The Evaluation of the Feminine Level of Speech

university of groningen

campus fryslân

Thesis

by

Yiqiu Wang

supervisor: dr. Shekhar Nayak
external supervisor: dr. Aki Kunikoshi

Acknowledgements

Completing my graduation thesis has felt like a remarkable journey, filled with challenges and rewards. It is with profound gratitude that I acknowledge the unwavering support of my friends and mentors, without whom my success would have remained unattainable.

I extend my heartfelt appreciation to **Aki**, whose guidance and assistance proved invaluable throughout the entirety of this endeavor. I also want to thank **Shekhar** and **Matt** for understanding and supporting me. I am thankful to **Iva** and **Elle**, for being part of the Motivation Group. We comforted and motivated each other. To all who provided their unwavering support, belief, and inspiration, my profound appreciation goes out.

Contents

ABSTRACT

The assessment of femininity and masculinity levels within speech constitutes a significant attribute perceptible to human listeners. This study delves into a novel methodology for quantifying these gender-related aspects of speech by leveraging the Deep Speaker model in tandem with Support Vector Machine (SVM) techniques. The outcome of this evaluation manifests as the "feminine level" of the speech. Through a comparative analysis between the predicted feminine level derived from the model and the outcomes of a listening test, which captures human perceptions of femininity and masculinity of speech, a discernible positive linear correlation emerges. Consequently, this study concludes that the proposed technique effectively predicts feminine levels, offering promising implications for applications centered around speech assessment and synthesis.

Keywords: feminine level; deep speaker model; SVM; human perception

1

INTRODUCTION

Voice technology (VT) has witnessed significant development over the past few decades, enabling humans to interact with computers, smartphones, and other devices with their voice rather than typing or touching. VT applications such as speech recognition, speech synthesis, and speaker verification (SV) have become widely used in our daily life and benefit us in various aspects. For example, Speech recognition systems that convert the input speech into text help users with either devices or their friends without evening touching the devices, while speech synthesis systems that convert the input text into speech improve user experience by making devices generate human-like speech. One of the most common applications of VT is in virtual assistants like Siri and Alexa. Such intelligent voice-activated assistants can answer questions, set reminders, send messages, control certain home devices, etc. In navigation systems, VT applications assist the driver to receive guidance and instructions on the direction without taking their hands off the steering wheel, which improves the safety of driving. In addition, the speech recognition technique is commonly used in transcription services, helping the user to transcribe audio content into text content efficiently and accurately. This benefits a wide range of individuals from video content creators to researchers. The technique is also utilized in interactive voice response systems in customer service to guide callers through automated menus and lead them to the department corresponding to their needs. On the other hand, speaker verification has become a crucial problem of voice-controlled applications, especially those involving sensitive private data or financial transactions. Smartphones, home devices, and online bank accounts make use of speaker verification techniques to make sure that only authorized users have access to these systems. Since voice-controlled applications play an increasingly important role in our life while voice spoofing is becoming used, we have paid much attention to speaker verification technique, which acts as a safeguard against potential fraud and identity theft and protect private information as well as financial assets. While various techniques have been developed to solve such tasks, most of them used speaker embeddings to represent the speaker's characteristics based on their speech those embeddings are able to contain a variety of information in addition to the speaker identification. In my thesis, I am going to use speaker embeddings to explore the gender information of speakers, and then use these embeddings to train a model for evaluating how feminine/masculine a speech is. Meanwhile, I am going to conduct a listening test to obtain the human perception of how feminine/masculine a speech sounds, and the effectiveness of my methods can be gained by analyzing the correlation of the results.

The thesis consists of five chapters. In the introduction part, I have made a brief picture of the applications in VT and explored the possibilities of speaker embeddings in multiple speaker-related tasks. In the second chapter, I will systematically address three key facets. Firstly, I will elucidate the linguistic correlates of gender, synthesizing a multitude of linguistic studies that unveil the manifestations of gender differences in voices through various factors such as physiology, behavioral habits, societal influences, and linguistic-cultural elements. Secondly, I will expound upon the concept of speaker embeddings, delineating a sequence of speaker embedding methodologies that traverse from earlier to more contemporary approaches. Lastly, I will undertake an analytical examination of the application of speaker embeddings within the realm of speaker recognition, with a specific focus on gender recognition. These research findings collectively furnish a robust theoretical foundation for substantiating the propositions central to my study. In the third chapter, I will give an in-depth description of the dataset used in this experiment and the models employed. In the fourth chapter, a comprehensive analysis of the experimental results will be explained, coupled with a detailed discussion of the limitations of my experiments and recommendations for future studies while in the last chapter, I will conclude this project.

2

Literature Review

This chapter provides a comprehensive overview of gender-related studies in speech from both linguistic and technical dimensions. Beginning with a linguistic perspective, the discussion delves into the nuanced gender-related information inherent in speech. Next, I illustrate what is speaker embeddings and review conventional and contemporary techniques employed in extracting these embeddings. The discourse then transitions to an exploration of research endeavors centered around gender recognition, wherein speaker embeddings are harnessed as pivotal tools. Within this context, the chapter culminates by presenting the formulated research question and hypothesis, setting the stage for the subsequent investigative work.

