Analysis of a Latent Prosody Space for Controlling Speaking Styles in Finnish End-to-End Speech Synthesis

Tuukka Törö

Master's thesis

Master's Programme in Linguistic Diversity and Digital Humanities

Phonetics

Faculty of Arts

University of Helsinki

May 2022

Abstract

Faculty: Faculty of Humanities

Study programme: Master's Programme in Linguistic Diversity and Digital Humanities

Study track: Phonetics

Author: Tuukka Törö

Title: Analysis of a Latent Prosody Space for Controlling Speaking Styles in Finnish End-to-End Speech Synthesis

Type: Master's thesis

Month and year: May 2022

Number of pages: 4+40

Keywords: Speech synthesis, speaking style, prosody, latent space, deep learning

Supervisors: Juraj Šimko, Martti Vainio

Where deposited: Helsinki University library

Abstract:

In recent years, advances in deep learning have made it possible to develop neural speech synthesizers that not only generate near natural speech but also enable us to control its acoustic features. This means it is possible to synthesize expressive speech with different speaking styles that fit a given context. One way to achieve this control is by adding a reference encoder on the synthesizer that works as a bottleneck modeling a prosody related latent space.

The aim of this study was to analyze how the latent space of a reference encoder models diverse and realistic speaking styles, and what correlation there is between the phonetic features of encoded utterances and their latent space representations. Another aim was to analyze how the synthesizer output could be controlled in terms of speaking styles. The model used in the study was a Tacotron 2 speech synthesizer with a reference encoder that was trained with read speech uttered in various styles by one female speaker.

The latent space was analyzed with principal component analysis on the reference encoder outputs for all of the utterances in order to extract salient features that differentiate the styles. Basing on the assumption that there are acoustic correlates to speaking styles, a possible connection between the principal components and measured acoustic features of the encoded utterances was investigated. For the synthesizer output, two evaluations were conducted: an objective evaluation assessing acoustic features and a subjective evaluation assessing appropriateness of synthesized speech in regard to the uttered sentence.

The results showed that the reference encoder modeled stylistic differences well, but the styles were complex with major internal variation within the styles. The principal component analysis disentangled the acoustic features somewhat and a statistical analysis showed a correlation between the latent space and prosodic features. The objective evaluation suggested that the synthesizer did not produce all of the acoustic features of the styles, but the subjective evaluation showed that it did enough to affect judgments of appropriateness, i.e., speech synthesized in an informal style was deemed more appropriate than formal style for informal style sentences and vice versa.

Tiivistelmä

Tiedekunta: Humanistinen tiedekunta

Koulutusohjelma: Kielellisen diversiteetin ja digitaalisten ihmistieteiden koulutusohjelma

Opintosuunta: Fonetiikka

Tekijä: Tuukka Törö

Työn nimi: Latentin prosodia-avaruuden analysointi ja puhetyylien hallinta suomenkielisessä end-to-end puhesynteesissä

Työn laji: Maisterin tutkielma

Kuukausi ja vuosi: Toukokuu 2022

Sivumäärä: 4+40

Avainsanat: Puhesynteesi, puhetyyli, prosodia, latentti avaruus, syväoppiminen

Ohjaaja tai ohjaajat: Juraj Šimko, Martti Vainio

Säilytyspaikka: Helsingin yliopiston kirjasto

Tiivistelmä:

Viime vuosina syväoppimisen saralla tapahtunut kehitys on mahdollistanut neuroverkkoihin perustuvan puhesynteesin, joka lähes luonnollisen puheen tuottamisen lisäksi sallii syntetisoidun puheen akustisten ominaisuuksien hallinnan. Tämä merkitsee sitä, että on mahdollista tuottaa eloisaa puhetta eri tyyleillä, jotka sopivat kyseiseen kontekstiin. Yksi tapa, jolla tämä voidaan saavuttaa, on lisätä syntetisaattoriin referenssi-enkooderi, joka toimii pullonkaulana mallintaen prosodiaan liittyvän latentin avaruuden.

Tämän tutkimuksen päämääränä oli analysoida kuinka referenssi-enkooderin latentti avaruus mallintaa moninaisia ja realistisia puhetyylejä, ja miten puheennosten akustiset ominaisuudet ja niiden latentin avaruuden representaatiot korreloivat keskenään. Toinen päämäärä oli arvioida kuinka syntetisoidun puheen tyyliä voi kontrolloida. Tutkimuksessa käytettiin referenssi-enkooderilla varustettua Tacotron 2 syntetisaattoria, joka oli koulutettu yhden naispuhujan luetulla puheella usealla puhetyylillä.

Latenttia avaruutta analysoitiin tekemällä pääkomponenttianalyysi puhedatan kaikista puheennoksista otetuille referenssivektoreille, jotta saataisiin esille puhetyylien keskeisimmät erot. Olettaen puhetyyleillä olevan akustisia korrelaatteja, tutkittiin pääkomponenttien ja mitattujen akustisten ominaisuuksien välillä olevaa mahdollista yhteyttä. Syntetisoitua puhetta analysoitiin kahdella tapaa: objektiivisella evaluaatiolla, joka arvioi akustisia ominaisuuksia ja subjektiivisella evaluaatiolla, joka arvioi syntetisoidun puheen sopivuutta liittyen puhuttuun lauseeseen.

Tulokset osoittivat, että referenssienkooderi mallinsi tyylillisiä eroja hyvin, mutta tyylit olivat monisyisiä ja niissä oli merkittävää sisäistä vaihtelua. Pääkomponenttianalyysi erotteli akustiset piirteet jossain määrin, ja tilastollinen analyysi osoitti yhteyden latentin avaruuden ja prosodisten ominaisuuksien välillä. Objektiivinen evaluaatio antoi ymmärtää, että syntetisaattori ei tuottanut tyylien kaikkia akustisia ominaisuuksia, mutta subjektiivinen evaluaatio näytti, että mallinnus riitti vaikuttamaan sopivuuteen liittyviin arvioihin. Toisin sanoen spontaanilla tyylillä syntetisoitua puhetta pidettiin formaalia sopivampana spontaaniin tekstityyliin ja päinvastoin.

Table of Contents

1 Introduction

Speaking styles have been researched both within phonetics and sociolinguistics, and more recently, in the world of speech technology. In recent years, developments in computing power have led to major advances in deep learning, which has broadened possibilities for controlling speech synthesis using latent spaces such as neural network bottlenecks. The aim of this thesis is to explore how the reference encoder of an end-to-end speech synthesizer trained with a diversity of speaking styles models prosodic features of speaking styles and how those styles can be controlled.

Controlling prosody is a hot topic within speech synthesis research, but the data used tends to be speech in highly controlled, stereotypical styles without much ecological validity. The styles are usually on the lines of “angry” and “happy” instead of more subtle and realistic ones. In this thesis, I will analyze speech elicited from less controlled and more diverse text styles ranging from fairy tales and Wikipedia texts to film subtitles and YouTube captions. After the statistical analysis, I will synthesize speech on a stylistic continuum and evaluate the synthesizer output. The evaluation is done both through subjective evaluation – in the form of a listening test – as well as by taking a peek inside the black box and extracting information on the acoustic features the reference encoder models.

The study incorporates theory of speaking style research from phonetics and sociolinguistics into practical speech synthesis and explores possibilities for controlling more ecologically valid speaking styles. My academic motivation for the work lies both in the implications that realistic synthesized speaking styles can have for accessibility and domains such as clinical speech synthesis as well as in the potential that these kinds of speech models could be used in phonetics research, e.g., for typological and sociophonetic studies.

The research questions are as follows: 1) Does the reference encoder model speaking styles from stylistically diverse speech data, and can the potential stylistic differences and their acoustic correlates be unearthed with statistical methods? 2) Is the synthesizer able to form utterances with specific speaking styles, and can this be controlled?

While the reference encoder does not model the human articulatory system per se – instead modeling structures in the corpus – the different acoustic features it takes as input are connected in the highly intertwined pathway between the lungs and the lips. Thus, we begin the background section with a short introduction into articulation, then proceed to acoustics, prosody, speaking styles, and finally to deep learning.

2 Background

2.1 Articulation and Speech Acoustics

Most human speech is produced by air blown from the lungs through the vocal tract. In voiced sounds, the vocal folds in the larynx are narrowed by the arytaenoid cartilage and tensed by the thyroid cartilage (Aaltonen et al., 2009, p. 142; Stevens 1998). The constrained air flow makes the folds vibrate, creating minute, propagating changes in air pressure called sound waves. This buzz-sound – a complex set of waves of different wavelengths – is called the source. The source sound is then filtered by the rest of the vocal tract, mostly by the natural resonances of the pharyngeal, oral and nasal cavities, as well as by the superposition of the sound waves (Benesty et al., 2008, p. 7).

Humans are able to shape parts of the vocal tract – mainly by moving their lips, tongue, lower jaw and velum (for opening and closing the nasal cavity) – to amplify different sound waves, resulting in a series of acoustic effects a listener can decode into meaningful messages (Aaltonen et al., 2009, pp. 65,136; Stevens 1998; Benesty et al., 2008, p. 7). While filtering gives us the qualities of different vowels, nasals and so on, the raw buzzing sound emanating from the vocal folds itself conveys many features needed for communication. It contains the information needed for the listener to perceive *pitch*, *intensity* and, consequently, *duration*.

2.1.1 Pitch

Pitch is the perceptual notion of the “highness” and “lowness” of a sound. An acoustic correlate of pitch is the fundamental frequency, f_0. Voiced sounds such as the vowel /a/ – as opposed to noise like the glottal fricative /h/ – have a harmonic structure, which means that there are relatively high-amplitude sound waves, or components, on frequencies that are integer multiplications of the f_0. Thus, f_0 denotes the lowest harmonic component in a complex sound (Rossing et al., 2007, p. 486, Benesty et al., 2008, p. 12).

The f_0 of a human adult, i.e., the natural resonance of the vocal folds, is around 100-300Hz depending on the size of the folds. The resonance can be manipulated by changing the stiffness and mass of the vocal folds by stretching them: When the vocal folds are long and thin the f_0 is higher, and when they are short and thick, the f_0 is lower (Stevens, 1998, p. 57, 73; Benesty et al., 2008, p. 12). An adult male can produce f_0 ranging between around 80 and 400Hz whereas women are able to produce f_0 between around 120 and 800Hz (Benesty et al., 2008, p. 12). While pitch is primarily tuned by stretching the vocal folds, the level of air pressure with which air is blown from the lungs has a secondary effect in that higher pressure tends to decrease the duration of vibration cycles in

the vocal folds thus increasing f_0 (Aaltonen et al., pp. 64,142; Stevens, 1998, pp. 76-77; Rossing et al., 2007, p. 675).

2.1.2 Intensity

Intensity is for loudness what fundamental frequency is for pitch. Loudness is something we perceive, and the acoustic measure behind it is intensity, which is determined by the amplitude of the sound wave. Changes in vocal intensity are mainly produced by changing air pressure below the glottis, i.e. the pressure of the air coming from the lungs, so much so that a doubling in the air pressure in the lungs increases sound pressure by fourfold, while every doubling of sound pressure quadruples intensity (Gick et al., 2013, p. 68; Sundberg et al., 2004; Rossing, 2014, p. 103). The human ear's perception of loudness is non-linear, and thus it is convenient to use a logarithmic scale for the perception of sound, the decibel scale, dB. The scale starts from 0dB which is determined by the threshold for hearing a 1000Hz sound wave, while 130dB is the threshold where we start feeling pain from the air pressure. Between the hearing and pain thresholds, intensity level increases 10^6, i.e. it becomes a million times higher (Berg, n.d.; Rossing, 2014, p. 103).

2.1.3 Duration

Apart from factors like the phonological features of a specific language – such as contrastive quantity – or whether the speaker reads from a paper or is having a conversation, there are intrinsic durational aspects to human speech production that stem from the mechanics of articulation. For example, the movement of an articulator from a constricted position to a vowel and back to the original position usually takes at least 200ms, and there can be differences between the velocity of movement of different parts of the same articulator. For example, the movement of the tip of the tongue for alveolar stops is faster than the movement of the tongue body for velar stops (Stevens, 1998). Thus, there are constraints to how fast we can speak. We perceive duration, for example, from single phonemes to whole utterances as well as how long pauses there are and how often they occur.

