# The Prosody of Emojis

**Anonymous ACL submission**

## Abstract

Prosodic features such as pitch, timing, and intonation are central to spoken communication, conveying emotion, intent, and discourse structure. In text-based settings, where these cues are absent, emojis act as visual surrogates that add affective and pragmatic nuance. This study examines how emojis influence prosodic realisation in speech and how listeners interpret prosodic cues to recover emoji meanings. Unlike previous work, we directly link prosody and emojis by analysing human speech data collected through a controlled elicited production task[1]. Using Bayesian multilevel modelling, we show that speakers systematically adapt their prosody based on emoji cues, and that listeners can recover intended meanings significantly above chance. Furthermore, our results reveal a clear hierarchy in prosodic shifts: greater semantic differences between emojis correspond to increased prosodic divergence. These findings suggest that emojis are meaningful carriers of prosodic intent that bridge the gap between digital text and spoken production.

## 1 Introduction

In spoken language, prosody, encompassing pitch, rhythm, and intonation, and other paralinguistic features, plays a crucial role in conveying linguistic nuances and socio-emotional cues (Cole, 2015; Hellbernd and Sammler, 2016; Ward and Levow, 2021). Acting as a dynamic carrier of information, prosody enriches spoken communication by modulating affect, signalling discourse structure, and shaping listener expectations. However, the shift toward computer-mediated communication presents challenges: text-based interaction strips away vital non-verbal elements like intonation, gesture, and facial expression, essential to holistic interpretation (Archer and Akert, 1977; Crystal, 2001).

To compensate for this loss, users of digital communication platforms increasingly rely on emojis, which enrich digital communication by offering a visual medium for expressing emotions (Gülşen, 2016) and conveying semantic meanings (Na'aman et al., 2017), transcending textual constraints. Their interpretive flexibility, however, raises intriguing questions about their relationship to prosody. Prior research has shown that prosodic features cannot be fully recovered from text alone (Wolf et al., 2023). This leads to a compelling inquiry: can emojis, embedded in otherwise static written language, serve as mediators for prosodic intent, compensating for the absence of vocal cues? Specifically, we investigate the following research questions:

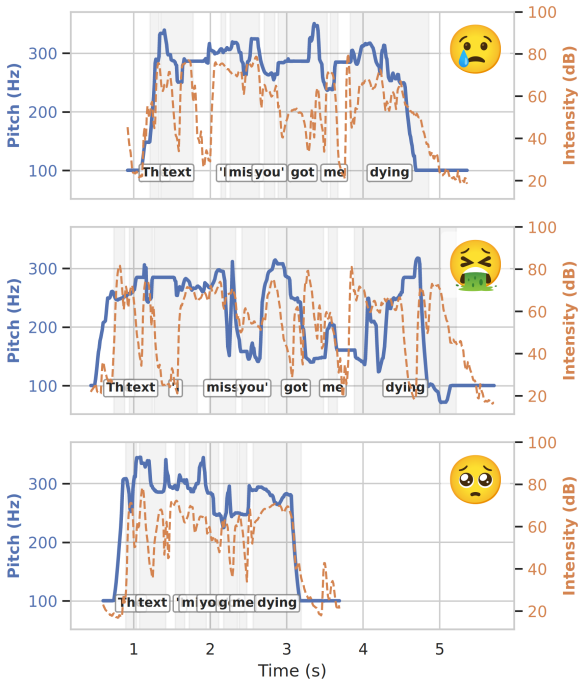



Figure 1: Prosodic variations for the sentence *"The text 'I miss you' got me dying"* across three emoji contexts. Each subplot displays the pitch (solid blue, Hz) and intensity (dashed orange, dB) contours.

**RQ1:** Do emojis influence how prosodic features are expressed in spoken language?

[1]The data will be publicly released for research purposes.

**RQ2:** Can listeners interpret prosodic cues to recover intended emoji meanings in speech?

**RQ3:** How does inter-speaker similarity in prosodic realisation relate to listener interpretation of emoji meaning?

**RQ4:** Is there a correlation between the semantic distance of emojis and the magnitude of the prosodic shift in their vocal delivery?

To investigate these questions, we conducted four online experiments to collect a dataset of emoji-enriched speech and listener perception data (Figure 1). We employed a design that holds lexical content constant across emoji conditions to isolate the influence of emojis on prosody, thereby preventing confounding effects from varying syntax or vocabulary. While this involves reading specific sentences rather than recording spontaneous conversations, we do not instruct participants on how to deliver the text or focus on the emojis. This approach captures a broad range of naturalistic expressive strategies while ensuring that observed prosodic shifts reflect authentic interpretations of emoji meaning.

Our results demonstrate that emojis consistently invite prosodic modulation across diverse speakers, even with constant verbal content. This variation is often sufficient for listeners to recover intended emoji meanings significantly above chance. We find that prosodic convergence, where different speakers independently produce similar contours for the same emoji, significantly boosts listener success. This suggests that shared prosodic intuitions facilitate robust interpretation. Finally, our analysis reveals a hierarchy in prosodic shifts. The magnitude of prosodic change correlates with emoji semantic distance, as greater semantic dissimilarity between emojis leads to systematically larger prosodic divergence.

## 2 Background

### 2.1 Emoji Interpretation

Emojis constitute a pervasive form of digital paralanguage: graphical cues that represent the nonverbal behaviours, such as facial expressions and gestures, intrinsic to face-to-face interaction (Alshenqeeti, 2016). From a psychological perspective, emojis do not merely decorate text but actively modulate affective processing. They have been shown to trigger cognitive responses in the receiver similar to those elicited by physical facial expressions, thereby influencing social attributions of the sender, such as perceived warmth, trust, and sincerity, and promoting a sense of social presence (Boutet et al., 2021).

Beyond their social impact, emojis serve critical pragmatic functions by acting as illocutionary force indicators. They allow users to clarify intent and manage face-threatening acts, such as softening a critique or signalling irony, which are tasks typically reserved for vocal nuance (Li and Yang, 2018; Holtgraves and Robinson, 2020). However, this interpretive process is fraught with ambiguity. While literal meanings are relatively stable across languages, figurative is closely intertwined with the sentiment of the context (Zhou et al., 2024a). Despite the presence of contextual cues intended to clarify intent, empirical evidence shows that misinterpretation persists, suggesting that emojis remain a high-variance channel that context alone cannot resolve (Miller et al., 2017; Czkestochowska et al., 2022).

### 2.2 Prosody in Speech Processing

Prosody encompasses the suprasegmental features of speech that operate beyond individual phonemes, such as intonation, rhythm, pitch, and timing that structure discourse and convey speaker intent (Lehiste and Lass, 1976; Gerken and McGregor, 1998). While speech processing systems often prioritise lexical cues, especially in cascade pipelines where prosodic nuance is lost during transcription (Zhou et al., 2024b; Tsiamas et al., 2024), research shows that these cues can independently support tasks like spoken question-answering, emotion recognition, and speaker profiling (Hirschberg, 2002; Chi et al., 2025).

This importance extends to speech synthesis, where naturalness depends on accurate prosodic modelling. Recent efforts have focused on this by implementing expressive prosody transfer (Skerry-Ryan et al., 2018) and modelling human-like rhythm to improve listener engagement in Text-to-Speech (TTS) systems (Kane et al., 2024).

### 2.3 Prosody and Emojis

Given that emojis function as pragmatic substitutes for non-verbal cues, they present a unique opportunity to recover the prosodic intent lost in text. Prior research suggests conceptual parallels, likening emojis to written intonation or visual facial expressions (James, 2017) that highlight information focus or compensate for a lack of vocal expressivity

(Wagner, 2016; Hu et al., 2019; Kaiser, 2021).

