Spring 6-11-2023

Say That Again: The role of multimodal redundancy in communication and context

Brandon Javier Dormes
Dartmouth College, brandon.j.dormes.23@dartmouth.edu

Say That Again:

The role of multimodal redundancy in communication and context

Brandon Dormes

Advised by Mark Thornton and Landry Bulls

Cognitive Science Senior Honors Thesis

Dartmouth College

22 May 2023

Acknowledgements

I am deeply indebted to my advisor and my mentor, Mark Thronton and Landry Bulls, who have offered generous amounts of both encouragement and direction. Without their insights and patience, this work would be nothing but scattered notes on a legal pad. Instead, a vague idea became smudged whiteboards, code, drafts, headaches, revelations, and results. Their belief in me far outstripped my own, and I cannot thank them enough.

I would also like to thank Chujun Lin, Amisha Vyas, and Lindsey Tepfer, all of whom have lent their technical and professional expertise to this project. As a whole, the SCRAP Lab has demonstrated the power of purposeful, compassionate curiosity. Their dedication to improving the experience of undergraduate research has been a tremendous boon to my time at Dartmouth.

Finally, I would like to thank my friends who kept me sane by listening to monologues on entropy, offering support in my most challenging moments, and saving me from my own laptop. I cherish the fact that this is not my work alone.

Abstract

With several modes of expression, such as facial expressions, body language, and speech working together to convey meaning, social communication is rich in redundancy. While typically relegated to signal preservation, this study investigates the role of cross-modal redundancies in establishing performance context, focusing on unaided, solo performances. Drawing on information theory, I operationalize redundancy as predictability and use an array of machine learning models to featurize speakers' facial expressions, body poses, movement speeds, acoustic features, and spoken language from 24 TEDTalks and 16 episodes of *Comedy Central Stand-Up Presents*. This analysis demonstrates that it is possible to distinguish between these performance types based on cross-modal predictions, while also highlighting the significant amount of prediction supported by the signals' synchrony across modalities. Further research is needed to unravel the complexities of redundancy's place in social communication, paving the way for more effective and engaging communication strategies.

1. Introduction

To do good science, researchers must often excise context. This is not a blind decision. It allows researchers to better explain causal patterns within their data, reducing the possible contributors to collected measures. Often, this is not because the surrounding measures are unimportant, but rather due to the feasibility of accounting for them. When a study is built on face stimuli, how can bangs be represented in the data? What of glasses, makeup, and setting? These features are meaningful, but they are difficult to capture and make sense of. Consequently, faces are hewn from their heads, sounds are canned and funneled through headphones, and pain is supplied by heating devices.

This is especially relevant to the field of social psychology, where subtle differences can color experiences and inferences. When surrounded by friends and preceded by some wry comment, the utterance "I can't stand you" might sound charming. In a stark, white room, with a completely even tone, the same phrase could be chilling. Turning to inference, one can imagine seeing a couple's argument through a diner window. The situation is easy to scan from a distance: shoulders are pinched back, their brows are furrowed, and their gestures are sharp — all without hearing a word of what is being said. Though, consider the same situation from a cook's perspective. Over the grill's sizzle, the tense atmosphere is still unambiguous. The cook would *hear* the argument's volume, tone, and choice of words to arrive at the same conclusion — all without seeing a thing. We can intuitively infer crucial information from multiple slices of the situation. Yet, nature is famously antagonistic to redundancy. The additional actions and coordination demand additional calories to support, and anything unnecessary ought to be hacked away. Why, then, is social communication so ridden with redundancy?

To interrogate this, I collect an array of data relevant to social communication at a high temporal resolution. I accomplish this by using a mix of traditional measures and machine

learning models to featurize and represent speakers' behavior from audiovisual recordings, frame-by-frame. Specifically, it estimates information concerning a speaker's facial expressions, their acoustic features, their use of language, and how they position and move their body. By collecting this spectrum of data, I also seek to support the study's central aim.

Motivated by information theory and literature from social psychology, I hypothesize that redundancy does more than mere signal preservation. Rather, it entails some part of the interaction's context. To interrogate this, I draw data from two sets of videos: TEDTalks and *Comedy Central Stand-Up Presents* performances. I operationalize redundancy as predictability, arguing that stable relationships and interpretations lie near the heart of redundancy. I then work to predict sets of signals across modalities, articulate the links between them, and investigate the differences between either context.

I begin with an overview of redundancy in information theoretic terms before moving to theories of multimodal communication. I then detail both the measures collected, the methods used to obtain them, and the motivations for my analysis. Lastly, I connect these findings to larger literature and argue that, as is often the case, the results are promising, but additional research is needed to fully describe redundancy's place in social communication.

1.1. Redundancy

Redundancy's origins in information theory are telling. The idea demands some subjective view on what a signal “means”. As a field, information theory arose with telecommunications. To cross long distances, messages needed a systematic way of being both reduced and interpreted. Its early framework was marshaled by four core concepts, namely senders, mediums, messages, and receivers. At the heart of it all is the bit: the smallest unit of information possible.

Take Paul Revere's ride and message: "One if by land, two if by sea." Gleick (2011) identifies this as a resonant example of how information was rearranged and communicated in the early days of telecommunication. Revere, in essence, communicated a single, binary bit, and it offers only the ability to discriminate between two possibilities. Though on any given night, an individual could have ridden with as few or as many lanterns as they would have liked without rousing a small army. What made the difference was the a priori agreement about how to interpret the signal. Paul Revere was the sender; the horse and lanterns were the medium; minutemen were the recipients; British approach was the message.

Suppose though, that Revere needed to plan for the possibility that the British would, in an anachronistic turn, be arriving by air. In this hypothetical, Revere's current method would be insufficient, as it only allows recipients to discriminate between two possibilities. To extend the space of messages which Revere could transmit, he could include stipulations for a *second* ride. In the first, a single lantern would alert the army to an attack by air, and no further rides would be needed. Seeing two lanterns, though, would alert the observer to a *set* of possibilities, and the subsequent ride would follow the structure of the historical one: one if by land, two if by sea. While this would be perfectly sufficient for overcoming the complications in this hypothetical, it introduces a new fold with theoretical implications.