2.1.4 Voice Quality

When we have accounted for pitch, intensity and duration, there is at least one important suprasegmental feature that is still unaccounted for, voice quality. Compared to pitch and loudness, voice quality is difficult to define, as it cannot be described with one acoustic characteristic such as fundamental frequency or intensity. Different voice qualities can be described as, for example,

breathy, creaky, soft or rough (Barstiers & De Bodt, 2015; Gordon & Lagdefoged, 2001), and voice quality also serves a role in gender identification of a speaker (Mendoza et al. 1996). One useful acoustic parameter for examining voice quality is *spectral tilt*. Spectral tilt is the measure of how much intensity decreases as frequency increases in the power spectrum, i.e., what amplitudes do sound waves have at different frequencies of a complex sound. Generally, the breathier voice, the more negative the tilt and creakier the voice, the more positive the tilt. Spectral tilt is computed by subtracting the amplitude of f_0 from the amplitudes of frequencies above it, e.g., the first and second harmonics, and comparing their difference (Gordon & Lagdefoged, 2001; Styler, 2022).

2.2 Prosody

Prosody evades a clear bordered definition, but we can think of it as a set of overlapping suprasegmental patterns in human speech. The most important features of prosody are intonation, emphasis and speaking rate (Aaltonen et al., 2009, p. 214) which are closely related to the aforementioned pitch, intensity, and duration.

Although we are not often aware of prosody, it has a wide variety of functions, from lexical and grammatical to conveying emotion and attitudes. It can help us infer information about the agenda, health, background, and emotional state of an interlocutor. It is used to bring focus to new information that is not shared by the interlocutors (Vainio & Järvikivi, 2007), to ask questions and imply refusal, to denote lexical differences, and to negotiate turn taking in a conversation (Aaltonen et al., 2009, pp. 49, 142, 277-278).

Even though prosody is not something we necessarily notice on the conscious level while we converse, it is important to understand it as a perceptual phenomenon. However, there are no clear cut parameters for prosodic features. While the three primary ways to perceive prosody are duration, emphasis and intonation, the acoustic features behind them are intertwined: Quantity is not only about perceiving the durational contrast between long and short phones but also f_0 chimes in. Similarly, emphasis is not only about intensity, but at least f_0 is involved in this as well. It may also be that the perception of emphasis is also based on duration and on precision of articulation just as it is about changes in intensity (Aaltonen et al., 2009, pp. 214-215). All of these prosodic phenomena take part in forming what we call speaking styles.

2.3 Defining Speaking Styles

Even though speaking styles have been researched in depth within both phonetics and sociolinguistics, neither discipline has offered a standard definition for style (Wagner et. al., 2015).

The different, often overlapping categories for labeling speaking styles, such as *lab speech, read speech, spontaneous speech,* and *professional speech* are vague and hard to delineate. After all, a parent reading a fable to a child and an English professor reciting Shakespeare in front of university students are surely not uttered in the same style, let alone a radio presenter reading marine weather forecasts. Similarly, spontaneous speech varies between situations and can be influenced by things such as the relationship between the interlocutors and by their respective backgrounds. For example, a middle-aged male politician might be more comfortable with speaking in an informal manner in public than a young female one because he is not under the same gendered looking glass when it comes to his abilities as a leader.

Phoneticians have tended to look at styles from a binary perspective where *spontaneous* – or "natural" – speech is opposed with *lab speech*. The terms are not clearly defined, but they can be described according to the level of control in the situation by the researcher (Wagner et. al., 2015). In its extreme end, lab speech constitutes words in isolation or in context of carrier sentences outside of realistic pragmatic or communicative contexts while spontaneous speech is elicited under more ecologically valid conditions (Llisterri, 1992; Wagner et. al., 2015). However, Wagner et. al. (2015) point out that in a wider sense, most research in phonetics use data that was recorded in a laboratory setting and could be categorized as lab speech. It is not necessarily even possible to achieve complete ecological validity and to get rid of all aspects of acting involved in a recording situation. They posit that instead of a dichotomous categorization between lab speech and "natural" speech, we should describe the recording context according to the level of control in the situation.

While Wagner et al. propose a more specified terminology and methodological grounding for analyzing speaking styles, the continuum still operates in the same domain as the lab speech vs. natural speech dichotomy. They use style as "a broad cover term for speaking situations leading to phonetic variation where the level of control is the main descriptive factor." However, this focus on the level of control might miss out on some aspects of stylistic variation. Let's look at an example from Wagner et al.: "a news broadcaster speaking in a conversational style would be as unacceptable for listeners as a friend talking to you in a news broadcasting style." Could one describe all the differences between these styles through level of control? Of course we might say that the style of a broadcaster is more controlled and follows rigid rules, but one merely needs to watch five minutes of the US news programmes NBC Nightly News and PBS Newshour – the former being on a commercial channel and the latter on a publicly owned one – to see that strikingly different speaking styles may be elicited between almost identical contexts.

Thus, if we are to analyze speaking styles that have the same level of control, e.g., read paragraphs from different types of work of fiction, we might want to look toward the social aspects

of recording events. Joos (1968) – speaking of English – says that if we are ever to define speaking styles, “the several styles will be found to be correlated to an equal or greater number of sociologically definable occasions.” Labov (1972) describes speaking styles by the situational contexts they can be elicited in. He differentiates between *reading style* (e.g. monologues, word lists, and minimal pairs), *careful speech, spontaneous speech, and casual speech*, which are set in a continuum with reading style on one end, and casual speech that occurs in informal situations on the other. Labov posits that casual and spontaneous speech styles cannot be elicited in the “social constraints of the interview situation” (1972, p. 79). Thus, he shares a similar dichotomy between lab speech and “natural” speech with phoneticians. According to Labov, casual speech is spoken in situations where no attention is given to the speech act itself while spontaneous speech is “excited, emotionally charged speech when the constraints of a formal situation are overridden” (1972, p. 86) . Compared to the binary opposition traditionally found in phonetics, Labov’s styles are more nuanced. “Natural” speech can encompass at least Labov’s *careful*, *casual*, and *spontaneous* styles and perhaps even more controlled styles, while *reading style*, with word lists and minimal pairs, best correspond to lab speech. Labov goes on to define five contexts where the aforementioned social constraints of the interview situation could be overcome: *speech outside the interview*, e.g. the participant chatting with the researcher before the interview begins, *speech with a third person, speech not in direct response to questions, childhood rhymes and customs,* and *danger of death* where the subject gives a subjective account of a near death experience.

While this thesis is not about sociolinguistics nor about eliciting speech without the constraints of the recording event, the “sociologically definable occasions” – as Joos puts it – will affect the styles, even if they are acted. For example, some differences, such as, between read speech and scripted “spontaneous” speech, might be connected with what Gregory and Carroll (1978) call “the relationship between speaker and the medium of transmission”, i.e. whether you are reading from a paper or improvising. Others, for example, the possible acoustic differences between reading adult prose, a fairy tale, or Wikipedia, will be connected to social expectations. According to Joos (1968), the social occasion and corresponding style correlate in two ways: “the speaker uses the style that suits the occasion” and simultaneously “defines the occasion for the listener (and for himself (sic.)) by his ‘choice’ of style” (p. 189).

The neural network architecture used in this study may not be able to model and disentangle minute variation between speaking styles, but if it can extract significant differences between, e.g., lab speech, prose style read speech and scripted spontaneous speech, there might be potential for developing these tools not only for speech synthesis but also for phonetic analysis. This could remedy the issue of using less controlled, more ecologically valid, data which are difficult to

analyze with traditional measurements. This is especially true for situations where lab recorded read speech is difficult to produce. These include studying children's speech, languages without a standardized orthography, and speakers who lack the ability to read fluently. In fact, as Wagner et al. (2015) state, if we only use those who are used to reading aloud, i.e., read fluently, it will lead to many groups of people being excluded from such studies.

For categorizing speaking styles, Gregory and Carroll (2019) have posited three main dimensions for stylistic variation: *fields of discourse, modes of discourse,* and *tenors of discourse*. The field of discourse entails what the speech act is about, such as topic and subject matter, while mode is concerned with the relationship between the speaker and the "medium of transmission", e.g., reading from a paper versus speaking spontaneously. The tenor of discourse covers both the personal tenor, the relationship between the speaker and her audience, and the functional tenor – the speakers agenda – such as whether she is trying to persuade the audience, to make them laugh, or to teach them (Gregory & Caroll, 2019, p. 8). These three dimensions can be combined into registers: expectations of style that stem from their repeated association to a specific situation type (Gregory & Caroll, 2019, p.72). A stereotypical example could be drawn from a hierarchical social system like the military where the expected style of speaking is derived from the topic of conversation, the formal expectations to dialogue within the ranks, and both the personal relationship between the interlocutors, and what is the speaker attempting to "do" to the audience, e.g., control them.

These three extralinguistic situational dimensions seem useful for accounting for more complex stylistic variation than a single continuum based on the level of control. Speech is always affected by a set of social variables which would imply that there are more than one dimension of continua involved. For example, to elicit ecologically valid speech, one might add some background noise, such as pub chatter, on the speaker's headphones, or build a recording studio that looks like a living room. As soon as you create an environment for a speech act, there are social variables that affect the way we speak, even if the environment is an anechoic chamber with zero ecological validity.

When it comes to ecological validity, one might question whether on not different styles of read speech are sufficient for modeling speaking styles. Here we encounter the general limitations to lab speech as well as the importance of contextual appropriateness.

When we recreate real life contexts in a laboratory, such as giving directions on a map, the participants' motivation is to solve the problem not because they are lost in the city, but because the researcher asked them to (Wagner et. al., 2015). Thus, Wagner et. al. argue that these types of data should be described as "mimicking a target speaking style" based on the assumption that they will always differ from data gathered in real life situations. This assumption is supported by evidence that there are differences in voice quality and f_0 contours which listeners can distinguish between

real and acted emotion (Jürgens, Hammerschmidt, & Fischer, 2011), and between read and spontaneous speech (Laan, 1997; Dellwo et. al. 2015). However, in the case of read vs. spontaneous speech this is a bit more complicated, as read speech can be very difficult to distinguish from spontaneous speech if the material is written in a conversational style and the subjects act as if they are conversing spontaneously (Mixdorff & Pfitzinger, 2004).

It has been argued that – outside of eavesdropping – there is no way to record completely authentic data (see "observer's paradox" in Labov, 1972). If we look at Labov's "danger of death" context which is used to elicit spontaneous speech: Asking someone to tell their account of a near death experience elicits a kind of simulation of those emotions in a controlled environment. Similarly, we can think that an actor reading a fairy tale simulates a relationship between the reader and audience – an adult and a child – even though she is reading the story in a laboratory environment. Instead of looking at authenticity, it might be better to judge styles by their contextual appropriateness. A casual spontaneous style does not have an innate "natural" quality compared to stylized read speech, but instead is more or less appropriate for a specific situation or useful for answering a specific research question (Wagner et al., 2015).

Even if we come to terms with not being able to create a perfect environment where the subjects speak like they would if there was no one recording, there are other issues to eliciting speaking styles in a laboratory setting. If the scenario is not controlled clearly enough, let's say the subject is merely asked to speak in a conversational style, we will not know all the contextual variables involved, such as who is she speaking to, or what she is trying to achieve? However, while it is in the scope of this study to analyze the differences between styles, the main motivation is not to find the best way to elicit the most "natural" styles but primarily to explore how realistic and diverse stylistic variation can be controlled in speech synthesis.

The data used in this study without a doubt is of the less-than-authentic kind, as it is scripted. However, if it is to yield positive results, the method could well be used for actual spontaneous speech. Furthermore, there are domains – such as human-machine interaction – in which real authenticity is not necessarily something to strive for. One might not want to train a computer to convey "real" emotion but instead a speaking style that is authentic enough that one can suspend their disbelief but is recognizable as acted might be more suitable. However, there are applications in clinical speech synthesis – such as creating voices for people with motor disorders – where it would be useful to convey as-real-as-possible spontaneous speech with expressiveness and emotion. Here, actual spontaneous speech would be the most appropriate source of data.

2.4 Acoustic Correlates of Speaking Styles

Within the domain of phonetics, speaking styles can be studied both from the perspective of segmental and suprasegmental features. Segmental features include things like *vowel quality*, *coarticulation*, and *vowel reduction*, which can be analyzed both through their acoustic and articulatory correlates. Suprasegmentals – which are in the center of this thesis – can be uncovered mainly by analyzing f_0, intensity, and duration. Duration can be extracted from phonemes, syllables, and words to whole utterances and analyzed, e.g., in terms of pauses and articulation rate. Analyzable aspects of f_0 include its mean, minimum and maximum, range, *SD* (standard deviation), and slope. Many of these phenomena may be intertwined, for example, articulation rate is likely to affect the slope, as faster articulation will also lead to steeper changes in f_0.

