequency, and thus no more than two of these poles are ever required.

[5] If we take a closer look at equation (2.53) for an all pole filter, we can see that if $p_k = r_k e^{i\omega}$ for some k with $r_k = 1 - \varepsilon$ for some small $\varepsilon > 0$ and no other poles have an argument close to ω, then the factor $|p_k - e^{i\omega}| = |r_k e^{i\omega} - e^{i\omega}| = |\varepsilon|$ becomes dominant in the denominator of $H(\omega)$, thus creating a peak at the frequency ω in the frequency spectrum.

As we previously noted in section 3.1.3, different vowels can be characterized by their formant structure, with the first few formants having the greatest impact on the created vowel. The general idea with the MCMC-GIF method is thus to start out with a reasonable guess for a vocal tract filter (obtained with the IAIF method) and then moving around the first few poles of the transfer function trying to improve the estimate. An estimate for the poles of the transfer function is obtained using MCMC, which we will explain in more detail in section 4.3.1. A more precise description of how the MCMC-GIF algorithm works will be given in section 4.3.3.

4.3.1 Bayesian inversion and Markov chain Monte Carlo

Bayesian inversion is a technique for estimating the solution for ill-posed inverse problems as in equation (4.1). The idea of Bayesian inversion is to use *a priori* information that we have of the problem in order to solve it. The information we have can be used to construct probability distributions, giving a clue to how probable different kinds of solutions are.

Let us start by introducing some basic concepts of probability theory, and then work our way toward a more precise formulation of Bayesian inversion and the usage of MCMC in order to solve the problem. In this discussion we will restrict ourselves to continuously derivable probability distributions.

Let X be an $\mathbb{R}^n$-valued random variable with the probability distribution p_X. The probability density function (PDF) $p_X : \mathbb{R}^n \to \mathbb{R}$ describes the relative likelihood of the random variable taking a given value. For a PDF it holds that

$$p_X(x) > 0$$

for all $x \in \mathbb{R}^n$ and

$$\int_{\mathbb{R}^n} p_X(x)\,\mathrm{d}x = 1.$$

The probability for a sample x' of the random variable X to be in a subset $A \subset \mathbb{R}^n$ is

$$\mathrm{P}(x' \in A) = \int_A p_X(x)\,\mathrm{d}x.$$

Let now X and Y be $\mathbb{R}^n$ and $\mathbb{R}^k$ valued random variables with the PDFs p_X and p_Y respectively. The joint probability density $p_{X,Y} : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$ is now also a PDF, defined as

$$p_{X,Y}(x, y) = p_{Y|X}(y|x)p_X(x) = p_{X|Y}(x|y)p_Y(y), \tag{4.6}$$

where $p_{Y|X}$ and $p_{X|Y}$ are the conditional probabilities of Y given $X = x$ and X

given $Y = y$ respectively. Now equation (4.6) yields the so called Bayes' formula

$$p_{X|Y}(x|y) = \frac{p_X(x)p_{Y|X}(y|x)}{p_Y(y)}, \tag{4.7}$$

where $p_Y(y) > 0$ is required.

Let us assume m and g from equation (4.1) to be random vector with continuous probability densities. Now according to Bayes' formula we acquire the so called *posterior distribution*

$$p_{G|M}(g|m) = \frac{p_G(g)p_{M|G}(m|g)}{p_M(m)}. \tag{4.8}$$

The PDF $p_{M|G}$ in equation (4.8) is the so called *likelihood distribution*. It is related to the noise in the measurements, giving high probabilities of the measurements m which are close to $A(x)$ and a low to those that are further away. In the case of Gaussian noise, as we assumed in our case, the likelihood distribution takes the form

$$p_{M|G}(m|g) = C \exp\left(-\frac{1}{2\sigma^2} \|A(g) - m\|_2^2\right), \tag{4.9}$$

where C is a normalization constant.

Now the likelihood distribution behaves as described. In our case we have the measurements m constant, meaning that we can inspect the probability for a given solution candidate g to yield the measurements m; if $A(g)$ (the perfect noiseless measurements assuming g) is close to m, then the probability is high for g being the actual solution, and if $A(g)$ differs a lot from m the probability for g being the solution is low.

The role of the *prior distribution* $p_G(g)$ in equation (4.8) is to include all the *a priori* information that we have about the solution. It should assign high probabilities to likely solutions and low probabilities to unlikely solutions, in light of the *a priori* knowledge.

The actual solution to the inverse problem in equation (4.1) according to Bayesian inversion is the posterior distribution defined in equation (4.8). This is however hard to visualize, and a point estimate of some sort is much rather presented. In the case of MCMC-GIF we will be using the *conditional mean estimate* defined as

$$g_{\mathrm{CM}} = \int_{\mathbb{R}^n} g\, p_{G|M}(g|m)\, \mathrm{d}g. \tag{4.10}$$

Another widely used point estimate is the *maximum a posteriori* estimate, defined as

$$g_{\mathrm{MAP}} = \arg\max_{g \in \mathbb{R}^n} \left(p_{G|M}(g|m)\right). \tag{4.11}$$

The conditional mean can be approximated using Markov chain Monte Carlo. We want to estimate the integral in equation (4.10) numerically as

$$\int_{\mathbb{R}^n} g\, p_{G|M}(g|m)\, \mathrm{d}g \approx \frac{1}{N} \sum_{m=1}^{N} g^{(m)}, \tag{4.12}$$

where the values $g^{(1)}, g^{(2)}, \ldots, g^{(N)}$ are distributed according to the posterior density $p_{G|M}$.