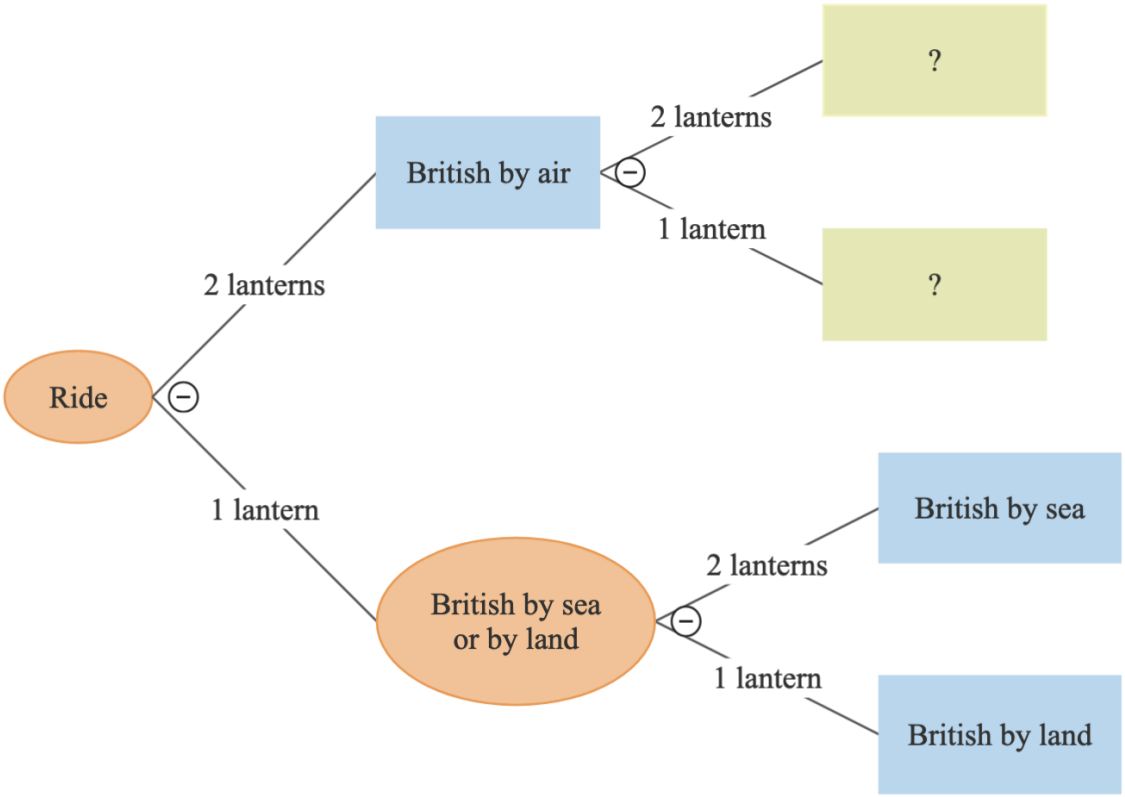


Figure 1. A diagram of Paul Revere's hypothetical ride(s), where the state of a bit (i.e. the lanterns) is transmitted in each ride. Orange circles indicate states where an additional ride is necessary, given the interpretation scheme. Blue squares represent states with established meanings. Green squares represent possible states which the scheme does not address. Note that each state can be represented by the sequence of lantern counts needed to reach that state. "British by sea" may be represented as [12], or scaling back the lantern counts by one: [01].

What to make of the phantom second ride? Within the scheme, people are already primed to observe two rides, and as a result, preparations for a second ride would be "wasted", given an aerial approach. To place this in information theoretic terms, Revere's unused bit would not allow its recipients to discriminate between any pair of possibilities. It would be redundant — a gap

between pragmatic necessities and optimal efficiency that introduces nothing new. In formal terms, this is rendered out as the formula:

$$R(X) = 1 - (H(X) / H_{max})$$

Where X is a sender which transmits sequences of bits according to some probability distribution, H(X) is the amount of information transmitted by the sender, and H_{max} is the theoretically greatest amount of information possible to transmit given the length of X's sequences. A sender which transmits the maximum amount of information will have a redundancy of 0. This is not intended to be a thorough account of the information theoretic construct, but a high-level description helps to demonstrate the relationship between information, a channel's limits, and redundancy. Traditionally, these redundant bits are used to brace and preserve the signals of the non-redundant bits. Hamming codes, for example, allow signals to self-correct when an error occurs in transmission.

Note again, though, that the salience of any message also depends on the interpretive scheme. That phantom ride could easily have been used to communicate whether Revere needed to buy more flour on his next grocery trip, but human constructs of relevance constrain the contents of the system. Redundancy reflects a highly theoretical relationship between what is maximally encodable and what content a system supports. This mode of thought dominated ideas about how information was packaged and transmitted, and they still remain as fixtures of telecommunications, semiotics, and computer science. However, Shannon (1948)'s foundational model was also accompanied by a fleet of assumptions and specific goals.

For instance, bits are not the *only* way to store information; the hypothetical Revere could have alerted the minutemen by using a third lantern. Bits' salience comes from the fact that almost any message can be represented in their form, not because all information *must* be

reduced to bits. It also assumes a unidirectional relationship between sender and receiver and a single channel through which information is introduced. Commissioned by Bell Labs, these theories sought commercial viability, scalability, and efficiency alongside scientific rigor. It succeeds in defining what is minimal: moving 0 from A to B, not in describing the full spectrum of rhetoric and communication.

One of the most aggressive assumptions made by Shannon's original model is that only one channel informs the message's identity. Placing this lens over human communication would irreparably distort the picture. It would imagine a brain which only attends to a single medium, and a single dimension within that medium. It might, for instance, look like an entire brain working only with the inputs from a singular cochlear hair cell. Meanwhile, the visual, tactile, olfactory, and residual auditory worlds glide past perception. This limitation carves out a meaningful boundary for the original model and situates it to respond best to information, not psychology.

1.2 Multimodal Communication

The multimodal world is messy, and it is flush with information. Though, making use of it demands violating the information theoretic framework. To maintain a throughline, I return to the auditory system. No single cell supplies the entire acoustic world. Groups of cells work in concert to provide that information. Together, they render complex, recognizable sounds. Songs, traffic, the incessant clicking of a pen — all grow from that collection of channels. These constructs and experiences are inaccessible to a lone receptor; multiple channels inform the construct's identity.

With multiple channels, we can also exploit regularities between channels to infer additional information. With an ear on either side of the head, we can extract the spatial location

of a sound's source. We compare the sound's onset across ears to infer this information, and a sound which hits the left ear before the right is likely to come from our left side (Middlebrooks & Green, 1991). This also imperils the use of the Shannon model by revealing multiple messages from the same signal.

However, with the introduction of two receptors of the same medium, redundancy seems to loom. In its first introduction, redundancy clung tightly to algorithm and deterministic rules. An unused bit in an exhaustively defined language that maps bits to meaning. However, humans have no such authoritative mapping. From birth, we must contend with ambiguity, form categories, and develop theories about the world. Lacking these clear associations between encodings and their meanings, this perspective on redundancy no longer holds. What may remain, however, is structural redundancy.

The patterns of low-level information encoded by sensory cells may be highly similar to the information encoded by another set of cells. In the case of binaural perception, we would say that one channel is highly redundant of the other: correcting for the offset and acoustic absorption, the signals should be the same (in most cases). This offers a slightly different perspective on redundancy: knowledge of what one stream’s contents may be used to predict the contents of another. When the coherence is high, so is the redundancy.

While unimodal redundancy relates to interpretations of signals within some scheme, multimodal redundancy can relate to both interpretation and structure between channels. Similarly, both forms of multimodal redundancy are able to help preserve signals, as in the unimodal case. Fire alarms produce visual and auditory signals, and this distribution makes their alarms perceptible even when one of the two modalities are unavailable, demonstrating interpretive redundancy. For those who can both see and hear the alarms, the signals’

synchronization allows them to understand the pair as having a unitary meaning, making use of structural redundancy. Signal preservation, however, is not redundancy's only purpose.

