university of groningen

campus fryslân

Identifying Acoustic Features that Enhance TTS Voice Intelligibility and Naturalness in Noisy Environments

Jingxuan Yue

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Identifying Acoustic Features that Enhance TTS Voice Intelligibility and Naturalness in Noisy Environments

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
Supervisor Dr. M. Coler (Voice Technology, University of Groningen)
with the second reader being
Supervisor (Voice Technology, University of Groningen)

Jingxuan Yue (S5657660)

June 29, 2024

Acknowledgements

In the process of completing this thesis, I have received generous help and support from many people. I would like to express my heartfelt gratitude to them here.

First and foremost, I would like to deeply thank my supervisor, Matt. Your professional insights and selfless guidance have been pivotal throughout my research journey. You always provided valuable suggestions and inspirational feedback promptly, helping me continually refine my thesis. Your encouragement and support have not only led me to explore the field of voice technology but have also made me more confident on my academic path.

Secondly, I would like to thank my family, partner, and friends. Your unconditional support and love have been my driving force. In times of difficulty and challenge, you have always given me boundless care and understanding, allowing me to fully immerse myself in my research. Your companionship and encouragement have helped me find balance in my busy academic life and pursue my dreams with determination.

I would also like to extend my sincere gratitude to all individuals who participated in my listening tests. Thank you for your immense contribution to this research. There are many more thanks that I cannot express individually, but I will transform all the kindness and love I have received into motivation to continue exploring my academic journey.

Abstract

With the continuous advancement of voice technology, the application of TTS (Text-to-Speech) in daily life has become increasingly widespread. However, the acoustic environments in practical application scenarios are complex and variable, filled with different levels of noise, which poses challenges to the intelligibility and naturalness of TTS voice. Research indicates that speech with Lombard speech characteristics has higher intelligibility in noisy environments. Therefore, to identify and understand which acoustic features can effectively enhance the intelligibility and naturalness of synthetic speech in noisy conditions, this study conducted systematic and comprehensive experiments. The results show that enhancing F0 independently can significantly improve the intelligibility of synthetic speech in noisy environments; while increasing duration independently can enhance intelligibility, it also decreases naturalness. On the other hand, typical Lombard speech and solely flattening spectral tilt have no effect on improving naturalness, providing valuable insights for the development of more adaptive and user-centered TTS systems.

Keywords: Text-to-Speech, Acoustic features, Intelligibility, Naturalness, Noisy environments

Contents

1 Introduction

In today's rapidly advancing world of voice technology, the intelligibility and naturalness of Text-to-Speech (TTS) systems have significantly improved, leading to a wider range of everyday applications. From public transportation announcements and in-car navigation to ubiquitous voice assistants, synthetic speech permeates various aspects of daily life. However, assessments of TTS intelligibility and naturalness are typically conducted in quiet environments. Real-world applications, on the other hand, are characterized by diverse acoustic environments filled with varying levels of noise, such as bustling train stations during peak hours or lively conversations in cafes. These ambient noises can diminish the intelligibility and naturalness of TTS output, negatively affecting the quality of user interaction with synthetic speech. Therefore, investigating ways to enhance the intelligibility and naturalness of synthetic speech in noisy environments holds significant practical value for improving user experience in voice technology-based interactions.

When discussing research on speech in noisy environments, an important concept that cannot be overlooked is the Lombard effect. This phenomenon refers to the natural adjustments humans make to their speech in noisy environments to enhance intelligibility (Lombard, 1911). Key characteristics of Lombard speech include increased intensity and fundamental frequency (F0), along with increased duration and flattened spectral tilt. Understanding these adjustments has laid the groundwork for improving TTS voice in noisy environments by emulating the Lombard effect, which has been shown to improve intelligibility in numerous studies.

However, most of these studies have primarily focused on enhancing intelligibility, while the impact of such modifications on the naturalness of synthetic speech remains relatively underexplored. For example, the perception of naturalness in Lombard speech and the impact of flattening spectral tilt on speech naturalness have not been thoroughly investigated. Moreover, current research often adjusts all typical features of Lombard speech uniformly to simulate its effects, with few studies independently modifying different acoustic features to explore their individual contributions. This gap suggests a need for a comprehensive investigation that not only assesses intelligibility but also evaluates the naturalness of synthetic speech in noisy settings, while also independently modifying individual acoustic features to clarify their specific contributions.

Therefore, this research aims to systematically investigate how variations in key acoustic features influence the intelligibility and naturalness of synthetic speech in noisy environments. Drawing on the Lombard effect, this study independently identifies key acoustic features and uses a state-of-the-art TTS model to systematically modify these features. The study utilizes environmental audio recordings from a bustling train station during peak hours to simulate the noisy conditions. These recordings are combined with different synthetic speech samples to conduct a comparative study, aiming to identify the impact of different acoustic features on enhancing the intelligibility and naturalness of synthetic speech in noisy environments. Detailed methods will be outlined in the methodology section.

As a result, this research contributes to a deeper comprehension of the interplay between acoustic features and user perception. It facilitates the development of more intelligible and natural synthetic speech experiences across real-world environments. Furthermore, the insights gained from this study can inform the design of future TTS systems, ensuring they are better equipped to handle the challenges posed by noisy environments, thereby improving user satisfaction and communication effectiveness in practical applications.

The thesis is structured as follows. Section 2 offers a comprehensive literature review that contex-

tualizes the research question and hypothesis within the broader field. Section 3 covers the methodology, detailing the steps taken throughout the research process. Section 4 presents the results obtained and analyzes the findings. Section 5 thoroughly discusses the result and indicates the limitation. Finally, Section 6 summarizes the key points, emphasizing significant conclusions.

2 Literature Review

This study conducted a comprehensive literature search using authoritative academic databases, including IEEE Xplore, JSTOR, SpringerLink, arXiv, and Google Scholar. Keywords used in the search included “Lombard effect”, “Lombard speech”, “synthetic speech” and “noisy environments”. Highly cited, peer-reviewed studies directly related to synthetic speech in noisy environments were selected and further filtered based on their relevance to the research questions. In total, over 30 key articles were chosen for in-depth analysis, covering methods, findings, and research gaps. This thorough literature review process ensures a comprehensive understanding of the current state of research and provides a crucial basis for the research questions and hypotheses.

This section’s literature review is organized into the following parts. Firstly, subsection 2.1 introduces the critical concept of the Lombard effect, which serves as the theoretical foundation for many TTS systems adapted to noisy environments. Following this, subsection 2.2 delves into the acoustic characteristics of speech generated in noisy environments—namely, Lombard speech—and the research on how the modification of these characteristics enhances speech intelligibility. Subsection 2.3 then focuses on the synthesis of Lombard speech, discussing studies and practices that aim to improve TTS voice intelligibility by simulating Lombard speech. By reviewing and analyzing these studies, this section identifies the major gaps in the current research, thereby leading to the formulation of the research questions and hypotheses for this study.

2.1 Lombard Effect

When studying speech characteristics in different environments, the Lombard effect remains a crucial concept. The Lombard effect is named after the French otolaryngologist Etienne Lombard, who investigated the impact of noisy environments on speech production. He discovered that speakers increase their vocal intensity in noisy settings (Lombard, 1911) . This phenomenon, where speakers adjust their vocal effort in response to noise, resulting in changes in their speech production, is known as the Lombard Effect. Subsequent researchers have built upon this concept to study specific changes in speech characteristics caused by the Lombard effect (Castellanos, Benedí, & Casacuberta, 1996; Junqua, 1993, 1996; Summers, Pisoni, Bernacki, Pedlow, & Stokes, 1988), laying the foundation for research on speech characteristics in various environments.

Researchers have different perspectives on the understanding of the Lombard effect. Lombard (1911) suggested that the changes in vocal production in noisy environments are an involuntary reflex. A second perspective posits that the Lombard effect is an active regulation mechanism, where speakers consciously adjust their vocal intensity based on auditory feedback to maintain speech intelligibility in noisy environments, thereby ensuring effective communication (Lane & Tranel, 1971) . Another perspective suggests that the Lombard effect is driven by a combination of reflexive mechanisms and conscious communicative strategies (Garnier, Henrich, & Dubois, 2010). Lau (2008) experiments further demonstrated the conscious communicative aspect of the Lombard effect by controlling the noise environment for both listeners and speakers. Even when only the listener is in a noisy environment, speakers in a quiet setting will adjust their vocal production to enhance speech intelligibility.

This conscious communicative strategy indicates that speakers actively adjust their vocal production in noisy environments to enhance speech intelligibility and ensure clear transmission of information. This spontaneous adjustment reflects the human need for high-quality communication

and our expectation for highly intelligible speech.

2.2 Lombard Speech

Based on the Lombard effect, researchers have conducted a series of studies on Lombard speech — speech produced in noisy environments. Research on Lombard speech can be divided into two main areas: first, the study of the acoustic characteristics of Lombard speech, specifically how these characteristics differ from normal speech produced in quiet environments. Second, the study of the intelligibility of Lombard speech, focusing on whether Lombard speech has improved intelligibility compared to normal speech produced in quiet environments, and which modifications in speech characteristics contribute to this improvement in intelligibility.

