

What Do Prosody and Text Convey? Characterizing How Meaningful Information is Distributed Across Multiple Channels

Aditya Yadavalli
UC San Diego
a1yadavalli@ucsd.edu

Tiago Pimentel
ETH Zürich
tiago.pimentel@inf.ethz.ch

Tamar I. Regev
MIT
tamarr@mit.edu

Ethan Gottlieb Wilcox
Georgetown University
ethan.wilcox@georgetown.edu

Alex Warstadt
UC San Diego
awarstadt@ucsd.edu

Abstract

Prosody – the melody of speech – conveys critical information often not captured by the words or text of a message. In this paper, we propose an information-theoretic approach to quantify how much information is expressed by prosody alone and not by text, and crucially, what that information is about. Our approach applies large speech and language models to estimate the mutual information between a particular dimension of an utterance’s meaning (e.g., its emotion) and any of its communication channels (e.g., audio or text). We then use this approach to quantify how much information is conveyed by audio and text about sarcasm, emotion, and questionhood, using speech from television and podcasts. We find that for sarcasm and emotion the audio channel – and by implication the prosodic channel – transmits over an order of magnitude more information about these features than the text channel alone, at least when long-term context beyond the current sentence is unavailable. For questionhood, prosody provides comparatively less additional information. We conclude by outlining a program applying our approach to more dimensions of meaning, communication channels, and languages.

1 Introduction

Language is more than just the words and phrases we speak, sign, or write down. Linguistic communication typically takes place over multiple modalities and channels, including facial expression, gesture, typography, and speech prosody (Aristotle, 1991; Kress, 2009; Patel-Grosz et al., 2023). Prosody, in particular, encompasses the suprasegmental features of speech such as pitch, tempo, and loudness, and plays an essential role in communication (Hirschberg, 2002). In written communication, the audio channel containing the prosody is completely removed, leaving behind mainly the lexical content, often referred to as

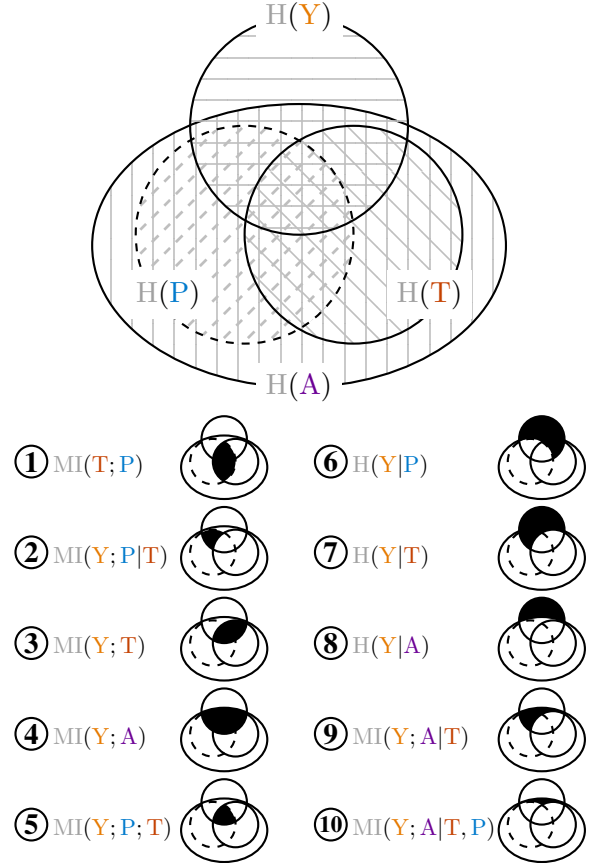


Figure 1: The informational relationship between linguistic channels – Text T , Prosody P , Audio A – and a meaningful feature Y . Features are discrete properties of a message, such as its syntax or the speaker’s affect.

the “segmental information”, but which we refer to as the **text**.¹ In this paper, we examine how much information about an utterance’s meaning is contained in that audio, and how much information is lost when only the text is available.

There is a growing trend of applying information

¹However, the information conveyed by prosody is so critical to successful communication that humans have developed numerous typographical conventions to recover much of that information loss, including commas, question marks, italics, and emojis (Chafe, 1988; Holtgraves and Robinson, 2020).

theory to questions in linguistics and language processing (Levy, 2008; Piantadosi et al., 2011; Futrell et al., 2020; Williams et al., 2021; Socolof et al., 2022). Particularly relevant, some recent work has investigated how information is distributed across the prosodic and textual channels (Wolf et al., 2023; Regev et al., 2025; Wilcox et al., 2025). However, this prior work gives an incomplete picture: Its focus has been entirely on using mutual information estimation to quantify the overall *redundancy* between text and prosody. While it is intriguing that text and prosody convey overlapping information, this methodology gives us no insight into how much *unique* information prosody conveys and what that information is *about*.

Here, we instead characterize how information about a particular meaningful feature (Y , e.g., whether an utterance is sarcastic) is distributed over distinct channels (prosody P , audio A , and text T). Formally, a feature Y ’s information is quantified by its entropy, $H(Y)$. As illustrated in fig. 1, this entropy can be decomposed as the sum of several information-theoretic terms ($H(Y) = \textcircled{3} + \textcircled{2} + \textcircled{10} + \textcircled{8}$). Thus, estimating these terms gives us a more complete picture of how information about Y is distributed across prosody and text. Following Pimentel et al. (2020), we estimate these information-theoretic terms by fine-tuning pretrained models to predict Y given one or more channels as input.²

As case studies, we apply our method to two features which we predict are strongly associated with prosodic signals – sarcasm and emotion – and to one feature we predict to be less strongly associated with prosody due to strong textual cues – questionhood. We find that the audio channel, including the prosody, conveys roughly an order of magnitude more information than text alone about the two first linguistic features – at least when the length of textual context is limited to a single sentence – but only 2.4 times as much for the latter feature.

More broadly, our approach introduces a framework that can be applied to many other research questions. Prosody conveys crucial information about syntax, lexical identity, speech act, and discourse – all of which can be investigated using our approach. Furthermore, any communication chan-

nel – be it the pitch contour, gesture, or emojis – can be similarly explored, as long as a classifier can be trained to accept input from that channel. Finally, languages differ significantly in the ways they use prosody (Hyman, 2008; Wilcox et al., 2025) as well as other communication channels. Our approach can give fine-grained insight into the diversity of strategies that can be used for distributing information across communication channels.

2 Background

2.1 Prosody

Prosody is the suprasegmental component of speech, including all the auditory features of the speech signal apart from those that signal to segmental features (i.e., phonemes). This encompasses pitch (fundamental frequency), tempo (duration and pauses), and loudness, but also more subtle spectral properties, such as harmonics-to-noise ratio or resonance changes affecting voice quality (Cheang and Pell, 2008). Prosodic cues can convey critical information, including the locations of word boundaries (Cutler, 2014), syntactic disambiguation (Pauker et al., 2011; Snedeker and Trueswell, 2003), information about lexical identity (Wilcox et al., 2025), grammatical marking (e.g., whether or not an utterance is a question; Cole, 2015; Hellbernd and Sammler, 2016), and information structure in a sentence (Breen et al., 2010; Roettger et al., 2019). Speakers can also use prosody to convey paralinguistic information, such as affect (Fernandez and Picard, 2011), intent, and signals that they are about to end their turn in a conversation (Cutler and Pearson, 2018). Importantly for this study, much of the information conveyed by prosody is absent in written language.