The connection between suprasegmental patterns and different speaking styles have been studied extensively, at least in the case of widely spoken Indo-European languages. Previous studies have found that compared to spontaneous speech, read speech has a lower speech rate (Laan, 1996; Koopmans & van Beinum, 1991), higher f_0 median (Koopmans & van Beinum, 1991), wider f_0 range (Blaauw, 1991; Wagner & Windman, 2015), and lower f_0 minimum (Wagner & Windmann, 2015). These findings, however, may not represent the "whole truth". According to Toivola and Lennes (2013) read speech is faster than spontaneous speech, and Wagner and Windmann (2015) also found in their study that read speech was faster than spontaneous speech. It seems there may be differences between languages or other factors than read vs. spontaneous speech involved in speaking rate.

When it comes to pitch patterns, there might not be a difference between read and spontaneous speech (Bruce & Touati,1991), but a study found a difference both in melodic patterns and in f_0 range between different styles of spontaneous speech (Bhatt & Leon, 1991). A study on Portuguese (Delgado-Martins & Freitas, 1991) found that there are differences in the relationship of pauses and spoken sequences between read, spontaneous and professional speaking styles. While these results cannot be expected to be transferable to Finnish, it shows there are distinguishable prosodic patterns at least between read and spontaneous speech as well as between different unscripted styles. It may also be that the function of prosody in read speech is different from spontaneous speech as there is less need to negotiate turn taking and to establish common ground between speakers. This could explain the lower f_0 minimum and wider f_0 range of read speech compared to spontaneous speech (Wagner & Windmann, 2015).

2.5 Deep Learning

If we want to use computers to model complex real-world phenomena, hard-coded "if-else" type systems quickly become insufficient. Instead, we need machine learning approaches where computers autonomously extract patterns from raw data through experience. Machine learning covers methods from simple naïve Bayesian algorithms that can assign data into categories, such as whether an email is spam or legitimate (Jurafsky & Martin, 2020, p. 55), to complex architectures of deep neural networks that can generate speech so natural that it is near indistinguishable from the real thing. Computers can achieve this by learning to "understand the world in terms of a hierarchy of concepts, with each concept defined through its relation to simpler concepts" (Goodfellow et al., 2016, p.1).

One way to train a machine learning algorithm is to give it a representation of the data as a set of features engineered by a human, which works well for tasks where features are easily identifiable. Another way is to use machine learning on raw data to find out those features. This is called *representation learning*, and it becomes handy in situations where questions become too complex for humans to know which features to extract (Goodfellow et al., 2016). It may be, for example, feasible to give an algorithm the main features that contribute to the likelihood of developing a certain disease, but it is harder to extract features that differentiate between a picture of a cat and one of a dog. The building blocks of representation learning are called *factors of variation*. They are the different – often unobserved – factors, or even abstract concepts, that influence the data we see (Goodfellow et al., 2016, p. 5). When researching speaking styles, there are factors we need to analyze to understand variation – such as whether the person is speaking spontaneously or reading from a paper – as well as factors we might want to disregard, such as gender, age, or accent. Thus, we need to be able to disentangle them; to extract features that are contributing to a specific concept. Goodfellow et al. (2016) say, to learn these representations, deep neural networks represent complex and abstract concepts by combining them from simple features. In an image recognition task – which can also be a graphical representation of sound such as a spectrogram – the first layer may recognize small patterns such as edges. As the data propagates through the layers, the patterns begin combining into more complex ones until finally a layer recognizes, e.g., a nose, eyes and ears, and the next one combines these features to recognize a human face (Goodfellow et al., 2016, p.5; Jurafsky and Martin, 2020, p. 127). This serves as an example of how neural networks understand the world as a "hierarchy of concepts".

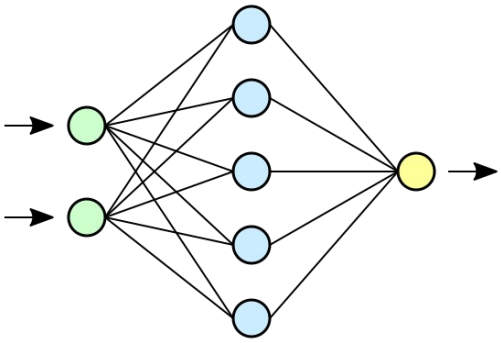

Figure 1: A simple fully-connected neural network. Source: Wikimedia Commons

A feed-forward deep neural network consists of layers of interconnected nodes that takes a vector of input values on its visible layer, followed by hidden layers, and outputs some new value(s) (Jurafsky & Martin, 2020). An individual node takes the weighted sum of the outputs, or activations, from the nodes on the previous layer, applies an activation function on the value, and outputs it for the nodes on the next layer to take as input. The connections between the nodes assign weights on the outputs which determine how much of an effect the signal has on the next node. The weights are adjusted to adapt to the data at hand as the network learns from experience. The weighted sum z can be represented as $z = w \cdot x + b$, where w is the weight, x the input and b a scalar bias. A non-linear activation function is then applied to z that maps the output to a range, such as *x is x when it is positive and otherwise it is 0*, as is the case with the ReLU function (Kröse et al., 1993, pp. 15-17, Jurafsky & Martin, 2020, pp. 128-129). Traditional fully connected networks are computationally heavy: the parameters rapidly proliferate as the amount of neurons and layers grow. They are also prone to overfitting, i.e., to learn idiosyncracies from the data that are not generalizable. A more suitable neural network architecture for time-series and image data is the convolutional neural network (CNN). CNNs are otherwise similar to fully connected neural networks but instead of connecting all of the neurons between layers and computing a matrix multiplication, neurons only connect to a small number of corresponding neurons on the previous layer, using convolution as a filter that recognizes features from the input. The output of the layer will then be pooled, i.e., downsampled to reduce the parameters needed to represent the data (O'Shea & Nash, 2015).

The way that a neural network can learn from experience, is by adjusting the weights and biases that determine the strength of the signal that a node outputs. When we train a feed-forward network we have a gold output, the correct result that the network should give, which the actual output of the network can be compared with. As the network is trained, its weights are adjusted so that the

network will approximate the gold output as close as possible. To be able to do this, a loss function is used to calculate the distance between the gold and the actual output. We want to find the optimal weights to minimize this distance. This can be done by iteratively updating them by *gradient descent*, which means finding "a minimum of a function by figuring out in which direction [weights as its parameters] ... the function's slope is rising the most steeply, and moving in the opposite direction" (Jurafsky & Martin, 2020, p. 137). To find this direction not only on the last layer, but through the network, we must calculate the gradient, $\nabla\theta J(\theta)$, of the loss function in relation to weights by *backpropagation*. Backpropagation, aptly named, is an algorithm that lets the information from the loss function to propagate backward through the hidden layers of the network, so that a gradient can be calculated and weights tuned accordingly (Goodfellow et al., 2016, p. 200; Jurafsky & Martin, 2020, p. 84, 137-139).

One type of deep learning concerns itself with dimensionality reduction. An example of these kinds of networks are autoencoders. They are encoder-decoder networks that have a bottleneck layer with a considerably lower dimensionality than the input and output layers. This forces them to prioritize what they copy by finding the most important patterns in the data. In an autoencoder, the input and target – or gold – output are the same. Thus, often it is not the output of the autoencoder that we are interested in but the bottleneck as we want to store the information in a dimensionally reduced representation. The output of the bottleneck, also called the latent space, is a vector that should contain the most salient structural information of the input signal. Autoencoders have been used in, for example, dimensionality reduction, feature learning, denoising, and image compression (Goodfellow et al., 2016, Spinner et al., 2018). This takes us to current research in speech synthesis.

2.6 Related Studies

As computing power and new innovations have made more complex phenomena possible to be modeled by neural networks, controlling prosody has become a hot topic in neural speech synthesis. Wang et al. (2018) proposed global style tokens, a speaking style and prosody related latent space trained together with Tacotron in "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis." A similar approach was taken with a simpler reference encoder in Skerry-Ryan et al. (2018) in "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron."

What comes to research using these kinds of methods, there are at least a few different focus points: some studies have focused on disentangling and controlling specific prosodic features such as Mohan et al. (2021), Shechtman & Sorin (2019) and Raitio et al. (2020); some on prosody

transfer from an unseen speaker such as Klimkov et al. (2019); and some controlling a continuous emotion space such as Sivaprasad et al., 2021. Another theme has been to use dimensionality reduction algorithms to disentangle the prosody space, such as Tits et al. (2019a & 2019b). In Tits et al. the styles disentangled well on the continuous latent space. However, the styles used were clear cut and highly controlled, as the same set of sentences were uttered in eight acted styles: *neutral, happy, sad, bad guy, from afar, proxy, whisper,* and *old man*.

Recently, a paper (Šimko et.al, submitted) I co-authored on controlling speech synthesis – using the same model and corpus as this thesis – shows the model can disentangle some prosodic features for controlling synthesis. This was done by fitting linear models between the reference vector outputs and phonetic features of the encoded utterances. Adjusting the reference encoder vector using the estimate coefficients from the linear models, we could control phonetic features. The method concentrated on controlling specific acoustic features such as f_0 mean or speaking rate and disentangling them from other features through orthogonalization. Anecdotally, this method can be coupled together with the method in this thesis, e.g., one can start with a specific speaking style and then generate that style with a slower speaking rate or a higher pitch. However, analyzing this was not in the scope of this study as it is focused on controlling the speaking styles themselves.

Looking at speaking styles from the perspective of, on one hand speech synthesis and on the other, phonetics and sociolinguistics, the different disciplines are talking about distinctively different things. However, in order to create speech synthesis that is close to the real thing and appropriate for domains such as clinical speech synthesis, tools from phonetics and sociolinguistics can prove useful for the speech synthesis field as well.

2.7 Network Architecture

Tacotron 2 is a neural network architecture used for speech synthesis proposed in “Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions” by Shen et al. (2018). The network can be trained unsupervised and solely on data which means that instead of engineered features, it takes audio samples with their corresponding texts without labels as input. It works by aligning characters from a text input with acoustic features of the accompanying sound signal (in the form of mel-spectrograms) and generates hidden representations of said character sequences (Jurafsky & Martin, 2020, p. 567). The spectrogram – essentially a visual depiction of frequencies and amplitudes of a sound – as an 80 dimensional vector enables the network to extract these hidden representations that reflect salient features and patterns in the sequence. This teaches the

network, not only how words should be pronounced, but can also model prosodic features (Shen et al., 2018).

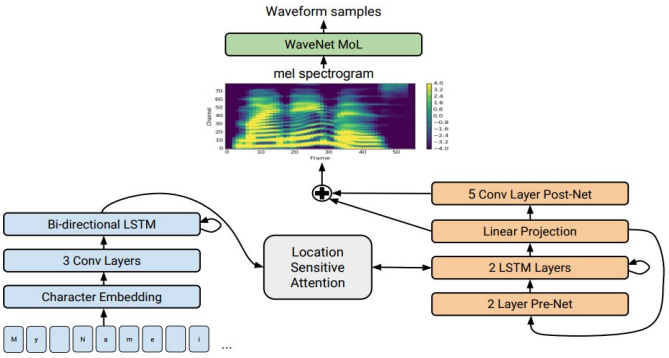

Figure 2: Original Tacotron 2 architecture (reprinted from Shen et al. 2018)

The original Tacotron 2 architecture consists of two parts: the first is a recurrent sequence-to-sequence network that predicts spectrogram frames from character sequences, while the second part is a vocoder that produces the waveform. The prediction network first creates 512-dimensional character embeddings of all the input characters. The three convolutional layers, present in figure 2, model a five character spanning context for the current character within the input sequence. Then the bi-directional LSTM produces the encoding that is fed to an attention network. Location based attention takes a weighted sum of all the encoder hidden states and the attention weights from its previous time-state (Jurafsky & Martin, 2020, p. 213). These weights allow the network to focus on the correct part of the spectrogram for a given character and to generate it into a fixed-length context vector of the whole character sequence for the decoder. The decoder, which is represented on the right-hand side of figure 2, takes the output of the encoder network and predicts the next frame in the mel-spectrogram. The decoder is an autoregressive recurrent neural network, meaning it generates its outputs by conditioning them by its own earlier output. In Tacotron 2 this is done by feeding the decoder its previous predicted mel-spectrum which together with the context vector from the attention layer are used to predict a new 80-dimensional spectrogram frame of 50 milliseconds with 12.5 millisecond frame hop (Shen et al., 2017, Jurafsky & Martin, 2020, pp. 176-181, 567).

The problem with the original Tacotron 2 is that it generates near natural speech but it generates a kind of median style of the corpus, rid of any expressiveness. In order to control that the synthesizer models prosody – and as such, speaking styles – I am using a Tacotron 2

implementation[1] based on Skerry-Ryan et al.'s (2018) "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron." The implementation includes a reference encoder, a text encoder, and a speaker encoder. The text encoder models the given input text and its relation to the acoustic signal, the speaker encoder takes into account speaker specific features. This leaves the reference encoder to deal with "the unexplained variation in the signal, i.e. the prosody and recording environment" (Skerry-Ryan et al., 2018). The reference encoder output is of relatively low dimensionality compared to the 512-dimensional vector of the text encoder, it does not use attention, and it has to be constant through the utterance. In essence, it works as a bottleneck similarly to an autoencoder (Skerry-Ryan et al., 2018).

