ANKARA
HACI BAYRAM VELİ ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

A Comparative Analysis of Auditory Recognition of Emotional Prosody Generated by Text-To-Speech AI Programs

Master's Thesis

Eda Şahin

Department of English Language and Literature

Ankara, 2025

A Comparative Analysis of Auditory Recognition of Emotional Prosody Generated by Text-To-Speech AI Programs

Eda ŞAHİN

Prof. Dr. Güven MENGÜ

Thesis Commitee

Prof. Dr. Güven MENGÜ

Assoc. Prof. Dr. Nazlı GÜNDÜZ

Assoc. Prof. Dr. İsmail Fırat ALTAY

MASTER'S THESIS

DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE

ENGLISH LINGUISTICS

Ankara - 2025

ETİK BEYAN

Bu tezi/projeyi, Ankara Hacı Bayram Veli Üniversitesi Tez ve Proje Yazım Kılavuzuna uygun olarak hazırladığımı; tezin/projenin tamamında akademik kurallara ve etik ilkelere uyduğumu ifade ederim. Yararlandığım eserlerin tamamını metin içinde referanslandırdığımı ve kaynakçada kılavuzda tanımlanan şekilde yer verdiğimi, haricindeki ifadelerin bana ait olduğunu, herhangi bir kaynaktan kopyalama yapmadığımı ya da yapay zeka aracılığı ile üretilmiş ifadelere metinde yer vermediğimi beyan ederim. Herhangi bir zamanda bu beyanıma uygun olmayan bir durumun tespit edilmesi halinde, aleyhime doğacak bütün hak kayıpları dahil tüm hukuki sonuçları kabul ettiğimi bildiririm.

Eda Şahin

19/06/2025

ONAY

Ankara Hacı Bayram Veli Üniversitesi Lisansüstü Eğitim Enstitüsü İngiliz Dili ve Edebiyatı Anabilim Dalı İngiliz Dilbilimi Programı Eda Şahin öğrencisi tarafından hazırlanan A Comparative Analysis of Auditory Recognition of Emotional Prosody Generated by Text-To-Speech AI Programs Başlıklı tez çalışması 19/06/2025 tarih ve saatinde yapılan tez savunma sınavında aşağıdaki jüri tarafından OY BİRLİĞİ ile YÜKSEK LİSANS TEZİ olarak KABUL edilmiştir.

	Kabul	Ret
Başkan:	☒	☐
Doç. Dr. İsmail Fırat ALTAY / Hacettepe Üniversitesi		
Üye:	☒	☐
Prof. Dr. Güven MENGÜ / Ankara Hacı Bayram Veli Üniversitesi		
Üye:	☒	☐
Doç. Dr. Nazlı GÜNDÜZ / Ankara Hacı Bayram Veli Üniversitesi		

Yapay Zeka Tabanlı Metin-Konuşma Sistemleriyle Üretilen Duygusal Bürünün İşitsel Algılanmasına Yönelik Karşılaştırmalı Bir İnceleme

Eda ŞAHİN

Yüksek Lisans Tezi

Danışman: Prof. Dr. Güven MENGÜ
T.C. Ankara Hacı Bayram Veli Üniversitesi, Lisansüstü Eğitim Enstitüsü
İngiliz Dili ve Edebiyatı Anabilim Dalı
2025, Ankara

ÖZET

Bu çalışma, yapay zeka tarafından üretilen konuşmalardaki duygusal bürünün algısal olarak açık ve insan konuşmasından ayırt edilebilir olup olmadığını incelemektedir. Katılımcılar hem yapay zeka hem de insan üretimi ses kayıtlarını dinleyerek iletilen duyguyu belirlemişlerdir. Pearson korelasyon analizi kullanılarak yapılan değerlendirme sonucunda, özellikle erkek ses çiftlerinde olmak üzere mutluluk ve üzüntü duyguları için algısal uyumun orta düzeyde olduğu görülmüştür. Ancak korku ve şaşkınlık gibi duyguların yapay zeka konuşmalarında tanınabilirliği düşük bulunmuştur. Bu bulgular, duygusal konuşma sentezi alanındaki ilerlemeleri ve devam eden zorlukları ortaya koymakta olup, duygusal bilişim (affective computing) bağlamında TTS sistemlerinin geliştirilmesine yönelik önemli çıkarımlar sunmaktadır.

Anahtar Kelimeler: Duygusal bürün, konuşma sentezi, yapay zeka, insan-bilgisayar etkileşimi, duygusal bilişim

A Comparative Analysis of Auditory Recognition of Emotional Prosody Generated by Text-To-Speech AI Programs

Eda ŞAHİN

Master's Thesis

Supervisor: Prof. Dr. Güven MENGÜ
Ankara Haci Bayram Veli University, Institute of Graduate Programs
Department of English Language and Literature
2025, Ankara

ABSTRACT

This study investigates the perceptual clarity and distinguishability of emotional prosody in AI-generated speech compared to human speech. After listening to recordings made by AI and humans, participants were asked to identify the feeling they were receiving. Results from Pearson correlation analysis revealed that some emotions, primarily happiness and sadness, had a moderate level of perceptual alignment, especially when male voice pairs were used. However, AI speech showed poor recognition of emotions like surprise and fear. These results provide insights for enhancing TTS systems in affective computing, highlighting both the advancements and enduring difficulties in emotional speech synthesis.

Keywords: Emotional prosody, AI-generated speech, speech synthesis, perceptual evaluation, affective computing

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my esteemed advisor, Prof. Dr. Güven MENGÜ, for his knowledge, expertise, and guidance throughout the preparation of this thesis.

I would also like to thank Prof. Dr. Selma ELYILDIRIM, Assoc. Prof. Dr. Nazlı GÜNDÜZ, and Assoc. Prof. Dr. Aydan IRGATOĞLU, who guided me with their insights and suggestions throughout my graduate studies.

I would also like to thank all the volunteers who participated in the data collection process, and Kendal Kutlu, who assisted me in conducting the research and provided me with support for the data obtained.

I am also grateful to my family, who have always been there for me throughout this process, supporting me with their patience and love, my friends who provided moral motivation, and my beloved cat, Gofret.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Sayfa

LIST OF ABBREVIATIONS

The following table presents the abbreviations used throughout this thesis, along with their corresponding explanations.

Abbreviation	Explanation
AI	Artificial Intelligence
VAS	Visual Analog Scale
TTS	Text-to-Speech

1 INTRODUCTION

Beckman (1996) refers to prosody as "the organizational structure of speech". The term "prosody" is used to describe rhythm, intonation, and stress in spoken sentences. Since they establish patterns that are primarily independent of a word or phrase's segmental composition (i.e., its vowel and consonant phones), intonation and rhythm are characterized as suprasegmental aspects of speech. Pitch, loudness, duration, and relative timing of phones, syllables, and other speech units are all considered suprasegmental features. Auditory qualities are influenced by the acoustic signal's time-varying properties, such as fundamental frequency (F0) and amplitude, as well as the duration of acoustic intervals for phonemes and syllables (Cole, 2015: 1-2). The study of prosody in processing must rely on ideas and methods created to explore spoken language in various fields, including phonetics, phonology, speech production and perception, and psycholinguistics.

Prosodic structure is formally defined in linguistic theories of autosegmental phonology and intonation, and it has measurable acoustic-phonetic correlates such as variation in fundamental frequency, amplitude, spectral information and the relative duration of sound and silence across and between utterances. (Speer & Blodgett, 2006) Cutler and Isard (1980) define prosody as “the sauce of a sentence” (pp. 245). It complements, enriches, or subtly alters the natural flavour. And, like a fine sauce, an understanding of a sentence's prosodic structure is a combination of various elements, none of which can be detected separately in the finished result. Prosody, or the sound characteristics of vocal expressions, transmits both language and paralinguistic information, including the speaker's emotional state (Banse & Scherer, 1996: 614) and intention. Thus, prosody plays an essential instrument for human communication by encoding feelings and intentions through acoustic features.

2 LITERATURE REVIEW

2.1 Background of the Study

Larrouy-Maestri et al. (2024: 6) present a comprehensive examination of almost three decades of study on emotional prosody. Emotional prosody is the adjustment of acoustic parameters such as pitch (f0), loudness, speech rate, and timbre to convey emotions in discourse (Banse & Scherer, 1996: 614; Larrouy-Maestri et al., 2024: 6). It is a fundamental component of spoken language, allowing speakers to express emotions that go beyond lexical content (Scherer, 2013: 41). Bänziger et al. (2015: 196) describe the role of emotional prosody through encoding-decoding processes. In the encoding process, the speaker expresses emotions by modifying the features of their voice. The decoding process, on the other hand, refers to the interpretation of listeners and their recognition of these emotional cues. Key auditory characteristics linked to various emotions were discovered in early research. Higher pitch is associated with fear, joy, and enthusiasm whereas lower pitch is found to be linked to sadness and disdain (Banse & Scherer, 1996; Juslin & Laukka, 2003: 770). In terms of loudness, joy and anger display greater amplitude and energy, sadness and shame have softer speech intensity (Eyben et al., 2016). Speech rate and duration is faster in anger, happiness and pride, and slower in sadness and boredom. Finally, speech that is joyful is usually resonant and clear. On the other hand, fearful speech is frequently harsh and breathy (Banse & Scherer, 1996; Goudbeek & Scherer, 2010: 1322). The precise relationship between auditory characteristics and emotional expression is still unclear though, after decades of research (Larrouy-Maestri et al., 2024: 6).

2.2 AI And Emotional Speech Synthesis

The parametric synthesis used by early TTS systems resulted in monotonous and robotic speech. By enabling more dynamic and human-like speech creation, deep learning—specifically, convolutional neural networks (CNNs), deep neural networks (DNNs), and recurrent neural networks (RNNs)—revolutionized speech synthesis (Barakat et al., 2024: 5). Although text-to-speech (TTS) technology has made major advances in producing speech that sounds natural, it still struggles to accurately replicate prosody, which are fluctuations in pitch, rhythm, and intensity that communicate meaning and emotion. By applying prosody embeddings and systematic modeling, contemporary TTS systems—such as those that employ deep learning and neural networks—have increased prosodic accuracy (Mondal et al., 2024: 12). According to research, synthesized speech becomes much more expressive and natural when explicit prosody control is used, such as feature-based modulation and emphasis tags (Wada et al., 2024: 8). Even said, a lot of systems still have trouble accurately representing the subtleties of human speech, especially when it comes to expressing emotional prosody and discourse-related stress (Korotkova et al., 2024: 15). Additionally, research shows that although end-to-end TTS systems perform better in prosody control than conventional concatenative and parametric methods, they still have trouble adjusting to various speakers and linguistic settings (Tsiamas et al., 2024: 22). Furthermore, one-to-many mapping issues in TTS have been addressed by diffusion-based hierarchical prosody modeling, which enables more adaptable and lifelike speech synthesis (Liu et al., 2024: 30). Even though existing models are promising, research is still being done to improve prosodic qualities even more so that TTS may be used in real-world scenarios like language acquisition, assistive technology, and AI-driven communication.

The creation of highly expressive and realistic-sounding synthetic speech is hampered by a number of significant issues with emotional prosody synthesis in text-to-speech (TTS) systems. The absence of high-quality, emotionally identified speech datasets is a significant obstacle because the labels in the emotional speech corpora that are currently available are frequently too basic to adequately represent the subtleties of human emotion (Bian et al., 2024: 3). Due to this restriction, TTS models are not as able to provide real emotional variances. The difficulties of modeling prosody at multiple levels are another important problem. Existing models frequently use utterance-level embeddings without taking into account the complex interactions between frame-level and word-level emotional cues, which results in a lack of organic variation in artificially produced speech (Tang et al., 2024: 7). According to Ngo et al. (2024), TTS models trained on particular speakers have trouble generalizing emotional expressions across many voices and languages, which further complicates cross-speaker and cross-linguistic emotion transfer. Comparative learning techniques are crucial for enhancing emotion distinction, according to recent research; yet they necessitate huge amounts of data and computational assets (Ngo et al., 2024: 12).

Future advances in AI-generated text-to-speech (TTS) and emotional prosody will be dependent on better algorithms for deep learning, multimodal integration of data, and real-time adaptive speech synthesis. One of the most promising directions is the integration of multimodal emotion synthesis, which enhances AI-generated speech by incorporating visual and contextual information, such as facial expressions and body language, to produce a more authentic and emotionally compatible speech output (James et al., 2023: 45). Also, reinforcement instruction and self-supervised training techniques are predicted to improve prosodic flexibility, allowing AI to modify intonation, speech tempo, and pitch modulation during real-time conversations (Kolekar et al., 2024: 78). Handling cross-speaker variability remains a significant problem, as existing models fail to generalize prosodic patterns across individuals and languages (Liu et al., 2023: 102). Transfer learning and multilingual emotional prosody mapping are being investigated by researchers to improve emotional expressiveness in a variety of linguistic settings (Nithin and Prakash, 2022: 56). Another notable innovation is prosody-aware text preparation, in which AI systems predict and integrate emotional indicators in synthetic speech before vocalization, resulting in a more seamless and expressive output (Ao & Yuan, 2024: 33).

The ability of artificial intelligence (AI) to mimic emotional speech has advanced significantly, yet it is still difficult to completely replicate human emotional expressiveness. Deep learning models like Tacotron and FastSpeech2 are now incorporated into AI-driven text-to-speech (TTS) systems, allowing for more precise control over rhythm, intonation, and pitch to improve realistic emotion (Roshan et al., 2024: 1). According to studies, AI-generated speech can accurately portray common emotions such as happiness, sadness, rage, and neutrality, scoring highly on subjective assessments of naturalness (Ao & Yuan, 2024: 2). However, current models have trouble capturing complex affective expressions, especially in cross-linguistic and multi-speaker scenarios (Kolekar et al., 2024: 3). This makes it difficult to capture delicate and mixed feelings. Furthermore, even while zero-shot and transfer learning techniques enable AI models to generalize across various voices and emotional tones, they frequently fall short in practical applications that call for precise emotional variation (Chaudhury et al., 2024: 1).

2.3 Human Perception and Evaluation of Ai Speech

The use of artificial emotional datasets, which are useful but fall short of accurately simulating the richness and diversity of genuine human speech, is another significant drawback (Khalifah et al., 2024: 15). Hybrid models that combine rule-based emotion cues with algorithms for deep learning have been offered as a solution to these gaps; however, they need to be further refined for scalable and real-time deployment. In spite of these obstacles, artificial intelligence (AI)-generated emotional speech is becoming better and is being effectively used in virtual personal assistants, audiobooks, and mobility tools indicating its potential for more realistic and expressive synthetic voices. When implemented to new speakers or languages, deep learning models trained on a particular voice sometimes lose their emotional authenticity, making cross-speaker emotion transmission problematic (Nithin & Prakash, 2022: 3). Moreover, high-quality emotional TTS systems demand substantial processing resources, which makes large-scale implementation challenging. Real-time delay and computational expense are still issues (Wushouer & Tuerhong, 2023: 7). Emotional prosody, characterized by variations in pitch, intensity, tempo, and rhythm, enables listeners to discern emotional intent and speaker affect (James et al., 2023: 22). However, studies suggest that while current TTS models can convincingly replicate certain emotional tones, such as happiness or sadness, they often fall short in expressing nuanced or mixed emotional states, which are more complex for listeners to decode (Ao & Yuan, 2024: 10). Moreover, research on perceptual evaluations reveals that listeners may struggle to distinguish between human and AI voices when the emotional cues are subtle, highlighting a perceptual gap in emotional authenticity (Kolekar et al., 2024: 5). This underscores the need for systematic assessment methods—such as listener-based evaluations and psychometric tools—to examine the intelligibility, naturalness, and emotional clarity of AI-generated speech. By understanding how humans interpret and respond to synthetic emotional prosody, researchers can develop more refined models that align with the subtleties of human communication.