In Markov chain Monte Carlo the value of the element $g^{(m)}$ only depends on the previous element $g^{(m-1)}$ (thus the term Markov chain, which corresponds to this particular property). In practice we cannot guarantee the initial guess $g^{(1)}$ to be close to the conditional mean, which means that the beginning of the chain might not be distributed properly. To avoid biasing due to this fact, we use a so called *burn-in period*, where the first M elements of the chain are discarded, and we assume that from the element $g^{(M+1)}$ on the chain is properly distributed. The estimate of the conditional mean in equation (4.12) then becomes

$$\int_{\mathbb{R}^n} g\, p_{G|M}(g|m)\, \mathrm{d}g \approx \frac{1}{N-M} \sum_{m=M+1}^{N} g^{(m)}. \tag{4.13}$$

We are still left with the task of acquiring the values $g^{(k)}$. This can be done with the Metropolis-Hastings algorithm, which will be explained next.

4.3.2 The Metropolis-Hastings algorithm

The idea behind the Metropolis-Hastings algorithm [14, 21] is to find samples that correspond to a probability distribution p using a Markov process. This Markov process should be generated so that it reaches asymptotically a stationary distribution π, such that $\pi = p$. We will now explain and derive the Metropolis-Hastings algorithm. The proofs regarding results of Markov processes are beyond the scope of this work, but we will give references to sources of the proofs.

A Markov process is a process with the Markov property, namely such a process where the next state $x^{(m)}$ only depends on the current state $x^{(m-1)}$. The process is therefore memoryless, meaning that it doesn't matter how we got to the state $x^{(m-1)}$, the probabilities for the process to move to any state remain the same. A Markov process can be uniquely defined by defining the transition probabilities $p(x \to x')$ for all the states x and x'.

Let us now examine a Markov process with the transition probabilities $p(x \to x')$ between the states x and x'. We want to know when the Markov process reaches asymptotically a unique stationary distribution. This is true when we can show that such a stationary distribution exists and that it is

unique. We get the following conditions (of which the first one, theorem 4.14, is a sufficient but not necessary condition, but will suffice in our case) [27]:

4.14 Theorem. *There exists a stationary distribution π to which the Markov process converges asymptotically, if the probability for being in a state x and moving to a state x' is equal to being at the state x' and moving to the state x, for all states x and x'. This can be expressed mathematically as*

$$\pi(x)p(x \to x') = \pi(x')p(x' \to x). \tag{4.15}$$

4.16 Theorem. *The stationary distribution π is unique, if the Markov process is (1) aperiodic and (2) positive recurrent.*

Now we would like to choose the transition probabilities in our Markov process in a way that the conditions in theorems 4.14 and 4.16 hold. We will do this by separating the transition probabilities into two steps: proposal of the next state and acceptance of the state. If we denote the *proposal probability* with q and the *acceptance probability* with a we can write

$$p(x \to x') = q(x \to x')a(x \to x'). \tag{4.17}$$

Combining equations (4.15) and (4.17) we get the condition

$$\frac{a(x \to x')}{a(x' \to x)} = \frac{p(x')q(x' \to x)}{p(x)q(x \to x')}. \tag{4.18}$$

The common choice for the acceptance probability is to choose

$$a(x \to x') = \min\left\{1, \frac{p(x')q(x' \to x)}{p(x)q(x \to x')}\right\}. \tag{4.19}$$

We still want to show that this choice satisfies the condition in equation (4.18). We see that either $a(x \to x') = 1$ or $a(x' \to x) = 1$ for all states x and x'. Assume now that $a(x \to x') = 1$. Then we know that

$$a(x' \to x) = \frac{p(x)q(x \to x')}{p(x')q(x' \to x)}$$

and substituting this into the LHS of equation (4.18) we get

$$\frac{a(x \to x')}{a(x' \to x)} = \frac{1}{p(x)q(x \to x')/(p(x')q(x' \to x))} = \frac{p(x')q(x' \to x)}{p(x)q(x \to x')}.$$

Thus we see the choice in equation (4.19) satisfies the condition in equation (4.18).

The proposal probability distribution q is still to be chosen. This remains as a free parameter to be chosen for each particular problem. In many cases q is

Algorithm 4.1 The Metropolis-Hastings algorithm

```
1: procedure METROPOLIS_HASTINGS
2:     choose x^(1) at random
3:     let m = 1
4:     for m = 1, ..., N do
5:         choose x' according to q(x^(m) → x')
6:         let r = p(x')/p(x^(m)) * q(x' → x^(m))/q(x^(m) → x')
7:         if rand(0, 1) < r then
8:             let x^(m+1) = x'
9:         else
10:            let x^(m+1) = x^(m)
11:        end if
12:    end for
13: end procedure
```

chosen to be symmetric (a multivariate normal distribution with the mean at the current state and a suitable variance is a popular choice when dealing with problems in $\mathbb{R}^n$), such that $q(x \to x') = q(x' \to x)$ for all x and x'. In this case the acceptance probability is reduced to

$$a(x \to x') = \min\left\{1, \frac{p(x')}{p(x)}\right\}. \tag{4.20}$$

As we see, the acceptance of the new state becomes a really easy task; if the new state has a higher posterior probability than the last state it is immediately accepted, if it has a lower probability the state is accepted with respect to how much less probable it is compared with the previous state.

The Metropolis-Hastings algorithm is described with step by step instructions in algorithm 4.1.

4.3.3 The MCMC-GIF algorithm

We will now describe how Markov chain Monte Carlo and the Metropolis-Hastings algorithm are used to solve the GIF problem. The process is done for a small frame of recorded speech, which allows us to assume that the vocal tract is time-invariant for the whole speech frame. According to our model of speech production explained in section 3 the problem can be written as a convolution

$$m = p * v + \varepsilon, \tag{4.21}$$

where $m = m(n)$ are the measurements, $p = p(n)$ is the glottal pressure signal, $v = v(n)$ is the impulse response of the vocal tract and ε is random noise.

The initial step in MCMC-GIF is to find an initial guess for the MCMC. As already mentioned earlier, the solution from the IAIF method is used as such.