2.2.1 Acoustic Features of Lombard Speech

A series of studies have pointed out that compared to speech produced in quiet environments, Lombard speech produced in noisy environments has the following main acoustic characteristics: increased vocal intensity, increased fundamental frequency (F0), increased duration, a shift in energy from low-frequency bands to middle or high bands, a shift in formant center frequencies, and flatter spectral tilting (Junqua, 1993, 1996; Lane & Tranel, 1971; Summers et al., 1988) .

The aforementioned studies on the characteristics of Lombard speech are primarily based on English. Additionally, Spanish (Castellanos et al., 1996) and Czech (Boril & Pollak, 2005) have also been used as reference languages for studying the acoustic characteristics of Lombard speech, with results consistent with those found in English Lombard speech. This indicates that the acoustic characteristics of Lombard speech have a cross-linguistic commonality.

Multimodal speech research and interaction studies have further validated the acoustic characteristics of Lombard speech. In studies of audio-visual Lombard speech that include visual cues, Davis, Kim, Grauwinkel, and Mixdorff (2006) for the English corpus and Garnier, Bailly, Dohen, Welby, and Loevenbruck (2006) for the French corpus both confirmed that the acoustic characteristics of multimodal Lombard speech are consistent with previous studies, further demonstrating the universality of Lombard speech characteristics. Experimental studies involving interaction and communication (Garnier & Henrich, 2014; Lau, 2008) also validated these common characteristics of Lombard speech, highlighting the importance and practicality of Lombard speech research in real-world speech interaction studies.

In studies on the synthesis of Lombard speech, many researchers have compared normal speech and Lombard speech using statistical data on acoustic characteristics (Cooke, Mayo, & Villegas, 2014; Novitasari, Sakti, & Nakamura, 2022; Valentini-Botinhao, Yamagishi, King, & Stylianou, 2013). These studies consistently confirmed the acoustic characteristics of increased intensity, F0, duration, and flattened spectral tilt in Lombard speech.

2.2.2 Intelligibility of Lombard Speech

Dreher and O'Neill (1957) reported that, when presented at a constant speech-to-noise ratio, speech produced in noisy environments is more intelligible than speech produced in quiet environments. This conclusion was validated by Summers et al. (1988), whose experiments showed that listeners in noisy environments had a higher recognition accuracy for speech produced in noise compared to

speech produced in quiet environments. This indicates that Lombard speech has higher intelligibility in noisy environments. Subsequent research further confirmed this conclusion (Godoy & Stylianou, 2012; Lu & Cooke, 2009; Pittman & Wiley, 2001).

These findings have prompted subsequent research to explore which changes in acoustic characteristics are key to improving the intelligibility of Lombard speech (Garnier & Henrich, 2014; Godoy & Stylianou, 2012; Lu & Cooke, 2009; Zorila, Kandia, & Stylianou, 2012), in order to understand the acoustic properties that effectively enhance intelligibility.

The key acoustic characteristic that has been clearly confirmed is spectral tilt. Studies by Cooke et al. (2014), Lu and Cooke (2009), and Zorila et al. (2012) all indicated that flatter spectral tilt significantly improves speech intelligibility. Flattening the spectral tilt effectively shifts the speech energy to higher frequency regions that are less likely to be masked by noise, thereby increasing the audible portions of the speech. The research by Godoy and Stylianou (2012) found a consistent increase in energy in the frequency range of 500-4500 Hz (which includes the formant regions), further supporting the notion that spectral adjustments and increased energy in the higher frequency regions are crucial for enhancing the intelligibility of Lombard speech.

However, other major characteristics of Lombard speech, such as duration and f0, have not been sufficiently proven to enhance intelligibility. The experimental results by Cooke et al. (2014) showed that speech with increased duration in Lombard speech tends to have flatter spectral tilts. This positive correlation between increased duration and flatter spectral tilts suggests the potential role of duration in enhancing intelligibility. However, this study did not directly verify the impact of duration on intelligibility, which requires further research for confirmation. Similarly, the f0 of Lombard speech has not been sufficiently proven to enhance intelligibility. The study by Lu and Cooke (2009) modulated F0 and spectral tilt separately and in combination respectively, finding that increasing F0 alone did not significantly improve speech intelligibility. Therefore, further research is needed to explore the impact of F0 on the intelligibility of Lombard speech in noise environments.

Based on the findings from the Lombard effect and Lombard speech studies, recent research has begun to apply these discoveries to the development and optimization of TTS systems in various acoustic environments, aiming to improve the intelligibility of synthetic speech and thereby enhance user interaction in different settings.

2.3 Lombard Speech Synthesis

In recent years, to enhance the performance of TTS voice in complex real-world environments, researchers have based their work on the Lombard effect, simulating changes in the acoustic characteristics of Lombard speech to improve the intelligibility of synthetic speech in noisy environments.

Early simulation work involved modifying synthetic speech by statistically analyzing the acoustic parameters of normal speech and Lombard speech. The main modifications included extending duration, increasing F0, and increasing intensity (Huang & Ong, 2010; Patel, Everett, & Sadikov, 2006). Additionally, the experiments by Cooke et al. (2013) and López, Seshadri, Juvela, Räsänen, and Alku (2017) involved reducing spectral tilt. These modifications achieved higher intelligibility compared to unmodified TTS systems. Moreover, some studies used the Hidden Markov Model (HMM) framework to modify speech parameters, similarly enhancing the intelligibility of synthetic speech in noisy environments by extending speech duration, increasing F0, and reducing spectral tilt (Raitio, Suni, Vainio, & Alku, 2011; Suni et al., 2013) .

Subsequently, deep neural networks (DNNs) have been applied to the synthesis of Lombard speech, achieving end-to-end synthesis. For example, Bollepalli, Juvela, and Alku (2019) utilized transfer learning to synthesize Lombard speech in a Seq2Seq TTS system combined with the WaveNet vocoder, effectively improving intelligibility. Paul, Shifas, Pantazis, and Stylianou (2020), based on Tacotron and WaveRNN, employed spectral shaping and dynamic range compression, and both objective and subjective measurements showed that this method enhanced intelligibility. Hu et al. (2021) , using an optimized DNN model—Tacotron 2, also implemented spectral shaping and dynamic range compression. Multiple experiments, including word error rate and subjective listening tests, indicated that the synthesized Lombard speech had significantly higher intelligibility in noisy environments compared to normal speech. Novitasari et al. (2022) and Novitasari, Sakti, and Nakamura (2021) further developed a dynamically adaptive TTS system for noisy environments by using data augmentation to train the TTS system with modifications in the intensity, F0, and duration of normal speech. Ultimately, the system with auditory feedback successfully produced highly intelligible speech, surpassing the performance of standard TTS systems.

Whether based on HMMs or DNNs, different model frameworks for simulating Lombard speech to improve intelligibility commonly involve modifications such as increasing intensity, increasing F0, increasing duration, and reducing spectral tilt. These modifications have achieved enhanced intelligibility of synthetic speech in noisy environments.

With the increasing application of synthetic speech in noisy environments, not only does intelligibility become important, but listeners' perception of naturalness is also crucial. However, research on naturalness is relatively insufficient. This leads to on of the research question: "When the intelligibility of speech is significantly enhanced, specifically through the combined adjustment of acoustic features and the flattening of spectral tilt, will the naturalness of the speech also improve?"

The study by Van Ngo, Kubo, and Akagi (2019) further prompted reflection on the contribution of different acoustic characteristics to the enhancement of intelligibility and naturalness. In their research, the simulated Lombard speech showed the best performance in terms of intelligibility and naturalness when only spectral tilt was adjusted. However, when other features (such as F0) were also adjusted, both intelligibility and naturalness decreased. This study not only further validated the findings of Cooke et al. (2014), Godoy and Stylianou (2012), and Lu and Cooke (2009), that spectral tilt can effectively improve intelligibility, but also highlights the complex relationships between different acoustic characteristics and the necessity of studying each feature independently for its contribution to intelligibility enhancement.

In the context where the effects of F0 and duration have not been fully verified, this research background and gap lead to another research question: "Can independently raising F0 or duration improve the intelligibility and naturalness of TTS voice in noisy environments?"

The literature review reveals that simulating Lombard speech can enhance the intelligibility of TTS voice in noisy environments. However, how these adjustments in acoustic features affect the perceived naturalness of TTS voice, especially in practical application environments, remains underresearched. This gap indicates the need for a more comprehensive investigation that not only evaluates intelligibility but also examines the naturalness of TTS voice. Meanwhile, there is a lack of studies that independently adjust specific acoustic features, such as F0 and duration, to determine their unique effects on intelligibility and naturalness in synthetic speech. The review indicates that the contributions of duration and F0 to intelligibility have not been fully verified and require further validation. Therefore, research should consider the acoustic environments of practical applications and modify individual acoustic features to clarify their specific contributions to both intelligibility

and naturalness.

This study proposes a systematic exploration of how changes in key acoustic features affect the intelligibility and naturalness of TTS voice in noisy environments, aiming to most effectively enhance the intelligibility and naturalness of synthetic speech, thereby improving user interaction in real-world applications. By establishing subsequent research questions and hypotheses, the impact of different acoustic features on enhancing speech intelligibility and naturalness can be further clarified, advancing the TTS field towards more adaptive and user-centered solutions.