Despite these theoretical links, empirical analysis of how emojis influence the acoustic realisation of speech is limited. While some work has proposed emojis as graphical surrogates for prosody based on surveys (Alnuzaili et al., 2024), or used emoji-enriched speech to increase the variance of TTS systems (Tuttösí et al., 2025), research has yet to characterise the systematic prosodic adaptations human speakers make in response to different emoji semantics. In contrast, our study examines how speakers naturally modulate their acoustic prosody when encountering emoji-enriched utterances. By analysing real human speech, we provide empirical evidence that emoji semantics directly influence prosodic expression, bridging the gap between symbolic digital cues and spoken production.

## 3 Methods

### 3.1 Data Collection

To investigate our research questions, we conducted four online experiments[2], with each experiment employing a unique pool of participants. To isolate the influence of emojis on prosody, we employed a design that holds lexical content constant across conditions. While this involves elicited speech rather than spontaneous conversation, this approach is necessary to generate the parallel data required to compare how the same lexical string is transformed by different emoji contexts, a phenomenon rarely captured in naturalistic corpora.

The experimental framework is built upon a base corpus of 230 unique utterances, each of which originally contained at least one emoji. These were manually selected and edited from publicly available user-generated textual content to cover a broad range of topics (e.g., music, gaming, film) and diverse emoji usage. All selected utterances were specifically screened to ensure they remained natural when spoken aloud.

**Experiment 1: Utterance-Emoji Stimulus**

The objective of Experiment 1 was to construct a collection of utterances with identical lexical content but varying emoji usage. Participants were presented with the base utterances (with original emojis removed) and instructed to create up to five variations per sentence by inserting different emojis to alter the emotional tone or interpretation. To ensure variety, they were advised against using emojis with highly similar meanings for the same utterance. Each utterance was annotated independently by two annotators, resulting in 1,595 unique emoji-augmented stimuli.

**Experiment 2: Emoji-Influenced Prosodic Data**

We collected a corpus of emoji-influenced speech by asking participants to read a subset of the stimuli annotated in Experiment 1. Each speaker recorded approximately 25 utterances, ensuring they produced at least two distinct variations of each base sentence (including a non-emoji control version). To minimise bias in prosodic expression, participants were not informed of the focus on emojis. Following the recordings, participants summarised the perceived meaning of each emoji in one word to verify their engagement and interpretation. The final dataset comprised 3,153 valid recordings from 129 speakers. To account for potential sources of variance in our analysis, we also collected demographic data, including age, gender, and self-reported general emoji usage (measured on a scale of 1–5). Additionally, participants reported their recording hardware (headset, mobile device, or built-in computer microphone).

**Experiment 3: Judging Prosodic Variation Within Speakers**

To test listeners' perception of prosodic changes within the same speaker, we sampled 645 recording pairs which had identical lexical content. These pairs represented two distinct contrast types: *Dual Emoji*, comparing two emoji renditions (e.g., *Text* + 😔 vs. *Text* + 🥰) and *Single Emoji*, comparing an emoji rendition against an emoji-free baseline. For each pair, listeners: judged whether the recordings differed in prosodic realisation (RQ1), and matched each recording to its corresponding emoji (RQ2). Listeners were instructed to focus on vocal expressiveness and ignore recording artefacts. Each pair received judgments from an average of three annotators. Consistent with Experiment 2, we collected demographic data from these listeners, including age, gender, and general emoji usage (scale 1-5).

**Experiment 4: Identifying Shared Prosody Across Speakers**

Experiment 4 investigated whether inter-speaker similarity in prosodic realisation, or prosodic convergence, facilitates the interpretation of emoji

[2]Appendix G for full instructions and Appendix B for data summaries.

| Predictor | Est. ($\beta$) | 95% CI |
|---|---|---|
| (Intercept) | -0.25 | [-0.66, 0.15] |
| Age | -0.04 | [-0.16, 0.07] |
| Gender (Female) | 0.16 | [-0.12, 0.44] |
| Mic (Headset) | 0.10 | [-0.26, 0.46] |
| Mic (Mobile) | -0.15 | [-0.42, 0.12] |
| Emoji Use | **0.18** | **[0.06, 0.30]** |
| Contrast Type (Dual) | **0.29** | **[0.03, 0.55]** |

Table 1: Model 1 results for perceived prosodic differentiation. Bold indicates 95% Credible Intervals that strictly exclude zero.

meaning (RQ3). To ensure sufficient expressive range, we excluded speakers who demonstrated low prosodic variation in Experiment 3 (see Section 4.2). The final stimulus set comprised 49 inter-speaker audio pairs, each consisting of recordings of the same sentence and emoji produced by two different speakers. Participants performed two tasks: judging the prosodic similarity of the recordings, and matching each recording to either the target emoji or a random distractor. Crucially, while both recordings in a pair shared the same target emoji, participants were not informed of this fact. Each pair received judgments from an average of four annotators. As in Experiments 2 and 3, we collected demographic data from these listeners, including age, gender, and general emoji usage.

### 3.2 Statistical Analysis

To evaluate our research questions, we implemented Bayesian linear mixed-effects regression models (Appendix A). This framework allows us to directly quantify the strength of evidence for our hypotheses by estimating the full probability distribution of our parameters. We modelled the influence of emoji cues on both vocal production and listener perception, including the accuracy of recovering intended meanings. By including random effects for speakers, listeners, and specific utterances, we ensured that our results account for individual expressive variability and the inherent prosodic differences between sentences.

## 4 Prosody in Emoji-Enriched Speech

### 4.1 Speaker Variation in Prosodic Expression

Our first inquiry examines whether emojis elicit prosodic variation (RQ1). We analysed data from Experiment 3, where listeners compared pairs of recordings produced by the same speaker. To quantify this effect, we modelled listener judgments using a Bayesian multilevel logistic regression:

**Model 1**: `PerceivedVariation` $\sim$ `Age + Gender + Mic + EmojiUse + ContrastType + (1 | Utterance) + (1 | Listener)`

In this model, `PerceivedVariation` is the binary response variable (whether the two recordings sounded different or not), while `ContrastType` (Single vs. Dual Emoji comparisons), self-reported `EmojiUse`, and speaker demographics serve as the fixed effects.

The results, detailed in Table 1, provide a nuanced answer to RQ1. The influence of emojis appears to be a capacity that varies significantly between individuals. While most speakers exhibit moderate variation, there is significant spread, with clear tails of highly expressive and highly monotonous speakers (see Appendix B.1). Model 1 helps explain this variance through `EmojiUse` ($\beta = 0.18, CI[0.06, 0.30]$). The positive effect suggests that for frequent emoji users, the parallel between emoji semantics and prosody is stronger, leading to more distinct acoustic signals. Furthermore, `ContrastType` systematically influenced detectability. The positive coefficient for Dual Emoji trials ($\beta = 0.29, CI[0.03, 0.55]$) indicates that listeners were significantly more likely to detect a difference when comparing two distinct emojis than when comparing an emoji utterance against a neutral baseline. This implies that the prosodic modifications speakers make are not binary. Instead, the signal is graded, becoming most acoustically distinct when speakers must actively differentiate between conflicting states. These findings verify the existence of the parallel proposed in RQ1: emojis can and do influence prosodic expression. However, this influence is not uniform. Instead, emojis function as an expressive resource that is modulated by the speaker's digital fluency and the communicative necessity of the contrast.

### 4.2 Listener Interpretation of Prosodic Variation

Beyond mere production, we investigated whether these variations effectively communicate symbolic intent and allow listeners to recover intended emoji meanings (RQ2). We tested this using a second model:

**Model 2**: `IdSuccess` $\sim$ `SpeakerExpressivity + ContrastType + (SpeakerEmojiUse` $\times$

| Predictor | Est. ($\beta$) | 95% CI |
|---|---|---|
| *Model 2: Intra-speaker Identification (RQ2)* | | |
| (Intercept) | -0.16 | [-0.45, 0.11] |
| Speaker Expressivity | **0.24** | **[0.10, 0.38]** |
| Contrast Type (Dual) | 0.03 | [-0.27, 0.32] |
| Speaker Emoji Use | 0.05 | [-0.09, 0.19] |
| Listener Emoji Use | -0.15 | [-0.31, 0.01] |
| Speaker×Listener Use | 0.09 | [-0.05, 0.23] |
| *Model 3: Inter-speaker Identification (RQ3)* | | |
| (Intercept) | -1.79 | [-3.38, -0.17] |
| Perceived Similarity (Yes) | **2.42** | **[1.66, 3.22]** |

Table 2: Model 2 (intra-speaker) and Model 3 (inter-speaker) results for emoji identification success. Bold indicates 95% Credible Intervals that strictly exclude zero.