A well-known model that organizes emotions in people into an organized structure and highlights their interrelationships and magnitudes is Plutchik's Wheel of Emotions (Plutchik, 1980: 7). It offers a thorough method for comprehending emotional expression and response, making it useful in studies on psychology, AI education, and human-machine interaction (Mondal & Gokhale, 2020: 2). Emotions, according to Robert Plutchik, are evolutionary adaptations that are essential for survival. The eight main emotions in his model—joy, trust, fear, surprise, sadness, anticipation, anger, and disgust—are organized in a wheel shape to show how they are related to one another (Plutchik, 1980: 7). A structural foundation for intricate emotional interactions is formed by the opposite pairs of each emotion, such as joy and sadness and anger and fear (Hans, 2021). In order to improve the model's capacity to explain complex emotional experiences, Plutchik also added emotional pairs, in which fundamental emotions combine to produce secondary emotions (Abukhodair et al., 2023: 5). In psychology, Plutchik's Wheel has been widely used for behavioral research, therapy, and emotion identification (Brkić, 2024: 52). Its systematic categorization aids in the diagnosis of emotional disorders and the development of methods for therapy. Furthermore, the model facilitates emotion identification in sentiment evaluation and social network mining, which helps to enhance AI's capacity for empathetic reaction in machine learning and artificial intelligence (Mondal & Gokhale, 2020:

2). In order to improve human-computer connection through subtle emotional expressiveness, recent developments integrate the Wheel into AI-powered emotional speech generation (Hans, 2021).

2.4 Rationale And Significance of the Study

The rationale for this study is grounded in the increasing presence of AI-generated speech in everyday technologies and the critical role of emotional prosody in ensuring effective and human-like communication. As artificial intelligence continues to evolve, especially in the realm of text-to-speech (TTS) synthesis, the ability of AI systems to convey authentic emotional cues through speech is essential for enhancing user engagement, trust, and comprehension in human-computer interactions (Liu, Liu, & Li, 2023: 5266). Despite advancements in deep learning-based TTS models, recent research suggests that synthetic speech often lacks the nuanced prosodic elements—such as pitch variability, speech rate, and intensity modulation—that characterize natural emotional expression (James et al., 2023). Furthermore, users may misinterpret or fail to recognize the intended emotion in AI speech, especially in subtle or secondary emotional states, limiting its applicability in emotionally sensitive contexts such as virtual therapy, education, or customer service (Kolekar et al., 2024). This study is significant because it addresses a key gap in current research: the perceptual evaluation of emotional prosody in AI-generated speech by human listeners. By investigating the recognizability and clarity of emotional cues in synthetic versus human speech, this research contributes to the refinement of TTS systems, ultimately supporting the development of more empathetic and context-aware AI voices (Ao & Yuan, 2024: 58). Moreover, these findings have practical implications for designing inclusive and emotionally intelligent AI applications in both commercial and assistive domains.

The primary objective of this study is to evaluate how well human listeners can recognize emotional prosody in text-to-speech (TTS) recordings produced by artificial intelligence. The study specifically intends to investigate how effectively people comprehend emotional speech created by artificial voices and whether they are able to consistently distinguish between speech produced by AI and speech produced by humans when emotions are incorporated into the prosodic elements of the utterance. One of the main goals is to assess how well AI systems can represent emotional prosody and how closely these synthetic representations resemble the emotional richness and genuineness of normal human speech. The study also aims to determine which emotions—such as happiness, sadness, anger, fear, and surprise—are more likely to be viewed as artificial intelligence (AI)-generated and which are more likely to be viewed as less expressive or unnatural. The study looks into the precise prosodic components—such as pitch, speech pace, rhythm, and intensity—that affect perception in addition to the accuracy of emotion recognition. Knowing these factors can help determine which audio characteristics stand out the most in listener assessments and aid in either successful identification or misunderstanding. Examining misclassification trends is a crucial part of the analysis, especially when AI-generated speech gets mixed up with human speech and vice versa. Finding the prosodic clues that could lead to perceptual ambiguity or confusion is part of this. Lastly, by providing suggestions for enhancing prosody modeling in synthetic voices, the study hopes to make a significant contribution to the continued advancement of AI speech synthesis. The results can direct future developments aiming at producing AI speech that is not only understandable but also believable human-like in its emotional depth and diversity by pointing out areas where existing TTS models lack emotional expressiveness.

2.5 Research Questions and Objectives

The research questions and objectives of this study are designed to explore the extent to which emotional prosody in AI-generated speech is perceptible and distinguishable by human listeners. Central to this investigation is the question of whether participants can accurately identify AI-generated emotional speech and differentiate it from human-produced recordings, particularly when the emotional content is embedded in the prosodic features of speech. A secondary aim is to assess how clearly these emotional cues are perceived, using a Visual Analogue Scale (VAS) to gauge the clarity of emotional expression. The study also seeks to determine which specific emotions are more convincingly synthesized by AI, and whether certain emotional tones (e.g., happiness, sadness, anger) are more prone to misclassification. These objectives are informed by current literature, which highlights both the advancements in deep learning-based speech synthesis and the ongoing challenges in replicating nuanced human emotions (Liu, Liu, & Li, 2023; James et al., 2023). Furthermore, the research aims to contribute to the growing field of affective computing by offering empirical insights into listener perception, which is critical for developing more expressive and socially intelligent AI voices (Kolekar et al., 2024). Ultimately, this study aims to bridge the gap between technical speech synthesis capabilities and the perceptual realities of human users, ensuring that future AI systems can communicate with greater emotional authenticity and effectiveness (Ao & Yuan, 2024).

2.6 Scope And Limitations

This study focuses on the perception of emotional prosody in AI-generated speech by human listeners. The research is conducted by using 100 recordings; 40 RAVDESS dataset's human production speech (20 male and 20 female) and 60 AI production recording (30 male and 30 female) which are obtained from Genny.ai and Murf.ai. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a popular multimodal dataset used to do research on emotional speech detection, affective computing, and human-computer interaction. The RAVDESS dataset, created by Livingstone and Russo (2018), consists of 7356 recordings made by 24 professional actors (12 males and 12 females), each performing a set of predetermined emotional expressions in two modalities: speech and song. The emotions are calm, happy, sad, angry, afraid, surprised, and disgusted, which are expressed at two intensity levels—except for neutral. The dataset was captured in a controlled environment using high-fidelity 48kHz WAV files with 16-bit resolution, which ensured constant sound quality while reducing environmental noise. Its balanced structure and consistent protocol make it ideal for training and testing machine learning algorithms that require precise emotional classification. RAVDESS has been extensively used as a benchmark dataset in speech emotion detection systems, AI-generated prosody performance testing, and emotional response modeling in emotional computing applications. Its adaptability and robustness have established its status as a key tool for computational emotion research (Livingstone & Russo, 2018)

Murf.ai and Genny.ai are popular platforms for AI-powered text-to-speech (TTS) and content production, each with unique capabilities targeted to different user demands. Murf.ai stands out as a versatile TTS generator, offering over 200 customizable voices in 20+ languages, allowing users to generate realistic and human-like voiceovers for e-learning, advertising, podcasts, and presentations. Its simple drag-and-drop interface enables users of different skill sets to utilize the platform without requiring technical knowledge, speeding the voiceover creation process (Murf.ai, 2024). Furthermore, Murf.ai's recent addition of the MultiNative function enables seamless transformation between languages inside a single voiceover, increasing its utility in multilingual scenarios (Murf.ai, 2024). Genny.ai, founded by LOVO, is a full video production platform that includes advanced TTS capabilities. Genny allows users to efficiently create and change films by including features like an AI writer for developing compelling narratives and an automatic subtitle generator for making dynamic subtitles with a click. Genny.ai's combination of TTS with video editing tools presents it as a reliable choice for creators looking at streamlining their material production workflows (Play.ht, 2024). Both platforms demonstrate improvements in AI technology, meeting the ever- increasing need for effective and high-quality material creation tools.

The study included 77 undergraduate students from the English Literature Department of Ankara Hacı Bayram Veli University. All participants were fluent in English, ensuring that they understood the emotional and linguistic implications of the speech stimuli. Participants were shown a series of WAV audio recordings with both AI-generated and human-produced speech samples expressing various emotions. Participants were asked to evaluate the authenticity of the speech, and the emotional clarity of each tape using a Visual Analogue Scale (VAS) without knowing where the recordings came from. They were also asked to define the emotional state that the recordings reflect among 5 emotions: anger, happiness, sadness, fear and surprise. The objective of this design was to investigate the subjective clarity of emotional expression in various production modes as well as the perceptual recognizability of emotional prosody in synthetic speech.

In psychological and behavioral research, the Visual Analogue Scale (VAS) is a psychometric assessment tool that is frequently used to sensitively record subjective experiences and internal states (Gift, 1989: 286; Wewers & Lowe, 1990: 227). The VAS allows respondents to mark a particular point along the continuum to indicate their perception. It is often displayed as a 100 mm horizontal or vertical line anchored by bipolar adjectives (e.g., “not at all clear” to “extremely clear”) (Huskisson, 1974: 1128). When evaluating subjective and non-observable dimensions, including emotional clarity, speech naturalness, or perceived authenticity, this approach works very well (Aitken, 1969: 990). VAS enables more accurate and customized responses than category rating scales, enabling in-depth statistical analysis using parametric techniques (Wewers & Lowe, 1990: 227). Additionally, its continuous style offers a more sophisticated perspective on participant judgments by mitigating response bias linked to preset categories. In order to compare the expressive capacities of AI-generated and human-produced recordings in-depth, the VAS was used in this psychoacoustic study to evaluate the clarity with which emotional prosody was expressed in speech samples.

3 MATERIALS AND METHODS

3.1 Participants

The participants in this study consisted of 77 undergraduate students enrolled in the English Literature Department at Ankara Hacı Bayram Veli University. These ranged in age from 18 to 24 years, with 24 identifying as male and 53 as female. With consideration for the language skills necessary to properly assess emotional prosody in English speech, this participant group was carefully chosen. All participants were native Turkish speakers who use English as a second language, and their English proficiency was verified through their successful completion of the university's entrance examination, which includes language assessment components. Additionally, students are often subjected to English in academic and communicative contexts since they are actively enrolled in an English-medium program.

Their participation in a perceptual investigation including subtle emotional cues in English speech was well-founded on this prolonged and thorough exposure to the language. For the purposes of this study, their comprehension and analysis of prosodic elements like intonation, rhythm, and emotional tone were deemed sufficient, especially since the task required them to distinguish minute variations between emotional speech recorded by humans and AI. Participants were also tested to make sure none had any cognitive disorders or hearing impairments that could impede speech perception or auditory processing. By ensuring a homogeneous group in terms of language background and cognitive-auditory health, this meticulous participant selection improved the validity of the perceptual data gathered for the study.

3.2 Materials

The study's audio stimuli were created to compare the emotional prosody of speech produced by AI (Genny.ai and Murf.ai) versus speech produced by humans in a methodical manner. Two linguistically neutral sentences, "Dogs are sitting by the door" and "Kids are talking by the door," were used in all recordings. These sentences were chosen for their simplicity, grammatical consistency, and lack of semantic bias, ensuring that participants' assessments were based more on prosodic cues than lexical content. Five different emotional states—happiness, sadness, wrath, fear, and surprise—were used to represent each sentence. The 40 WAV files of AI-generated recordings were split equally between Genny AI and Murf AI. Each of the two target statements was created by 5 male and 5 female voices from each platform, generating 20 recordings per platform. To guarantee a balanced gender representation and emotional coverage, these voices were pre-selected from the platforms' emotional speech libraries.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), a verified and popular database in emotional speech research, provided the 30 WAV files that made up the human-produced recordings. In particular, three different male and three different female actors provided five recordings for each of the two emotional categories in each line, resulting in 60 recordings per sentence and guaranteeing actor diversity and gender balance. A consistent and regulated comparison across stimuli was made possible by matching the AI-generated files' prosodic variation and tone to the emotional renditions from RAVDESS. To guarantee technical consistency, all recordings were converted to the same format (WAV, 48kHz, 16-bit) and normalized for loudness. A thorough examination of the rendering and perception of emotional prosody in synthetic versus actual human speech was made possible by this carefully selected stimulus set.

This study used Google Forms, a popular online survey tool known for its dependability, accessibility, and convenience of use in academic research settings, to facilitate data collection. Because of its ability to effectively gather quantitative data and display auditory stimuli in a structured fashion, Google Forms has become more and more popular in experimental designs including perceptual and linguistic investigations (Moser & Korstjens, 2018). After each audio stimulus was placed in a standardized digital form, participants were given questions that evaluated the recording's emotional clarity as well as the perceived authenticity of the voice. All participants were instructed to wear headphones during the survey in order to guarantee consistent and ideal listening conditions. This measure improved prosodic perception accuracy and reduced ambient noise interference, which is particularly important for tasks involving small emotional differences in speech (Woods et al., 2017). Google Forms' ease of dissemination has led to its growing use in postgraduate and doctoral studies, especially in situations involving distant or asynchronous data collecting where in-person interactions are scarce (Suryani, Sari, & Fitriani, 2023). Apart from its practical benefits, Google Forms complies with conventional ethical requirements for managing participant information by ensuring data security through encrypted transmission and storage (Google Support, n.d.). Together, these characteristics support the platform's validity and dependability as a survey tool for scholarly research, including its successful use in dissertation projects that call for scalable and effective data collection techniques.

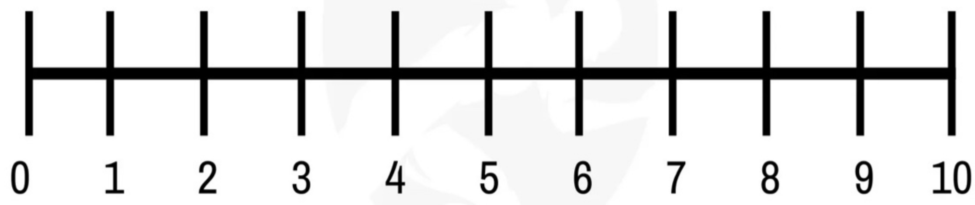


Image 3.1 VAS Scale (2025)

Source: Bookbush Institue (2025)

The Visual Analogue Scale (VAS), a psychometric instrument renowned for its sensitivity and capacity to capture subtle perceptual judgments, was used in the study to assess emotional clarity subjectively. Participants indicated their answers on a horizontal continuum that was anchored by opposing descriptions, such as "not at all clear" and "extremely clear," that made up the VAS. When compared to categorical rating systems, VAS is preferred due to its high sensitivity and resilience to ceiling and floor effects. It has been verified in several fields for measuring internal and subjective states, including perceptual clarity (Gift, 1989; Wewers & Lowe, 1990). It is especially appropriate for studies involving the perception of emotions because of its continuous character, which also enables more accurate parametric statistical analysis. The VAS and Google Forms together offered a methodologically sound and repeatable framework for assessing participant answers in a flexible yet controlled online setting.