2.4 Research Question and Hypothesis

Based on the previous discussion, it is known that combining adjustments to the typical acoustic characteristics of Lombard speech and independently adjusting spectral tilt can effectively enhance the intelligibility of TTS voice in noisy environments. However, the perceived naturalness of voices with these adjustments in practical noisy environments needs further research. This leads to the first two research questions.

Furthermore, the contributions of individual acoustic feature adjustments to both intelligibility and naturalness have not been fully studied. The specific effects of F0 and duration on intelligibility and naturalness still require further verification. This leads to the subsequent two research questions.

Therefore, my summary of the research questions is as follows:

1. Does the combined adjustment of acoustic features (increasing intensity, raising F0, increasing duration, and flattening spectral tilt) improve the naturalness of TTS voice in noisy environments?

2. Can independently flattening spectral tilt improve the naturalness of TTS voice in noisy environments?

3. Can independently raising F0 improve the intelligibility and naturalness of TTS voice in noisy environments?

4. Can independently increasing duration improve the intelligibility and naturalness of TTS voice in noisy environments?

Current research on the naturalness of Lombard speech is limited, but the few existing studies suggest a potential positive correlation between naturalness and intelligibility (Van Ngo et al., 2019). Therefore, this study hypothesizes that acoustic features which enhance the intelligibility of synthetic speech in noisy environments can also improve naturalness. Based on this, it is hypothesized that typical Lombard speech and independently adjusting spectral tilt, both of which significantly enhance intelligibility, can also improve naturalness.

Furthermore, Cooke et al. (2014) found that speech with longer duration typically has flatter spectral tilts. Given this potential positive correlation, and in reference to spectral tilt, this study hypothesizes that increasing duration independently can also enhance both the intelligibility and naturalness of TTS voice in noisy environments. Regarding F0, based on the findings of Lu and Cooke (2009), which indicate that solely increasing F0 does not significantly enhance intelligibility, this study hypothesizes that F0 cannot improve either intelligibility or naturalness of synthetic speech in noisy settings.

Therefore, the hypotheses corresponding to the research questions are as follows:

- **H1:** Combined adjustments of acoustic features will help enhance the perceived naturalness of TTS voice in noisy environments
- **H2:** Independently flattening spectral tilt will help improve the perceived naturalness of TTS voice in noisy environments.
- **H3:** Independently raising F0 will not improve the intelligibility and naturalness of TTS voice in noisy environments.
- **H4:** Independently increasing duration in noisy environments will improve both the intelligibility and naturalness of TTS voice in noisy settings.

Through these research questions and hypotheses, the impact of changes in key acoustic features on TTS voice in practical application environments can be systematically explored. This approach determines which acoustic features and their adjustment methods (combined adjustments or individual adjustments) can enhance the intelligibility and naturalness of synthetic speech. Ultimately, important evidence will be provided for more efficiently optimizing the intelligibility and naturalness of TTS systems, thereby enhancing the user interaction experience with synthetic speech.

3 Methodology

This section details the experimental design, including the model selection and pre-trained resources, stimuli creation, intelligibility and naturalness evaluation, and data analysis. It specifically elaborates on the experimental variables and control mechanisms used to explore the specific effects of different acoustic features on intelligibility and naturalness.

3.1 Model Selection and Pre-trained Resources

In this study, selecting the appropriate model and utilizing pre-trained resources is crucial for achieving accurate and reliable results. The chosen models and resources must effectively facilitate the manipulation and control of various acoustic features. Detailed below is the process of model selection and the specific pre-trained resources leveraged to ensure the robustness of the experimental outcomes.

3.1.1 Text-to-Speech Model: FastSpeech 2

This experiment employs the advanced end-to-end TTS model, FastSpeech 2, for speech synthesis. Firstly, FastSpeech 2 introduces unique duration, energy, and pitch predictors, enabling the model to flexibly control acoustic features such as duration, intensity, and F0, and directly generate speech with varying acoustic characteristics. Therefore, utilizing FastSpeech 2 allows for independent or combined adjustments of F0, duration, and intensity, which directly meets the conditions required by the research questions of this study. This is the primary motivation for selecting the FastSpeech 2 model.

Additionally, coupled with the optimized length regulator and improved loss functions, FastSpeech 2 can better capture subtle speech features, producing more natural and higher-quality speech. FastSpeech 2 also directly predicts speech features for all time steps, avoiding the attention misalignment issues commonly found in attention-based models, thus enhancing the stability and robustness of the generated speech. These characteristics ensure that FastSpeech 2 can generate stable and high-quality speech, providing a solid foundation for exploring the intelligibility and naturalness of synthetic speech in noisy environments in this study.

3.1.2 Vocoder: HiFi-GAN

For the vocoder, this study employs the advanced neural vocoder HiFi-GAN. HiFi-GAN is based on a generative adversarial network (GAN) architecture. Through adversarial training, the generator produces more realistic and natural speech. This optimization of audio quality fidelity results in more intelligible and natural speech, which is crucial for studying the subtle effects of acoustic feature modifications on intelligibility and naturalness in noisy environments. Its robust architecture ensures that the detailed control over acoustic features provided by FastSpeech 2—such as pitch, duration, and energy adjustments—are accurately reflected in the final synthesized speech. This compatibility allows precise manipulation and evaluation of these features to understand their impact on intelligibility and naturalness. These advantages make HiFi-GAN well-aligned with the objectives of this research.

By using HiFi-GAN, this study ensures that the synthesized speech maintains a high level of quality and naturalness, which is critical for accurately assessing the impact of acoustic feature adjustments on intelligibility and naturalness in noisy environments. The combination of FastSpeech 2 and HiFi-GAN provides a powerful and flexible framework for generating and analyzing synthetic speech, supporting a detailed exploration of the research questions and hypotheses.

3.1.3 Pre-trained Checkpoints and Dataset: LJ Speech

This study uses an open-source pre-trained model[1] trained on the LJ Speech dataset with 900,000 steps for the following reasons. Firstly, the LJ Speech dataset offers high audio quality, with clear recordings and no significant background noise. This is crucial for training high-performance FastSpeech 2 models, ensuring that the generated speech is both natural and intelligible. The high quality of the synthetic speech allows for precise manipulation of acoustic features while maintaining high intelligibility and naturalness, even in noisy environments. This quality is beneficial for conducting research in noisy environments as required by our study. Additionally, the dataset consists of 13,100 audio clips, all read by the same female speaker, providing highly consistent speech data that helps the model learn stable and rich speech features. This consistency minimizes variability caused by speaker differences and enhances the reliability of the experimental results.

3.2 Stimuli Generation

This section details the process of generating stimuli for the experiments. The process includes selecting appropriate noisy environments and preparing the audio samples necessary for evaluating the TTS system's performance. By carefully designing the stimuli, the experimental conditions closely replicate real-world scenarios, thereby enhancing the practical relevance of the findings.

3.2.1 Noisy Environment Selection

To accurately simulate real-world application scenarios of TTS voice, it is essential to select appropriate acoustic environments for evaluating intelligibility and naturalness. Therefore, this study uses soundscape audio from the open-source soundscape library freesound.org[2], focusing on a train station environment (80 dB), which is one of the most common noisy TTS application environments. This choice ensures that the findings are relevant and applicable to practical use cases.

3.2.2 Synthetic Speech Modification

To test the intelligibility and naturalness of different acoustic features in the selected noisy environment, specific modifications were made to the modules of FastSpeech 2 to allow precise control of pitch, duration, and intensity. Specifically, standardization was added in the VarianceAdaptor class to account for the mean and standard deviation of pitch and energy, and adjustments were made in calculating the control parameters. Details of these modifications can be seen in Figure 1, where the specific changes have been highlighted. For a complete view of the code, please refer to this

[1] https://github.com/ming024/FastSpeech2

[2] https://freesound.org/

repository[3]. These enhancements enable more accurate control over variations in pitch and energy, ensuring that the generated speech aligns more closely with the intended adjustments. These modifications are crucial for generating synthetic speech samples that accurately reflect adjustments in different acoustic features.

```python
class VarianceAdaptor(nn.Module):
    """Variance Adaptor"""

    def __init__(self, preprocess_config, model_config):
        super(VarianceAdaptor, self).__init__()
        self.duration_predictor = VariancePredictor(model_config)
        self.length_regulator = LengthRegulator()
        self.pitch_predictor = VariancePredictor(model_config)
        self.energy_predictor = VariancePredictor(model_config)

        self.pitch_feature_level = preprocess_config["preprocessing"]["pitch"][
            "feature"
        ]
        self.energy_feature_level = preprocess_config["preprocessing"]["energy"][
            "feature"
        ]
        assert self.pitch_feature_level in ["phoneme_level", "frame_level"]
        assert self.energy_feature_level in ["phoneme_level", "frame_level"]

        pitch_quantization = model_config["variance_embedding"]["pitch_quantization"]
        energy_quantization = model_config["variance_embedding"]["energy_quantization"]
        n_bins = model_config["variance_embedding"]["n_bins"]
        assert pitch_quantization in ["linear", "log"]
        assert energy_quantization in ["linear", "log"]
        with open(
            os.path.join(preprocess_config["path"]["preprocessed_path"], "stats.json")
        ) as f:
            stats = json.load(f)
            pitch_min, pitch_max = stats["pitch"][:2]
            energy_min, energy_max = stats["energy"][:2]
            self.pitch_mean, self.pitch_std = stats["pitch"][2:4]
            self.energy_mean, self.energy_std = stats["energy"][2:4]

    def get_pitch_embedding(self, x, target, mask, control):
        prediction = self.pitch_predictor(x, mask)
        if target is not None:
            embedding = self.pitch_embedding(torch.bucketize(target, self.pitch_bins))
        else:
            prediction = ((prediction * self.pitch_std +self.pitch_mean) * control - self.pitch_mean)/self.pitch_std
            embedding = self.pitch_embedding(
                torch.bucketize(prediction, self.pitch_bins)
            )
        return prediction, embedding

    def get_energy_embedding(self, x, target, mask, control):
        prediction = self.energy_predictor(x, mask)
        if target is not None:
            embedding = self.energy_embedding(torch.bucketize(target, self.energy_bins))
        else:
            prediction = ((prediction * self.energy_std +self.energy_mean) * control - self.energy_mean)/self.energy_std
            embedding = self.energy_embedding(
                torch.bucketize(prediction, self.energy_bins)
            )
        return prediction, embedding
```