The reference encoder consists of a 6-layer convolutional neural network with 3x3 filters and 2x2 stride, SAME padding, ReLU activation and batch normalization. To create the low dimensional output vector, the encoder uses a recurrent neural network with a Gated Recurrent Unit layer with 128 dimensions. This resulting embedding vector is called the *prosody space* (Skerry-Ryan et al., 2018).

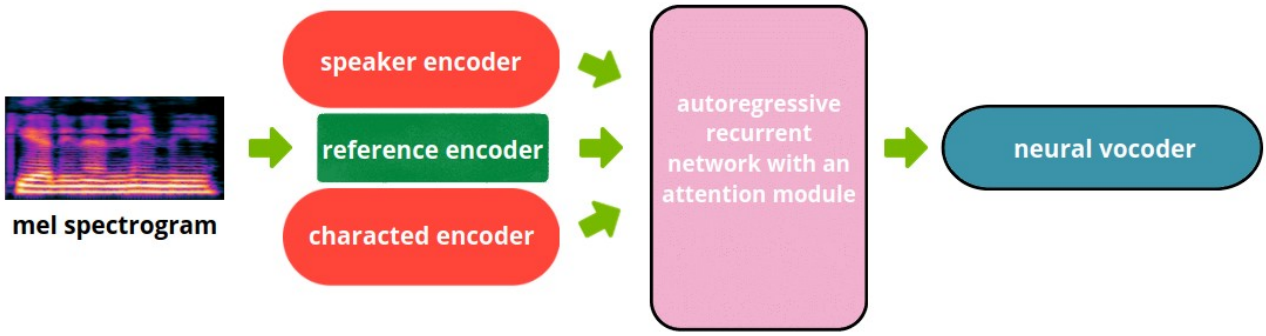

Figure 3: Architecture of Tacotron 2 with a reference encoder

The reference encoder may be able to learn disentangled factors, i.e., specific dimensions of the vector will be responsible for certain variation and only that variation from the corpus. This would mean that, in addition to successfully copying the prosody of a given utterance and being able to recreate it, specific acoustic phenomena could be controlled by adjusting values in the prosody space. Often the problem with using deep neural networks is that they are essentially black boxes, and it is difficult to extract what acoustic features the network models. If the reference encoder disentangles them, the underlying acoustic features could be derived from the vector representations. Thus, we could both visualize how speaking styles position on the latent prosody

1 The synthesizer model was provided by Antti Suni.

space and find out what acoustic features contribute to the differences, say, between formal speech and informal speech.

If the reference encoder is able to learn continuous latent representations of speaking styles that have relatively subtle differences, e.g., read formal speech and acted spontaneous speech, it could both be useful for speech synthesis but also for giving more scientifically reproducible results for speaking style research.

3 Research Questions

The aim of the thesis is to analyze how the reference encoder models speaking styles and how those styles can be controlled, as well as to evaluate how this control carries over to the output of the network architecture, the synthesized speech. The research questions are:

- **Research question 1:** Does the reference encoder model speaking styles from stylistically diverse speech data, and can the potential stylistic differences and their acoustic correlates be unearthed with statistical methods?
  This will be answered by a principal component analysis on the reference encoder output for encoded utterances spoken in different speaking styles. Also, basing on the assumption that speaking styles correlate with specific phonetic features, I will investigate a possible correspondence between the representations extracted from the reference encoder and phonetic characteristics of the encoded utterances.

- **Research question 2:** Is the synthesizer able to form utterances with specific speaking styles, and can this be controlled?
  This will be answered by synthesizing speech with styles derived from the encoded utterances. An objective evaluation is conducted by analyzing possible correspondence between the reference encoder representations and acoustic characteristics of the synthesized speech. The same analysis will also be done on the original utterances from the speech corpus, and the two results will be compared. A subjective evaluation will also be conducted with participants judging synthesized speech produced with different styles on its appropriateness given the contents of the synthesized sentence.

4 Material and Method

4.1 Material

The material used is from one female native Finnish professional speakers. It was recorded during the autumn of 2021 at the University of Helsinki for a national project for developing Finnish AI and speech synthesis.

The speaker was recorded in a highly treated recording booth for a total of 25-30 hours. The texts were read in 12 – 35 min sessions, and she was allowed to adjust her speaking to match the text style. The microphone used was an AKG C414 XLS with omni pattern and no pad or hi-pass filter on. The speaker wore Sennheiser HD250 Linear II closed back headphones when recording primary material to monitor her own speech.

The corpus contains read material spoken in several different styles. The speaker was given a text, and she chose the appropriate speaking style according to the contents of the text.

The text styles were as follows:

- YouTube captions from videos with different themes, e.g., an interview about social media and a talk about illegal nightclubs
- Rich text, e.g., news headlines and film subtitles
- Parliamentary speeches
- Discussion forum posts
- Prose texts from novels by writers such as Dumas, Gogol and Chekhov
- Children’s prose and fairy tales
- Wikipedia articles
- Blog texts
- Other fact texts

In addition to these, I am using a smaller *reference data set* of utterances from the same speaker that the synthesizer was not trained with for evaluating how the reference encoder models styles. This data set consists of 320 sentences spoken in five styles for a total of 1613 utterances (a few sentences were uttered a couple of times if the speaker or recorder were not satisfied with the resulting style). The elicited styles were loud, soft, negative, positive, and neutral, and the sentences were taken from “Developing a speech intelligibility test based on measuring speech reception thresholds in noise for English and Finnish” (Vainio et al., 2005).

4.2 Method

4.2.1 Preprocessing

The recordings were split so that every sentence was considered its own utterance while very long sentences were cut into multiple utterances at silence. The audio and transcripts were forced aligned and then manually checked and fine tuned. Repetitions and disfluencies were cut, apart from those scripted in the spontaneous style texts. The audio was loudness normalized to -21 LUFS and run through a high-pass filter at 60Hz, 24dB in Audacity (Audacity Team, 2021).

4.2.2 Network Training

The synthesizer used in the thesis was an implementation based on NVIDIA's Mellotron (Valle et al., 2020). The reference encoder and synthesizer were trained in an unsupervised manner, meaning that the utterances fed to the network were not accompanied by labels that would tell the network the expected speaking style. It was given the full signal of the 12910 utterances as 80 dimensional mel-spectrograms. The reference encoder output was set at 128 dimensions. After training, any sound file containing speech could be fed to the synthesizer which would output a reference vector for the utterance.

4.2.3 Principal Component Analysis

Principal component analysis (PCA) is a linear transformation technique that can be used for dimensionality reduction. The idea behind PCA is that often the number of variables in the data we aim to analyze is high but the underlying structures may be simple (Shlens, 2014). Thus, we want to reduce the dimensionality of the data, keeping as much statistical information intact as possible (Jolliffe & Cadima, 2016).

PCA is derived using linear algebra. We want to find variables that are linear functions of the original ones that maximize variance and are uncorrelated, so that the first components represent most of the variation in the original data (Jolliffe, 2002; Jolliffe & Cadima, 2016). This is similar to neural networks that use bottlenecks – in that we want to find a low-dimensional representation of the data – but we can also use PCA on the latent space of a neural network to extract its most important (i.e. principal) components.

PCA is computed by transforming space through the linear transformation of a zero-centered vector based on its covariances (the relationship between vector measurements in how they vary from the mean). By multiplying the vector measurements by their covariance, we find eigenvectors

that are perpendicular to the other vectors and only scale in length from the origin as they are multiplied. The idea of the eigenvector is that it fits a line that best represents the data at hand. The amount that the eigenvectors are scaled by the matrix tells us which of the vectors best describes the data, giving us the 1st, 2nd … *n*th principal component (PC). We take, e.g., the first two eigenvectors and multiply their transpose by the transpose of the original zero-centered vector. This transforms the original data from its dimensions into two dimensions in terms of the eigenvectors (Smith, 2002). Thus, we now space where *x* and *y* axes are based on PC1 and PC2 instead of a Cartesian plane. We hope that the first few PCs explain as much as possible of the variation of the data, so we are able to use just two or three of them and visualize how the data points (individual utterances) plot on a plane.

In this study, after the synthesizer was trained with all of the 12910 utterances, and 128 dimensional latent space vectors were extracted from the reference encoder, I used the PCA module from the Python Scikit-Learn (Pedregosa et al., 2011) library to derive the most salient features of the latent space. As the styles are diverse and multiple unknown variables are expected to affect them, I wanted to first establish that the reference encoder is able to model stylistic variation in general. This was done by extracting reference encoder vectors from the *reference set* that consists of the 320 controlled sentences and 5 speaking styles. After analyzing the *reference set* I proceeded to analyze the main data set.

In the results section, I will use the PCs to analyze and visualize differences between the text styles. Using only the first three PCs, the utterances can be visualized both on a one dimensional histogram and as data points in a three dimensional scatter plot. Then, I can both use the original text styles as color coding as well as categories I have derived by the level of formality and spontaneity.

However, the PCs themselves do not tell us anything about the acoustic correlates of the latent space. To connect these two domains, I needed to analyze the utterances with traditional acoustic measurements as well.

4.2.4 Acoustic Analysis

I extracted f_0 *mean*, f_0 *SD*, f_0 *range*, f_0 *minimum*, f_0 *maximum*, f_0 *slope*, *spectral tilt*, *onset time*, *articulation rate*, and *pauses-to-speech-ratio* from the 12910 utterances. These acoustic analyses can then be correlated with PC values to find out which PC is connected to which acoustic phenomenon. Onset time – when speaking starts in the audio file – has nothing to do with speaking

style, but it was extracted because the reference encoder will most likely model it, and thus it will affect the latent space.

I took the f_0 measurements and spectral tilt by calling Praat (Boersma & Weenick, 2022) with Python's Parselmouth (Jadoul, Thompson & de Boer, 2018) library:

- f_0 *mean,* f_0 *SD,* f_0 *slope* (the mean of absolute change in semitones/second), f_0 *range,* f_0 *minimum,* and f_0 *maximum* were extracted in semitones with the pitch floor at 75 Hz, ceiling at 400 Hz and a time step of 0.1.
- *Spectral tilt* was computed by analyzing the power spectrum to Long-Term Average Spectrum with a bandwidth of 100 Hz and computing its slope with the low band between 0 and 1 kHz and the high band between 1 and 4 kHz, using energy as the averaging method.

Temporal features were extracted by first running Aalto University's automatic forced aligner for Finnish (Leinonen, Virpoja & Kurimo, 2021) to obtain time information for words within the utterances:

- *Onset* was calculated by extracting time when the first transcribed word of the utterance begins.
- *Articulation rate* was calculated by taking the duration of the spoken sections and dividing it by the orthographic length of the utterance – which works well for Finnish with its highly phonetic orthography. Thus, higher the value faster the articulation rate, disregarding pauses between words.
- *Pauses-to-speech-ratio* was calculated by dividing the duration of pauses between words by the duration of the spoken sections. Higher the value, the more pause time there is in the utterance.

Finally, I ran a cross-correlation between all the measurements in R (R Core Team, 2022). f_0 maximum and range correlated almost one-to-one (~0.97), so I omitted the f_0 maximum variable from the study. All the other variables correlated less than 0.65 with each other.

4.2.5 Multiple Linear Regression

Linear regression is a statistical model that describes relationships between observations by assuming a linear relationship between them and finds the best possible straight line to represent it.

It predicts the response of the explained (dependent) variable to the value of the explanatory (independent) variables (Ross, 2017).

The equation for multiple linear regression is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

where Y represents the expected value of the dependent variable, β_0 is Y when the independent variables are zero, i.e. the intercept, X_1 to X_k are the independent variables, and β_1 to β_k are the estimated coefficients – that is, how many steps does Y move when an independent variable moves (LaMorte, 2016). While the coefficient gives us the strength of the relationship, p-value gives us its statistical significance; telling us the probability of the null hypothesis being true. The null hypothesis is the status quo that expects there to be no linear relationship between a dependent variable and independent variables. The alternative hypothesis, then, is that there is a linear relationship between them. A p-value of 0.05 means there is a 5% chance that we would get the result if the null hypothesis were true. Thus, if we are satisfied with the significance level of 0.05, we can reject the null hypothesis if the p-value is <0.05. Common levels of significance used are 0.1, 0.05, and 0.01 (Ross, 2017).

