university of groningen

campus fryslân

Sarcastic speech synthesis in Dutch using voice-transformation

Tessa Zwart

university of groningen

campus fryslân

University of Groningen

Sarcastic speech synthesis in Dutch using voice-transformation

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
Prof. Dr. M. Coler (Voice Technology, University of Groningen)
and
X. Gao (Voice Technology, University of Groningen)

Tessa Zwart (s3683850)

July 15, 2023

Contents

Acknowledgments

First of all, many thanks to my supervisors Matt Coler and Xiyuan Gao for supervising this project and helping me when needed. Special thanks to Phat Do, who helped me a lot with the technical part of the study. Without them, I do not think I would have managed to finish this study.

Abstract

This research delves into the creation of a voice that sounds sarcastic and examines the recognition of sarcasm in synthetic speech. Sarcasm, known for its ability to convey meaning beyond literal interpretation, plays a significant role in our everyday interactions. Our main focus lies in identifying the acoustic features relevant to Dutch sarcasm, utilizing the FastSpeech2 model for synthesizing and manipulating speech. To evaluate the synthesized speech, a survey was conducted, which revealed that the sarcastic synthetic voice has a limited recognition accuracy of up to 35%. This suggests that the factors responsible for conveying sarcasm are not collectively perceived as sarcastic by listeners. Nevertheless, when the manipulation of the speech files is doubled, the accuracy rises to as high as 43%, suggesting that increasing the degree of manipulation leads to a higher likelihood of recognizing the speech as sarcastic. Additionally, we investigated three sentence types (tag-questions, declaratives, and wh-exclamatives) and found no significant difference in recognizing sarcasm based on sentence type. This implies that sentence structure does not influence people's perception of sarcasm. Potential explanations for the low recognition rates include the need for further modifications to synthetic speech or the incorporation of facial expressions to enhance sarcasm recognition.

1 Introduction

Sarcasm, an expressive form of communication, is often met with varying degrees of comprehension among individuals. While some struggle to understand its intended meaning, others effortlessly command attention through its usage. Sarcasm manifests itself when a neighbour remarks, 'What great weather!' when it is raining outside, or a friend says, 'This is a delicious meal you made!' when at the same time feeding it to their dog. In essence, sarcasm surpasses the literal interpretation by not merely conveying the opposite sentiment but by intensifying it. The meal, for example, is not 'not delicious', but it is actually 'awful'.

Sarcasm synthesis is both a very relevant and under-researched area within the field of speech technology. Work on this topic stands to provide advancements in expressive communication. While previous research has primarily focused on sarcasm recognition and the analysis of speech features associated with sarcasm, there remains a notable gap in the area of sarcasm synthesis, particularly in the Dutch language. By synthesizing sarcastic speech, we open up a valuable avenue for investigating the distinct acoustic properties associated with sarcastic expressions. This exploration not only enables scientists to gain a deeper understanding of the nuances inherent in this expressive form but also offers potential benefits to individuals who struggle with sarcasm comprehension.

To fill in the gap of current research in terms of sarcastic speech synthesis, we aim to synthesize voices and subsequently manipulate the synthetic voice into a voice that is recognized as sarcastic. We propose that by manipulating the acoustic features that are significant for sarcasm, listeners will be able to identify sarcasm in the modified synthetic speech. This research aims to advance the field of sarcastic speech synthesis by delving into the acoustic properties associated with sarcasm and developing suitable algorithms for voice synthesis and emotional manipulation. Through these endeavours, we aim to expand the existing knowledge base and contribute to the progress of sarcasm synthesis in speech technology.

This thesis is structured as follows: In Section 2, a comprehensive literature review is presented that explores the scientific foundation of this research, including the essence of sarcasm, factors that are relevant to sarcasm production across languages, as well as the recognition of sarcasm and voice transformation techniques. Furthermore, the section covers previous work concerning the FastSpeech2 model, which holds a significant role in this study. Section 3 details the proposed methodology I employed to conduct this study, followed by the experiment details in Section 4 and the evaluation strategy in Section 5. Subsequently, an in-depth presentation of the results is given in Section 6, and Section 7 encompasses a thorough discussion of my findings, including any limitations encountered throughout the study. Finally, in Section 8, I provide a conclusive summary of our study's key insights and their implications.

2 Literature review

The literature review consists of four subsections. First, we delve into the use of sarcasm (Section 2.1), exploring its purposes and significance, thereby establishing the relevance of this study. Next, we start on a cross-linguistic comparison (Section 2.2), uncovering how different languages employ distinct acoustic cues for sarcasm production. Then, we transition to the challenges of sarcasm recognition (Section 2.3), highlighting its complexity and the potential benefits of investigating synthetic speech. Lastly, we delve into text-to-speech models (Section 2.4), with a specific focus on the FastSpeech2 model, which will be utilized in our research. By examining these interconnected subsections, we lay the groundwork for a comprehensive understanding of sarcasm synthesis and its implications.

The research exploration commenced with an exploration of papers concerning individuals who face difficulties in understanding sarcasm, particularly sourced from Google Scholar[1]. This initial investigation paved the way for a deeper exploration of the problem and the quest for potential solutions in Section 2.3. Subsequently, the research expanded to unravel the underlying reasons behind the use of sarcasm, starting with papers providing descriptions of sarcasm and progressing towards more recent studies in Section 2.1. For an understanding of sarcasm in different languages, the Dutch paper served as a starting point, leading to the discovery of additional papers in various languages in Section 2.2. Lastly, the exploration of voice manipulation techniques led us to the FastSpeech2 model, which is elucidated in papers related to the model, as detailed in Section 2.4.

2.1 The use of sarcasm

Understanding sarcasm can be challenging, as its meaning deviates from the literal interpretation of words spoken (see Section 2.3). Nevertheless, sarcasm is frequently used in everyday conversations, according to Gibbs (2000), who conducted a thorough investigation of sarcasm usage by reviewing sixty-two 10-minute conversations. The research findings revealed that sarcasm was employed in 8% of all conversational turns. In the recorded conversations, an average of 4.7 instances of sarcasm were documented, indicating that sarcasm was utilized approximately every 2 minutes in everyday language. These findings underscore its frequent occurrence in daily life.

However, the question remains: why do we continue to use sarcasm even when its meaning is not always fully understood? There are multiple explanations for the human use of sarcasm. One possible explanation is its function in reducing confrontation. According to Filik et al. (2016), sarcasm can mitigate the emotional impact of both criticism and praise. Their research demonstrated that sarcastic criticism was perceived as less negative compared to its literal counterpart. Similarly, positive comments expressed sarcastically were found to be less positive than their literal versions. Dews et al. (1995) also found similar results in their study on sarcastic criticism. Sarcastic criticism was rated as less insulting compared to literal criticism, while sarcastic compliments were perceived as more insulting. Furthermore, they observed that the use of sarcasm provided a protective shield for both the speaker and the addressee. It made criticism less aggressive and afforded the speaker a greater sense of control. Interestingly, the speaker-addresser relationship was found to be less affected when sarcastic criticism was employed compared to literal criticism. Therefore, sarcasm can serve as a means to indirectly express negative emotions while being less confrontational and aggressive towards the addressee.

Another explanation is the use of sarcasm in a humorous way. For instance, television shows often incorporate sarcasm to elicit laughter from the audience. Castro et al. (2019) even created a multimodal data set of TV show sarcasm, mostly featuring sarcasm from the shows Friends and The Big

[1] Website of Google Scholar: `https://scholar.google.com`

Bang Theory, containing the text of the sarcasm comments, the facial expressions, and the utterance itself. Moreover, Dews et al. (1995) discovered in one of their experiments that sarcastic remarks were rated funnier than non-sarcastic remarks. However, Dynel (2014) stated in their paper that sarcastic remarks are never only funny; sarcastic remarks always carry something serious and only sometimes contain humour.

However, variations in the frequency of sarcasm usage across different regions have been observed. Dress et al. (2008) conducted a study comparing the use of sarcasm among students in the northern region of the United States, the State University in New York, to students of the University of Memphis that live more in the south. The students of their 'Northern sample' mostly came from cities around New York, including Syracuse, Albany, and Rochester. The 'Southern sample' consisted primarily of students from Memphis and its suburbs in western Tennessee. The results of their research indicate that the 'Northern sample' used sarcasm more often than the 'Southern sample'. Additionally, the students in the northern region perceived sarcasm as funnier in comparison to their southern counterparts.

In related research, the examination of comic-style markers and the dark triad (consisting of Machiavellianism, psychopathy, and narcissism) was conducted (Dionigi et al., 2022). Machiavellianism is a personality whereby the person is very manipulative and indifferent to morality, and narcissism is characterized by excessive focus on self and a great sense of self-importance. Lastly, psychopathy is characterized by antisocial behaviour and a lack of empathy. In their research, they found that psychopathy is the strongest positive correlated to sarcasm, and they argue that they do this to lower other statuses without considering their feelings. On the other side, using sarcasm also seem to improve the creativity of people and thus, creative people use sarcasm more often (Huang et al., 2015). The researchers highlight in their study that sarcasm can boost creativity for both the speaker and the addressee, as it necessitates abstract thinking to understand sarcasm. Hence, the utilization of sarcasm appears to be influenced not only by regional factors but also by individual personality traits.

