Emotion Perception in Synthetic Speech: The Influence of Pitch and Semantics

Yogie Permana
yogie@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Sabika Amalina
sabika@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Pin Ju Huang
pjhuang@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Oskar Lövgren
oskarlov@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Wei Kang Wong
wkwo@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

***Abstract* — This study examines the interaction between prosody, specifically pitch, and semantic content in the perception of emotion in synthetic speech. Using a controlled within-subject experimental design, six conditions were tested by combining two pitch levels (high/low) with three semantic valence categories (positive, neutral, negative). Linear Mixed Model analysis revealed that low pitch significantly reduced emotional ratings in positive and neutral contexts, supporting hypotheses predicting ratings below neutrality. High pitch conditions, however, showed no significant deviations from neutral ratings, providing limited support for hypotheses predicting ratings above neutrality. These results highlight the stronger influence of pitch in diminishing emotional ratings and offer insights for improving emotional expressiveness in synthetic speech systems.**

Keywords — Emotion Perception, Synthetic Speech, Semantics, Prosody, Pitch

I. INTRODUCTION

With the rapid advancement of artificial intelligence and human-computer interaction, understanding how emotions are perceived in synthetic voices is more relevant than ever. As technology increasingly incorporates voice interfaces, from virtual assistants to social robots, the ability to convey and interpret emotions in speech plays a critical role in creating natural and engaging user experiences [1].

Emotional perception hinges on the interplay of prosodic features-like pitch, intonation, and rhythm—and semantic content, the emotional tone of the content (e.g. positive, neutral, or sad). For synthetic speech to effectively convey the intended emotional message, semantics and prosody must work in harmony.

When certain emotions are perceived in speech, the influence of prosody has been shown to be particularly important. Perceived sadness in speech especially seems to be dominated by prosodic features rather than semantic information. However, in the case of second-language English speakers, there also appears to exist a reliance mainly on prosody when perceiving happiness, yet not as much when perceiving anger [2]. Further, the relationship between pitch and emotion has particularly well-established associations of lower pitch with sadness and higher pitch with happiness. Previous studies [2, 3] have however been conducted using recordings of real human voices as stimuli, this delimits an area of interest for our study.

Our study aims to explore whether prosody, specifically pitch, can override semantic valence in synthetic speech. We hypothesize that higher pitch will consistently convey happiness, both when paired with neutral and negative semantic content, also lower pitch will signal sadness regardless of the semantic sentiment, neutral and positive. The outcomes of this research have significant implications for the development of more natural, emotionally expressive synthetic voices, particularly for applications in robotics and artificial intelligence, where effective communication is paramount.

II. BACKGROUND

Understanding how emotions are perceived in speech requires examining both prosodic and semantic aspects of communication. Much previous research has explored the influence of prosody and specifically of pitch in human voices and also made comparisons between the influence of voice versus the influence semantics. In the context of synthetic voices, accurately replicating these elements remains a significant challenge, as their interplay is crucial for producing natural and emotionally expressive communication.

A. *Prosodic Aspects*

Research shows that vocal features such as pitch, intonation, and tempo are key markers of emotion in human speech [4]. These features allow listeners to perceive the emotional state of a speaker, whether they are happy, sad, angry, or calm. The So-To-Speak [5] interface allows for a general specification of the pitch of the generated voice. Pitch is the perceptual property that corresponds to the physical property of fundamental frequency [6]. The average F0 (fundamental frequency) of speech is therefore a very relevant acoustical property to explore the implications of while conducting research with So-To-Speak.

The relationship between emotion and pitch in speech and the impact of average F0 on the perceived emotional valence and arousal of a voice has been subject to much research and discussion. Regarding the expression of happiness and sadness, several studies show that happiness (but also other types of high-arousal speech, such as anger) is commonly associated with a higher average F0, while a sad voice often is associated with a lower average F0 [7, 8, 9, 10, 11].