2.2 Spoken Language Models

Transformers (Vaswani et al., 2017) have made efficient processing of sequential data possible. While they were initially used for text processing (Devlin et al., 2019; Radford et al., 2019), they have also enabled many recent advancements in spoken language modeling. The speech representations output by Transformer-based models pretrained on unlabeled audio data such as wav2vec (Schneider et al., 2019; Baevski et al., 2020), HuBERT (Hsu et al., 2021), and many others (e.g., Conneau et al., 2020; Babu et al., 2021; Pratap et al., 2024) have been successfully applied to tasks such as speech recognition, speaker identification, and affect

²Choosing Y to be a discrete variable, MI estimation largely consists of training simple classifiers and measuring their cross-entropy loss on a dataset labeled for Y . When using A as our channel, our classifiers are based on spoken language models (SLM); when using T , they are based on language models (LMs).

classification with a limited amount of labeled training data. In particular, the Whisper models (Radford et al., 2022) were trained on 680,000 hours of audio data across speech recognition and other speech-to-text tasks and, at the time of their release, they showed state-of-the-art performance on many of these tasks across accents, languages, acoustic conditions, and speaker variability, leading to its widespread adoption (Radford et al., 2022; Olatunji et al., 2023; Javed et al., 2023; Bhogale et al., 2023; Talafha et al., 2023).

These advances in spoken language modeling have significant implications for human-computer interaction. As Large Language Models (LLMs) are increasingly being used in user-facing applications, they require capabilities beyond text understanding – particularly the ability to detect users’ affect and intents from vocal cues. While text-only models can infer affect only from orthography, audio-based models have access to crucial paralinguistic features like tone, pitch variations, and speech rate that often convey more reliable affective signals. This is why spoken language understanding, which integrates these audio capabilities with classical language understanding approaches, has emerged as a critical field bridging the gap between spoken language processing advancements and practical applications in conversational AI systems (Shon et al., 2022a,b, 2024; Arora et al., 2024). In addition to serving many practical purposes, the goal of our contribution is to demonstrate how speech models can also enable us to address fundamental scientific questions about language.

3 Information Content of Prosody

Formally, our objective is to study how information is distributed among four random variables: **T**, a **text**-valued random variable; **P**, **prosody**-valued; **A**, which is **audio**-valued; and **Y**, which assumes values of a **target feature** of interest (e.g., affect or sarcasm). We represent each of these random variables as upper-case letters (e.g., **T**), and their instances with lower-case letters (e.g., **t**). We also use **X** to denote any random variable representing a communication channel (or a combination thereof). For our purposes, a text **t** is an orthographic representation of a single utterance. A prosody-valued instance **p** is a joint representation of all the values of each prosodic feature across an entire utterance. An audio-valued instance **a** represents the

full waveform for an utterance. Finally, a target feature **y** represents the value this feature takes in an utterance, for instance, whether the entire utterance is sarcastic. All utterances we will consider contain a single sentence in English.

3.1 Mutual Information Estimation

The **mutual information** (MI) between random variables **X** and **Y**, with supports \mathcal{X} and \mathcal{Y} , is defined as:

$$\text{MI}(\mathbf{X}; \mathbf{Y}) = \sum_{\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}} p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}. \quad (1)$$

Mutual information is an information-theoretic measure that quantifies (e.g., in bits) how much the uncertainty about one random variable can be reduced when another random variable is observed (Shannon, 1948, 1951). More informally, MI is the amount of overlap in information between two variables (e.g., between sarcasm and prosody).

Due to the difficulty of estimating the joint distribution $p(\mathbf{X}, \mathbf{Y})$, Pimentel et al. (2020) propose to analyze mutual information as a difference of two entropies, given the following identity,

$$\text{MI}(\mathbf{X}; \mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y} | \mathbf{X}), \quad (2)$$

where these Shannon entropies H (Shannon, 1948) are defined as follows:

$$H(\mathbf{Y}) = - \sum_{\mathbf{y} \in \mathcal{Y}} p(\mathbf{y}) \log p(\mathbf{y}), \quad (3)$$

$$H(\mathbf{Y} | \mathbf{X}) = - \sum_{\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}} p(\mathbf{x}, \mathbf{y}) \log p(\mathbf{y} | \mathbf{x}). \quad (4)$$

Wolf et al. (2023) recently adopted this mutual information decomposition to measure $\text{MI}(\mathbf{T}; \mathbf{P})$, the redundancy between text **T** and prosody **P**.³ While quantifying this redundancy can provide some insight, it does not tell us what this information is actually *about*.

3.2 Our Approach

Here, we seek to quantify how much information *about* a specific type of meaning is shared by text and prosody, and how much is conveyed by them separately. The central idea is that, rather than estimate the mutual information of prosody and text directly $\text{MI}(\mathbf{T}; \mathbf{P})$, we can estimate their mutual information with some meaningful feature **Y** instead.

³Wolf et al. formulate **P** as ranging over the prosody of a single word. This difference does not affect our discussion.

Our approach is summarized in fig. 1. This diagram illustrates how different information-theoretic quantities of interest relate to one another. As discussed above, the quantity measured by Wolf et al. (2023), $MI(\mathbf{T}; \mathbf{P})$, is ①. We are interested in a larger set of quantities. We ultimately aim to characterize the relationship of the text \mathbf{T} , prosody \mathbf{P} , and some meaningful feature \mathbf{Y} . To estimate these values, however, we train classifiers to predict \mathbf{Y} from channel \mathbf{X} . While we can effectively train audio-based classifiers (using existing audio encoder models, see Section 2.2), no such architecture exists for prosody (see Section 8 for further discussion). We thus choose to estimate these terms using the full audio signal \mathbf{A} instead of \mathbf{P} .

Fortunately, we can estimate many (but not all) of the relevant quantities for prosody by studying just the text and the full audio signal. The information conveyed by audio $H(\mathbf{A})$ contains all of the information conveyed by text $H(\mathbf{T})$ and prosody $H(\mathbf{P})$. Thus, ④ $MI(\mathbf{Y}; \mathbf{A})$ contains information about \mathbf{Y} that can be obtained from either text, prosody, or some other aspect of the audio. We estimate this quantity using the following identity

$$MI(\mathbf{Y}; \mathbf{A}) = H(\mathbf{Y}) - H(\mathbf{Y} | \mathbf{A}). \quad (5)$$

In the same way, we can estimate ③ $MI(\mathbf{Y}; \mathbf{T})$ as

$$MI(\mathbf{Y}; \mathbf{T}) = H(\mathbf{Y}) - H(\mathbf{Y} | \mathbf{T}), \quad (6)$$

Then, we can approximate ② $MI(\mathbf{Y}; \mathbf{P} | \mathbf{T})$ given the above MIs. This quantity tells us how much information prosody conveys about \mathbf{Y} that is not conveyed through the text. In practice, though, we compute a different conditional MI,⁴

$$MI(\mathbf{Y}; \mathbf{A} | \mathbf{T}) = MI(\mathbf{Y}; \mathbf{A}) - MI(\mathbf{Y}; \mathbf{T}). \quad (7)$$

However, we argue that the quantity ⑨ $MI(\mathbf{Y}; \mathbf{A} | \mathbf{T})$ is a good approximation for ② $MI(\mathbf{Y}; \mathbf{P} | \mathbf{T})$. This is because ⑨ overestimates ② only by the term ⑩ $MI(\mathbf{Y}; \mathbf{A} | \mathbf{T}, \mathbf{P})$, where ⑩ is all the information in the audio signal about \mathbf{Y} that is not conveyed by either text or prosody. We reason that this quantity is minimal: Most of the information in audio that does not fall under text or prosody is likely to be irrelevant information about background noise, accent, or other qualities

⁴Following the definition of MI, technically $MI(\mathbf{Y}; \mathbf{A} | \mathbf{T}) = MI(\mathbf{Y}; \mathbf{A}, \mathbf{T}) - MI(\mathbf{Y}; \mathbf{T})$. However, this can be simplified to eq. (7) under our assumption that \mathbf{A} fully determines \mathbf{T} .

of speaker identity.⁵ Thus, we can approximately isolate the unique contribution of the prosodic channel to communication about a particular kind of meaning, such as sarcasm, without requiring access to a specialized pretrained prosody model.

4 Estimating Mutual Information

Prior work (Pimentel et al., 2020; Williams et al., 2021; Wolf et al., 2023) has used neural network-based classifiers to estimate the mutual informations above. In particular, Pimentel et al. (2020) propose a method to estimate $MI(\mathbf{X}; \mathbf{Y})$ given a dataset \mathcal{D} , composed of pairs $(\mathbf{x}_i, \mathbf{y}_i)$ sampled from the joint distribution $p(\mathbf{X}, \mathbf{Y})$. We follow their approach here.