`ListenerEmojiUse) + (1 | Utterance) + (1 | Listener)`

In this model, `IdSuccess` denotes the correct assignment of emoji meanings to both recordings in the pair. Key predictors included `SpeakerExpressivity` (a proxy for expressive bandwidth derived from Experiment 3), `ContrastType` (Dual vs. Single), and the interaction between speaker and listener `EmojiUse` to test for shared digital fluency effects. Model 2 (Table 2) revealed a baseline identification probability of $P(\text{IdSuccess}) \approx 0.46$ (Intercept = $-0.16$). Although the intercept coefficient includes zero (indicating a probability near 50%), this value sits significantly above the 25% chance level[3], confirming that the task was generally performable.

The primary driver of communicative success was Speaker Expressivity: for every standard deviation increase in a speaker's prosodic differentiation, the log-odds of correct emoji identification increased reliably ($\beta = 0.24, CI[0.10, 0.38]$). Interestingly, while Contrast Type was a major factor in the perception of differences (Model 1), it did not reliably impact identification success in Model 2 ($\beta = 0.03, CI[-0.27, 0.32]$). This suggests that once a prosodic difference is detected, the intended meaning is equally recoverable whether comparing two different emojis or an emoji against a neutral baseline. Additionally, the lack of a reliable interaction between speaker and listener emoji familiarity indicates that this communication is mutually intelligible regardless of digital background, relying instead on universal expressive mechanisms. These findings address RQ2 by demonstrating that while listeners can interpret prosodic cues to recover emoji meanings significantly above chance, the success of this communication is not universal; rather, it is a high-variance channel that depends largely on the individual speaker's ability to encode pragmatic intent into vocal variation.

[3]Chance performance is 25% based on the four possible emoji assignment combinations available to the listener (Log-odds $\approx -1.1$).

### 4.3 Prosodic Convergence and Emoji Interpretation

RQ3 shifts focus to the inter-speaker dynamics explored in Experiment 4. Specifically, we investigated whether prosodic convergence, defined as instances where different speakers independently produce similar contours for the same emoji, facilitates higher listener accuracy. To test this, we modelled the relationship between the listener's perception of similarity and their ability to correctly identify the emoji using a third Bayesian model:

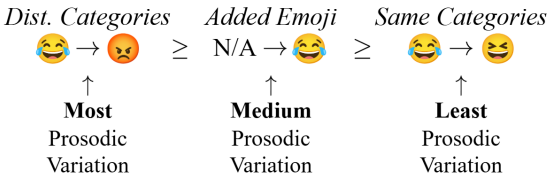
**Model 3**: $\texttt{IdSuccess} \sim \texttt{PerceivedSimilarity} + (1 \mid \texttt{Utterance}) + (1 \mid \texttt{Listener}) + (1 \mid \texttt{Speaker}_A) + (1 \mid \texttt{Speaker}_B)$

The fixed effect `PerceivedSimilarity` is a binary predictor indicating whether the listener judged the recordings from the two speakers as sounding prosodically similar. Random effects account for variance across utterances, listeners, and the two unique speakers involved in each pair ($\texttt{Speaker}_A$ and $\texttt{Speaker}_B$). The regression results, presented as Model 3 in Table 2, confirm that perceived similarity is a robust predictor of successful emoji recovery. The model found a significant positive effect for similarity ($\beta = 2.42, 95\%$ CI $[1.66, 3.22]$). This result suggests that listeners are approximately 11 times more likely to achieve discriminative success when speakers converge on a shared prosodic signature. Even after controlling for random variation, the evidence for this effect remained extreme with a posterior probability of 1.0. These findings provide a direct answer to RQ3 by demonstrating that inter-speaker similarity in prosodic realisation is fundamentally tied to listener interpretation success. When speakers independently converge on a shared prosodic strategy, listener recovery of intended meaning improves significantly. However, communicative success is notably more fragile when prosodic strategies diverge. In such cases, listeners often perceive the recordings as sounding different and prefer to assign distinct emoji mean-

ings to each, reflecting the increased uncertainty caused by a lack of prosodic agreement (further details in Appendix B.3). Thus, while emoji vocalisation reflects a stable and shared paralinguistic code that can ensure semantic intent remains recoverable across different voices, its effectiveness is largely contingent on whether speakers align their prosodic intuitions.

## 5 Prosodic Distance and Emoji Semantic Variations

This section investigates how emoji semantics influence prosodic variation. To address RQ4, we test the hypothesis that prosodic differences are greater when associated emojis belong to distinct semantic categories than when they share a category. Additionally, we examine whether the inclusion of an emoji in a previously emoji-free sentence induces greater prosodic change than switching between emojis within the same semantic group.




### 5.1 Emoji Semantic Categories

Accurately categorising emojis poses a challenge because standard groupings, such as Unicode emoji groups and subgroups[4], do not always reflect the communicative or semantic similarities perceived by users (e.g., 🥲 and 🥰 belong to the *"face-affection"* subgroup). To address this, we derived emoji embeddings from participant intents (Experiment 1) and interpretations (Experiment 2). Despite individual variation, these embeddings showed strong correlations between intents and interpretations, indicating a substantial shared understanding of the communicative roles of emojis (see Appendix E for detailed analysis).

To identify meaningful semantic groups, we performed unsupervised clustering on the interpretation embeddings using HDBSCAN (Campello et al., 2013) with a minimum cluster size of 3 (see Appendix F for further discussion). Based on these clusters, we established three experimental conditions for our prosodic analysis: recording pairs where emojis belong to different semantic clusters (Diff); recording pairs where emojis belong to the same semantic cluster. (Same); pairs consisting of an emoji-free recording matched with an emoji-enriched version of the same utterance (Added).

All relevant recording pairs were extracted and categorised accordingly, excluding speakers with minimal prosodic differentiation ($\leq 30\%$, as identified in Section 4.1).

### 5.2 Methodology

For each recording, we extracted multiple prosodic features to capture variation across several dimensions. Pitch contours were estimated using the FCNF0++ model (Morrison et al., 2023), and intensity was calculated via the Praat Parselmouth interface (Jadoul et al., 2018). To account for varying recording lengths, we employed Dynamic Time Warping (DTW) to align pitch and intensity contours. From these paths, we derived a Temporal Warping metric defined as the percentage of non-diagonal steps in the alignment to quantify rhythmic divergence and local temporal distortion. Additionally, we calculated the absolute difference in speech rate (characters per second) as a measure of global articulation speed via the Stopes toolkit (Seamless Communication et al, 2023).

Beyond raw acoustic features, we extracted prosodic embeddings from the 7th layer of the Masked Prosody Model (Wallbridge et al., 2025). These frame-wise, 256-dimensional representations are explicitly trained to model pitch, energy, and voice activity. Distances between these sequences were computed using multivariate DTW to capture holistic prosodic divergence across these latent dimensions.

