university of groningen

campus fryslân

Synthesis of sarcastic speech: Research on adjusting pitch and energy at keyword level using FastSpeech2

Weihao Jiang

university of groningen

campus fryslân

University of Groningen - Campus Fryslân

Synthesis of sarcastic speech: Research on adjusting pitch and energy at keyword level using FastSpeech2

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Zhu Li (Voice Technology, University of Groningen)
with the second reader being

Weihao Jiang(S5464781)

June 11, 2024

Acknowledgements

First of all, I would like to thank my supervisor Zhu Li. I have received his careful guidance on both the technical aspects and the writing of this paper. I am very grateful to him. The second thing I want to thank is Matt Coler and Phat Do. They gave me valuable opinions and guidance in the early stages of the thesis. Moreover, I would like to thank the Habrok provided by the University of Groningen, which made my training and inference stages very smooth. Finally, I would like to thank Yinqiu Wang, who is in the same group with me, for constantly working with me on modifying the model, as well as for her support and encouragement.

Abstract

Sarcasm is one of the most common and important rhetorical techniques in daily life, and the synthesis of sarcastic speech is also one of a crucial aspect of emotional speech synthesis that warrants attention. While previous research has extensively focused on the detection and recognition of sarcasm, there has been less emphasis on the synthesis of sarcastic speech. Therefore, this thesis explores how to use the LLM (Large language model), ChatGPT, to predict sarcastic keywords within sentences in the model and synthesizes sarcastic speech by precisely controlling the pitch and energy of these keywords using the FastSpeech2. This research can effectively fill a gap in the field of sarcastic speech synthesis. Additionally, an evaluation involving 22 native or second-language English speakers validated the practicality and effectiveness of this method in enhancing the recognition and synthesis of sarcastic tones. Experimental results demonstrated that controlling the acoustic features of keywords alone within a sentence can significantly improve the perception of sarcasm in listeners, compared to global-level pitch and energy control. The results of this experiment can be viewed on the GitHub page: `https://weihaohaoao.github.io/weihao.github.io`. The specific method of synthesizing sarcastic audio and the modification of the model in this study is open-sourced, which can be found on my github page `https://github.com/Weihaohaoao/Synthesis-sarcastic-voice/tree/main`

Key words: Sarcasm; synthesis; keywords; detection; FastSpeech2

Contents

1 Introduction

The speaker's emotional state plays an important role in natural language understanding, and the sarcastic state is often used in our daily communication and also in human-computer interaction. Although deep learning technology continues to promote the development of speech, emotions such as sarcasm and exaggeration cannot be expressed accurately, which may lead to bad effect and quality of speech synthesis to a certain extent. For example, ”My cell phone battery is awesome, it ran out of power in only ten minutes.” People can easily tell that ”awesome” here is not its original literal intention, but a sarcastic way of expressing the bad performance of the phone battery. Therefore, the literal meaning of sarcasm is contrary to the speaker's intention to complain, or to criticize something. Although in most cases one can determine whether a sentence is sarcastic by only reading the text, such as the example mentioned above. However, in some other cases, it is impossible to judge by text alone. For example, ”Your haircut is great.” If the word ”great” is in a sarcastic way in the speaker's original intention, which means the sentence means bad comment it cannot be defined by text alone, but to combine audio and even video to further get the result.

Previous researches in sarcastic recognition and synthesis were almost all focused on the acoustic features of the entire sentence, but according to the examples above, we can find that whether the sarcasm that can be seen directly from the text is in the sentence, or on the contrary, the sarcasm in the sentence emotions are often not reflected at the level of the entire sentence, but can be made sarcastic through changes in the acoustic characteristics of a few specific keywords. In this paper, I will synthesize and evaluate sarcastic emotion sounds by modifying the Fastspeech2 where the acoustic features are controlled so that they act on predicted sarcastic keywords. This study can fill the gap in sarcastic synthesis research, and study on emotional synthesis from a new perspective, the perspective of keywords, providing new opportunities and directions for subsequent research.

The paper is organized as follows: Section 1 introduces a general situation of research background and content. Section 2 provides a comprehensive literature review that explores the scientific basis of this study, including the development of speech synthesis, sarcasm detection, and expressive speech techniques. In addition, this section introduces related research on the FastSpeech2 model, which plays an important role in this study. Section 3 describes the research methods in detail, Section IV introduces the experimental details, including how to change model, and Section 5 explains the evaluation strategy. Then, Section 6 presents the research results in depth, and Section 7 discusses the research findings, including any limitations encountered during the research process. In the last part, Section 8, summarizes the main insights and significance of this study.

1.1 Research Question and Hypothesis

In the Fastspeech2 model, specific acoustic feature adjustments (such as pitch, energy, etc.) for sarcastic keywords in sentences can improve the recognition effect of sarcastic sentences and the naturalness perceived by listeners. Compared with the traditional whole sentence adjustment method, this method has obvious advantages in simulating sarcastic tone. This study will evaluate the effect of this sarcastic audio synthesis method by designing a control experiment, comparing the effect of keyword adjustment, and using subjective testing methods.

2 Literature Review

In this part, I will first fully elaborate on the concept and importance of speech synthesis, sarcasm and conduct a detailed analysis based on the acoustic characteristics of speech and the trend of sarcastic speech features in different languages. Secondly, the milestones of expressive speech synthesis and sarcasm detection are outlined to let readers understand the main process of sarcastic discourse detection. Finally, I described in detail the emotion detection capabilities and achievements of large language models, especially ChatGPT in various fields.

2.1 Speech synthesis

Speech Synthesis, which is synonymous with Text-to-Speech (TTS) in most cases, is a technology that converts text into speech. The development of speech synthesis can be roughly divided into three stages: features synthesis, statistical parametric synthesis and neural speech synthesis. In this section, a detailed introduction of the development process in all three stages will be presented.

2.1.1 Features synthesis

Articulator synthesis generates speech by imitating the movement of human vocal organs, such as the most common parts, lips and tongues. This method tries to reproduce the human speech generation process as naturally as possible. Formant synthesis generates speech by controlling a simplified source-filter model through a set of rules. These rules imitate the formant structure and other spectral characteristics of speech without relying on large speech datasets, but the generated speech is usually not natural enough. Concatenation synthesis generates speech by concatenating pre-recorded speech fragments, which range from whole sentences to syllables. The system finds and concatenates matching speech units based on the input text, although it cannot generate high-fidelity speech.

Allen, Hunnicutt, Carlson, and Granstrom (1979) made a study used formant synthesis technology, which made the system to accurately realize the formant synthesis of speech through digital technology, and combined abbreviation conversion, vocabulary analysis, and letter-to-sound rule conversion, realizing the conversion from English text to synthesized speech for the first time. The system converts the phoneme information in the text into the parameters required for speech synthesis through phoneme-to-parameter conversion, generating natural and fluent speech output.

In Moulines and Charpentier (1990)'s research, the main algorithm used in the paper relies on the pitch-synchronized overlap-add (PSOLA) method to modify speech prosody and concatenate speech waveforms. The modification of speech signals can be performed in the frequency domain (FD-PSOLA) using fast Fourier transform or directly in the time domain (TD-PSOLA). The frequency domain method can flexibly adjust the spectral characteristics of the speech signal, while the time domain method improves the efficiency of speech synthesis. Although this algorithm essentially process by splicing speech fragments, there is still many aspects can be improved.

In Clark, Richmond, and King (2004)'s study, this is a general unit selection speech synthesis engine that uses a unit selection algorithm to dynamically match the most appropriate phoneme unit from a speech database, which has been already designed, to minimize the target and connection costs, and generate high-quality speech. The core technology of the system include prediction of the target sentence structure, pre-selection of phoneme units, and Viterbi algorithm of phoneme unit sequences.