Figure 1: Modifications to VarianceAdaptor for Precise Control of Pitch and Energy

After modifying the FastSpeech 2 code, duration, F0, and intensity can be precisely controlled via the command line. For another critical parameter in our study, spectral tilt, adjustments are made by increasing the energy in the high-frequency region using Python libraries including numpy, librosa, and scipy. The adjustment ratios are based on the parameter changes observed between Lombard speech and normal speech in the studies by Novitasari et al. (2022) and Valentini-Botinhao et al. (2013). Finally, this study generates five types of synthetic speech with the following parameter configurations:

- **Config1:** Synthetic speech without any parameter adjustments (Baseline)
- **Config2:** Duration increased by 15%
- **Config3:** F0 increased by 20%
- **Config4:** Energy in the high-frequency region (greater than 2000 kHz) increased by 15%
- **Config5:** Duration increased by 15%, F0 increased by 20%, high-frequency region (greater than 2000 kHz) energy increased by 15%, and overall intensity increased by 5%

These specific adjustments allow us to systematically investigate the impact of different acoustic features on speech intelligibility and naturalness.

[3] https://github.com/jingxuan16/Thesis-Project

3.2.3 Detailed Stimuli Preparation

With the parameter configurations established, the stimuli are generated. Following the standard of Harvard sentences, three sentences that simulate actual announcements made in train stations are used, making the experiment more relevant to real-world applications. For each sentence, the above five configurations are applied, resulting in a total of 15 audio samples. The specific parameters for the three sets of sentences are shown in Tables 1, 2, and 3 respectively.

Table 1: Configurations of Speech Features for Sentence 1

Config	**intensity (dB)**	**duration (s)**	**mean f0 (Hz)**	**spectral tilt (dB/octave)**
1	76.87	2.80	212.59	-1.33
2	76.98	3.20	213.29	-1.31
3	76.88	2.80	253.53	-1.28
4	79.36	2.80	212.59	-0.85
5	82.86	3.20	252.36	-1.02

Table 2: Configurations of Speech Features for Sentence 2

Config	**intensity (dB)**	**duration (s)**	**mean f0 (Hz)**	**spectral tilt (dB/octave)**
1	76.44	3.40	188.99	-1.55
2	76.72	3.89	188.65	-1.57
3	76.89	3.40	229.28	-1.53
4	77.84	3.40	188.99	-1.07
5	82.2	3.89	227.47	-1.31

Table 3: Configurations of Speech Features for Sentence 3

Config	**intensity (dB)**	**duration (s)**	**mean f0 (Hz)**	**spectral tilt (dB/octave)**
1	75.29	3.00	191.16	-1.85
2	75.07	3.46	190.42	-1.86
3	75.18	3.00	228.23	-1.76
4	76.36	3.00	191.16	-1.24
5	81.06	3.46	226.31	-1.47

Next, to create the final experimental stimuli, the Audacity tool is utilized to combine the 15 generated sentences with the environmental audio. The environmental audio is edited to ensure

consistency, with each stimulus beginning and ending with 1.5 seconds of ambient noise to create a more immersive environment and help listeners adjust to the setting. This process ensures that each synthetic speech sample is tested in a realistic noisy environment, providing the 15 audio samples required for the experiment. Each combined audio file is carefully reviewed to ensure the quality and consistency of the stimuli.

3.3 Intelligibility and Naturalness Evaluation

For the aforementioned generated stimuli, this study utilizes both subjective and objective methods to evaluate their intelligibility and naturalness.

3.3.1 Subjective Evaluation

In the subjective testing phase, this study employs an online listening test. 65 participants, all with normal hearing and proficient in English, are recruited to complete the test in a quiet environment while wearing headphones. These criteria ensure that participants can accurately perceive and evaluate the intelligibility and naturalness of the speech samples without bias from hearing impairments or language difficulties. Conducting the test in a quiet environment with headphones minimizes external noise interference, ensuring the reliability and consistency of the test results.

For the questionnaire design, the study intersperses stimuli from different sets, presenting various sentence contents alternately. Additionally, three audio samples that are not included in the final score are inserted to reduce the promising effect. After listening to each audio sample, participants are asked to rate the intelligibility and the human-likeness of the audio on a scale from 1 to 5.

3.3.2 Objective Evaluation

For the objective evaluation phase, this study employs the advanced automatic speech recognition (ASR) model Whisper[4] to transcribe the stimuli. Whisper's high accuracy and robustness in handling various speech types make it an ideal choice for this experiment.

The transcription results from Whisper are then analyzed using the Python library JiWER[5], which calculates the Word Error Rate (WER) and Character Error Rate (CER). WER measures the percentage of incorrectly recognized words, while CER measures the percentage of incorrect characters. These metrics provide a quantitative assessment of the intelligibility of the synthetic speech samples in noisy environments.

By comparing the WER and CER values across different configurations, this study objectively determines the impact of each acoustic adjustment on speech intelligibility. This complements the subjective testing phase, providing a comprehensive assessment of the synthetic speech's performance.

3.4 Data Analysis

To systematically evaluate the impact of different acoustic features and their adjustment methods on the intelligibility and naturalness of TTS voice, this study conduct comprehensive data analysis on

[4] https://github.com/openai/whisper

[5] https://github.com/jitsi/jiwer

the experimental results.

3.4.1 Subjective Evaluation Analysis

For the subjective testing phase, the analysis begins with collecting intelligibility and naturalness scores from 65 participants for the five stimuli in each set. The analysis includes evaluations at both the individual sentence set level and the combined level. It encompasses three main aspects:

1. Descriptive Statistics: Calculate the mean, standard deviation, and quartiles for intelligibility and naturalness scores, providing a summary of the distribution and central tendency.

2. t-Test: Compare the means of intelligibility and naturalness scores between different configurations and the baseline to determine statistically significant differences.

3. Wilcoxon Signed-Rank Test: Use this non-parametric test to compare median differences in scores, providing an alternative validation method when data does not follow a normal distribution.

By conducting analyses at both the individual set level and the combined level, this study ensures a comprehensive evaluation of the effects of different acoustic feature adjustments on the intelligibility and naturalness of synthetic speech samples. These analyses provide robust statistical evidence to support the research hypotheses.

3.4.2 Objective Evaluation Analysis

The data for objective intelligibility analysis was obtained by inputting the synthetic speech of different configurations into Whisper, the ASR system, generating corresponding transcription texts. The CER and WER for each configuration and sentence were calculated using the JiWER library. The statistical results will be visualized to show the CER and WER performance of each configuration across the three sentences, clearly illustrating the trends relative to the baseline configuration.

The results of the both subjective result and objective evaluation result are detailed in Section 4, aiding in understanding the effectiveness of various configurations in enhancing TTS voice intelligibility. These specific results are compared with the subjective test outcomes to provide a comprehensive evaluation of the configurations.

3.5 Ethical Considerations

Ethical considerations are paramount in this study to ensure the protection and respect of all participants. The research adheres to the following ethical guidelines:

Informed Consent: Participants were fully informed about the study's nature and purpose through an information sheet and were asked to sign an informed consent form. This ensured participants were aware of their rights, including the right to withdraw from the study at any time without consequences.

Voluntary Participation: Participation was entirely voluntary. They could choose not to answer any questions and could withdraw from the study at any point without providing a reason.

Privacy and Confidentiality: The study adhered to strict privacy and confidentiality guidelines. All collected data were anonymized and securely stored in accordance with the GDPR rules of the University of Groningen. Personal information was not disclosed to anyone outside the study team, and names were not published in any reports or publications.

Data Management: Data, including consent forms, recordings, and transcripts, were securely stored on the University of Groningen server for five years, as required by university GDPR legislation. Anonymized data may be used for further research or academic publications, ensuring ongoing confidentiality.

These ethical practices ensured that the study was conducted with the highest level of integrity and respect for participants, aligning with the ethical standards required for academic research. By adhering to these guidelines, the study not only protected participants but also enhanced the validity and reliability of the research findings.

This concludes the methodology section that provides a high-level overview of the methods employed in this research. In the next section, the results obtained by applying the detailed methodologies will be presented.