To determine if the magnitude of the prosodic shift varies systematically according to emoji semantic relationships (RQ4), we employed Bayesian linear mixed-effects regression. We fitted separate models for each acoustic feature using the following specification:

**Model 4**: `ProsodicDistance` ∼
`SemanticRelation + (1 + SemanticRelation | Speaker) + (1 | Utterance)`

The dependent variable, `ProsodicDistance`, represents the calculated acoustic difference between the recordings. To satisfy normality assumptions and facilitate direct comparison across features, all distance metrics were log-transformed and Z-standardised prior to fitting. The fixed effect, `SemanticRelation`, categorises the pairing

[4] https://unicode.org/Public/emoji/16.0/emoji-test.txt

| Hypothesis | Est. ($\beta$) | 95% CI | ER |
|---|---|---|---|
| *Embeddings* | | | |
| Distinct > Added | **0.10** | **[0.05, 0.14]** | >100 |
| Same < Added | 0.05 | [-0.03, 0.13] | 0.2 |
| Distinct > Same | 0.05 | [-0.02, 0.12] | 6.2 |
| *Intensity* | | | |
| Distinct > Added | **0.06** | **[0.02, 0.11]** | 63.0 |
| Same < Added | -0.01 | [-0.09, 0.08] | 1.2 |
| Distinct > Same | 0.07 | [-0.01, 0.15] | 12.1 |
| *Character Speech Rate* | | | |
| Distinct > Added | 0.09 | [0.00, 0.19] | 21.2 |
| Same < Added | -0.01 | [-0.20, 0.18] | 1.1 |
| Distinct > Same | 0.10 | [-0.07, 0.28] | 5.1 |
| *Temporal Warping* | | | |
| Distinct > Added | 0.04 | [-0.04, 0.12] | 3.3 |
| Same < Added | -0.08 | [-0.22, 0.06] | 5.1 |
| Distinct > Same | 0.12 | [-0.01, 0.25] | 12.9 |
| *Pitch* | | | |
| Distinct > Added | 0.02 | [-0.05, 0.09] | 2.2 |
| Same < Added | -0.01 | [-0.13, 0.11] | 1.3 |
| Distinct > Same | 0.03 | [-0.08, 0.15] | 2.2 |

Table 3: Model 4 results for prosodic shifts across semantic conditions relative to the 'Added' baseline. ER = Evidence Ratio. Bold indicates 95% credible intervals that strictly exclude zero (certainty of non-zero magnitude). Underline indicates ER > 3 (substantial evidence for directionality).

as *Added Emoji* (Baseline), *Same Category* (Congruent), or *Distinct Category* (Divergent). The random effects structure included random intercepts for speakers and utterances, as well as by-speaker random slopes for `SemanticRelation` to capture individual variability in prosodic strategies.

### 5.3 Results

Table 3 summarises the posterior estimates and evidence ratios for each acoustic feature. Given the high inter-speaker variability inherent in prosodic realisation, we provide both 95% Credible Intervals (CI) and Evidence Ratios (ER) to distinguish between certainty of effect magnitude and probability of effect direction. Following Jeffreys' scale (Jeffreys, 1998), we interpret an ER $> 3$ as "substantial" and $> 10$ as "strong" evidence for a directional shift, allowing for the identification of consistent trends in cases where high variance causes the 95% CI to overlap zero.

We observed a clear hierarchy where prosodic divergence was largest for Distinct Category pairings, while Same Category pairings clustered near the

| Feature | 😢 - 🥺 | 🤮 - 🥺 | 😢 - 🤮 |
|---|---|---|---|
| Embeddings | 6.03 | 6.14 | **6.18** |
| Intensity | **0.58** | **0.58** | 0.53 |
| Speech Rate | 1.28 | **1.68** | 0.58 |
| Temp. Warp | 4.38 | **4.48** | 4.45 |
| Pitch | 1.49 | **2.67** | 2.60 |

Table 4: Prosodic distances between three recordings of the sentence *"The text 'I miss you' got me dying"*, spoken by a single participant with different emoji cues. Bolded values indicate the highest distance within each row.

Added Emoji baseline. This "conflict effect" was supported by very strong directional evidence in the Prosodic Embeddings ($\beta = 0.10, \text{ER} > 100$), driven by a robust increase in Intensity ($\beta = 0.06, \text{ER} = 63.0$) and a high directional probability for increased Speech Rate ($\text{ER} = 21.2$) relative to the baseline. Crucially, speakers also differentiated conflict from congruence: Distinct pairs exhibited greater Temporal Warping ($\text{ER} = 12.9$) and Intensity ($\text{ER} = 12.1$) than Same pairs, indicating moderate to strong directional evidence for prosodic differentiation. Conversely, synonymous emojis functioned as prosodic stabilisers, showing negligible deviation from the baseline ($\text{ER} \approx 1.2$) and even a potential reduction in temporal warping ($\beta = -0.08, \text{ER} = 5.1$). Notably, while raw Pitch Distance showed the expected directionality ($Distinct > Added$), the evidence was inconclusive due to high variance ($\text{ER} \approx 2.2$). This stands in contrast to Intensity, which provided a robust univariate signal, and the Prosodic Embeddings, which showed very strong differentiation. This discrepancy likely stems from the acoustic fragility of raw pitch tracking: features like creaky voice or irregular phonation, which are common in expressive speech, introduce noise that obscures the DTW signal.

## 6 Discussion

### 6.1 Prosodic Differentiation Across Emojis

Figure 1 illustrates prosodic modulation (RQ1) for the sentence *"The text 'I miss you' got me dying"* across three emoji contexts.

The 😢 and 🥺 versions exhibit comparable high, sustained pitch contours, though 🥺 is more compressed in duration. In contrast, the 🤮 version diverges with a rising pitch on *"you"* and a sarcastic profile reinforced by a loudness spike on *"got"*.

These observations align with the log-transformed distances in Table 4. The semantically similar 😢 – 🥺 pair shows minimal prosodic distance, while the 🤮 – 🥺 pair exhibits the highest divergence in temporal alignment and pitch contours. These results align with the broader hierarchy identified in Model 4 regarding the overall magnitude of divergence. While the general model indicates that distinct categories drive larger shifts globally, this specific case illustrates the divergence primarily through pitch and timing.

### 6.2 Semantic Overlap and Expressive Limitation

While many speakers modulate prosody in response to emojis, others produce similar patterns across different emoji conditions. These overlaps raise questions about whether they stem from meaningful semantic similarities or individual expressive limitations, a distinction that is particularly challenging when lexical content implies a dominant tone.

Semantic redundancy often explains this lack of contrast. For instance, in utterances like *"Happy birthday, have a great day"* or *"He started salsa dancing,"* the inherent positivity of the text aligns closely with emojis such as 😁 , 😍 , or 💃 . In these cases, the lack of prosodic differentiation likely reflects shared affective intent or semantic alignment rather than a failure to encode the cue. Conversely, other examples suggest expressive constraints. The sentence *"Welp, this is gonna hurt to watch"* was read in an emotionally flat manner for both 😬 and 😔 despite their divergent tones of nervous tension and sadness. Such instances may reflect a difficulty or unwillingness to encode nuanced emotional distinctions in prosody, or a mismatch between the emoji and the speaker's personal interpretation.

Together, these examples highlight a central challenge for RQ1: similar prosody across conditions may reflect semantic overlap, speaker limitations, or the deliberate downplaying of emoji cues. Disentangling these factors is methodologically difficult and would require detailed manual inspection alongside speaker-level analysis.

### 6.3 Prosodic Divergence and Interpretation Failure

Finally, we consider a case where prosodic strategies fail to support accurate listener interpretation (RQ2/RQ3) involving the sentence *"I'm not surprised* 😱 ", where 😱 is the intended emoji and 🤮 the random alternative in Experiment 4.

The first speaker produced a lowering pitch contour with a subtle lengthening on "surprised", giving the utterance a flat, resigned tone signalling emotional detachment rather than shock. Post-hoc interpretation revealed the speaker viewed the emoji as representing "shame," indicating a misalignment between their internal representation of 😱 and its conventional use. The second speaker used a faster tempo and sharper accents that were perceived as sarcastic rather than alarmed. Listeners in both cases favoured the 🤮 emoji, likely due to the mismatch between these prosodic cues and the semantics of the target emoji.