B. *Semantic Aspects*

The impact of pitch on emotional interpretation becomes more complex when combined with the semantic content of speech. Sentiment, the emotional tone conveyed by the meaning of the words, can be categorized into positive, neutral, or negative (sad) sentiments. A study by Kikutani and Ikemoto [2] touches on the comparison between prosody and semantic information and their roles in conveying emotion in speech. Expanding on a similar study by Ben-David et al.[3] which results showed that prosodic information is prioritized over semantic content when the two are incongruent, Kikutani and Ikemoto investigated the relative dominance of prosody versus semantic content in both L1, L2 and unfamiliar languages. Japanese participants were tasked with assessing whether voice recordings conveyed happiness, sadness, or anger, when the phrases “I'm pleased,” “I'm sad,” and “I'm angry” were articulated with emotional tones intentionally varied in their native language (Japanese), their second language (English) and in unfamiliar languages (Khmer and Swedish). Especially when expressing sadness, the impact of voice was dominant in comparison to semantics, regardless of language. Semantic dominance was found for both happiness and anger in the native language. In the second language however, voice dominance was found only for happiness and sadness, not anger.

C. *Emotion in Synthetic Voices*

One of the major challenges in developing emotionally expressive synthetic voices is accurately translating these vocal features [8]. Unlike humans who can instinctively adjust their vocal tone to match their emotions, synthetic voices rely on programmed algorithms to do so. As a result, we can have limited or rigid emotional expressions. If these cues are poorly implemented, they can lead to the “uncanny valley” effect, where synthetic emotions seem unnatural or unsettling [12].

D. *Aims and Delimitations*

In this project, we aim to explore how pitch and semantic content affect how people perceive emotions in synthetic speech. We will focus on understanding how pitch and semantics can influence emotional perception by testing different pitch levels (high and low) with different semantic contents (positive, neutral, and negative phrases). We will not study other factors like the relationship between the speaker and listener, the impact of language familiarity, or other voice features like rhythm or speech rate. By keeping the focus narrow, we can deeply analyze the role of pitch and semantic tone in shaping emotional understanding.

III. METHOD

This study investigates the perception of emotion in synthesized speech by examining how pitch (high/low) and semantic content (positive, neutral, negative) influence listeners' emotion classifications. The experiment uses a within-subject design, where each participant is exposed to six conditions: combinations of high/low pitch and positive/neutral/negative sentiment. The speech samples are generated using the So-To-Speak interface, which allows precise manipulation of pitch and the input of custom sentences.

A. *Experimental Design*

To minimize participant fatigue and potential pattern recognition, each participant is exposed to 30 randomized trials, with 5 stimuli per condition. The stimuli are drawn from a predetermined pool of sentences for each sentiment category, ensuring balanced representation across participants.

B. *Survey*

The survey section of our research aimed to evaluate the emotions experienced when participants listening to a specific sound. They rated their emotional response on a Likert scale ranging from 1 to 7, where 1 represented "Sad," 4 represented "Neutral," and 7 represented "Happy." This scale allowed us to quantify the general emotional tone participants associated with the sound, offering a structured way to capture the intensity of their emotional responses.

The data collection process involved test participants answering questions via a Google Form. Since it was not possible to directly embed the generated voice files into the form, the audio was uploaded to YouTube, and the resulting video was embedded into the Google Form. This workaround ensured that participants could listen to the generated audio and respond to all questions within a single survey form.

C. *Participants*

The study involved 29 participants, selected to ensure sufficient data for statistical analysis.

D. *So-to-Speak*

In this study, we utilize So-to-Speak [5], an interactive tool designed to showcase the capabilities of controllable Text-to-Speech (TTS) systems. So-to-Speak provides a flexible platform to generate, synthesize, and explore hundreds of synthetic speech samples simultaneously, offering control over both low-level prosodic features (e.g., pitch, speech rate) and higher-level speaking styles (e.g., emotional or conversational speech). This tool displays samples on an interactive grid, where prosody and style controls are easily adjusted, allowing users to explore the relationships between these features and their impact on speech perception.