4 Results

In this section, the results of the analysis based on data collected from subjective listening tests and objective evaluation of synthetic speech in noisy environments are presented. The results are structured into two main parts: subjective evaluation result and objective evaluation result.

4.1 Subjective Evaluation Results

The subjective test results are derived from the evaluations provided by 65 participants on the intelligibility and naturalness of the stimuli. Each sentence has five different configurations, and the results for each sentence are presented separately. Finally, the combined results of all three sentences are analyzed to provide a comprehensive overview.

4.1.1 Subjective Evaluation for Sentence 1

First, the descriptive statistics for the intelligibility and naturalness scores of the five different configurations of Sentence 1 are presented.

Table 4: Descriptive Statistics for Intelligibility in Sentence 1

Config	Mean	Median	Max	Min	Std	Var
1	2.89	3.00	5.00	1.00	0.95	0.91
2	3.09	3.00	5.00	1.00	1.10	1.21
3	3.71	4.00	5.00	1.00	1.03	1.05
4	4.08	4.00	5.00	2.00	0.91	0.82
5	4.49	5.00	5.00	2.00	0.73	0.54

Table 4 presents the descriptive statistics for intelligibility scores of Sentence 1. The mean values of all configurations are higher than the baseline (Config 1), suggesting that the different acoustic feature adjustments in this study can potentially improve intelligibility. Among the various configurations, typical Lombard speech (Config 5) has the most significant impact on enhancing intelligibility, as it has the highest mean and median scores, and the smallest variance, indicating robust performance. However, further statistical tests are necessary to confirm whether these different configurations can reliably enhance intelligibility.

Table 5: Descriptive Statistics for Naturalness in Sentence 1

Config	Mean	Median	Max	Min	Std	Var
1	3.58	4.00	5.00	1.00	1.12	1.25
2	3.05	3.00	5.00	1.00	1.15	1.33
3	3.62	4.00	5.00	1.00	1.11	1.24
4	3.60	4.00	5.00	1.00	1.14	1.31
5	3.25	3.00	5.00	1.00	1.25	1.56

Table 5 illustrates the distribution of naturalness scores for the sentence 1. Comparing the means with the baseline, modifying F0 (Config 3) and spectral tilt (Config 4) slightly improves naturalness, while modifying duration (Config 2) and typical Lombard speech (Config 5) slightly decreases it.

However, no clear trend is evident when considering other statistical indicators. Similarly, further statistical tests are required to confirm whether these different configurations affect the enhancement of naturalness.

To rigorously test the statistical significance of different configurations on enhancing intelligibility and naturalness, t-tests were conducted. Given that some of the data deviates from normality, Wilcoxon signed-rank tests were also performed to ensure the robustness and reliability of the results. In t-tests, if the p-value is less than 0.05, the difference between the groups is considered statistically significant. Additionally, the t-value represents the difference in means between the two groups, with larger absolute t-values indicating greater differences. Similarly, in Wilcoxon tests, if the p-value is less than 0.05, the difference between the groups is considered statistically significant. If both the t-test and Wilcoxon test p-values are less than 0.05, it can be concluded that there is a statistically significant difference between the two groups.

Table 6: t-Test and Wilcoxon Test Results for Sentence 1

Aspect	Comparison Group	Test Type	t/w-value	p-value
Intelligibility	C2 vs C1	t	1.11	0.270
		Wilcoxon	275.5	0.142
	C3 vs C1	t	4.69	**6.88e-06**
		Wilcoxon	91.5	**6.78e-07**
	C4 vs C1	t	7.26	**3.39e-11**
		Wilcoxon	42.0	**9.61e-10**
	C5 vs C1	t	10.73	**1.49e-19**
		Wilcoxon	13.0	**1.35e-11**
Naturalness	C2 vs C1	t	-2.71	**0.008**
		Wilcoxon	196.5	**0.001**
	C3 vs C1	t	0.16	0.875
		Wilcoxon	407.0	0.966
	C4 vs C1	t	0.08	0.938
		Wilcoxon	341.0	0.871
	C5 vs C1	t	-1.63	0.106
		Wilcoxon	412.5	0.064

Table 6 presents the statistical test results for Sentence 1. In terms of intelligibility, for Configurations 3, 4, and 5 compared to the baseline Config 1, both the t-test and Wilcoxon test p-values are less than 0.05, and the t-values are positive. This indicates that these three adjustments significantly improve intelligibility. Additionally, the t-values show that typical Lombard speech (Config 5) has the most substantial effect on enhancing intelligibility (with the highest t-value of 10.73).

In terms of naturalness, for Configuration 2 compared to the baseline, both the t-test and Wilcoxon test p-values are less than 0.05, and the t-value is negative, indicating a statistically significant decrease in naturalness. This means that increasing duration reduces naturalness. The other three configurations do not significantly affect naturalness, but it is noteworthy that for Lombard speech (Config 5), the t-value is negative, and the p-value is close to 0.05, approaching significance in decreasing naturalness. Therefore, potential negative impacts of Lombard speech on naturalness should be considered.

4.1.2 Subjective Evaluation for Sentence 2

Similarly, the descriptive statistics for subjective evaluations of Sentence 2, along with the results of the statistical tests, are presented in the tables below.

Table 7: Descriptive Statistics for Intelligibility in Sentence 2

Config	Mean	Median	Max	Min	Std	Var
1	2.98	3.00	5.00	1.00	0.94	0.89
2	3.58	4.00	5.00	1.00	0.98	0.97
3	4.28	4.00	5.00	2.00	0.82	0.67
4	4.05	4.00	5.00	2.00	0.82	0.67
5	4.45	5.00	5.00	2.00	0.66	0.44

Table 8: Descriptive Statistics for Naturalness in Sentence 2

Config	Mean	Median	Max	Min	Std	Var
1	3.63	4.00	5.00	1.00	0.98	0.96
2	3.45	3.00	5.00	1.00	1.08	1.16
3	3.58	4.00	5.00	1.00	1.21	1.47
4	3.68	4.00	5.00	1.00	1.21	1.47
5	3.65	4.00	5.00	1.00	1.10	1.20

Table 9: t-Test and Wilcoxon Test Results for Sentence 2

Aspect	Comparison Group	Test Type	t/w-value	p-value
Intelligibility	C2 vs C1	t	3.55	**0.00054**
		Wilcoxon	105.0	**3.45e-06**
	C3 vs C1	t	8.34	**1.04e-13**
		Wilcoxon	22.0	**1.42e-09**
	C4 vs C1	t	6.85	**2.75e-10**
		Wilcoxon	64.0	**1.10e-08**
	C5 vs C1	t	10.22	**2.69e-18**
		Wilcoxon	13.5	**4.52e-11**
Naturalness	C2 vs C1	t	-1.02	0.308
		Wilcoxon	126.0	0.113
	C3 vs C1	t	-0.24	0.811
		Wilcoxon	360.0	0.662
	C4 vs C1	t	0.24	0.812
		Wilcoxon	363.5	0.916
	C5 vs C1	t	0.08	0.933
		Wilcoxon	430.0	0.995

From the descriptive statistics in Table 7, the mean and median values for the four different configurations are all higher than the baseline, which initially suggests that the different adjustments can improve intelligibility in Sentence 2. Similarly, typical Lombard speech (Config 5) has the highest

mean, median, and the smallest variance, demonstrating its optimal performance in enhancing intelligibility. Table 8 presents the statistical data for the naturalness evaluation of Sentence 2. The mean and median values for Configuration 2 are both lower than the baseline, indicating its potential negative impact on naturalness. Overall, no clear trend is evident across the various statistical indicators.

Table 9 further shows the results of the statistical tests. For intelligibility, the p-values for both the t-tests and Wilcoxon tests are less than 0.05 for all four configurations, and the t-values are positive, indicating that these four different adjustments can significantly enhance speech intelligibility. Notably, Configuration 3 has a t-value second only to Configuration 5, suggesting that solely modifying F0 has an improvement in intelligibility that is almost as effective as Lombard speech. In terms of naturalness, since all p-values are greater than 0.05, there is no statistical significance, suggesting that the four different adjustments do not have a significant impact on naturalness.

4.1.3 Subjective Evaluation for Sentence 3

In this subsection, the descriptive statistics for subjective evaluations of Sentence 3, along with the results of the statistical tests, are presented in the Tables 10, 11, and 12.

Table 10: Descriptive Statistics for Intelligibility in Sentence 3

Config	Mean	Median	Max	Min	Std	Var
1	3.82	4.00	5.00	1.00	0.83	0.68
2	4.08	4.00	5.00	1.00	0.89	0.79
3	4.51	5.00	5.00	2.00	0.71	0.50
4	4.51	5.00	5.00	2.00	0.75	0.57
5	4.85	5.00	5.00	4.00	0.36	0.13

Table 11: Descriptive Statistics for Naturalness in Sentence 3

Config	Mean	Median	Max	Min	Std	Var
1	3.63	4.00	5.00	1.00	1.01	1.02
2	3.43	3.00	5.00	1.00	1.06	1.12
3	3.74	4.00	5.00	1.00	1.19	1.41
4	3.62	4.00	5.00	1.00	1.19	1.43
5	3.51	4.00	5.00	1.00	1.30	1.69

Table 10 presents the descriptive statistics for intelligibility evaluations of Sentence 3. From the mean values, all four different adjustments improved the baseline, initially indicating a trend of enhanced speech intelligibility. Considering the median and variance values, typical Lombard speech (Config 5) remains the most effective in improving intelligibility. Additionally, adjusting F0 (Config 3) shows robust performance in enhancing intelligibility, with mean and variance values second only to Config 5. In terms of naturalness, Table 11 shows that the mean and median values for adjusting duration (Config 2) and Lombard speech (Config 5) are lower than the baseline, suggesting a potential negative impact on naturalness from these two adjustments. No clear trend is observed for the other configurations.