This example highlights several key challenges. Prosodic strategies vary widely between speakers and may not always align with shared semantic expectations. Furthermore, listener interpretation depends not only on prosodic cues but also on culturally or contextually anchored understandings of emoji meaning. Even when prosody is used expressively, its communicative intent can remain opaque without a clear shared mapping between prosodic form and emoji semantics. These failures reinforce our Model 3 findings, demonstrating that successful communication is contingent on speakers and listeners aligning their paralinguistic intuitions.

## 7 Conclusion

This work provides the first systematic evidence linking emoji semantics to prosodic expression and interpretation in speech. We establish that emojis act as mediators of prosodic intent, allowing speakers to encode pragmatic nuance into otherwise static text. Our results reveal a hierarchy in this process: the magnitude of prosodic divergence is systematically governed by the semantic distance between emojis, with intensity and timing serving as primary acoustic carriers of emoji-driven intent. Finally, we demonstrate that these modulations enable listeners to recover intended meanings with accuracy significantly above chance.

Beyond these findings, this study highlights emojis as a structured paralinguistic layer bridging the gap between digital text and spoken production. Future work could leverage these emoji–prosody mappings to enhance multimodal embeddings or drive acoustic emoji prediction, positioning emojis as vital tools for modelling the nuances of human communication.

## Limitations

While our empirical analysis establishes a robust baseline for emoji-driven prosody, this study represents an initial exploration rather than an exhaustive account. First, we prioritised lexical control via elicited reading over spontaneous conversation. Although participants received no instructions to focus on emojis, the resulting prosody may be more deliberate than in casual speech. Second, untrained speakers exhibited significant variation in expressivity. While our Bayesian models accounted for individual baselines, we observed that the clarity of prosodic cues is graded and varies by speaker, rather than being uniform across the population. Furthermore, our curated stimuli may introduce selection bias, and we did not control for cross-platform emoji rendering. Finally, as this work is restricted to British English, findings may not generalise to languages with distinct prosodic or pragmatic norms.

## Ethical Considerations

This research was approved by a departmental ethics board (details are omitted to preserve anonymity), with all participants providing explicit informed consent for the use and release of their recordings and annotations for research purposes. Recruitment was conducted via Prolific at a compensation rate of £9/hour, adhering to fair pay and living wage standards. To maintain privacy, data was anonymised at the source using Prolific IDs, ensuring no direct identifiers were collected. To verify data integrity and safety, we utilized an Automatic Speech Recognition (ASR) model (Whisper) to ensure recordings matched the target stimuli and contained no personally identifying information. This automated process was supplemented by manual audits of a subset of recordings and annotations to verify the absence of harmful content.

## References


Ehab Saleh Alnuzaili, Muhammad Waqar Amin, Sami Saad Alghamdi, Nazir Ahmed Malik, Abdulbasit A. Alhaj, and Asad Ali. 2024. Emojis as graphic equivalents of prosodic features in natural speech: evidence from computer-mediated discourse of whatsapp and facebook. *Cogent Arts & Humanities*, 11(1):2391646.

Hamza Alshenqeeti. 2016. Are emojis creating a new or old visual language for new generations? a socio-semiotic study. *Advances in language and Literary Studies*, 7(6).

Dane Archer and Robin M Akert. 1977. Words and everything else: Verbal and nonverbal cues in social interpretation. *Journal of personality and social psychology*, 35(6):443.

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

Isabelle Boutet, Megan LeBlanc, Justin A Chamberland, and Charles A Collin. 2021. Emojis influence emotional communication, social attributions, and information processing. *Computers in Human Behavior*, 119:106722.

Paul-Christian Bürkner. 2021. Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, 100(5):1–54.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Jie Chi, Maureen de Seyssel, and Natalie Schluter. 2025. The role of prosody in spoken question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8468–8479, Albuquerque, New Mexico. Association for Computational Linguistics.

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

David Crystal. 2001. *Language and the Internet*. Cambridge University Press.

Justyna Czkestochowska, Kristina Gligorić, Maxime Peyrard, Yann Mentha, Michał Bień, Andrea Grütter, Anita Auer, Aris Xanthos, and Robert West. 2022. On the context-free ambiguity of emoji. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1388–1392.

LouAnn Gerken and Karla McGregor. 1998. An overview of prosody and its role in normal and disordered child language. *American Journal of Speech-Language Pathology*, 7(2):38–48.

Tüge T Gülşen. 2016. You tell me in emojis. In *Computational and cognitive approaches to narratology*, pages 354–375. IGI Global.

Nele Hellbernd and Daniela Sammler. 2016. Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, 88:70–86.

Julia Hirschberg. 2002. Communication and prosody: Functional aspects of prosody. *Speech Communication*, 36(1-2):31–43.

Thomas Holtgraves and Caleb Robinson. 2020. Emoji can facilitate recognition of conveyed indirect meaning. *PloS one*, 15(4):e0232361.

Jiaxiong Hu, Qianyao Xu, Limin Paul Fu, and Yingqing Xu. 2019. Emojilization: An automated method for speech to emoji-labeled text. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2:193–218.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

Allan James. 2017. Prosody and paralanguage in speech and the social media: The vocal and graphic realisation of affective meaning. *Linguistica*, 57(1):137–149.

Harold Jeffreys. 1998. *The theory of probability*. OuP Oxford.

Elsi Kaiser. 2021. Focus marking with emoji: On the relation between information structure and expressive meaning. In *Colloque de syntaxe et sémantique á Paris (CSSP), Paris, France*.

Joseph Kane, Michael N Johnstone, and Patryk Szewczyk. 2024. Voice synthesis improvement by machine learning of natural prosody. *Sensors*, 24(5):1624.

Ilse Lehiste and Norman J Lass. 1976. Suprasegmental features of speech. *Contemporary issues in experimental phonetics*, 225:239.

Li Li and Yue Yang. 2018. Pragmatic functions of emoji in internet-based communication—a corpus-based study. *Asian-Pacific Journal of Second and Foreign Language Education*, 3:1–12.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Hannah Miller, Daniel Kluver, Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication. In *Eleventh international AAAI conference on web and social media*.

Max Morrison, Caedon Hsieh, Nathan Pruyne, and Bryan Pardo. 2023. Cross-domain neural pitch and periodicity estimation. In *arXiv preprint arXiv:2301.12258*.

Noa Na'aman, Hannah Provenza, and Orion Montoya. 2017. Varying linguistic purposes of emoji in (Twitter) context. In *Proceedings of ACL 2017, Student Research Workshop*, pages 136–141, Vancouver, Canada. Association for Computational Linguistics.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Seamless Communication et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Ioannis Tsiamas, Matthias Sperber, Andrew Finch, and Sarthak Garg. 2024. Speech is more than words: Do speech-to-text translation systems leverage prosody? In *Proceedings of the Ninth Conference on Machine Translation*, pages 1235–1257, Miami, Florida, USA. Association for Computational Linguistics.

Paige Tuttösí, Shivam Mehta, Zachary Syvenky, Bermet Burkanova, Gustav Eje Henter, and Angelica Lim. 2025. Emojivoice: Towards long-term controllable expressivity in robot speech. *arXiv preprint arXiv:2506.15085*.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854.

Michael Wagner. 2016. How to be kind with prosody. *in Speech prosody*, 1:250–1253.

Sarenne Wallbridge, Christoph Minixhofer, Catherine Lai, and Peter Bell. 2025. Prosodic structure beyond lexical content: a study in self-supervised learning. In *proceedings of Interspeech 2025*.

Nigel Ward and Gina-Anne Levow. 2021. Prosody: Models, methods, and applications. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 26–28, Online. Association for Computational Linguistics.

Lukas Wolf, Tiago Pimentel, Evelina Fedorenko, Ryan Cotterell, Alex Warstadt, Ethan Wilcox, and Tamar Regev. 2023. Quantifying the redundancy between prosody and text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9765–9784, Singapore. Association for Computational Linguistics.