3.3 Procedure

This study's methodology was carefully designed to guarantee impartiality, reduce prejudice, and enable a trustworthy evaluation of emotional prosody in both human- and AI-generated speech. Google Forms, a popular platform for academic data collecting because of its accessibility, response variety, and secure data management, was used to direct participants to listen to a series of randomized audio recordings (Raju & Harinarayana, 2016; Google Support, n.d.). A series of three structured questions followed each audio clip that was included in the form. Initially, participants were asked to score the recording's authenticity. Second, they were asked to use a Visual Analogue Scale (VAS) to score the speech's emotional clarity. This gave them the opportunity to express how much the intended emotion was understood. The VAS, which is renowned for its sensitivity and accuracy in capturing subjective perceptual evaluations, was shown as a horizontal line with ends marked "not at all clear" and "extremely clear" (Gift, 1989; Wewers & Lowe, 1990).

Third, participants were asked to choose from five pre-established emotional categories—happiness, anger, fear, sadness, and surprise—in order to determine the precise emotion that was expressed in the audio. All recordings were randomized using an online randomization tool prior to being uploaded to the form in order to maintain objectivity and avoid any ordering effects. Crucially, participants were not told whether the recordings were made by humans or by AI, maintaining the objectivity of their assessments and removing expectation bias. In order to facilitate effective data compilation and export for subsequent statistical analysis, every response was automatically captured and safely kept within the Google Forms interface.

3.4 Data Analysis

Participants' impressions of emotional prosody in both human and AI-generated speech were evaluated using both descriptive and inferential statistical approaches during the study's data analysis phase. The mean score on the Visual Analogue Scale (VAS) was the main indicator used to assess emotional clarity and authenticity. A continuous 0–100 scale was used to quantify each VAS score; higher scores indicated that the emotion in the provided audio recording was regarded to be more clear and closer to the human production of emotional prosody. To provide an overall measure of perceived emotional expressiveness, these ratings were averaged across recordings for each emotion and production type (human or AI). The central tendencies and dispersions of the VAS ratings for each category were compiled using descriptive statistics like means and standard deviations.

The statistical program IBM SPSS Statistics (Version 27.0), which is well-known in the academic and professional research world for its strong skills in data administration, statistical analysis, and output visualization, was utilized for inferential analysis (Field, 2018). Specifically, the association between participants' assessments of authenticity and their perceived emotional clarity across several recordings was investigated using the Pearson product-moment correlation coefficient. A parametric statistical technique called the Pearson test is used to assess the direction and magnitude of a linear relationship between two continuous variables (Field, 2018). This test highlighted the influence of prosody on perceived naturalness by revealing whether recordings that were thought to be more emotionally clear were also more likely to be regarded as real (i.e., human-produced). Participants' associations between emotional expressiveness and speaker authenticity were examined for trends and possible perceptual biases using the Pearson correlation analysis results. The study's findings about the detectability of emotional prosody in synthetic speech were validated by the use of SPSS (27.0.1), which guaranteed a trustworthy, standardized, and reproducible analysis framework.

4 RESULTS

This study examined how emotional prosody in AI-generated and human-produced speech recordings related to one another using Pearson correlation analyses. Participants listened to recordings created by Genny AI and Murf AI for artificial voices and the RAVDESS dataset for human vocal emotions as part of the data collection process, which was conducted through Google Forms. On a Visual Analogue Scale (VAS), which measures perceived emotional intensity, participants scored each tape after listening in terms of the recording's authenticity, clarity and they were asked to guess the emotion the recording reflects. The collected responses were exported to Excel and subsequently analyzed using SPSS. For every matched AI–human recording pair, correlations have been calculated to evaluate how similar the perceived emotional expressions were.

Table 4.1

Correlations

		Emotion3	Emotion87
Emotion3	Pearson Correlation	1	,076
	Sig. (2-tailed)		,510
	N	77	77
Emotion87	Pearson Correlation	,076	1
	Sig. (2-tailed)	,510	
	N	77	77

Table 3.1. shows that for the emotion "happy" and the sentence "Kids are talking by the door," Emotion3 (AI Female) and Emotion87 (Human Female) showed a weak positive correlation ($r = 0.076$, $p = 0.510$), which was not statistically significant.

Table 4.2

Correlations

		Emotion4	Emotion15
Emotion4	Pearson Correlation	1	-,025
	Sig. (2-tailed)		,832
	N	77	77
Emotion15	Pearson Correlation	-,025	1
	Sig. (2-tailed)	,832	
	N	77	77

Table 3.2. shows that for the emotion "sad" and the sentence "Dogs are sitting by the door," Emotion4 (AI Female) and Emotion15 (Human Female) showed a very weak negative correlation (r = -0.025, p = 0.832), which was not statistically significant.

Table 4.3

Correlations

		Emotion8	Emotion55
Emotion8	Pearson Correlation	1	-,068
	Sig. (2-tailed)		,554
	N	77	77
Emotion55	Pearson Correlation	-,068	1
	Sig. (2-tailed)	,554	
	N	77	77

The correlation result in table 3.3. shows a very weak negative relationship between the male human-produced sad recording (Emotion8) and the male AI-generated sad recording (Emotion55) for the sentence *"Dogs are sitting by the door."* The Pearson correlation is r=−0.068 with a p-value of p=0.554, indicating no statistically significant correlation.

Table 4.4

Correlations

		Emotion10	Emotion48
Emotion10	Pearson Correlation	1	-,100
	Sig. (2-tailed)		,387
	N	77	77
Emotion48	Pearson Correlation	-,100	1
	Sig. (2-tailed)	,387	
	N	77	77

Table 3.4. shows a weak negative correlation between the male human-produced (Emotion10) and AI-generated (Emotion48) angry recordings for the sentence *"Kids are talking by the door."* The Pearson correlation coefficient is r=−0.100, with a p-value of p=0.387, indicating the result is not statistically significant.

Table 4.5

Correlations

		Emotion11	Emotion12
Emotion11	Pearson Correlation	1	,124
	Sig. (2-tailed)		,281
	N	77	77
Emotion12	Pearson Correlation	,124	1
	Sig. (2-tailed)	,281	
	N	77	77

This correlation results in Table 3.5. shows a very weak positive relationship between the male AI-generated (Emotion11) and human-produced (Emotion12) happy recordings for the sentence *"Kids are talking by the door."* The Pearson correlation coefficient is r=0.124, with a p-value of p=0.281, indicating no statistically significant correlation.

Table 4.6

Correlations

		Emotion27	Emotion11
Emotion27	Pearson Correlation	1	,280*
	Sig. (2-tailed)		,014
	N	77	77
Emotion11	Pearson Correlation	,280*	1
	Sig. (2-tailed)	,014	
	N	77	77

*. Correlation is significant at the 0.05 level (2-tailed).

Table 3.6. reveals a moderate positive correlation between the AI-generated male recording (Emotion11) and the human-produced male recording (Emotion27) for the emotion *happy* and the sentence *"Kids are talking by the door."* The Pearson correlation coefficient is r=0.280, and the result is statistically significant with p=0.014. This indicates that participants' emotional perception of the AI and human recordings in this case was significantly aligned.

Table 4.7

Correlations		Emotion13	Emotion43
Emotion13	Pearson Correlation	1	,057
	Sig. (2-tailed)		,621
	N	77	77
Emotion43	Pearson Correlation	,057	1
	Sig. (2-tailed)	,621	
	N	77	77

Table 3.7. shows a very weak positive correlation between the male AI-generated (Emotion13) and human-produced (Emotion43) happy recordings for the sentence *"Dogs are sitting by the door."* The Pearson correlation is r=0.057, with a p-value of p=0.621, indicating no statistically significant relationship.

Table 4.8

Correlations		Emotion17	Emotion19
Emotion17	Pearson Correlation	1	,095
	Sig. (2-tailed)		,410
	N	77	77
Emotion19	Pearson Correlation	,095	1
	Sig. (2-tailed)	,410	
	N	77	77

Table 3.8. displays a very weak positive correlation (r=0.095) between the male human (Emotion17) and male AI (Emotion19) recordings for the emotion *anger* with the sentence *"Dogs are sitting by the door."* The p-value (p=0.410) indicates that the correlation is not statistically significant, suggesting low agreement among participants in recognizing similar emotional prosody between the two recordings.

Table 4.9

Correlations		Emotion18	Emotion42
Emotion18	Pearson Correlation	1	,138
	Sig. (2-tailed)		,232
	N	77	77
Emotion42	Pearson Correlation	,138	1
	Sig. (2-tailed)	,232	
	N	77	77

Table 3.9. indicates a weak positive correlation (r=0.138) between the male AI (Emotion18) and male human (Emotion42) recordings for the emotion *sad* using the sentence *"Kids are talking by the door."* The correlation is not statistically significant (p=0.232), suggesting that participants did not consistently perceive a similar emotional expression between the AI and human recordings in this instance.

Table 4.10

Correlations

		Emotion19	Emotion46
Emotion19	Pearson Correlation	1	-,010
	Sig. (2-tailed)		,933
	N	77	77
Emotion46	Pearson Correlation	-,010	1
	Sig. (2-tailed)	,933	
	N	77	77

In Table 3.10. the Pearson correlation coefficient between Emotion19 (AI Male) and Emotion46 (Human Male) is -0.010, indicating virtually no linear relationship. The p-value is 0.933, which is far from statistically significant.

Table 4.11

Correlations

		Emotion20	Emotion26
Emotion20	Pearson Correlation	1	,225*
	Sig. (2-tailed)		,049
	N	77	77
Emotion26	Pearson Correlation	,225*	1
	Sig. (2-tailed)	,049	
	N	77	77

*. Correlation is significant at the 0.05 level (2-tailed).

In Table 3.11. the Pearson correlation coefficient between Emotion20 (AI Female) and Emotion26 (Human Female) is 0.225, indicating a weak positive relationship. The p-value is 0.049, which is statistically significant at the 0.05 level (2-tailed), suggesting that participants' responses to the AI and human recordings of anger had a modest but meaningful association.

Table 4.12

Correlations

		Emotion21	Emotion36
Emotion21	Pearson Correlation	1	,045
	Sig. (2-tailed)		,696
	N	77	77
Emotion36	Pearson Correlation	,045	1
	Sig. (2-tailed)	,696	
	N	77	77

In Table 3.12. the Pearson correlation coefficient between Emotion21 (AI Female) and Emotion36 (Human Female) is 0.045, indicating a very weak positive relationship. The p-value is 0.696, which is not statistically significant. This suggests there is no meaningful correlation between participants' responses to the AI and human female recordings of anger for the sentence "Kids are talking by the door."

Table 4.13

Correlations

		Emotion25	Emotion93
Emotion25	Pearson Correlation	1	,341**
	Sig. (2-tailed)		,002
	N	77	77
Emotion93	Pearson Correlation	,341**	1
	Sig. (2-tailed)	,002	
	N	77	77

**. Correlation is significant at the 0.01 level (2-tailed).

In table 3.13. the Pearson correlation coefficient between Emotion25 (AI Female) and Emotion93 (Human Female) is 0.341, indicating a moderate positive relationship. The p-value is 0.002, which is statistically significant at the 0.01 level. This means there is a meaningful and significant correlation between participants' responses to the AI and human female recordings of the emotion "happy" for the sentence "Kids are talking by the door,".

Table 4.14

Correlations

		Emotion28	Emotion64
Emotion28	Pearson Correlation	1	,067
	Sig. (2-tailed)		,564
	N	77	77
Emotion64	Pearson Correlation	,067	1
	Sig. (2-tailed)	,564	
	N	77	77

In Table 3.14. the Pearson correlation coefficient between Emotion28 (AI Female) and Emotion64 (Human Female) is 0.067, indicating a very weak positive relationship. The p-value is 0.564, which is not statistically significant. This suggests there is no meaningful correlation between participants' responses to the AI and human female recordings of the emotion "anger" for the sentence "Kids are talking by the door," based on a sample size of 77.

Table 4.15

Correlations

		Emotion33	Emotion55
Emotion33	Pearson Correlation	1	,356**
	Sig. (2-tailed)		,001
	N	77	77
Emotion55	Pearson Correlation	,356**	1
	Sig. (2-tailed)	,001	
	N	77	77

**. Correlation is significant at the 0.01 level (2-tailed).

Table 3.15. shows that for the emotion "sad" and the sentence "Dogs are sitting by the door," Emotion33 (AI Male) and Emotion55 (Human Male) showed a moderate positive correlation ($r = .356$, $p = .001$), indicating a statistically significant similarity in participant responses.

Table 4.16

Correlations

		Emotion34	Emotion4
Emotion34	Pearson Correlation	1	-,128
	Sig. (2-tailed)		,268
	N	77	77
Emotion4	Pearson Correlation	-,128	1
	Sig. (2-tailed)	,268	
	N	77	77

Table 3.16. indicates that for the emotion "sad" and the sentence "Dogs are sitting by the door," Emotion34 (Human Female) and Emotion4 (AI Female) showed a weak negative correlation (r = –.128, p = .268), which was not statistically significant.

Table 4.17

Correlations

		Emotion37	Emotion75
Emotion37	Pearson Correlation	1	-,170
	Sig. (2-tailed)		,140
	N	77	77
Emotion75	Pearson Correlation	-,170	1
	Sig. (2-tailed)	,140	
	N	77	77

Table 3.17. shows that for the emotion "happy" and the sentence "Kids are talking by the door," Emotion37 (AI Male) and Emotion75 (Human Male) had a weak negative correlation (r = –.170, p = .140), which was not statistically significant.

Table 4.18

Correlations

		Emotion39	Emotion86
Emotion39	Pearson Correlation	1	,247*
	Sig. (2-tailed)		,030
	N	77	77
Emotion86	Pearson Correlation	,247*	1
	Sig. (2-tailed)	,030	
	N	77	77

*. Correlation is significant at the 0.05 level (2-tailed).

Table 3.18. shows that for the emotion "happy" and the sentence "Dogs are sitting by the door," Emotion39 (Human Female) and Emotion86 (AI Female) had a weak positive correlation (r = .247, p = .030), which was statistically significant.

Table 4.19

Correlations

		Emotion43	Emotion99
Emotion43	Pearson Correlation	1	,302**
	Sig. (2-tailed)		,008
	N	77	77
Emotion99	Pearson Correlation	,302**	1
	Sig. (2-tailed)	,008	
	N	77	77

**. Correlation is significant at the 0.01 level (2-tailed).

Table 3.19. shows that for the emotion "happy" and the sentence "Dogs are sitting by the door," Emotion43 (AI Male) and Emotion99 (Human Male) showed a moderate positive correlation (r = .302, p = .008), which was statistically significant.