Table 12: t-Test and Wilcoxon Test Results for Sentence 3

Aspect	Comparison Group	Test Type	t/w-value	p-value
Intelligibility	C2 vs C1	t	1.736	0.085
		Wilcoxon	65.0	**0.00159**
	C3 vs C1	t	5.121	**1.09e-06**
		Wilcoxon	52.5	**2.16e-07**
	C4 vs C1	t	4.991	**1.92e-06**
		Wilcoxon	48.0	**1.55e-07**
	C5 vs C1	t	9.198	**8.72e-16**
		Wilcoxon	0.0	**6.29e-10**
Naturalness	C2 vs C1	t	-1.102	0.273
		Wilcoxon	192.5	0.163
	C3 vs C1	t	0.557	0.579
		Wilcoxon	286.0	0.435
	C4 vs C1	t	-0.079	0.937
		Wilcoxon	327.0	0.920
	C5 vs C1	t	-0.603	0.548
		Wilcoxon	408.0	0.417

Further statistical test results can be found in Table 12. For intelligibility, the t-test and Wilcoxon test p-values for Configurations 3, 4, and 5 are all less than 0.05, and the t-values are positive, indicating a significant impact on improving intelligibility. However, for Configuration 2, only the Wilcoxon test p-value is less than 0.05, which does not fully confirm that this adjustment significantly improves intelligibility. Similarly, the t-values indicate that Lombard speech (Config 5) and adjusting F0 (Config 3) show excellent performance in enhancing intelligibility. In terms of naturalness, since all p-values are greater than 0.05, none of the adjustments have a significant impact on naturalness.

4.1.4 Subjective Evaluation for All Sentences

After presenting the descriptive statistics and statistical test results for each individual sentence, the data for the same adjustments across the three sentences are aggregated for further analysis, allowing for a holistic assessment of the effects of various acoustic modifications.

Table 13: Descriptive Statistics for Intelligibility Across All Sentences

Config	Mean	Median	Max	Min	Std	Var
1	3.23	3.00	5.00	1.00	1.00	0.99
2	3.58	4.00	5.00	1.00	1.07	1.14
3	4.16	4.00	5.00	1.00	0.92	0.85
4	4.21	4.00	5.00	2.00	0.85	0.72
5	4.59	5.00	5.00	2.00	0.63	0.40

Table 13 presents the descriptive statistics for intelligibility scores of the same adjustments across all sentences. From the mean and median values, all four adjustments improved the baseline intelligibility, with Lombard speech still showing the best performance in enhancing intelligibility,

Table 14: Descriptive Statistics for Naturalness Across All Sentences

Config	Mean	Median	Max	Min	Std	Var
1	3.62	4.00	5.00	1.00	1.03	1.06
2	3.31	3.00	5.00	1.00	1.11	1.22
3	3.65	4.00	5.00	1.00	1.17	1.36
4	3.63	4.00	5.00	1.00	1.18	1.39
5	3.47	4.00	5.00	1.00	1.22	1.50

consistent with the results for individual sentences. The descriptive statistics for naturalness scores across all sentences can be seen in Table 14. From the mean and median values, Configurations 2 and 5 performed slightly worse than the baseline in terms of naturalness, indicating potential negative impacts. Subsequently, t-tests and Wilcoxon tests were conducted to examine the statistical significance of the overall data, and the results can be found in Table 15.

Table 15: t-Test and Wilcoxon Test Results for All Sentences

Aspect	Comparison Group	Test Type	t/w-value	p-value
Intelligibility	C2 vs C1	t	3.38	**0.00079**
		Wilcoxon	1288.0	**1.29e-07**
	C3 vs C1	t	9.60	**9.91e-20**
		Wilcoxon	486.0	**9.57e-21**
	C4 vs C1	t	10.44	**1.19e-22**
		Wilcoxon	450.0	**7.51e-23**
	C5 vs C1	t	16.16	**2.81e-45**
		Wilcoxon	81.0	**3.15e-29**
Naturalness	C2 vs C1	t	-2.84	**0.00473**
		Wilcoxon	1504.0	**0.00017**
	C3 vs C1	t	0.28	0.78282
		Wilcoxon	3272.5	0.85420
	C4 vs C1	t	0.14	0.89093
		Wilcoxon	3075.0	0.91944
	C5 vs C1	t	-1.30	0.19509
		Wilcoxon	3688.5	0.09967

In terms of intelligibility, the p-values for both the t-tests and Wilcoxon tests for all adjustments are less than 0.05. Combined with the positive t-values, this indicates that these four adjustments have a significant impact on improving overall speech intelligibility. Notably, Lombard speech (Config 5), with the highest t-value (16.16), once again proves that combined adjustments can maximally enhance intelligibility.

In terms of naturalness, Configuration 2 has p-values less than 0.05 for both the t-test and Wilcoxon test, and since the t-value is negative, it demonstrates that increasing duration (Config 2) has a significant impact on reducing the naturalness of the baseline speech. It is also noteworthy that Configuration 5 shows a trend towards decreasing naturalness (with a t-value of -1.3), and its p-value is close to 0.05, approaching significance in decreasing naturalness. This implies that while typical Lombard speech significantly enhances intelligibility, there is a potential risk of reducing naturalness, which needs to be carefully balanced in practical applications.

This approach allows comparisons within the same sentence to precisely assess the impact of different configurations and across different sentences to evaluate the consistency of these effects in varied contexts. By aggregating the results across all sentences, broader trends can be identified and more generalized conclusions can be drawn about the effectiveness of different acoustic adjustments.

4.2 Objective Evaluation Results

In the following analysis, the objective metrics of Character Error Rate (CER) and Word Error Rate (WER) are examined for different configurations across three sentences. Line graphs are used to illustrate the performance of each configuration in terms of objective intelligibility. The results are shown in the figure below.

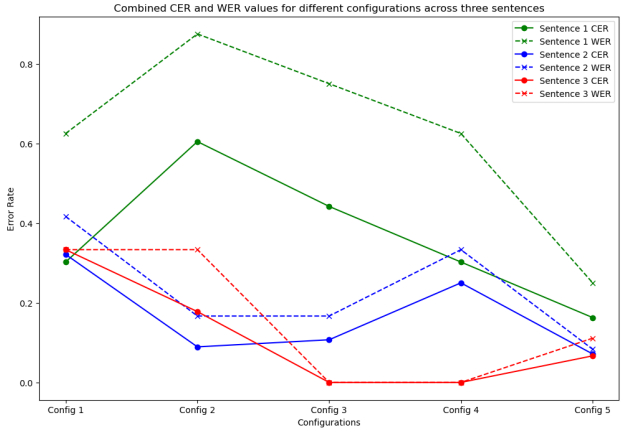

Figure 2: Combined CER and WER Values for Different Configurations Across Three Sentences

As illustrated in Figure 2, Sentence 2 and Sentence 3 exhibit a consistent trend where various configurations reduce WER and CER to varying degrees compared to the baseline Configuration 1, indicating higher objective intelligibility. This result aligns with the subjective evaluation results for Sentence 2 and Sentence 3, suggesting that all four modifications effectively enhance intelligibility. Sentence 1 is an exception where adjusting duration (Config 2) and F0 (Config 3) show increased WER and CER compared to the baseline, indicating lower objective intelligibility. This is inconsistent with the subjective evaluation results for Sentence 1, highlighting the differences between subjective and objective evaluations. This also underscores the necessity of complementing subjective and objective evaluations.

Considering the three sentences as a whole, typical Lombard speech (Config 5) demonstrates the best overall performance in terms of CER and WER, indicating higher objective intelligibility. This is consistent with the statistical performance in subjective evaluation, suggesting that typical Lombard speech significantly enhances TTS voice intelligibility in noisy environments. Similarly, adjusting spectral tilt (Config 4) also shows relatively stable performance, with error rates equal to or significantly lower than the baseline, thus improving intelligibility. This is also consistent with the subjective tests. For adjusting F0 (Config 3) and duration (Config 2), lower error rates compared to the baseline are observed in Sentence 2 and Sentence 3, which is consistent with the subjective test results in improving intelligibility in these two sentences.

Through the presentation and brief analysis of subjective test results and objective intelligibility

results, a preliminary discussion has been conducted regarding the research question and hypotheses. In the next section, the results will be thoroughly discussed, detailing how they answer the research question and whether the hypotheses are validated.

5 Discussion

Based on the data analysis results from the previous section, clear answers to the research questions and hypotheses have emerged. This section will delve into these results, examining the validation of each hypothesis to gain a more comprehensive understanding and interpretation of their significance. Additionally, potential influencing factors and explanations will be discussed.