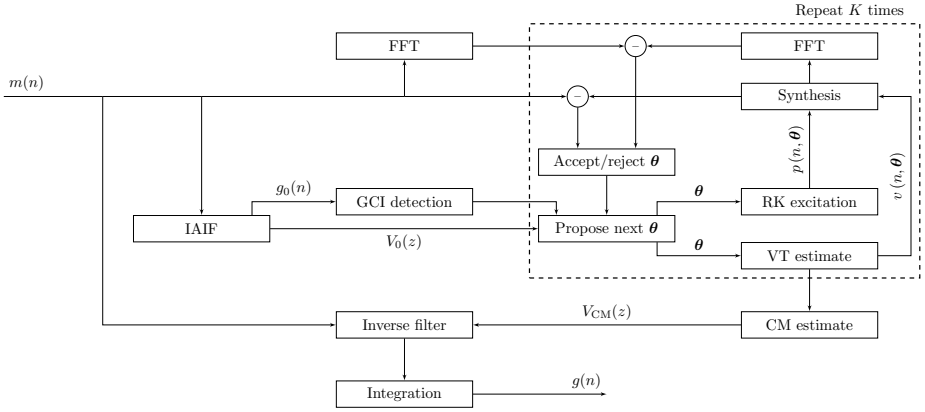


Figure 4.4: The MCMC-GIF algorithm shown as a flow diagram.

From IAIF we acquire an estimate for the glottal flow, $g_0(n)$, as well as the parameters for the all-pole filter describing the vocal tract, $V_0(z)$. Once this is done we can shift the frame into position.[6] This is done by searching for the *glottal closure instant* (GCI) using a simple peak finding algorithm. The Klatt-parameter is initialized usually in the range $q_0 \in [0.4, 0.6]$. We then have a parametrized IAIF estimate for our problem, with q_0 being the Klatt-parameter for the glottal flow and $v = v(z_1, \ldots, z_N)$ being the vocal tract filter parametrized by the poles z_k, $k = 1, \ldots, N$. Let us assume that the poles are ordered by the absolute value of their argument, $|\arg(z_k)| < |\arg(z_l)|$ for $k < l$. We will disregard the purely real poles from the IAIF result, as they do not correspond to any actual formant frequencies.[7]

Next we want to compute the radii and arguments of the poles, which makes it easier to put constraints on the numerical values of the numbers when simulating the MCMC. As we only regard the purely complex poles, we get the vector $\boldsymbol{\theta} = (r_1, \varphi_1, \ldots, r_M, \varphi_M, q)$, where q is the Klatt-parameter, r_k and φ_k correspond to the radii and arguments of the k:th conjugate pair of poles,

$$z_{2k-1} = r_k \exp(i\varphi_k)$$

and

$$z_{2k} = r_k \exp(-i\varphi_k),$$

[6] This is an important part in MCMC-GIF, as it is important that the glottal source signal created with the RK-model has the correct phase that matches the original signal.

[7] In the original work [5] by Auvinen *et al.* the proposed method is to completely discard the real poles. It is still up to discussion if this is the best thing to do, as the real poles still shape the frequency response, even though they do not correspond to any actual formant frequencies. Another way to go about this is to keep the real poles in the calculations keeping them unaltered during the whole process. This method is used in the results presented in section 5.

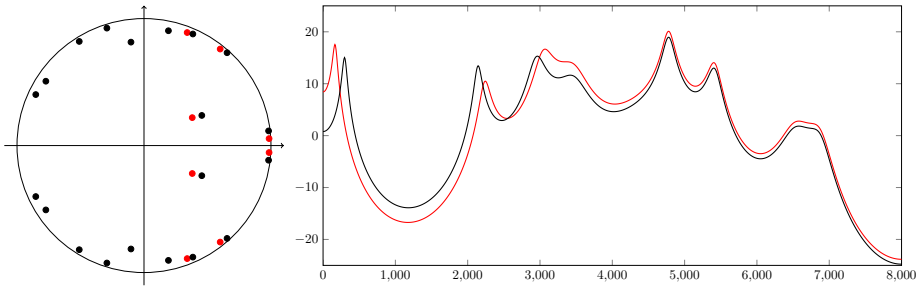


Figure 4.5: An example of the effect on the frequency response of the movement of the first four poles.

$k = 1, \ldots, M$ and $M = N/2$. The forward solution can now be written as

$$m_{\boldsymbol{\theta}}(n) = p(n, \boldsymbol{\theta}) * v(n, \boldsymbol{\theta}),$$

where $m_{\boldsymbol{\theta}}$ is the forward solution, $p(n, \boldsymbol{\theta}) = p(n, q)$ is the glottal pressure described by the RK-model and $v(n, \boldsymbol{\theta})$ is the impulse response of the all-pole filter $V_{\boldsymbol{\theta}}(z)$ defined by the poles $z_1, \ldots, z_N$.

Our aim is now to find the values for $\boldsymbol{\theta}$ that correspond best to our measurements m. The idea is to improve the transfer function by moving the poles and the Klatt-parameter to fit the measurements. Because we know that a high dimension of the free variables complicates the numerical stability of the problem, and we know that the first few formants are most important in defining the vocal tract for vowels, we will choose the four first poles of the transfer function to be free variables while we fix the rest. This means that the problem will take place in the space $\mathbb{R}^9$ (the Klatt-parameter and four poles). An example of the effect of the movement of the first four poles can be seen in figure 4.5. As it can be seen, the poles correspond approximately to the formant peaks, as we have mentioned earlier.

As explained in section 4.3.1, the Bayesian inversion solution to the described problem is

$$\pi(\boldsymbol{\theta}|m) = \frac{\pi(\boldsymbol{\theta})\,\pi(m|\boldsymbol{\theta})}{\pi(m)}, \tag{4.22}$$

where $\pi(m|\boldsymbol{\theta})$ is the likelihood distribution, $\pi(\boldsymbol{\theta})$ is the prior distribution and $\pi(m)$ is a normalizing constant.