This reliance on multiple streams opens the door to parallelization. Rather than serially interpreting a stream of bits, as in the cochlear cell hypothetical, we can use multiple brain areas to simultaneously process the outputs of sensory cells. This theory, elaborated on by Rumelhart & McClelland (1986), arose from early connectionist literature. It also went on to spawn the philosophical strands which, today, support much of machine learning, and it finds support in a number of behavioral studies. Miller (1982) found that bimodal (audio-visual) signals were detected faster than either unimodal variant. Similar results are also borne out in user design and experience research. Oviatt (1995), in a task simulating the role of a real estate broker, found that 95% of users preferred interacting with a map multimodally, using both a pen interface and speech recognition to navigate. Additionally, those who used the multimodal system were about 10% faster in completing the task, and they produced 36% fewer errors. Politis et al. (2014) also found that, in driving simulations, participants responded faster when presented with multimodal signals, as compared to their unimodal constituents. Returning to social psychology, Holler et al. (2018) found that questions accompanied by a gesture saw shorter response times during freeform conversation. The conclusion that multimodal signals confer response time reductions, however, are not uncontested (Rubi & Stephens, 2016; Gielen et al. 1983). While multimodal signals do not always demonstrate these reductions, they are consistently less error-prone, and they do not increase reaction time. Though their interaction with detection speed is clear, these observations are constrained to the effect of multimodal signals on the listener. They do not, for instance, explore the role of multimodality in speaker choices.

1.3 Research Question and Aims

Multimodal signals clearly offer more than signal preservation. They provide faster response times, reduced errors, and the potential for an expanded space of interpretations. However, the role of multimodality in speaker choices remains underexplored. Following from the extended use of redundancy and multimodal signals, I seek to explore whether these concepts may allow for an additional inference: context. Given redundancy's ability to preserve signals, high redundancy may flag that the information is of greater importance.

To examine this, I work from two videosets: TEDTalks and *Comedy Central Stand-Up Presents*. Both center on solo performances, but with distinct aims. TEDTalks are educational and often center around an argument relating to technology, entertainment, design, business, or science. Their videos detail more abstract concepts and relationships. As a consequence, I hypothesize that these videos, on average, will demand more redundancy, reflecting the precision and complexity of its content. Stand-up videos, I suggest, will favor surprising combinations of signals, and their levels of redundancy will be lower.

To extract channels' contents, I make use of a spectrum of machine learning models. I capture information reflecting the speaker's facial expressions, body pose, speed of movements, acoustic features, and use of spoken language. These modalities will be used to predict one another, and the success of each will act as a proxy for the amount of cross-modal redundancy. As these models will serve as the proxies for human perception and extraction of features, I proceed to detail their architectures. This description serves to articulate each operations' assumptions and context while also motivating these specific models' selection for the present study.

2. Methodology

2.1 Data Collection and Processing

To perform this analysis, I collected two video sets: TEDTalks and *Comedy Central Stand-Up Presents* performances. Given the large number of TEDTalks available online, I randomly selected 24 from 2009 onwards, when 720p video became more commonplace online. Videos were only removed from the dataset if they did not have a 720p encoding. The corpus of Comedy Central Stand-Up videos, however, was more narrow, and so I merely collected 16 of the available videos.

This imbalance was driven by the assumption that TEDTalk videos would include more variation, and this may reduce the amount of recoverable measures from the data. Comedy Central videos, for instance, are more heavily produced and have more consistent editing, camera angles, and settings. Additionally, these videos are not unbroken performances, but 4 separate 'acts' during which performers deliver their sets. However, the analysis I apply, generalized linear model, is atemporal and agnostic to continuity between frames.

2.2 Featurization

Investigating the role of redundancy across modalities demands accurate measures of those modalities. Previous research into multi-channel communication, though, has struggled to collect precise estimates of ephemeral social signals. D’Mello & Graesser (2010) collected a similar spectrum of modalities to those proposed in the present study: facial action units, body pose, and conversational cues. Their collection scheme relied upon a tutoring computer program called AutoTutor, which used user inputs to track participant knowledge and probe for understanding.

While participants interacted with the system, users' faces were recorded by webcam before being manually rated, body pose was constrained to a seated position and measured with a pressure-sensitive mat, and conversational cues were derived from written responses to AutoTuror's questions. While rigorous, these schemes are difficult to scale, and the temporal resolution is constrained to the gaps between user input prompts. However, current-day machine learning models have made considerable progress in accuracy, featurization, and ease of use. In this section, I introduce four such models and their target data: Py-Feat for facial action units (Cheong et al., 2023), PARE for body pose (Kocabas et al., 2021), WhisperX for transcripts production (Bain et al., 2023), and a Homomorphic Projective Distillation transformer for semantic embedding (Zhao et al., 2022).

2.3.1.1 Facial Action Units

Originally introduced to psychology by Ekman & Friesen (1978), facial action units have become the predominant method for collecting and quantifying data from facial expressions. Rather than canonizing some set of expressions, the Facial Action Coding System (FACS) decomposes facial expressions into constituent muscle contractions. It allows researchers to articulate small differences in face states, and has been used in a spectrum of subjects, including consumer interest, psychopathology, and social psychology (Clark et al., 2020; Dursun et al., 2010; Zuckerman et al. 1981). While the measures make facial expressions precise and concrete, they are arduous to collect. Official training in the data collection procedure can entail 50 to 100 hours of self-directed study (Paul Ekman Group LLC). At 24 frames per second, it might take a coder one hour to annotate 4 seconds of video (Donato et al., 1999). While human coders do not often annotate individual video frames, this study aims to featurize all considered signals with a

high temporal resolution. Though datasets of video with FACS annotations are available (Dhall et al, 2012), these videos are not constrained in their setting or count of speakers.

2.3.1.2 Py-Feat

Py-Feat is a machine learning based library of models for producing accurate, frame-wise estimates of facial action units from a single 2D video. The tool is best understood as a composite, seeking to capture five separate facets of detection and analysis: face detection, landmark detection, head pose, facial action unit estimation, and emotion. It is important to note that this model decomposes videos into still frames for sequential processing, treating each frame in isolation. Additionally, not all adult face action units are predicted by the model. The authors attribute this both to the fact that most FAC models predict a similar subset, and that datasets' selections constrict which action units are available for training.

Py-Feat is also highly modular, allowing different model selections for different stages of processing and analysis. In its first stage, this model identifies whether a face is present within the image and the location of a bounding box to envelope it. Here, I chose to use the img2pose-unconstrained model.

Introduced by Albiero et al. (2020) and trained on more than 130,000 annotated faces, img2pose has maintained persistent, strong rankings in 3D human pose estimation (Papers with Code n.d *img2pose*). To locate faces, it employs an RNN with a Region Proposal Network — a resolution-hungry approach. This typically filters input images into feature maps (and therefore reduces their resolutions) before using those abstractions to infer objects' bounding boxes. As a consequence, it struggles with identifying faces at smaller scales, where too many filters will wash away all signal. To compensate, it is buttressed by a Feature Pyramid Network, which coheres high-level, low-resolution feature maps from the "end" of the RNN with low-level,

high-resolution representations from early layers, learning the object structure at various scales. By relying on the Feature Pyramid Network's scale-robust representations, the Region Proposal Network is better able to identify faces' locations in space. Choosing this model additionally demanded a determination on which variant to use: constrained or unconstrained.

While img2pose-constrained *can* provide more accurate estimates than the selected model, its performance suffers when working with faces skewed too far from the camera. My video sets do offer high resolution and central framing of the speaker, but these recordings often include dynamic camera angles to maintain viewer attention, increasing the prevalence of such skews. Head turns, additionally, represent a junction between body pose and facial expression, making these moments particularly valuable for an investigation of multimodality in social signals.

Following face detection, landmarks are identified using a MobileFaceNet, which demonstrates the greatest reliability. These landmarks include areas such as vertices located at the chin, cheeks, eyes, nose, and mouth especially. These landmarks are then used to isolate faces before applying a filter, rendering histograms of oriented gradients. These filtered images then provide the basis for the model of greatest interest, trained by Py-Feat, which generates estimates of facial activation units. In this category, I chose the XGBoost classifier, as it provides magnitude estimates of each facial action unit, rather than binary activation encodings. This model was trained on over 400,0000 expert-annotated frames, and achieves robust results. The data collected for this modality was then stored in .CSV files.