2.1.2 Statistical parametric synthesis

Statistical Parametric Speech Synthesis (SPSS) is a synthesising method for generating speech based on statistical models, which depends on possibilities. It learns speech parameters and dynamic features from a large amount of speech data by training Hidden Markov Models (HMM) and other methods. When synthesizing speech, SPSS generates parameters based on these models and then converts these parameters into speech waveforms through a vocoder. Compared with traditional concatenated synthesis, SPSS is more flexible and can effectively control the emotion and rhythm of speech. However, the generated speech may appear slightly smooth and unnatural.

Yoshimura, Tokuda, Masuko, Kobayashi, and Kitamura (1999)'s research, proposes a speech synthesis system based on HMM, which jointly models the spectrum, pitch, and state duration under a unified GMM-HMM framework. The pitch and state duration are modeled by multi-dimensional probability distribution HMM respectively. Subsequently, these parameters and states are independently clustered using a contextual clustering technique based on the decision tree algorithm. Finally, the synthesized speech is generated by a speech parameter generation algorithm based on HMM and a vocoder technique based on ”Mel cepstrum”. Tokuda, Yoshimura, Masuko, Kobayashi, and Kitamura (2000) derived a new speech synthesis algorithm for HMM, where the generated speech parameter sequence consists of observation vectors generated by HMM, which contain spectral parameters and their dynamic characteristics. In this method, it is assumed that the state sequence or part of it is an invisible hidden state. Therefore, the algorithm step by step calculates through a forward-backward algorithm and a parameter generation algorithm based on a given state sequence. Experimental results show that compared with a single hybrid HMM, this algorithm can generate a clearer resonance peak structure from multiple hybrid HMM.

2.1.3 Neural Speech Synthesis

Neural Speech Synthesis is a method that uses deep neural network technology (such as recurrent neural network, which is RNN, convolutional neural network, which is CNN and Transformer, etc.) to generate natural speech. In an end-to-end model, the method generates speech waveforms directly from text, eliminating the need for traditional intermediate processing steps. By learning large-scale speech data, the model can capture the subtle features of speech and generate high-quality speech.

WaveNet van den Oord et al. (2016) is a deep neural network model that generates raw audio waveforms. It performs probabilistic modeling of each audio sample and uses extended causal convolution technology (the calculation of the convolution kernel only relies on the current and previous time step data) to effectively handle long-term dependencies and ensure that the generated speech is consistent. Compared with traditional speech synthesis methods, WaveNet can handle the speech features of multiple languages and different speakers, showing its broad application potential in tasks such as speech synthesis and speech recognition. It improves the quality of text-to-speech synthesis by generating samples that are closer to natural speech.

Deep Voice Arik et al. (2017) is a real-time text-to-speech synthesis system based on deep neural networks that achieves high-quality speech generation through modern neural network technology. Different from traditional speech synthesis methods, Deep Voice uses some main components to build its end-to-end speech synthesis process: the letter-to-phoneme conversion model is responsible for converting written text into phoneme sequences; the phoneme boundary detection model uses connected temporal classification, which is also called CTC, to determine the boundaries of phonemes. The phoneme duration prediction model estimates the duration of each phoneme in speech. The CTC can solve the problem of inconsistent lengths between the input sequence and the output sequence, by using probabilities modeling the entire sequence, it can find the best match between the input sequence and the output sequence without precise annotation.

Wang et al. (2017) did a model called Tacotron is an innovative end-to-end text-to-speech synthesis model. It uses a sequence-to-sequence (seq2seq) framework and an attention mechanism to directly generate speech spectrum from characters, and then converts the spectrum into a speech waveform through the Griffin-Lim algorithm. The model includes a feature extraction module and a bidirectional GRU for encoding character sequences and aligning characters with the spectrum through an attention mechanism. Compared with traditional methods, Tacotron does not require manual feature engineering and complex phoneme-level annotation, and can efficiently learn from text and speech paired data through random initialization, thereby generating speech with high-quality naturalness and fluency. However, the disadvantage of this model is that it is autoregressive and takes a lot of time to synthesize audio.

2.2 Expressive Speech synthesis

Expressive speech synthesis is a technology that generates speech with natural human characteristics in speech. It makes the synthesized speech sound more natural and realistic by incorporating various emotional and prosodic features. For example, it can synthesize emotions like happiness, sadness, anger, etc.

Wang et al. (2018) did research an unsupervised model called Global Style Tokens (GST) is proposed, which is applied to style modeling, control and transfer in end-to-end speech synthesis. The GST model is based on the Tacotron architecture, which is introduced in last part. By training with unlabeled data, it generates a set of embedding vectors that can capture a variety of acoustic features and generate soft labels, achieving and enabling control over speech rate and speech style, reproducing a single word throughout the text. To the speech style of the audio, GST can not only train on unlabeled noise data and single factors such as noise and speaker identity, but also provide direction for large-scale and stable speech synthesis. Experimental results show that GST can not only effectively capture and generate different speech styles without labeling, but also achieve flexible control and migration of different speech styles through simple weighting operations.

Zhang, Pan, He, and Ling (2018) introduces an end-to-end speech synthesis model based on variational autoencoders (VAE) for unsupervised learning of latent representations of speech style. By introducing UVAE, which is by compressing high-dimensional data into a low-dimensional latent space, and then reconstructing the original data from this space, the model can then learn the latent representation of speaking style under label-free conditions and control style transfer. This research also combines the Tacotron2 architecture to extract style representations from reference audio through a recognition network and use them to check the style of synthesized speech. The model can be trained effectively on noisy unlabeled data, exhibits good style transfer performance, and good in various speech generation and style checking.

Skerry-Ryan et al. (2018) proposes an end-to-end prosody transfer model is proposed, still based on the extended Tacotron architecture, using unsupervised learning to extract prosody features from reference audio and embed them into a latent space. The model consists of a reference encoder composed of a convolutional neural network (CNN) and a gated recurrent unit (GRU) to process the reference audio and generate prosodic embedding vectors. These embedding vectors are then combined with text features to generate spectrograms and the final speech is synthesized via a neural vocoder such as WaveNet or WaveRNN. This method also uses variational autoencoders (VAE) to learn and control prosodic features without labeling, and provides precise control of prosodic changes such as pitch and stress. The experimental results show that the model is capable of generating high-fidelity, natural, fluent and expressive speech in multi-speaker situations.

2.3 The overview of sarcasm

Sarcasm is understood as a reversal of meaning, which is not limited merely to the semantic content but also encompasses force and evaluative attitudes. Here are four different subtypes of sarcasm defined by Camp (2012): Propositional Sarcasm, Lexical Sarcasm, Illocutionary Sarcasm, and ”Like”-Prefixed Sarcasm. The distinctions among these categories lie in the targets of inversion and the complex roles they play in actual communication. Specifically, sarcasm is one of the most common forms of communication in our lives. Kreuz and Roberts (1993) analyzed conversations among 149 university students and their friends, finding that 8 percent of conversational turns contained sarcasm, suggesting that sarcasm occurs nearly every two minutes in our everyday interactions. Similarly, in online environments, speakers often convey special messages to their audiences by controlling linguistic and prosodic information, such as tone of voice. Therefore, prosodic features often play a significant role in sarcastic communication.