The prior distribution should map likely values of $\boldsymbol{\theta}$ to high probabilities and unlikely values to low probabilities. This should not take into account the measurements, but only correspond to the *a priori* knowledge we have of the values. As we know from theorem 3.61, all the poles of our transfer function should be strictly inside the unit circle. The poles' radii can also be restricted to certain ranges based on the IAIF result; it is highly unlikely that the values

would change drastically from the initial estimate. The poles' arguments also follow the ordering $\varphi_1 < \varphi_2 < \ldots < \varphi_M$.[8] The Klatt-parameter q can also be restricted to a certain range. All of the above qualities are represented in the prior distribution.

As we assumed that the measurement noise is Gaussian white noise with some standard deviation $\sigma > 0$, as explained in section 4.3.1, the likelihood model takes the form

$$\pi(m|\boldsymbol{\theta}) = \exp\left(-\frac{1}{2\sigma^2}\left\|p(\boldsymbol{\theta}) * v(\boldsymbol{\theta}) - m\right\|_2^2\right).$$

However, in the case of GIF it is important to measure the error in both the time domain and the frequency domain. This is due to the fact that characteristics of signal are often better visible in the frequency domain representation. As proposed by Auvinen *et al.* in [5], this results in the likelihood distribution

$$\pi(m|\boldsymbol{\theta}) = \exp\left(-c_t\Delta_t(m,\boldsymbol{\theta}) - c_f\Delta_f(m,\boldsymbol{\theta})\right), \tag{4.23}$$

where

$$\Delta_t(m,\boldsymbol{\theta}) = \left\|p(\boldsymbol{\theta}) * v(\boldsymbol{\theta}) - m\right\|_2^2$$

is the time domain squared norm,

$$\Delta_f(m,\boldsymbol{\theta}) = \left\|\left|\mathrm{FFT}(p(\boldsymbol{\theta}) * v(\boldsymbol{\theta})\right| - \left|\mathrm{FFT}(m)\right|\right\|_2^2$$

is the frequency domain squared norm and $c_t, c_f > 0$ are parameters to be evaluated experimentally.

With the given specifications we can generate a sequence $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(K)}$ using the Metropolis-Hastings algorithm described in section 4.3.2 for a predefined number of samples K.[9] After computing the sequence we can acquire a point estimate calculating the conditional mean of the sequence as in equation (4.13), with a burn-in period of K_0 samples. The conditional mean is then

$$\boldsymbol{\theta}_{\mathrm{CM}} \approx \frac{1}{K - K_0}\sum_{k=K_0+1}^{K}\boldsymbol{\theta}^{(k)}. \tag{4.24}$$

When we have acquired the conditional mean estimate, we can finally calculate our new glottal flow estimate. This is done by inverse filtering the measurements

[8] This would otherwise not be needed as a constraint, but because all the poles correspond to the transfer function equally, a change of order of the poles' angles could possibly result in disrupting the estimate; when taken the conditional mean of the results two poles would be mixed up resulting in unexpected behaviour.

[9] In both this work and the work [5] by Auvinen *et al.* the method used to obtain the sequence is actually a modern variation of the Metropolis-Hastings algorithm known as DRAM [13]. The implementation of DRAM can be found in the Matlab package [19].

with the filter $V_{\mathrm{CM}}(z)$ described by the poles in the conditional mean estimate and then integrating the result.

A diagram of the MCMC-GIF algorithm is shown in figure 4.4.

5 Numerical results

In this section we will present some numerical results obtained by using the MCMC-GIF method for solving the glottal inverse filtering problem. All the results presented are obtained from synthetic data, created for the purpose of example. The results are not to be regarded as any kind of proof for the performance of MCMC-GIF in a general case, but merely as examples to demonstrate how the algorithm performs in typical cases. Synthetic data is used to ensure objective analysis of the results; we have a reference signal for the correct answer. The results of MCMC-GIF will compared with the results of the IAIF method in each case.[10]

We will first present some earlier results of the performance of the MCMC-GIF algorithm in section 5.1. In section 5.2 we will present new results of the performance of MCMC-GIF in some example cases.

5.1 Earlier results

A comprehensive set of results using the MCMC-GIF method was presented by Auvinen *et al.* in [5]. In the article, the authors showed using a large set of synthetic vowels, that MCMC-GIF performs better than other existing GIF methods in almost all cases.[11] It was noted that the errors for all methods tend to grow larger with higher fundamental frequencies, but this also depends on which vowels are used. It was also noted that although the MCMC-GIF method performed better in most cases, the most prominent improvements compared to the IAIF method were received with low fundamental frequencies.

Although the results by Auvinen *et al.* were promising, the authors wanted to stress that the results are still preliminary and that the method can most probably be further improved. In particular, the authors suggested that further study is needed to understand how different prior distributions affect the results.

5.2 Numerical examples

Let us now present some results of the performance of MCMC-GIF acquired by the author. As already mentioned before, the results presented are not to be regarded as any kind of proof of the algorithm's performance, but merely as examples of how the algorithm works.

[10] As MCMC-GIF uses the IAIF result as an initial guess, the results actually tell how much MCMC-GIF was able to improve (or worsen) the results obtained by the IAIF method.

[11] The authors of the article want to note that synthetic vowels were used in order to obtain a reference signal for the results, which is not possible with natural recorded vowels. The synthetic vowels were, however, created using physical modelling techniques of the vocal folds and the vocal tract, in order to ensure that the results are not biased by the data being created with the same source-filter theory than what the GIF methods are based on.