Lastly, the Py-Feat toolkit also allows users to generate emotion estimates, based on human-coded classifications. These estimates, while perhaps too abstracted from the complete multimodal picture we often rely on to infer affective state, were computationally inexpensive to

compute and were included in the final model. However, these estimates are not a major focus of this work. Rather, it offers a convenient test for the coherence of these prototypes to predict dynamic signals. If these expression prototypes are meaningful, then these high-level inferences may prove predictive of the dynamics of the complete multimodal picture.

Lastly, this study does not incorporate the landmark vertices the model produces. This data does not generalize well across camera angles, and the face state is better represented by facial action unit activation.

2.3.2.1 Body pose

To featurize the full spectrum of available social channels, this study aims to include body pose. While non-verbal communication is a respected notion, its study is often focused on facial expressions. de Gelder (2009) noted that body pose had been severely under investigated in spite of its recognizability and the poverty of stimulus from focus on facial expressions. Additionally, pose offers a strong indication of what physical actions an individual may be preparing to perform, and pose may have affective loadings (Wallbott, 1998). Unfortunately, there is no dominant, manual methodology for collecting pose information similar to the FAC system to draw from. However, there are indications about what levels of abstraction may be meaningful to humans. I turn to biological motion, introduced by Johansson (1973).

Biological motion removes a significant amount of information from its raw visual input, abandoning body contour, clothing, and other surface features, but it retains the structure of human anatomy. Shoulders, elbows, wrists, hips, and other key joints are tracked and represented with only dots. Even at its reduced level of information, these representations are easily interpretable even to young infants (Kuhlmeier et al., 2010).

2.3.2.2 PARE

To collect body pose data, I used Kocabas et al. (2021)'s Part Attention Regressor (PARE). This choice was largely motivated by the model's ability to infer the position of key joints, even if the input video does not depict all parts of the body. This study would struggle to collect a sufficient amount of data if its analysis was fully constrained to static-camera video videos. Its output is a data object with "tracks": unbroken streams of frame-wise estimates for a given person's pose. Each pose is defined as a set of coordinates in 3D space.

PARE produces more than coordinates. It creates multiple representations of body pose in its processing pipeline, with the goal of creating a 3D mesh of the target in the video. For present purposes, this mesh data was far too cumbersome, as it includes manifold variables to account for non-linear deformations and estimates of body shape. As these variables are static for each performer, they fail to reflect channel dynamics which center this investigation.

The model's architecture begins with a residual CNN to derive pixel-level features from an image still. From here, the final hidden states are passed to two separate branches, one focused on 2D estimates and the other on 3D. Both branches begin with their own untrained series of convolutions to reduce the dimensionality of the hidden state, and both branch outputs inform the final pose estimate used in this study.

The 2D branch is trained to assign each pixel 25 separate values, representing the probability that the pixel falls on one of 24 different body parts, or if it is instead capturing the background. The operations in the 3D branch are more opaque. It refines the residual network's final hidden states, and loads a value for each pixel based on each final hidden state's set of dimensions. A 600 x 900 image would produce a matrix of size 600 x 900 x C, with C representing the number of output dimensions from the CNN. While the 2D branch is trained at

this step on hand-annotated body part maps, the 3D matrix is only trained by downstream errors from the final mesh output.

Lastly, these branches are united by a part attention layer. Each pixel's set of body part probabilities acts as a softmax on each pixel's set of channel loadings from the 3D layer. The result is a kind of attention weight, denoting the importance of each channel for each body part. This final featurized matrix is then passed to a multilayer perceptron for each body part's joint angle with respect to its connecting part. 3D coordinates are retrieved by passing these angle-based joints through a simple pretrained linear layer. While the model generates coordinate points relative to the midhip, it relies on the target's skew relative to the camera to orient the rest of the body. To compensate for this, I rendered the body pose as pairwise distances. Additionally, I removed a number of comparisons which do not meaningfully vary, such as any connections between face parts or connections between respective limb joints. These include the distance between the left and right ears and between an elbow and its respective wrist. Additionally, given that a significant number of camera zooms capture only the mid-torso and upwards, I chose to remove all data points below the midhip. The end result is a profile of distances between 13 key points for each frame (minus the low-quality distances).

Lastly, to capture the speed of movements, I simply used the frame-wise difference for each joint pair, resulting in an equal number of dimensions for this separate modality. For frames without immediately preceding estimates, I simply copied the next frame's speed, assuming that these movements are smooth and continuous.

2.3.3 Semantic Encoding

As humans, it makes intuitive sense to conceive of language use as a unitary construct. However, natural language processing (NLP) has remained an essential lodestar in the field of

machine learning precisely because of the significant challenges this domain has posed, even when decomposed to subtasks and subdomains. NLP models, consequently, have often tracked the latest developments in architecture design and are often used as popular benchmarks for gauging the current capabilities of the field. The most recent advancement in machine learning architecture has been the introduction of transformer architectures, originally introduced in Vaswani et al. (2017). Some of the most notable models from this class include Alphabet's family of BERT models and the various iterations of OpenAI's GPT. Ultimately, this study made use of a length-60 vector to the semantic content of speakers' utterances. The significant complexity entailed in producing these representations warrants a discussion of its details, and how it diverges from more well-known semantic representations.

At a broad view, trained NLP models perform a number of sub-processes to create representations of human inputs, perform transformations on those representations, and return them to human-interpretable text. These models near-universally begin with a process which decomposes text into sub-word tokens, such as "we" and "ll" to form "well". Then, an embedding layer essentially serves as a lookup-table, storing a vector with a fixed number of dimensions for each token in a model's vocabulary. Transformers diverge from more traditional convolutional and recurrent neural networks, however, by the inclusion of "self-attention mechanisms."

This difference begins with the inclusion of positional encodings in the input embeddings. These positional encodings are simply an additional dimension — its value characterized by miniscule steps along sinusoidals as a function of the token's index within the larger input. The periodicity of that value allows models to establish regularities within and between whole inputs, and to generalize patterns across inputs of varying length. A layer's

attention head then creates, for each token embedding, three vectors: a key, a query, and a value. Each is a fraction of the size of the original embedding. Using the keys and queries within an input, the models create scores that relate each token embedding to every other token embedding. Then, a token's value vector incorporates the relational scores associated with its query vector to produce a weighted sum, representing the overall importance of that token. Multiple attention heads perform this operation at the same step, creating multiple weighted value vectors, each with different parameters to attend to different relationships. These value vectors are then concatenated (hence the fractional size), and combined — often with element-wise addition — with the original token embedding to create a more contextualized embedding for that token. Within a model, these embeddings are passed through multiple feedforward and attention layers. Ultimately, the transformer produces a final contextualized embedding (also called a final hidden state) for each input token, ostensibly representing its semantic content. These embeddings can then be tailored and used for specific tasks, such as classification, search, translation, and others. While representing profound advances in the field, these semantic representations greatly diverge from humans'.

Though the final hidden states are semantically rich, evidenced by transformer models' robust performance, they do not create a unified embedding. The count of contextualized embeddings is derived from the count of input tokens; natively, there is no single vector that captures the complete input's semantic meanings. Within the context of this study, where the goal is to represent each frame's featurized state with a fixed number of dimensions, token-wise embeddings are incompatible. To overcome this, I make use of the HPD-MiniLM-F128 model (Zhao et al., 2022).