Table 4.20

Correlations

		Emotion48	Emotion76
Emotion48	Pearson Correlation	1	-,014
	Sig. (2-tailed)		,907
	N	77	77
Emotion76	Pearson Correlation	-,014	1
	Sig. (2-tailed)	,907	
	N	77	77

According to Table 3.20., for the emotion "anger" and the sentence "Kids are talking by the door," Emotion48 (AI Male) and Emotion76 (Human Male) showed a very weak negative correlation (r = –.014, p = .907), which was not statistically significant.

Table 4.21

Correlations

		Emotion1	Emotion47
Emotion1	Pearson Correlation	1	,058
	Sig. (2-tailed)		,613
	N	77	77
Emotion47	Pearson Correlation	,058	1
	Sig. (2-tailed)	,613	
	N	77	77

Table 3.21. indicates that for the emotion "surprised" and the sentence "Dogs are sitting by the door," Emotion1 (AI Female) and Emotion47 (Human Female) showed a very weak positive correlation (r = .058, p = .613), which was not statistically significant.

Table 4.22

Correlations

		Emotion2	Emotion56
Emotion2	Pearson Correlation	1	-,100
	Sig. (2-tailed)		,388
	N	77	77
Emotion56	Pearson Correlation	-,100	1
	Sig. (2-tailed)	,388	
	N	77	77

In Table 3.22., for the emotion "fear" and the sentence "Kids are talking by the door," Emotion2 (Human Male) and Emotion56 (AI Male) exhibited a weak negative correlation ($r = -.100$, $p = .388$), which was not statistically significant.

Table 4.23

Correlations

		Emotion5	Emotion45
Emotion5	Pearson Correlation	1	,036
	Sig. (2-tailed)		,753
	N	77	77
Emotion45	Pearson Correlation	,036	1
	Sig. (2-tailed)	,753	
	N	77	77

In Table 3.23., for the emotion "fear" and the sentence "Dogs are sitting by the door," Emotion5 (Human Male) and Emotion45 (AI Male) showed a very weak positive correlation ($r = .036$, $p = .753$), which was not statistically significant.

Table 4.24

Correlations

		Emotion7	Emotion32
Emotion7	Pearson Correlation	1	,013
	Sig. (2-tailed)		,911
	N	77	77
Emotion32	Pearson Correlation	,013	1
	Sig. (2-tailed)	,911	
	N	77	77

According to Table 3.24., for the emotion "surprised" and the sentence "Dogs are sitting by the door," Emotion7 (Human Male) and Emotion32 (AI Male) demonstrated a negligible correlation ($r = .013$, $p = .911$), indicating no statistically significant relationship.

Table 4.25

Correlations

		Emotion9	Emotion22
Emotion9	Pearson Correlation	1	,080
	Sig. (2-tailed)		,492
	N	77	77
Emotion22	Pearson Correlation	,080	1
	Sig. (2-tailed)	,492	
	N	77	77

Table 3.25. shows that for the emotion "fear" and the sentence "Kids are talking by the door," Emotion9 (AI Female) and Emotion22 (Human Female) showed a weak, non-significant correlation (r = .080, p = .492).

Table 4.26

Correlations

		Emotion14	Emotion50
Emotion14	Pearson Correlation	1	-,134
	Sig. (2-tailed)		,245
	N	77	77
Emotion50	Pearson Correlation	-,134	1
	Sig. (2-tailed)	,245	
	N	77	77

Table 3.26. indicates that for the emotion "surprised" and the sentence "Kids are talking by the door," Emotion14 (AI Female) and Emotion50 (Human Female) showed a weak, non-significant negative correlation (r = -0.134, p = .245).

Table 4.27

Correlations

		Emotion16	Emotion23
Emotion16	Pearson Correlation	1	-,031
	Sig. (2-tailed)		,787
	N	77	77
Emotion23	Pearson Correlation	-,031	1
	Sig. (2-tailed)	,787	
	N	77	77

In Table 3.27., for the emotion "surprised" and the sentence "Kids are talking by the door," Emotion16 (Human Male) and Emotion23 (AI Male) showed a very weak, non-significant negative correlation (r = -0.031, p = .787).

Table 4.28

Correlations

		Emotion22	Emotion60
Emotion22	Pearson Correlation	1	,256*
	Sig. (2-tailed)		,025
	N	77	77
Emotion60	Pearson Correlation	,256*	1
	Sig. (2-tailed)	,025	
	N	77	77

*. Correlation is significant at the 0.05 level (2-tailed).

In Table 3.28., for the emotion "fear" and the sentence "Kids are talking by the door," Emotion22 (Human Female) and Emotion60 (AI Female) showed a statistically significant positive correlation ($r = .256$, $p = .025$).

Table 4.29

Correlations

		Emotion23	Emotion40
Emotion23	Pearson Correlation	1	-,111
	Sig. (2-tailed)		,335
	N	77	77
Emotion40	Pearson Correlation	-,111	1
	Sig. (2-tailed)	,335	
	N	77	77

Table 3.29. indicates that for the emotion "surprised" and the sentence "Kids are talking by the door," Emotion23 (AI Male) and Emotion40 (Human Male) showed a weak negative correlation ($r = -0.111$, $p = 0.335$), which was not statistically significant.

Table 4.30

Correlations

		Emotion24	Emotion51
Emotion24	Pearson Correlation	1	-,008
	Sig. (2-tailed)		,942
	N	77	77
Emotion51	Pearson Correlation	-,008	1
	Sig. (2-tailed)	,942	
	N	77	77

In Table 3.30., for the emotion "surprised" and the sentence "Kids are talking by the door," Emotion24 (AI Female) and Emotion51 (Human Female) had a near-zero negative correlation ($r = -0.008$, $p = 0.942$), which was not statistically significant.

Table 4.31

Correlations

		Emotion29	Emotion65
Emotion29	Pearson Correlation	1	,028
	Sig. (2-tailed)		,811
	N	77	77
Emotion65	Pearson Correlation	,028	1
	Sig. (2-tailed)	,811	
	N	77	77

Table 3.31. shows that for the emotion "fear" and the sentence "Dogs are sitting by the door," Emotion29 (Human Female) and Emotion65 (AI Female) showed a very weak positive correlation (r = 0.028, p = 0.811), which was not statistically significant

Table 4.32

Correlations

		Emotion30	Emotion91
Emotion30	Pearson Correlation	1	-,014
	Sig. (2-tailed)		,906
	N	77	77
Emotion91	Pearson Correlation	-,014	1
	Sig. (2-tailed)	,906	
	N	77	77

According to Table 3.32., for the emotion "fear" and the sentence "Kids are talking by the door," Emotion30 (Human Male) and Emotion91 (AI Male) showed a negligible negative correlation (r = -0.014, p = 0.906), which was not statistically significant.

Table 4.33

Correlations

		Emotion31	Emotion9
Emotion31	Pearson Correlation	1	,067
	Sig. (2-tailed)		,562
	N	77	77
Emotion9	Pearson Correlation	,067	1
	Sig. (2-tailed)	,562	
	N	77	77

In Table 3.33., for the emotion "fear" and the sentence "Kids are talking by the door," Emotion31 (Human Female) and Emotion9 (AI Female) showed a very weak positive correlation (r = 0.067, p = 0.562), which was not statistically significant.

Table 4.34

Correlations

		Emotion35	Emotion70
Emotion35	Pearson Correlation	1	-,017
	Sig. (2-tailed)		,886
	N	77	77
Emotion70	Pearson Correlation	-,017	1
	Sig. (2-tailed)	,886	
	N	77	77

Table 3.34. shows that for the emotion "surprised" and the sentence "Kids are talking by the door," Emotion35 (AI Male) and Emotion70 (Human Male) showed a negligible negative correlation (r = -0.017, p = 0.886), which was not statistically significant.

Table 4.35

Correlations

		Emotion38	Emotion79
Emotion38	Pearson Correlation	1	,059
	Sig. (2-tailed)		,608
	N	77	77
Emotion79	Pearson Correlation	,059	1
	Sig. (2-tailed)	,608	
	N	77	77

Table 3.35. indicates that for the emotion "fear" and the sentence "Dogs are sitting by the door," Emotion38 (AI Male) and Emotion79 (Human Male) had a very weak positive correlation (r = 0.059, p = 0.608), which was not statistically significant.

Table 4.36

Correlations

		Emotion44	Emotion77
Emotion44	Pearson Correlation	1	,036
	Sig. (2-tailed)		,756
	N	77	77
Emotion77	Pearson Correlation	,036	1
	Sig. (2-tailed)	,756	
	N	77	77

In Table 3.36., for the emotion "surprised" and the sentence "Dogs are sitting by the door," Emotion44 (AI Female) and Emotion77 (Human Female) had a very weak positive correlation (r = 0.036, p = 0.756), which was not statistically significant.

Table 4.37

Correlations

		Emotion45	Emotion95
Emotion45	Pearson Correlation	1	-,096
	Sig. (2-tailed)		,407
	N	77	77
Emotion95	Pearson Correlation	-,096	1
	Sig. (2-tailed)	,407	
	N	77	77

Table 3.37. shows that for the emotion "fear" and the sentence "Dogs are sitting by the door," Emotion45 (AI Male) and Emotion95 (Human Male) showed a weak negative correlation ($r = -0.096$, $p = 0.407$), which was not statistically significant.

Table 4.38

Correlations

		Emotion49	Emotion94
Emotion49	Pearson Correlation	1	,046
	Sig. (2-tailed)		,692
	N	77	77
Emotion94	Pearson Correlation	,046	1
	Sig. (2-tailed)	,692	
	N	77	77

Table 3.38. shows that for the emotion "happy" and the sentence "Dogs are sitting by the door," Emotion49 (AI Male) and Emotion94 (Human Male) had a very weak positive correlation ($r = 0.046$, $p = 0.692$), which was not statistically significant.

Table 4.39

Correlations

		Emotion52	Emotion72
Emotion52	Pearson Correlation	1	,190
	Sig. (2-tailed)		,098
	N	77	77
Emotion72	Pearson Correlation	,190	1
	Sig. (2-tailed)	,098	
	N	77	77

Table 3.39. indicates that for the emotion "sad" and the sentence "Dogs are sitting by the door," Emotion52 (AI Female) and Emotion72 (Human Female) showed a weak positive correlation ($r = 0.190$, $p = 0.098$), which was not statistically significant.

Table 4.40

Correlations

		Emotion53	Emotion62
Emotion53	Pearson Correlation	1	-,031
	Sig. (2-tailed)		,790
	N	77	77
Emotion62	Pearson Correlation	-,031	1
	Sig. (2-tailed)	,790	
	N	77	77

In Table 3.40., for the emotion "sad" and the sentence "Kids are talking by the door," Emotion53 (AI Male) and Emotion62 (Human Male) exhibited a very weak negative correlation (r = -0.031, p = 0.790), which was not statistically significant.

Table 4.41

Correlations

		Emotion54	Emotion14
Emotion54	Pearson Correlation	1	-,094
	Sig. (2-tailed)		,417
	N	77	77
Emotion14	Pearson Correlation	-,094	1
	Sig. (2-tailed)	,417	
	N	77	77

According to Table 3.41., for the emotion "surprised" and the sentence "Kids are talking by the door," Emotion54 (Human Female) and Emotion14 (AI Female) showed a weak negative correlation (r = -0.094, p = 0.417), which was not statistically significant.

Table 4.42

Correlations

		Emotion57	Emotion59
Emotion57	Pearson Correlation	1	,152
	Sig. (2-tailed)		,186
	N	77	77
Emotion59	Pearson Correlation	,152	1
	Sig. (2-tailed)	,186	
	N	77	77

In Table 3.42., for the emotion "sad" and the sentence "Kids are talking by the door," Emotion57 (Human Female) and Emotion59 (AI Female) had a weak positive correlation (r = 0.152, p = 0.186), which was not statistically significant.

Table 4.43

Correlations

		Emotion58	Emotion65
Emotion58	Pearson Correlation	1	-,118
	Sig. (2-tailed)		,308
	N	77	77
Emotion65	Pearson Correlation	-,118	1
	Sig. (2-tailed)	,308	
	N	77	77

Table 3.43. shows that for the emotion "fear" and the sentence "Dogs are sitting by the door," Emotion58 (Human Female) and Emotion65 (AI Female) showed a weak negative correlation (r = -0.118, p = 0.308), which was not statistically significant.

Table 4.44

Correlations

		Emotion59	Emotion71
Emotion59	Pearson Correlation	1	,164
	Sig. (2-tailed)		,154
	N	77	77
Emotion71	Pearson Correlation	,164	1
	Sig. (2-tailed)	,154	
	N	77	77

In Table 3.44., for the emotion "sad" and the sentence "Kids are talking by the door," Emotion59 (AI Female) and Emotion71 (Human Female) showed a weak positive correlation (r = 0.164, p = 0.154), which was not statistically significant.

Table 4.45

Correlations

		Emotion60	Emotion81
Emotion60	Pearson Correlation	1	,198
	Sig. (2-tailed)		,085
	N	77	77
Emotion81	Pearson Correlation	,198	1
	Sig. (2-tailed)	,085	
	N	77	77

Table 3.45. shows that for the emotion "fear" and the sentence "Kids are talking by the door," Emotion60 (AI Female) and Emotion81 (Human Female) had a weak positive correlation (r = 0.198, p = 0.085), which was not statistically significant.

Table 4.46

Correlations

		Emotion61	Emotion20
Emotion61	Pearson Correlation	1	,223
	Sig. (2-tailed)		,052
	N	77	77
Emotion20	Pearson Correlation	,223	1
	Sig. (2-tailed)	,052	
	N	77	77

Table 3.46. indicates that for the emotion "anger" and the sentence "Dogs are sitting by the door," Emotion61 (Human Female) and Emotion20 (AI Female) showed a weak positive correlation (r = 0.223, p = 0.052), which approached but did not reach statistical significance.

Table 4.47

Correlations

		Emotion63	Emotion74
Emotion63	Pearson Correlation	1	,085
	Sig. (2-tailed)		,461
	N	77	77
Emotion74	Pearson Correlation	,085	1
	Sig. (2-tailed)	,461	
	N	77	77

According to the corralation shown in Table 3.47., for the emotion "surprised" and the sentence "Dogs are sitting by the door," Emotion63 (Human Male) and Emotion74 (AI Male) demonstrated a very weak positive correlation (r = 0.085, p = 0.461), which was not statistically significant.

Table 4.48

Correlations

		Emotion65	Emotion69
Emotion65	Pearson Correlation	1	,113
	Sig. (2-tailed)		,329
	N	77	77
Emotion69	Pearson Correlation	,113	1
	Sig. (2-tailed)	,329	
	N	77	77

Table 3.48., shows that for the emotion "fearful" and the sentence "Dogs are sitting by the door," Emotion65 (AI Female) and Emotion69 (Human Female) showed a very weak positive correlation (r = 0.113, p = 0.329), which was not statistically significant.

Table 4.49

Correlations

		Emotion66	Emotion74
Emotion66	Pearson Correlation	1	-,013
	Sig. (2-tailed)		,914
	N	77	77
Emotion74	Pearson Correlation	-,013	1
	Sig. (2-tailed)	,914	
	N	77	77

As shown in Table 3.49., for the emotion "surprised" and the sentence "Dogs are sitting by the door," Emotion66 (Human Male) and Emotion74 (AI Male) showed a negligible negative correlation (r = -0.013, p = 0.914), which was not statistically significant.