2.2 Cues of sarcasm in speech

It is interesting that sarcastic utterances are produced differently across languages. Numerous studies have been conducted worldwide, predominantly focusing on comparisons of pitch, duration, intensity, and vocal noise. For instance, sarcasm in English is characterized by a lower mean fundamental frequency (F0), a narrower F0 range, and a longer duration of the utterance, according to Cheang and Pell (2008). However, their study did not extensively explore vocal noise and intensity levels to draw definitive conclusions about these features. Nevertheless, Rockwell (2000) supports these findings and adds that the intensity is higher for sarcastic utterances compared to sincere ones.

When examining the distinctions between sarcastic speech and sincere speech in German, research indicates that German speakers exhibit similar voice modifications to their English counterparts (Niebuhr, 2014). Like English speakers, Germans tend to lower the mean F0, reduce the F0 range, and lengthen the duration of the utterance when being sarcastic. However, one notable difference is that Germans lower the intensity of their voice during sarcastic speech, whereas English speakers tend to raise it. Niebuhr (2014) also investigated the voice quality and found that sarcastic utterances were produced with more breath and a tenser voice than sincere utterances.

In addition, Mexican Spanish speakers use a lower mean F0 and a lower duration when making a sarcastic comment (Rao, 2013). The intensity level is also lowered in Mexican Spanish, although this effect was only shown for male and not for female speakers. Similarly, a narrower F0 range was identified among male speakers, whereas no significant difference was found for female speakers.

However, not all languages distinguish a sarcastic from a sincere speech by decreasing the mean

F0. For example, Italian speakers higher the mean F0 rather than decrease it (Anolli et al., 2002). Next to that, like English speakers, they reduce the F0 range and higher the intensity. An interesting observation regarding Italian speakers is that there appears to be no significant difference in the duration of the utterance, which contrasts with the trend of longer durations in sarcastic utterances observed across other languages. However, Anolli et al. (2002) found a difference in the rate of articulation, with sarcastic speech characterized by a longer rate of articulation compared to sincere speech. They also experimented with variations in the number of pauses, pauses length, and duration of spoken segments, but none of those seemed to be significantly different.

Next to Italian speakers, Cantonese speakers higher the mean F0 and reduce the F0 range (Cheang and Pell, 2009). Although Cheang and Pell (2009) did not specifically investigate intensity differences, they did find that the duration of sarcastic comments is longer than that of sincere comments, while the voice quality is higher. The measurement of voice quality was conducted using the Harmonics-to-Noise Ratio (HNR), where a higher HNR indicates a more harmonic voice, while a lower HNR suggests a more asthenic voice (Murphy and Akande, 2005). As can be noticed, this is different than in German, where speakers seem to lower the voice quality (Niebuhr, 2014).

Similarly, French speakers employ a distinct prosodic pattern in sarcastic speech, characterized by an increase in the mean fundamental frequency (F0) and an extension of the F0 range (Lœvenbruck et al., 2013). As expected, they also make the utterances longer when speaking sarcastically compared to sincerely. However, the paper by Lœvenbruck et al. (2013) does not provide specific information regarding voice quality and intensity in the context of sarcasm.

Upon delving into various languages, we will refocus our attention on the Dutch language. Jansen et al. (2020) examined the features that seem to be important for sarcasm in Dutch. Participants of the study made an imaginary telephone conversation with a friend and were asked to react to this imaginary friend in either a sincere or a sarcastic manner. The available response options included tag-questions (e.g., ‘Your sister is really sweet, isn’t she?’), wh-exclamatives (e.g., ‘What an amazing result!’), and declaratives (e.g., ‘He is absolutely hilarious’).

In their study, containing 10 male and 10 female participants, Jansen et al. (2020) found that the mean duration is longer, the mean intensity is lower, and there is less vocal noise (higher HNR) when a sentence is sarcastically said than when this sentence is sincerely spoken. Moreover, the mean F0 and the F0 range are different, but only for female speakers. Specifically, the mean F0 is higher for female speakers for the tag-questions (232 Hz for sarcastic utterances vs 227 Hz for sincere utterances) and for wh-exclamatives (233 Hz for sarcastic vs 262 Hz for sincere). Additionally, the F0 range for female speakers is bigger when producing sarcasm (14.5 semitones in sarcastic speech vs 12.7 semitones in sincere speech).

Also, the mean duration is consistently longer across all different sentences. Declarative sentences show the biggest difference (1516 ms for sarcastic vs 1317 ms for sincere speech), followed by wh-exclamatives (1397 ms vs 1207 ms) and tag-questions (1424 ms vs 1338 ms). However, the overall differences are larger for males (1449 ms for sarcastic vs 1275 ms for sincere speech) than for females (1442 ms vs 1299 ms).

Continuing with the mean intensity, this is lower in all conditions. Wh-exclamations show the biggest differences between sarcastic utterances (60.0 dB) compared to sincere utterances (61.2 dB). Also, the declarative sentences (59.5 db for sarcastic vs 60.1 dB for sincere utterances) and tag-questions (60.6 dB vs 61.0 dB) show differences in intensity.

Finally, Jansen et al. (2020) found a difference in voice quality, as a higher HNR was measured (13.3 dB in sarcastic speech vs 13.0 dB in sincere speech).

An overview of the finding above is given in Table 1. The features that were not investigated in the research papers that are discussed are indicated as N.E. (not examined).

Language	Pitch (F0)		Duration	Intensity	Voice quality
	Mean	Range			
Dutch[1]	Higher (female)	Bigger (female)	Longer	Lower	Higher
English[2]	Lower	Smaller	Longer	Higher	*N.E.*
German[3]	Lower	Smaller	Longer	Lower	Lower
(Mexican) Spanish[4]	Lower	Smaller (male)	Longer	Lower (male)	*N.E.*
Italian[5]	Higher	Smaller	-	Higher	*N.E.*
Cantonese[6]	Higher	Smaller	Longer	*N.E.*	Higher
French[7]	Higher	Bigger	Longer	*N.E.*	*N.E.*

Table 1: Sarcastic speech compared to sincere speech across different languages. The 'N.E.' notation signifies that the specific feature was not examined in the respective research paper, and therefore, no differences were reported. When gender is indicated within brackets, it denotes that the observed difference was specific to that gender. The information is found in the following papers: 1. Jansen et al. (2020), 2. Cheang and Pell (2008), 3. Niebuhr (2014), 4. Rao (2013), 5. Anolli et al. (2002), 6. Cheang and Pell (2009), 7. Lœvenbruck et al. (2013).

2.3 The recognition of sarcasm

As sarcasm differs in languages, it is hard for second language (L2) speakers to learn sarcasm in a different language (Dolan, 2015). Dolan (2015) examined this by looking at the difference between native English speakers and L2 speakers of English in interpreting sarcastic cues. Their results indicate that native English speakers displayed a better understanding of sarcastic comments compared to L2 speakers of English. Also, the duration that L2 speakers of English are surrounded by English in their lives seems to be an important factor, as people who encounter English for a longer period seem to be better at producing sarcasm in this L2. Additionally, Smorenburg (2015) found out that explicit training in sarcasm in English can improve the production of sarcasm for L2 speakers. So, although L2 learners may initially face challenges in comprehending sarcasm in their L2 compared to native speakers, L2 speakers seem to have the capability to learn sarcasm in L2.

L2 speakers are not the only people who are having trouble understanding sarcasm. Persicke et al. (2013) tried to teach three children with autism, who had difficulties understanding sarcasm, how to use sarcasm in daily life using training. In this research, the children had a teacher that explained to the children what sarcasm is, and at the end of the training, these three children demonstrated the ability to understand sarcasm and were even able to produce sarcastic remarks themselves.

Another example of people that seem to have trouble with understanding sarcasm is older people (Phillips et al., 2015). Phillips et al. (2015) examined the difference in understanding sarcasm comparing different ages. They create three groups: a group of 40 young adults aged from 18 to 39 (27 females, 13 males), a group of 40 middle-aged adults aged from 40 to 64 (21 females, 19 males), and a group of 36 older adults aged from 65 to 86 (19 females, 17 male). The participants had to do two different tasks: a verbal stories task and a video task. In both tasks, the group of older adults scored poorer than the other two groups in understanding sarcastic interactions, while there was no difference between the groups in understanding sincere interactions. The differences between the groups were smaller for the video task, and Phillips et al. (2015) argue that this might be the case as in the video task, the facial movements were visible, and in the verbal stories task, this was not the case. Nevertheless, the older people did understand sarcasm worse than the middle-aged adults and the young adults in these experiments.

Further exploration of speech synthesis can contribute to understanding the crucial features involved in generating sarcasm. This research not only benefits scientists seeking a deeper comprehension of sarcasm but also individuals who encounter challenges in using or comprehending sarcasm.

2.4 Voice-transformation

With the arrival of Deep Learning, a lot of research has been done in the field of text-to-speech (TTS). Examples of TTS models are Tacotron (Wang et al., 2017), Tacotron 2 (Shen et al., 2018), Deep Voice 3 (Ping et al., 2018). These models create mel-spectrograms from text input autoregressively first before the speech is synthesized by using these mel-spectrograms using a vocoder. However, because these models are autoregressive, this gave a few problems (Ren et al., 2019). Firstly, the autoregressive models are slow, as all mel-spectrograms are generated one by one based on the previously generated mel-spectrograms. Next to that, the models are fragile because of error propagation (uncertain propagation) and wrong attention alignments between text and speech. This makes the model sometimes skip and repeat words. Finally, the models are hard to control, as the mel-spectrograms are generated automatically without given alignments between text and speech, and thus it is hard to control the voice speed and prosody in these models. Ren et al. (2019) noticed those problems and created a non-autoregressive model: FastSpeech.