4.2.6 Speech Synthesis

The material used in this study is quite different from the ones cited in the background section. The text styles are diverse, and features modeled by the reference encoder are probably more nuanced than lab speech vs. natural speech or a continuum based on the level of control or ecological validity. For speech synthesis, these distinctions are not necessarily even a fruitful approach, but instead the appropriateness of style for a given context is important.

To evaluate how the synthesizer can emulate speaking styles, I conducted both an objective evaluation where I analyzed the acoustic features of synthesized sentences and a subjective evaluation that focused on the appropriateness of synthesized stimuli given the content of the synthesized sentences. I chose two styles that would serve as opposites on a read/formal vs. spontaneous/informal axis. In this case read/formal means read factual texts in “kirjakieli”, the standard variety of Finnish used in written texts, and spontaneous/informal means a casual style written with a conversational voice and non-standard orthography.

On the read/formal end I derived the style from Wikipedia texts on geography, and on the informal/spontaneous end I chose Youtube captions on social media. There would have been pure

lab speech available as well, but for practical speech synthesis use, reading Wikipedia text is much more valid than sentences uttered without any context. For these two styles, the field of discourse, mode of discourse and the tenor of discourse are all different: one consists of factual texts read to no one while the other – though scripted and acted – is based on a spontaneous text style on an informal subject in a casual interview situation.

I computed reference encoder vectors for the styles by taking the vector measurements from all of the utterances of both text styles and calculated their means. I now had two 128 dimensional embedding vectors, one for each style. To get a “mean style” embedding, i.e. a style midway between the two styles, I calculated the mean of these two vectors. I also wanted to synthesize extreme points of the two styles. Instead of extrapolating linearly from the two vectors, I computed extremes from the two style vectors separately, multiplying both style vectors by 1.3 (anything higher would create unwanted features, like mumbling at the end of the utterance). This means the extremity is in regards to the origin (when all the reference encoder dimensions would be set at zero) instead of the other style. This resulted in 5 embedding vectors corresponding to extra-formal, formal, mean style, spontaneous, and extra-spontaneous speaking styles.

For the stimuli, I came up with 9 sentences with varying text styles that the synthesizer had never seen: 3 in a formal style, 3 in a neutral style, and 3 in a spontaneous style. I then synthesized all of these sentences with all of the 5 speaking styles resulting in 45 synthesized stimuli. The synthesizer is non-deterministic, meaning that every time it synthesizes a sentence – even with the same reference encoder vector – the iteration will be different. Thus, by chance, one sentence synthesized in a specific style might sound better than another. To control this, I synthesized five iterations of all the 45 stimuli, meaning that I would have 5 versions of the same sentence in the same style. I then randomized them so a participant could get different iterations of the same stimuli during the subjective test.

The sentences used for synthesis were as follows, English translations in brackets:

Spontaneous:

- Me oltiin tavallaan sit niinku työkavereita, ja hän kerto myöski jostain taustastaan ja silleen (We were then kinda like colleagues, and she told me also some stuff about her background and what not).
- Joo – siis joku salaattihan olis ihana – siis rakentaa kunnon niinku ruokaisa salaatti. Mut en mä jaksa. (Yeah – like some salad would be wonderful – like to build like a proper nutritious salad. But I can’t be bothered.)

- Mä en nyt enää muista sen yhen naisen nimee, joka muutaman kerran kävi siinä. Se on niinku kirjottanu jonku kirjanki. (I can’t remember that one woman’s name right now, who visited a few times. She’s like written some book too.)

Neutral:

- Arvoisat kolleegat - olen seurannut huolella viimeaikaisia keskusteluja. (Dear colleagues, I’ve followed recent conversations with concern.)
- Raastepöydästä voi aloittaa, ja ruokajuomiin kuuluvat: vesi, maito ja kotikalja. (You can start with the grated root vegetable buffet, and beverages include: water, milk and kvass)
- Mökille pääsee omalla autolla - tai taksilla Ivalon lentoasemalta. (The cabin can be reacher by car – or with a taxi from the Ivalo airport.)

Formal:

- Tunturipöllön latinankielinen nimi on Bubo scandiacus. (Snowy owl’s Latin name is Bubo Scandiacus.)
- Antilooppi on nimitys, jota käytetään monista erilaisista onttosarvisten heimoon kuuluvista sorkkaeläinlajeista. (The term antilope is used to refer to many species of even-toed bovidae.)
- Palvelusektori tuottaa seitsemänkymmentä yksi prosenttia bruttokansantuotteesta, ja työllistää seitsemänkymmentä prosenttia työvoimasta. (The service sector produces seventy one percent of the gross domestic product, and employs seventy one percent of the workforce.)

The synthesized stimuli can be listened at: https://tuukkaot.github.io/styleTTSdemo

4.2.7 Objective Evaluation of Synthesis

After synthesizing the sentences, I ran the same acoustic measurements on the resulting 225 utterances as I had for the whole corpus. These were f_0 mean, f_0 *SD*, f_0 range, f_0 minimum, f_0 maximum, f_0 slope, spectral tilt, articulation rate, and pauses-to-speech-ratio. I fitted a linear model with speaking style as a numeric value between -1.3 and 1.3 as the dependent variable and the aforementioned acoustic measurements as independent variables. The model would show if there was a linear relationship between the continuum and acoustic variables.

4.2.8 Subjective Evaluation of Synthesis

While the objective evaluation will show us how controlling the synthesizer on the axis affects the acoustic measurements, it does not really tell us if the synthesizer can produce appropriate styles for given texts. Furthermore, the acoustic features involved in human speech are quite complex while traditional acoustic measures are rather simple. One cannot assume all stylistic variation to be grasped with measures such as average pitch, the rate it changes, or how much pause time there is. Thus, to really know how the prosody space models style, we need a subjective evaluation.

The aim of the subjective evaluation was to test the appropriateness of the utterances produced in different styles, as modeled by the latent space. This was done by giving participants pairs of audio stimuli with the same sentence synthesized in two styles and asking them which of the stimuli was more appropriate for the textual content of the sentence. If the participants would deem the synthesized speech in a formal style more appropriate for the formal texts and spontaneous style for the spontaneous texts, it would suggest that the latent space models stylistic variation and is able to produce appropriate styles.

To conduct the evaluation, I created a listening test on the Web Audio Evaluation Tool (Jillings et al., 2015) for which I recruited 10 adult native Finnish speakers as participants. The evaluation consisted of a short preliminary test measuring *mean opinion score* (MOS), while the main evaluation was conducted as a forced choice AB test. In the former, the participants listened to seven versions of the same utterance and scored them from 1 to 5 on quality of the synthesis, 5 being the highest quality. The utterances consisted of the sentence: “Kotiin pitää mennä kaupan kautta” (One must to go home via the shop). The sentence was synthesized with the five speaking style values on the formality/spontaneity axis, while the remaining two versions were actual utterances from the corpus, one in a positive style and one in a neutral style. Again, five iterations were synthesized of all the versions and the participants were given random iterations of the utterance.

Even though the synthesizer does produce near natural speech, testing the quality of synthesis in and of itself was not in the scope of this study; instead the MOS serves to give us an inkling on whether or not synthesizing with the extreme values has a clear effect on the quality that would skew the results of the AB test. To assess quality itself, the MOS test would have needed to consist of a range of sentences with different kinds of phonetic phenomena involved instead of a single sentence.

The AB test consisted of all possible pairs of matching sentences from the 45 synthesized utterances in three different styles (total of 90 pairs) in random order. The samples were randomly selected from the 5 iterations as explained in section 3.2.7, so the participant would not listen to the

exact same utterance every time when, say, the value 1.3 of a sentence was compared to another value. For every pair, the participants were asked not to focus on the quality of the utterances but instead to choose which of the two was more appropriate for the given sentence. After the test, I asked the participants to rate the style of each sentence as formal, neutral or informal.

5 Results

5.1 Results of PCA

As the unexplained variance left for the reference encoder by the speaker encoder and text encoder includes the recording environment and settings, I wanted to make sure they are not a major source of variation on the prosody space. The speaker, recording setting and equipment were the same throughout the recordings. Thus, I needed to take into account the specific times of the recording events. I extracted the dates for all of the recorded utterances and whether they were recorded before or after 12pm and ran a cross-correlation with all of the PCs. The highest correlations with recording date were with PC3 (0.26), PC16 (0.2), and PC9 (0.19), while none of the components had more than a 0.2 correlation with the time of day – PC9 being the highest (0.18). There is some correlation especially with the date but it is not particularly strong, so we can ignore them.

I also wanted to see if there is some relationship between the PCs and the orthographic length of the spoken sentences. On figure 3, the values of PC1 (y-axis) are plotted against the text length (x-axis). I ran a cross correlation between text length and the PCs. PC1 and text length correlate by -0.82. Thus, there is a very strong relationship between them. This may have an effect on the synthesis output. More on that later.

Figure 4: A scatterplot with PC1 and text length, showing a clear negative relationship.

Next, I proceeded to analyze reference encoder outputs for the *reference set* utterances the synthesizer had not been trained with. These were the 320 sentences uttered in five styles for a total of 1613 utterances.

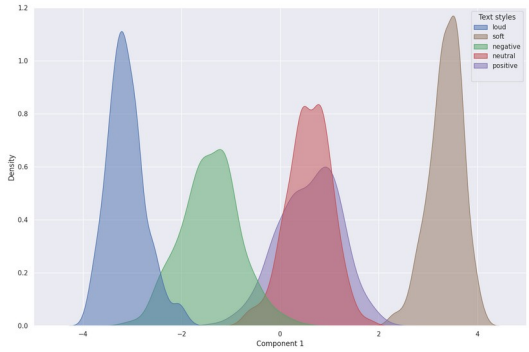

Figure 5: Reference data set. Text styles on PC1. Neutral and positive overlap with each other, while other styles are clearly distinguished.

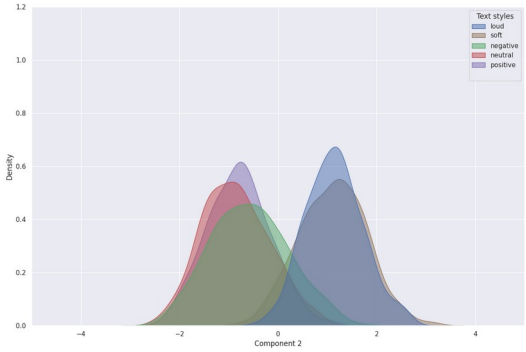

Figure 6: Reference data set. Text styles on PC2 are divided into two groups: loud and soft on one end, and the rest on the other end.

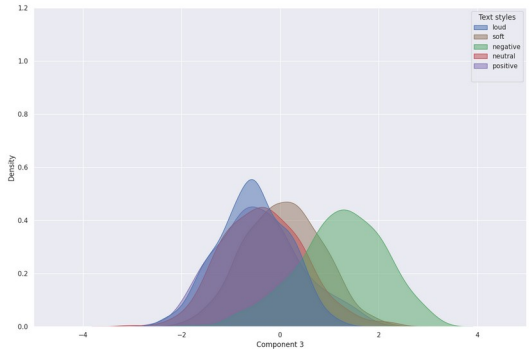

Figure 7: Reference data set. Text styles on PC3 are spread more evenly and the only one that clearly separates is negative.

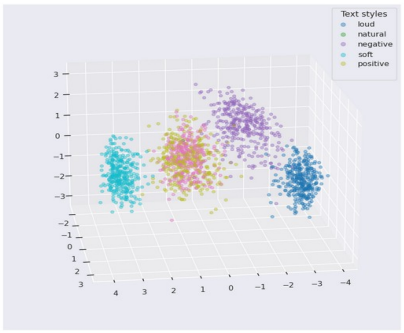

Figure 8: Reference data set. The 3d scatter plot with PC1, PC2, and PC3 shows the individual utterances are mostly in their respective areas, apart from the neutral and positive styles.

Figures 5, 6 and 7 have the principal component on x-axis and kernel density estimate inferred from the data points on y-axis. As we can see, the styles separate on PC1, and there is little internal variation. Neutral and positive almost completely overlap, while all of the other styles are clearly different from one another. On PC2, loud and soft, seemingly opposite styles, overlap while all the others group on the other side. On PC3, negative is on its own while the other styles flock together. The scatter plot in figure 8 shows there are almost no data points mixed in with another group apart from neutral and positive overlapping. The differences between styles become almost nonexistent after the first three PCs, and the are no discernible differences after the first eight PCs.

These utterances were controlled for clear cut, even stereotypical, styles and with the same sentences for all styles similarly to Tits et al. (2019). Next, I analyzed the reference embeddings for the whole data set which the synthesizer was trained with. The scatter plot in figure 9 below shows that the data points do group somewhat, e.g., parliament text in turquoise seems to group as its own

cluster on bottom-left and captions in orange are mainly in bottom-right corner. However, there are a lot of data points outside their group and the styles overlap all over.