	/a/ 100 Hz	/a/ 200 Hz	/i/ 200 Hz	/i/ 250 Hz
H1-H2				
IAIF	0.04 dB	0.44 dB	6.80 dB	6.39 dB
MCMC-GIF	0.02 dB	0.01 dB	0.16 dB	0.56 dB
NAQ				
IAIF	3.2 %	0.6 %	41.3 %	32.8 %
MCMC-GIF	0.9 %	1.8 %	2.9 %	2.8 %

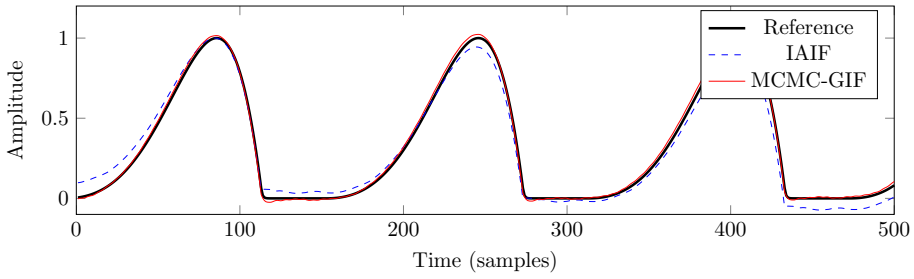
Table 5.1: Results of the numerical experiments.

5.2.1 Experiment setup

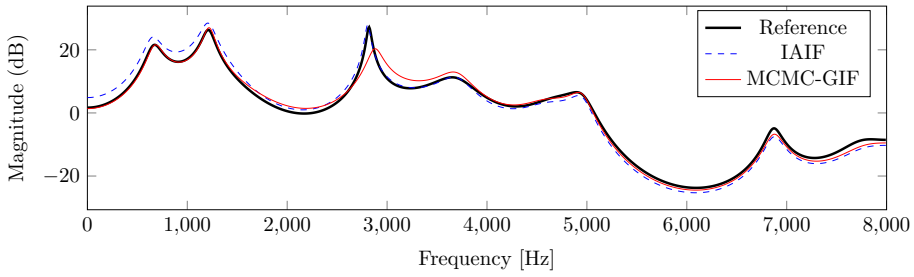
All the synthetic test data were created in the following way. The glottal excitation signal was created using the LF-model, described in section 3.3.2. The parameters for the LF-model were chosen to represent viable excitation signals and the fundamental frequency f_0 was chosen separately for each case. The glottal excitation signal was then filtered with a vocal tract filter previously recovered with some existing GIF method to create the synthetic vowel data. The sampling frequency of the speech data was chosen to be 16 kHz in each experiment, with the speech frame of the length 25 ms (400 samples), except for the case with the fundamental frequency of $f_0 = 100$ Hz, where the frame was 31.25 ms (500 samples) long.

It is important to note, that the data was created with a different model than what MCMC-GIF uses for solving the problem; the data is created with the LF-model and MCMC-GIF uses the RK-model. By doing this we avert the so called *inverse crime*, where the inspected method gets an unfair advantage of using the same model as the data it is tested on. In other words, if the data is created with the same model than what is used in solving the problem, the data is "too easy" for the method to solve. By choosing the data wisely we get comparable results.

The IAIF algorithm was run on the data frame to receive a 20:th order LP estimate for the vocal tract filter. A total of nine parameters were then estimated with the MCMC-GIF algorithm; the Klatt-parameter q and the radii r_k and angle shift $\Delta\varphi_k$ of the four first poles. The initial value of the Klatt-parameter was chosen as $q_0 = 0.5$, with a uniform prior distribution in the range $[0.2, 0.9]$. The radii were initialized to the values of the IAIF estimate, R_k, with a uniform prior distribution in the range $[R_k - 0.1, 0.99]$. The angle shift was initialized to 0 for all angles (resulting in the IAIF estimate angle for all poles) with a uniform prior distribution in the range $[-\pi/16, \pi/16]$. The parameters c_t and c_f for the likelihood distribution were chosen to give equal weights to both the time domain and frequency domain errors. The MCMC-GIF algorithm was run with a total of $K = 100\,000$ simulations, using a burn-in period of $K_0 = 50\,000$.



(a)



(b)

Figure 5.1: The (a) glottal flow estimates and (b) vocal tract filter estimates for the vowel /a/ at $f_0 = 100$ Hz.

5.2.2 Results

The numerical results of four different GIF cases are shown in table 5.1. A solution for each case was calculated using both the IAIF method and the MCMC-GIF method. Two different error estimates were calculated for all the results, one in both the time domain and the frequency domain.

The first error estimate, H1-H2, is a frequency domain error, measuring the magnitude difference between the first and second harmonics of the glottal flow [32]. The second error estimate, NAQ (Normalized Amplitude Quotient), is a time domain error, measuring the ratio between the amplitude of the glottal flow and the negative peak amplitude of the glottal pressure, normalized in respect with the length of the fundamental period [4]. Both methods are widely used in determining the vocal quality.

The results of the glottal inverse filtering using both IAIF and MCMC-GIF are shown in figures 5.1–5.4. The glottal flow comparison between IAIF and MCMC-GIF in the time domain for each case is shown in figures 5.1a and 5.2–5.4. The frequency response of the recovered vocal tract filter for one of the cases is also shown in figure 5.1b.

The computation time required for the MCMC-GIF varied from 6 to 7 hours using a single core on a laptop computer.

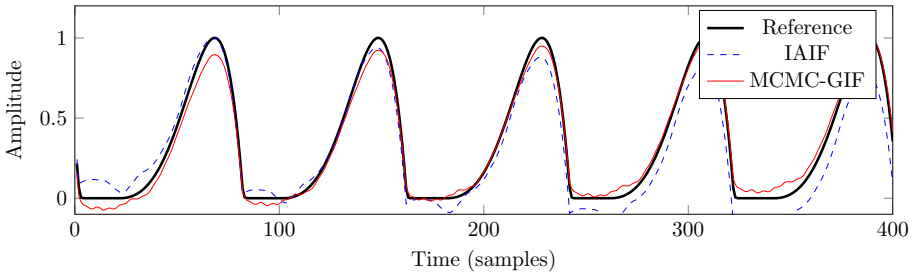


Figure 5.2: The glottal flow estimates for the vowel /a/ at $f_0 = 200$ Hz.