Table 4.50

Correlations

		Emotion67	Emotion86
Emotion67	Pearson Correlation	1	,175
	Sig. (2-tailed)		,129
	N	77	77
Emotion86	Pearson Correlation	,175	1
	Sig. (2-tailed)	,129	
	N	77	77

Table 3.50. shows that for the emotion "happy" and the sentence "Dogs are sitting by the door," Emotion67 (Human Female) and Emotion86 (AI Female) showed a small positive correlation (r = 0.175, p = 0.129), which was not statistically significant.

Table 4.51

Correlations

		Emotion68	Emotion82
Emotion68	Pearson Correlation	1	,274*
	Sig. (2-tailed)		,016
	N	77	77
Emotion82	Pearson Correlation	,274*	1
	Sig. (2-tailed)	,016	
	N	77	77

*. Correlation is significant at the 0.05 level (2-tailed).

Table 3.51. indicates that for the emotion "anger" and the sentence "Dogs are sitting by the door," Emotion68 (AI Male) and Emotion82 (Human Male) showed a moderate positive correlation (r = 0.274, p = 0.016), which was statistically significant at the 0.05 level.

Table 4.52

Correlations

		Emotion69	Emotion84
Emotion69	Pearson Correlation	1	-,216
	Sig. (2-tailed)		,060
	N	77	77
Emotion84	Pearson Correlation	-,216	1
	Sig. (2-tailed)	,060	
	N	77	77

Table 3.52 shows that for the emotion "fear" and the sentence "Dogs are sitting by the door," Emotion69 (Human Female) and Emotion84 (AI Female) had a negative correlation (r = -0.216, p = 0.060), which approached but did not reach statistical significance.

Table 4.53

Correlations

		Emotion73	Emotion78
Emotion73	Pearson Correlation	1	,264*
	Sig. (2-tailed)		,020
	N	77	77
Emotion78	Pearson Correlation	,264*	1
	Sig. (2-tailed)	,020	
	N	77	77

*. Correlation is significant at the 0.05 level (2-tailed).

According to the correlation results in Table 3.53., for the emotion "sad" and the sentence "Dogs are sitting by the door," Emotion73 (AI Male) and Emotion78 (Human Male) had a statistically significant positive correlation (r = 0.264, p = 0.020).

Table 4.54

Correlations

		Emotion80	Emotion6
Emotion80	Pearson Correlation	1	,072
	Sig. (2-tailed)		,533
	N	77	77
Emotion6	Pearson Correlation	,072	1
	Sig. (2-tailed)	,533	
	N	77	77

Table 3.54. shows that for the emotion "anger" and the sentence "Dogs are sitting by the door," Emotion80 (Human Female) and Emotion6 (AI Female) showed a weak positive correlation (r = 0.072, p = 0.533), which was not statistically significant.

Table 4.55

Correlations

		Emotion83	Emotion21
Emotion83	Pearson Correlation	1	,064
	Sig. (2-tailed)		,577
	N	77	77
Emotion21	Pearson Correlation	,064	1
	Sig. (2-tailed)	,577	
	N	77	77

Table 3.55. indicates that for the emotion "anger" and the sentence "Kids are talking by the door," Emotion83 (Human Female) and Emotion21 (AI Female) showed a very weak positive correlation (r = 0.064, p = 0.577), which was not statistically significant.

Table 4.56

Correlations

		Emotion85	Emotion53
Emotion85	Pearson Correlation	1	,133
	Sig. (2-tailed)		,248
	N	77	77
Emotion53	Pearson Correlation	,133	1
	Sig. (2-tailed)	,248	
	N	77	77

In Table 3.56., for the emotion "sad" and the sentence "Kids are talking by the door," Emotion85 (Human Male) and Emotion53 (AI Male) had a weak positive correlation (r = 0.133, p = 0.248), which was not statistically significant.

Table 4.57

Correlations

		Emotion88	Emotion96
Emotion88	Pearson Correlation	1	,033
	Sig. (2-tailed)		,779
	N	77	77
Emotion96	Pearson Correlation	,033	1
	Sig. (2-tailed)	,779	
	N	77	77

Table 3.57. shows that for the emotion "sad" and the sentence "Kids are talking by the door," Emotion88 (Human Female) and Emotion96 (AI Female) had a very weak positive correlation (r = 0.033, p = 0.779), which was not statistically significant.

Table 4.58

Correlations

		Emotion89	Emotion90
Emotion89	Pearson Correlation	1	,156
	Sig. (2-tailed)		,177
	N	77	77
Emotion90	Pearson Correlation	,156	1
	Sig. (2-tailed)	,177	
	N	77	77

Table 3.58. indicates that for the emotion "happy" and the sentence "Dogs are sitting by the door," Emotion89 (AI Female) and Emotion90 (Human Female) showed a weak positive correlation (r = 0.156, p = 0.177), which was not statistically significant.

Table 4.59

Correlations

		Emotion92	Emotion100
Emotion92	Pearson Correlation	1	,100
	Sig. (2-tailed)		,387
	N	77	77
Emotion100	Pearson Correlation	,100	1
	Sig. (2-tailed)	,387	
	N	77	77

According to the correlation results in Table 3.59., for the emotion "anger" and the sentence "Kids are talking by the door," Emotion92 (Human Male) and Emotion100 (AI Male) showed a weak positive correlation (r = 0.100, p = 0.387), which was not statistically significant.

Table 4.60

Correlations

		Emotion97	Emotion1
Emotion97	Pearson Correlation	1	-,019
	Sig. (2-tailed)		,873
	N	77	77
Emotion1	Pearson Correlation	-,019	1
	Sig. (2-tailed)	,873	
	N	77	77

Table 3.60. indicates that for the emotion "surprised" and the sentence "Dogs are sitting by the door," Emotion97 (Human Female) and Emotion1 (AI Female) showed a negligible negative correlation (r = -0.019, p = 0.873), which was not statistically significant.

Table 4.61

Correlations

		Emotion98	Emotion25
Emotion98	Pearson Correlation	1	,081
	Sig. (2-tailed)		,484
	N	77	77
Emotion25	Pearson Correlation	,081	1
	Sig. (2-tailed)	,484	
	N	77	77

Table 3.61. shows that for the emotion "happy" and the sentence "Kids are talking by the door," Emotion98 (Human Female) and Emotion25 (AI Female) showed a weak positive correlation (r = 0.081, p = 0.484), which was not statistically significant.

5 DISCUSSION

5.1 Overview

The purpose of this study was to compare the perceptual clarity and recognizability of emotional prosody in human and AI-generated speech. Participants were asked to determine the emotional tone of a series of AI and human-generated statements, chosen from anger, happiness, fear, surprise, and sadness, while remaining blind to the recordings' source. Each participant chose the feeling they recognized in response to the recordings, without knowing if they were made by humans or by artificial intelligence. The alignment between AI and human interpretations of identical phrases across emotions was evaluated by analyzing the results of statistical analysis using Pearson correlation. This study attempted to determine whether listeners could reliably and consistently recognize the emotions conveyed by synthetic voices and whether these feelings corresponded to the equivalent human renditions of identical sentences using a series of Pearson correlation analyses ($n = 61$). The results have immediate implications for the development of artificial intelligence that is emotionally expressive and offer subtle insights into the strengths and weaknesses of existing Text-to-Speech (TTS) systems.

5.2 Key Findings

5.2.1 Happiness

Happiness produced the most consistent perceptual alignment of any emotional category. Several AI-human pairs showed statistically significant correlations. For instance, there was a moderately positive connection ($r = 0.280$, $p = 0.014$) between the pairs Emotion11 (AI Male) and Emotion27 (Human Male), but a stronger association ($r = 0.341$, $p = 0.002$) between Emotion25 (AI Female) and Emotion93 (Human Female). These results imply that modern TTS systems are better at synthesizing happiness, an emotion that is frequently linked to certain prosodic characteristics like higher pitch and more intensity. Furthermore, Emotion43 (AI Male) and Emotion99 (Human Male) showed a statistically significant connection ($r = 0.302$, $p = 0.008$) in the phrase "Dogs are sitting by the door," indicating similar success. These findings suggest that, particularly when presented in male voices, happiness is the emotion in synthetic speech that is most consistently perceived.

5.2.2 Sadness

In male voice pairings, sadness showed less consistent results, with two statistically significant relationships found. There was a significant correlation between Emotion33 (AI Male) and Emotion55 (Human Male) ($r = 0.356$, $p = 0.001$), as well as between Emotion73 (AI Male) and Emotion78 (Human Male) ($r = 0.264$, $p = 0.020$). Despite being less intense than happiness, these findings imply that male synthetic voices can effectively portray sadness.

Nevertheless, a large number of female voice pairs produced no noteworthy outcomes. There may be inconsistencies in the modeling of this emotion, as evidenced by the small and non-significant correlation between Emotion4 (AI Female) and Emotion15 (Human Female) ($r =-0.025$, $p = 0.832$).

5.2.3 Anger

There was limited perceptual alignment for the feeling of anger. Emotion20 (AI Female) and Emotion26 (Human Female) ($r = 0.225$, $p = 0.049$) and Emotion68 (AI Male) and Emotion82 (Human Male) ($r = 0.274$, $p = 0.016$) were the only two statistically significant associations found. Both occurrences took place within the sentence "Dogs are sitting by the door," indicating that the recognition of anger in synthetic speech may be impacted by context-specific wording. Otherwise, most pairs had little to no association, suggesting a more general difficulty in capturing the subtleties of angry prosody.

5.2.4 Fear

The emotion that proved to be the most challenging for AI synthesis was fear. Emotion22 (human female) and Emotion60 (AI female) showed the only significant link ($r = 0.256$, $p = 0.025$). Weak or negative correlations were found in the remaining pairings, indicating inconsistent perceptual alignment. For instance, a non-significant correlation of $r = 0.036$ ($p = 0.753$) was found between Emotion5 (Human Male) and Emotion45 (AI Male), indicating the difficulties of synthesizing fear and the likely need for elaborate auditory modulations such breathiness or temporal jitter.

5.2.5 Surprise

The emotion that was least effectively expressed was surprise. There were no statistically significant associations found in any of the combinations. Emotion1 (AI Female) and

Emotion47 (Human Female) ($r = 0.058$, $p = 0.613$), as well as Emotion23 (AI Male) and Emotion40 (Human Male) ($r = -0.111$, $p = 0.335$), all had correlation coefficients that were close to zero or negative. This might be a reflection of how difficult it is to synthesize surprise, which usually entails abrupt changes in speech rate, dynamic range, and pitch—elements that existing AI TTS models might not be able to adequately represent.

5.3 Implications

The study's conclusions have a number of significant implications for the advancement of artificial intelligence voice synthesis that is emotionally expressive. It was demonstrated that human and AI voices perceptually matched emotions like happiness and, to a lesser extent, sadness, especially when male voices were employed. This implies that existing TTS (text-to-speech) systems are better able to capture these emotions, a finding that is corroborated by earlier studies on acoustic emotion modeling. According to recent research, certain prosodic characteristics—like higher pitch and a faster tempo for happiness and reduced pitch fluctuation and a slower rate for sadness—are easier for neural TTS models to duplicate because they are more consistent (Kolekar et al., 2024). According to empirical studies of TTS systems, male voices are more effective than female voices at producing perceptually clear emotional speech because their pitch contours are more consistent and their acoustic variance is lower (Kolekar et al., 2024).

In addition, the findings validate current affective computing initiatives to enhance emotional authenticity in interactions between humans and computers. More user engagement, trust, and communication success can be facilitated by improved emotional expressiveness in AI-driven platforms including educational tools, digital assistants, and therapeutic bots (Liu et al., 2023; Ao & Yuan, 2024). This study thus adds to the larger conversation about moral and socially conscious AI by highlighting the necessity of emotion-aware design that is perceptually verified by human listeners.

5.4 Limitations

A number of limitations must be noted, even if the study offers insightful information about the perceptual identification of emotional prosody in AI-generated speech. Initially, the sample size was limited to a convenient subset of English literature students at Ankara Hacı Bayram Veli University. This homogeneous participant pool might restrict the findings' applicability to larger groups with more varied cultural and linguistic backgrounds, which could have an impact on how emotional prosody is interpreted by different demographic groups.

Second, a controlled laboratory setting was not possible for the experiment. Participants used Google Forms to complete the listening task remotely while in a variety of acoustic environments and with their own audio equipment, such as speakers or headphones. The credibility of the results may have been impacted by the lack of environmental control, which introduces heterogeneity in audio perception and focus even when headphones are instructed to be used.

Third, only two publicly available TTS systems—Genny AI and Murf AI—were used to synthesize emotional speech because of resource limitations. This restricted the range of technological representation and diversity in AI voices. It's feasible that the emotional clarity and recognizability of various commercial or research-grade TTS models will differ.

Lastly, the 300 stimuli and long, repeating format of the questionnaire might have eventually caused participant weariness and concentration decrease. Especially in the second half of the trial, this cognitive strain might have added pollution to the results. Quicker, more engaging layouts or adaptive testing methods that reduce exhaustion without reducing data richness would be advantageous for future studies.

5.5 Recommendations for Future Research

In order to improve our comprehension of emotional prosody in AI-generated speech, a number of research directions are suggested, building on the results and constraints of the current study. First, future research should focus on increasing the participants' language and demographic variety. Conclusions on the interpretation of speech emotional signals by various listener populations could be more broadly drawn by including a larger sample, both geographically and culturally. Second, acoustically controlled laboratory settings should be used for future research. The reliability of perceptual judgments would be improved, and external variability would be decreased by standardization replay devices, loudness, and environmental circumstances. Furthermore, to offer multimodal insights into emotional perception, physiological and neurocognitive measurements like eye tracking, galvanic skin response, or EEG could be included. Third, it would be beneficial to evaluate a broader variety of TTS systems, such as innovative research-based models like DiffStyleTTS and commercial platforms like Amazon Polly, Google WaveNet, or Microsoft Azure (Liu et al., 2024). These kinds of comparisons might be useful in identifying the precise computational or auditory elements that support effective emotional synthesis.

Furthermore, emphasis should be placed on creating adaptive AI voices that alter their emotional prosody in response to user and environmental input. This strategy is in line with current developments in affective computation and human-computer interaction, where user engagement and trust are greatly influenced by emotional expressiveness (Ao & Yuan, 2024).

6 CONCLUSION

The purpose of the study was to investigate a crucial area of affective computing: the clarity and perceptual recognizability of emotional prosody in AI-generated speech in comparison to utterances produced by humans, using five core emotions; happiness, sadness, anger, fear, and surprise. Also, two different sentence contexts; "Kids are talking by the door." and "Dogs are sitting by the door.". By means of Pearson correlation studies on 61 matched pairs and listener assessments using a Visual Analogue Scale (VAS), this study offered a thorough analysis of current Text-to-Speech capabilities and emotional expressiveness. The results show that current Text-to-Speech (TTS) programs have both notable shortcomings and encouraging advances.