The FastSpeech model can generate the mel-spectrograms in parallel, which makes the model faster than the autoregressive models. Additionally, a phoneme duration predictor is added to this model to ensure hard alignments between each phoneme and its mel-spectrograms. This way, not many words are skipped or repeated, as the model avoids error propagation and wrong attention alignments. The last problem is solved because of the use of the duration regulator. The duration can easily be controlled and changed by the regulator. Also, the prosody can be adjusted as the model can add breaks between phonemes.

Despite the advancements made by the FastSpeech model compared to autoregressive models, Ren et al. (2020) proposed a novel model known as FastSpeech2. They noted that there are several flaws in the FastSpeech model: 1) the combination of the duration regulator and the model itself is a complicated pipeline, 2) the duration regulator predicts the duration not accurately enough, and 3) since the FastSpeech model reduces the data variance to simplify the data, the voice quality and prosody gets limited. Ren et al. (2020) tackled the first problem by letting the model train with a ground-truth target instead of using the complicated pipeline. Next to that, more variance information of speech (pitch, energy, and more accurate duration) is added in the model to close the gap of information between the input (the sequence of text) and the target output (the mel-spectrograms). In training, this is trained using the mel-spectrograms, and during inference, predictors are used that are trained together with the model. As this model consists of three different predictors, it is easier to control the pitch, energy, and duration. This is the reason why we are using this model because it is easy to manipulate the speech and try to make it sound sarcastic.

The FastSpeech 2 model is shown in Figure 1(a). The phoneme sequence is the input, and the encoder converts this input to a hidden sequence. The variance adaptor adds variance information like the duration, pitch and energy in the hidden sequence. Then, the mel-spectrogram decoder decodes the modified hidden sequences in parallel into mel-spectrogram sequences. In addition, Ren et al. (2020) proposed the FastSpeech 2s model. This model generates the waveform directly from the hidden sequence and thus the speech directly from the text. This last part happens in the waveform decoder.

In Figure 1(b), the variance adaptor is shown in more detail. As can be seen, the variance adaptor

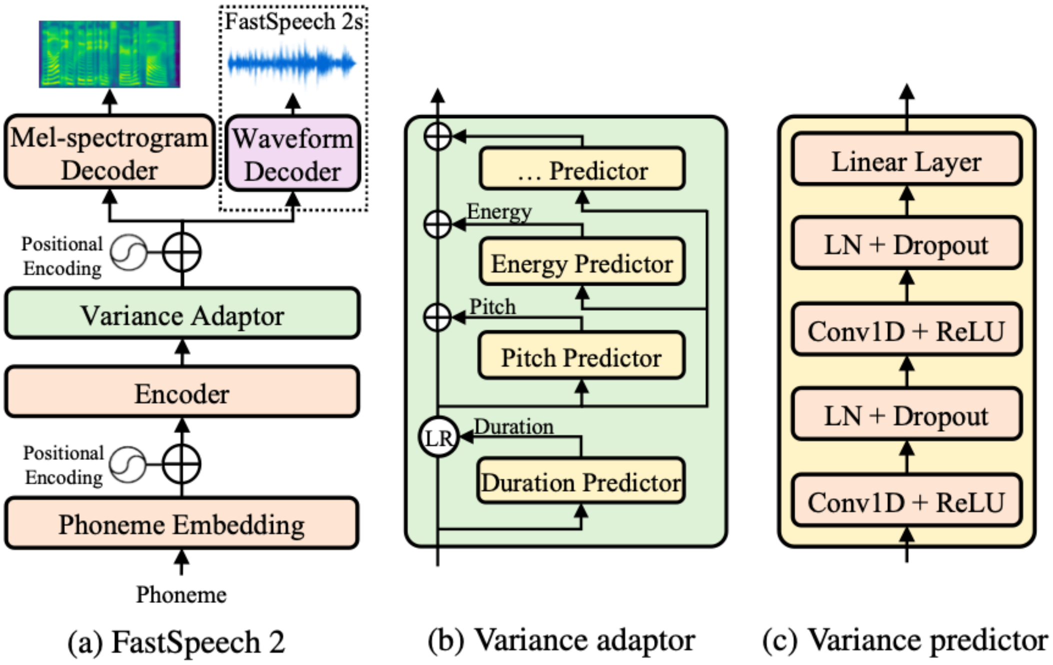


Figure 1: The FastSpeech 2 model in more detail. In (a), the overview of the whole model is shown. (b) shows the variance adaptor in more detail, and (c) shows each variance predictor in detail (pitch, energy, and duration). This figure is retrieved from Ren et al. (2020).

consists of an energy predictor, a pitch predictor, and a duration predictor. In training, the ground-truth value of energy, pitch, and duration is extracted from the recordings. The target speech is predicted by using these ground-truth values in the hidden sequences. During inference, the different variance predictors are used to synthesize the target speech. Figure 1(c) shows the variance predictors in more detail. Each predictor has 2 1D-convolutional layers with ReLU activation, followed by a layer normalization and dropout layer. The final layer of the predictors is a linear layer to create from the hidden states the output sequence.

For this research, PyTorch implementation of FastSpeech is used from Chien et al. (2021)[2].

2.5 Summary of literature review

In summary, sarcasm is used a lot in daily life, especially to be less confrontational towards others (Filik et al., 2016; Dews et al., 1995) or to be funny (Castro et al., 2019; Dews et al., 1995). Furthermore, an individual's personality (Dionigi et al., 2022; Huang et al., 2015) and place of residence (Dress et al., 2008) appear to influence the frequency of sarcasm usage. However, certain individuals, including those with autism (Persicke et al., 2013), non-native language speakers (Dolan, 2015) or elderly people (Phillips et al., 2015), may struggle to understand sarcasm. Around the world, people use different acoustic cues to produce sarcasm, so it is important to focus on the Dutch findings (Jansen et al., 2020). These findings will be used to manipulate the speech in order to see if this speech can be recognized as sarcastic. Therefore, this research aims to examine whether speech manipulation using the FastSpeech2 model (Ren et al., 2020), which allows for synthesizing and modifying speech parameters such as duration, pitch, and energy, can create sarcastic speech. The research question

[2] GitHub repository: `https://github.com/ming024/FastSpeech2`

guiding this study is: Can the synthesized and modified speech be recognized as sarcastic? The hypothesis suggests that through the manipulation of acoustic properties related to sarcasm, the synthetic speech will convey identifiable sarcasm to listeners.

3 Proposed methods

To create a sarcastic-sounding speech, speech will be synthesized, and subsequently manipulated. The synthesis and manipulation of the speech will be done using the FastSpeech2 model. The specific details of this process are elaborated in Section 4.2. This section will provide an overview of the training data that will be used for the model and outline the specific manipulations that will be applied.

3.1 Data

The Dutch data that is used is from Park and Mulc (2019)[3]. They created a single-speaker data set for 10 languages, and only the Dutch data is used. This data is from the book ‘20.000 mijlen onder zee’ (‘Twenty Thousand Leagues Under the Seas’ in English) from Verne (1876) voiced by Bart de Leeuw. It is a male voice and contains over 14 hours of speech data, split into 6494 WAV files.

3.2 Voice manipulation

The required manipulations for Dutch sarcasm, as compared to sincere speech, have been identified in a study by Jansen et al. (2020). We will focus on the differences for male speakers, as the voice of the training data is male. The differences are presented in Table 2. In the first column, the different features are shown that can be manipulated in the FastSpeech model. The ‘–pitch-control’ controls the pitch of the voice, the ‘–duration-control’ the length of the utterance and the ‘–energy-control’ controls the intensity of the voice. The second column denotes what type of sentence, and the third and fourth columns show the values that are found for sarcastic and sincere utterances. The ‘Factor (sarc)’ column indicates the rate of manipulation required to transform sincere speech into sarcastic speech. For example, in the model, the ‘–duration-control’ is multiplied by the factor to generate from the sincere speech the sarcastic speech. Notably, the intensity values in the table are initially expressed in decibels (dB), but they are converted to a linear scale before calculating the factors. In addition, to enlarge the difference between sincere and sarcastic speech, the factor of sarcastic speech is doubled. This is added because only the mean values of duration and intensity are found in literature (Jansen et al., 2020). The manipulation values for this category are added in the sixth column. This speech that is manipulated by these factors is referred to as *extra_sarcastic*. No differences in pitch were found comparing male sarcastic and sincere speech according to Jansen et al. (2020), so no manipulation in pitch is done. Using the factors shown in the table, there are three types of manipulation used for the synthesis: *sincere* utterances (no manipulation), *sarcastic* utterances (manipulation according to the findings of Jansen et al. (2020)), and *extra_sarcastic* utterances (manipulation like the *sarcastic* utterance but doubled).