2.1. Linguistic Correlates of Gender

In the book “Gender: Linguistic Aspects”, which was written by A.V. Kirilina in 1999, it was discussed that the rise of the concept of "gender" meant a series of social and cultural expectations imposed by society upon individuals, based on their biological sex. At first, the research on gender characteristics of speech took place in the West, and the initial systematic explanations of the features of the speech of different genders were carried out based on languages from Romance and Germanic language families. The term "gender" was employed to paint a picture of the social, cultural, and psychological aspects of "feminine" in comparison with "masculine", and that is, “while highlighting everything that forms traits, norms, stereotypes, roles, typical and desirable for those whom society defines both women and men”. Meanwhile, the term was regarded by some scholars as a concept shaped by society and through languages. Women and men speak differently and listeners are able to distinguish the gender of the speaker in many ways, such as the quality of voice, the utterance style, the language forms, and so on (Rustamov et al., 2021). The information conveyed through speech includes not only the contents but also the speaker’s nationality, emotional status, gender, etc. The voice of a male and a female is different and various studies have found that listeners are able to correctly detect the gender of a randomly-picked normal adult speaker at a rate nearly reaching 100%. Many linguistic researchers have focused on gender distinction and studied the connection between language and sex. One of the main reasons for the male-female difference is the biophysical difference such as the vocal fold and the vocal tract. In the articulation of speech, the air leaves the lungs and passes through the vocal fold and causes it to vibrate or in other words, it opens and closes in very quick succession. The frequency of vibration or fundamental frequency (F0) is related to the perceived pitch of the voice. In most cases, a male has a longer and thicker vocal fold that vibrates more slowly at a lower frequency than a female. The average F0 of male speakers of languages like German and English is 100-120 Hz (Hertz, cycles per second) while the average F0 of female speakers is almost twice the male F0 (200-220 Hz) (Simpson & Weirich, 2020). The length of the vocal tracts between males and females differs as well with a typical adult male vocal tract being 17 cm, and the female vocal tract at 14 cm according to Goldstein in her thesis (Goldstein, 1980). The length gap results in a lower resonance frequency in males in comparison to females, for example, the formants of vowels produced by females are roughly 20% higher than those articulated by males (Fant, 1966). The anatomical difference between males and females causes the fundamental frequency of speech sounds produced by different genders to vary to a considerable extent. Moreover, the F0 of speakers of the same gender may also differ according to their age, and in fact, even for a specific person, the voice he makes may slightly vary through aging. In addition to these anatomical differences between the two genders that cause phonetic variability in speech, behavioral and social aspects also have influences on speech. The speaker's gender has an impact on his or her phonation type, giving rise to varied characteristics of voice. Voices produced by females are considered more breathy in many instances due to having a more significant glottal open quotient -GOQ than the other gender. On the contrary, voices produced by males, especially by American English speakers, are more creaky in most cases due to having a particularly low GOQ (Pépiot, 2014). Nonetheless, it's important to note that though such a voice was found to be used much more regularly by men in the past, in recent times, it has been identified as a phonetic signal passing on social interpretation in females as well. A study revealed a common creaky voice occurrence among females from northern California and eastern Iowa. This vocal characteristic was associated with perceptions of being well-educated, urban-focused, and having upward mobility. Furthermore, in sociological aspects, since female speech is expected to be clearer, there are several characteristics in them such as longer utterance durations, longer segment durations, slower speech rates, a larger vowel space, and larger contrasts between some consonants (Simpson & Weirich, 2020). Nonetheless, the majority of the findings mentioned earlier were carried out among English speakers, when we take into account the data from other languages we can discover some curious facts. Even though differences in mean F0 between males and females are universal, van Bezooijen (Van Bezooijen, 1995) found that women having a high pitch are regarded as more attractive to Japanese listeners but not to Dutch listeners. Meanwhile, another study has shown that the differences between the genders in mean F0 in German speakers are more significant than in Swedish speakers because of the considerably lower mean values for Swedish women than for German women (Weirich et al., 2019). As another example, a study presented that in the Chinese Wu dialect, such as Shanghainese, the mean F0 of both males and females was nearly equivalent. Actually, there are several parallel instances in many other languages, and the cross-gender differences show variations from one language to another, voice differences resulting from gender are relatively smaller in Danish compared to Russian. Besides the gender, age, and language factors that result in voice difference, especially the difference in F0, some other factors such as daily habits also have impacts on it and it is found that the F0 of speech articulated by smokers is generally brought to a lower level. In addition to the F0, many studies have shown that the various vowel and consonant formants tend to be placed at higher frequencies in female speakers. These factors discussed earlier are unlikely to stem from physiological and anatomical variations between males and females. Instead, socially constructed behaviors can provide the underlying reason (Pépiot, 2014). I have fully and comprehensively discussed the gender-related information delivered through a speech from a linguistic perspective and explained the difference resulting from varied biophysical articulation systems and the behavioral and social divergence between males and females, and noted the perception difference in cross-language and cross-culture situations.

2.2. Speaker Embeddings

Speaker embeddings are representations of speaker characteristics in a multidimensional space that are derived from speech signals using different methods such as the deep neural network model. They are typically obtained using techniques like i-vector, d-vector, s-vector, or x-vector. These embeddings encode various properties related to the speaker, such as pitch, gender, accent, speaking style, and speaking rate and therefore, they are able to capture the unique voice characteristics of an individual speaker and can be used to discriminate between different speakers. They have a wide application in various tasks such as speaker verification. In these tasks, these embeddings are used to encode and compare speaker-related information and determine whether the input speaker identity is the target speaker identity or to detect the input speaker identity among the dataset (Wang et al., 2017).