The study produced some interesting results. Multiple statistically significant correlations showed that happiness was the most reliably synthesized emotion, with male voices in particular exhibiting perceptual clarity and alignment. There were moderate and statistically significant positive correlations between emotions like sadness and happiness. In particular, male voice pairs showed sadness (up to $r = 0.36$) while male voice pairings showed happiness ($r \approx 0.28$–0.34). These findings imply that the related auditory markers—higher pitch and speed for happy, lower pitch and slower tempo for sadness—are accurately described in existing AI systems and show that these emotions can be successfully synthesized by AI in specific situations. These findings provide important insights into the strengths of existing TTS technology by indicating that specific emotional states with distinct prosodic indicators, such pitch and tempo, can be successfully replicated in AI-generated speech. However, there was inconsistent or non-significant alignment with human-produced speech for more complex emotions like fear, surprise, and even anger. This difference highlights the need for more complex modeling methods that take into consideration context-sensitive emotional triggers, dynamic prosodic shifts, and cross-linguistic variations in emotional presentation.

Additionally, the inconsistent results for anger—significant alignment was only observed in very specific sentence contexts—indicate that linguistic framing, structure of sentences, and even gender-specific vocal characteristics have an impact on emotional intelligibility in addition to TTS algorithmic competence. These discoveries have important wider ramifications. The ability of machines to accurately represent and interpret emotional states will become a fundamental need for applications ranging from virtual assistants to therapeutic agents and educational technologies as artificial intelligence becomes more and more incorporated into human communication. If affective computing is to promote empathy, trust, and user happiness, it must include emotional fluency in addition to technical fluency. Future research ought to be based on multidisciplinary cooperation and integrate knowledge from signal processing, psychology, languages, and human-computer interaction in order to do this. Training and assessing next-generation TTS systems will require investments in inclusive experimental designs, real-time feedback mechanisms, and high-quality emotional speech corpora. Furthermore, listener-focused evaluation techniques—like the one used in this study—will continue to be essential for determining how useful emotionally expressive AI is in the actual world.

There was a clear gender-based trend: male AI voices continuously outperformed female voices in terms of perceptual congruence with human voices. This pattern highlights gender differences in prosody modeling, which may be caused by male speech's more stable acoustic profiles and smaller pitch range, which make them simpler to algorithmically reproduce. These results are consistent with previous research on emotional constancy in synthetic speech and TTS voice modeling.

These findings have a variety of implications. First, they show that although some emotions, like happiness and sadness, may be rendered by existing TTS systems (particularly in male voices), displays of fear, surprise, and subtle kinds of anger still present considerable difficulties. This implies that emotion-specific training and improvement in TTS processes are in great need, with a particular emphasis on context-aware prosody, pitch dynamics, and timbral variations.

Second, the gender-related differences indicate that multi-speaker training regimens or gender-adaptive modeling methodologies which ensure expressiveness of emotion equivalency across voice types should be used in future systems. In addition to boosting user confidence, achieving balance in emotional intelligibility promotes more inclusive AI communication practices. The results also have applications in the developing field of affective computing. In human–computer interaction, emotional prosody has a major impact on coherence, empathy, and user engagement. Thus, thoughtful advancements in emotion-rich TTS have the potential to improve user experience in virtual assistants, assistive technology, therapeutic applications, and educational platforms.

The study's limitations, which include a homogeneous sample group, uncontrolled listening surroundings, potential participant fatigue from 300 stimuli, and dependence on only two free AI voices, must be taken into consideration when drawing these conclusions. To more thoroughly validate emotional perception, future studies should include lab-based controls or multimodal metrics, increase participant diversity, and incorporate more reliable commercial TTS platforms.

In conclusion, this study not only develops the connection between acoustic modeling and user-focused evaluation, but also lays out the current state of emotional prosody in AI voice synthesis, highlighting what works, what doesn't, and why. In order to develop voices that are not only understandable but also emotionally resonant and humanlike, the next wave of emotionally grounded TTS systems will be guided by revealing the areas in which AI succeeds and fails in emotional conveyance.

EXTENDED ABSTRACT

Yapay Zeka Tabanlı Metin-Konuşma Sistemleriyle Üretilen Duygusal Bürünün İşitsel Algılanmasına Yönelik Karşılaştırmalı Bir İnceleme

Yapay zeka tarafından üretilen konuşmanın kendini ifade etme yeteneği, yapay zekanın (AI) günlük yaşama giderek daha fazla entegre olmasıyla birlikte artan eleştirilere maruz kalmaktadır. Duygusal bürün, yani sözcüksel içeriğin ötesinde duygusal nüansları iletmek için ses perdesinin, ritminin, yoğunluğunun ve tınısının değiştirilmesi, gerçek insan-bilgisayar etkileşimi için olmazsa olmazdır. Bu tezde, dinleyicilerin yapay zeka tarafından üretilen konuşmadaki duygusal bürünü ve insan konuşmasını ayırt etme ve tanımlama yeteneği incelenmektedir. Özellikle yapay zekâ metinden sese (TTS) programlarının mutluluk, üzüntü, korku, öfke ve şaşkınlık gibi duyguları etkili bir şekilde ifade edip edemeyeceği ve bu performansın duygu türüne, cinsiyete ve cümle içeriğine göre nasıl farklılık gösterdiği araştırılmaktadır

Duygusal bürün, konuşmaya duygusal anlam ve pratik bağlam kazandırarak, iletişimin temel bir bileşeni olarak hizmet eder (Banse ve Scherer, 1996; Scherer, 2013). Derin öğrenmedeki gelişmeler sayesinde TTS sistemleri artık daha doğal konuşmalar üretebilse de insan duygusal ifadesinin nüanslarını taklit etmek hâlâ zordur (Ao & Yuan, 2024; James et al., 2023; Liu et al., 2024). Bu çalışma, iki ücretsiz TTS motoru olan Genny AI ve Murf AI'nın dinleyici tabanlı derecelendirmeler kullanarak insan konuşmasıyla algılanan benzerliğe ulaşıp ulaşmadığını incelemektedir. Yapay zeka tarafından üretilen duygusal bürünün tutarlılığını ve tanınabilirliğini değerlendirmek için dinleyici tepkileri Görsel Analog Skala (VAS) kullanılarak analiz edilmiştir.

Bu çalışmanın amacı, dinleyicilerin yapay zekâ metinden sese (TTS) sistemleri tarafından üretilen duygusal bürünü ne ölçüde tanımlayıp anlayabildiklerini araştırmaktır. Çalışma, dinleyicilerin özellikle yapay zekâ tarafından üretilen konuşma ile mutluluk, üzüntü, korku, öfke veya şaşkınlık gibi duygular arasında doğru bir şekilde ayrım yapıp yapamadıklarını incelemektedir. Göze çarpan farklılıkları tespit etmek için, yapay zekâ tarafından üretilen ifadenin algılanan duygusal netliği, insan konuşmasının netliğiyle de karşılaştırılmaktadır. Ayrıca, "Kids are talking by the door" ve "Dogs are sitting by the door," ifadelerine verilen yanıtları karşılaştırarak, söylenen cümlenin içeriğinin duyguların algılanmasını etkileyip etkilemediği de araştırılmaktadır. Duygusal ipuçlarının anlaşılabilirliği üzerinde erkek veya kadın seslerinin bir etkisinin olup olmadığının incelenmesi, ses cinsiyetinin olası düzenleyici etkisine ilişkin bir diğer araştırma alanıdır. Son olarak, çalışma, yanlış anlaşılma olasılığı daha yüksek olan duyguları ve yapay zeka sistemleri tarafından hangilerinin daha uygun şekilde ifade edildiğini belirlemeyi amaçlamaktadır.

Kayıtlar, kontrollü bir çevrimiçi ortamda, Ankara Hacı Bayram Veli Üniversitesi'nde öğrenim gören, tamamı ana dili Türkçe olan ve İngilizce konuşabilen 77 lisans öğrencisi tarafından değerlendirildi. Katılımcılara kulaklık aracılığıyla rastgele bir sırayla sunulan 100 adet sesli uyaran dinletildi (RAVDESS korpusundan alınan 60 gerçek kayıt (duygu ve cinsiyete göre dengelenmiş cümle başına 30 kayıt) ve yapay zeka tarafından oluşturulan 40 kayıt (duygu ve cinsiyete göre dengelenmiş, her sistemden 20 cümle)). Dinleyicilerden Google Forms kullanılarak VAS üzerinden duygusal netlik puanı vermeleri ve hissettikleri duygu türlerini seçmeleri istendi.

Yanıtlar Microsoft Excel'e aktarıldıktan sonra IBM SPSS kullanılarak değerlendirildi. Cinsiyet, duygu ve metin açısından eşleştirilen AI-insan çiftlerinin VAS puanlarıyla eşleşen Pearson korelasyon katsayıları analizin ana odağıydı. Duygu, cümle ve cinsiyete ilişkin 61

eşleşme ayrı ayrı değerlendirildi. Algısal uyumu saptamak için istatistiksel anlamlılık .05 ve .01 düzeylerinde değerlendirildi. Mutluluk, özellikle erkek ses örneklerinde en tutarlı algısal uyumu gösterdi. Örneğin, "Çocuklar" ifadeleri için Duygu11–Duygu27 ($r = .280, p = .014$) ve Duygu25–Duygu93 ($r = .341, p = .002$) ve "Köpekler" ifadeleri için Duygu43–Duygu99 ($r = .302, p = .008$) anlamlı, orta düzeyde pozitif ilişkiler gösterdi. Duygu33–Duygu55, $r = .356$'ya ($p = .001$) ulaştı; bu, özellikle erkek AI seslerinde üzüntüyle anlamlı bir uyum olduğunu gösteriyor. Öfke ifadesi düzensizdi. Orta düzeyde uyum sadece az sayıda erkek ve kadın ses eşleşmesinde bulundu (Duygu68–Duygu82: $r = .274, p = .016$; Duygu20–Duygu26: $r = .225, p = .049$). Yapay zeka korku duygusunu üretme konusunda zayıftı. Sadece bir kadın ses kaydı çifti, $r = .256$'da ($p = .025$), (Duygu60–Duygu22) istatistiksel olarak anlamlılığa ulaştı. Benzer olarak, şaşkınlık duygusu da zayıf bir performans gösterdi; tüm eşleştirmeler düşük ve önemsiz korelasyonlar üretti (genellikle $r = .10$'un altında), bu da yapay zekanın şaşkınlığın ve korku duygusunun altında yatan işitsel özellikleri yeterli bir şekilde yeniden üretemediğini gösteriyor. Ayrıca, kadın sesleriyle karşılaştırıldığında erkek yapay zeka seslerinin, insan sesleriyle duygusal netlik açısından daha güçlü korelasyonlar ürettiği görüldü. TTS sistemlerinde cinsiyet kapsayıcı modellemeye olan ihtiyaç, erkek konuşmasının daha dar perde değişimi ve akustik kararlılığı ile açıklanabilecek olan bu cinsiyete dayalı eşitsizlikle daha da desteklenmektedir.

Sonuçlar, mevcut ücretsiz AI TTS sistemlerinin, özellikle erkek sesleri kullanıldığında, tanınan sevinç ve üzüntüyü üretmede yalnızca kısmen başarılı olduğunu göstermektedir. Bu duyguların güçlü akustik tanımlayıcıları (üzüntü için daha düşük perde ve daha yavaş tempo, mutluluk için daha yüksek perde ve daha hızlı tempo) TTS sistemlerinin bu duyguları tanımlamakta ve taklit etmekte daha başarılı olmasının bir sebebi olabilir (James et al., 2023; Kolekar et al., 2024). Korku ve şaşkınlık gibi dinamik ve hızlı bürün gerektiren duyguları sentezlemedeki yetersizlik, duygusal tonal modellemenin eksikliklerini ortaya koymaktadır (Ao & Yuan, 2024; Liu et al., 2024). Duygu aktarımındaki düzensizlik, belirli bürünsel müdahalelerinin gerekliliğini vurgulamaktadır (Liu et al., 2024).

Cinsiyetler arasındaki eşitsizlik, erkek seslerinin daha küçük perde aralıkları nedeniyle modellenmesinin daha kolay olduğu anlamına gelmektedir. Cinsiyete duyarlı ses modellemesi bu nedenle TTS sistemlerinin hedefi olmalıdır. Duygusal bilişim bu gerçekleşmelerden faydalanabilir. Sanal asistanlarda, terapi botlarında ve eğitim-öğretim arayüzlerinde duygusal anlaşılırlık, kullanıcılar arasındaki etkileşimi, güvenilirliği ve duygusal tutarlılığı artıracaktır. (Ao & Yuan, 2024; Liu et al., 2024).

Önemli kısıtlamalar arasında katılımcı homojenliği, çevresel değişkenlik, ve TTS sistemlerinin değişkenliği yer almaktadır. Öncelikle örnekleme sadece Türk lisans öğrencileri dahil edildiğinden kültürler arası genellenebilirlik azalmıştır. İkinci olarak, uzaktan uygulanan test, dikkat dalgalanmalarını olası hale getirmektedir. Son olarak, yorumlama yalnızca iki ücretsiz sistem ile sınırlıdır; gelişmiş ticari veya özel tasarımlar farklı performanslar sergileyebilir.

Gelecekteki çalışmalarda farklı dil veya kültürel geçmişlere sahip ve demografik çeşitlilikteki katılımcıların katılımı artırılmalı ve fizyolojik ölçümler (EEG, GSR) ve kontrollü işitsel ortamlar kullanılarak laboratuvar tabanlı çalışmalar yürütülmelidir. Daha kapsamlı bir değerlendirme için daha geniş çeşitlilikte yenilikçi TTS sistemleri dahil edilebilir. Eşitsizlikleri azaltmak için cinsiyete uyarlanabilir modelleme teknikleri incelenmeli ve veri bütünlüğünü optimize etmek ve dinleyici yorgunluğunu azaltmak için etkileşimli ve uyarlanabilir testler kullanılabilir. Ayrıca, karmaşık duyguların tasvirini geliştirmek için gelecekteki araştırmalarda çok yönlü duygu ipuçları (görsel, anlamsal bağlam) kullanılabilir.

Bu çalışma, yapay zekanın ürettiği konuşmanın, mutluluk ve üzüntü gibi bazı duyguları erkek sesleriyle etkili bir şekilde ifade edebilmesine rağmen, şaşkınlık, korku ve öfke konusunda hâlâ sorun yaşadığını gösteriyor. Sonuçlar, TTS sistemlerini geliştirmek için duyguya özgü hazırlık, cinsiyet ayrımı gözetmeyen modelleme ve çok modlu tekniklerin gerekli olduğunu göstermektedir. Pearson korelasyonları gibi dinleyici merkezli değerlendirmelerin uygulanması, yapay zeka seslerinin algısal gerçekliğini değerlendirmek için güçlü bir çerçeve sağlayacaktır. Yapay zeka aracılı iletişimin duygusal açıdan etkileyici etkileşimler üretebilmesi için bu alanlarda daha fazla ilerleme kaydedilmesi gerekmektedir.

REFERENCES

Abukhodair, N., Song, M., Pekçetin, S., & DiPaola, S. (2023). Designing a wheel-based assessment tool to measure visual aesthetic emotions. *Cognitive Systems Research*.

Aitken, R. C. B. (1969). Measurement of feelings using visual analogue scales. *Proceedings of the Royal Society of Medicine, 62*(10), 989–993.