Cheang and Pell (2008) conducted an acoustic analysis of native English speakers and explored the acoustic features of sarcastic speech, discovering that a sarcastic tone typically manifests as an overall decrease in fundamental frequency (F0) and Harmonics-to-Noise Ratio (HNR), as well as a reduction in the standard deviation of F0. These results suggest that specific vocal patterns play a crucial role in conveying sarcasm. Attardo, Eisterhold, Hay, and Poggi (2003), in their analysis of sarcastic discourse in American comedy, tracked pitch variations and found that sarcasm involves extreme pitch changes, such as starting with a high pitch and then shifting to a very low range, or maintaining a minimal pitch variation, resulting in a flat tone. Facial expressions also contribute to conveying sarcasm, often characterized by a ”blank face.” Similar patterns have been researched in many other languages. A study on Cantonese sarcasm Cheang and Pell (2008) found that Cantonese speakers demonstrated higher average F0, but smaller amplitude and F0 range in sarcastic versus sincere, humorous, and neutral tones. This not only significantly differed from sincere discourse but also showed an opposite trend to that of English in F0 variations.

Additionally, a study on Spanish sarcasm analyzed the prosodic features of 19 bilingual English-Spanish speakers residing in the American mid-west, using Praat (a software for speech analysis that allows for analyzing and manipulating speech) to analyze average frequency (F0), range, and other features. The results indicated that in English, sarcastic speech was slower than sincere speech. Moreover, when speaking Spanish, the mean and range of F0 were higher Rao, Ye, and Butera (2022). An acoustic analysis of 14 female speakers founded that sarcastic criticisms, compared to literal comments, had a lower average fundamental frequency (F0), higher energy levels, and longer vowel duration Scharrer and Christmann (2011). Furthermore, research by Rao et al. (2022) on the characteristics of French sarcasm involved 12 native French speakers who performed in specific contexts. Their sarcastic and literal intonations were recorded for further acoustic analysis using

Praat. This study found that sarcastic tone often features higher F0 levels, a wider pitch range, and longer duration, emphasizing the role of acoustic parameters in expressing sarcastic intent.

To conclude, sarcasm, as a prevalent mode of communication, exhibits complexity not only in linguistic content but also in vocal expression. Various studies demonstrate that sarcastic speech possesses distinct acoustic features across languages, such as decreased fundamental frequencies, pitch variation, and duration. These acoustic markers enable speakers to convey sarcastic intent without explicit articulation, while also challenging the listener's interpretative skills. Therefore, understanding and analyzing these features is not only meaningful for linguistic research but also has practical implications for enhancing the capability of artificial intelligence systems to handle complex human communications.

2.4 Recognition of Sarcasm

The task of detecting sarcasm, which involves predicting whether a text contains sarcasm, is a crucial step in sentiment analysis. To date, there have been three milestones in this field: semi-supervised pattern extraction for identifying implicit emotions, supervised identification using hashtag-based labeling, and incorporating context beyond the target text Joshi, Bhattacharyya, and Carman (2018).

2.4.1 Semi-Supervised Pattern Extraction for Implicit Emotions

This stage focuses on using semi-supervised learning methods to identify and extract language patterns that imply hidden emotions. To be more clear, this approach relies on large amounts of unlabeled data and a smaller set of labeled data to train models to recognize underlying negative or sarcastic sentiments. It involves using natural language processing techniques to detect expressions that are positive but have a negative undertone. Features extracted from texts, such as emotional contrasts within sentences, specific vocabulary and phrase, and other linguistic markers.

2.4.2 Supervised Learning Based on Hashtag Labels

With the growing influence of social media platforms like TikTok and Twitter, researchers have begun using tweets with specific ”tags” to automatically create training datasets. For instance, many users append hashtags like “sarcasm” to their sarcastic tweets. This method allows researchers to quickly collect and label large volumes of sarcastic data without the need for manual annotation, avoiding much unnecessary and meaningless work. This not only speeds up the data collection process but also enhances the training efficiency and coverage of sarcasm detection models. The collected data are then used to train classification models, such as Support Vector Machines (SVM) or neural networks, to automatically identify sarcastic content in unlabeled data.

2.4.3 Incorporating Context Beyond the Target Text

This research direction recognizes that many texts can convey multiple meanings depending on the context, and it is often difficult to accurately judge whether a text is sarcastic based solely on the target text itself because sarcasm heavily depends on context. Therefore, this approach focuses on how to incorporate additional information related to the target text into the model. This includes the text's conversational history, even the social-cultural background knowledge related to specific topics, and multi-modal information about user's behavior. For example, understanding the flow of conversation behind a tweet can help the model detect subtle expressions of sarcasm, thereby improving the accuracy of sarcasm detection.

2.5 Key word spotting

There has always been a strong connection between keywords and speech, but it is more reflected in speech recognition. Keyword Spotting (KWS) technology is a technology that identifies specific keywords in an audio streamLópez-Espejo, Tan, Hansen, and Jensen (2022). The system usually includes steps such as speech feature extraction, acoustic modeling and post-processing, including the use of specific model to distinguish keywords and non-keywords. Deep learning methods such as convolutional neural networks (CNN), recurrent neural networks (RNN), time delay neural networks, and deal with open vocabulary keyword spotting tasks through connection time classification (CTC) and sequence to sequence (Seq2Seq) models.

Transformer architecture, as a representative of self-attention models, has been successfully applied in many fields. In KWS, the self-attention mechanism is usually used in combination with convolutional or recurrent encoders. However, the Keyword Transformer (KWT) proposed by Berg, O'Connor, and Cruz (2021) is a completely self-attention-based architecture that can surpass existing state-of-the-art models on multiple tasks without pre-training or additional data. KWT simplifies the model structure by applying self-attention in the temporal domain and achieves 98.6 percent and 97.7 percent accuracy on the 12 and 35 command tasks of the Google Speech Commands dataset, setting new benchmark records.

2.6 Emotion detection by large language models

Emotion detection technologies are extensively applied across various fields, and large language models (LLMs) like GPT-3.5 have shown promising future in many natural language processing (NLP) applications. Carneros-Prado et al. (2023) conducted a comparative analysis of sentiment and emotion classification using GPT-3.5 and IBM Watson (a technology platform based on big data and machine learning) on a dataset containing 30,000 tweets related to the Covid-19 pandemic. The results indicated that, with right and precise prompting, these models could adapt to various NLP tasks beyond their initial training objectives. Despite not being explicitly trained for these tasks, GPT-3.5, with the right contextual prompts, was able to achieve competitive performance against IBM Watson's emotion classification capabilities, particularly good in detecting subtle emotions like sarcasm.

In a study by Boitel, Mohasseb, and Haig (2024), the performance of GPT and BERT models in emotion recognition tasks was compared. Although GPT demonstrated superior performance across multiple tasks, this study questioned its advantage over other models specifically trained for particular functions. The comparison examined the performance of these models on specific tasks and general AI tasks, exploring the potential reasons behind GPT's performance and why it may not outperform specialized models in emotion recognition. The conclusion suggested that newer models like GPT-4 could perform even better than one-task models.

Furthermore, AI in healthcare has seen rapid development and extensive research. A study evaluated the effectiveness of LLM-based GPT-Neo-125M and GPT-2 in three binary text-based mental health classification tasks: stress, depression, and suicide. To our knowledge, this study demonstrated that GPT-2 provided higher overall accuracy, achieving accuracy rates exceeding 0.98 across all test datasets Jain, Goyal, and Sharma (2024). This highlights the potential of LLMs in healthcare applications, specifically in mental health diagnostics. and a letter error rate (LER) of 15.05% when using the large language model.