This model reflects a lineage of challenges and proposed solutions in attempting to reduce semantics to numerics. The first abstraction from a basic transformer model is the disregard for a majority of final hidden state vectors. Rather than averaging all dimensions across output embeddings, this model takes the first token from classic BERT models: the “invisible” [CLS] token. Originally used for classification tasks, this token represents a broad picture of the output embeddings, addressing the problem of needing multiple vectors to represent a single utterance, rather than encoding some word/token. However, there is significant room for improvement in refining the token's content. This was addressed by SimCSE (Gao et al., 2021), which attaches a multi-layer perceptron to the output of the [CLS] token. The SimCSE model then trains the production of [CLS] tokens by using contrastive learning and dropout. From a batch of text sentences, the model passes one sentence through the encoder twice, each time introducing dropout. In the case of “He hasn’t surfed in years”, the model might pass “He hasn’t surfed _ years” and “_ hasn’t surfed in years”. The other sentences in the batch are undisturbed. The model is trained to make representations of dropout sentences more similar and to increase the contrast with undisturbed sentences, producing semantic representations which are more agnostic to random input noise. This is especially important in the present study, given that transcripts are represented as iteratively-expanding utterances, where one sub-utterance should, in some way, relate to the complete utterance. While this does raise the quality of and unify embeddings into a single vector, the output dimensions climb to 1,024, far outweighing the number of dimensions used in other modalities (the next greatest is body pose, composed of 63 dimensions).

One method to address this is to use a student-teacher paradigm (Wang et al., 2020). In this method, a transformer with fewer dimensions and layers is trained to mimic the final queries,

keys, and values of the larger models' final hidden state. These MiniLMs are further improved with "teacher assistants", which are intermediary models between the teacher model's size and target student's. Trained with a BERT base, this TinyLM can produce inferences twice as fast as its teacher while retaining up to 99% of its performance on some tasks. While smaller, these models still produce 756 dimensions, a 26% reduction which still far exceeds the other modalities. Finally returning to Zhao et al. (2022)'s work, the Homomorphic Projective Distillation (HPD) model directly attempts to decrease dimensionality, rather than simplify architecture.

The technique resembles the teacher-student model by training a smaller pre-trained model (here, the MiniLM) on a larger pre-trained model (SimCSE). Instead of training to mimic attention outputs, the model instead trains to a reduced SimCSE final hidden state. To reduce the dimensionality of the teacher's final hidden states, the technique applies a principal component analysis, bringing he count down to 128 dimensions. The pre-trained MiniLM, however, has an output size of 512 and so feeds its final hidden state to a single linear layer, with the resulting output being of size 128. The model's loss function merely minimizes the distance between the two 128-length vectors. While still considerably smaller than base transformer models' 1,024 dimensions, this reduced vector still falls out of parity with the rest of the modalities. To compensate, I apply an additional PCA, using the dimensionality-reducing script from SBERT (Reimers & Gurevych, 2019), which includes the ability to evaluate the resulting representation on the STS benchmark. The benchmark only focuses on models' abilities to relate semantically similar sentences and does not consider a spectrum of potential downstream tasks, as is the case for GLUE or SUPERGLUE benchmarks. While the PCA performed does not capture 95% of the variance, as is typical of most academically-geared data reductions, all variance is not inherently

semantically rich. The final, 60-length embedding is sufficiently representative to achieve 0.814 Pearson's correlation on the benchmark, while the base HPD model with 128 dimensions performs at 0.836. While models from 2019 are able to achieve scores well above this, this method's passable performance on balance with a vector size less than one tenth the length makes this suitable for the present study.

2.3.4 Traditional Audio Features

I also sought to capture traditional acoustic features which may be socially relevant. I used Librosa, an audio analysis and editing package for Python, to produce Mel-frequency cepstra coefficients, root-mean-square, and zero-crossing rates analysis. One important complication to note is that video data and audio data are stored at different rates. Where video tends to be stored at 24 or 30 frames per second, audio data is stored at a rate between 16k and 44.1k samples per second. Each sample captures the decibel amplitude at the time of recording, a ratio between a reference-amplitude (chosen by a recording's audio engineers) and the amplitude recorded by the microphone. These analyses are convenient in part because they inherently reduce the sample rate by deriving measures which represent spans of the data, rather than returning one measure per sample. Once such analysis is the production of Mel-frequency cepstra coefficients, which occupy 20 of the 34 featurized dimensions dedicated to traditional audio features.

The Mel-frequency is an alternative "scale" for quantifying frequencies, where each step is defined as being perceptually equal to all other steps, climbing the pure hertz scale logarithmically (Vergin et al., 2002). Focused on human perceptions, this measure has even been shown to be useful in a variety of machine learning tasks (Radford et al., 2022). The analysis begins with a "pre-emphasis" filter which simply takes a slice (whose length is user defined, but

typically 512 samples) and subtracts it from the following slice. This reduces both the temporal resolution of the measure and the presence of low-salience signals in the sample's lower spectrum bands. What follows is a classic Fast-Fourier Transform (FFT), which decomposes slices of the sample into power loadings. These represent the contributions of various sinusoidal frequencies in explaining a slice; combining the frequencies, scaled by their power loadings, should closely reproduce the original signal. Before returning the original scale, these power loadings are binned and consolidated into mel-frequency bands, converting the loadings from hertz. This is accomplished by convolving the power spectrum with triangle kernels, where the first kernel peaks at 1Hz, and the final kernel peaks at the frequency equal to half of the sampling rate. The rest of the triangle filters (whose count is determined by the number of MFCC desired from the analysis) are arranged linearly in the mel scale, with each triangle's peak coinciding with the previous triangle's fade to 0 and the next triangle's rise from 0. As higher mel-frequency steps are larger, this also reduces the resolution of the spectrum's higher frequencies. Given that perceptual volume scales with the log of power, the power loadings are then passed through a logarithm. Lastly, a Discrete Cosine Transform (DCT) renders each mel-frequency cepstral coefficient (MFCC), combining the power loadings of all mel filters and weighting each loading by a cosine function. The *i*th MFCC weights the *i*th mel filter most heavily. This decorrelates the coefficients and ensures that each MFCC contains at least some information about the entire spectrum.

Zero-crossing rates and root-mean-square and are far less complex. Both are rolling window averages, the former operating on sign change counts, and the latter on the window's amplitude (though, RMS then takes the square root of the average). Both use a window of length 2048 samples and a roll length of 512 samples. The final result is a set of 20 coefficients

representing perceptually-salient frequencies across time, a measure of power or volume, and zero-crossing rates, serving as a marker for speaker detection and identification, which may help in discriminating between the moments of audience response and speaker performance (Wasson & Donaldson, 1975; Kathirvel et al., 2011). However, these measures still far exceed the sample rate for visual measures. To retain parity with the rest of the data, I lastly downsampled these measures to match each video's respective framerate.