Ao, Y., & Yuan, M. (2024). *Analysis of Subjective Evaluation of Al Speech Synthesis Emotional Expressiveness*. 56–61. https://doi.org/10.1109/snpd61259.2024.10673914

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, *70*(3), 614.

Bänziger, T., Hosoya, G., & Scherer, K. R. (2015). Path models of vocal emotion communication. *PLOS ONE, 10(9)*, Article e0136675. https://doi.org/10.1371/journal .pone.0136675

Barakat, H., Turk, O., & Demiroglu, C. (2024). Deep learning-based expressive speech synthesis: A systematic review of approaches, challenges, and resources. *Eurasip Journal on Audio, Speech, and Music Processing*.

Beckman, M. E. (1996). The parsing of prosody. *Language and Cognitive Processes, 11*(1-2), 17–67. https://doi.org/10.1080/016909696387213

Bian, W., Zhou, Y., Zhang, K., & Gu, X. (2024). *EmoSpeech: A Corpus of Emotionally Rich and Contextually Detailed Speech Annotations*. https://doi.org/10.48550/arxiv.2412.06581

Brkić, G. (2024). Vocabulary building in EFL using Plutchik's Wheel of Emotions. *European Journal of English Language Teaching*.

Brookbush Institute. (n.d.). *Visual Analogue Scale (VAS)*. Retrieved April 4, 2025, from https://brookbushinstitute.com/glossary/visual-analogue-scale-vas

Chaudhury, R., Godbole, M., Garg, A., & Seo, J. H. (2024). Humane Speech Synthesis through Zero-Shot Emotion and Disfluency Generation. *arXiv.Org*, *abs/2404.01339*. https://doi.org/10.48550/arxiv.2404.01339

Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience*, *30*(1-2), 1-31.

Cutler, A.; Isard, S.D. 1980, Part of book or chapter of book (B. Butterworth (ed.), Language Production; vol. 1: Speech and talk, pp. 245-269)

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications.

Gift, A. G. (1989). Visual analogue scales: Measurement of subjective phenomena. *Nursing Research, 38*(5), 286–288. https://doi.org/10.1097/00006199-198909000-00014

Google Support. (n.d.). *How to ensure that information submitted in Google Forms is secure, encrypted, and private*. Retrieved from https://support.google.com/docs/thread/222150923

Google Support. (n.d.). *How to ensure that information submitted in Google Forms is secure, encrypted, and private*. Retrieved from https://support.google.com/docs/thread/222150923

Goudbeek, M., Goldman, J. P., & Scherer, K. R. (2009). Emotion dimensions and formant position. In *10th Annual Conference of the International Speech Communication Association (INTERSPEECH-2009)* (pp. 1575–1578). International Speech Communication Association.

Hans, A. (2021). Replication data for: Reddit posts related to Artificial Intelligence categorized using Plutchik's Wheel of Emotions. *Dataset*.

Huskisson, E. C. (1974). Measurement of pain. *The Lancet, 304*(7889), 1127–1131. https://doi.org/10.1016/S0140-6736(74)90884-8

IBM Corporation. (2021). *IBM SPSS Statistics for Windows, Version 27.0* [Computer software]. Armonk, NY: IBM Corp.

James, J., BalamuraliB, T., Watson, C., & Mixdorff, H. (2023). Exploring Prosodic Features Modelling for Secondary Emotions Needed for Empathetic Speech Synthesis. *Sensors*, *23*(6), 2999. https://doi.org/10.3390/s23062999

Khalifah, M. J., Ptaszyński, M., & Masui, F. (2024). Emotional Text-To-Speech in Japanese Using Artificially Augmented Dataset. *IEEE Access*, 1. https://doi.org/10.1109/access.2024.3495694

Kolekar, S. S., Richter, D. J., Bappi, M. I., & Kim, K. (2024). *Advancing AI Voice Synthesis: Integrating Emotional Expression in Multi-Speaker Voice Generation*. 193–198. https://doi.org/10.1109/icaiic60209.2024.10463204

Kolekar, S. S., Richter, D. J., Bappi, M. I., & Kim, K. (2024). Advancing AI voice synthesis: Integrating emotional expression in multi-speaker voice generation. In *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 193-198). IEEE.

Korotkova, Y., Kalinovskiy, I., & Vakhrusheva, T. (2024). *Word-level Text Markup for Prosody Control in Speech Synthesis*. 2280–2284. https://doi.org/10.21437/interspeech.2024-715

Liu, J., Liu, Z., Hu, Y., Gao, Y., Zhang, S., & Ling, Z. (2024). *DiffStyleTTS: Diffusion-based Hierarchical Prosody Modeling for Text-to-Speech with Diverse and Controllable Styles*. https://doi.org/10.48550/arxiv.2412.03388

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE, 13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Mondal, A., & Gokhale, S. S. (2020). Mining emotions on Plutchik's wheel. *Proceedings of SNAMS*.

Mondal, A., Bharadwaj, R., Mallela, J., Vuppala, A. K., & Yarra, C. (2024). *A Preliminary Analysis of Automatic Word and Syllable Prominence Detection in Non-Native Speech With Text-to-Speech Prosody Embeddings*. https://doi.org/10.48550/arxiv.2412.08283

Moser, A., & Korstjens, I. (2018). Series: Practical guidance to qualitative research. Part 3: Sampling, data collection and analysis. *European Journal of General Practice, 24*(1), 9–18. https://doi.org/10.1080/13814788.2017.1375091

Murf.ai. (2024, March 6). *Murf AI launches MultiNative, text-to-speech voices that can seamlessly switch between any language; completes brand revamp on 4-year anniversary*. PR Newswire. https://www.prnewswire.com/news-releases/murf-ai-launches-multinative-text-to-speech-voices-that-can-seamlessly-switch-between-any-language-completes-brand-revamp-on-4-year-anniversary-302310259.html

Murf.ai. (2024). *Free AI voice generator: Versatile text to speech software | Murf AI*. https://murf.ai/

Ngo, B., Rohmatillah, M., & Chien, J. (2024). *Learning Contrastive Emotional Nuances in Speech Synthesis*. 1–6. https://doi.org/10.1109/o-cocosda64382.2024.10800372

Nithin, S. K., & Prakash, J. (2022). Emotional Speech Synthesis using End-to-End neural TTS models. *International Computer Engineering Conference*, *1*, 1–7. https://doi.org/10.1109/ICENCO55801.2022.10032463

Play.ht. (2024). *Lovo vs Murf AI: Compare samples, price & features – PlayHT*. https://play.ht/blog/ai-apps/vs/lovo-vs-murf-ai/

Plutchik, R. (1980). Emotion: A psychoevolutionary synthesis. Harper & Row.

Raju, V., & Harinarayana, N. S. (2016). Online survey tools: A case study of Google Forms. Retrieved from https://www.researchgate.net/publication/326831738_Online_survey_tools_A_case_study_of_Google_Forms

Roshan, R., Roshan, R. K., Sohan, M. H., Raj, S., & Prasad, V. R. B. (2024). *Sentient Sound waves: Elevating Emotional Communication with AI-Generated Speech Technology*. https://doi.org/10.1109/incet61516.2024.10593200

Scherer, K. R. (2013). Emotion in action, interaction, music, and speech. In A. Arbib (Ed.), *Language, music, and the brain: A mysterious relationship* (pp. 107–139). MIT Press. https://doi.org/10.7551/mitpress/9780262018104.003.0005

Speer, S., & Blodgett, A. (2006). Prosody. In *Handbook of psycholinguistics* (pp. 505-537). Academic Press.

Suryani, A., Sari, R. N., & Fitriani, D. A. (2023). Students' perspective toward the implementation of Google Forms as a digital learning assessment tool. *English Education and Applied Linguistics Journal,* *7*(2). Retrieved from https://journal.institutpendidikan.ac.id/index.php/eeal/article/view/1624

Tang, H., Zhang, X., Cheng, N., Xiao, J., & Wang, J. (2024). *ED-TTS: Multi-Scale Emotion Modeling Using Cross-Domain Emotion Diarization for Emotional Speech Synthesis*. https://doi.org/10.1109/icassp48485.2024.10446467

Tsiamas, I., Sperber, M., Finch, A., & Garg, S. (2024). *Speech is More Than Words: Do Speech-to-Text Translation Systems Leverage Prosody?* https://doi.org/10.48550/arxiv.2410.24019

Wada, T., Hara, S., & Abe, M. (2024). *Explicit Prosody Control to Realize Discourse Focus in End-to-End Text-to-Speech*. 1–6. https://doi.org/10.1109/mlsp58920.2024.10734738

Wewers, M. E., & Lowe, N. K. (1990). A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in Nursing & Health, 13*(4), 227–236. https://doi.org/10.1002/nur.4770130405

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Wushouer, M., & Tuerhong, G. (2023). Semi-Supervised Learning for Robust Emotional Speech Synthesis with Limited Data. *Applied Sciences*, *13*(9), 5724. https://doi.org/10.3390/app13095724

APPENDICES

APPENDIX-1. Etik Kurul İzni

Evrak Tarih ve Sayısı: 17.07.2024-280096

T.C.
ANKARA HACI BAYRAM VELİ ÜNİVERSİTESİ
Etik Komisyonu

Sayı : E-11054618-302.08.01-280096
Konu : Bilimsel ve Eğitim Amaçlı

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ MÜDÜRLÜĞÜNE

İlgi : 07.06.2024 tarih ve E.271991 sayılı yazı.

İlgi yazınız ile göndermiş olduğunuz, Enstitünüz İngiliz Dili ve Edebiyatı Anabilim Dalı **Yüksek Lisans öğrencisi Eda ŞAHİN, Prof.Dr. Güven MENGÜ** danışmanlığında yürüttüğü **"Yazıyı Sese Çeviren Yapay Zeka Programları Tarafından Üretilen Duygusal Prozodinin Tanınabilirlik Bakımından Karşılaştırmalı Analizi (A Comparative Analysis of Auditory Recognition of Emotional Prosody Generated by Text-to-Speech Al Programs)"** adlı tez çalışması ile ilgili konu Komisyonumuzun 17.07.2024 tarih ve 08 sayılı toplantısında görüşülmüş olup,

Etik Komisyonunca onaylanan ilgilinin çalışmasının, yapılması planlanan yerlerden izin alınması koşuluyla yapılmasında etik açıdan bir sakınca bulunmadığına oybirliği ile karar verilmiş; karara ilişkin katılım listesi ve onaylanan çalışmalar ekte gönderilmiştir.

Bilgilerinizi ve gereğini rica ederim.

Araştırma Kod No: 2024/308

Prof. Dr. Kürşat GÖKTÜRK
Komisyon Başkanı

Ek:
1- Katılımcı Listesi
2- Onaylı Çalışma

Bu belge, güvenli elektronik imza ile imzalanmıştır.

ANKARA HACI BAYRAM VELİ ÜNİVERSİTESİ
ETİK KOMİSYONU KATILIM LİSTESİ

TOPLANTI TARİHİ :17.07.2024
TOPLANTI SAYISI : 08

ADI-SOYADI	İMZA
Prof.Dr. Kürşat GÖKTÜRK BAŞKAN	KATILDI
Prof.Dr. M.Fadıl YILDIRIM Başkan Yrd.	KATILDI
Prof.Dr. Neşe Yaşar ÇEĞİNDİR	KATILDI
Prof.Dr. Funda YURDAKUL	KATILDI
Prof.Dr. Serdar ÖZTÜRK	KATILDI
Prof.Dr. Özlem ALTUNÖZ	KATILDI
Prof.Dr. Yalçın ARSLANTÜRK	KATILDI
Prof.Dr. Emine ÖNER KAYA	KATILDI
Prof.Dr. Aydan ÖZSOY	KATILDI
Prof.Dr. Ayşe CANATAN	KATILDI

APPENDIX-2. Araştırma Anketi

T.C.
ANKARA HACI BAYRAM VELİ ÜNİVERSİTESİ

İNGİLİZ DİL BİLİMİ YÜKSEK LİSANS TEZİ ARAŞTIRMA ANKETİ

Aşağıdaki görsel analog ölçeklerini dinlediğiniz ses kayıtlarına göre işaretleyiniz.

(0: en az, 10: en çok)

Circle the visual analog scale under every question according to the recordings you hear. (0: least relevant, 10: most relevant)

1. Dinlediğiniz ses kaydı sizin için anlaşılır mı? / *Is the recording you are listening clear to you? (0: Not clear at all, 10: Very clear)*

0 1 2 3 4 5 6 7 8 9 10

2. Dinlediğiniz ses kaydı aşağıdaki duygulardan hangisini yansıtıyor? / *Which emotion does the recording reflect?*

A) Mutluluk / *Happiness* B) Üzüntü / *Sadness* C) Kızgınlık / *Anger*

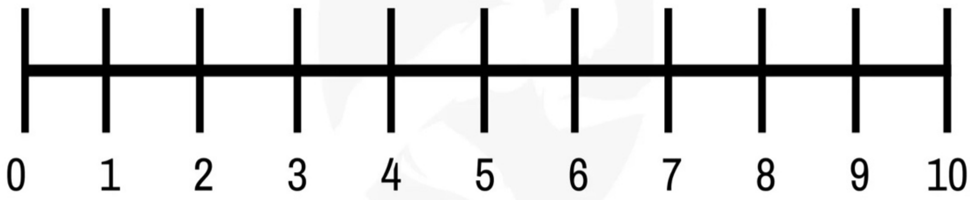
D) Korku / *Fear* E) Şaşkınlık / *Surprise*

3. Dinlediğiniz ses kaydında duygu ne kadar doğal yansıtılıyor? (0: Hiç doğal değil, 10: Çok doğal) / *Evaluate the recording in terms of how authentically it conveys an emotion. (0: Not natural at all, 10: Very natural)*

0 1 2 3 4 5 6 7 8 9 10

ETİK KOMİSYONU ONAYI

APPENDIX-3. Visual Analog Scale



APPENDIX-4. Pearson Correlation Tables

Correlations

		Emotion1	Emotion47
Emotion1	Pearson Correlation	1	,058
	Sig. (2-tailed)		,613
	N	77	77
Emotion47	Pearson Correlation	,058	1
	Sig. (2-tailed)	,613	
	N	77	77

Correlations

		Emotion2	Emotion56
Emotion2	Pearson Correlation	1	-,100
	Sig. (2-tailed)		,388
	N	77	77
Emotion56	Pearson Correlation	-,100	1
	Sig. (2-tailed)	,388	
	N	77	77

Correlations

		Emotion3	Emotion87
Emotion3	Pearson Correlation	1	,076
	Sig. (2-tailed)		,510
	N	77	77
Emotion87	Pearson Correlation	,076	1
	Sig. (2-tailed)	,510	
	N	77	77

Correlations

		Emotion4	Emotion15
Emotion4	Pearson Correlation	1	-,025
	Sig. (2-tailed)		,832
	N	77	77
Emotion15	Pearson Correlation	-,025	1
	Sig. (2-tailed)	,832	
	N	77	77

Correlations

		Emotion5	Emotion45
Emotion5	Pearson Correlation	1	,036
	Sig. (2-tailed)		,753
	N	77	77
Emotion45	Pearson Correlation	,036	1
	Sig. (2-tailed)	,753	
	N	77	77

Correlations

		Emotion7	Emotion32
Emotion7	Pearson Correlation	1	,013
	Sig. (2-tailed)		,911
	N	77	77
Emotion32	Pearson Correlation	,013	1
	Sig. (2-tailed)	,911	
	N	77	77

Correlations

		Emotion8	Emotion55
Emotion8	Pearson Correlation	1	-,068
	Sig. (2-tailed)		,554
	N	77	77
Emotion55	Pearson Correlation	-,068	1
	Sig. (2-tailed)	,554	
	N	77	77

Correlations

		Emotion9	Emotion22
Emotion9	Pearson Correlation	1	,080
	Sig. (2-tailed)		,492
	N	77	77
Emotion22	Pearson Correlation	,080	1
	Sig. (2-tailed)	,492	
	N	77	77

Correlations

		Emotion10	Emotion48
Emotion10	Pearson Correlation	1	-,100
	Sig. (2-tailed)		,387
	N	77	77
Emotion48	Pearson Correlation	-,100	1
	Sig. (2-tailed)	,387	
	N	77	77

Correlations

		Emotion11	Emotion12
Emotion11	Pearson Correlation	1	,124
	Sig. (2-tailed)		,281
	N	77	77
Emotion12	Pearson Correlation	,124	1
	Sig. (2-tailed)	,281	
	N	77	77

Correlations

		Emotion27	Emotion11
Emotion27	Pearson Correlation	1	,280*
	Sig. (2-tailed)		,014
	N	77	77
Emotion11	Pearson Correlation	,280*	1
	Sig. (2-tailed)	,014	
	N	77	77

*. Correlation is significant at the 0.05 level (2-tailed).