[3] The data is retrieved from `https://www.kaggle.com/datasets/bryanpark/dutch-single-speaker-speech-dataset`.

Male	**Type sentence**	**Sarcastic**	**Sincere**	**Factor (sarc)**	**Factor (extra sarc)**
–pitch-control	Declaratives	-	-	-	-
	Wh-exclamatives	-	-	-	-
	Tag-questions	-	-	-	-
–duration-control	Declaritives	1516.0 ms	1317.0 ms	1.15	1.30
	Wh-exclamatives	1397.0 ms	1207.0 ms	1.16	1.32
	Tag-questions	1424.0 ms	1338.0 ms	1.06	1.12
–energy-control	Declaratives	59.5 dB	60.1 dB	0.87	0.74
	Wh-exclamatives	60.0 dB	61.2 dB	0.76	0.52
	Tag-questions	60.6 dB	61.0 dB	0.91	0.82

Table 2: The differences in sarcastic and sincere speech for males per sentence. The ‘Factor (sarc)’ column shows the factor that is multiplied by the duration and energy values of the sincere speech to convert it to sarcastic speech. The last column shows the manipulation that is used for the *extra_sarcastic* synthesized voice. This manipulation is the manipulation of the *sarcastic* voice doubled.

4 Experimental setup

This section describes in detail how the speech is synthesized and manipulated using the FastSpeech2 (Ren et al., 2020) model and the Montreal Forcerd Aligner (McAuliffe et al., 2017). As demo, a GitHub repository[4] is created including all the steps that are taken to synthesize the speech.

4.1 Data preprocess and alignment

The first step is data preprocessing. The Dutch single-speaker data set is first trimmed, so that the noises at the beginning and the end of the audio files are removed. This ensures that the FastSpeech2 model is not trained on irrelevant silences, resulting in cleaner and clearer synthetic speech compared to using untrimmed data. Additionally, all the text of each speech fragment is extracted from the transcript file and each text is stored in a LAB file with a name corresponding to the audio (WAV) file. The combination of these LAB and WAV files form the corpus.

Then, the speech data is aligned using the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017). This aligner is used, as it can get the ground-truth phoneme-level duration, and these will be used in the FastSpeech2 model. These durations are more accurate than the teacher model of Fastspeech (Ren et al., 2020). The aligner utilizes a pronunciation dictionary to align the orthographic transcription and the corresponding phonetic sequence of each sentence, generating TextGrids. For this experiment, the dictionary of Doherty (2019) is used, which includes a wide range of Dutch words with their pronunciations in the International Phonetic Alphabet (IPA). The stress marks were removed from the dictionary, as it deemed insignificant.

Next, the MFA was used to create an acoustic model from the dictionary and the corpus. Due to the high number of Out of Vocabulary (OOV) items, a new grapheme-to-phoneme (G2P) model is constructed from the dictionary using the MFA. This new model enables the creation of an extended dictionary that covers a larger vocabulary. Then, the revised dictionary and the corpus were used to generate the aligned TextGrids. As a final step, in the phoneme sequence in the TextGrids, a silent token was added when no phoneme was specified and thus there was a silence in speech. These TextGrids, along with the revised dictionary and the corpus, are used in the FastSpeech2 model.

4.2 Voice manipulation with FastSpeech2

When the alignment is complete, it can be used in the FastSpeech2 model. Prior to training, the model requires preparation, including data preprocessing at the phoneme level. Also, English text cleaners are used in preprocessing, as they are similar to Dutch in terms of using the same alphabet. For example, the cleaners converted the text to lowercase and extraneous whitespaces are collapsed. Additionally, since the dictionary uses IPA symbols, the IPA symbols are incorporated as additional symbols in the model. Then, the model can train.

The training is conducted on an Nvidia A100 GPU, running on Hábrók (Center for Information Technology of the University of Groningen, 2023). The model undergoes training for 200,000 steps using optimizer parameters detailed in Table 3. The training took 16 hours and 54 minutes and used 6.56GB of memory.

[4] The GitHub repository to the demo: `https://github.com/TessaZwart/Sarcastic_Speech_Synthesis`

Parameter	**Value**
batch_size	16
betas	[0.9, 0.98]
eps	0.000000001
weight_decay	0.0
grad_clip_thresh	1.0
grad_acc_step	1
warm_up_step	4000
anneal_steps	[300000, 400000, 500000]
anneal_rate	0.3

Table 3: The values of each optimizer parameter from the trained model.

4.3 Pilot

A pilot test involving three female participants was conducted, consisting of 32 questions that exclusively featured ‘sarcastic’ and ‘sincere’ utterances from the sentences provided in Appendix B.1. The results of the pilot test revealed the participants’ difficulty in distinguishing between sarcastic and sincere utterances, with correct answers ranging from 10 to 14 out of the 32 questions. Of particular significance, participants achieved only 2, 1, and 6 correct answers, respectively, for questions where the expected response was sarcastic.

Following the pilot test, adjustments were made to the survey, including the addition of extra questions. The values obtained from literature (Jansen et al., 2020) were doubled to create more noticeable distinctions between the sincere and new sarcastic utterances. Since the mean values of intensity and duration were only provided in the paper, it was logical to introduce utterances with increased differences in duration and intensity. Consequently, the survey was expanded by 16 additional questions, encompassing ‘sincere’ utterances, ‘sarcastic’ utterances, and ‘extra sarcastic’ utterances.

5 Evaluation

The final evaluation incorporates all categories of manipulation: *sincere*, *sarcastic*, and *extra_sarcastic*. This way, participants can be tested if they recognize the sarcastic and sincere speech. A survey is conducted and will be analyzed in this section.

5.1 Participants

A total of 36 participants completed the survey. The participants were family and friends, who found the link to the survey online or in a text message. Also, in the Whatsapp group chat of Campus Fryslân, a message was sent with the question to fill in the survey. Two of the participants provided the same answer for all questions, so the results of those two participants were excluded as outliers from the study. 32 out of the remaining 34 participants filled in that Dutch was their first language. The other two participants filled in Frisian, but both filled in Dutch as another language they understand. As Frisians live in the Netherlands, they were not excluded from this survey. The effect of the first language will be analyzed in the results. 6 of the participants were male, and the other 28 were female. It is worth noting that only one participant missed answering a single question, which was disregarded during the data analysis phase.

5.2 Survey

A survey[5] was conducted using Qualtrics (2005) to evaluate the synthetic speech. The survey is added to Appendix A. Participants were requested to provide consent in compliance with the GDPR regulations. The consent statement outlined their ability to halt the survey at any point, permitted the utilization of their responses for research purposes, and ensured that the data would be stored anonymously. Additionally, some questions were asked about the linguistic background of the participants. First, the native language was asked. Because sarcasm is found to be different across languages (see section 2.2), it was important for this research that only native Dutch speakers were asked to fill in the survey. Participants were also inquired about their comprehension of other languages and their gender. The latter was asked because, in the production of sarcasm in Dutch, differences were found between male and female Jansen et al. (2020), which might indicate that there is also a difference in hearing and recognizing sarcasm. The difference in the amount of male and female participants was unfortunately too big to examine this difference.

As intensity is important in producing sarcasm in Dutch, the next task the participants had to do, was check if their speech files were good to hear. An example sentence was given to check the volume of the files. After setting the volume of their own devices, it was mentioned not to change the volume anymore. Additionally, three more sentences were given to the participants to hear what the basic sound of the synthetic voice was. This way, the participants could get familiar with the voice as well, and then they are also able to hear the difference between sincere and sarcastic speech because sincere speech was known. The sentences that are chosen as base sentences were found in the paper of Hemmer (2018). They investigated the emotional charge of sentences. The sentences that were used as base sentences were neutral sentences, not emotionally charged. The following sentences were used:

[5] The survey can be found here: `https://rug.eu.qualtrics.com/jfe/preview/previewId/e2272825-9b32-40a0-9148-b95e7e61d38f/SV_1Ug61gScEydSeG2?Q_CHL=preview&Q_SurveyVersionID=current`

- ‘Waarom staat de zon in de winter lager dan in de zomer?’ *(Why is the sun lower in the sky during winter compared to summer?*)
- ‘Piet daalde geduldig de trappen af.’ *(Piet patiently descended the stairs.)*
- ‘Heb je uiteindelijk je zin gekregen?’ *(Did you eventually get what you wanted?)*

After this, 48 questions were asked about the emotion of the speaker. Each question contained a speech file with the text of one of the 16 sentences shown in Appendix B.1 either not manipulated (*sincere*), manipulated with the feature found in the literature (*sarcastic*) or manipulated with double the manipulations of *sarcastic* (*extra_sarcastic*). The sentences of Appendix B.2 were also generated, but they were not understandable enough, so we did not include them in this study. The sentences were used by Jansen et al. (2020) as well and were retrieved from Chen and Boves (2018) and translated from English to Dutch. Three types of sentences were used: declaratives, wh-exclamatives, and tag-questions. The tag-questions are translated with ‘hè’ at the end of the sentence as the translation for ‘wasn’t he’, ‘weren’t you’, ‘hasn’t he’, etc. Jansen et al. (2020) argued that this word seems to be the best translation, because the semantics, syntactic position, and frequency of this tag are similar to the English tag-questions. This is why the sentences in this research are translated accordingly.