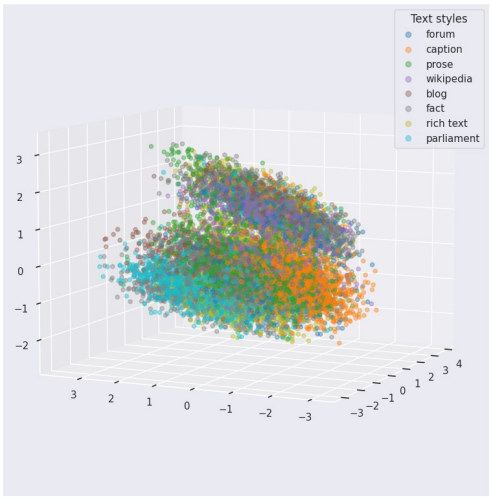

Figure 9: Full data set. The individual utterances overlap all over even though some concentrations are observable.

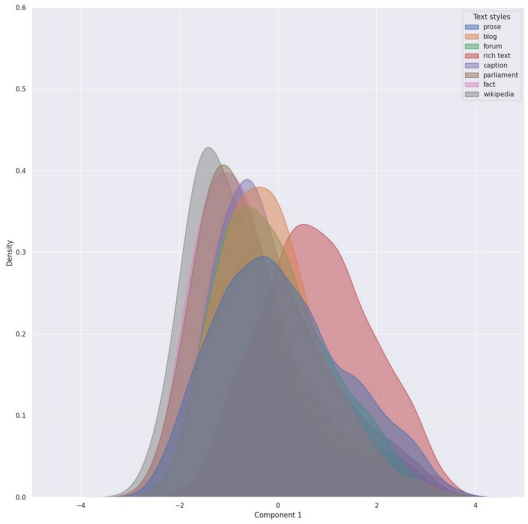

Figure 10: Full data set. The styles overlap and are all spread wide over the PC1 axis. Wikipedia, fact and parliament are on the left while rich text is on the opposite end.

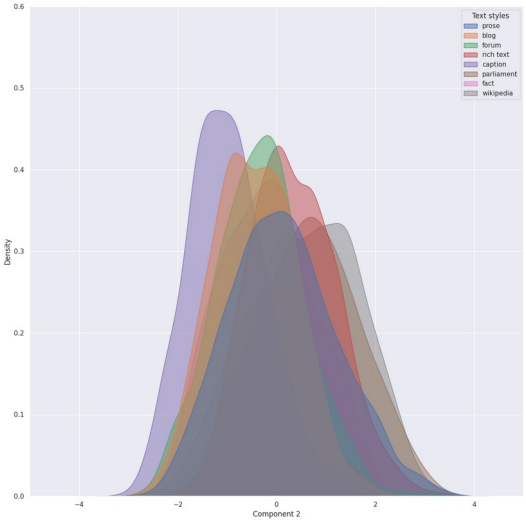

Figure 11: Full data set. The styles differentiate somewhat on PC2, but there is also a lot of overlap between styles. Captions are on the extreme left and Wikipedia and parliament are its opposites. Forum, blog, rich and prose are in between.

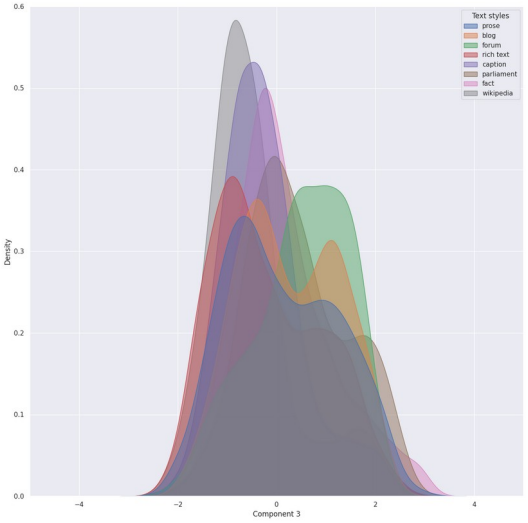

Figure 12: Full data set. On PC3 there are multiple peaks within the styles, showing internal variation. Wikipedia, captions and fact are defined peaks, i.e., there is not much internal variation for them on PC3, while blog has two distinct peaks.

The histograms in figures 10, 11, and 12 above show that the story with the more realistic speaking styles is much more complex. While the styles do separate somewhat on the three first

components, there is a lot of internal variation; all of the text styles have utterances that set on both extremes of the components and everything in between. Figure 12 on PC3 also shows interesting internal variation, e.g., blog style has two clear peaks.

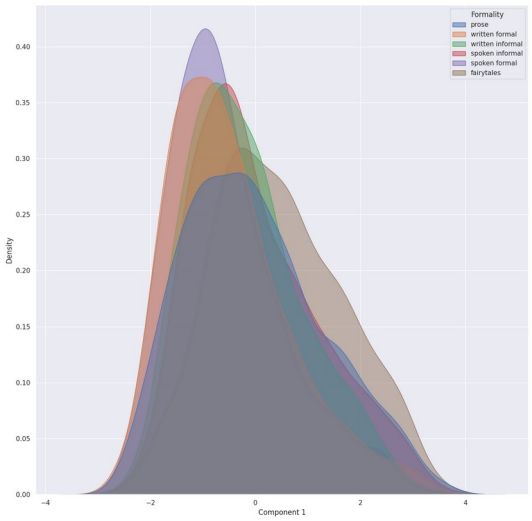

Figure 13: Full data set. The styles overlap on PC1. Spoken formal is centered the most and fairy tales are slightly to the right from the other styles.

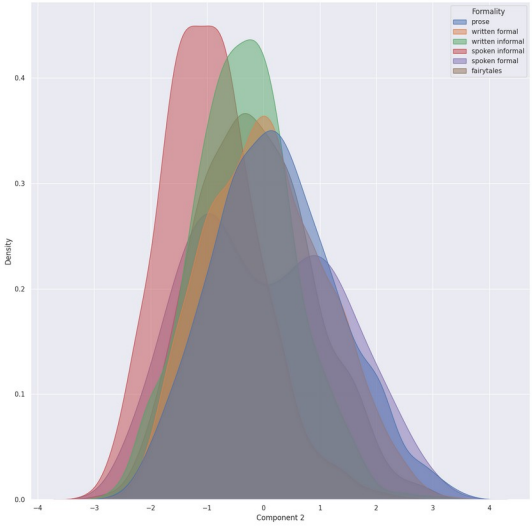

Figure 14: Full data set. Written informal and spoken informal congregate the most on PC2. Spoken informal is the leftmost peak while the others are positioned more or less toward the right from it. Spoken formal has two peaks: one on the left and one one the right.

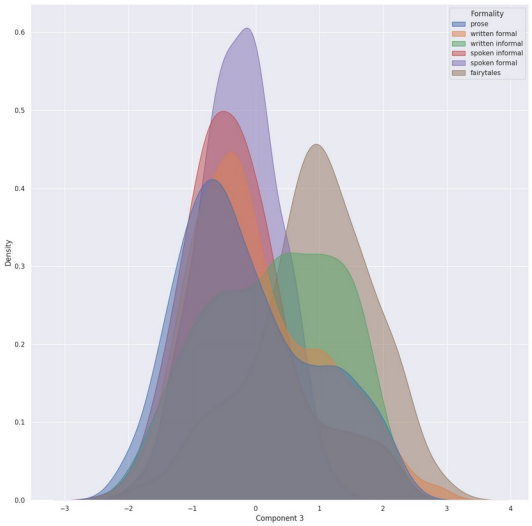

Figure 15: Full data set. Spoken formal is the most concentrated speaking style on PC3. Prose, spoken informal, and written formal are mostly on the left from it. Fairy tales peak on the opposite side with some written informal texts.

Thus, it seems the styles have more nuanced sub-styles within them. To make sense of the variation, I proceeded to divide the utterances based on perceived formality and (mimicked) medium of transmission instead of the text styles.

Figures 13, 14, and 15 show the specific sub-styles, e.g., “YouTube captions on social media” and “Wikipedia texts on geography”, divided by their formality and medium of transmission. Prose and children’s fairy tales are in their own groups as there will be different styles within them that cannot be separated, e.g., narration and dialogue.

Looking at figures 13, 14, and 15, the styles do not separate almost at all on PC1 and PC2, except that spoken informal and spoken formal are on the opposite ends on PC2. On PC3, “fairy tales” and “written informal” group together somewhat, although fairy tales are more clearly positioned on the PC. Spoken formal is the clearest peak on both PC1 and PC2 showing the least internal variation.

While in the reference data set, the first three PCs explain 40.97% (PC1 explains 27.69%) of the variation, in the main data set, they only explain 21.27% (PC1 explains 8.86%). It is expected, then, that there are PCs later on that account for important features in the corpus. Furthermore, as the PCs are arranged in descending order based on how much variance they explain, we must remember that the latent space does not model human articulation but features of the corpus. As such, if there are features that only exhibit themselves in a small number of utterances, while they might be important features for human verbal communication, the PC that explains them might not be among the first ones.

Component	F0 mean	F0 STD	F0 minimum	F0 range	F0 slope	tilt	pause ratio	articulation rate	onset
1	0.0442	0.0399	0.3110	0.4854	0.2370	0.0694	0.4360	0.1630	0.3382
2	0.0125	0.0723	0.0063	0.0407	0.0336	0.0733	0.1762	0.1292	0.5591
3	0.0704	0.0211	0.0295	0.0546	0.0686	0.0177	0.2447	0.0565	0.1854
4	0.0343	0.0389	0.0719	0.0831	0.0399	0.0196	0.3933	0.2125	0.4507
5	0.1705	0.0359	0.1980	0.1292	0.0306	0.0130	0.1091	0.2342	0.2478
6	0.3323	0.2691	0.0853	0.1867	0.1106	0.0598	0.0800	0.0996	0.1090
7	0.4603	0.1734	0.1660	0.1554	0.2240	0.0147	0.0531	0.1407	0.0518
8	0.1371	0.2268	0.0006	0.1243	0.1640	0.0461	0.1475	0.0516	0.1284
9	0.0142	0.0230	0.0783	0.0087	0.0432	0.0722	0.0469	0.0369	0.0115
10	0.3298	0.0865	0.1369	0.0307	0.1527	0.0208	0.0660	0.1380	0.0486
11	0.0141	0.0706	0.0669	0.0649	0.0253	0.0236	0.2282	0.0890	0.0071
12	0.1452	0.1636	0.0024	0.0757	0.1324	0.0071	0.1983	0.0518	0.0045
13	0.0859	0.1281	0.1478	0.0818	0.0070	0.0203	0.0169	0.0650	0.0733
14	0.0203	0.0560	0.0188	0.0288	0.0537	0.0478	0.1651	0.0492	0.0252
15	0.0346	0.0020	0.0354	0.0232	0.0072	0.0842	0.0262	0.0282	0.0328
16	0.1096	0.1020	0.0150	0.0714	0.0614	0.2652	0.0879	0.0324	0.0135
17	0.1617	0.1731	0.0327	0.1004	0.0961	0.0627	0.0346	0.0021	0.0174
18	0.2148	0.4383	0.0572	0.2366	0.2008	0.1088	0.0546	0.0865	0.0060
19	0.0294	0.1905	0.0191	0.1035	0.0738	0.0233	0.0045	0.0048	0.0416
20	0.0765	0.0861	0.0715	0.0613	0.0353	0.0268	0.0332	0.0822	0.0064
21	0.1571	0.1424	0.0687	0.0581	0.0899	0.0687	0.0140	0.0218	0.0015
22	0.2270	0.1085	0.0159	0.0605	0.0694	0.0331	0.0152	0.0596	0.0374
23	0.0622	0.0130	0.0200	0.0285	0.0033	0.1021	0.0115	0.0981	0.0279

Table 1: Cross-correlation of acoustic features and PCs; color-coded by correlation with the top three correlates for every acoustic feature emphasized with borders.

To take this into account, I extracted f_0 mean, f_0 *SD*, f_0 minimum, f_0 range, f_0 slope, spectral tilt, articulation rate, pauses-to-speech-ratio, and onset time from the all of the 12910 utterances. I then cross-correlated the acoustic features and the 128 PCs to determine which PCs correlate the most with which acoustic features. The first 23 PCs explain over 70% of the variation (71.31%) and the

top three correlates for every acoustic feature are within them. Thus, they should give us a sufficient picture into how the PCs disentangle acoustic features. Table 1 above shows the first 23 PCs color coded by how much they correlate with a given acoustic feature and top three correlations for a given acoustic feature are emphasized with borders.