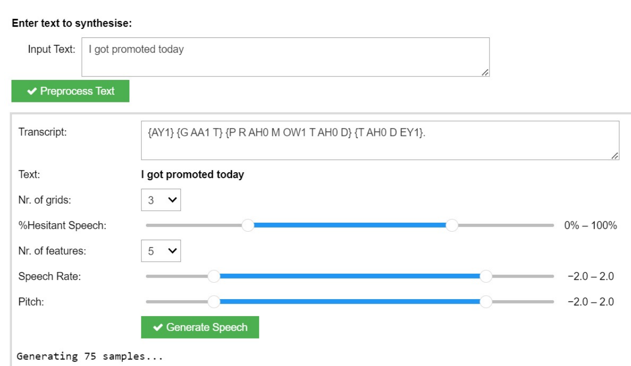

Figure 1: So-to-Speak Interface

An automatic estimation of Mean Opinion Scores (MOS) is included for each sample, providing a measure of synthesis quality. This feature helps users evaluate variations in speech quality as they manipulate prosodic and stylistic features, making it especially useful for research on emotion perception in synthetic speech. In our study, we use So-to-Speak to generate speech samples with varying emotional tones and analyze how pitch and semantic content affect emotion perception. The flexibility of the platform and the visual interface make it an ideal tool for exploring how different features of speech influence emotional expression in synthetic speech.

E. *Stimuli*

The So-to-Speak interface was used to generate synthetic speech stimuli with precise control over key prosodic features, especially pitch, as well as the semantic content of the sentences. The pitch of the speech samples was manipulated across two distinct levels: high and low. These pitch variations were selected based on their common associations with emotional expression in speech—a high pitch is often linked to emotions like happiness and excitement, while a low pitch is typically associated with emotions like sadness [3, 7].

It was important to balance pitch levels to maintain the naturalness of the synthetic speech. In order to ensure the generated voices did not sound artificial or strained, we fine-tuned the pitch adjustments within a range that was perceptually distinct but still realistic. Through an iterative process, we adjusted the pitch until we achieved an optimal balance between emotional expressiveness and naturalness.

The semantic content of the stimuli varied across three valence categories: positive, neutral, and negative. Each sentence was carefully constructed to ensure the emotional meaning was clear and consistent while maintaining similar sentence length and structure across the three categories. To create the six experimental conditions, we combined high pitch and low pitch with the three semantic valence categories.

Semantic	Pitch: High	Pitch: Low
Positive	I got promoted today. We had a huge success today. What a beautiful day! My son has finally graduated high school! She won the lottery yesterday.	I got promoted today. We had a huge success today. What a beautiful day! My son has finally graduated high school! She won the lottery yesterday.
Neutral	I'll be here tomorrow. We did our job. It is snowing The meeting starts at noon He opened the window	I'll be here tomorrow. We did our job. It is snowing The meeting starts at noon He opened the window
Negative	I am fired. I lost my job. What a total failure. I'm having a bad day! Her cousin just dropped out of university. He lost a huge sum of money.	I am fired. I lost my job. What a total failure. I'm having a bad day! Her cousin just dropped out of university. He lost a huge sum of money.

Figure 2: Matrix of Condition

Five different sentences were generated for each of these six conditions, each with slightly varied wording or phrasing, but maintaining the same emotional tone and semantic category. This variability ensured that the results were not biased by any single sentence. As a result, the total number of stimuli in the experiment was 30 samples (6 conditions × 5 sentences per condition), like presented in the following matrix of combinations in **Figure 2**.

F. *Statistical Analysis*

A linear mixed model (LMM) was employed to analyze the data, accounting for the repeated measures structure of the study, where each participant (i) provided ratings (x_{ij}) across multiple conditions (j). The LMM was specified to include fixed effects for pitch (high/low), sentiment (positive, neutral, negative), and their interaction, alongside random intercepts to account for individual variability. The model is expressed as:

$$x_{ij} = \beta_0+\beta_1\text{Pitch}_j+\beta_2\text{Sentiment}_j+\beta_3(\text{Pitch}_j\times\text{Sentiment}_j)+u_i+\epsilon_{ij}$$

where:

- x_{ij} is the rating provided by participant i for condition j,
- β_0 is the fixed intercept, representing the overall mean rating across all participants and conditions,
- $\beta_1, \beta_2, \beta_3$ are the fixed effects for pitch, sentiment, and their interaction, respectively,
- Pitch_j and Sentiment_j are categorical variables encoding the pitch (high/low) and sentiment (positive/neutral/negative) of condition j,
- $u_i \sim \mathcal{N}(0, \sigma_u^2)$ is the random intercept for participant i, modeling individual variability, and
- $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is the residual error for each observation.

The model assumes that the random effects (u_i) and residuals (ϵ_{ij}) are independent and normally distributed.