We decided to ask about the emotion of the speech, as we did not want the priming effect to take place (Janiszewski and Wyer Jr, 2014). This can cause the participants to be primed towards the sarcastic answer, as this is the most prominent one. To include decoy answers this attention is drawn to these answers as well. Two decoy answers were added (‘happy’ and ‘sad’) to convince the participants that the experiment was about emotion in speech and not only sarcasm in speech. The participants could go back to the previous question if they changed their minds. The visual of these questions is shown in Appendix C. The 48 questions were asked in random order.

The participants filled in the survey voluntarily on any device they preferred. The survey was spread online and was conducted without supervision.

6 Results

The sections show the results that are collected with the survey. The results are divided into multiple sections to get a better overview.

6.1 Differences between recognizing sarcastic and sincere speech

Firstly, a comparison between the recognition of sarcastic and sincere speech is made. The correct answers per type of manipulation (*sincere, sarcastic, extra_sarcastic*) were calculated and plotted in a boxplot to see the differences.

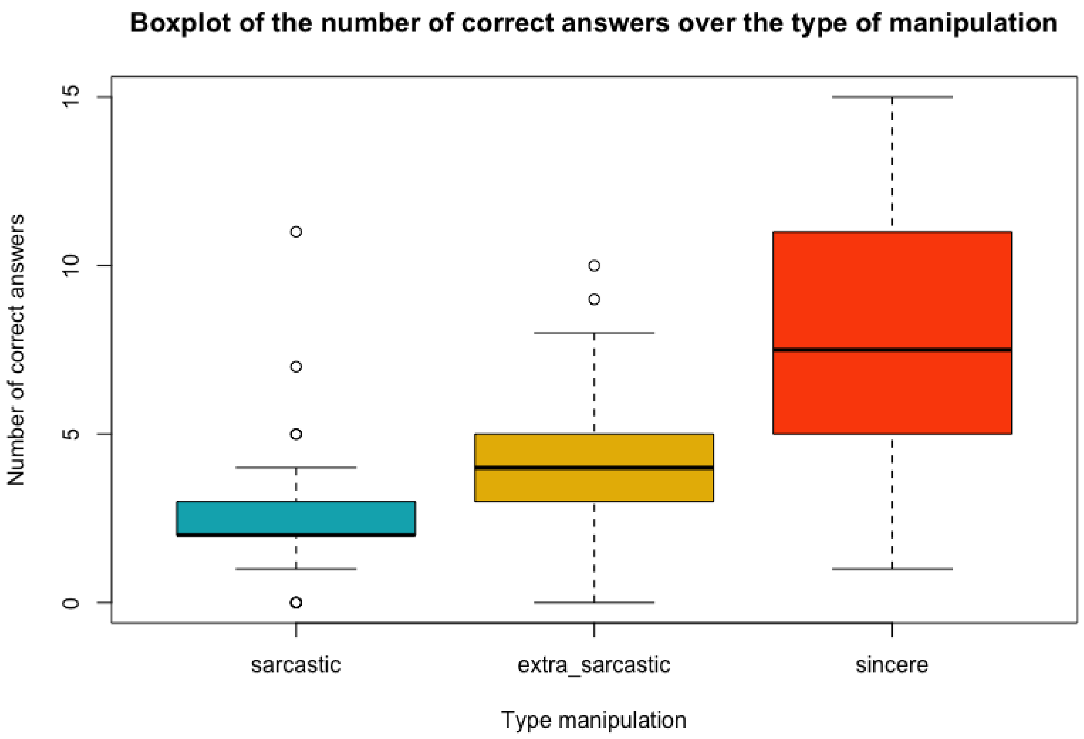


Figure 2: Boxplot with the number of correct answers divided into the different manipulations. As can be seen, the speech fragments that had the *sincere* manipulation were most often recognized correctly. The *sarcastic* manipulation caused the least correct answers.

As can be seen in the boxplot of Figure 2, it seems that the *sincere* manipulation is recognized the most, and the *sarcastic* manipulation the least. Also, it seems that the *extra_sarcastic* manipulation is better recognized than the *sarcastic* manipulation, which can indicate that the manipulation of the *sarcastic* speech was not recognizable enough, while the speech with the *extra_sarcastic* manipulation was recognized as sarcastic. A one-way ANOVA test was performed to see if there was a significant difference in correct recognized answers between the three groups. This test showed that there was a statistically significant difference in the number of correct answers between at least two groups (F(2, 99) = 33.37, p = 8.36e-12). Then, a Tukey's HSD Test for multiple comparisons was done to examine what groups are different and found that there was a significant difference between the *sincere* and *sarcastic* manipulation (p = 0.00e+00, 95% C.I. = [3.64, 6.89]) and between the *sincere* and *extra_sarcastic* manipulation (p = 0.00e+00, 95% C.I. = [2.58, 5.83]).

The final step that is taken to compare the difference between recognizing sarcastic and sincere speech is by calculating the precision, recall, and accuracy (Powers, 2020). Precision is the proportion of all predicted positive cases and the actual positive cases. The recall is how many real positive cases are predicted as positive. The accuracy is a combination of both. The precision, recall and accuracy are calculated as follows:

$$Precision = \frac{TP}{(TP+FP)} \tag{1}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{2}$$

$$Accuracy(F1) = \frac{2*precision*recall}{(precision+recall)} \tag{3}$$

- TP: True Positive, positive cases that are correctly predicted as positive
- FP: False Positive, negative cases that are incorrectly predicted as positive
- FN: False Negative, positive cases that are incorrectly predicted as negative

In Table 4, the precision, recall and accuracy are shown for both sincere and sarcastic, including the decoy answers or not, and including the *extra_sarcastic* data or not. In general, the precision of the *sincere* voice is worse (35% to 51%) than the precision of the *sarcastic* voice (54% to 74%). For the recall, this is the other way around (sincere: 50% to 77%, sarcastic: 17% to 30%). Overall, for none categories, the accuracy (F1) is measured to be higher than 62%, which is for the sincere category where both the *extra_sarcastic* data and the decoy answers were ignored. The precision of sarcasm is higher when the *extra_sarcastic* data is included. This means that the accuracy of only the *extra_sarcastic* data shows a higher precision than solely the *sarcastic* data. For *sincere*, the precision drops when the *extra_sarcastic* data is added. However, the recall for sincere and sarcastic is higher when the decoy answers are not included and are both comparable when the *extra_sarcastic* data is included or not. Finally, the accuracy for sincere is in all categories higher than sarcastic.

Decoy answers	**extra_sarcastic**	**Sincere or sarcastic**	**Precision**	**Recall**	**Accuracy (F1)**
Yes	Yes	Sincere	0.35	0.50	0.41
Yes	Yes	Sarcastic	0.74	0.20	0.32
Yes	No	Sincere	0.51	0.50	0.51
Yes	No	Sarcastic	0.54	0.17	0.26
No	Yes	Sincere	0.35	0.77	0.48
No	Yes	Sarcastic	0.74	0.30	0.43
No	No	Sincere	0.51	0.77	0.62
No	No	Sarcastic	0.54	0.26	0.35

Table 4: Table containing the precision, recall, and accuracy (F1) for both sincere and sarcastic, and including decoy answers or not and including the *extra_sarcastic* data or not.