There are several different kinds of speaker embeddings. To begin with, the super-vector is a high- and fixed-dimensional representation of an utterance that combines many smaller-dimensional vectors into a single vector. It can be created by stacking the mean vectors of a Gaussian Mixture Model - Universal Background Model (GMM-UBM). In speaker recognition systems super-vectors are used as inputs to Support Vector Machines (SVMs) for classification. Next, the i-vector is derived from the i-vector framework. It is a fixed- and low-dimension vector that models the speaker and channel variability in a low-dimensional space. In the i-vector framework, a Universal Background Model (UBM) is trained using a large amount of speech data. Next, for each speaker, a Total Variability Matrix (T) is estimated to capture the speaker-specific information. Then, the i-vector is obtained by projecting the speaker's speech data onto the T matrix. The formula is: $M = m + Tw$. The "M" represents the speaker- and session-dependent super-vector while the "m" is a speaker and session-independent super-vector. The "T" is a low-rank matrix that captures speaker and session variability and the i-vector is the posterior mean of "w". To add to this, the d-vector is generated using deep learning techniques such as deep neural networks (DNNs) or convolutional neural networks (CNNs). The process of extracting the d-vector mainly includes four parts. First, in the audio signal segmentation procedure, the input audio signal is divided into multiple segments. Second, the system extracts the feature by extracting filter bank energy features from each segment. These features typically represent the energy distribution of the audio in different frequency bands. Then, in DNN training, a DNN is trained using the extracted features, during which the input is the filter bank energy features, and the output is the activations of the last hidden layer of the DNN. In the end, the output vectors of the hidden layer obtained from each audio segment are averaged to generate the d-vector. The averaging process combines the information from multiple segments into a comprehensive representation to capture the characteristics of the entire audio signal. By comparing the i-vector with the d-vector, we can find the differences between them. The i-vector is known for its ability to encode speaker identity to a large extent and can also encode speech content using GMM-UBM modeling. However, it cannot encode the order of words. On the other hand, the d-vector is effective in capturing discriminative speaker information and has been shown to perform well in speaker verification tasks (Wang et al., 2017).

Nevertheless, a new Deep Speaker model of extracting speaker embeddings and conducting speaker verification tasks has been established and results have shown that it outperformed both i-vector and d-vector in the task. In the model, the embeddings are learned directly from speech utterances using DNNs, while in the two traditional techniques, the embeddings are derived from statistical models. It's important to note that there is a significant variance between the dimensionality of this model and the other models. In fact, Deep Speaker embeddings have dimensions in the range of hundreds to thousands, while i-vector and d-vector embeddings are typically lower-dimensional, often around 400-600 dimensions. As a result, Deep Speaker embeddings have shown better adaptability to different languages and speaker variations compared to i-vector and d-vector embeddings because they can transfer well across different languages and can be adapted to small datasets using transfer learning techniques. In essence, the model is a flexible and ideal approach for generating speaker representations that capture the voice characteristics from raw speech data (Li et al., 2017).[1]

2.3. Gender Recognition Using Speaker Embeddings

Speaker Verification is a biometric technique that uses a speaker's voice's unique features and characteristics such as pitch and intonation to verify their identity. This technique has multiple applications in different fields, such as security systems and smart home devices. The goal of this technique is to ensure the security and authenticity of the identity of the user by their voice. There are two main branches of speaker verification: speaker recognition (SR) and speaker verification (SV). Speaker recognition, also known as speaker identification, is the task to determine the identity of a speaker according to their voice. In a speaker recognition system, there is a database of voice samples from multiple speakers, and a one-to-many comparison is conducted between the input voice of the speaker and the database to find the best match. Unlike speaker recognition, speaker verification focuses on verifying whether a speaker is the target speaker or not. The SV system also has a database of voice samples from multiple target speakers. It

[1] The code for this model is available here.

does one-to-one matching by comparing the voice of a speaker with the target speaker in the database (Togneri & Pullella, 2011). While there are a number of studies focusing on the speaker verification task using speaker embeddings, some people saw the potential of using these embeddings in other speaker-related feature tasks. In a recent study, a relatively high accuracy of 99.60% for gender recognition has been reached when x-vector-based utterance embedder, as well as a d-vector-based system, were extended using transfer learning schemes and a QuartzNet embedder. In their model, the processed waveform of the utterance first goes through a feature extractor and then a DNN embedder. The gender classifier is then able to classify the gender of the speech based on the utterance embeddings.

2.4. Research Question and Hypothesis

The speech of males and females has a great difference and humans are able to perfectly distinguish the gender of a speaker according to the voice (Simpson & Weirich, 2020). Linguistic scholars have extensively examined the distinctions between male and female speech patterns, considering physiological, behavioral, societal, and cross-cultural perspectives. In contemporary times, a plethora of technological advancements have been devised to accurately discern the gender of a speaker, yielding notable achievements. It is worth noting, however, that even within the same gender category, the degree of femininity or masculinity inherent in individual voices varies. In other words, listeners possess the ability to discern and evaluate which voice exhibits a greater degree of femininity or masculinity. Such evaluations are inherently subjective, and while conclusions may diverge among individuals, a degree of consensus tends to emerge. For instance, voices with higher pitch are generally perceived as more feminine. Intriguingly, no endeavors have been undertaken thus far to explore the application of AI models in gauging the degree of femininity or masculinity in voices. The gap leads to the research question and hypothesis of this study (Kwasny & Hemmerling, 2021).