Relying on the decomposition in eq. (2), we must estimate two entropies. First, the unconditional entropy $H(\mathbf{Y})$ is estimated from the dataset non-parametrically using a plug-in estimation:

$$p(\mathbf{y}) \approx \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \mathbb{1}_{\mathbf{y}}(\mathbf{y}_i). \quad (8)$$

Second, the conditional entropy $H(\mathbf{Y} | \mathbf{X})$ is estimated using a neural network with parameters θ optimized on a separate dataset $\mathcal{D}_{\text{train}}$ to predict each \mathbf{y}_i from each \mathbf{x}_i . This model learns a conditional distribution $p_{\theta}(\mathbf{Y} | \mathbf{X})$, which can be used to compute the cross-entropy $H_{\theta}(\mathbf{Y} | \mathbf{X})$ as follows:

$$H_{\theta}(\mathbf{Y} | \mathbf{X}) \approx -\frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{i=1}^{|\mathcal{D}_{\text{test}}|} \log p_{\theta}(\mathbf{y}_i | \mathbf{x}_i). \quad (9)$$

As $H_{\theta}(\mathbf{Y} | \mathbf{X}) \geq H(\mathbf{Y} | \mathbf{X})$, we want to obtain the tightest upper bound on $H(\mathbf{Y} | \mathbf{X})$ by training the best model as possible. In our case, this means fine-tuning all the weights of a large foundation model to predict \mathbf{Y} from \mathbf{X} .

Comparison to Prior Work Estimating MI with Prosody.

Our estimation technique largely consists of training simple neural network classifiers to predict meaningful features like sarcasm from some input channel like text or audio. Prior information-theoretic approaches to studying prosody (Wolf et al., 2023; Wilcox et al., 2025; Regev et al., 2025) estimate $MI(\mathbf{T}; \mathbf{P})$ by decomposing it into $H(\mathbf{P})$ and $H(\mathbf{P} | \mathbf{T})$, which requires predicting the continuous variable \mathbf{P} . By treating

⁵This quantity may arguably include relevant content such as laughter, filler words, and small variations in phoneme production which may or may not be defined as prosodic.

the discrete variable Y as the dependent variable, our method avoids two key pitfalls of prior work:

First, the entropy of continuous variables, known as **differential entropy**, can be negative and does not have an intuitive interpretation. While the difference of $H(P | T)$ and $H(P)$ can still be interpreted as the MI ,⁶ the differential entropies on their own are not meaningful. By contrast, the entropies we measure, namely $H(Y)$ and $H(Y | X)$ can be directly interpreted as the expected information (in bits) gained when learning the value of Y for a given sentence-length utterance, with or without access to the channel X .

Second, learning probability density functions for continuous variables introduces several complications. These prior works estimate $H(P)$ using Gaussian kernel density estimator (KDE) models (Sheather, 2004), and they estimate $H(P | T)$ by fitting a neural network to output a parametrized distribution $p_\theta(P | t)$ given an input text t . Because the KDE for $p_\theta(P)$ is nonparametric, it can fit highly complex density functions, while $p_\theta(P | t)$ is a parametric probability density function. This can lead to overestimates of $H(P | T)$ and even negative estimates for MI , unless specialized methods like Mixture Density Networks (Bishop, 1994) are used (Wilcox et al., 2025). On the other hand, our approach simply requires training standard neural network classifiers with cross-entropy loss.

The main advantage of the prior approach (Wolf et al., 2023; Wilcox et al., 2025; Regev et al., 2025) is that their models are text-only LMs and their data consists of prosodic features automatically extracted from aligned text-audio data. Our approach requires models that accept input from the relevant channels (e.g., speech models, or multimodal text/prosody models which are currently unavailable), as well as aligned text-audio data with gold labels for the target feature Y .

5 Case Studies

We apply our mutual information estimation methods to three tasks: sarcasm, affect, and question-hood classification. We choose the first two tasks because, besides being meaning domains that are expected to depend on prosody, they both have

high-quality pre-existing datasets that pair audio, transcriptions, and labeled class data. For the third task, we curate a new dataset from existing resources.

5.1 Tasks & Datasets

Sarcasm Detection Sarcasm, sometimes referred to as “verbal irony,” plays an important role in daily communication. It can be defined as an individual expressing the opposite of what they believe (Haverkate, 1990), therefore setting up incongruity between a context, an utterance, and how it is expressed (i.e., its prosody; Matsui et al., 2016). Sarcasm serves several purposes, including modulating in-group status through humor, and its use varies between social groups (Dress et al., 2008).

One property of sarcasm that is useful for our purpose is that its relationship with text is highly ambiguous: Some sentences are obviously sarcastic from their text (and world knowledge) alone, while others are not easy to categorize without additional prosodic information. For example, the sentence *Wow, the dog smells amazing after rolling around in the garbage!* can easily be inferred to be sarcastic. However, the sentence *I really love this place!* could be sarcastic or not, and context, gesture, or prosody can help to disambiguate this. Such variance makes sarcasm an ideal testing bed for studying how information is distributed through different communicative channels.

There is extensive prior work on sarcasm detection (Farabi et al., 2024), but much of it was conducted in the text-only domain (Kreuz and Caucci, 2007; Filatova, 2012; Joshi et al., 2016; Abercrombie and Hovy, 2016). While initial contributions have been made using multimodal models (Tepperman et al., 2006; Cheang and Pell, 2008; Rakov and Rosenberg, 2013), they relied on smaller datasets. In our study, we use the publicly available Multimodal Sarcasm Detection Dataset (MUSARD) released by Castro et al. (2019). The dataset includes 690 utterances annotated for the presence of sarcasm, drawn from a variety of situational comedy TV shows. The classes are balanced.

Affect Classification Affect is conveyed through multiple communicative channels, including text, audio, and vision (Frick, 1985; Mozziconacci, 2002; Larrouy-Maestri et al., 2024). Several high-quality datasets for affect detection exist (Burkhardt et al., 2005; Busso et al., 2008; Bänziger et al., 2012; Shah et al., 2013; Russo and Living-

⁶In such cases where one measures the MI of a continuous and a discrete variable, the equality in eq. (2) requires additional definitions and assumptions (namely, the good mixed pair assumption; see Wolf et al., 2023 for details). This caveat applies not only to prior work, but also to our work in estimating quantities such as $MI(Y; A|T)$.

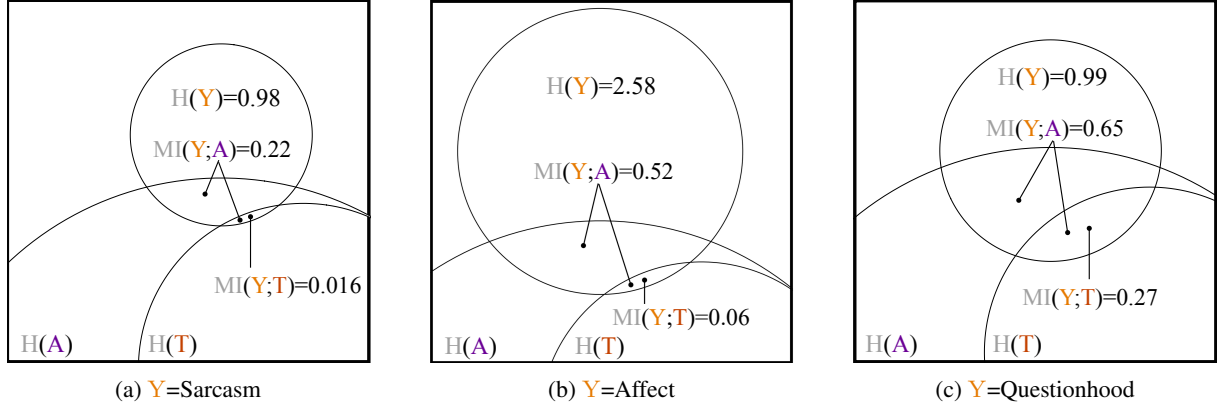


Figure 2: Proportional-area diagrams for sarcasm (left), affect (middle), and questionhood (right) features. In all plots, A = Audio and T = Text. In the case of multiple arrows, the quantity refers to the sum of the areas indicated.

stone, 2015; Busso et al., 2017; Lotfian and Busso, 2019; Vidal et al., 2020). We select the MSP-Podcast dataset (version 1.12) released by Lotfian and Busso (2019) due to its naturalistic sources, speaker diversity, and large size. This dataset formulates affect recognition as a classification task over 10 emotional categories. We refer the readers to Busso et al. (2025) for more details regarding the categories and their distributions in the latest version of the dataset, which was released subsequent to the completion of our work.