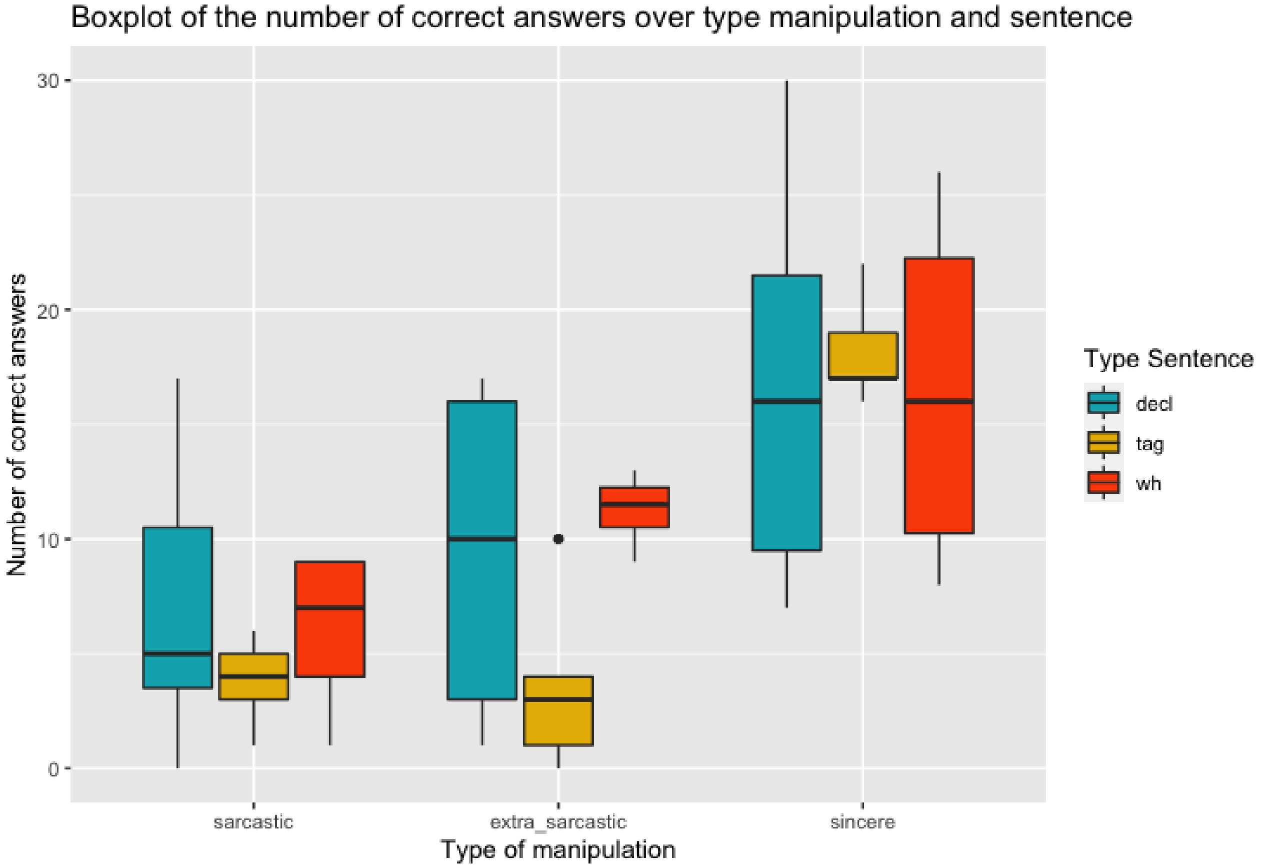


Figure 3: Boxplot of the number of correct answers divided into type of manipulation and type of sentence.

6.2 Differences between the various sentences

In this research, three types of sentences are investigated: declaratives, wh-exclamatives, and tag-questions. A boxplot was created and presented in Figure 3 to assess whether there was a variation in answering the questions in the survey correctly based on the sentence type.

In the boxplot, there seem to be some differences between the number of correct answers comparing the type of sentence. Especially in the manipulation *extra_sarcastic*, the wh-exclamatives seems to be recognized as sarcasm more often than the tag-questions. However, when performing a two-way ANOVA test of the effect of sentence type and the effect of manipulation type on the number of correct answers, only a significant difference was found for manipulation type (F(2, 43) = 16.97, p = 3.69e-06). This finding was also found before in section 6.1. Also, no interaction effect between the manipulation type and sentence type is found (F(4, 39) = 0.971, p = 0.43). So, there is no significant difference between the type of sentences on the number of correct answers to the survey questions, even when the type sentences are differently manipulated, as shown in Table 2.

Additionally, the precision, recall, and accuracy are calculated for all three types of sentences. The results are shown in Appendix D. There are no big differences between the sentences overall. Only the accuracy for the tag-questions seems to be lower for recognizing sarcasm compared to the other type of sentences.

6.3 Difference in hearing a manipulation or not

As participants filled in the decoy answers (‘happy’ and ‘angry’) 547 times out of 1632 total answers, we are interested to see if this happens more often with the manipulation voice (*sarcastic* and *extra_sarcastic* manipulation) than non-manipulated voice (*sincere*). If this is the case, participants might have heard a difference in the voice but did not recognize it as sarcastic. We collected the data where the decoy answers were given and compared it with how often it was chosen for each type of manipulation. This resulted in six categories: 2 decoy answers times 3 types of manipulation. A one-way ANOVA test shows that there was a significant difference between the category (F(5, 198) = 38.57, p = <2e-16). However, a Tukey’s HSD Test for multiple comparisons found that this difference only occurs between the categories of the decoy answer ‘happy’ and the categories of the decoy answer ‘angry’, so there are no significant differences between the manipulation type considering one decoy answer.

6.4 Difference between Frisian and non-Frisian

One of the participants that did the pilot of the experiment noticed that the synthetic voice sounded ‘Brabants’, which is a dialect of Dutch in the south of the Netherlands. As 14 out of 34 participants answered that they have Frisian as their native language or understand the Frisian language, this might be an influence on recognizing the sarcastic and sincere utterances. A two-way ANOVA test was performed to compare the effect of understanding Frisian or not on the correct number of answers in the survey. The test revealed that there was no significant difference between the two groups (F(1, 96) = 1.38, p = 0.24). Also, no interaction effect between understanding Frisian and the type of manipulation is not found (F(2, 96) = 0.64, p = 0.53).

7 Discussion and limitations

In this section, we will delve into the discussion of the results and establish the limitations associated with these findings. As written in section 6.1, a significant difference is found between recognizing sarcastic manipulated synthetic speech as sarcastic and non-manipulated synthetic speech as sincere. Also, a significant difference is found between recognizing the extra-manipulated synthetic speech as sarcastic and non-manipulated synthetic speech as sincere. This variation in correct responses across different manipulation types may be explained by examining the frequency distribution of each answer choice. The number of times 'Neutraal' (sincere) is chosen is 783, 'Sarcastisch' (sarcastic) is chosen only 301. Notably, for both the *sarcastic* and *extra_sarcastic* manipulations, the correct answer is exclusively 'Sarcastisch'. This means that both manipulations together have less than half of the answers that are correct for the *sincere* manipulation. This makes the chance of having the correct answer for the non-manipulated speech (*sincere*) higher than for the manipulated speech (*sarcastic* and *extra_sarcastic*).

A similar explanation can be applied when examining precision, recall, and accuracy measures. The precision for recognizing the *sarcastic* manipulated speech as sarcastic is higher than recognizing the non-manipulated speech as sincere. The recall is, however, lower than the non-manipulated speech, as the selection of the "Sarcastisch" answer is less frequent. This means that the answer 'Sarcastisch' is given less than 'Neutraal', but when it is chosen, it is more often correct. Conversely, for non-manipulated speech, the recall is higher than for manipulated speech, but the precision is lower. This suggests that 'Neutraal' is chosen frequently but is also frequently incorrect. In general, the accuracy of recognizing sarcastic with the *sarcastic* or *extra_sarcastic* manipulation does not reach higher than 43%. This means that in the optimal case (when the decoy answers are not considered, and the questions of the *extra_sarcastic* manipulated speech is included), the sarcastic speech generated in this research is recognized as sarcastic only with an accuracy of 43%.

Moreover, there appears to be a slight distinction in recognizing speech with the *sarcastic* manipulation and speech with the *extra_sarcastic* manipulation. This difference does not show to be significant, but as can be seen in Table 4, including the *extra_sarcastic* data, the accuracy for recognizing sarcasm slightly improves compared to when the *extra_sarcastic* data is not included. It is possible that a more pronounced manipulation could enhance the synthetic speech to sound more like sarcastic speech. However, more research needs to be done here; as for now, no significant difference is found between the two types of manipulation.

Section 6.2 presents the results regarding the differences between sentences. No significant distinction is observed among the sentences, despite being manipulated with varying factors. When looking at accuracy (F1), the accuracy of the tag-questions seems to be slightly lower compared to the other sentences. This can be the case because these kinds of sentences were manipulated the least, as indicated in Table 2.

Given the limited recognition of sarcastic utterances, reaching only up to 43% accuracy, it suggests that the visual aspect of conveying sarcasm may play a crucial role alongside speech. Attardo et al. (2003) highlighted the significance of facial cues, suggesting that a blank face can indicate that a speaker is being sarcastic. A blank face is an emotionless face in which no muscle is used. More specifically, each muscle is in its 'default value'. Additionally, Phillips et al. (2015) stated that elderly people were better at recognizing sarcasm when sarcasm was presented including a video. These findings underscore the importance of considering visual elements, such as facial expressions, alongside speech in order to enhance sarcasm recognition.

It is noteworthy to mention that pitch modification was not employed in this research to create sarcastic sentences. As Jansen et al. (2020) did not find a difference in pitch for male speakers between sincere and sarcastic speech, the pitch is not manipulated, while it seems to be an important factor in other languages (see Table 1). More research can be done to investigate more thoroughly the pitch influence on speech to ensure whether the pitch is important or not in Dutch sarcasm. Also, a feminine synthetic voice can be manipulated with pitch differences because Jansen et al. (2020) did find a difference in pitch between sincere and sarcastic speech for female speakers of Dutch.

Next, the decoy answers are examined in section 6.3. There is a significant difference found between the number of times the decoy answer 'Blij' (happy) (497 times) and the decoy answer 'Boos' (angry) (50 times) was chosen. When we look into the literature, this is hard to explain. For example, Zhu (2013) stated in their paper that the emotion anger is characterized by a longer duration, a higher intensity, and a slightly lower F0 mean. Only the duration matches with sarcasm in Dutch. On the other side, Zhu (2013) noted that happiness is characterized by shorter duration, ad slightly higher intensity and a higher F0 mean. All of those manipulations are not audible in sarcasm. The reason that people still answered 'happy' more often than 'angry' cannot be solely attributed to the acoustic features.

No variations were observed between the types of manipulation for each individual decoy answer. This means that the decoy answers are not chosen more often in manipulated speech. This could have indicated that the participants did hear a difference in emotion compared to the sincere speech but did not recognize it as sarcastic. However, as shown in the results, this is not the case.