G. *Model Evaluation*

The interaction term (β_3) in the model was included to assess whether the effect of pitch on ratings varied across sentiment categories. Significance tests for fixed effects were performed using z-statistics, with the null hypothesis (H_0) stating that the corresponding fixed effect coefficient (β) equals zero.

Additionally, pairwise comparisons using Tukey's Honest Significant Difference (HSD) test were conducted as a post-hoc analysis to identify specific differences between conditions (e.g., high pitch and neutral sentiment vs. low pitch and neutral sentiment). These comparisons allowed for more granular insight into the effects of pitch, sentiment, and their interaction.

H. *Hypotheses Testing*

To evaluate whether ratings for specific conditions were significantly greater or less than neutrality ($x = 4$), a series of pairwise post-hoc comparisons were conducted between conditions. Here, x represents ratings on a Likert scale ranging from 1 (Sad) to 7 (Happy), with 4 indicating Neutrality. The following hypotheses were tested:

H_1 : High pitch with negative semantic context conveys happiness, i.e., $\mu_{\text{High, Negative}} > 4$,

H_2 : Low pitch with positive semantic context conveys sadness, i.e., $\mu_{\text{Low, Positive}} < 4$,

H_3 : High pitch with neutral semantic context conveys happiness, i.e., $\mu_{\text{High, Neutral}} > 4$,

H_4 : Low pitch with neutral semantic context conveys sadness, i.e., $\mu_{\text{Low, Neutral}} < 4$.

Baseline conditions were included to validate the experimental design:

B_1 : High pitch with positive semantic context conveys happiness, i.e., $\mu_{\text{High, Positive}} > 4$,

B_2 : Low pitch with negative semantic context conveys sadness, i.e., $\mu_{\text{Low, Negative}} < 4$.

I. *Statistical Significance*

Significance was assessed at $\alpha = 0.05$. For the post-hoc pairwise comparisons, adjusted p-values using Tukey's HSD were reported. Conditions with adjusted p-values less than 0.05 were considered significantly different. The results of these analyses allowed us to assess both the main effects and interaction effects of pitch and sentiment on emotional ratings.

By including both the fixed effects of pitch and sentiment and their interaction, alongside participant-level random effects, this model provided robust estimates of the variability attributable to each factor while accounting for individual differences.

IV. RESULTS

Table 1: Results of the Linear Mixed Model (LMM) Analysis

Condition	$\bar{x}$ (Mean Rating)	p-Value
High pitch, negative sentiment (H_1)	3.98	0.558
Low pitch, positive sentiment (H_2)	3.17	< 0.001
High pitch, neutral sentiment (H_3)	4.02	0.402
Low pitch, neutral sentiment (H_4)	3.54	< 0.001
High pitch, positive sentiment (B_1)	4.96	< 0.001
Low pitch, negative sentiment (B_2)	2.70	< 0.001

The results of the Linear Mixed Model (LMM) analysis revealed distinct effects of pitch, sentiment, and their interaction on perceived emotion ratings ($p < 0.001$). Specifically: - Low pitch paired with positive sentiment (H_2) and neutral sentiment (H_4) produced significantly lower ratings compared to neutral ($x = 4$), supporting the hypotheses for sadness perception. - High pitch paired with negative sentiment (H_1) and neutral sentiment (H_3) failed to produce significant deviations from neutrality, suggesting that high pitch alone does not override semantic valence to convey happiness.

Baseline conditions further validated the experimental design. B_1 (High pitch with positive sentiment) produced significantly higher ratings compared to neutral ($x = 4$, $p < 0.001$), confirming its association with happiness. Conversely, B_2 (Low pitch with negative sentiment) yielded significantly lower ratings ($p < 0.001$), consistent with its association with sadness.

Post-hoc pairwise comparisons provided additional insights into the interaction effects. Conditions involving low pitch consistently demonstrated significant differences when compared to high pitch counterparts, particularly in neutral and positive sentiment contexts ($p < 0.05$). B_1 and B_2 also significantly differed from most other conditions, reinforcing the validity of the model.