Research Question: What is the potential efficacy of training an SVM algorithm to achieve precise predictions of the perceived level of femininity in human voices?

1 Is it possible to train an SVM algorithm to effectively classify the speaker's gender based on their speaker embeddings?

2 Is it practical and reliable to define the feminine level of either a speech or a speaker based on the distance between the speaker embeddings and the hyperplane calculated by the SVM?

3 Does the generated feminine level match the human perception of how feminine/masculine a speech is?

Hypothesis: Using a large amount of speaker embeddings data to train the SVM, I can obtain a criterion of feminine level that aligns with human perception.

A couple of studies have proven that the gender information of the speaker can be explored and represented by the extracted embeddings. Kwasny and Hemmerling proposed a gender classification architecture with an accuracy level of 99.6% in their 2021 research Kwasny and Hemmerling (2021) while another group successfully embedded the gender information using an advanced algorithm to improve the performance in speaker identification (Tang et al., 2022). It is reasonable to predict that by embedding the gender information in detail, I can train a model to determine the feminine level of a speech that matches such level from human perception.

Bibliography

Fant, G. (1966). A note on vocal tract size factors and non-uniform f-pattern scalings. *Speech Transmission Laboratory Quarterly Progress and Status Report*, *1*, 22–30.

Goldstein, U. G. (1980). *An articulatory model for the vocal tracts of growing children* (Doctoral dissertation). Massachusetts Institute of Technology.

Kwasny, D., & Hemmerling, D. (2021). Gender and age estimation methods based on speech using deep neural networks. *Sensors*, *21*(14), 4785.

Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., Cao, Y., Kannan, A., & Zhu, Z. (2017). Deep speaker: An end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*.

Pépiot, E. (2014). Male and female speech: A study of mean f0, f0 range, phonation type and speech rate in parisian french and american english speakers. *Speech Prosody 7*, 305–309.

Rustamov, D., Shakhabitdinova, S., Solijonovc, S., Mattiyev, A., Begaliyev, S., & Fayziev, S. (2021). Research of peculiarities of speech of male and female on phonetic and lexical levels of language. *Journal of Language and Linguistic Studies*, *17*(1), 421–430.

Simpson, A. P., & Weirich, M. (2020). Phonetic correlates of sex, gender and sexual orientation.

Tang, Y., Liu, C., Leng, Y., Zhao, W., Sun, J., Sun, C., Wang, R., Yuan, Q., Li, D., & Xu, H. (2022). Attention based gender and nationality information exploration for speaker identification. *Digital Signal Processing*, *123*, 103449.

Togneri, R., & Pullella, D. (2011). An overview of speaker identification: Accuracy and robustness issues. *IEEE circuits and systems magazine*, *11*(2), 23–61.

Van Bezooijen, R. (1995). Sociocultural aspects of pitch differences between japanese and dutch women. *Language and speech*, *38*(3), 253–265.

Wang, S., Qian, Y., & Yu, K. (2017). What does the speaker embedding encode? *Interspeech*, 1497–1501.

Weirich, M., Simpson, A. P., Öjbro, J., & Ericsdotter Nordgren, C. (2019). The phonetics of gender in swedish and german. *Fonetik 2019, Stockholm, Sweden, 10-12 June, 2019*, 49–53.

3

Methodology

This chapter entails a comprehensive examination of the dataset employed in the training, evaluation, and listening test phases, coupled with a thorough explication of the methodological framework adopted throughout the study.

3.1. Data

The extraction of speaker embeddings in this thesis is facilitated through the utilization of the CSTR VCTK Corpus (Centre for Speech Technology Voice Cloning Toolkit). This corpus comprises a diverse compilation of speech data contributed by 110 English speakers, encompassing an array of accents including English, Irish, Scottish, American, and Australian. Within the dataset, each individual speaker is represented by an approximately 400-sentence corpus. These sentences have been meticulously curated from reliable sources such as newspapers, the rainbow passage, and an elicitation paragraph derived from the speech accent archive. Notably, the entire corpus of speech data adheres to a standardized recording protocol, meticulously captured using a dual-microphone setup. This setup encompasses an omnidirectional microphone (DPA 4035) and a small diaphragm condenser microphone with an expansive bandwidth (Sennheiser MKH 800). Recording procedures were meticulously conducted within a controlled environment: a hemi-anechoic chamber situated at the University of Edinburgh. This environment ensured the mitigation of external interference, thereby guaranteeing the purity of the captured speech data. The recordings were undertaken at a sampling frequency of 96kHz, along with a resolution of 24 bits. Collectively, these meticulous recording measures underscore the high-fidelity nature of the speech data, further augmenting the credibility and efficacy of the subsequent speaker embedding extraction process.

During the training phase, I harnessed speech data sourced from all 22 American English speakers within the dataset, comprising 12 females and 5 males. From each of these speakers, I randomly extracted 100 utterances, aggregating to a total of 2200 speaker embeddings for training purposes. For the evaluative component, my focus shifted to 33 speakers possessing an English accent, evenly balanced between genders with 18 females and 15 males. To mitigate the influence of various factors—such as the speaker's identity, utterance content, and recording equipment—on the speaker embeddings, I specifically chose speech data recorded utilizing the omnidirectional microphone (denoted as "mic1" in the dataset) for this analysis. These recordings encompassed identical utterances produced by the selected speakers.