5.1 Validation of the First Hypothesis

Hypothesis 1 posits that comprehensively adjusted speech (Config 5), which represents typical Lombard speech (with increased intensity, F0, duration, and flatter spectral tilt), can enhance the naturalness of synthetic speech in noisy environments. However, based on the descriptive statistics and correlation tests for naturalness in the subjective tests, Hypothesis 1 is rejected.

From the results of the t-tests and Wilcoxon tests, in terms of naturalness, whether analyzing individual sentences or the combined analysis of the three sentences, the correlations for Config 5 compared to the baseline do not reach the significance level (p-values are all greater than 0.05). Therefore, it cannot be concluded that comprehensively adjusted speech features can improve the naturalness of synthetic speech in noisy environments.

Additionally, it is noteworthy that in the descriptive statistics for Sentences 1 and 3, the mean naturalness scores for Config 5 are lower than those for the baseline. Moreover, in the statistical test data for these two sentences, the t-values for Config 5 compared to the baseline are negative, indicating that Config 5's naturalness performance is weaker than the baseline. Although the p-values do not indicate significant correlations, this suggests a potential negative impact of typical Lombard speech on naturalness that warrants attention.

5.2 Validation of the Second Hypothesis

Hypothesis 2 posits that solely flattening the spectral tilt can improve the naturalness of synthetic speech in noisy environments. However, based on the descriptive statistics and correlation tests for naturalness in the subjective tests, Hypothesis 2 is invalidated.

From the results of the t-tests and Wilcoxon tests, in terms of naturalness, whether analyzing individual sentences or the combined analysis of the three sentences, the p-values for speech with flattened spectral tilt (Config 4) compared to the baseline are all greater than 0.05. This indicates that the correlations do not reach the significance level. Therefore, it is evident that solely flattening the spectral tilt can not improve the naturalness of synthetic speech in noisy environments.

Additionally, the descriptive statistics and t-test values for Sentences 1 and 2 show that the naturalness of Config 4 is slightly better than the baseline, but it is slightly worse than the baseline in Sentence 3. This instability, combined with the t-test results, further confirms that solely flattening the spectral tilt does not have a significant impact on improving the naturalness of synthetic speech in noisy environments, thereby refuting Hypothesis 2.

5.3 Validation of the Third Hypothesis

Hypothesis 3 posits that independently increasing F0 will not effectively improve intelligibility and naturalness of synthetic speech in noisy environment.

In terms of subjective intelligibility, both the individual sentence and overall descriptive statistics show that the mean values of increased F0 (Config 3) are superior to the baseline. The t-tests and Wilcoxon signed-rank tests also demonstrate significant correlations (p-values far less than 0.05), effectively confirming that solely increasing F0 can significantly enhance the intelligibility of synthetic speech in noisy environments. In terms of objective intelligibility, Config 3 shows significantly lower error rates compared to the baseline in Sentences 2 and 3, indicating a clear improvement in objective intelligibility. However, in Sentence 1, the objective intelligibility of Config 3 is slightly lower than the baseline. Combining the subjective results and most of the objective intelligibility results, it can be concluded that solely increasing F0 has the strong potential to significantly improve speech intelligibility in noisy environments.

Regarding naturalness, although Config 3's naturalness scores are slightly higher than the baseline in Sentences 1 and 3 and slightly lower in Sentence 2, the t-tests and Wilcoxon signed-rank tests do not reach significance (p-values greater than 0.05). This indicates that solely increasing F0 does not have a significant effect on the naturalness of speech in noisy environments.

In conclusion, independently increasing F0 can effectively improve speech intelligibility in noisy environments but does not have a significant effect on naturalness. Therefore, Hypothesis 3 is partially validated.

5.4 Validation of the Forth Hypothesis

Hypothesis 4 posits that independently increasing duration will effectively improve intelligibility and naturalness of synthetic speech in noisy environment.

In terms of subjective intelligibility, the descriptive statistics indicate that the mean values of increased duration (Config 2) are superior to the baseline across both individual sentences and overall analysis. Significant correlations were observed in Sentences 2 and 3 as well as in the overall data, suggesting that solely increasing duration can significantly enhance the subjective intelligibility of speech in noisy environments. For objective intelligibility, Config 2 outperformed the baseline in Sentences 2 and 3, though it was slightly lower than the baseline in Sentence 1. Combining both subjective and objective intelligibility results, it can be concluded that solely increasing duration has the potential to significantly improve intelligibility.

Regarding naturalness, the descriptive statistics for individual sentences and overall analysis show that the mean and median values of Config 2 are lower than those of the baseline. Significant negative correlations were found in Sentence 1 and the overall data, indicating that increasing duration significantly reduces naturalness in these cases. While significant correlations were not observed in Sentences 2 and 3, the results nearly reached significance, suggesting a potential negative impact of increased duration on naturalness

In summary, solely increasing duration has the potential to significantly improve intelligibility but also negatively impacts naturalness. Therefore, Hypothesis 4 is partially validated: while it does not consistently improve intelligibility or reduce naturalness in all cases, it demonstrates significant effects on both aspects.

5.5 Summary of Research Questions and Hypotheses

The above discussion provides a detailed analysis of the experimental results in relation to the research questions and hypotheses. For clearer understanding, the summary is as follows:

The first research question addresses whether the combined adjustment of acoustic features, specifically typical Lombard speech, can enhance the naturalness of speech in noisy environments. The hypothesis posited that it would. However, experimental results disproved this hypothesis, indicating that in practical applications, typical Lombard speech is insufficient to significantly enhance the naturalness of speech in noisy environments, and other methods need to be explored. Notably, although not directly related to the research question, the experimental results showed that Lombard speech significantly improved speech intelligibility in noisy environments, corroborating previous studies.

The second research question examines whether merely flattening the spectral tilt can improve the naturalness of speech in noisy environments. The hypothesis posited that it would. However, experimental results invalidated this hypothesis, suggesting that flattening the spectral slope independently cannot achieve improvements in naturalness in practical applications, and further methods need to be investigated. Additionally, the results support the notion that flattening the spectral tilt can effectively enhance speech intelligibility.

The third research question concerns F0 and whether solely increasing F0 can improve speech intelligibility and naturalness in noisy environments. The hypothesis posited that it would not. The experimental results partially validated this hypothesis, showing that while increasing F0 does not significantly affect naturalness, it can enhance speech intelligibility in noisy environments. This finding suggests that F0 is a crucial speech feature for improving intelligibility in practical noisy environments.

The fourth research question explores whether independently increasing duration can improve speech intelligibility and naturalness in noisy environments. The hypothesis posited that it would. The experimental results partially validated this hypothesis, indicating that although increasing duration can enhance speech intelligibility in noisy environments, it also reduces naturalness. This finding suggests that in practical applications, balancing intelligibility and naturalness is essential when adjusting the duration of speech to achieve optimal results.

In summary, whether it is independently increasing F0, increasing duration, flattening spectral tilt, or the combined adjustment of typical Lombard speech, all these methods can enhance speech intelligibility in noisy environments. However, for naturalness, the negative impact of solely increasing duration should be noted, while other acoustic features do not have a significant effect. Further exploration is needed to find effective ways to improve speech naturalness in noisy settings.

This summary has comprehensively enhanced the understanding of the contributions of different acoustic features to speech intelligibility and naturalness in noisy environments. The next section will summarize and highlight the key findings, while further discussing future research directions.

6 Conclusion

In the conclusion section, a summary of the main contributions of the research will be presented first, along with an examination of the limitations and future research directions. Through this approach, the thesis aims to provide valuable insights into the field and lay the groundwork for future investigations.

6.1 Summary of the Main Contributions

In this study, the effectiveness of various acoustic feature modifications on the intelligibility and naturalness of TTS voice in noisy environments was rigorously evaluated. This evaluation was conducted through a combination of subjective listening tests and objective metrics, providing a comprehensive analysis of how these modifications impact user perception and recognition accuracy.

Through detailed data analysis and comprehensive discussion, this study first partially validates previous research on Lombard speech, confirming that both typical Lombard Speech and spectral tilt flattening can effectively improve the intelligibility of synthetic speech in noisy environments. This is clearly demonstrated in the descriptive statistics, the correlation analyses of subjective intelligibility, as well as the objective intelligibility data.

Secondly, this study addresses and validates the research questions and hypotheses based on identified gaps, providing novel findings. These findings include: enhancing F0 independently can significantly improve the intelligibility of TTS voice in noisy environments; increasing duration alone can enhance intelligibility but also decreases naturalness, thus requiring a balance in practical applications; and adjustments to other acoustic features did not significantly impact naturalness, indicating that future research on naturalness should consider more dimensions.

Overall, this study contributes to understanding how modifications of specific acoustic features affect the intelligibility and naturalness of synthetic speech in noisy environments, offering valuable insights into the interaction between acoustic features and user perception. This, in turn, facilitates the provision of more intelligible and natural synthetic voices in real-world environments.

6.2 Limitations and Future Research

While the findings enhance understanding of how different acoustic characteristics affect synthetic speech intelligibility and naturalness in noisy environments, it is crucial to discuss the limitations that could provide valuable insights for future research.

Firstly, this study selected a train station as the representative noisy environment, using 80dB of ambient noise as the test condition, to validate the intelligibility and naturalness of synthetic voice in such an environment. However, this choice of noise environment has certain limitations. Future studies should explore a wider range of noise levels and more diverse noise environments, such as babble noise, to comprehensively verify the impact of different acoustic parameters on the intelligibility and naturalness of TTS voice across various noise conditions.