Emotional dynamics modeling is crucial in conversational emotion recognition. Existing studies mainly use recurrent neural network (RNN) models, but these models fail to fully utilize the latest pre-training strategies and cannot effectively distinguish interlocutor dependencies and emotional influences. Yang and Shen (2021) proposed a series of BERT-based models, using BERT instead of RNN to enrich the tag representation, and using flat structure BERT (F-BERT) to directly connect discourses and hierarchical structure BERT (H-BERT) to distinguish interlocutors. In addition, spatiotemporal structured BERT (ST-BERT) is proposed to capture the emotional influence between interlocutors. Experimental results show that the author's method improves the state-of-the-art baselines by about 5 percent and 10 percent on the conversational emotion recognition benchmarks, respectively.

Acheampong, Nunoo-Mensah, and Chen (2020) conducted a comparative analysis of the performance of pre-trained transformation models such as BERT, RoBERTa, DistilBERT, and XLNet in text emotion recognition. The study used the ISEAR dataset to analyze the performance of each model in identifying seven emotions (anger, disgust, fear, guilt, happiness, sadness, and shame) by adjusting the same hyperparameters. The results show that the RoBERTa performs best in all emotion categories, followed by XLNet, while BERT and DistilBERT perform weaker on some emotion categories. The study shows that while all models perform well in natural language processing tasks, RoBERTa outperforms other models in emotion recognition tasks.

3 Methodology

Overall, I conducted keyword extraction and made adjustments and additions to the variance adaptor. I will describe in detail how I experimented and tested this part. In the keyword extraction section, I wrote a function that utilizes the my own API provided by OpenAI to invoke and use GPT within the model before the synthesis stage. Specifically, the model specifies the GPT type as ”gpt-3.5-turbo,” along with ”temperature” to control creativity and ”n” to control the number of generated responses. To standardize GPT's responses, I first specified the format of GPT's answers in the input command as ”1. /2. /3. ” to pinpoint the three most likely sarcastic keywords in a sentence. Subsequently, I used code to extract and normalize the keywords from the responses, resulting in a Python list, which allows for further uniform processing of the text. The specific processes and examples, see Figure 1 and 2.

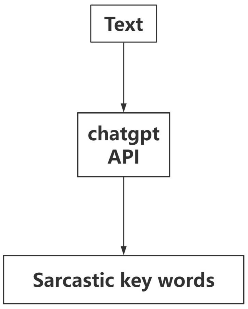


Figure 1: specific process

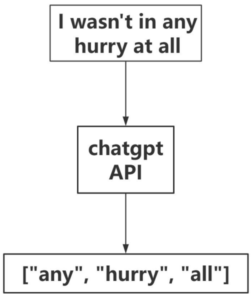


Figure 2: example

Additionally, the original model step involved breaking down the text into words, then using a dictionary and Grapheme-to-Phoneme (G2P) conversion tool to transform each word into a phoneme sequence. The obtained phoneme sequence was then used as part of the model input. In this part, I added a filtering process, which involves extracting the positions of the phonemes of the words in the keyword list from the entire sentence to serve as input for subsequent operations. These keyword lists were then used as inputs to the variance adaptor's “get pitch embedding” and “get energy embedding” sections. Originally, the function was to multiply the entire prediction (generated by the variance predictor) by a specified variable in the command, but now this multiplication only applies to the sequence of these keywords, thereby achieving the effect of altering the pitch and energy of the sarcastic keywords. These parameters were then passed on to relevant parts, such as ”forward”. I then controlled variables, continuously adjusting the pitch and energy coefficients, and the final test results showed that when the pitch control factor was 1.5 and energy was 4, the audio quality was optimal, with no noticeable noise, and the sarcasm was most pronounced. The specific processes can be seen in Figure 3.

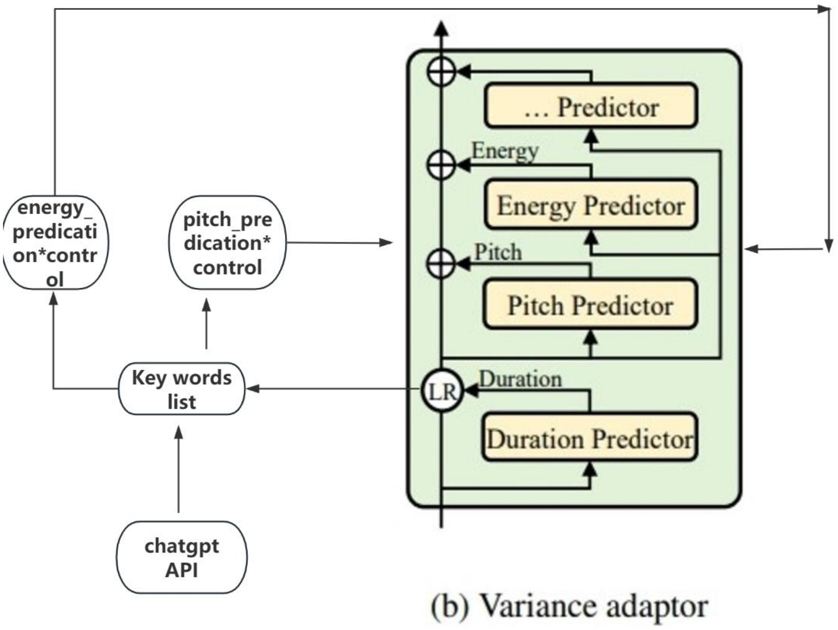


Figure 3: modified process

For the specific generation of sarcastic audio, I used GPT to produce ten sarcastic statements, five of which were textually determinable, such as ”My computer crashed again, exactly what I need right now.” The other five were not determinable based solely on text, such as ”Let’s do some more work, I’m not busy at all.” The purpose of this was to test the synthesized sarcastic audio from these two dimensions separately.

3.1 pilot study

At the initial stage, to test the feasibility of this method, I conducted a pilot feasibility test. I used a command to specify keywords and obtain sequences, similar in content and principle to the above description. Specifically, during the input command, I added a ”–keys” to specify keywords, instead of using GPT to predict them, and I tested with sarcastic sentences. I gathered some examples and had three classmates test them, and the results showed that they had a definite sarcastic effect.

3.2 Evaluation

This experiment uses the survey method to first conduct a horizontal comparison and test the synthesis effect of sarcastic speech sentences at the keyword level, and then conduct a correlation study on different groups of sarcastic utterances.

3.2.1 Participants

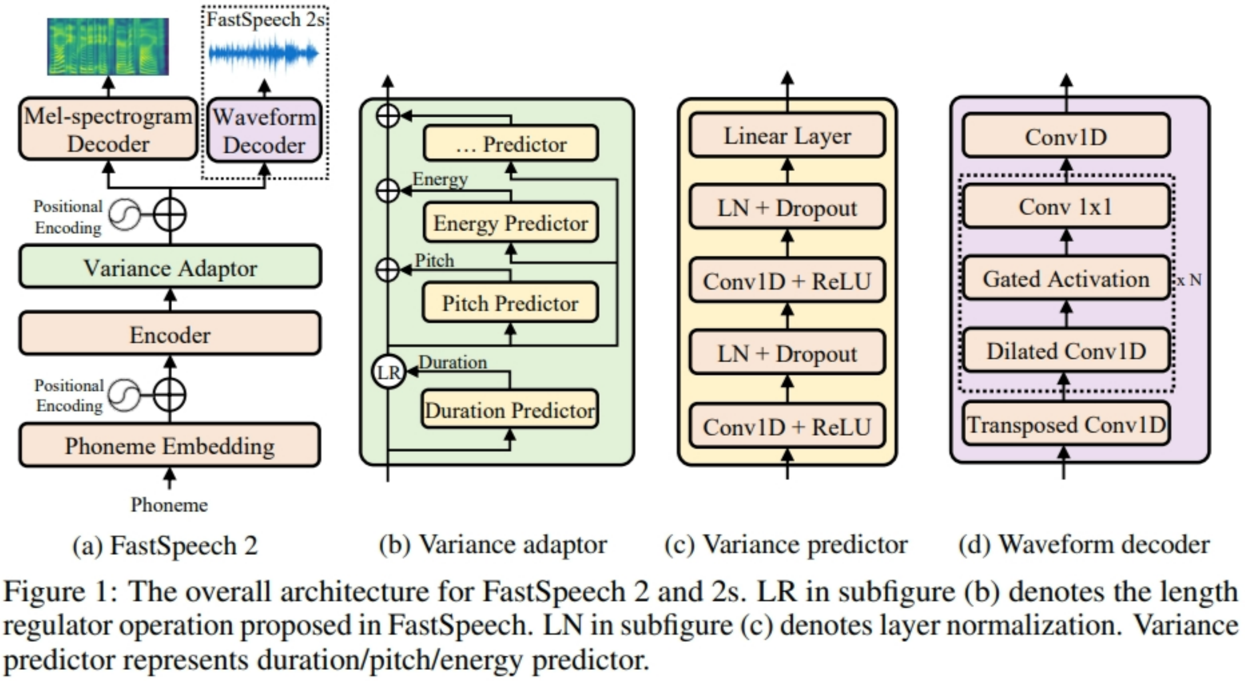
A total of 22 students participated in this survey. They are all classmates and friends, all of whom are native or second language speakers of English and can communicate in English as a language in daily life.