2.3.5 WhisperX

Though highly produced and intended for large audiences, both TEDTalks and Comedy Central videos suffer from a dearth of high-grade transcripts. TEDTalks often rely on automated transcribing tools, and Comedy Central offers no way to download its transcripts. Additionally, these transcripts, regardless of quality or accessibility, have a low temporal resolution. Consumer SRT files often are blocked in phrases. It represents only the phrase onset time and displays the entire phrase transcription at once. While this makes transcripts easier to read, they do not reflect how humans experience spoken language, which is central to the content of this dataset. To produce high-quality transcripts, I sought to use Whisper (Radford et al., 2022), a speech processing model produced by OpenAI designed to create transcriptions of English audio, translate audio to English, and derive timestamps for English transcripts. While it demonstrates impressive performance in extracting text, the original model suffers from poor timestamp accuracy. While the transcript absent of these markers may be useful, it would have no place in the present study. Its inclusion would come in the form of a single semantic embedding, represented uniformly and statically throughout all indices of the featurized matrix.

To overcome this, I used WhisperX (Bain et al., 2023), a modified version of the Whisper model explicitly designed to address its poor timestamp estimates. The original model uses a

simple chunk-and-stride method which reduces the impact of important context in creating transcription inferences. Whisper was trained on 30 second clips of audio. While the paper does not detail how longer audio was dealt with during training, Whisper's inference operations break longer audio files into sub-audios when they exceed this maximum. When phrases extend past the boundaries of a 30 second window, however, the model shifts that window to begin at the phrase's model-estimated start time. Bain et al., (2023) cite this reliance on the model's own inferences, alongside its heavy training on timestampless transcripts, as major sources of its poor timestamp performance. Errors in early windows compound with later errors, progressively skewing inferences. To address this, the authors introduce three additions to the base Whisper architecture.

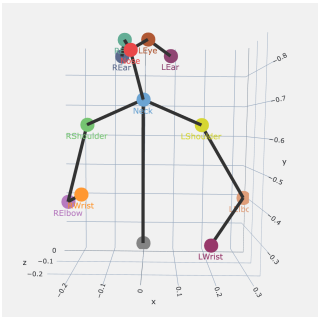
To avoid relying on Whisper's timestamp predictions to accommodate long sequences, WhisperX begins with a voice activity detection analysis to truncate non-speech. In removing wordless segments, this approach is able to perform fewer passes through the internal Whisper model. Additionally, marking these boundaries allows WhisperX to prevent mid-phrase segmentations, improving transcription quality by maintaining phrases' contexts. Once a transcript's text has been retrieved, the original text is passed through a phoneme-based model.

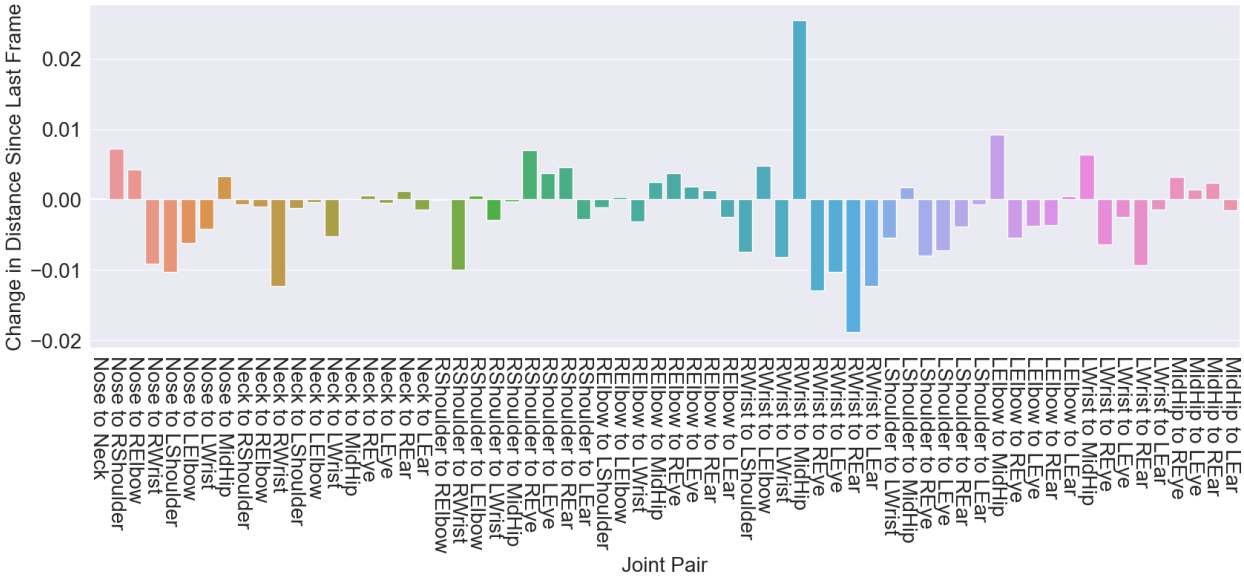
This phoneme-based model considers only the words which appear in the resulting transcript. It creates a matrix progressing through 20ms samples along one axis and the sequence of phonemes in the sample along the other. Each cell represents the logits of that phoneme being produced during that sample. WhisperX takes this matrix and performs Dynamic Time Warping (DTW). This defines the most likely path through time and the phoneme set, aligning the phonemes to timestamps. Phonemes at the beginning and ends of words help define that word's onset and durations. With improved timestamp estimates, I was able to accurately set the

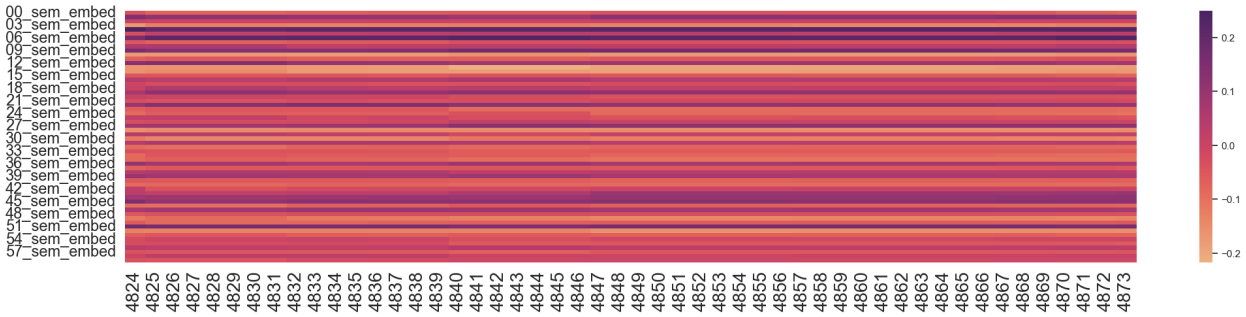
semantic embeddings to follow a time course mirroring the audience experience of each performance.

Within the featurized matrix, these transcripts are used to create iteratively expanding utterances, or grouped strings of words. Whisper might decompose “I went to see that movie you recommended to me the other day, and I don’t know if I fully understood the plot.” into separate phrases: one ending with “day” and the other beginning with “and”. Due to its training on human-generated transcripts, it breaks at semantically meaningful clefts rather than blind rolling windows. I take advantage of this to ensure that the encoded words represent something more akin to how the audience may segment and digest the performance.

While the semantic embeddings’ dimension size stays fixed at 60 in any given frame, a new embedding is generated at each word’s start time. The input text for each new embedding is derived from the most recent complete utterance, accompanied by a truncated version of the current utterance which ends with the most recent word spoken at that frame. This method allowed me to represent the dynamics of speech interpretation, both in attending to the most recent words and in grouping the words into semi-naturalistic segments. Additionally, the model which produces semantic embeddings in this study is suitable for dealing input text of this type. Recall that SimCSE (which was used to create the HPD model) was trained to represent input text with dropouts similarly (Gao et al., 2021). This ensures that the embeddings of sub-utterances are meaningfully similar to the embeddings of their complete utterances.