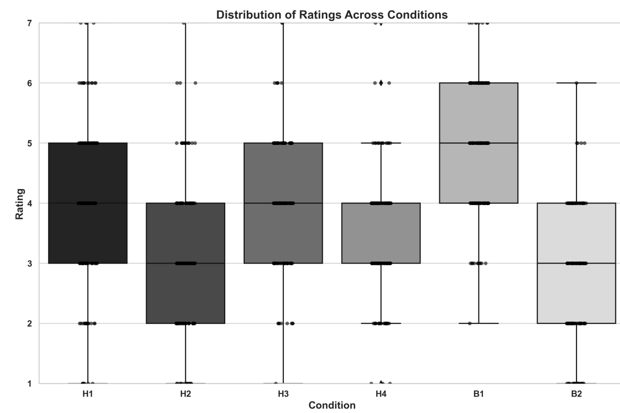

Figure 3: Distribution of Ratings Across Conditions. This figure illustrates the mean ratings for each condition, highlighting the significant differences observed in post-hoc comparisons.

These findings confirm that low pitch strongly influences sadness perception across varying semantic contexts, while high pitch's ability to convey happiness depends on alignment with positive semantic valence. The baseline conditions B_1 and B_2 provide validation for the experimental design by aligning with expected emotional ratings.".

V. DISCUSSION

A. *Strengths and Limitations of the Study*

This study uses synthetic speech to precisely manipulate pitch and semantics, ensuring strict control over the independent variables. Unlike human-generated speech, this approach eliminates variability and provides a more consistent experimental setup.

However, there are limitations. Synthetic speech lacks natural characteristics, such as nuanced intonation patterns and high-quality prosody, which could reduce the ecological validity of the results. Consequently, the findings may not fully generalize to human speech perception. Additionally, the study focuses on a narrow range of emotions (happiness and sadness) and does not consider a broader spectrum of emotions, such as the six basic emotions identified by Paul Ekman [13] or James Russell's Circumplex Model of Affect [14]. Omitting valence and arousal as variables further limits the understanding of high-pitch emotional perception.

B. *Implications and General Discussion*

The question, "Does pitch override semantic valence in synthetic speech?" yields mixed results. Low pitch appears to dominate over semantic valence, consistently resulting in sadness ratings, even when paired with positive or neutral semantics. High pitch, however, does not consistently override semantics. For instance, high pitch paired with negative semantics (H_1) resulted in ratings close to neutrality, suggesting that neither pitch nor semantics dominated in this scenario.

The findings raise the question of whether there is such a thing as a universally "happy voice." Our results suggest that a low-pitch voice reliably conveys sadness across all semantic contexts, potentially characterizing it as a "sad voice." However, the reverse is not true for high pitch. A higher pitch does not reliably convey happiness, aligning with previous research that associates high pitch with high-arousal emotions such as fear, anger, and happiness [7, 8, 9, 10, 11]. This ambiguity was reflected in the survey responses, where participants frequently reported secondary emotions like ambiguity or sarcasm for high-pitch conditions.

Prior research also indicates that intonation may have a greater influence on joy perception than pitch alone. For example, a lower-pitch voice with significant tonal variation may be perceived as more joyful than a high-pitch voice with limited variation [15]. The synthetic speech used in this study may have constrained tonal variability, particularly for high-pitch samples, thereby limiting the perception of happiness.

VI. CONCLUSION

A. *Key Findings*

This study demonstrates that low pitch consistently dominates over semantics, leading to the perception of sadness, even in positive or neutral contexts. High pitch, on the other hand, does not dominate semantics. Only high pitch paired with positive semantics (B_1) reliably conveys happiness. These findings reveal an asymmetry in how pitch influences emotional perception, particularly between sadness and happiness.

B. *Future Work*

Future studies should explore additional speech production parameters, such as speech rate, hesitation, and intonation, to better understand their role in shaping emotional perception. Investigating emotional valence and arousal through models like Russell's Circumplex Model [14] could also provide deeper insights into high-pitch emotional perception.

Expanding the range of emotions beyond happiness and sadness to include emotions such as anger, fear, and surprise would provide a more comprehensive understanding. Additionally, incorporating human speech alongside synthetic voices could enhance ecological validity and allow for meaningful comparisons between human and synthetic emotional expressiveness.