3.2.2 Survey

The Qualitrix website of the University of Groningen was used to design the questionnaire. At the beginning, a GDPR consent form will be provided. The respondent will only conduct subsequent audio testing after choosing to agree. There are ten specific test questions in total. Each question includes five audios, one for adjusting only pitch at key-word level, one for adjusting only energy at key-word level, one for adjusting both pitch and energy at key-word level, one for which no adjustment is made, and two variable controls at whole-sentence level. And the first five questions are sarcastic sentences that can be seen from the text, and the second half are sarcastic sentences that cannot be seen from the text only, but they become sarcastic sentences through the modification of the phonetic level.

4 Experimental Setup

4.1 Fastspeech2



(a) FastSpeech 2 (b) Variance adaptor (c) Variance predictor (d) Waveform decoder

Figure 1: The overall architecture for FastSpeech 2 and 2s. LR in subfigure (b) denotes the length regulator operation proposed in FastSpeech. LN in subfigure (c) denotes layer normalization. Variance predictor represents duration/pitch/energy predictor.

Figure 4: distribution of answers

Fastspeech2 is an advanced speech synthesis model designed based on the transformer architecture and employs a non-autoregressive approach, compared with Tacotron2, it has much more fast synthesising rate. This allows for the parallel generation of speech independent of previous outputs, enhancing synthesis efficiency. Fastspeech2 predicts key acoustic features crucial for speech synthesis through independent network layers, such as duration, energy, and pitch. Accurate prediction of these features is vital for achieving naturally sounding speech, especially in handling complex vocal expressions like sarcasm or emphasis, making it particularly important.

Moreover, the design of Fastspeech2 offers substantial advantages in processing long-text speech synthesis, avoiding the latency and potential error accumulation inherent in traditional autoregressive models. Through its parallel processing mechanism, Fastspeech2 not only speeds up the speech synthesis process but also ensures the coherence and fluency of the generated speech, making the synthetic voice sound more natural and closer to human speech. Furthermore, an enhanced attention mechanism has been incorporated, enabling the model to more accurately simulate human speech patterns in statements with complex intonations and rhythms. For instance, when mimicking the rise or fall in pitch, Fastspeech2 adjusts the tone more naturally, effectively conveying the emotional intent and nuances of the statements.

In FastSpeech2, there are two main parts that are most noteworthy and most relevant to this research. They are the variance predictor and variance adaptor, which work together to adjust and optimize the pitch, energy and duration of the generated speech. To be more clear, in the variance predictor, the phoneme features passed in from the previous transformer network layer are usually used as input. Each predictor outputs a sequence corresponding to the number of input phonemes, representing the predicted feature (pitch, energy, or duration). Then the variance adapter is passed in as a tensor and the pitch or energy of the entire sentence can be adjusted based on control parameters and based on input from external commands. In this model, users can control the energy and pitch of the entire sentence through ”–energy control” and ”–pitch control” on the command line, that is, multiply the phoneme tensor of the entire sentence by a parameter. The above two commands parameters to get the effect of controlling the pitch and energy of a sentence.

4.2 The training process of the model

In this experiment, I used the well-trained checkpoint of FastSpeech2, and trained a total of 900,000 steps. The following content will present the specific steps on how to train this checkpoint.

4.2.1 Database and MFA forced alignment

The English data set uses the LJSpeech database, which contains approximately 24 hours of clear English speech recordings from a single female speaker. Each recording is equipped with precise text annotation. The file format is that each wav file corresponds to a lab file, which is a text annotation. The alignment method used here is MFA forced alignment. This alignment technology analyzes the speech signal using phoneme and acoustic models and matches it with the predetermined text transcription, ensuring that the exact position of each word and phoneme in the audio is correctly marked.

4.2.2 Deal with oov

In speech synthesis, a dictionary is usually used to process the vocabulary involved, but it is a challenge if the words involved are out of vocabulary (OOV). Using the Grapheme-to-Phoneme (G2P) conversion method, unknown vocabulary can be converted from text form to phoneme form. For example, for an unknown word ”apple”, the G2P model converts it into a phoneme sequence, and by adding these phoneme sequences to the vocabulary, the system can correctly pronounce these words. In this way, G2P technology effectively solves the OOV problem and improves the accuracy of the speech synthesis system. This requires us to process the OOV words selected by the model through the G2P model in the first step of the ”process” stage, and then add them to the dictionary again for more refined synthesis and then train the model.

4.3 Generation of sarcastic samples

For the generation of specific sarcastic audio, I used gpt to randomly generate and select ten sarcastic sentences. Five of them can be directly determined from the text, such as ”My computer crashed again, exactly what I need right now.” The other five cannot be determined based only on the text, like ”Let’s do some more work, I’m not busy at all.” The purpose of this is to conduct a more comprehensive test of the synthesized sarcastic audio from these two dimensions respectively.

5 Results

In this part, I will introduce a detailed statistical analysis of the data obtained from the questionnaire. In order to enhance the readability of readers, the five letters ”ABCDE” in different colors in this Figure 5 correspond to specific audios, ”A” Represents the audio that only modifies the keyword energy; ”B” represents the audio that only modifies the keyword pitch; ”C” represents the audio that modifies both energy and pitch at key-word level; ”D” represents the audio that modifies energy and pitch at the whole-sentence level; and ”E” stands for audio without any modification on pitch and energy. The specific table is below.

Letters	**Corresponding audio**
A (blue)	Audio with only the keyword energy modified
B (orange)	Audio with only the keyword pitch modified
C (green)	Audio with both energy and pitch of the keywords modified
D (red)	Audio with energy and pitch of the entire sentence modified
E (purple)	Audio with no modifications

Table 1: Corresponding audio modifications for each letter

5.1 Intuitive analysis

In this bar chart figure 5 of the questionnaire survey, the horizontal axis 1-10 represents the ten questions, while the vertical axis shows the specific number of people for each option. The legend in the upper right corner clearly identifies the color of each option. It is obvious from the chart that option E (indicated in purple), which is the audio without any pitch and energy adjustment, has the lowest sarcastic effect, and in many cases the number of people is even zero. On the contrary, those audios with both pitch and energy adjusted at the keyword level were considered to have the best sarcastic effect in most questions, especially in questions 4, 5, 7, and 10, where their advantages were particularly significant. When pitch and energy adjustments are applied to the entire sentence, this combination also shows a better satirical effect in some questions, but from an overall distribution perspective, this effect is not as strong as adjusting keywords. The remaining two options only adjust the pitch or energy of the keyword, respectively. Next, I will use Anova, which is analysis of variance and comparative analysis of different categories of sarcastic sentences.