Within the ambit of the listening test, conducted to discern human perception and explore the alignment between the crafted feminine level criteria and actual human perception of voices, I elected to work with a subset of 10 speakers from the larger pool of 33. Initial exclusions involved omitting the speaker identified as p227 due to her age (38) deviating from the specified age range of 20 to 24 years that other participants adhered to. Subsequently, I stratified the remaining 32 speakers into gender-based groups and reorganized them based on their feminine level scores. From the male group of 15 participants, I systematically selected every third speaker, commencing with the individual possessing the highest feminine level. This culminated in the selection of the 1st, 4th, 7th, 11th, and 14th speakers. Similarly, within the female participant cohort of 17 individuals, I adopted a similar approach, handpicking every fourth speaker from the pool, beginning with the participant possessing the lowest feminine level. This resulted in the inclusion of the 1st, 5th, 9th, 13th, and 17th speakers. In the data about these 10 chosen speakers, I methodically culled three utterances from each, ensuring uniformity in the content of the selected utterances across the cohort.

* Sentence 1: Ask her to bring these things with her from the store.

* Sentence 2: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. Sentence

* Sentence 3: She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

3.2. Models

In the course of this experiment, I leveraged the utilization of two distinct models, the Deep Speaker model and the Support Vector Machine (SVM). The Deep Speaker model was employed to extract the speaker embeddings from each meticulously selected WAV file. These embeddings subsequently played a pivotal role in both the training and evaluation phases of the experiment.

3.2.1. Deep Speaker Model

The model I chose to use in this experiment to extract speaker embeddings is the Deep Speaker model. It is a neural speaker embedding system that maps a speech to a hypersphere where speaker similarity is measured by cosine similarity. The system utilizes deep residual CNN (ResCNN) and gated recurrent unit (GRU) architectures to extract acoustic features and employs mean pooling to generate speaker embeddings at the utterance level. The model is trained using a triplet loss based on cosine similarity. The speaker embeddings extracted by the Deep Speaker model can be then utilized in many tasks, such as speaker identification, speaker verification, and speaker clustering. Experimental results demonstrate that Deep Speaker outperforms the DNN-based i-vector baseline system in speaker verification and identification tasks. The model is available on GitHub. The architecture consists of 5 steps. First, it preprocesses the raw audio, and then, it extracts features using a feed-forward DNN with ResNet-style deep CNN or Deep Speech 2 (DS2)-style architecture. After converting frame-level input to an utterance-level speaker representation, it maps temporally-pooled features to a speaker embedding using an affine layer and length normalization layer. Finally, it applies the triplet loss layer to maximize cosine similarities between embeddings from the same speaker and minimize those from different speakers (Li et al., 2017).

3.2.2. Support Vector Machine (SVM)

Subsequently, the extracted speaker embeddings of the American English speakers were harnessed as training data for the development of an SVM model. The model is a supervised machine learning algorithm used mostly for classification and regression tasks. Therefore, they are particularly helpful in solving binary classification problems, in which the goal is to categorize data into two groups based on certain features. In my experiment, I trained an SVM to discern the subtle distinctions between the embeddings associated with male and female speakers. In this way, I obtained the hyperplane that best differentiates between the speaker embeddings of male and female speech. Following the successful derivation of this pivotal hyperplane, I proceeded to define the concept of "feminine level" as the distance from the target speaker embeddings to the calculated hyperplane. After that, I systematically computed the feminine level for each utterance spoken by every British English speaker in the dataset. Moreover, I calculated the average feminine level corresponding to each individual speaker. This comprehensive assessment enabled me to gain insights into the nuanced dimensions of the feminine level across different speech contents and individual speakers within the British English speaker subset.

3.2.3. Demonstrator

The demonstrator of the whole experiment is in this GitHub folder.[1]

3.3. Listening Test

In the listening test, the participants were first asked questions about their gender, age range, and the region they spent most of their life in. Instead of asking about the listener's native language, I asked: "In which country have you spent most of your life?" because, about gender recognition, I assume that cultural influence brings greater influence than that resulting from the native language. For example, there may be a difference between Dutch speakers in the Netherlands and Dutch speakers in Belgium about gender recognition and perception. At the same time, to avoid intra-rater variability, each WAV file is evaluated twice by the same person to reach consistency.

Bibliography

[1] https://github.com/YiqiuWangVT/deep-speaker/tree/master/deep_speaker/notebook

Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., Cao, Y., Kannan, A., & Zhu, Z. (2017). Deep speaker: An end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*.

4

Results & Discussion

This chapter contains the results analysis as well as the discussion part of my experiment. I will thoroughly explore how the model was trained and how I defined and evaluated the feminine level of each speech and each speaker. Additionally, I will focus on the results obtained from the careful execution of the listening test, a crucial part of this research. This evaluation not only includes the main outcomes of the listening test but also takes into account important participant details like their gender, age, and primary country of residence. Furthermore, I will compare the "feminine score" results from the listening test with the "feminine level" data generated by the model. Lastly, I will reflect on the limitations and drawbacks that appeared during the course of this experiment and give some recommendations for further improved and bettered studies. This candid self-examination aims to extract valuable insights that can guide future studies toward improvements. By thoughtfully considering the experiment's limitations, we lay the groundwork for refining and advancing future research endeavors.