REar
LEye
LEar
RShoulder
LShoulder
RElbow
RWrist
LElbow
LWrist


Change in Distance Since Last Frame
0.02
0.01
0.00
−0.01
−0.02
Nose to Neck
Nose to RShoulder
Nose to RElbow
Nose to RWrist
Nose to LShoulder
Nose to LElbow
Nose to LWrist
Nose to MidHip
Neck to RShoulder
Neck to RElbow
Neck to RWrist
Neck to LShoulder
Neck to LElbow
Neck to LWrist
Neck to MidHip
Neck to REye
Neck to LEye
Neck to REar
Neck to LEar
RShoulder to RElbow
RShoulder to RWrist
RShoulder to LElbow
RShoulder to LWrist
RShoulder to MidHip
RShoulder to REye
RShoulder to LEye
RShoulder to REar
RShoulder to LEar
RElbow to LShoulder
RElbow to LElbow
RElbow to MidHip
RElbow to REye
RElbow to LEye
RElbow to REar
RElbow to LEar
RWrist to LShoulder
RWrist to LElbow
RWrist to LWrist
RWrist to MidHip
RWrist to REye
RWrist to LEye
RWrist to REar
RWrist to LEar
LShoulder to MidHip
LShoulder to REye
LShoulder to LEye
LShoulder to REar
LShoulder to LEar
LElbow to MidHip
LElbow to REye
LElbow to LEye
LElbow to REar
LElbow to LEar
LWrist to MidHip
LWrist to REye
LWrist to LEye
LWrist to REar
LWrist to LEar
MidHip to REye
MidHip to LEye
MidHip to REar
MidHip to LEar
Joint Pair


00_sem_embed
03_sem_embed
06_sem_embed
09_sem_embed
12_sem_embed
15_sem_embed
18_sem_embed
21_sem_embed
24_sem_embed
27_sem_embed
30_sem_embed
33_sem_embed
36_sem_embed
39_sem_embed
42_sem_embed
45_sem_embed
48_sem_embed
51_sem_embed
54_sem_embed
57_sem_embed
4824 4825 4826 4827 4828 4829 4830 4831 4832 4833 4834 4835 4836 4837 4838 4839 4840 4841 4842 4843 4844 4845 4846 4847 4848 4849 4850 4851 4852 4853 4854 4855 4856 4857 4858 4859 4860 4861 4862 4863 4864 4865 4866 4867 4868 4869 4870 4871 4872 4873
0.2
0.1
0.0
−0.1
−0.2

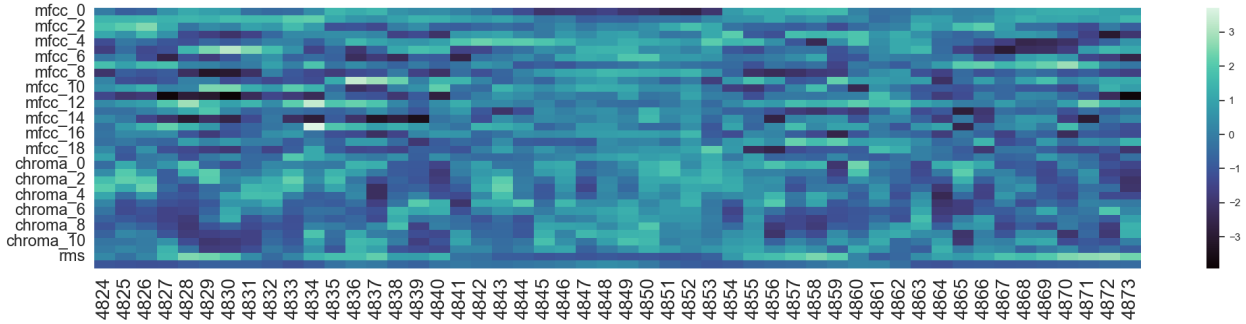

mfcc_0
mfcc_2
mfcc_4
mfcc_6
mfcc_8
mfcc_10
mfcc_12
mfcc_14
mfcc_16
mfcc_18
chroma_0
chroma_2
chroma_4
chroma_6
chroma_8
chroma_10
rms
4824 4825 4826 4827 4828 4829 4830 4831 4832 4833 4834 4835 4836 4837 4838 4839 4840 4841 4842 4843 4844 4845 4846 4847 4848 4849 4850 4851 4852 4853 4854 4855 4856 4857 4858 4859 4860 4861 4862 4863 4864 4865 4866 4867 4868 4869 4870 4871 4872 4873
3
2
1
0
−1
−2
−3

Figure 2. A visual representation of the data collected. The top column, from left to right, contains the pose estimation from PARE (though, these coordinates are represented as distance matrices in the final analysis), frame 4844 from Season 1, Episode 7 of *Comedy Central Stand-Up Presents*, and the facial action units estimated by Py-Feat. The second column contains traditional audio features from the same frame, and surround frames. The third column contains a similar visualization of semantic embeddings, obtained through WhisperX and the HPD model. The final column displays speed data, described as changes in the distance between key points. Together, these estimates aim to featurize important facets of social communication: facial expression, pose, speed, voice, and language.

3. Analysis and Results

One complication before performing this analysis is the sporadic ability of each model to produce estimates. For each frame, PARE attempts to infer the poses of as many bodies as are on screen. This means that in side-facing and crowd-facing shots, the model may estimate the pose for a large number of non-speakers. This is especially prevalent in TEDTalk videos, given their varied venues and performance structure. Slideshows, for instance, often include photos of humans which generate additional pose estimates, and throughout the set, audiences are seated at variable distances to the speaker. The sheer number of these instances in the data makes this worth addressing before analysis, but there is no reliable heuristic for selecting a speaker automatically. Rather than manually selecting the appropriate target in these instances, data where two poses are preset in a given frame are discarded. I justify this decision by noting that Comedy Central videos, without exception, are discontinuous due to their segments, meaning that the temporal dynamics are already disrupted. I follow a similar logic for PyFeat: if more than one face is detected within a single frame, the frame is discarded.

Audio features and semantic embeddings, however, are fully resistant to these complications. Traditional audio features are not dependent on any specific construct such as pose; instead, they describe and summarize aspects of the entire modality. Semantic embeddings, meanwhile, are represented continuously by design. Utterances' embeddings persist until a new utterance begins because, unlike pose and facial expressions, no activity occurs in this modality during the interim. While the performer may be making new facial expressions or poses while the camera films from an odd angle, audio data is only collected from a single microphone, meaning I can be confident that the relevant audio stream is fully captured. Additionally, WhisperX operates on an entire time series, whereas PARE and PyFeat do not use contextual

data (i.e. adjacent frames in the video stream). I therefore extend the semantic embeddings to reflect the entire time series and allow for disruptions in the visual data.

However, these decisions confer additional challenges to analysis. A generalized linear model assumes that all expected variables are present in order to function well across multiple samples. Consequently, I discarded all frames from all videos where one or more features were unavailable. Therefore, every retained sample is fully featurized — including pose, face, speed, audio features, and semantic embeddings. While this approach excludes a significant amount of data, this methodology also *creates* a significant amount of data. Following these exclusions, the TEDTalk dataset retained a total of 108,237 frames, or about 75 minutes of video. Comedy Central Videos retained 128,493 frames of fully featurized data, or about 89 minutes of video. To compensate for the difficulties in comparing measures with different magnitudes, relationships, and ranges, I standardized each video's data by converting the raw values into z-scores within each dimension. This was intended to control for the specifics of each speaker, including body size and acoustic profile.