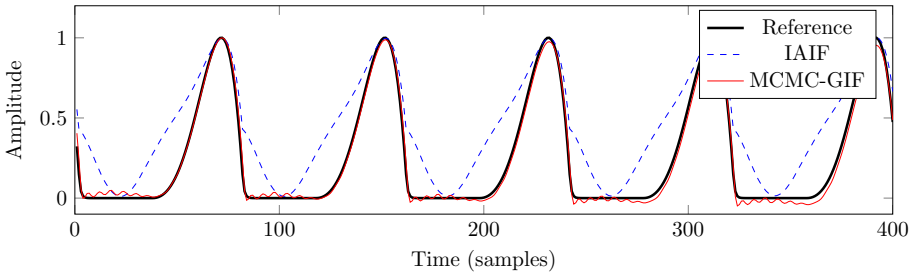


Figure 5.3: The glottal flow estimates for the vowel /i/ at $f_0 = 200$ Hz.

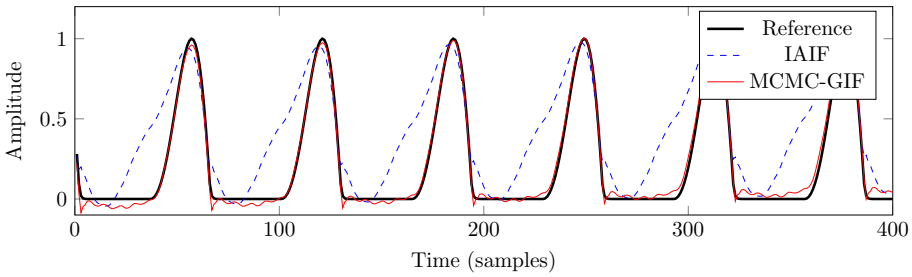


Figure 5.4: The glottal flow estimates for the vowel /i/ at $f_0 = 250$ Hz.

5.2.3 Discussion

The presented examples show that the MCMC-GIF method improves most of the results obtained by the IAIF method. In three out of the four examples better results were acquired with the MCMC-GIF method in both H1-H2 and NAQ sense, while one (200 Hz /a/) was better in H1-H2 sense but worse when measured with NAQ. It can be seen that the errors are clearly higher for both examined methods with higher frequencies, as already noted in section 5.1.

The chosen vowel also clearly affects the errors; the results using IAIF were good in both cases using the vowel /a/, but bad with the vowel /i/. This might be the result of the first formant of the vowel /i/ being around 240 Hz, whereas the first formant for the vowel /a/ is around 850 Hz. This can greatly affect the IAIF method, as the fundamental frequency might get "mixed up" with the first formant frequency.

From figure 5.1b we can see that the frequency response (or the poles of the transfer function) acquired by the MCMC-GIF method are not necessarily always better for the whole spectrum. We can clearly see that the frequency response is almost perfect for the frequencies 0–1500 Hz, but that the third formant at about 3 kHz is clearly misplaced. However, this does not seem to affect the final result of the glottal flow estimate, as we can see in figure 5.1a and from the results in table 5.1; the glottal flow estimate is almost perfect. The reason to why the misplaced third formant does not affect the result that much might be that the impact of the formants on the vowel sound is greatest at the first two formants, and as frequency of the third formant for this particular vowel is quite high it does not affect the result as much as the two first formants.

Even though the results obtained with MCMC-GIF are better than those acquired with the IAIF method, especially for higher frequencies and certain vowels, the computation time required is much longer. As the IAIF method only needs seconds to compute the result, it takes several hours for the MCMC-GIF to reach the results, making the algorithm quite impractical to use. However, the run time of MCMC-GIF could be shortened with a couple of changes.

The implementation of MCMC used in this work is a single threaded algorithm, which means that the computational time could be lowered drastically by using parallel algorithms for the MCMC calculations. Also, the number of simulations used in this work is quite high, being 100 000 compared to the 40 000 used in the original work by Auvinen *et al.*. The number of simulations was increased because it was noted that the results improved further when exceeding the suggested 40 000 simulations. The chosen number of 100 000 simulations in this work might though be a bit overkill, and the same level of results might easily be achieved with a smaller number of simulations, decreasing the computation time

of the algorithm even further. The computation time of the algorithm could thus possible be decreased to well within an hour with the right implementation and choice of the number of simulations used.

References

[1] Tuncay Aktosun. Inverse scattering for vowel articulation with frequency-domain data. *Inverse Problems*, 21(3):899, 2005.

[2] Paavo Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech communication*, 11(2):109–118, 1992.

[3] Paavo Alku. Glottal inverse filtering analysis of human voice production—a review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana*, 36(5):623–650, 2011.

[4] Paavo Alku, Tom Bäckström, and Erkki Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *The Journal of the Acoustical Society of America*, 112(2):701–710, 2002.

[5] Harri Auvinen, Tuomo Raitio, Manu Airaksinen, Samuli Siltanen, Brad H Story, and Paavo Alku. Automatic glottal inverse filtering with the Markov chain Monte Carlo method. *Computer Speech & Language*, 2013.

[6] Baris Bozkurt, Boris Doval, Christophe D'Alessandro, and Thierry Dutoit. Zeros of z-transform representation with application to source-filter separation in speech. *Signal Processing Letters, IEEE*, 12(4):344–347, 2005.

[7] Baris Bozkurt, Laurent Couvreur, and Thierry Dutoit. Chirp group delay analysis of speech signals. *Speech Communication*, 49(3):159–176, 2007.