Looking at the top three highest correlating PCs for every feature, PC1 has an effect on the f_0 minimum, range, and slope, as well as on articulation rate, pauses-to-speech-ratio, and onset time. PC2 is mostly responsible for onset and disentangled from other features. Having onset time on one of the first PCs is expected as the synthesizer needs to know when to start and stop generating sound for every utterance. PC 3 is quite disentangled as well, and it seems to have a connection to pauses in the utterances. PCs 6, 7, 8 and 10 are connected to f_0 but not temporal features. Spectral tilt, which corresponds to voice quality, is somewhat disentangled. Its main linear relationships are with PCs 16, 18 and 23. PC 18 it shares with f_0 *SD*, f_0 range and f_0 slope. The three f_0 measurements can be expected to overlap with each other as all have to do with the amount or rate of change in f_0.

5.2 Results of Objective Evaluation of Speech Synthesis

I fitted a linear model with the 225 synthesized utterances with value – between -1.3 and 1.3 on a formal/read vs. informal/spontaneous axis – as the dependent variable, and the acoustic measurements (except onset time) as independent variables without interactions. Table, 2, below shows the linear relationships.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.563001	1.184615	-0.475	0.6351	
F0 mean	0.090989	0.105541	0.862	0.3896	
F0 SD	-0.047675	0.156434	-0.305	0.7608	
F0 minimum	-0.03457	0.032934	-1.05	0.2951	
F0 range	-0.002172	0.028821	-0.075	0.94	
F0 slope	-0.04135	0.01737	-2.38	0.0182	*
Spectral tilt	0.107736	0.026481	4.069	6.65E-05	***
Articulation rate	0.272871	0.033077	8.25	1.61E-14	***
Pauses-to-speech	-3.496852	1.483406	-2.357	0.0193	*

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' ' 1

Table 2: Results of the multiple linear model with the formal – spontaneous axis as the dependent variable and acoustic values extracted from the synthesizer utterances as independent variables.

In this data set, f_0 slope, spectral tilt, articulation rate and pause-to-speech ratio all have a statistically significant linear relationships with the formality axis with a p-value below 0.05.

To compare this with the original utterances from the two sets, I fitted a binomial regression model with the Wikipedia utterances and social media captions – total of 534 utterances – as values 0 and 1, respectively, and acoustic measurements of the utterances as independent variables.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-5.155672	1.356608	-3.8	0.000144	***
F0 mean	0.714672	0.098556	7.251	4.12E-13	***
F0 SD	0.545971	0.157361	3.47	0.000521	***
F0 minimum	0.034921	0.06263	0.558	0.577131	
F0 range	-0.116227	0.040259	-2.887	0.003889	**
F0 slope	-0.001787	0.0131	-0.136	0.8915	
Spectral tilt	0.170286	0.048136	3.538	0.000404	***
Articulation rate	0.189866	0.052465	3.619	0.000296	***
Pauses-to-speech	0.120751	2.101178	0.057	0.954172	

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' ' 1

Table 3: Results of the binomial logistic regression model with "Wikipedia texts on geography" (0) and "YouTube captions on social media" (1) as the binary dependent variable and acoustic values extracted from the corresponding utterances as independent variables.

Table 3 shows statistically significant relationships with f_0 mean, f_0 *SD*, f_0 range, spectral tilt, and articulation rate. The statistical relationships in the synthesized data set that do not show in the original utterances are f_0 slope and pauses-to-speech ratio. Also articulation rate has a stronger relationship with style in the synthesized utterances than in the original data. f_0 slope is probably a by product of the faster articulation rate. As the synthesized style is more spontaneous, articulation rate rises, and thus the same changes in f_0 are bound to happen on a steeper slope.

On the other side, the significant relationships in the original utterances that are not there on the synthesized data set, are f_0 mean, f_0 *SD*, and f_0 range. It seems from the acoustic features that the reference encoder has modeled voice quality well but less so with the temporal and f_0 features. One explanation could be that I derived the reference embeddings for the extreme styles by moving away from the origin instead of the other style. To make sure this was not the case, I fitted the same linear model for the synthesized sentences but this time omitting the -1.3 and 1.3 values and merely keeping the two styles and a mean style. The results followed suit with the model in table 2. Next, the subjective evaluation tells us how these styles are judged by human listeners.

5.3 Results of Subjective Evaluation of Speech Synthesis

In the first part of the listening test, the participants were asked to assess the quality of an utterance on a scale from 1 to 5, and the means of their answers were computed into a mean opinion score. Table 4 shows that, as expected, the hidden references (real recordings) fare better than the

synthesized versions, and that there is no apparent link between the extremity of the reference embedding values and quality of the synthesis, i.e., the synthesis does not seem to break even on the -1.3 and 1.3 values. However, a proper MOS test with multiple sentences would be needed to verify this.

Stimulus	MOS
true neutral	4.3
true positive	3.5
-1.3	2.8
-1	2.7
0	2.6
1	2.35
1.3	2.6

Table 4: MOS scores of seven versions of the utterance: "Kotiin pitää mennä kaupan kautta."

The second part of the subjective evaluation consisted of a forced choice AB test. One of the participants notified me after the test that they had not understood the instructions correctly and had evaluated something else than what was asked. Thus, the results from participant number 6 were omitted before analysis.

As explained in section 3.2.7, the participants were given 90 pairs of audio stimuli with the same sentence in both stimuli. The participants were asked to choose which stimulus was more appropriate for the given text. After I received the results from the tests, I fitted a linear model with the value of the chosen stimuli – formality/spontaneity between -1.3 and 1.3 – as the dependent variable *choice* and formality of the text style as the independent variable *text style*. The formality was rated according to the participants judgments and set for every stimulus pair as -1, 0 or 1 on the text style variable.

Choice ~ text style	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.08636	0.03419	-2.526	0.0117	*
text style	0.36804	0.04042	9.106	<2E-016	***

Table 5: Summary of the linear model with choice on the read/formal – spontaneous/informal axis as the dependent variable and text style as the independent variable explaining it.

Table 5 shows a highly significant positive relationship between the text style and stimulus choice, i.e., the more informal and spontaneous the text style, the more spontaneous stimulus participants tended to pick – and vice versa. The coefficient of the positive relationship is quite strong, ~0.35. However, as we see from the intercept, the participants tended to prefer more formal sentences in general in a statistically highly significant manner.

Next, I looked at linear models with the same variables, fitted for each participant separately.

Choice ~ text style	Estimate	Std. Error	t value	Pr(>|t|)	
Participant 1					
(Intercept)	0.13358	0.11307	1.181	0.241	
text style	-0.08005	0.15448	-0.518	0.606	
Participant 2					
(Intercept)	-0.12623	0.09689	-1.303	0.196098	
text style	0.46566	0.11901	3.913	0.000181	***
Participant 3					
(Intercept)	-0.36111	0.09824	-3.676	0.000407	***
text style	0.175	0.12032	1.454	0.149374	
Participant 4					
(Intercept)	-0.27	0.1047	-2.578	0.0116	*
text style	0.2683	0.1283	2.092	0.0393	*
Participant 5					
(Intercept)	0.1644	0.1037	1.585	0.1165	
text style	0.3692	0.1176	3.141	0.0023	**
Participant 6					
(Intercept)	-0.03585	0.09772	-0.367	0.715	
text style	0.50956	0.12001	4.246	5.43E-05	***
Participant 7					
(Intercept)	-0.1575	0.1016	-1.549	0.125	
text style	0.4275	0.1016	4.206	6.24E-05	***
Participant 8					
(Intercept)	0.04045	0.0985	0.411	0.682	
text style	0.51333	0.11996	4.279	4.81E-05	***
Participant 9					
(Intercept)	-0.06855	0.10182	-0.673	0.503	
text style	0.48306	0.11545	4.184	6.76E-05	***

Table 6: Linear models fitted for each participant separately.

Shown on table 6, all participants apart from participant 1 have some statistically significant tendency for choice, either without the explaining variable, with it, or both. Participant 1 shows no statistical significance either way, and the coefficients are in the opposite direction compared to almost all other participants. The intercept (choice without the explaining variable text style) is towards informality, and when text style becomes more informal, choice moves toward formality. However, as there is no statistical significance, we cannot derive anything from participant 1's choice variable.

Two of the participants, 3 and 4, have a general preference that is statistically significant: they both lean toward formality. All participants apart from 1 and 3 have significant relationships between choice and text style, all of them positive.

6 Discussion

The reference encoder models speech styles quite well, however realistic speaking styles are much more complex than the acted stereotypical ones. These styles are hard to delineate and they may have significant internal variation. This variation can take place within, say, one conversation, or within a whole style with internal variation depending on topic and context. This pinpoints the complex character of speaking styles.

PCA does not completely disentangle the different acoustic features in the realistic speaking styles. Especially PC1 explains a variety of acoustic features. This was somewhat expected as there probably is a diverse set of hidden variables involved, which I attempted to unearth by correlating the PCs with simplistic acoustic measurements. It is also likely that some features cannot be disentangled from the latent space: acoustic and articulatory phenomena are so intertwined that they might be impossible to completely separate as the network will never see such data where they are not related. For example, there are mechanical constraints to the speed of human articulation, and a higher articulation rate may also lead to a steeper f_0 slope even if the style is not any more expressive than another. Faster speech than what the network has been trained with would require it to extrapolate to unseen space, which in the scope of this study is not interesting as it aimed to produce realistic styles – even though it may be interesting for AI researchers.

The objective evaluation yielded some promising results. The acoustic features the reference encoder seemed to grasp were articulation rate and tilt. The increased articulation rate, however, could be due to how the vector was computed instead of the reference encoder modeling articulation rate correctly. There are many short utterances in the corpus for the spontaneous style, e.g., filled pauses or short agreement words such as “joo” (yes). Because PC1 is heavily connected to text length (as established in the results section), there are latent embedding dimensions that correspond to it. I created the styles for synthesis by computing the means of all the embeddings from the two styles. As such, the synthesizer may have tried to squeeze the sentences into the average length of utterances from the spontaneous style data set. Thus, articulation rate increases as the spontaneity of the synthesis increases, even though there might not be such a strong connection to this in the data set itself. On the practical side, this heightened articulation rate also lead to the synthesizer sometimes omitting words and articulating in a less clear manner. If the manner of computing the vectors explains the articulation rate variable, then the only measured acoustic feature that the reference encoder modeled is spectral tilt. As the subjective evaluation showed us that there are discernible differences between the styles, either the acoustic measurements are too

simple for analyzing stylistic variation or spectral tilt itself was enough to deem an utterance more appropriate than another.

In terms of the latent space, another way of calculating the extreme styles would have been to linearly extrapolate them from the two embedding vectors. While the formality and informality axis worked in a practical sense, extrapolation would have created a linear continuum between the styles, instead of having the extreme values in relation to the origin. This would have benefited the objective evaluation as the statistical method fits a linear model, but it might have been detrimental for the subjective evaluation. Computing the extreme styles as stronger versions of the styles themselves will give us more formal or more spontaneous compared to a neutral style, while linear extrapolation would have given us a more opposite style in terms of the other style.

The subjective evaluation showed that the participants preferred a more formal speaking style for the formal sentences and a spontaneous speaking style for the spontaneous sentences. This suggests that the reference encoder models the styles at least somewhat and the synthesizer can be controlled on a read/formal-spontaneous/informal axis. In general, the participants tended to prefer more formal speaking styles. Reasons behind this may be numerous, from expectations on synthesized speech and quality of the synthesis to social attitudes towards women speaking in an informal voice. Anecdotally, the extremely informal voice often included distinct creaky voice and a rising tone at the end of some utterances, which at least in the anglophone context are seen as gendered female traits and often frowned upon (Habasque, 2021). Creaky voice is connected to spectral tilt, so spectral tilt may be enough to affect the evaluation of appropriateness. Also, the rising tone would not necessarily show in the acoustic measurements used in this thesis.

7 Conclusions

There are some obvious limitations to analyzing and synthesizing speaking styles with this method. Phenomena that are rare in the corpus may still be significant in regards to human communication, as the neural architecture does not model articulation but a specific corpus with its limitations. Furthermore, controlling all of the important variables present may be difficult and the network may model unintended features. To remedy this, one could analyze and disentangle unwanted variation from other features. For example, for this method, text length related embedding dimensions would need to be extracted and adjusted according to the length of the given sentence instead of using an average value from the style. In general, it is important to account for as many unwanted variables as possible, from recording and pre-processing to when we analyze the latent

space. In the future, it would be interesting to use more diverse data such as actual spontaneous speech or a multi-speaker corpus. Using a multi-speaker corpus with people of different genders, ages, and dialects could enable highly adaptable speech style synthesis with different dialects and even more realistic styles. Then, the method could also be more useful for phonetics research, as the data would be more varied and the prosody space more diverse. However, as the network is a black box, there are an unknown number of hidden variables that can affect the output, and getting reliable and analyzable results from the reference encoder may turn out to be a difficult task.