Questionhood Classification Questions in English can be signaled by both specific prosodic contours and by a unique syntactic form. In American English, yes-no questions are typically marked by rising pitch at the end, while *wh*-questions are marked by falling pitch (Pierrehumbert and Hirschberg, 1990). In text, questions are usually marked by subject auxiliary inversion (e.g., Are you coming?), and *wh*-questions additionally contain a question word at the beginning of the sentence. Thus, prosody and syntax convey significantly redundant information about whether an utterance is a question. An exception is echo questions, such as *You’re leaving already?*, where the prosody is the only cue to questionhood (after removing punctuation). Consequently, we predict that prosody will add comparatively little information in addition to text when determining the questionhood of an utterance.

We formulate questionhood classification as a binary classification task, classifying utterances as questions or non-questions (assertions, imperatives, etc.).⁷ The data is sampled from the MSP-Podcast

dataset (Lotfian and Busso, 2019). We segment the transcripts by sentence using spaCy (Honnibal et al., 2020), align the audio using the Montreal Forced Aligner (McAuliffe et al., 2017), and remove all utterances that are shorter than 2 seconds. Any utterance with a transcript ending in a question mark is labeled as a question, and all remaining utterances are labeled as non-questions. Finally, we downsample non-questions so that the dataset contains an approximately equal number of questions and non-questions.⁸ The train, development, and test splits have 13842, 1538, and 3845 samples, respectively. All occurrences of [.,?] are removed prior to training.

5.2 Experimental Setup

As described in Section 4, we estimate conditional entropy $H(Y | X)$ as the cross-entropy of a classifier trained to predict Y given X . We train classifiers by fine-tuning pretrained Transformer models. We base our code on the Transformers library pipeline Wolf et al. (2019). We experiment with four different model sizes for each model family and update all model parameters during training. We fine-tune the model using the cross-entropy loss and Adam optimizer (Kingma and Ba, 2014). We train for a fixed number of epochs and select the checkpoint that has the lowest loss on our development split. To select hyperparameters, we perform a sweep of 20 runs for each model size and select the top model based on the test loss value. For sarcasm detection, we implement 5-fold cross-validation due to the dataset’s small size. For

classes to interpret the results easily.

⁸We also downsample non-questions in such a way that the distribution of durations remains the same as that of the questions.

⁷This task is related to dialogue act classification (e.g., Stolcke et al., 2000), but we adopt a vastly simplified set of

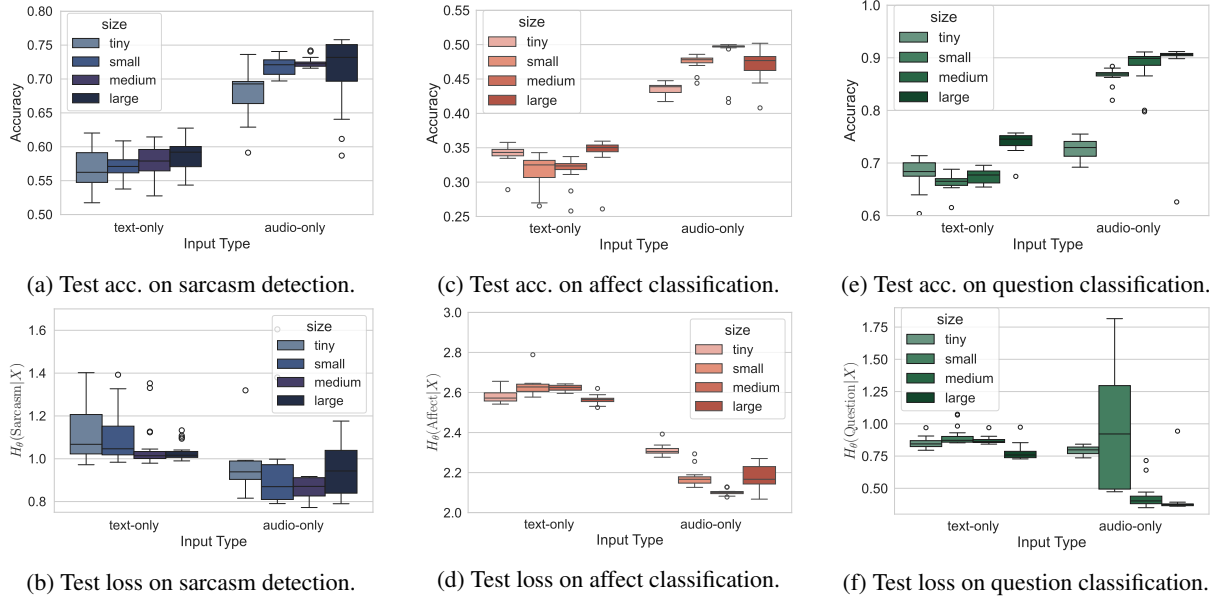


Figure 3: Test set performance on sarcasm (left) and affect (middle), and questionhood (right) classification tasks.

affect and questionhood classification, we use train, development, and test splits. More information on our hyperparameters is provided in Appendix A.

Text-only models For the text-only models, we fine-tune the GPT-2 suite of models (Radford et al., 2019) on the transcript portion of the data for each task. We use small, medium, large, and xl models.⁹ We train a classification head on top of this model. For sarcasm detection we fine-tune the entire model, and for affect and questionhood classification, we use low-rank adaptation (LoRA) (Hu et al., 2022) to reduce training time.

Audio-only models For our audio-only models, we use Whisper (Radford et al., 2022) and wav2vec 2.0 (Baevski et al., 2020). Whisper is a state-of-the-art multilingual speech-to-text model based on a Transformer (Vaswani et al., 2017) encoder-decoder architecture. The encoder takes the log mel spectrogram as input and the decoder uses text tokens as input and output. Since our tasks involve classifying an audio recording, as opposed to generating text, we fine-tune only the *encoder* portion to get the audio representations to pass onto a classification layer. We use Whisper tiny, small, medium, and large sized models. We also experiment with wav2vec (Baevski et al., 2020) for fine-tuning on audio tasks, as it is pretrained on an audio-only masking objective, unlike Whisper,

⁹For consistency in naming conventions across modalities, we subsequently refer to these models as tiny, small, medium, and large, respectively.

which is trained mainly to perform speech recognition on paired text-audio data. Finally, although we assume that the audio \mathbf{A} contains all the information about the text \mathbf{T} , in case audio models in practice fail to encode some information about \mathbf{T} , we conduct additional experiments fine-tuning audio+text models accepting both audio and text input. These models are described in Appendix B.

6 Results

Information Quantification The information diagrams for sarcasm, affect, and questionhood classification are shown in fig. 2. They show proportional areas for $H(\mathbf{Y})$, $MI(\mathbf{Y}; \mathbf{A})$, and $MI(\mathbf{Y}; \mathbf{T})$. The regions $H(\mathbf{A})$ and $H(\mathbf{T})$ are not estimated and are purely illustrative.