Giulio Zhou, Sydelle De Souza, Ella Markham, Oghenetekevwe Kwakpovwe, and Sumin Zhao. 2024a. Semantics and sentiment: Cross-lingual variations in emoji use. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18698–18712, Miami, Florida, USA. Association for Computational Linguistics.

Giulio Zhou, Tsz Kin Lam, Alexandra Birch, and Barry Haddow. 2024b. Prosody in cascade and direct speech-to-text translation: a case study on Korean wh-phrases. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 674–683, St. Julian's, Malta. Association for Computational Linguistics.

## A Model Specifications

All models were implemented in R (R Core Team, 2025, version 4.5.2) using the `brms` package (Bürkner, 2021), which interfaces with the Stan probabilistic programming platform. Binary outcomes for interpretation and perception trials (RQ1, RQ2, and RQ3) were modelled using a Bernoulli distribution with a logit link function. Continuous prosodic distance measures (RQ4) were modelled using Gaussian distributions.We applied weakly informative priors to the fixed effects ($Normal(0, 0.5)$) to provide regularisation and ensure stable posterior convergence. This prior choice reflects a conservative assumption that large effect sizes are relatively unlikely. Centred around a null effect of zero, these priors help prevent the models from over-fitting to noise in the high-variance speech data. Following the principle of the maximal random effects structure (Barr et al., 2013), we included random intercepts for speakers, listeners, and utterances to account for the non-independence of observations. Where experimental design permitted, specifically in Model 4, we included random slopes for semantic conditions by speaker. This accounts for individual variability in how speakers adapt their prosodic strategies to different emoji cues, ensuring that our population-level estimates are robust to speaker-level idiosyncratic behaviour.

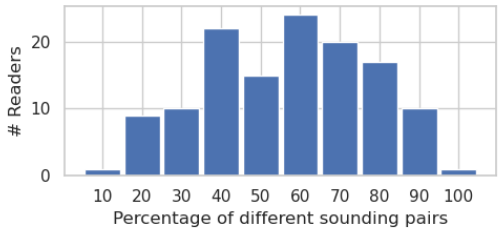



Figure 2: Readers (Experiment 2) grouped by the proportion of recordings judged as prosodically different (Experiment 3). Each bin represents a 10% range (e.g., 0%–9%), from least to most prosodically varied speakers.

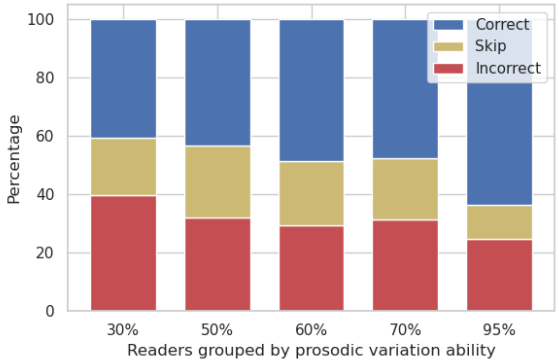



Figure 3: Listener (Experiment 3) accuracy in emoji assignment for prosodically distinct pairs, grouped by speaker variation. Bars show correct, incorrect, and skipped responses from least to most prosodically varied speakers.

## B Data statistics

### B.1 Speaker Expressivity

The distribution in Figure 2 categorises speakers by the proportion of their recording pairs judged as prosodically distinct. While 20 readers exhibited minimal differentiation (0–30%), the majority (81 speakers) showed moderate variation (31–70%), and a notable subset (28 speakers) consistently produced distinct renditions (71–100%). These results demonstrate that while individual variability is present, the prevalence of distinct recordings supports the conclusion that emoji cues systematically influence how prosody is realised.

### B.2 Listener Accuracy - Intra Speaker

Figure 3 illustrates how a speaker's expressiveness directly impacts communicative success. A Spearman's rank correlation revealed a significant positive relationship between speaker variation and correct emoji identification ($\rho = 0.90, p = .037$)

| Response | Overall | Correct | Partial | Incorrect |
|---|---|---|---|---|
| **Same** | 34.74% | 66.67% | 6.06% | 27.27% |
| **Different** | 65.26% | 14.75% | 74.59% | 10.66% |

Table 5: Results from the inter-speaker experiment showing how listeners judge prosodic similarity between two recordings and how these judgments affect their accuracy in assigning each recording to the correct emoji.

| Predictor | Est. ($\beta$) | 95% CI |
|---|---|---|
| (Intercept) | -1.867 | $[-2.279, -1.492]$ |
| Speaker Expressivity | -0.182 | $[-0.368, 0.003]$ |
| Contrast Type (Single) | 0.196 | $[-0.203, 0.591]$ |
| Speaker Emoji Use | -0.045 | $[-0.233, 0.135]$ |
| Listener Emoji Use | 0.151 | $[-0.184, 0.481]$ |
| Speaker $\times$ Listener Use | -0.063 | $[-0.257, 0.126]$ |

Table 6: Bayesian multilevel logistic regression results predicting the probability of a "semantic tie" (assigning the same emoji to both recordings)

and a significant negative correlation for incorrect assignments ($\rho = -0.90, p = .037$). For the most expressive speakers (95% bin), accuracy reached 62.50%, significantly exceeding both the incorrect rate (25.52%) and a random-choice baseline. Conversely, in the lowest variation group (30%), correct (41.09%) and incorrect (39.11%) rates were nearly equal, indicating that listeners were reduced to random guessing. While skip rates dropped from 20% to 11.98% in the highest bin, this trend was not statistically significant ($\rho = -0.10, p = .87$).

### B.3 Listener Accuracy - Inter Speaker

Results in Table 5 show that 34.74% of inter-speaker pairs were judged as sounding similar, while 65.26% were judged as different. This suggests that speakers often employ flexible prosodic strategies for the same emoji. In this experiment, listeners were presented with two recordings and two emoji options, allowing for four possible assignment combinations. We define Correct as the proper identification of the target emoji for both recordings, while Partial correctness occurs when a listener assigns the target emoji to only one of the two recordings. Incorrect reflects a failure to identify the target emoji in either instance. When listeners perceived prosodic similarity, they achieved full correctness 66.67% of the time, indicating strong interpretive agreement when prosody converges. For pairs judged as different, the high partial correctness rate (74.59%) reflects increased uncertainty; listeners often successfully "recovered" the emoji for one speaker while failing for the other. However, even when prosody differed, the full correctness rate (14.75%) still exceeded the incorrect rate (10.66%), suggesting that divergent realisations can still convey intended meaning, albeit with less reliability.

## C Analysis of Perceptual Overlap - RQ2

To further investigate the mechanics of communicative success, we analysed the likelihood of "semantic ties." These represent instances where a listener assigned the same emoji to both recordings within a single trial, modelled as a "Skip" in Figure 3. A high rate of semantic ties indicates perceptual overlap, where the speaker's prosodic modulation is insufficient to signal a change in symbolic intent. We modelled the probability of such ties using a Bayesian multilevel logistic regression:

**Model 5**: `SemanticTie` $\sim$
`SpeakerExpressivity + ContrastType +`
`(SpeakerEmojiUse` $\times$ `ListenerEmojiUse) +`
`(1 | Utterance) + (1 | Listener)`

The results in Table 6 show a strongly negative intercept ($\beta = -1.87, CI[-2.28, -1.49]$), indicating that the baseline probability of a semantic tie is very low ($P \approx 0.13$). This suggests that listeners are naturally predisposed to seek out prosodic distinctions rather than perceiving the recordings as identical. The primary driver of further reducing this perceptual ambiguity was Speaker Expressivity ($\beta = -0.18, CI[-0.37, 0.003]$). While the credible interval narrowly overlaps zero, the negative trend suggests that higher expressive bandwidth effectively helps listeners to recognise distinct meanings. By producing more distinctive prosodic signals, highly expressive speakers successfully disambiguate the potential semantic tie, making it easier for listeners to identify that two different meanings were intended. This finding reinforces the conclusion that the interpretability of emoji-enriched speech is directly contingent on the speaker's ability to vocalise pragmatic contrasts.