[8] John Warren Dettman. *Applied Complex Variables*. Courier Corporation, 1965.

[9] Thomas Drugman, Baris Bozkurt, and Thierry Dutoit. Complex cepstrum-based decomposition of speech for glottal source estimation. In *Interspeech*, pages 116–119, 2009.

[10] Thomas Drugman, Baris Bozkurt, and Thierry Dutoit. A comparative study of glottal source estimation techniques. *Computer Speech & Language*, 26 (1):20–34, 2012.

[11] G. Fant. *Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations*. Description and Analysis of Contemporary Standard Russian. De Gruyter, 1971.

[12] J.L. Flanagan. *Speech analysis, synthesis and perception*. Kommunikation und Kybernetik in Einzeldarstellungen. Springer-Verlag, 1972.

[13] Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. Dram: Efficient adaptive mcmc. *Statistics and Computing*, 16(4):339–354, 2006. ISSN 0960-3174.

[14] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[15] Takashi Kako and Kentarou Touda. Numerical approximation of dirichlet-to-neumann mapping and its application to voice generation problem. In *Domain Decomposition Methods in Science and Engineering*, pages 51–65. Springer, 2005.

[16] Takashi Kako and Kentarou Touda. Numerical method for voice generation problem based on finite element method. *Journal of Computational Acoustics*, 14(01):45–56, 2006.

[17] Dennis H Klatt. Software for a cascade/parallel formant synthesizer. *the Journal of the Acoustical Society of America*, 67(3):971–995, 1980.

[18] Dennis H Klatt and Laura C Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *the Journal of the Acoustical Society of America*, 87(2):820–857, 1990.

[19] Marko Laine. MCMC Toolbox for Matlab. `http://helios.fmi.fi/~lainema/mcmc/`, 2013. Accessed: 2015-08-17.

[20] Julien Mauprivez, Edson Cataldo, and Rubens Sampaio. Artificial neural networks applied to the estimation of random variables associated to a two-mass model for the vocal folds. *Inverse Problems in Science and Engineering*, 20(2):209–225, 2012.

[21] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[22] Richard L Miller. Nature of the vocal cord wave. *The Journal of the Acoustical Society of America*, 31(6):667–677, 1959.

[23] Alan V Oppenheim, Ronald W Schafer, John R Buck, *et al. Discrete-time signal processing*, volume 2. Prentice-hall Englewood Cliffs.

[24] D. O'Shaughnessy. *Speech communications: human and machine*. Institute of Electrical and Electronics Engineers, 2000.

[25] Douglas O'Shaughnessy. Linear predictive coding. *Potentials, IEEE*, 7(1): 29–32, 1988.

[26] L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall signal processing series. Prentice-Hall, 1978.

[27] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer New York, 2005.

[28] Aaron E Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *The Journal of the Acoustical Society of America*, 49(2B):583–590, 1971.

[29] Ian Stewart and David Tall. *Complex analysis*. Cambridge University Press, 1983.

[30] Nicolas Sturmel, Christophe d'Alessandro, and Boris Doval. A comparative evaluation of the zeros of z transform representation for voice source estimation. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[31] Ingo R. Titze. *Principles of Voice Production*. Prentice Hall, 1994.

[32] Ingo R Titze and Johan Sundberg. Vocal intensity in speakers and singers. *the Journal of the Acoustical Society of America*, 91(5):2936–2946, 1992.

[33] Kentarou Touda. *Study on numerical method for voice generation problem*. PhD thesis, The University of Electro-Communications, 2007.

[34] Jacqueline Walker and Peter Murphy. Advanced methods for glottal wave extraction. In *Nonlinear analyses and algorithms for speech processing*, pages 139–149. Springer, 2005.

Appendices

A Proofs of the DTFT and z-transform properties

In this appendix we will present the proofs for the properties of the DTFT and the z-transform, presented in tables 2.1 and 2.2 in section 2.

We will present all the proofs only for the z-transform case, because the z-transform is just a generalization of the DTFT. In other words, the DTFT can be acquired from the z-transform by only looking at values on the complex unit circle, i.e. $X_{\mathscr{F}}(\omega) = X_{\mathscr{Z}}\left(e^{i\omega}\right)$, where $X_{\mathscr{F}}$ is the DTFT and $X_{\mathscr{Z}}$ is the z-transform.

A.1 Theorem. *The z-transform is linear, i.e.*

$$\mathscr{Z}\{c_1x(n) + c_2y(n)\} = aX(z) + bY(z),$$

where $x(n)$ and $y(n)$ are sequences, $X(z)$ and $Y(z)$ their respective z-transforms and $c_1, c_2 \in \mathbb{R}$ constants.

Proof. With a direct calculation we get

$$\begin{aligned}\mathscr{Z}\{c_1x(n) + c_2y(n)\} &= \sum_{k=-\infty}^{\infty} (c_1x(k) + c_2y(k))z^{-k} \\ &= c_1 \sum_{k=-\infty}^{\infty} x(k)z^{-k} + c_2 \sum_{k=-\infty}^{\infty} y(k)z^{-k} \\ &= c_1X(z) + c_2Y(z).\end{aligned}$$

□

A.2 Theorem. *Let $x(n)$ be a sequence, $X(z)$ its z-transform, and $n_0 \in \mathbb{Z}$. Now the z-transform of the delayed sequence is*

$$\mathscr{Z}\{x(n - n_0)\} = z^{-n_0}X(z).$$

Proof. Using the substitution $m = k - n_0$ we get with a direct calculation that

$$\begin{aligned}\mathscr{Z}\{x(n - n_0)\} &= \sum_{k=-\infty}^{\infty} x(k - n_0)z^{-k} = \sum_{m=-\infty}^{\infty} x(m)z^{-(m+n_0)} \\ &= z^{-n_0} \sum_{m=-\infty}^{\infty} x(m)z^{-m} = z^{-n_0}X(z).\end{aligned}$$