For both sarcasm and affect, we observe that the mutual information between the feature of interest and the audio $MI(\mathbf{Y}; \mathbf{A})$ is larger than its mutual information with the text only $MI(\mathbf{Y}; \mathbf{T})$ by over an order of magnitude. This suggests that the audio contains substantial information about both of these features above and beyond the text, both confirming our intuitions and corroborating previous findings (Tepperman et al., 2006; Cheang and Pell, 2008). While $MI(\mathbf{Y}; \mathbf{A})$ values for sarcasm and affect are quite different, we note that the uncertainty coefficients $\frac{MI(\mathbf{Y}; \mathbf{A})}{H(\mathbf{Y})}$ (Press et al., 2007) are quite similar – 0.22 and 0.20 respectively – telling us that the audio signal conveys a similar proportion of the total information in these features.

By contrast, for questionhood, $MI(Y; A)$ is only about 2.4 times greater than $MI(Y; T)$. This aligns with our prediction that prosody would provide relatively less additional information compared to other tasks, reflecting the fact that questions are often indicated by syntax, in addition to prosody.

We also estimate the information that prosody *uniquely* conveys about sarcasm, affect, and questionhood, $MI(Y; P | T)$, as discussed in Section 3.2. Subtracting the mutual information of the features with the text from its mutual information with the audio provides this estimate. This quantity is 0.20 and 0.46 bits for sarcasm and affect, respectively, making up a large fraction of the $MI(Y; A)$. It is 0.38 bits for questionhood, making up a smaller portion of the $MI(Y; A)$.

We observe a large difference in $H(Y)$ for sarcasm and questionhood on the one hand and affect on the other, where for affect it is larger. While this matches intuitions – affect is inherently higher-dimensional than sarcasm or questionhood – this is also a consequence of the labeling scheme: The sarcasm and questionhood datasets have binary classes, while the affect dataset has ten classes.

Classifier Performance We find several patterns of interest in the performance of the classifiers trained to make the above estimates. Since cross-entropy provides an upper bound on entropy, i.e., $H_{\theta}(Y | X) \geq H(Y | X)$, we use the lowest-loss runs to estimate mutual information values in fig. 2. For $MI(Y; T)$, we select the best text-only model run across all model sizes, and for $MI(Y; A)$, we use the lowest loss run among the audio-only models and sizes.

We report the distribution of losses and accuracies for GPT-2 and Whisper models over 20 runs in fig. 3. To calculate the accuracy for one sarcasm run, we take the average accuracy across all 5-folds for each run. In the Appendix, we report additional results for wav2vec 2.0 (fig. 4) and for audio+text models (fig. 5). First, we find that our classifiers have strong performance relative to baselines: Our best sarcasm classifier shows a 10-point improvement over the baseline released by Castro et al. (2019), and our best emotion classifier is competitive with other models reported by Gonçalves et al. (2024, Table 5).

Second, we observe that the figures show several similar trends: Larger audio-only models tend to have stronger performance (though Whisper large does not consistently outperform Whisper

medium), while model size has a limited impact on the performance of text-only models, though exceptions exist. In all cases, audio-only models have higher accuracy than the text-only models. In fig. 4 and fig. 5, we see that audio-only models based on Whisper medium or -large have stronger performance than wav2vec 2.0 and audio+text models.

7 Discussion and Conclusion

The approach we introduce in Section 3.2 enables us to characterize how a particular kind of information is distributed over multiple communication channels. As case studies, we investigated how information about sarcasm, affect, and questionhood are conveyed through prosody and text. Our results provide nuance on the importance of prosody for these tasks: We find that the amount of information conveyed by the audio signal can be over an order of magnitude larger than the information conveyed by text alone, but this is highly dependent on the feature of interest. This result highlights the importance of treating language as a multi-channel communication system, and suggests that important properties of language will be overlooked if it is analyzed only in one channel (i.e., text) as is common.

An important caveat is that we restrict both audio and text to a single sentence. If we define T' and P' to be random variables over the text and prosody for the rest of the discourse, it stands to reason that $MI(Y; T, T')$ would be significantly larger than $MI(Y; T)$ for both sarcasm and affect, i.e., the larger context adds significantly more information about the feature of interest than one can get from the sentence alone. Indeed, Castro et al. (2019) show that previous utterance information improves sarcasm detection. On the other hand, $MI(Y; P, P')$ might be similar to $MI(Y; P)$, as prior work suggests that the unique information conveyed by prosody, relative to its past text, is mostly local (Regev et al., 2025). Thus, our results show that while *local* textual information is often insufficient to judge sarcasm or affect, *local prosody* often is sufficient. This conclusion has implications for discourses lacking substantial textual context: For example, long-term context may not be available for a person reading a short text message or for an automated speech-to-text customer service system responding to a customer over the phone. A fine-grained comparison of the information contributions of text and prosody in

short and long contexts is a promising direction for future work.

This work’s main contribution is to introduce a general framework, which future work can extend to other domains of meaning (Y), communicative channels, and languages. Regarding other domains of meaning, there is an extensive prior literature on the use of prosody for syntactic disambiguation (Selkirk, 2011; Pauker et al., 2011) and word boundary detection (Cutler, 2014), with important implications for language acquisition (Soderstrom et al., 2003). Prosody also provides important cues about turn-taking (Ruiter et al., 2006; Gravano and Hirschberg, 2011; Cutler and Pearson, 2018) and back-channels (Clark et al., 2025).

Other communicative channels that one might extend the approach to are specific prosodic features and combinations thereof. This will require architectural and representational innovations, as it will be necessary to train classifiers that take prosodic representations as input. Using vision language models (e.g., Radford et al., 2021), it will also be possible to investigate information conveyed through *visual* channels, such as facial expressions and hand gestures.

Finally, whereas we limit the present study to English, future work should compare a more diverse set of languages. Recent work (Wilcox et al., 2025) has suggested that prosodic typologies can be investigated from an information-theoretic lens. Our approach can provide a finer-grained technique to identify how typologically distinct languages use the prosodic channel to convey different aspects of linguistic information.

8 Limitations

In Section 3.2, we argue that ⑨ $MI(Y; A | T)$ serves as a reasonable approximation of ② $MI(Y; P | T)$, since the only information captured by ⑨ but not by ② reflects mostly irrelevant factors such as speaker identity or background noise. However, to more precisely estimate the contribution of the prosodic channel to a particular linguistic phenomenon – and, by extension, to human communication – we would ideally measure $MI(Y; P | T)$ directly. This would also allow us to estimate two additional quantities of interest: (1) the unique contribution of text to Y, $MI(Y; T | P)$, and (2) the redundant contributions of text and prosody to Y, $MI(Y; T; P)$. However, current spoken language models do not accept only prosodic features as in-

put, making this estimation impossible. We leave it to future work to tackle the substantial problem of designing a new architecture that can take representations of isolated prosodic features as input, enabling us to measure all of these quantities directly, either for individual prosodic features such as pitch, or the entire prosodic channel.

It is important to acknowledge that our feature estimates are heavily influenced by idiosyncrasies of the languages, datasets, and models we study. As we study these tasks in English only, our specific quantitative findings may not generalize to typologically different languages such as tone or pitch accent languages. For sarcasm detection, we rely on a balanced dataset in which sarcastic and non-sarcastic utterances are equally represented. However, this setup does not mimic our real-world human conversations – sarcastic utterances are relatively rare. In the case of affect recognition, we work in a highly constrained setting where each utterance is labeled with a single emotion. However, a single utterance may convey multiple emotions simultaneously or to varying degrees. Moreover, the dataset includes only ten discrete emotions – a small subset of the full range emotions that humans can express. While these specifics may arguably have relatively little impact on our estimates of ratio of $MI(Y; A)$ and $MI(Y; T)$, they strongly affect the magnitude of our estimates, as more imbalanced labels will lead to lower unconditional entropy, and more fine-grained labeling schemes will lead to higher unconditional entropy.

Finally, we acknowledge that we do not fine-tune the most powerful text LLMs available. Fine-tuning models larger than GPT-2 XL could lead to better estimations of $MI(Y; T)$. However, we opt not to work with these models due to the computational demands of fine-tuning. While it is possible that Whisper large would achieve state-of-the-art performance on these tasks, there is rapid progress in spoken language modeling, meaning better estimates may be achievable in the future. We leave it to future work to more thoroughly investigate the impact of the specific model choice on MI estimation.