| | Ambiguity | Correct | Partial | Incorrect |
|---|---|---|---|---|
| ❓ | 0.1577 | 25.0 | 75.0 | 0.0 |
| 😉 | 0.1707 | 83.3 | 16.7 | 0.0 |
| 🤛 | 0.2131 | 0.0 | 100.0 | 0.0 |
| 😢 | 0.2443 | 66.7 | 33.3 | 0.0 |
| 😱 | 0.3750 | 0.0 | 20.0 | 80.0 |
| 😕 | 0.4630 | 100.0 | 0.0 | 0.0 |
| 😏 | 0.5158 | 66.7 | 0.0 | 33.3 |
| 😶 | 0.5410 | 22.2 | 44.4 | 33.3 |
| 👍 | 0.6593 | 0.0 | 50.0 | 50.0 |

Table 7: Examples illustrating emoji ambiguity and Experiment 4 listener interpretation accuracy (in percentages), sorted from low to high ambiguity.

## D Listener Accuracy and Emoji Ambiguity

Table 7 presents a subset of emojis from Experiment 4 alongside listener interpretation accuracy and ambiguity scores derived from the semantic variation metrics reported by Czkestochowska et al. (2022). These scores reflect cross-participant variability in emoji interpretation when presented without context, where higher values indicate greater semantic ambiguity. Although the sample size precludes statistical generalisation, the examples illustrate a qualitative trend: emojis with lower ambiguity scores tend to yield fewer incorrect interpretations, whereas highly ambiguous emojis result in greater listener disagreement.

Specific emojis revealed distinct patterns within the data. Emojis such as 🙌 and 😕 demonstrated clear prosodic convergence alongside high accuracy. In contrast, emojis like 🤔 and 🤗 were reliably recognised despite divergent prosody, suggesting that inherent semantic context may at times override prosodic mismatch. Conversely, for 😂 , 😱 , and 😶 , prosodic similarity did not consistently ensure accurate interpretation.

Qualitative comparison with existing ambiguity scores suggests that while semantic ambiguity does not dictate the degree of prosodic convergence, emojis with lower inherent ambiguity tend to result in fewer incorrect interpretations. These tentative patterns suggest that higher semantic ambiguity may reduce the recoverability of emoji intent from prosody alone, as listeners lack a stable semantic "anchor" to which they can map vocal variations.

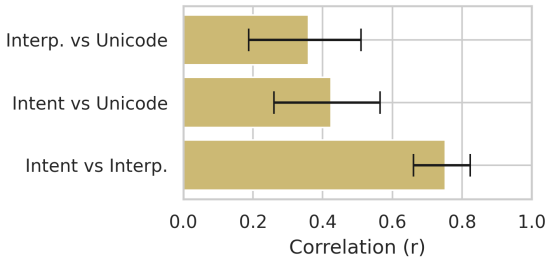



Figure 4: Mean Pearson correlations between emoji meaning spaces (Intent vs Interpretation vs Unicode Descriptions). Error bars indicate 95% confidence intervals computed using Fisher z-transformation.

## E Emoji Meaning Analysis

To analyse the semantic consistency of emoji usage, we compared three meaning representations for each emoji: the *intent* (from Experiment 1), the *interpretation* (from Experiment 2), and the official Unicode *description*.

For each emoji, we aggregated the most frequent annotations and merged them into a representative string per meaning type. We then generated sentence embeddings using the `sentence-transformers/all-mpnet-base-v2` model (Reimers and Gurevych, 2019; Song et al., 2020). This yielded one embedding per emoji for each semantic category.

We applied UMAP (McInnes et al., 2018) to reduce the embeddings to two dimensions for visualisation. Emojis were plotted directly in 2D space using their glyphs, after normalising representations (e.g., removing skin tone and gender modifiers) to avoid redundancy and improve clarity (Figure 5).

We evaluated the alignment between the three emoji meaning spaces by computing pairwise Pearson correlation coefficients between corresponding emoji embeddings. To assess the reliability of these correlations across the dataset, we applied the Fisher z-transformation, which stabilises the variance of correlations and enables us to compute statistically sound confidence intervals for the mean. Figure 4 shows the mean correlation and associated 95% confidence intervals across all emojis with at least five annotations. We observe the strongest alignment between intent and interpretation embeddings (**mean $r$ = 0.752**, CI: [0.660, 0.823]), reflecting high agreement between what users intend and how others interpret emoji use. In contrast, the unicode description space shows significantly

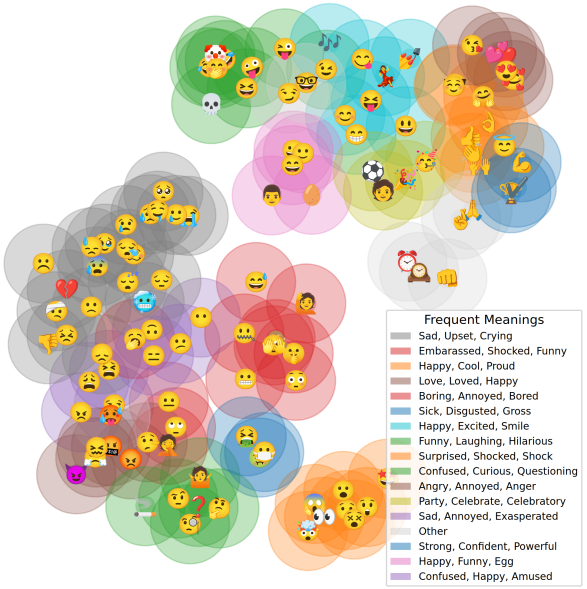



Figure 5: Emoji embeddings reduced with UMAP. Clusters computed with HDBSCAN (on 49 components). Labels are top 3 keywords among the words used to describe the emojis.

weaker alignment with both intent ($r$ **= 0.424**, CI: [0.260, 0.564]) and interpretation ($r$ **= 0.359**, CI: [0.188, 0.510]), reinforcing prior findings that Unicode definitions often fail to capture how emojis are actually used in communicative context.

## F Emoji Clustering

To explore emergent semantic groupings among emojis, we performed unsupervised clustering on the embeddings derived from Experiments 1 and 2, as well as from the emoji Unicode descriptions. We first applied Principal Component Analysis (PCA) to determine the intrinsic dimensionality of the embedding space. Based on the explained variance, we selected 49 dimensions to preserve approximately 90% of the variance.

Clustering was performed using HDBSCAN (Campello et al., 2013, Hierarchical Density-Based Spatial Clustering of Applications with Noise), a density-based algorithm that can identify an optimal number of clusters based on the density distribution of the data. HDBSCAN does not require the number of clusters to be predefined and is robust to noise and outliers. Emojis that did not belong to any cluster were labelled as noise. We set `min_cluster_size` to 3 to allow for finer-grained clusters, and used Euclidean distance as the similarity metric.

Figure 5 shows the two-dimensional projections of the resulting embeddings and clusters. The Figure shows labels for each cluster derived from the most frequent emoji interpretations within the group. These derived clusters serve as our semantic categories in subsequent analyses.

| Cluster Comparison | ARI | NMI |
|---|---|---|
| Int. vs Inter. | 0.249 | 0.613 |
| Int. vs Desc. | 0.077 | 0.431 |
| Inter. vs Desc. | 0.062 | 0.443 |

Table 8: Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) between clusterings based on different semantic spaces.

To assess the alignment between emoji clusters derived from different semantic spaces (Intent, Interpretation, and Description), we computed the *Adjusted Rand Index* (Hubert and Arabie, 1985, ARI) and *Normalized Mutual Information* (Vinh et al., 2010, NMI) between each pair of clusterings. These metrics quantify the similarity of cluster assignments while accounting for chance (ARI) and shared information (NMI).