4.1. Analysis of the listeners

First, I will have a short description of the listeners in the listening test on their gender, age range, and region. Here, the region refers to the country where the listener has spent most of their life. The reason for taking the region into account is that such cultural differences may have a considerable influence on gender recognition and perception according to the studies I have discussed previously. The listening test garnered participation from a total of 32 volunteers. Among this group, it is noteworthy that the female participants outnumbered their male counterparts, constituting a majority at 61.54%, whereas males comprised a comparatively smaller proportion of 34.62% of the total participants. Meanwhile, it is necessary to mention that 3.85% of the participants fall in the gender of "other". In conclusion, the gender distribution underscores the higher representation of females in the study, indicating a potentially meaningful gender-related aspect within the context of the test.

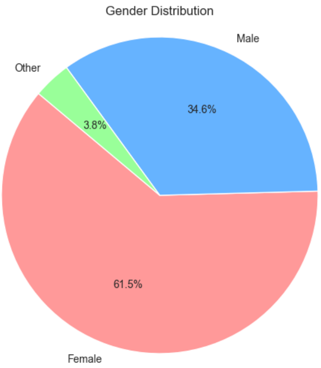

Figure 4.1: Gender Composition Distribution

In addition, the age composition of the participants in the study reflects a diverse spectrum. A significant portion, specifically 57.69%, falls within the age range of 18 to 30 years, indicating a notable presence of younger individuals. The listen aged between 41 and 50 years constitutes a proportion of 23.08%, showcasing a meaningful but comparatively smaller segment of the participants. Participants aged 51 to 60 years are less represented, accounting for 11.54% of the total, suggesting a relatively lower participation rate among individuals in this age bracket. Lastly, the age group spanning from 31 to 40 years encompasses 7.69% of the participants, indicating a modest yet distinct presence within the overall distribution.

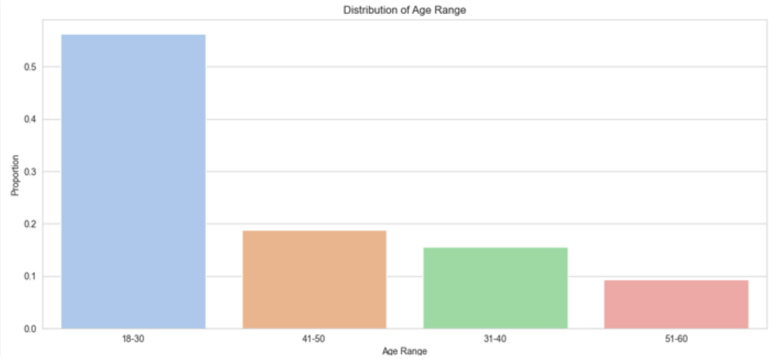

Figure 4.2: Participant Distribution by Age Range

The participants in this test exhibit a fascinating array of regions from around the world, offering an intricate and diverse global perspective. Leading the cohort is Japan, commanding a substantial presence at 23.08% of the total participants, thereby reflecting a robust engagement from this vibrant East Asian nation. Directly trailing this, we find China, which contributes 15.38% of the participants. Further enriching the international tapestry are the United States of America and the Netherlands, both contributing an equal distribution of 11.54%. This dual presence underscores the test's commitment to encompassing global viewpoints. Additionally, we observe Italy and Germany each accounting for 3.85%, which contributes a European perspective to the overall composition. Amongst the other contributing countries, each at 3.85%, we find Switzerland, Canada, India, the Plurinational State of Bolivia, Romania, Norway, and Bulgaria. This ensemble of nations underscores the test's embrace of diversity, capturing insights across a myriad of continents and cultures. To sum up, the inclusion of this diverse array of regions paints a vivid picture of the test's multinational and multicultural character, thereby providing a comprehensive and far-reaching understanding of the listening test's results from a truly cross-language and cross-culture standpoint.

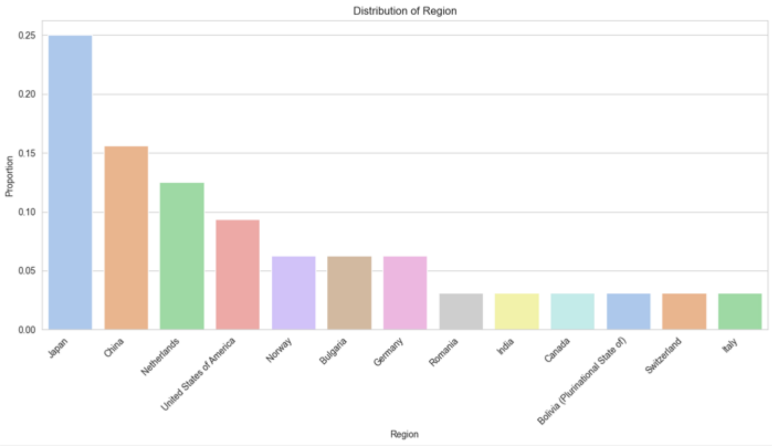

Figure 4.3: Participant Distribution by Country

4.2. Listening Test Results

First of all, as I have mentioned in the methodology chapter, in the listening test, each WAV file is evaluated twice by the same person to check consistency. Therefore, if the same test subject evaluated the same speech by the same speaker with a large difference in the feminine score (greater than or equal to 2), the result could be regarded as unreliable. I removed those unreliable results before analyzing and visualizing the data. Illustrated in Figure 4.4 is a meticulously designed scatterplot, meticulously curated to effectively portray the intricate interplay between two pivotal variables, namely the "feminine level" and the "feminine score." Within this graphical representation, the former entity epitomizes the outcomes emanating from the Support Vector Machine (SVM) model's intricate calculations, whereas the latter encapsulates the outcomes gleaned from the meticulous listening test conducted. The "feminine score," a pivotal metric of interest, is meticulously assessed on a meticulously calibrated scale extending from the numerical magnitude of 1 to the discernible magnitude of 5. This graded scale facilitates the nuanced assessment of participants' perceptions, elucidated subsequent to their auditory exposure to a randomized subset of speeches drawn from the experimental dataset. The options, presented in a systematic hierarchy, encompass a comprehensive spectrum of potential attributions that participants can select from, including:

1 very masculine

2 somewhat masculine

3 ambiguous masculine / feminine

4 somewhat feminine

5 very feminine

The careful structuring of these options offers participants an opportunity to not only make a binary selection but also to discern and articulate nuanced gradations of perception. This meticulously structured framework enriches the granularity of insights gathered from the listening test, facilitating a more nuanced comprehension of the intricate dynamics governing the alignment between the "feminine level" and the human-perceived "feminine score." As depicted in Figure 4.4, this graphic interface serves as a conduit for effectively conveying the complex relationship between these variables, enriching our understanding of the interplay between automated classification and human perceptual judgment.

After that, a meticulous calculation of the Pearson correlation coefficient ensued—a pivotal statistical metric renowned for its efficacy in quantifying the magnitude and orientation of the linear interrelation between two continuous variables. This statistical measure, which traverses a continuum ranging from -1 to 1, encapsulates a spectrum of correlations. The scalar value of "1" denotes a pristine positive linear correlation, indicative of a proportional augmentation in both variables as one variable increments. Conversely, the value of "-1" signifies an immaculate negative linear correlation, wherein an increase in one variable is reciprocated by a commensurate decrement in the other. Meanwhile, the value "0" signifies a conspicuous absence of linear correlation, attesting to the absence of any perceptible linear relationship between the variables under scrutiny. In the specific context of the interplay between "feminine level" and "feminine score," the calculated Pearson correlation coefficient is discerned as 0.9216057365496465. This outcome markedly underscores a robust and positive linear correlation between the two variables. This implies that the articulated criteria governing the determination of feminine levels intimately coincide with the collective human perception of the extent of femininity or masculinity conveyed by a given speech sample. To further expedite a comprehensive exploration into the model's efficacy in prognosticating the nuanced attributes of femininity as perceived by humans, an in-depth linear regression analysis was systematically conducted. The resultant linear regression equation stands as a testament to the meticulous analysis:

Feminine Score =5.8112 × Feminine Level + 2.7372

Integral to this regression analysis, the R-squared value registers at an impressive 84.94%. This pivotal statistic conveys that a substantial 84.94% of the variability in the output of the feminine score is intricately accounted for by the input parameter of the feminine level. This profound level of explanation lends credence to the model's precision in prognosticating the feminine attributes inherent within speech samples. Collectively, these analytical facets collectively substantiate the model's adeptness in discerning and articulating the feminine nuances imbued within the vocal expressions, thereby unveiling its proficiency in accurately predicting the feminine orientation of speech.