Acknowledgments

We thank Sarenne Wallbridge for our early conversations about this paper. This work used resources available through the National Research Platform (NRP) at the University of California, San

Diego (Weitzel et al., 2025). NRP has been developed, and is supported in part, by funding from National Science Foundation, from awards 1730158, 1540112, 1541349, 1826967, 2112167, 2100237, and 2120019, as well as additional funding from community partners.

References

- Gavin Abercrombie and Dirk Hovy. 2016. [Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of Twitter conversations](#). In *Proceedings of the ACL 2016 Student Research Workshop*, pages 107–113, Berlin, Germany. Association for Computational Linguistics.
- Aristotle. 1991. *On Rhetoric: A Theory of Civic Discourse*, 2nd edition. Oxford University Press.
- Siddhant Arora, Ankita Pasad, Chung-Ming Chien, Jionghao Han, Roshan Sharma, Jee-weon Jung, Hira Dhamyal, William Chen, Suwon Shon, Hung-yi Lee, et al. 2024. On the evaluation of speech foundation models for spoken language understanding. *arXiv preprint arXiv:2406.10083*.
- Arun Babu, Changan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.
- Tanja Bänziger, Marcello Mortillaro, and Klaus R. Scherer. 2012. [Introducing the geneva multimodal expression corpus for experimental research on emotion perception](#). *Emotion*, 12 5:1161–79.
- Kaushal Santosh Bhogale, Sai Sundaresan, Abhigyan Raman, Tahir Javed, Mitesh M Khapra, and Pratyush Kumar. 2023. Vistaar: Diverse benchmarks and training sets for indian language asr. *arXiv preprint arXiv:2305.15386*.
- Christopher M Bishop. 1994. Mixture density networks.
- Mara Breen, Evelina Fedorenko, Michael Wagner, and Edward Gibson. 2010. Acoustic correlates of information structure. *Language and cognitive processes*, 25(7-9):1044–1098.
- Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. 2005. [A database of german emotional speech](#). In *Interspeech*.
- C. Busso, S. Parthasarathy, A. Burmania, M. Abdel-Wahab, N. Sadoughi, and E. Mower Provost. 2017. [MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception](#). *IEEE Transactions on Affective Computing*, 8(1):67–80.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. [Iemocap: interactive emotional dyadic motion capture database](#). *Language Resources and Evaluation*, 42:335–359.
- Carlos Busso, Reza Lotfian, Kusha Sridhar, Ali N Salman, Wei-Cheng Lin, Lucas Goncalves, Srinivas Parthasarathy, Abinay Reddy Naini, Seong-Gyun Leem, Luz Martinez-Lucas, et al. 2025. [The MSP-Podcast Corpus](#). *arXiv preprint arXiv:2509.09791*.
- Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. [Towards multimodal sarcasm detection \(an _Obviously_ perfect paper\)](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.
- Wallace Chafe. 1988. Punctuation and the prosody of written language. *Written communication*, 5(4):395–426.
- Henry S. Cheang and Marc D. Pell. 2008. [The sound of sarcasm](#). *Speech Communication*, 50(5):366–381.
- Thomas Hikaru Clark, Moshe Poliak, Tamar Regev, Amanda J Haskins, Edward Gibson, and Caroline Robertson. 2025. [The relationship between surprisal, prosody, and backchannels in conversation reflects intelligibility-oriented pressures](#). *PsyArXiv preprints*.
- Jennifer Cole. 2015. Prosody in context: a review. *Language, Cognition and Neuroscience*, 30:1–31.
- Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.
- Anne Cutler. 2014. [Prosody and the word boundary problem](#). In *Signal to syntax*, pages 87–99. Psychology Press.
- Anne Cutler and Mark Pearson. 2018. [On the analysis of prosodic turn-taking cues](#). In *Intonation in discourse*, pages 139–156. Routledge.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

- Megan L. Dress, Roger J. Kreuz, Kristen E. Link, and Gina M. Caucci. 2008. [Regional variation in the use of sarcasm](#). *Journal of Language and Social Psychology*, 27(1):71–85.
- Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. 2024. A survey of multimodal sarcasm detection. *arXiv preprint arXiv:2410.18882*.
- Raul Fernandez and Rosalind Picard. 2011. [Recognizing affect from speech prosody using hierarchical graphical models](#). *Speech Communication*, 53(9):1088–1103. Sensing Emotion and Affect - Facing Realism in Speech Processing.
- Elena Filatova. 2012. [Irony and sarcasm: Corpus generation and analysis using crowdsourcing](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 392–398, Istanbul, Turkey. European Language Resources Association (ELRA).
- Robert W Frick. 1985. Communicating emotion: The role of prosodic features. *Psychological bulletin*, 97(3):412.
- Richard Futrell, Edward Gibson, and Roger P Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive science*, 44(3):e12814.
- Lucas Goncalves, Ali N Salman, Abinay R Naini, Laureano Moro Velazquez, Thomas Thebaud, Leibny Paola Garcia, Najim Dehak, Berrak Sisman, and Carlos Busso. 2024. [Odyssey 2024-speech emotion recognition challenge: Dataset, baseline framework, and results](#). In *The Speaker and Language Recognition Workshop (Odyssey 2024)*, pages 247–254.
- Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634. Publisher: Elsevier.
- Henk Haverkate. 1990. [A speech act analysis of irony](#). *Journal of Pragmatics*, 14(1):77–109.
- 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.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength natural language processing in python](#).
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Larry M. Hyman. 2008. [Universals in phonology](#). *The Linguistic Review*, 25(1-2):83–137.
- Tahir Javed, Sakshi Joshi, Vignesh Nagarajan, Sai Sundaresan, Janki Nawale, Abhigyan Raman, Kaushal Bhogale, Pratyush Kumar, and Mitesh M Khapra. 2023. Svarah: Evaluating english asr systems on indian accents. *arXiv preprint arXiv:2305.15760*.
- Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, and Mark James Carman. 2016. [Harnessing sequence labeling for sarcasm detection in dialogue from tv series ‘friends’](#). In *Conference on Computational Natural Language Learning*.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Gunther Kress. 2009. *Multimodality: A social semiotic approach to contemporary communication*. Routledge.
- Roger Kreuz and Gina M. Caucci. 2007. [Lexical influences on the perception of sarcasm](#).
- Pauline Larrouy-Maestri, David Poeppel, and Marc D Pell. 2024. The sound of emotional prosody: Nearly 3 decades of research and future directions. *Perspectives on Psychological Science*, page 17456916231217722.
- Roger Levy. 2008. [Expectation-based syntactic comprehension](#). *Cognition*, 106(3):1126–1177.
- R. Lotfian and C. Busso. 2019. [Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings](#). *IEEE Transactions on Affective Computing*, 10(4):471–483.
- Tomoko Matsui, Tagiru Nakamura, Akira Utsumi, Akihiro T. Sasaki, Takahiko Koike, Yumiko Yoshida, Tokiko Harada, Hiroki C. Tanabe, and Norihiro Sadato. 2016. [The role of prosody and context in sarcasm comprehension: Behavioral and fmri evidence](#). *Neuropsychologia*, 87:74–84.
- Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. [Montreal forced aligner: Trainable text-speech alignment using kaldi](#). In *Interspeech*, volume 2017, pages 498–502.