5.2 Anova analysis

ANOVA, also known as analysis of variance, is a statistical method that determines whether there is a significant difference in the means of multiple groups. The purpose of this method is to explore whether the influence of independent variables on dependent variables is statistically significant. It works by decomposing the total variability of the data, specifically divided into Within-group variability and Between-group variability, which reflect the deviation of individual data points from the

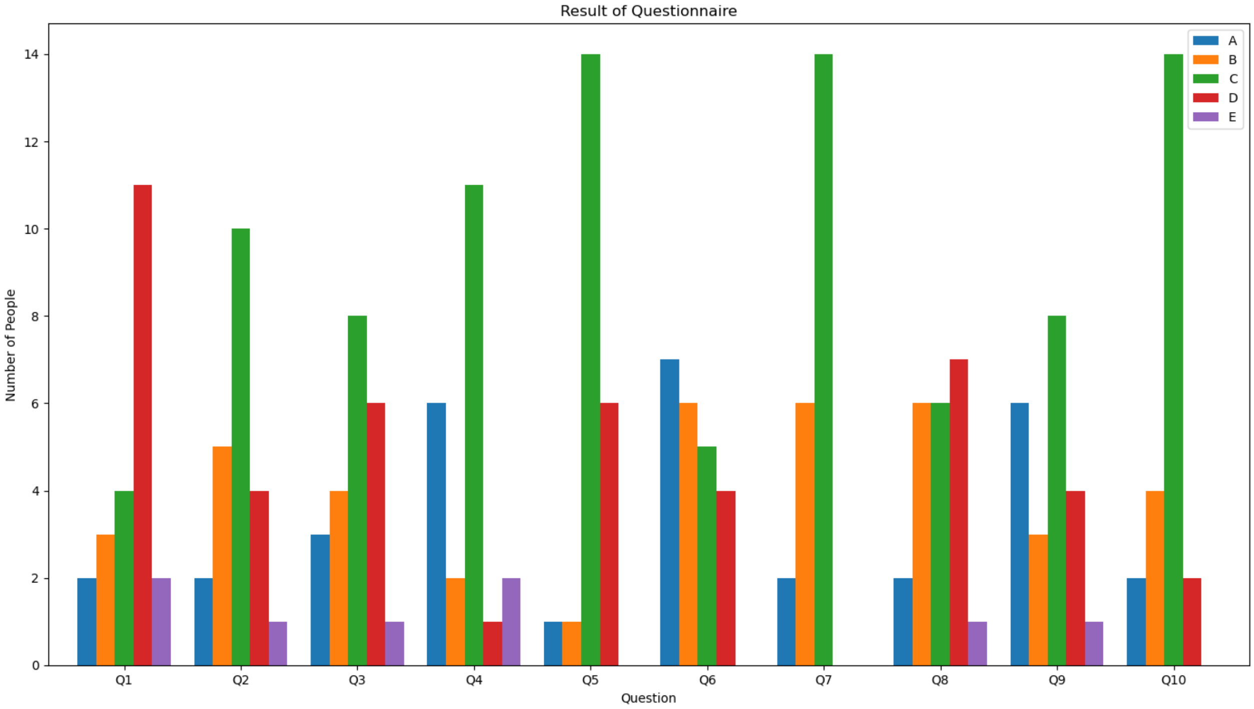


Figure 5: distribution of answers

group mean within the same group and the mean difference between different groups respectively. If this difference is greater than random error may indicate that the treatment effect is significant. The F value is a statistic used to measure the ratio of Within-group variability to Between-group variability in ANOVA. When the F value is significantly greater than 1, it indicates that the treatment or conditions in the experiment have a significant impact on the results, that is, the variation between groups is significantly greater than the variation within the group. In this case, an increase in the F-statistic is usually accompanied by a decrease in the p-value, which is used to determine whether the difference is statistically significant.

The questionnaire results show that the F-value is 14.82. A high F-value indicates that the between-group variability is significantly greater than random error (Between-group variability), which means that there are significant statistical differences in the scores of different options on different questions. A very small p-value (8.45e-08 is much less than 0.05), that is, the situation where all options have the same score in all questions does not exist at all. In other words, there are significant differences in the scores of options on different questions. Combined with the proportions in the figure, it can be considered that adjusting the pitch and energy of sarcastic sentences from the keyword level is indeed effective.

5.3 Comparison between groups

As introduced before, the structure of the ten questions in the questionnaire is that the sarcastic sentences in the first five questions are sentences that can be judged to be sarcastic utterances from the text, and the last five sentences are sarcastic sentences that cannot be judged from the text. In this part, I will use Pearson correlation coefficient and p-value to test the correlation of the five options

(Option)	(Correlation)	p-value
A	0.459	0.436
B	0.791	0.111
C	0.779	0.120
D	0.179	0.774
E	0.612	0.272

Table 2: Pearson correlation coefficient and p-value

in questions 1-5 and 6-10, so as to further explore the specific situation of the two sentences using this research method. If they are related If the coefficient is close to 1, there is a strong correlation, and if it is close to 0, there is no correlation. The results obtained are shown in the table below.

Here I analyze each option individually and summarize. Audio with energy adjusted only at the keyword level shows a moderate positive correlation (0.459) between the two parts, but its p-value is 0.436, indicating that this correlation is not statistically significant. The audio with pitch adjusted only at the keyword level and the audio with pitch and energy applied at the keywords level have higher correlations (0.791 and 0.779) respectively, but their p-values (0.111 and 0.120) also more than the conventional ones. Significance test threshold (0.05), therefore these high correlation results cannot be considered statistically significant, which may mean that although strong correlations are observed, these results are insufficient due to insufficient sample size or large variability. The correlation is not sufficient to reach statistical significance. When the two control the entire sentence, the correlation is the lowest (0.179), and its p-value is 0.774, showing a very weak positive correlation and not statistically significant at all. The correlation between the two parts of the audio without controlling any pitch and energy is 0.612, which is at a medium level, but its p-value is 0.272, which also indicates that it is not statistically significant. To conclude, although some options showed strong correlations, the p-values for all options did not fall below the conventional significance threshold of 0.05, indicating that regardless of the strength of the correlation, from a statistical perspective, it is not directly It was determined that there is a significant linear correlation between questions 1 to 5 and questions 6 to 10. In other words, there is no difference in the listener's reaction in this experiment between sentences that can be judged to be ironic in the text and those that cannot be judged, that is, the experimental method has a similar effect on the two kinds of ironic sentences.

In this survey, through detailed statistical analysis, it can be seen that changes in pitch and energy at the keyword level will have the most positive impact on the satire effect. The results of ANOVA analysis of variance show that there are significant differences in scores between different options on different questions, which proves that the treatment method does have an impact on the sarcastic effect. However, the results of the Pearson correlation test showed that this effect was not always statistically significant.