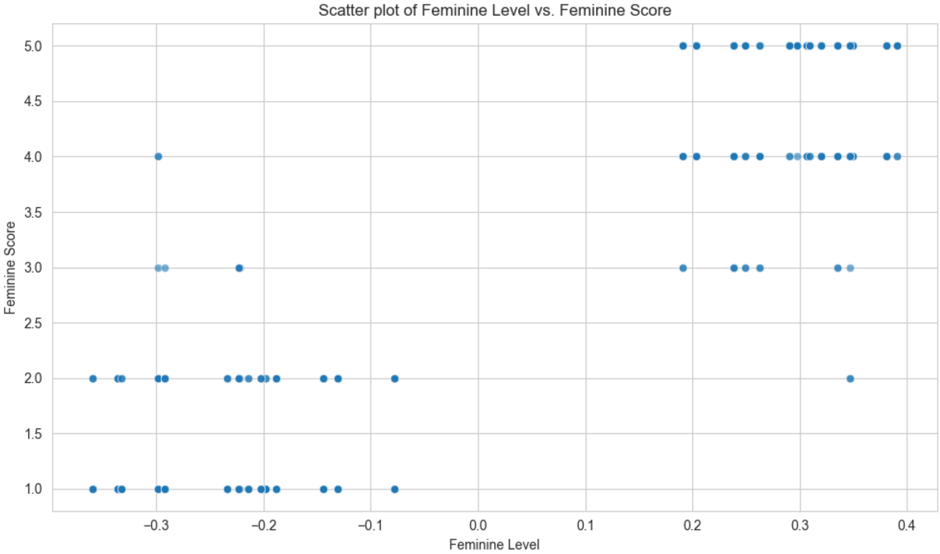

Figure 4.4: Scatter plot of Feminine Level vs. Feminine Score

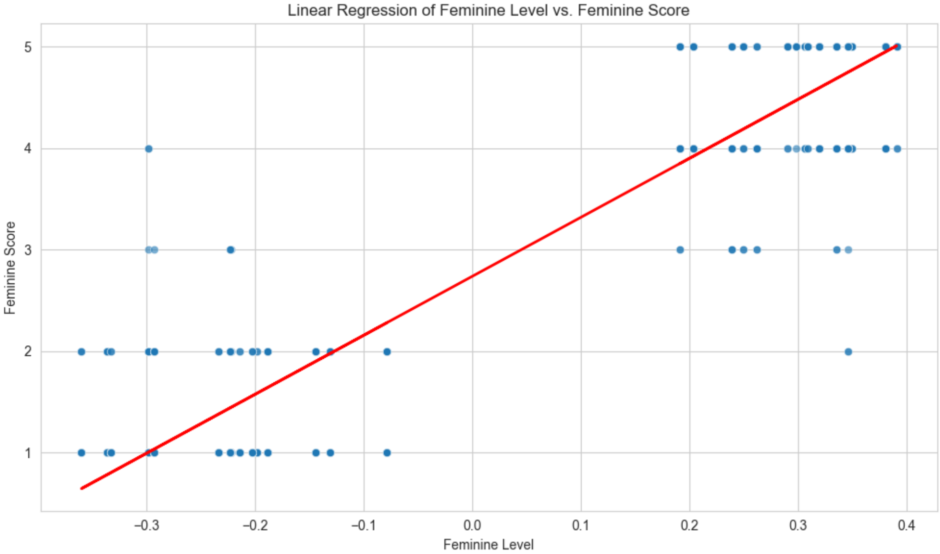

Figure 4.5: Linear Regression of Feminine Level vs. Feminine Score

4.3. Limitations and Recommendations

While the selection of the Deep Speaker model and utilization of the VCTK dataset for the present experiment offer a suitable foundation for research, it is imperative to acknowledge the prevailing limitations inherent within this methodology. Primarily, due to temporal constraints, a total of 2200 wav files from American English speakers exclusively were incorporated into the model training process. Consequently, the training data set does not encompass the broader spectrum of English speakers characterized by accents such as Irish, Scottish, Australian, and Canadian. The potential divergence in speech attributes, including pitch, among English speakers with diverse accents introduces the prospect of variance. This, in turn, holds the capacity to attenuate the accuracy and efficacy of the model in predicting feminine levels congruent with human perception. A plausible strategy for enhancing the model's performance in subsequent iterations involves the utilization of a more expansive dataset for training purposes. Indeed, the exclusive reliance on an English-language dataset within the confines of this experiment introduces a notable caveat that could potentially undermine the precision of the resultant findings. As expounded upon within the comprehensive review of existing literature, the acoustic attributes of speech exhibit noteworthy variations across different languages. Foremost among these disparities is the salient feature of vocal pitch, alongside the average pitch differentials observed between distinct genders (Weirich et al., 2019). The proposition emerges that a robust dataset incorporating diverse languages holds the promise of yielding heightened accuracy in the model's discernment, thus engendering a more refined rubric for the determination of feminine levels. It is reasonable to contend that the integration of a profusion of data across a gamut of languages would furnish the researcher with an enriched and more sophisticated framework capable of capturing and elucidating the nuanced degrees of femininity in speech. This broader dataset amalgamation augments the potential for the model to transcend the confines of English and facilitate more precise projections pertaining to the feminine attributes of speakers across various languages. In tandem with the constraints associated with the training data, the constraint of participant numbers within the listening test merits consideration. Notably, the demographics of the listening test participants indicate that over 50% fall within the 18 to 30 age bracket, while those older than 40 account for less than 10% of the total. This demographic distribution introduces the potential for a certain lack of objectivity in the feminine score ascribed by human perception, as the viewpoints of individuals beyond the younger generation are not well represented. As we embark on the forthcoming sequence of experiments, a promising avenue for mitigating this limitation lies in the expansion of our recruitment endeavors for the listening test. This strategic expansion entails proactive engagement with a more extensive and diverse cadre of volunteers, encompassing a wider panorama of age groups and cultural affiliations that span a multitude of countries. The deliberate cultivation of this diversified volunteer cohort bears the potential to furnish us with an amplified and intricate tapestry of human perceptual insights, intricately woven into the nuanced dimensions of speech characteristics, particularly concerning the dichotomy of its masculine and feminine attributes.

BIBLIOGRAPHY

Weirich, M., Simpson, A. P., Öjbro, J., & Ericsdotter Nordgren, C. (2019). The phonetics of gender in swedish and german. *Fonetik 2019, Stockholm, Sweden, 10-12 June, 2019*, 49–53.

5

Conclusion

This project explores the utilization of the Deep Speaker model for the purpose of extracting speaker embeddings. Subsequently, these extracted embeddings are employed to train an SVM, resulting in a model that effectively discriminates between different gender-based vocal characteristics. Moreover, this model demonstrates the capacity to accurately calculate the distance from given speaker embeddings to the calculated hyperplane, a metric referred to as the "feminine level" within this context. As the results of the evaluation part, the study encompasses the computation of both the average feminine level of each of the 33 British English speakers within the VCTK dataset and the individual feminine levels associated with each utterance produced by every speaker. The last stage of experimentation entails a listening test, wherein an analysis of the feminine scores assigned by 32 participants is compared with the feminine levels predicted by the model. Through this analysis, a linear relationship between these two sets of results is established, indicating a positive correlation between them. The ultimate findings of this study validate the initial research hypothesis. By leveraging a substantial volume of speaker embeddings data for SVM training, I have successfully derived a criterion for assessing feminine levels that is in concurrence with human perceptual judgments. These outcomes signify the feasibility of employing the model to quantify and compute the degree of femininity or masculinity within speech. This model offers the potential to assess the gender-related attributes of synthesized speech, thereby enabling the evaluation of the quality of generated audio through an analysis of the generated speech's feminine level. This application holds particular significance for personalized Text-to-Speech (TTS) systems, where it can provide valuable assistance in refining the quality and naturalness of synthesized voices.