Correlations

		Emotion13	Emotion43
Emotion13	Pearson Correlation	1	,057
	Sig. (2-tailed)		,621
	N	77	77
Emotion43	Pearson Correlation	,057	1
	Sig. (2-tailed)	,621	
	N	77	77

Correlations

		Emotion14	Emotion50
Emotion14	Pearson Correlation	1	-,134
	Sig. (2-tailed)		,245
	N	77	77
Emotion50	Pearson Correlation	-,134	1
	Sig. (2-tailed)	,245	
	N	77	77

Correlations

		Emotion16	Emotion23
Emotion16	Pearson Correlation	1	-,031
	Sig. (2-tailed)		,787
	N	77	77
Emotion23	Pearson Correlation	-,031	1
	Sig. (2-tailed)	,787	
	N	77	77

Correlations

		Emotion17	Emotion19
Emotion17	Pearson Correlation	1	,095
	Sig. (2-tailed)		,410
	N	77	77
Emotion19	Pearson Correlation	,095	1
	Sig. (2-tailed)	,410	
	N	77	77

Correlations

		Emotion18	Emotion42
Emotion18	Pearson Correlation	1	,138
	Sig. (2-tailed)		,232
	N	77	77
Emotion42	Pearson Correlation	,138	1
	Sig. (2-tailed)	,232	
	N	77	77

Correlations

		Emotion19	Emotion46
Emotion19	Pearson Correlation	1	-,010
	Sig. (2-tailed)		,933
	N	77	77
Emotion46	Pearson Correlation	-,010	1
	Sig. (2-tailed)	,933	
	N	77	77

Correlations

		Emotion20	Emotion26
Emotion20	Pearson Correlation	1	,225*
	Sig. (2-tailed)		,049
	N	77	77
Emotion26	Pearson Correlation	,225*	1
	Sig. (2-tailed)	,049	
	N	77	77

*. Correlation is significant at the 0.05 level (2-tailed).

Correlations

		Emotion21	Emotion36
Emotion21	Pearson Correlation	1	,045
	Sig. (2-tailed)		,696
	N	77	77
Emotion36	Pearson Correlation	,045	1
	Sig. (2-tailed)	,696	
	N	77	77

Correlations

		Emotion22	Emotion60
Emotion22	Pearson Correlation	1	,256*
	Sig. (2-tailed)		,025
	N	77	77
Emotion60	Pearson Correlation	,256*	1
	Sig. (2-tailed)	,025	
	N	77	77

*. Correlation is significant at the 0.05 level (2-tailed).

Correlations

		Emotion23	Emotion40
Emotion23	Pearson Correlation	1	-,111
	Sig. (2-tailed)		,335
	N	77	77
Emotion40	Pearson Correlation	-,111	1
	Sig. (2-tailed)	,335	
	N	77	77

Correlations

		Emotion24	Emotion51
Emotion24	Pearson Correlation	1	-,008
	Sig. (2-tailed)		,942
	N	77	77
Emotion51	Pearson Correlation	-,008	1
	Sig. (2-tailed)	,942	
	N	77	77

Correlations

		Emotion25	Emotion93
Emotion25	Pearson Correlation	1	,341**
	Sig. (2-tailed)		,002
	N	77	77
Emotion93	Pearson Correlation	,341**	1
	Sig. (2-tailed)	,002	
	N	77	77

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

		Emotion28	Emotion64
Emotion28	Pearson Correlation	1	,067
	Sig. (2-tailed)		,564
	N	77	77
Emotion64	Pearson Correlation	,067	1
	Sig. (2-tailed)	,564	
	N	77	77

Correlations

		Emotion29	Emotion65
Emotion29	Pearson Correlation	1	,028
	Sig. (2-tailed)		,811
	N	77	77
Emotion65	Pearson Correlation	,028	1
	Sig. (2-tailed)	,811	
	N	77	77

Correlations

		Emotion30	Emotion91
Emotion30	Pearson Correlation	1	-,014
	Sig. (2-tailed)		,906
	N	77	77
Emotion91	Pearson Correlation	-,014	1
	Sig. (2-tailed)	,906	
	N	77	77

Correlations

		Emotion31	Emotion9
Emotion31	Pearson Correlation	1	,067
	Sig. (2-tailed)		,562
	N	77	77
Emotion9	Pearson Correlation	,067	1
	Sig. (2-tailed)	,562	
	N	77	77

Correlations

		Emotion33	Emotion55
Emotion33	Pearson Correlation	1	,356**
	Sig. (2-tailed)		,001
	N	77	77
Emotion55	Pearson Correlation	,356**	1
	Sig. (2-tailed)	,001	
	N	77	77

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

		Emotion34	Emotion4
Emotion34	Pearson Correlation	1	-,128
	Sig. (2-tailed)		,268
	N	77	77
Emotion4	Pearson Correlation	-,128	1
	Sig. (2-tailed)	,268	
	N	77	77

Correlations

		Emotion35	Emotion70
Emotion35	Pearson Correlation	1	-,017
	Sig. (2-tailed)		,886
	N	77	77
Emotion70	Pearson Correlation	-,017	1
	Sig. (2-tailed)	,886	
	N	77	77

Correlations

		Emotion37	Emotion75
Emotion37	Pearson Correlation	1	-,170
	Sig. (2-tailed)		,140
	N	77	77
Emotion75	Pearson Correlation	-,170	1
	Sig. (2-tailed)	,140	
	N	77	77

Correlations

		Emotion38	Emotion79
Emotion38	Pearson Correlation	1	,059
	Sig. (2-tailed)		,608
	N	77	77
Emotion79	Pearson Correlation	,059	1
	Sig. (2-tailed)	,608	
	N	77	77

Correlations

		Emotion39	Emotion86
Emotion39	Pearson Correlation	1	,247*
	Sig. (2-tailed)		,030
	N	77	77
Emotion86	Pearson Correlation	,247*	1
	Sig. (2-tailed)	,030	
	N	77	77

*. Correlation is significant at the 0.05 level (2-tailed).

Correlations

		Emotion43	Emotion99
Emotion43	Pearson Correlation	1	,302**
	Sig. (2-tailed)		,008
	N	77	77
Emotion99	Pearson Correlation	,302**	1
	Sig. (2-tailed)	,008	
	N	77	77

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

		Emotion44	Emotion77
Emotion44	Pearson Correlation	1	,036
	Sig. (2-tailed)		,756
	N	77	77
Emotion77	Pearson Correlation	,036	1
	Sig. (2-tailed)	,756	
	N	77	77

Correlations

		Emotion45	Emotion95
Emotion45	Pearson Correlation	1	-,096
	Sig. (2-tailed)		,407
	N	77	77
Emotion95	Pearson Correlation	-,096	1
	Sig. (2-tailed)	,407	
	N	77	77

Correlations

		Emotion48	Emotion76
Emotion48	Pearson Correlation	1	-,014
	Sig. (2-tailed)		,907
	N	77	77
Emotion76	Pearson Correlation	-,014	1
	Sig. (2-tailed)	,907	
	N	77	77

Correlations

		Emotion49	Emotion94
Emotion49	Pearson Correlation	1	,046
	Sig. (2-tailed)		,692
	N	77	77
Emotion94	Pearson Correlation	,046	1
	Sig. (2-tailed)	,692	
	N	77	77

Correlations

		Emotion52	Emotion72
Emotion52	Pearson Correlation	1	,190
	Sig. (2-tailed)		,098
	N	77	77
Emotion72	Pearson Correlation	,190	1
	Sig. (2-tailed)	,098	
	N	77	77

Correlations

		Emotion53	Emotion62
Emotion53	Pearson Correlation	1	-,031
	Sig. (2-tailed)		,790
	N	77	77
Emotion62	Pearson Correlation	-,031	1
	Sig. (2-tailed)	,790	
	N	77	77

Correlations

		Emotion54	Emotion14
Emotion54	Pearson Correlation	1	-,094
	Sig. (2-tailed)		,417
	N	77	77
Emotion14	Pearson Correlation	-,094	1
	Sig. (2-tailed)	,417	
	N	77	77

Correlations

		Emotion57	Emotion59
Emotion57	Pearson Correlation	1	,152
	Sig. (2-tailed)		,186
	N	77	77
Emotion59	Pearson Correlation	,152	1
	Sig. (2-tailed)	,186	
	N	77	77

Correlations

		Emotion58	Emotion65
Emotion58	Pearson Correlation	1	-,118
	Sig. (2-tailed)		,308
	N	77	77
Emotion65	Pearson Correlation	-,118	1
	Sig. (2-tailed)	,308	
	N	77	77

Correlations

		Emotion59	Emotion71
Emotion59	Pearson Correlation	1	,164
	Sig. (2-tailed)		,154
	N	77	77
Emotion71	Pearson Correlation	,164	1
	Sig. (2-tailed)	,154	
	N	77	77

Correlations

		Emotion60	Emotion81
Emotion60	Pearson Correlation	1	,198
	Sig. (2-tailed)		,085
	N	77	77
Emotion81	Pearson Correlation	,198	1
	Sig. (2-tailed)	,085	
	N	77	77

Correlations

		Emotion61	Emotion20
Emotion61	Pearson Correlation	1	,223
	Sig. (2-tailed)		,052
	N	77	77
Emotion20	Pearson Correlation	,223	1
	Sig. (2-tailed)	,052	
	N	77	77

Correlations

		Emotion63	Emotion74
Emotion63	Pearson Correlation	1	,085
	Sig. (2-tailed)		,461
	N	77	77
Emotion74	Pearson Correlation	,085	1
	Sig. (2-tailed)	,461	
	N	77	77

Correlations

		Emotion65	Emotion69
Emotion65	Pearson Correlation	1	,113
	Sig. (2-tailed)		,329
	N	77	77
Emotion69	Pearson Correlation	,113	1
	Sig. (2-tailed)	,329	
	N	77	77

Correlations

		Emotion66	Emotion74
Emotion66	Pearson Correlation	1	-,013
	Sig. (2-tailed)		,914
	N	77	77
Emotion74	Pearson Correlation	-,013	1
	Sig. (2-tailed)	,914	
	N	77	77

Correlations

		Emotion67	Emotion86
Emotion67	Pearson Correlation	1	,175
	Sig. (2-tailed)		,129
	N	77	77
Emotion86	Pearson Correlation	,175	1
	Sig. (2-tailed)	,129	
	N	77	77

Correlations

		Emotion68	Emotion82
Emotion68	Pearson Correlation	1	,274*
	Sig. (2-tailed)		,016
	N	77	77
Emotion82	Pearson Correlation	,274*	1
	Sig. (2-tailed)	,016	
	N	77	77

*. Correlation is significant at the 0.05 level (2-tailed).

Correlations

		Emotion69	Emotion84
Emotion69	Pearson Correlation	1	-,216
	Sig. (2-tailed)		,060
	N	77	77
Emotion84	Pearson Correlation	-,216	1
	Sig. (2-tailed)	,060	
	N	77	77

Correlations

		Emotion73	Emotion78
Emotion73	Pearson Correlation	1	,264*
	Sig. (2-tailed)		,020
	N	77	77
Emotion78	Pearson Correlation	,264*	1
	Sig. (2-tailed)	,020	
	N	77	77

*. Correlation is significant at the 0.05 level (2-tailed).

Correlations

		Emotion80	Emotion6
Emotion80	Pearson Correlation	1	,072
	Sig. (2-tailed)		,533
	N	77	77
Emotion6	Pearson Correlation	,072	1
	Sig. (2-tailed)	,533	
	N	77	77

Correlations

		Emotion83	Emotion21
Emotion83	Pearson Correlation	1	,064
	Sig. (2-tailed)		,577
	N	77	77
Emotion21	Pearson Correlation	,064	1
	Sig. (2-tailed)	,577	
	N	77	77

Correlations

		Emotion85	Emotion53
Emotion85	Pearson Correlation	1	,133
	Sig. (2-tailed)		,248
	N	77	77
Emotion53	Pearson Correlation	,133	1
	Sig. (2-tailed)	,248	
	N	77	77

Correlations

		Emotion88	Emotion96
Emotion88	Pearson Correlation	1	,033
	Sig. (2-tailed)		,779
	N	77	77
Emotion96	Pearson Correlation	,033	1
	Sig. (2-tailed)	,779	
	N	77	77

Correlations

		Emotion89	Emotion90
Emotion89	Pearson Correlation	1	,156
	Sig. (2-tailed)		,177
	N	77	77
Emotion90	Pearson Correlation	,156	1
	Sig. (2-tailed)	,177	
	N	77	77

Correlations

		Emotion92	Emotion100
Emotion92	Pearson Correlation	1	,100
	Sig. (2-tailed)		,387
	N	77	77
Emotion100	Pearson Correlation	,100	1
	Sig. (2-tailed)	,387	
	N	77	77

Correlations

		Emotion97	Emotion1
Emotion97	Pearson Correlation	1	-,019
	Sig. (2-tailed)		,873
	N	77	77
Emotion1	Pearson Correlation	-,019	1
	Sig. (2-tailed)	,873	
	N	77	77

Correlations

		Emotion98	Emotion25
Emotion98	Pearson Correlation	1	,081
	Sig. (2-tailed)		,484
	N	77	77
Emotion25	Pearson Correlation	,081	1
	Sig. (2-tailed)	,484	
	N	77	77

CURRICULUM VITAE

Personal Information

Surname, Name: ŞAHİN, Eda

Education

Bachelor's Degree	English Language Teaching	TED University	2021

İş Deneyimi

2022	DFF Dil Akademisi	English Language Teacher
2023-2024	Anka Bilim Koleji	English Language Teacher
2024-2025	Doğa Okulları	English Language Teacher

ankara HBV