- Sylvie Mozziconacci. 2002. Prosody and emotions. In *Speech prosody*, volume 2002, pages 1–9.
- Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure FP Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, et al. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.
- Pritty Patel-Grosz, Salvador Mascarenhas, Emmanuel Chemla, and Philippe Schlenker. 2023. Super linguistics: an introduction. *Linguistics and Philosophy*, 46(4):627–692.
- Efrat Pauker, Inbal Itzhak, Shari R. Baum, and Karsten Steinhauer. 2011. [Effects of cooperating and conflicting prosody in spoken english garden path sentences: Erp evidence for the boundary deletion hypothesis](#). *Journal of Cognitive Neuroscience*, 23(10):2731–2751.
- Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. [Word lengths are optimized for efficient communication](#). *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.
- Janet Pierrehumbert and Julia Hirschberg. 1990. [The meaning of intonational contours in the interpretation of discourse](#). In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*. The MIT Press.
- Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. *arXiv preprint arXiv:2004.03061*.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.
- William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. 2007. Numerical recipes 3rd edition. *Cambridge: New York*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). *arXiv preprint*.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Rachel Rakov and Andrew Rosenberg. 2013. [“sure, i did the right thing”: a system for sarcasm detection in speech](#). In *Interspeech 2013*, pages 842–846.
- Tamar I Regev, Chiebuka Ohams, Shaylee Xie, Lukas Wolf, Evelina Fedorenko, Alex Warstadt, Ethan G Wilcox, and Tiago Pimentel. 2025. The time scale of redundancy between prosody and linguistic context. *arXiv preprint arXiv:2503.11630*.
- Timo B Roettger, Tim Mahrt, and Jennifer Cole. 2019. Mapping prosody onto meaning—the case of information structure in american english. *Language, Cognition and Neuroscience*, 34(7):841–860.
- Jan-Peter De Ruiter, Holger Mitterer, and N. J. Enfield. 2006. [Projecting the End of a Speaker’s Turn: A Cognitive Cornerstone of Conversation](#). *Language*, 82(3):515–535.
- Frank A. Russo and Steven R. Livingstone. 2015. [The ryerson audio-visual database of emotional speech and song](#).
- Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.
- Elisabeth Selkirk. 2011. The syntax-phonology interface. *The handbook of phonological theory*, pages 435–484.
- Miraj Shah, David G. Cooper, Houwei Cao, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 2013. [Action unit models of facial expression of emotion in the presence of speech](#). In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 49–54.
- Claude E. Shannon. 1948. [A mathematical theory of communication](#). *Bell Syst. Tech. J.*, 27:623–656.
- Claude E Shannon. 1951. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64.
- Simon J Sheather. 2004. Density estimation. *Statistical science*, pages 588–597.
- Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan Sharma, Wei-Lun Wu, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. 2022a. Slue phase-2: A benchmark suite of diverse spoken language understanding tasks. *arXiv preprint arXiv:2212.10525*.
- Suwon Shon, Kwangyoun Kim, Yi-Te Hsu, Prashant Sridhar, Shinji Watanabe, and Karen Livescu. 2024. Discreteluu: A large language model with self-supervised discrete speech units for spoken language understanding. *arXiv preprint arXiv:2406.09345*.
- Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J. Han. 2022b. [Slue: New benchmark tasks for spoken language](#)

- understanding evaluation on natural speech. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7927–7931.
- Jesse Snedeker and John Trueswell. 2003. Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, 48(1):103–130.
- Michaela Socolof, Jacob Louis Hoover, Richard Futrell, Alessandro Sordoni, and Timothy J. O’Donnell. 2022. [Measuring morphological fusion using partial information decomposition](#). In *Proceedings of the 29th international conference on computational linguistics*, pages 44–54, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Melanie Soderstrom, Amanda Seidl, Deborah G Kemler Nelson, and Peter W Jusczyk. 2003. The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language*, 49(2):249–267.
- Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. [Dialogue act modeling for automatic tagging and recognition of conversational speech](#). *Computational Linguistics*, 26(3):339–373.
- Bashar Talafha, Abdul Waheed, and Muhammad Abdul-Mageed. 2023. N-shot benchmarking of whisper on diverse arabic speech recognition. *arXiv preprint arXiv:2306.02902*.
- Joseph Tepperman, David R. Traum, and Shrikanth S. Narayanan. 2006. ["yeah right": Sarcasm recognition for spoken dialogue systems](#). In *Interspeech*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- A. Vidal, A. Salman, W.-C. Lin, and C. Busso. 2020. [MSP-face corpus: A natural audiovisual emotional database](#). In *ACM International Conference on Multimodal Interaction (ICMI 2020)*, pages 397–405, Utrecht, The Netherlands.
- Derek Weitzel, Ashton Graves, Sam Albin, Huijun Zhu, Frank Wuerthwein, Mahidhar Tatineni, Dmitry Mishin, Elham Khoda, Mohammad Sada, Larry Smarr, Thomas DeFanti, and John Graham. 2025. [The national research platform: Stretched, multi-tenant, scientific kubernetes cluster](#). In *Practice and Experience in Advanced Research Computing 2025: The Power of Collaboration*, PEARC ’25, New York, NY, USA. Association for Computing Machinery.
- Ethan Gotlieb Wilcox, Cui Ding, Giovanni Acampa, Tiago Pimentel, Alex Warstadt, and Tamar I Regev. 2025. Using information theory to characterize prosodic typology: The case of tone, pitch-accent and stress-accent. *arXiv preprint arXiv:2505.07659*.
- Adina Williams, Ryan Cotterell, Lawrence Wolf-Sonkin, Damián Blasi, and Hanna Wallach. 2021. [On the relationships between the grammatical genders of inanimate nouns and their co-occurring adjectives and verbs](#). *Transactions of the Association for Computational Linguistics*, 9:139–159.
- Lukas Wolf, Tiago Pimentel, Evelina Fedorenko, Ryan Cotterell, Alex Warstadt, Ethan Wilcox, and Tamar Regev. 2023. [Quantifying the redundancy between prosody and text](#). *Preprint*, arXiv:2311.17233.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Hyperparameters

We use Bayesian hyperparameter optimization to run our sweeps, optimizing to minimize the test loss. For all experiments, we search for the optimal learning rate using a log-uniform distribution. In the rest of this section, we share the task-specific hyperparameter choices we have made and the compute used.

Affect Classification Table 1 shows the hyperparameters used for this task. All experiments use cosine learning rate decay for smooth convergence and better generalization. Our preliminary experiments on audio-text models show that our choice of batch size and number of epochs does not affect task performance. Therefore, we choose values that optimize training time.

We use V100 GPUs and A100 GPUs to perform our text-only and audio-only affect recognition sweeps respectively. To conduct our audio-text tiny, small, and medium affect recognition model sweeps we use 8 A10, 4 A6000, and 2 A100 GPUs respectively. Our text-only model sweeps took less than 2 days each. Our audio-only sweeps took 3 to 7 days depending on the model size. Finally, our audio-text sweeps took 5 to 10 days depending on the model size.

Hyperparameters	text-only	audio-only	audio-text
Epochs	5, 10, 15	5, 10, 15	5
Batch Size	8, 16, 32, 64	8, 16, 32, 64	64
Weight Decay	0, 0.01, 0.1	0, 0.01, 0.1	0, 0.01, 0.1
Learning Rate	[1e-6, 1e-4]	[1e-6, 1e-4]	[1e-6, 1e-4]
LoRA r	4, 8, 16	–	–
LoRA α	8, 16, 32	–	–

Table 1: Affect Classification Hyperparameters

Sarcasm Detection Table 2 shows the hyperparameters used for this task. We use A100 GPUs to train all our sarcasm detection models. No sweep across all modalities and sizes ran for more than 2 days.

Hyperparameters	All Models
Epochs	5, 10, 15
Batch Size	8, 16, 32, 64
Weight Decay	0, 0.01, 0.1
Learning Rate	[1e-6, 1e-4]

Table 2: Sarcasm Detection Hyperparameters

Questionhood Classification With the exception of Whisper large, Table 3 shows the hyperparameters used for all the models for this task. For Whisper large, to maximize the usage of the available GPU VRAM and thereby reduce the training time, we run our sweep on the following (effective) batch sizes: 16, 32, 64, 128. Our preliminary experiments on Whisper large show that our choice of number of epochs does not affect task performance. Therefore, we limit the number of epochs for this model to be 5. The rest of the hyperparameters are the same as the other models. We use A100 and V100 GPUs to train all our questionhood classification models. No sweep across all modalities and sizes ran for more than 3 days.