6 Discussion

By adjusting the acoustic features of sarcastic speech in the Fastspeech2 model and adjusting pitch and energy at the keyword level, this method achieves more precise sarcastic emotion synthesis. Experimental results support this hypothesis, showing that synthesized sarcastic speech significantly improves effect when pitch and energy are adjusted at the keyword level. Specifically, pitch and energy adjustments to sarcastic keywords significantly enhanced listeners' perception of sarcasm more than whole-sentence adjustments or no adjustments. Audio samples that adjust the pitch and energy of keywords always have the highest ratings for the effectiveness of sarcastic tone, especially in questions 4, 5, 7, and 10 of the questionnaire, where keyword-level adjustments are significantly better than other methods. ANOVA analysis further confirmed the statistical significance of these results, with an F value of 14.82 and a p value significantly less than 0.05, indicating that the differences in the perception of irony effects between different treatment methods are not generated randomly.

Although whole-sentence adjustments also showed some improvement in sarcasm perception, the overall effect was not as good as keyword-level adjustments, which is consistent with the hypothesis that sarcastic tone is usually conveyed through specific keywords rather than evenly distributed throughout the sentence. Through testing of sentences that can be directly judged as sarcastic from the text and sentences that cannot be judged from the text alone, the results show that the effect of sarcastic sentences adjusted at the keyword level is more obvious in both types of sentences. Although some options show higher correlations, the p-values of all options are not lower than the traditional significance threshold of 0.05, which indicates that there is no statistically significant difference in audience responses to visible versus invisible sarcastic sentences in the text. There is a significant difference, that is, this experimental method has similar effects on the two types of sarcastic sentences. The effectiveness of this method was verified through experimental settings and questionnaire survey results. Statistical analysis of questionnaire data shows that adjusting the pitch and energy of keywords can significantly improve the satirical effect of the audio compared to making no adjustments or just adjusting the entire sentence, especially when the satirical intent cannot be individually identified from the text. This is especially true under the circumstances. Although this study has achieved certain results, there are also some limitations, such as the lack of rigorous scientific comparison of specific acoustic feature adjustment values and relying only on subjective auditory perception for evaluation. Future research can further explore on this basis to optimize the synthesis and recognition technology of sarcastic tone through more precise experimental design and larger-scale data analysis, which may also include the adaptability of sarcastic expressions in different cultural and linguistic backgrounds. Research.

7 Conclusion

In this section, I will summarize the entire process of this study, and will explain the shortcomings of this study and what further research can be done in the future.

7.1 Summary

In this paper, by adjusting the acoustic characteristics of sarcastic speech and adjusting pitch and energy at the keyword level through the Fastspeech2 model, a more accurate synthesis of sarcastic emotions is achieved. This method shows obvious advantages compared with the traditional whole sentence adjustment method, allowing the sarcastic keywords in a sentence to be more effectively emphasized to achieve a more natural and effective conveying of the sarcastic tone, thus improving the quality of the sarcastic statement. Recognition rate. Through the experimental setting and the results of the questionnaire survey, this study verified the effectiveness of this method. Statistical analysis of questionnaire data shows that adjusting the pitch and energy of keywords can significantly improve the sarcastic effect of the audio compared to making no adjustments or adjusting the whole sentence alone, as well as adjusting one of the two alone, especially when it cannot be recognized from the text alone. This is even more obvious in the case of satirical intent.

In addition, the large-scale language models used in the study (such as GPT-3.5) have shown strong potential in sentiment analysis and sarcasm detection, illustrating the application prospects of deep learning technology in understanding and generating complex language expressions, such as sarcastic tone. Although this study has achieved certain results, there are also limitations, and I will illustrate more in next part.

7.2 Limitation and future work

I did not have many opportunities to further expand the research because of the time shortage, but I am here to provide three main future research in this section. First of all, in this study, only the API of the large model gpt was called to predict the sarcastic keywords of a sentence. Future research can be based on the existing sarcastic data set and manually mark the keywords in the sarcastic discourse as a In the new dimension, a model such as SVM or neural network is then used to train a model that selects sarcastic keywords from sarcastic utterances. This kind of functional model may have better performance. Morever, this experiment does not strictly compare or divide the added value of specific pitch and energy, but only makes judgments based on the sense of hearing. Therefore, through this aspect, future research can test more coefficients and make larger and more rigorous questionnaires to conduct a more comprehensive design of keyword-level sarcastic discourse synthesis. Finally, according to research and judgment, the energy and pitch manipulation of the FastSpeech2 model are not related to multiples in the auditory sense, which affects further emotion synthesis and control to a certain extent. Therefore, in future research, we can use the tensor to Processing and control of normalization can achieve a multiple increase in human hearing.

References

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2020). Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th international computer conference on wavelet active media technology and information processing (iccwamtip)* (pp. 117–121). IEEE. Retrieved from `https://doi.org/10.1109/ICCWAMTIP51612.2020.9317379` doi: 10.1109/ICCWAMTIP51612.2020.9317379

Allen, J., Hunnicutt, S., Carlson, R., & Granstrom, B. (1979). Mitalk-79: The 1979 mit text-to-speech system. *Journal of the Acoustical Society of America*, *65*(S1), S130. Retrieved from `https://doi.org/10.1121/1.2017051` doi: 10.1121/1.2017051

Arik, S. O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., ... Shoeybi, M. (2017). Deep voice: Real-time neural text-to-speech. *arXiv preprint arXiv:1702.07825*. Retrieved from `https://arxiv.org/abs/1702.07825`

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor: International Journal of Humor Research*, *16*(2), 243–260. Retrieved from `https://doi.org/10.1515/humr.2003.012` doi: 10.1515/humr.2003.012

Berg, A., O'Connor, M., & Cruz, M. T. (2021). Keyword transformer: A self-attention model for keyword spotting. *arXiv preprint arXiv:2104.00769*. doi: 10.48550/arXiv.2104.00769

Boitel, E., Mohasseb, A., & Haig, E. (2024). A comparative analysis of gpt-3 and bert models for text-based emotion recognition: Performance, efficiency, and robustness. In N. Naik, P. Jenkins, P. Grace, L. Yang, & S. Prajapat (Eds.), *Advances in computational intelligence systems: Ukci 2023, lecture notes in networks and systems* (Vol. 1453, p. xx-xx). Springer, Cham. Retrieved from `https://doi.org/10.1007/978-3-031-47508-5_44` doi: 10.1007/978-3-031-47508-5_44

Camp, E. (2012). Sarcasm, pretense, and the semantics/ pragmatics distinction. *Noûs*, *46*(4), 587-634. doi: 10.1111/j.1468-0068.2010.00706.x

Carneros-Prado, D., Villa, L., Johnson, E., Dobrescu, C. C., Barragán, A., & García-Martínez, B. (2023). Comparative study of large language models as emotion and sentiment analysis systems: A case-specific analysis of gpt vs. ibm watson. In J. Bravo & G. Urzáiz (Eds.), *Proceedings of the 15th international conference on ubiquitous computing & ambient intelligence (ucami 2023), lecture notes in networks and systems* (Vol. 842, p. 367-379). Springer, Cham. Retrieved from `https://doi.org/10.1007/978-3-031-48642-5_22` doi: 10.1007/978-3-031-48642-5_22

Cheang, H. S., & Pell, M. D. (2008). The sound of sarcasm. *Speech Communication*, *50*(5), 366-381. Retrieved from `https://doi.org/10.1016/j.specom.2007.11.003` doi: 10.1016/j.specom.2007.11.003

Clark, R. A., Richmond, K., & King, S. (2004). Festival 2: Build your own general purpose unit selection speech synthesiser. In *Festival 2.* CSTR, The University of Edinburgh. Retrieved from `https://www.academia.edu/18596920/Festival_2_build_your_own_general_purpose_unit_selection_speech_synthesiser`

Jain, B., Goyal, G., & Sharma, M. (2024). Evaluating emotional detection & classification capabilities of gpt-2 & gpt-neo using textual data. In *Proceedings of the 14th international conference on cloud computing, data science & engineering (confluence)* (p. 12-18). Noida, India. Retrieved from `https://doi.org/10.1109/Confluence60223.2024.10463396` doi: 10.1109/Confluence60223.2024.10463396