Secondly, to control variables, this study generated five different acoustic configurations for each sentence, all with the same content. During the listening tests, despite efforts to mitigate the priming effect by randomizing the stimuli, inserting non-scoring distractors, and instructing listeners not to repeat the audio content, it was still impossible to completely avoid this effect, which can enhance intelligibility due to repeated exposure to the same sentence content. Therefore, future research and

experimental design should consider using stimuli with different content while controlling variables. For example, narrowing the scope of variables to generate only two different parameter configurations per sentence and placing them farther apart in the questionnaire, while increasing the diversity of the overall sentence content. By designing experiments more meticulously, the priming effect on the results can be minimized to the greatest extent possible.

Lastly, in evaluating naturalness, this study primarily focused on the dimension of human-likeness. The findings indicated that merely increasing the duration significantly diminished human-likeness, whereas other adjustments had minimal impact. Future research on naturalness assessment should consider a more comprehensive evaluation framework, encompassing multiple dimensions: the naturalness of intonation to assess the alignment of synthetic speech intonation patterns with those of natural speech; the fluency of prosody to examine the rhythm, stress, and overall fluency of the speech; and the naturalness of emotional expression to evaluate how effectively synthetic speech conveys emotions. By incorporating these dimensions, future studies can more thoroughly investigate the impact of various acoustic features on speech naturalness, thereby providing stronger guidance for enhancing the naturalness of TTS voice in noisy environments.

By summarizing the contributions of this research and looking towards future research directions, this thesis aims to contribute to and lay a foundation for TTS technology research in noisy environments. This study has identified key acoustic features that enhance the intelligibility and naturalness of synthetic speech in real-world noisy settings, providing a clear direction for further optimization of TTS systems. It is hoped that these efforts will promote the development of more user-centered synthetic speech, especially in addressing the challenges of real-world noise environments. Future research will explore a broader range of acoustic conditions and a wider spectrum of evaluation dimensions to develop more adaptive, user-centered TTS systems, continuously enhancing the interactive experience of synthetic speech in real-world noise environments.

References

Bollepalli, B., Juvela, L., & Alku, P. (2019). Lombard speech synthesis using transfer learning in a tacotron text-to-speech system. In *Interspeech 2019* (pp. 2833–2837). doi: 10.21437/Interspeech.2019-1333

Boril, H., & Pollak, P. (2005). Design and collection of czech lombard speech database. In *Interspeech 2005* (pp. 1577–1580). doi: 10.21437/Interspeech.2005-461

Castellanos, A., Benedí, J.-M., & Casacuberta, F. (1996). An analysis of general acoustic-phonetic features for spanish speech produced with the lombard effect. *Speech Communication*, *20*(1-2), 23–35. doi: 10.1016/S0167-6393(96)00042-8

Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., & Tang, Y. (2013). Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, *55*(4), 572–585. doi: 10.1016/j.specom.2013.01.001

Cooke, M., Mayo, C., & Villegas, J. (2014). The contribution of durational and spectral changes to the lombard speech intelligibility benefit. *The Journal of the Acoustical Society of America*, *135*(2), 874–883. doi: 10.1121/1.4861342

Davis, C., Kim, J., Grauwinkel, K., & Mixdorff, H. (2006). Lombard speech: Auditory (a), visual (v) and av effects. In *Speech prosody 2006* (p. paper 252-0). doi: 10.21437/SpeechProsody.2006-88

Dreher, J., & O'Neill, J. (1957). Effects of ambient noise on speaker intelligibility for words and phrases. *Journal of the Acoustical Society of America*, *29*(12), 1320–1323.

Garnier, M., Bailly, L., Dohen, M., Welby, P., & Loevenbruck, H. (2006). An acoustic and articulatory study of lombard speech: Global effects on the utterance. In *Interspeech 2006* (p. paper 1862-Thu1A3O.6-0). doi: 10.21437/Interspeech.2006-323

Garnier, M., & Henrich, N. (2014). Speaking in noise: How does the lombard effect improve acoustic contrasts between speech and ambient noise? *Computer Speech & Language*, *28*(2), 580–597. doi: 10.1016/j.csl.2013.07.005

Garnier, M., Henrich, N., & Dubois, D. (2010). Influence of sound immersion and communicative interaction on the lombard effect. *Journal of Speech, Language, and Hearing Research*, *53*(3), 588–608. doi: 10.1044/1092-4388(2009/08-0138)

Godoy, E., & Stylianou, Y. (2012). Unsupervised acoustic analyses of normal and lombard speech, with spectral envelope transformation to improve intelligibility. In *Interspeech 2012* (pp. 1472–1475). doi: 10.21437/Interspeech.2012-417

Hu, Q., Bleisch, T., Petkov, P., Raitio, T., Marchi, E., & Lakshminarasimhan, V. (2021). Whispered and lombard neural speech synthesis. *arXiv*. Retrieved from `http://arxiv.org/abs/2101.05313` (arXiv:2101.05313)

Huang, D.-Y., & Ong, E. P. (2010). Lombard speech model for automatic enhancement of speech intelligibility over telephone channel. In *2010 international conference on audio, language and image processing* (pp. 429–434). doi: 10.1109/ICALIP.2010.5684545

Junqua, J.-C. (1993). The lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, *93*(1), 510–524. doi: 10.1121/1.405631

Junqua, J.-C. (1996). The influence of acoustics on speech production: A noise-induced stress phenomenon known as the lombard reflex. *Speech Communication*, *20*(1-2), 13–22. doi: 10.1016/S0167-6393(96)00041-6

Lane, H., & Tranel, B. (1971). The lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, *14*(4), 677–709. doi: 10.1044/jshr.1404.677

Lau, P. (2008). The lombard effect as a communicative phenomenon. *UC Berkeley PhonLab Annual Report*, *4*. Retrieved from `https://escholarship.org/uc/item/19j8j0b6` doi: 10.5070/P719j8j0b6

Lombard, (1911). Le signe de l'élévation de la voix [the sign of voice raising]. *Annales des Maladies de l'Oreille et du Larynx*, *37*, 101–119.

Lu, Y., & Cooke, M. (2009). The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, *51*(12), 1253–1262. doi: 10.1016/j.specom.2009.07.002

López, A. R., Seshadri, S., Juvela, L., Räsänen, O., & Alku, P. (2017). Speaking style conversion from normal to lombard speech using a glottal vocoder and bayesian gmms. In *Interspeech 2017* (pp. 1363–1367). doi: 10.21437/Interspeech.2017-400

Novitasari, S., Sakti, S., & Nakamura, S. (2021). Dynamically adaptive machine speech chain inference for tts in noisy environment: Listen and speak louder. In *Interspeech 2021* (pp. 4124–4128). doi: 10.21437/Interspeech.2021-946

Novitasari, S., Sakti, S., & Nakamura, S. (2022). A machine speech chain approach for dynamically adaptive lombard tts in static and dynamic noise environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2673–2688. doi: 10.1109/TASLP.2022.3196879

Patel, R., Everett, M., & Sadikov, E. (2006). Loudmouth: Modifying text-to-speech synthesis in noise. In *Proceedings of the 8th international acm sigaccess conference on computers and accessibility* (pp. 227–228). doi: 10.1145/1168987.1169028

Paul, D., Shifas, M. P. V., Pantazis, Y., & Stylianou, Y. (2020). Enhancing speech intelligibility in text-to-speech synthesis using speaking style conversion. In *Interspeech 2020* (pp. 1361–1365). doi: 10.21437/Interspeech.2020-2793

Pittman, A. L., & Wiley, T. L. (2001). Recognition of speech produced in noise. *Journal of Speech, Language, and Hearing Research*, *44*(3), 487–496. doi: 10.1044/1092-4388(2001/038)

Raitio, T., Suni, A., Vainio, M., & Alku, P. (2011). Analysis of hmm-based lombard speech synthesis. In *Interspeech 2011* (pp. 2781–2784). doi: 10.21437/Interspeech.2011-696

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, *84*(3), 917–928. doi: 10.1121/1.396660

Suni, A., Karhila, R., Raitio, T., Kurimo, M., Vainio, M., & Alku, P. (2013). Lombard modified text-to-speech synthesis for improved intelligibility: Submission for the hurricane challenge 2013. In *Interspeech 2013* (pp. 3562–3566). doi: 10.21437/Interspeech.2013-766

Valentini-Botinhao, C., Yamagishi, J., King, S., & Stylianou, Y. (2013). Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of hmm-based synthetic speech in noise. In *Interspeech 2013* (pp. 3567–3571). doi: 10.21437/Interspeech.2013-767

Van Ngo, T., Kubo, R., & Akagi, M. (2019). Evaluation of the lombard effect model on synthesizing lombard speech in varying noise level environments with limited data. In *2019 asia-pacific signal and information processing association annual summit and conference (apsipa asc)* (pp. 133–137). doi: 10.1109/APSIPAASC47483.2019.9023227

Zorila, T.-C., Kandia, V., & Stylianou, Y. (2012). Speech-in-noise intelligibility improvement based

on spectral shaping and dynamic range compression. In *Interspeech 2012* (pp. 635–638). doi: 10.21437/Interspeech.2012-197