In future research, manipulated speech to create happy speech and angry speech can be included. By comparing sarcastic manipulated speech to different emotional expressions in speech, participants can gain a better understanding of how the speaker communicates in various contexts and become more familiar with the speaker's style. Pexman and Zvaigzne (2004) examined if people that have a closer relationship are better at understanding sarcastic comments, and it was found that familiarity does play a role in comprehension. This means that if the participants are more familiar with the synthetic voice, they may also recognize sarcastic speech earlier. Additionally, it is worth considering that the similarity between sincere and sarcastic speech may be contributing to the recognition challenges. So comparing the sarcastic speech to different emotions in speech might give different results.

Furthermore, it is possible that the presence of decoy answers caused confusion among participants, leading them to choose these options at least occasionally. Participants might have been overly focused on identifying all four emotions, which could have influenced their perception of happiness and anger as well. Additionally, it is worth considering that the sincere voice used in the study may not have been perceived as sincerely as intended. If the manipulation was applied to speech that was not explicitly sincere, this could have resulted in a speech that did not effectively convey sarcasm. However, a lot of times, the answer 'Neutraal' (sincere) was given, so it seems that the speech that was not manipulated sounded sincere.

The language that people speak also influences the use of sarcasm, as outlined in Section 2.2. Given that 14 out of 34 participants indicated either their first language was Frisian or they understand Frisian, a comparison in correct answers between people that understand Frisian and people who do not was made. One of the participants noticed that the synthetic voice sounded 'Brabants', which is a dialect in the south of the Netherlands, while Frisians live in the north of the Netherlands. When we compared the two groups, we did not see a significant difference in correctly categorizing the speech with different manipulations. This suggests that the differences between the 'Brabants' dialect and the Frisian language may not significantly impact the perception of sarcasm. Also, 12 out of 14 participants that understand Frisian did not fill Frisian in as their native language. Thus, these participants might not speak Frisian at all and only Dutch, and 'Brabants' is also Dutch, just a dialect.

Lastly, it is important to consider the training data used for the model, which originated from an old book dating back to 1878 (Verne, 1876). This book primarily consists of outdated vocabulary, as newer words did not exist during that time period. If the FastSpeech2 model is not trained on those newer words, it might be the case that those words are synthesized worse. Additionally, the model does not synthesize a fluent human voice in terms of fluency and naturalness. The participants might not be familiar with synthetic speech, and the model could benefit from extended training beyond the 200,000 steps to let the synthetic speech become more like human voices.

8 Conclusion

In this paper, we examined whether synthesized and modified speech can be recognized as sarcasm. We hypothesized that through the manipulation of acoustic properties related to Dutch sarcasm, the synthetic speech will convey identifiable sarcasm to listeners.

Sincere synthesized speech is manipulated using the FastSpeech2 model with the aim of sounding sarcastic. The results reveal that the majority of the manipulated synthetic speech is not easily identifiable as sarcastic, with the accuracy of sarcasm recognition not surpassing 43%. However, when excluding the questions where the manipulation was doubled, the accuracy decreases to 35%, suggesting that increased manipulation leads to a higher recognition of sarcasm. It is worth noting that the participants often failed to perceive the manipulations, as the manipulated speech was frequently identified as sincere rather than sarcastic or the decoy options of happy or angry. Next to that, no significant differences are found by comparing the different sentence types. Even though the manipulation found in literature was already doubled in this research, more manipulation might be needed to recognize the sarcastic voice. Additionally, the combination of facial expression, text and way of speaking is important to convey sarcasm.

With this research, we hope we are a step closer to synthesizing sarcasm and that there is more insight into what features are important in Dutch sarcasm, which can be helpful for people that have difficulties using sarcasm.

Bibliography

Anolli, L., Ciceri, R., and Infantino, M. G. (2002). From “blame by praise” to “praise by blame”: Analysis of vocal patterns in ironic communication. *International Journal of Psychology*, 37(5):266–276.

Attardo, S., Eisterhold, J., Hay, J., and Poggi, I. (2003). Multimodal markers of irony and sarcasm.

Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., and Poria, S. (2019). Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*.

Center for Information Technology of the University of Groningen (2023). `https://wiki.hpc.rug.nl/habrok/start`.

Cheang, H. S. and Pell, M. D. (2008). The sound of sarcasm. *Speech communication*, 50(5):366–381.

Cheang, H. S. and Pell, M. D. (2009). Acoustic markers of sarcasm in Cantonese and English. *The Journal of the Acoustical Society of America*, 126(3):1394–1405.

Chen, A. and Boves, L. (2018). What’s in a word: Sounding sarcastic in British English. *Journal of the International Phonetic Association*, 48(1):57–76.

Chien, C.-M., Lin, J.-H., Huang, C.-y., Hsu, P.-c., and Lee, H.-y. (2021). Investigating on Incorporating Pretrained and Learnable Speaker Representations for Multi-Speaker Multi-Style Text-to-Speech. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8588–8592.

Dews, S., Kaplan, J., and Winner, E. (1995). Why not say it directly? The social functions of irony. *Discourse processes*, 19(3):347–367.

Dionigi, A., Duradoni, M., and Vagnoli, L. (2022). Humor and the dark triad: Relationships among narcissism, machiavellianism, psychopathy and comic styles. *Personality and Individual Differences*, 197:111766.

Doherty, L. (2019). ipa-dict - Monolingual wordlists with pronunciation information in IPA.

Dolan, J. (2015). How Second Language English Learners Interpret Sarcasm in English: A Survey. *Schwa: Language and Linguistics*, 13:11–26.

Dress, M. L., Kreuz, R. J., Link, K. E., and Caucci, G. M. (2008). Regional variation in the use of sarcasm. *Journal of Language and Social Psychology*, 27(1):71–85.

Dynel, M. (2014). Isn’t it ironic? Defining the scope of humorous irony. *Humor*, 27(4):619–639.

Filik, R., Ţurcan, A., Thompson, D., Harvey, N., Davies, H., and Turner, A. (2016). Sarcasm and emoticons: Comprehension and emotional impact. *Quarterly Journal of Experimental Psychology*, 69(11):2130–2146.

Gibbs, R. W. (2000). Irony in talk among friends. *Metaphor and symbol*, 15(1-2):5–27.

Hemmer, A. (2018). Codering van de affectieve functie van taal: een onderzoek naar de emotionele lading van afwijkend taalgebruik. B.S. thesis.

Huang, L., Gino, F., and Galinsky, A. D. (2015). The highest form of intelligence: Sarcasm increases creativity for both expressers and recipients. *Organizational Behavior and Human Decision Processes*, 131:162–177.

Janiszewski, C. and Wyer Jr, R. S. (2014). Content and process priming: A review. *Journal of consumer psychology*, 24(1):96–118.

Jansen, N., Chen, A., et al. (2020). Prosodic encoding of sarcasm at the sentence level in Dutch. In *Proc. 10th International Conference on Speech Prosody 2020*, pages 409–413. ISCA-INST SPEECH COMMUNICATION ASSOC.

Lœvenbruck, H., Jannet, M. B., d'Imperio, M., Spini, M., and Champagne-Lavau, M. (2013). Prosodic cues of sarcastic speech in French: slower, higher, wider. In *Interspeech 2013-14th Annual Conference of the International Speech Communication Association*, pages 3537–3541.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Murphy, P. J. and Akande, O. O. (2005). Cepstrum-based estimation of the harmonics-to-noise ratio for synthesized and human voice signals. In *Nonlinear Analyses and Algorithms for Speech Processing: International Conference on Non-Linear Speech Processing, NOLISP 2005, Barcelona, Spain, April 19-22, 2005, Revised Selected Papers*, pages 150–160. Springer.

Niebuhr, O. (2014). A little more ironic–voice quality and segmental reduction differences between sarcastic and neutral utterances. In *7th International Conference on Speech Prosody, Dublin, Ireland, Proceedings*, pages 608–612.

Park, K. and Mulc, T. (2019). CSS10: A Collection of Single Speaker Speech Datasets for 10 Languages. *Interspeech*.

Persicke, A., Tarbox, J., Ranick, J., and Clair, M. S. (2013). Teaching children with autism to detect and respond to sarcasm. *Research in Autism Spectrum Disorders*, 7(1):193–198.

Pexman, P. M. and Zvaigzne, M. T. (2004). Does irony go better with friends? *Metaphor and symbol*, 19(2):143–163.

Phillips, L. H., Allen, R., Bull, R., Hering, A., Kliegel, M., and Channon, S. (2015). Older adults have difficulty in decoding sarcasm. *Developmental psychology*, 51(12):1840.

Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J., and Miller, J. (2018). Deep voice 3: Scaling text-to-speech with convolutional sequence learning.

Powers, D. M. (2020). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Qualtrics (2005). `https://www.qualtrics.com`.

Rao, R. (2013). Prosodic consequences of sarcasm versus sincerity in Mexican Spanish. *Concentric: Studies in Linguistics*, 39(2):33–59.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2019). Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32.

Rockwell, P. (2000). Lower, slower, louder: Vocal cues of sarcasm. *Journal of Psycholinguistic research*, 29(5):483–495.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. (2018). Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Smorenburg, B. (2015). The effect of explicit training on the prosodic production of L2 sarcasm by Dutch learners of English. B.S. thesis.

Verne, J. (1876). *20.000 mijlen onder zee*.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Zhu, Y. (2013). *Expression and recognition of emotion in native and foreign speech: The case of Mandarin and Dutch*. Netherlands Graduate School of Linguistics.