Hyperparameters	text-only	audio-only
Epochs	5, 10, 15	5, 10, 15
Batch Size	8, 16, 32, 64	8, 16, 32, 64
Weight Decay	0, 0.01, 0.1	0, 0.01, 0.1
Learning Rate	[1e-6, 1e-4]	[1e-6, 1e-4]
LoRA r	4, 8, 16	–
LoRA α	8, 16, 32	–

Table 3: Questionhood Classification Hyperparameters

B Audio+Text models

We conduct additional experiments with models that accept both audio and text input for sarcasm and affect classification. The motivation is that, while the audio [A](#) theoretically contains all the information about the text [T](#), audio models may fail to encode some of this information. If this is the case, we can mitigate this problem by inputting the full text in addition to the audio.

For our audio+text models, we use representations from the encoder part of the Whisper model to get audio embeddings and representations from the GPT-2 model to get the text embeddings. We concatenate both these embeddings and use the combination as input to a classification layer. We choose this architecture so that the results would be easily comparable to the audio-only and text-only models. We pair GPT-2 small and Whisper tiny, GPT-2 medium and Whisper small, and GPT-2 large and Whisper medium. We refer to these pairs as audio+text tiny, small, and medium respectively.

In fig. [5d](#), we present losses and accuracies for models.

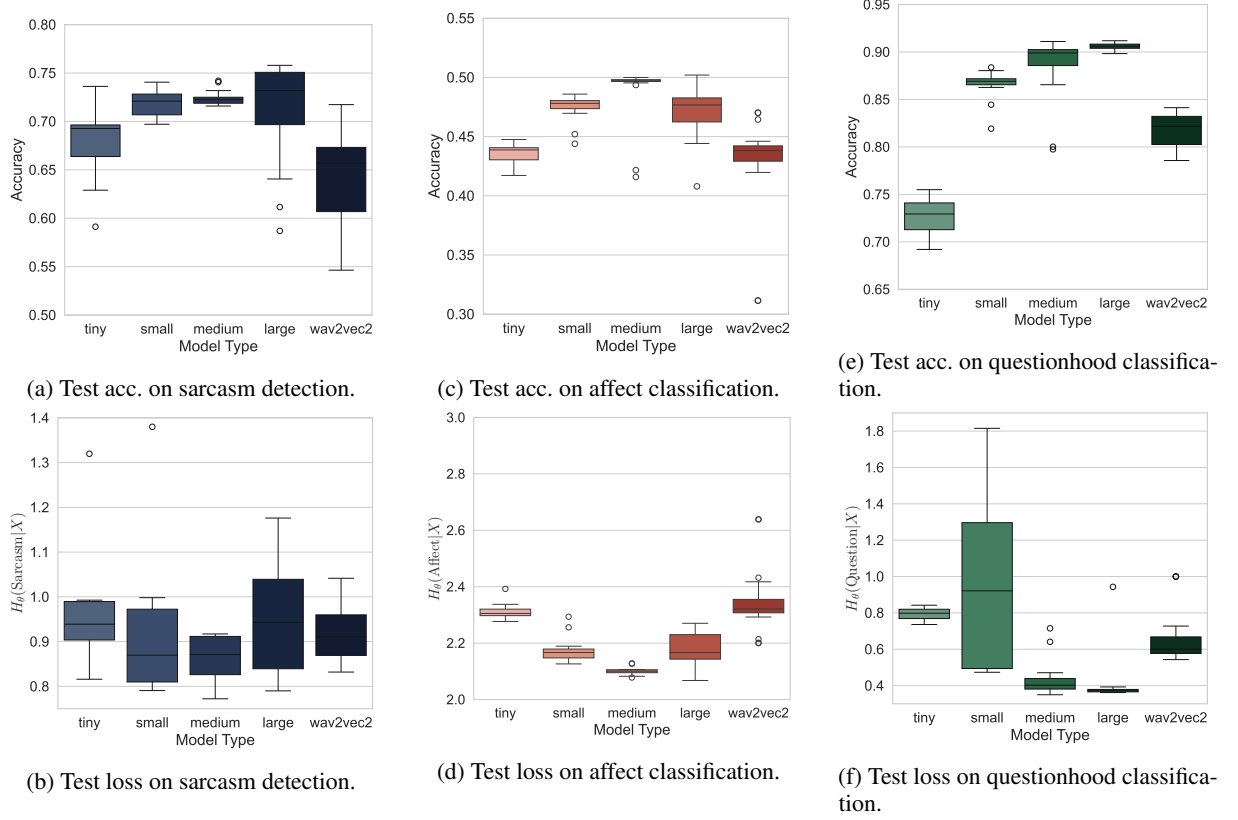


Figure 4: audio-only (Whisper + wav2vec) performance on sarcasm (left), affect (middle), and questionhood (right) classification

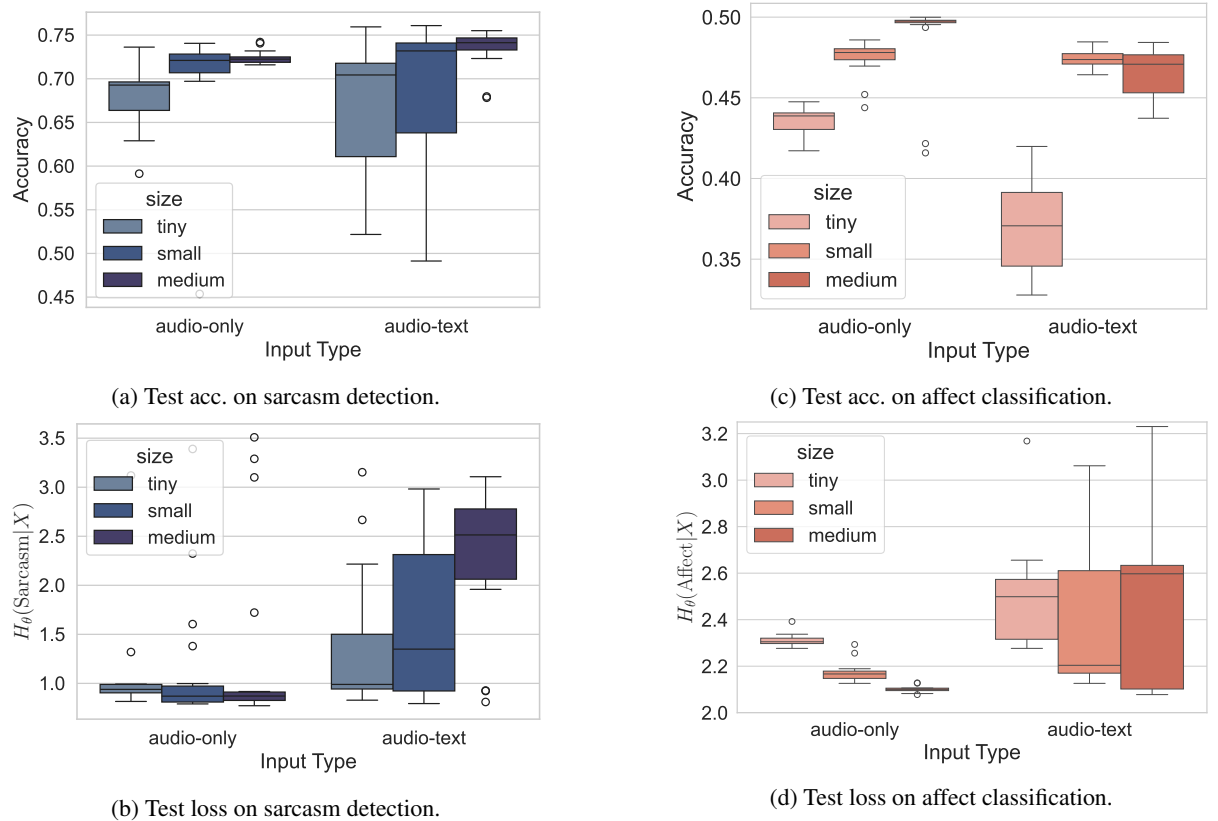


Figure 5: audio-only and audio-text performance on sarcasm (left) and affect (right)