□

A.3 Theorem. *Let $x(n)$ be a sequence, $X(z)$ its z-transform, and $z_0 \in \mathbb{C}$. Now the z-transform of the new sequence $y(n) = z_0^n x(n)$ is*

$$\mathcal{Z}\left\{z_0^n x(n)\right\} = X(z/z_0).$$

Note. The theorem is equivalent to the modulation property of the DTFT. If you restrict $|z_0| = 1$, and write $z_0 = e^{i\omega_0}$, where $\omega_0 = \arg(z_0)$, the theorem will take the form

$$\begin{aligned}\mathscr{F}\left\{e^{i\omega_0 n}x(n)\right\} = \mathcal{Z}\left\{e^{i\omega_0 n}x(n)\right\} &= \mathcal{Z}\left\{\left(e^{i\omega_0}\right)^n x(n)\right\} \\ &= \mathcal{Z}\left\{z_0^n x(n)\right\} \overset{A.3}{=} X_{\mathcal{Z}}\left(z/z_0\right) = X_{\mathcal{Z}}\left(e^{i\omega}/e^{i\omega_0}\right) \\ &= X_{\mathcal{Z}}\left(e^{i(\omega-\omega_0)}\right) = X_{\mathscr{F}}\left(\omega-\omega_0\right).\end{aligned}$$

Proof. Let $x(n)$ be a sequence, $X(z)$ its z-transform, and $z_0 \in \mathbb{C}$. Now we get

$$\begin{aligned}\mathcal{Z}\left\{z_0^n x(n)\right\} &= \sum_{k=-\infty}^{\infty} z_0^k x(k) z^{-k} = \sum_{k=-\infty}^{\infty} x(k)(1/z_0)^{-k} z^{-k} \\ &= \sum_{k=-\infty}^{\infty} x(k)(z/z_0)^{-k} = X(z/z_0).\end{aligned}$$

□

A.4 Theorem. *Let $x(n)$ be a sequence and $X(z)$ its z-transform. Then we have*

$$\mathcal{Z}\left\{\overline{x(n)}\right\} = \overline{X\left(\overline{z}\right)}.$$

Note. If we again restrict $|z| = 1$ and write $z = e^{i\omega}$, we get for the DTFT the equivalent expression

$$\begin{aligned}\mathscr{F}\left\{\overline{x(n)}\right\} = \mathcal{Z}\left\{\overline{x(n)}\right\} &\overset{A.4}{=} \overline{X_{\mathcal{Z}}\left(\overline{z}\right)} = \overline{X_{\mathcal{Z}}\left(\overline{e^{i\omega}}\right)} \\ &= \overline{X_{\mathcal{Z}}\left(e^{-i\omega}\right)} = \overline{X_{\mathscr{F}}\left(-\omega\right)},\end{aligned}$$

as mentioned in table 2.1 .

Proof. Let $x(n)$ be a sequence and $X(z)$ its z-transform. Now we get

$$\mathcal{Z}\left\{\overline{x(n)}\right\} = \sum_{k=-\infty}^{\infty} \overline{x(k)} z^{-k} = \overline{\sum_{k=-\infty}^{\infty} \overline{x(k)} z^{-k}} = \overline{\sum_{k=-\infty}^{\infty} x(k)\overline{z^{-k}}}$$

$$= \overline{\sum_{k=-\infty}^{\infty} x(k)\overline{z}^{-k}} = \overline{X\left(\overline{z}\right)}.$$

□

A.5 Theorem. *Let $x(n)$ be a sequence and $X(z)$ its z-transform. Now it holds for the z-transform of the time-reversed signal that*

$$\mathcal{Z}\left\{x(-n)\right\} = X\left(z^{-1}\right).$$

Proof. Let $x(n)$ be a sequence and $X(z)$ its z-transform. Now we get with the substitution $m = -k$ that

$$\begin{aligned}\mathcal{Z}\left\{x(-n)\right\} &= \sum_{k=-\infty}^{\infty} x(-k)z^{-k} = \sum_{m=-\infty}^{\infty} x(m)z^{m} \\ &= \sum_{m=-\infty}^{\infty} x(m)\left(z^{-1}\right)^{-m} = X\left(z^{-1}\right).\end{aligned}$$

□

A.6 Theorem (The convolution theorem for z-transforms)**.** *Let $x(n)$ and $y(n)$ be sequences, and $X(z)$ and $Y(z)$ their respective z-transforms. Now it holds for the z-transform of the discrete convolution that*

$$\mathcal{Z}\left\{(x * y)(n)\right\} = X\left(z\right)Y\left(z\right).$$

Proof. Let $x(n)$ and $y(n)$ be sequences, and $X(z)$ and $Y(z)$ their respective z-transforms. Now we get with the substitution $n = k - m$ that

$$\begin{aligned}\mathcal{Z}\left\{(x * y)(n)\right\} &= \sum_{k=-\infty}^{\infty} (x * y)(k)z^{-k} \\ &= \sum_{k=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} x(m)y(k-m)z^{-k} \\ &= \sum_{m=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} x(m)y(k-m)z^{-k} \\ &= \sum_{m=-\infty}^{\infty} x(m)\sum_{k=-\infty}^{\infty} y(k-m)z^{-k} \\ &= \sum_{m=-\infty}^{\infty} x(m)\sum_{n=-\infty}^{\infty} y(n)z^{-(n+m)} \\ &= \sum_{m=-\infty}^{\infty} x(m)\sum_{n=-\infty}^{\infty} y(n)z^{-n}z^{-m}\end{aligned}$$

$$= \sum_{m=-\infty}^{\infty} x(m) z^{-m} \sum_{n=-\infty}^{\infty} y(n) z^{-n}$$
$$= X(z) Y(z).$$

□