Appendices

A Appendix: Survey

Start survey:

> Hoi! Bedankt voor het invullen van deze enquête!
>
> In dit experiment zijn we geïnteresseerd in emoties in spraak. U zult luisteren naar 48 spraakfragmenten. Dit zal ongeveer 9 minuten duren. Er zullen vragen gesteld worden in welke emotie deze tekst gesproken wordt. Op elke pagina krijgt u een spraakfragment te zien. Na het afspelen van het spraakfragment kiest u een van de 4 emoties die het meest overeenkomt met het afgespeelde spraakfragment.
>
> Door verder te gaan naar de volgende pagina geeft u ons toestemming om uw resultaten te gebruiken in ons onderzoek. U kunt dit onderzoek op elk moment stoppen door het venster te sluiten en uw deelname te beëindigen. Door de hele enquête in te vullen gaat u vrijwillig akkoord met deelname. Dit onderzoek is anoniem, u kunt niet worden geïdentificeerd door de antwoorden die u in dit onderzoek hebt gegeven en niemand zal weten dat u deelgenomen heeft.
>
> Als u enige vragen heeft over dit onderzoek, kunt u een email sturen naar t.zwart.1@student.rug.nl.
>
> Geeft u toestemming en wilt u doorgaan met de enquête?
>
> - Ja, ik geef toestemming
> - Nee, ik geef geen toestemming

Demographic questions:

> Wat is uw moedertaal?
>
> *Text entry*
>
> Welke andere talen begrijpt u?
>
> *Text entry*
>
> Wat is uw geslacht?
>
> - Man
> - Vrouw
> - Non-binair / derde geslacht
> - Zeg ik liever niet

Audio instructions:

> Voor dit experiment is het belangrijk dat u het volume van uw telefoon/computer niet verandert tijdens het experiment. Zet het volume op een comfortabel niveau zodat u het audiobestand kunt horen.

Controleer het volume met het volgende spraakfragment:

Audio fragment of example sentence

Nu volgen 48 vragen. Ter referentie, dit is normaal gegenereerde spraak van deze spreker:

Audio fragment of three example sentences, not emotionally charged

Main question, this questions is asked 48 times, all 16 sentences using the three types of manipulations:

Wat is volgens u de emotie van de spreker?

Audio fragment of synthesized sentence

- Blij
- Boos
- Sarcastisch
- Neutraal

B Appendix: Sentences

B.1 Synthesized used sentences

Type sentence	Sentence	Translation to English
Declarative	Ze is een sierlijke danspartner.	She’s a graceful dance partner.
	Ik heb een fantastische tijd.	I’m having a great time.
	Het is een prachtige dag buiten.	It’s a beautiful day outside.
	Dat klinkt comfortabel.	That sounds comfortable.
	Ze is een gezonde vrouw.	She’s a healthy lady.
	De bediening is erg goed hier.	The service’s really good here.
	Die grap is hilarisch.	That joke’s hilarious.
Tag-question	Hij heeft goed geholpen, hè?	He was a big help, wasn’t he?
	Je was blij om haar te zien, hè?	You were pleased to see her, weren’t you?
	Hij heeft prachtig werk verricht, hè?	He’s done a beautiful job, hasn’t he?
	Je bent een goede buurman, hè?	You’re a nice neighbour, aren’t you?
	Ze bijten goed dit seizoen, hè?	They’re biting this season, aren’t they?
Wh-exclamative	Wat een fantastisch resultaat!	What an amazing result!
	Wat een fantastische rijder!	What a great driver!
	Wat een dappere man!	What a brave man!
	Wat een aangrijpende film!	What a gripping film!

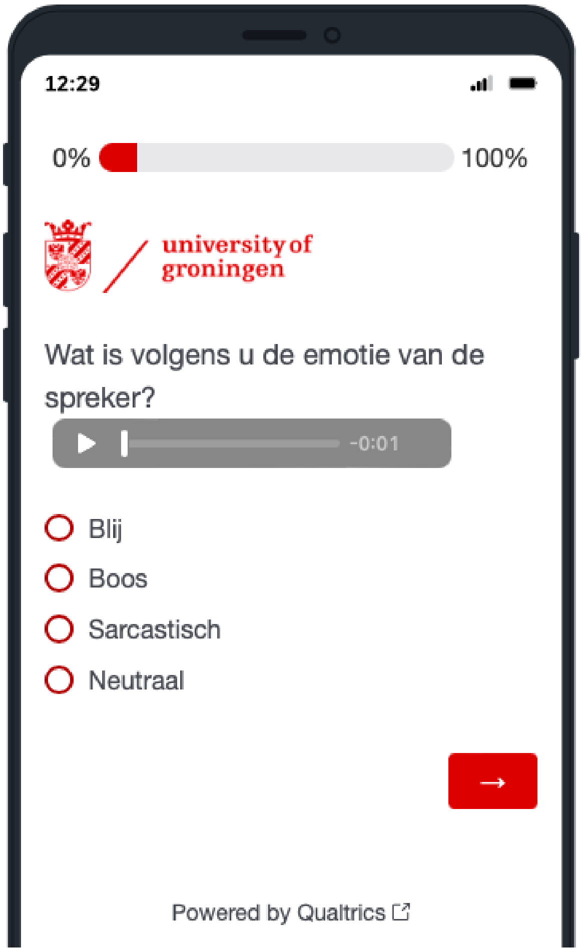
Table 5: Sentences that are generated by the FastSpeech2 model and used in the survey. Also, the type of sentences are shown and the translation in English.

B.2 Synthesized non-used sentences

Type sentence	**Sentence**	**Translation to English**
Declarative	Ryanair is altijd betrouwbaar.	Ryanair's always reliable.
Tag-question	Je zusje is lief, hè?	Your sister's sweet, isn't she?
	Jullie zijn in shape, hè?	You lot are in shape, aren't you?
	Hij is bescheiden, hè?	He's modest, isn't he?
Wh-exclamative	Wat een talentvolle chef!	What an accomplished chef!
	Wat een leuke verassing!	What a nice surprise!
	Wat een attente vriend!	What a considerate friend!
	Wat een boeiend hoorcollege!	What an engaging lecture!

Table 6: Sentences that are generated by the FastSpeech2 model but were not understandable enough. These sentences were not used in the survey.

C Appendix: Visuals of survey



Wat is volgens u de emotie van de spreker?

○ Blij
○ Boos
○ Sarcastisch
○ Neutraal

Figure 4: The visual of the survey on the phone. The question (*What, in your opinion, is the speaker's emotion?*) was asked for all generated speech files. The following options could be given as answers: 'blij' (*happy*), 'boos' (*angry*), 'sarcastisch' (*sarcastic*), 'neutraal' (*neutral/sincere*).

D Appendix: Precision, recall, accuracy for the different sentence types

D.1 Declaratives

Decoy answers	extra_sarcastic	Sincere or sarcastic	Precision	Recall	Accuracy (F1)
Yes	Yes	Sincere	0.34	0.48	0.40
Yes	Yes	Sarcastic	0.74	0.24	0.37
Yes	No	Sincere	0.50	0.48	0.49
Yes	No	Sarcastic	0.55	0.21	0.30
No	Yes	Sincere	0.34	0.74	0.47
No	Yes	Sarcastic	0.74	0.35	0.47
No	No	Sincere	0.50	0.74	0.60
No	No	Sarcastic	0.55	0.31	0.39

Table 7: Table containing the precision, recall, and accuracy (F1) for both sincere and sarcastic, and including decoy answers or not and including the *extra_sarcastic* data. The data that is used here is only from the declarative sentences.

D.2 Wh-exclamatives

Decoy answers	extra_sarcastic	Sincere or sarcastic	Precision	Recall	Accuracy (F1)
Yes	Yes	Sincere	0.37	0.49	0.42
Yes	Yes	Sarcastic	0.80	0.25	0.39
Yes	No	Sincere	0.52	0.49	0.50
Yes	No	Sarcastic	0.59	0.18	0.27
No	Yes	Sincere	0.37	0.80	0.51
No	Yes	Sarcastic	0.80	0.38	0.52
No	No	Sincere	0.52	0.80	0.63
No	No	Sarcastic	0.59	0.29	0.38

Table 8: Table containing the precision, recall, and accuracy (F1) for both sincere and sarcastic, and including decoy answers or not and including the *extra_sarcastic* data. The data that is used here is only from the wh-exclamatives.

D.3 Tag-questions

Decoy answers	**extra_sarcastic**	**Sincere or sarcastic**	**Precision**	**Recall**	**Accuracy (F1)**
Yes	Yes	Sincere	0.34	0.54	0.41
Yes	Yes	Sarcastic	0.64	0.11	0.19
Yes	No	Sincere	0.48	0.11	0.18
Yes	No	Sarcastic	0.34	0.81	0.48
No	Yes	Sincere	0.64	0.17	0.27
No	Yes	Sarcastic	0.64	0.17	0.27
No	No	Sincere	0.51	0.81	0.63
No	No	Sarcastic	0.48	0.18	0.26

Table 9: Table containing the precision, recall, and accuracy (F1) for both sincere and sarcastic, and including decoy answers or not and including the *extra_sarcastic* data. The data that is used here is only from the tag-questions.