The ARI scores suggest moderate agreement between *Intent* and *Interpretation* clusters (ARI $=$ 0.249), whereas clusterings involving *Description* show weak alignment (ARI $<$ 0.1). NMI results are consistent with this pattern: the Intent and Interpretation spaces share more information (NMI $=$ 0.613) than either does with the Description space (NMI $\approx$ 0.43). These findings indicate that user-generated intentions and interpretations exhibit a shared semantic structure that is not well captured by official Unicode descriptions.

## G Instructions and Trial Samples

### G.1 Experiment 1

Participants were given the following instructions:“

1. *Select ONE emoji and place it in any suitable position within the provided sentence. Only use emojis and spaces, avoiding inserting any other elements.*

2. *Describe how the chosen emoji modifies the sentence. Keep it short! One word should be enough.*

3. *Repeat until you have filled out at* ***least two forms*** *with a suitable emoji and its corresponding descriptions.*

*Avoid choosing multiple emojis that are too similar.*

*For instance, if you select a happy emoji, consider avoiding another happy emoji to add diversity to your selections."*

Figure 6a shows an example of a trial page for Experiment 1.

### G.2 Experiment 2

Participants were asked to complete the Recording Tasks first, and then provide their interpretation of the encountered emojis. Here, the instructions given to the participants.

#### G.2.1 Recording Task

*"Click on the mic icon to start recording audio, then click again to stop.*

*You will be able to playback your recording. You can confirm by pressing 'Continue' or try again by pressing 'Record again' (it will start recording immediately).*

*Try to record in a quiet environment."*

Figure 6b shows the recording interface.

#### G.2.2 Annotation Task

*"In this second (and last) part, you will be asked to describe how you interpreted the emoji used in the previous section.*

*Keep it short!* ***One word*** *should be enough.*

*Leave blank if there wasn't any emoji."*

### G.3 Experiment 3

Figure 6c shows an example of the interface for Experiment 3. The same interface was presented for experiments with Inter and Intra-Speaker Data.

#### G.3.1 Inter-Speaker Data

*"In this study, you will listen to pairs of recorded speech. Each pair will contain the same words, but they may differ in how they are spoken. For example, in rhythm, intonation, intensity, use of pauses and so on. Your task is to determine whether there are intentional variations between the two recordings. Please disregard any differences that do not seem deliberate. For example, small lexical variations or stutter.*

*You will be provided with transcriptions of the recordings, which may contain different emojis or none at all. Your task is to match each recording to the transcription that best corresponds to it.*

*Try to assign each recording to a different transcription first!"*

#### G.3.2 Intra-Speaker Data

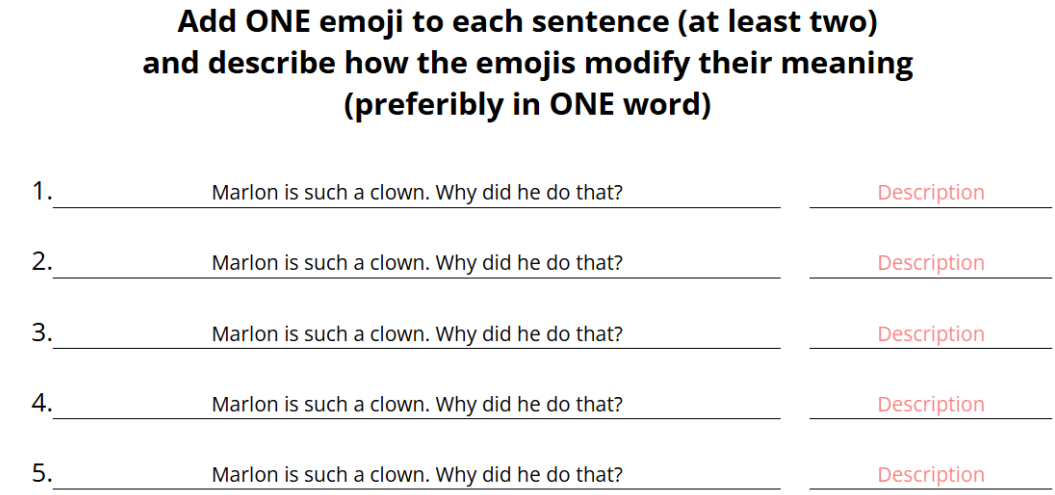
*"In this study, you will listen to pairs of recorded speech from* ***two different speakers****. Each pair will contain the* ***same words, but they may differ in how they are spoken****, for example, in rhythm, intonation, intensity, use of pauses and so on. Your task is to determine whether there are intentional variations between the two recordings. Please disregard any differences that do not seem deliberate. For example, small lexical variations or stutter.*

*You will hear* ***two recordings of the same sentence, spoken by different speakers.***

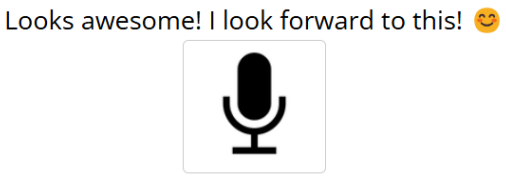
*Your task is to listen carefully to each recording and choose the emoji that best matches its tone.*

*Each speaker produced their recording based on an emoji prompt.*

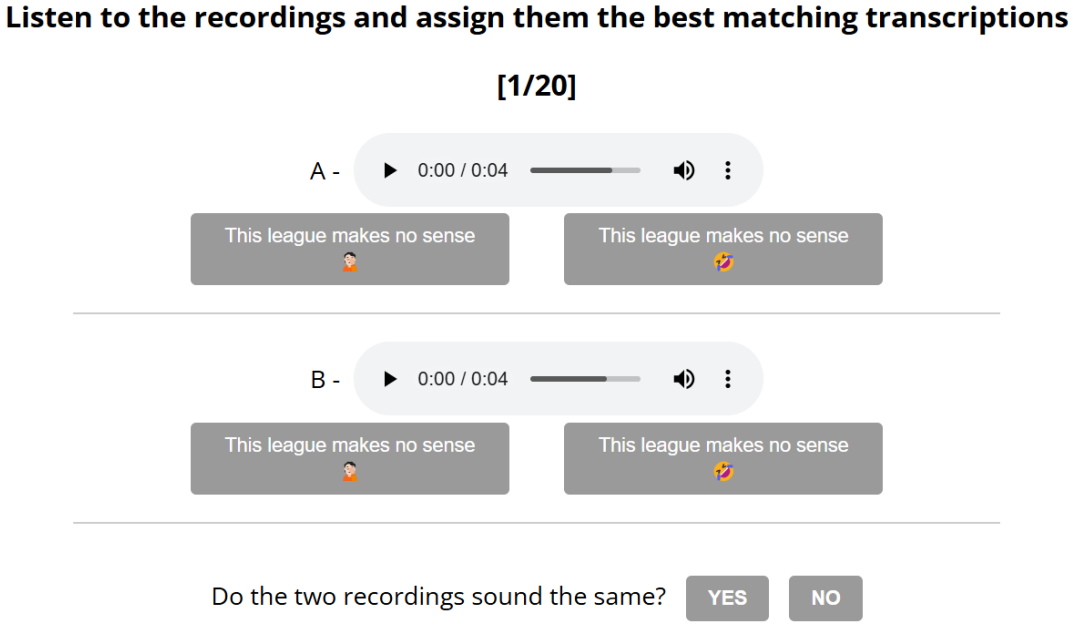
*You can assign the* ***same emoji to both recordings****, or a* ***different one to each****, depending on what you perceive."*




(a) Experiment 1 - Emoji Annotation Task




(b) Experiment 2 - Recording Task




(c) Experiment 3 and 4 - Listening Task

Figure 6: Example of trials' main page for the online experiments.