References

[1] Suryakant Tyagi and Sándor Szénási. "Semantic speech analysis using machine learning and deep learning techniques: a comprehensive review". In: *Multimedia Tools and Applications* 83.29 (Sept. 1, 2024), pp. 73427–73456. ISSN: 1573-7721. DOI: 10.1007/s11042-023-17769-6. URL: https://doi.org/10.1007/s11042-023-17769-6 (visited on 01/09/2025).

[2] Mariko Kikutani and Machiko Ikemoto. "Detecting emotion in speech expressing incongruent emotional cues through voice and content: investigation on dominant modality and language". In: *Cognition and Emotion* 36 (Jan. 3, 2022), pp. 1–20. DOI: 10.1080/02699931.2021.2021144.

[3] Boaz M. Ben-David et al. "Prosody and semantics are separate but not separable channels in the perception of emotional speech: Test for Rating of Emotions in Speech". In: *Journal of Speech, Language, and Hearing Research* 59.1 (2016). Place: US Publisher: American Speech-Language-Hearing Assn, pp. 72–89. ISSN: 1558-9102. DOI: 10.1044/2015_JSLHR-H-14-0323.

[4] Charles Callaway and Khalil Sima'an. "Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship". In: *Computational Linguistics* 32 (Sept. 1, 2006), pp. 451–452. DOI: 10.1162/coli.2006.32.3.451.

[5] Éva Székely, Siyang Wang, and Joakim Gustafsson. "So-to-Speak : an exploratory platform for investigating the interplay between style and prosody in TTS". In: 24th International Speech Communication Association, Interspeech 2023, August 20-24, 2023, Dublin, Ireland. International Speech Communication Association, 2023, pp. 2016–2017. URL: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-337833 (visited on 01/13/2025).

[6] A. Klapuri and M. Davy. *Signal Processing Methods for Music Transcription*. Springer US, 2007. ISBN: 978-0-387-32845-4. URL: https://books.google.se/books?id=AF30yR41GIAC.

[7] Gangamohan Paidi, Sudarsana Kadiri, and B. Yegnanarayana. "Analysis of Emotional Speech—A Review". In: Mar. 22, 2016, pp. 205–238. ISBN: 978-3-319-31056-5. DOI: 10.1007/978-3-319-31056-5_11.

[8] Łukasz Stolarski. "Pitch Patterns in Vocal Expression of "Happiness" and "Sadness" in the Reading Aloud of Prose on the Basis of Selected Audiobooks". In: *Research in Language* 13 (June 1, 2015). DOI: 10.1515/rela-2015-0016.

[9] Marc Schröder et al. *Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis*. Vol. 1. Pages: 90. Sept. 3, 2001. 87 pp. DOI: 10.21437/Eurospeech.2001-34.

[10] Ccile Pereira. "Dimensions Of Emotional Meaning In Speech". In: *ISCA Workshop on Speech and Emotion* (Sept. 10, 2000).

[11] Jo-Anne Bachorowski. "Vocal Expression and Perception of Emotion". In: *Current Directions in Psychological Science* 8.2 (Apr. 1, 1999). Publisher: SAGE Publications Inc, pp. 53–57. ISSN: 0963-7214. DOI: 10.1111/1467-8721.00013. URL: https://doi.org/10.1111/1467-8721.00013 (visited on 01/09/2025).

[12] Frank Pollick. *In Search of the Uncanny Valley*. Vol. 40. Journal Abbreviation: User Cent. Med Pages: 78 Publication Title: User Cent. Med. Dec. 9, 2009. 69 pp. ISBN: 978-3-642-12629-1. DOI: 10.1007/978-3-642-12630-7_8.

[13] Paul Ekman Group. *Universal Emotions | What are Emotions? | Paul Ekman Group*. URL: https://www.paulekman.com/universal-emotions/ (visited on 01/13/2025).

[14] James Russell. "A Circumplex Model of Affect". In: *Journal of Personality and Social Psychology* 39 (Dec. 1, 1980), pp. 1161–1178. DOI: 10.1037/h0077714.

[15] Emma Rodero. "Intonation and Emotion: Influence of Pitch Levels and Contour Type on Creating Emotions". In: *Journal*

of Voice 25.1 (Jan. 1, 2011), e25–e34. ISSN: 0892-1997. DOI: 10.1016/j.jvoice.2010.02.002. URL: https://www.sciencedirect.com/science/article/pii/S0892199710000378 (visited on 01/13/2025).