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2018). Automatic sarcasm detection: A survey. *ACM Computing Surveys*, *50*(5), 1-22. Retrieved from `https://doi.org/10.1145/3124420` doi: 10.1145/3124420

Kreuz, R. J., & Roberts, R. M. (1993). The empirical study of figurative language in literature. *Poetics*, *22*(1-2), 151-169. Retrieved from `https://doi.org/10.1016/0304-422X(93)90026-D` doi: 10.1016/0304-422X(93)90026-D

López-Espejo, I., Tan, Z.-H., Hansen, J. H. L., & Jensen, J. (2022). Deep spoken keyword spotting: An overview. *IEEE Access*, *10*, 4169-4199. doi: 10.1109/ACCESS.2021.3139508

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5–6), 453-467. Retrieved from `https://www.sciencedirect.com/science/article/pii/016763939090021Z` doi: 10.1016/0167-6393(90)90021-Z

Rao, R., Ye, T., & Butera, B. (2022). The prosodic expression of sarcasm vs. sincerity by heritage speakers of spanish. *Languages*, *7*(1), 17. Retrieved from `https://doi.org/10.3390/languages7010017` doi: 10.3390/languages7010017

Scharrer, L., & Christmann, U. (2011). Voice modulations in german ironic speech. *Language and Speech*, *54*(4), 435-465. Retrieved from `https://doi.org/10.1177/0023830911402608` doi: 10.1177/0023830911402608

Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., ... Saurous, R. A. (2018). Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. *arXiv preprint arXiv:1803.09047*. Retrieved from `https://doi.org/10.48550/arXiv.1803.09047`

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). Speech parameter generation algorithms for hmm-based speech synthesis. In *2000 ieee international conference on acoustics, speech, and signal processing. proceedings (cat. no.00ch37100)* (Vol. 3, pp. 1315–1318). Retrieved from `https://ieeexplore.ieee.org/document/859907` doi: 10.1109/ICASSP.2000.859907

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*. Retrieved from `https://doi.org/10.48550/arXiv.1609.03499`

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*. Retrieved from `https://doi.org/10.48550/arXiv.1703.10135`

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R., Battenberg, E., Shor, J., ... Saurous, R. A. (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017*. Retrieved from `https://doi.org/10.48550/arXiv.1803.09017`

Yang, H., & Shen, J. (2021). Emotion dynamics modeling via bert. *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8. Retrieved from `https://doi.org/10.1109/IJCNN52387.2021.9533860` doi: 10.1109/IJCNN52387.2021.9533860

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. In *Eurospeech.* Retrieved from `https://api.semanticscholar.org/CorpusID:8037054`

Zhang, Y.-J., Pan, S., He, L., & Ling, Z.-H. (2018). Learning latent representations for style control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1812.04342*. Retrieved from https://doi.org/10.48550/arXiv.1812.04342

Appendices

A Text

As shown in Table-4, all ten sentences are numbered Q1 to Q10. The first five sentences cannot be predicted to be sarcastic from the text alone, while the last five sentences can be predicted from the text alone.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
A	2	2	3	6	1	1	7	2	6	2
B	3	5	4	2	1	1	6	6	3	4
C	4	10	8	11	11	14	5	14	6	14
D	11	4	6	1	6	6	4	7	8	2
E	2	1	1	2	0	0	0	1	1	0

Table 3: Survey Results

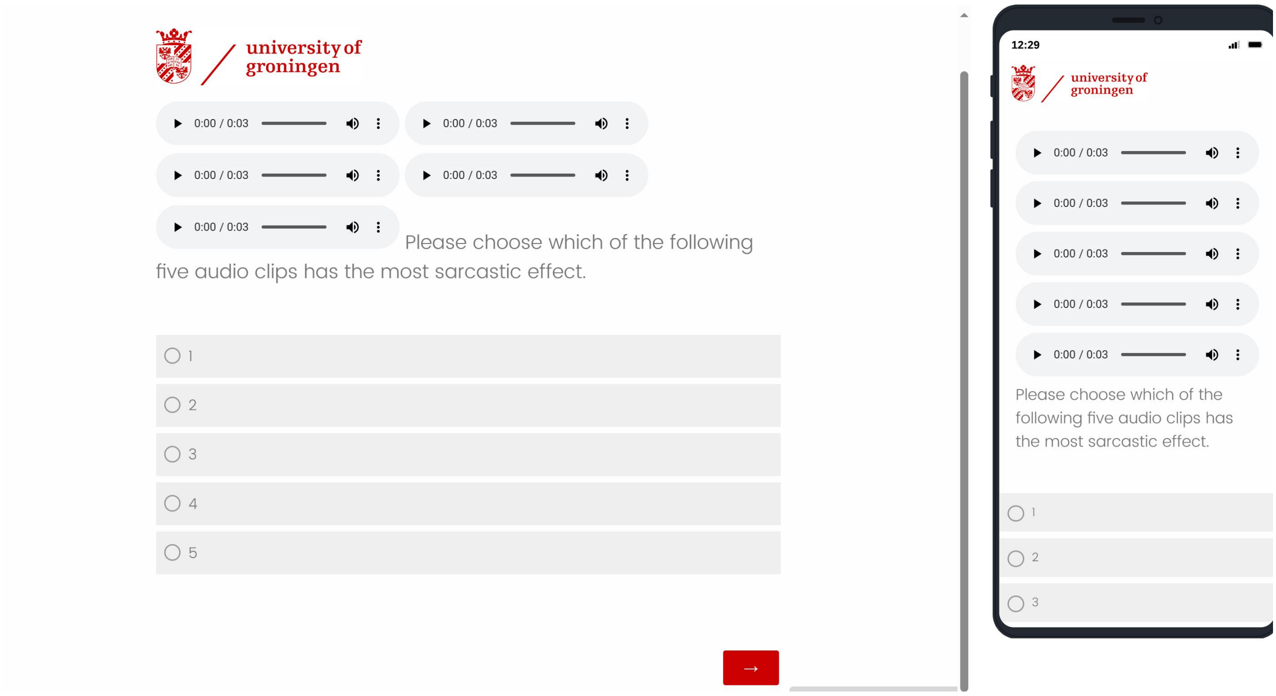
Question Number	Question Text
Q1	Let’s do some more work, I’m not busy at all.
Q2	I can’t wait to start cleaning
Q3	Take your time, I wasn’t in any harry at all.
Q4	Oh, your new haircut is just, great!
Q5	Watching paint dry is my favourite thing to do
Q6	You finally responded to my message, I just waited a century.
Q7	My computer crushed again, exactly what I need right now.
Q8	They raised their keyboards, thinking they were kings.
Q9	Late again, you are such a master of time management.
Q10	Without beautiful facial features, who would have the time feel you heart.

Table 4: text

B Questionnaire

Table 3 shows the distribution of the results of the questionnaire survey. The red mark represents the option with the most people. ABCDE respectively represent adjusting pitch only in the keyword dimension, adjusting energy only in the keyword dimension, adjusting pitch and energy in the keyword dimension, adjusting pitch and energy in the whole sentence dimension, and adjusting pitch and energy at the whole sentence level.

Figure 6 shows the specific format of the online questionnaire. The left side is the computer version, and the right side is the mobile version.



Please choose which of the following five audio clips has the most sarcastic effect.

○ 1

○ 2

○ 3

○ 4

○ 5

Please choose which of the following five audio clips has the most sarcastic effect.

○ 1

○ 2

○ 3

Figure 6: Questionnaire