Outside of practical applications, style controlled speech synthesis could be interesting for sociolinguistics research, for example, regarding language attitudes. It is difficult to control styles with real human speakers if the text style is discongruent with the speaking style. With synthesized speech, one could control these aspects when producing stimuli for listening tests.

Based on this study, it would be interesting to continue researching speech synthesis from the point of view of appropriateness. Appropriateness is a socially loaded question and as such may signify different things for different people. Also, while the stylistic differences were somewhat hazy in the objective evaluation, they were clear in the subjective evaluation. It would be interesting to go deeper into what are the specific acoustic correlates for appropriateness for a given text. As stated in the background section, we used this same model for disentangling and controlling phonetic features for synthesis. A similar subjective evaluation could be done where instead of interpolating between styles, we could adjust specific acoustic values, such as spectral tilt or f_0 mean, and analyze if there are specific phonetic features that affect the judgment of appropriateness. This could also have a practical relevance as we look into what types of synthesized speech will become mainstream.

References

Aaltonen, O., Aulanko, R., Iivonen, A., Kilpi, A., Vainio, M. (Eds.) (2009). *Puhuva ihminen.* Otava.

Audacity Team (2021). *Audacity(R): Free Audio Editor and Recorder* [Computer application]. Version 3.0.4 from https://audacityteam.org/

Barsties, B. De Bodt, M. (2015). *Assessment of voice quality: Current state-of-the-art.* Auris Nasus Larynx, Volume 42, Issue 3. https://doi.org/10.1016/j.anl.2014.11.001.

Benesty, J., Sondhi M. M., Huang, Y. (Eds.) (2008). *Springer Handbook of Speech Processing.* Springer-Verlag.

Berg, R. E. (n.d.). *Sound*. Encyclopedia Britannica. https://www.britannica.com/science/sound-physics

Boersma, P. & Weenink, D. (2022). *Praat: doing phonetics by computer* [Computer program]. Available at http://www.praat.org/

Delgado-Martins, M.R., Freitas, M.J. (1991) *Temporal structures of speech: "reading news on TV."* Proc. ESCA Workshop on Phonetics and Phonology of Speaking Styles, paper 019

Dellwo, V., Leemann, A., & Kolly, M.-J. (2015). *The recognition of read and spontaneous speech in local vernacular: The case of Zurich German dialect.* Journal of Phonetics, 48, 13–28 10.1016/j.wocn.2014.10.011

Drolet, M., Schubotz, R., & Fischer, J. (2011). Authenticity affects the recognition of emotions in speech: Behavioral and fMRI evidence. Cognitive, Affective & Behavioral Neuroscience, 12, 140–150.

Gick, B., Wilson, I., Derrick, D. (2013). *Articulatory Phonetics.* Wiley-Blackwell.

Gregory, M. & Carroll, S. (2019). *Language Situation. Language Varieties and Their Social Contexts* (# 3rd. ed.). Routledge.

Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning.* MIT Press.

Gordon, M. and Ladefoged, P. (2001). *Phonation types: a cross-linguistic overview.* Journal of Phonetics, 29(4):383–406. https://doi.org/10.1006/jpho.2001.0147

Habasque, P. (2021). *Is Creaky Voice a Valley Girl Feature? Stancetaking & Evolution of a Linguistic Stereotype*. Anglophonia. French Journal of English Linguistics. https://doi.org/10.4000/anglophonia.4104

Hiovain, K., Suni, A., Kakouros, S., & Šimko, J. (2020). *Comparative analysis of majority language influence on North Sámi prosody using WaveNet-based modeling.* Language and Speech, [0023830920983591]. https://doi.org/10.1177/0023830920983591

Jadoul, Y., Thompson, B., & de Boer, B. (2018). *Introducing Parselmouth: A Python interface to Praat*. Journal of Phonetics, 71, 1-15. https://doi.org/10.1016/j.wocn.2018.07.001

Jillings, N., De Man, B., Moffat, D. & Reiss, J.D. (2015). *Web Audio Evaluation Tool: A Browser-Based Listening Test Environment*. 12th Sound and Music Computing Conference.

Joos, M. (1968). The Isolation of Styles. In J.A. Fishman (Ed.), *Readings in the Sociology of Language* (pp. 185-191). De Gruyter Mouton.

Jolliffe I. T. (2002). *Principal Component Analysis.* Springer.

Jolliffe I. T. and Cadima Jorge. (2016). *Principal component analysis: a review and recent developments.* Phil. Trans. R. Soc. A. 374: 20150202. http://doi.org/10.1098/rsta.2015.0202

Jurafsky, D., Martin, J. H. (2020). *Speech and Language Processing: An Introduction to Natural Language Processing,Computational Linguistics, and Speech Recognition.* https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Jürgens R, Hammerschmidt K and Fischer J (2011) *Authentic and play-acted vocal emotion expressions reveal acoustic differences.* Front. Psychology **2**:180. doi: 10.3389/fpsyg.2011.00180

Klimkov, V., Ronanki, S., Rohnke, J., & Drugman, T. (2019). *Fine-grained robust prosody transfer for single-speaker neural text-to-speech. ArXiv, abs/1907.02479.*
Koopmans-van Beinum, Florien J. (1991): *Spectro-temporal reduction and expansion in spontaneous speech and read text: focus words versus non-focus words*, In PPoSpSt-1991, paper 036.
Kröse, Ben & Krose, B. & van der Smagt, Patrick & Smagt, Patrick. (1993). *An introduction to neural networks*. J Comput Sci. 48.
Laan, G. (1997). *The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style*. Speech Communication, 22,
43–65.
Labov, W. (1972). *Sociolinguistic Patterns*. Basil Blackwell.
Leinonen, J., Virpioja, S. and Kurimo, M. (2021). *Grapheme-Based Cross-Language Forced Alignment: Results with Uralic Languages*. NoDaLiDa.
Lieberman, P., & Blumstein, S. (1988). *Speech Physiology, Speech Perception, and Acoustic Phonetics* (Cambridge Studies in Speech Science and Communication). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139165952
LaMorte, W. (2016/5/31). *The Multiple Linear Regression Equation.* https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704-ep713_multivariablemethods/bs704-ep713_multivariablemethods2.html
LLISTERRI, J. (1992) *Speaking styles in speech research*, ELSNET/ESCA/SALT Workshop on Integrating Speech and Natural Language. Dublin, Ireland, 15-17 July 1992.
Mendoza, E. & Valencia, N. & Muñoz López, J. & Trujillo Mendoza, H. (1996). *Differences in Voice Quality Between Men and Women: Use of the Long-Term Average Spectrum (LTAS)*. Journal of voice : official journal of the Voice Foundation. 10. 59-66. DOI: 10.1016/S0892-1997(96)80019-1.
Mixdorff, H. & Pfitzinger, H.R. (2005). *Analysing fundamental frequency contours and local speech rate in map task dialogs. Speech Commun., 46, 310-325.*
Mohan, D.S.R, Hu, V., Teh, T.H., Torresquintero, A., Wallis, C.G.R., Staib, M., Foglianti, L., Gao, J. and King, S. (2021). *Ctrl-P: Temporal Control of Prosodic Variation for Speech Synthesis.* ArXiv. DOI: 10.48550/ARXIV.2106.08352
O'Shea, K., & Nash, R. (2015). *An Introduction to Convolutional Neural Networks*. ArXiv, abs/1511.08458.
Pedregosa et al. (2011). *Scikit-learn: Machine Learning in Python.* JMLR 12, pp. 2825-2830.
Raitio, T., Rasipuram, R., & Castellani, D. (2020). *Controllable neural text-to-speech synthesis using intuitive prosodic features. INTERSPEECH.*
R Core Team (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/.
Ross, S.M. (2018). *Introductory Statistics* (4th edition). Academic Press.
Rossing, T. D., Dunn, F., Hartmann, W. M., Murray Campell, D., Fletcher, N. H. (Eds.) (2007). *Springer Handbook of Acoustics.* Springer-Verlag New York Inc.
Sivaprasad, S., Kosgi, S., Gandhi, V. (2021). *Emotional Prosody Control for Speech Generation.* Interspeech 2021. DOI: 10.21437/interspeech.2021-307
Shechtman, S., Sorin, A. (2019) *Sequence to Sequence Neural Speech Synthesis with Prosody Modification Capabilities*. Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10), 275-280, doi: 10.21437/SSW.2019-49
Shechtman, S., Sorin, A. (2019) *Sequence to Sequence Neural Speech Synthesis with Prosody Modification Capabilities*. Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10), 275-280, doi: 10.21437/SSW.2019-49
Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y.,

Skerry-Ryan, R.J., Battenberg, E., and Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R.J., Clark, R. and Saurous, R.A. (2018) *Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron.* DOI: 10.48550/ARXIV.1803.09047

Skerry-Ryan, R.J., Saurous, R.A., Agiomyrgiannakis, Y., & Wu, Y. (2018). *Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions.* 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4779-4783.

Shlens, J. (2014). *A Tutorial on Principal Component Analysis.* ArXiv, abs/1404.1100.

Šimko, J., Törö, T., Vainio, M., & Suni, A (2022). *Prosody under control: A method for controlling prosodic characteristics in text-to-speech synthesis by adjustments in latent reference space* [submitted for publication].

Šimko, J., Vainio, M., & Suni, A. (2020). *Analysis of speech prosody using WaveNet embeddings: The Lombard effect.* In Proceedings of 10th International Conference on Speech Prosody 2020, Tokyo, Japan (pp. 910-914). (Speech prosody). ISCA. *https://doi.org/10.21437/SpeechProsody*.

Smith, L. I. (2002). *A Tutorial on Principal Component Analysis.* (Computer Science Technical Report No. OUCS-2002-12). Retrieved from http://hdl.handle.net/10523/7534

Spinner, T., Körner, J., Gortler, J., & Deussen, O. (2018). *Towards an Interpretable Latent Space – An Intuitive Comparison of Autoencoders with Variational Autoencoders*

Stevens, K. N. (1998). *Acoustic Phonetics*. The MIT Press.

Styler, W. (2022). *Using Praat for Linguistic Research.* Version 1.9.1. Retrieved from: *https://wstyler.ucsd.edu/praat/UsingPraatforLinguisticResearchLatest.pdf*

Sundberg J, Fahlstedt E, Morell A. *Effects on the glottal voice source of vocal loudness variation in untrained female and male voices.* J Acoust Soc Am. 2005 Feb;117(2):879-85. Doi: 10.1121/1.1841612.

Suni, A., Wlodarczak, M., Vainio, M., & Simko, J. (2019). *Comparative analysis of prosodic characteristics using WaveNet embeddings.* In G. Kubin, T. Hain, B. Schuller, D. El Zarka, & P. Hodl (Eds.), 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019) : Crossroads of Speech and Language (pp. 2538-2542). (Interspeech). ISCA. https://doi.org/10.21437/Interspeech.2019-2373

Tits, N.& Wang, F. & El Haddad, K. & Pagel, V. & Dutoit, T. (2019). *Visualization and Interpretation of Latent Spaces for Controlling Expressive Speech Synthesis Through Audio Analysis.* 4475-4479. 10.21437/Interspeech.2019-1426.

Toivola, M., & Lennes, M. (2012). Speech rate. In P. Robinson (Ed.), *The Routledge Encyclopedia of Second Language Acquisition* Routledge.

Vainio, Martti & Järvikivi, Juhani. (2007). *Focus in production: Tonal shape, intensity and word order*. The Journal of the Acoustical Society of America. 121. EL55-61. 10.1121/1.2424264.

Vainio, M., Suni, A., Järveläinen, H., Järvikivi, J., & Mattila, V-V. (2005). *Developing a speech intelligibility test based on measuring speech reception thresholds in noise for English and Finnish.* Journal of the Acoustical Society of America, 118, 1742-1750. https://doi.org/10.1121/1.1993129

Valle, R., Li, J., Prenger, R.J., & Catanzaro, B. (2020). *Mellotron: Multispeaker Expressive Voice Synthesis by Conditioning on Rhythm, Pitch and Global Style Tokens*. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6189-6193.

Wagner, P., & Windmann, A. (2015). *Acted and Spontaneous Conversational Prosody — Same or Different?*

Wagner, P., Trouvain, J., Zimmerer, F. (2015). *In defense of stylistic diversity in speech research.* Journal of Phonetics 48: 1-12.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R.J., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., & Saurous, R.A. (2018). *Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis*. ICML.

Zhang, Y., Pan, S., He, L., & Ling, Z. (2019). *Learning Latent Representations for Style Control and Transfer in End-to-end Speech Synthesis.* ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6945-6949.