Each GLM was defined by its video source, its target modality, a target dimension within that modality, and its predictor modality. For example, I collected a separate R^2 value for each dimension in the face modality: AU1, AU2, AU4, ... AU42, where estimates for each AU were informed by the entire semantic vector for that single frame. This means that there was a variable number of GLMs performed for each target-predictor modality pair, determined by the number of dimensions in the target modality. Additionally, this set of GLMs was performed for each video, meaning the total number of R^2 values collected is equal to the product the number of total dimensions across all modalities measured (249), the number of potential predictor modalities for each dimension (4, given that a target cannot be its own predictor) and the number of videos

analyzed (40). This comes to 39,840 total R^2 values, each constituted from a varying number of predictors due to variation in each modality's number of and frames.

This high number of R^2 values, however, need to be contextualized. I began by adjusted the R^2 values, using their number of predictors and the number of samples available to the GLM. This was necessary as the number of dimensions vary by modality, and each video contains a different number of fully featurized frames. To address this study's central question — whether TEDTalk or stand-up performances demonstrate more cross-modal redundancy — I performed a two-tailed, independent samples, weighted t-test, where each R^2 value was weighted by the number of samples in its video, producing $t(35.181) = 1073.569$, $p = .002$, *Cohen's d* = .987. On average, the R^2 values in the TEDTalk videoset (average $R^2 = 0.113$) were greater than those in the stand-up videoset (average $R^2 = 0.0824$).

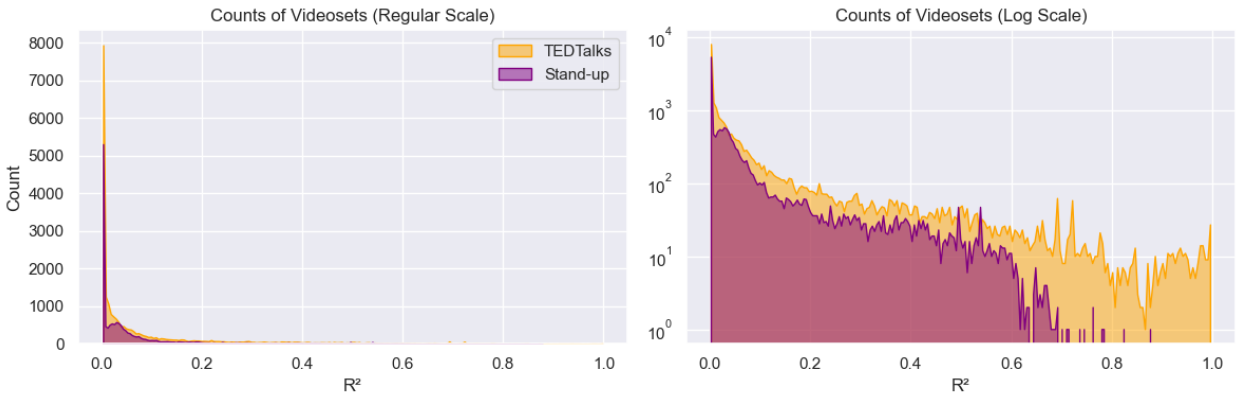


Figure 3. Histograms of the R^2 values from the TEDTalk (n = 23,904) and Stand-up (n = 15,936) videosets. While the Stand-up videos contain more fully featurized frames, they also contain fewer total videos, resulting in fewer R^2 values.

While the two video sets are separable, there is a question of whether modalities can genuinely predict other modalities greater than chance, given the data provided to the model.

Some of the strongest structure in the data comes from its synchrony. This demands probing for the role of stable structures within each modality's data, and whether the relationships between modalities' structures may be informing the predictions. If measures of the same frame are not more predictive than measures from different frames, this would discount the idea that speakers make active use of redundancy.

To address this, I created a null distribution by performing the same GLM analysis, but with circular shifting. Thus, this null distribution preserves the role of sequence, while disrupting the data's synchrony. When predicting the dimensions of a given target modality, those target dimensions were uniformly shifted by a random amount. All measures originating from the same frame remain in the same row, though offset from their predictors. Frame 2215's audio data, for instance, may now be predicting frame 4931's face data. After shifting the data, I performed the same GLM as I performed with the real data, and I conducted this shift-predict operation 5,000 times for each video. To mirror the analysis I performed with my real data, I averaged these R^2 values into 5,000 null grand-means. When compared with this null distribution, the data reveal stable structures' weak role: no circular-shifted R^2 means are greater than the true average.

$$\begin{bmatrix} AU1 & AU2 & \dots & AU43 \\ frame\,1 & frame\,1 & frame\,1 & frame\,1 \\ frame\,2 & frame\,2 & frame\,2 & frame\,2 \\ frame\,3 & frame\,3 & frame\,3 & frame\,3 \\ frame\,4 & frame\,4 & frame\,4 & frame\,4 \end{bmatrix} [\Theta] = \begin{bmatrix} mfcc1 & mfcc2 & \dots & zcr \\ frame\,1 & frame\,1 & frame\,1 & frame\,1 \\ frame\,2 & frame\,2 & frame\,2 & frame\,2 \\ frame\,3 & frame\,3 & frame\,3 & frame\,3 \\ frame\,4 & frame\,4 & frame\,4 & frame\,4 \end{bmatrix} \Rightarrow \begin{bmatrix} mfcc1 & mfcc2 & \dots & zcr \\ R^2 & R^2 & R^2 & R^2 \\ R^2 & R^2 & R^2 & R^2 \\ R^2 & R^2 & R^2 & R^2 \\ R^2 & R^2 & R^2 & R^2 \end{bmatrix}$$

$$\begin{bmatrix} AU1 & AU2 & \dots & AU43 \\ frame\,1 & frame\,1 & frame\,1 & frame\,1 \\ frame\,2 & frame\,2 & frame\,2 & frame\,2 \\ frame\,3 & frame\,3 & frame\,3 & frame\,3 \\ frame\,4 & frame\,4 & frame\,4 & frame\,4 \end{bmatrix} [\Theta] = \begin{bmatrix} mfcc1 & mfcc2 & \dots & zcr \\ frame\,4 & frame\,4 & frame\,4 & frame\,4 \\ frame\,1 & frame\,1 & frame\,1 & frame\,1 \\ frame\,2 & frame\,2 & frame\,2 & frame\,2 \\ frame\,3 & frame\,3 & frame\,3 & frame\,3 \end{bmatrix} \Rightarrow \begin{bmatrix} mfcc1 & mfcc2 & \dots & zcr \\ R^2 & R^2 & R^2 & R^2 \\ R^2 & R^2 & R^2 & R^2 \\ R^2 & R^2 & R^2 & R^2 \\ R^2 & R^2 & R^2 & R^2 \end{bmatrix} \Rightarrow \begin{bmatrix} audio \\ average\ frame\,4\ R^2 \\ average\ frame\,1\ R^2 \\ average\ frame\,2\ R^2 \\ average\ frame\,3\ R^2 \end{bmatrix}$$

Figure 4. Demonstration of R^2 value collection scheme for true (top) and null distributions (bottom), using face action units as predictors and acoustic features as targets. In the null calculations, sequence is preserved, but offsets are applied to the indices of target modalities.

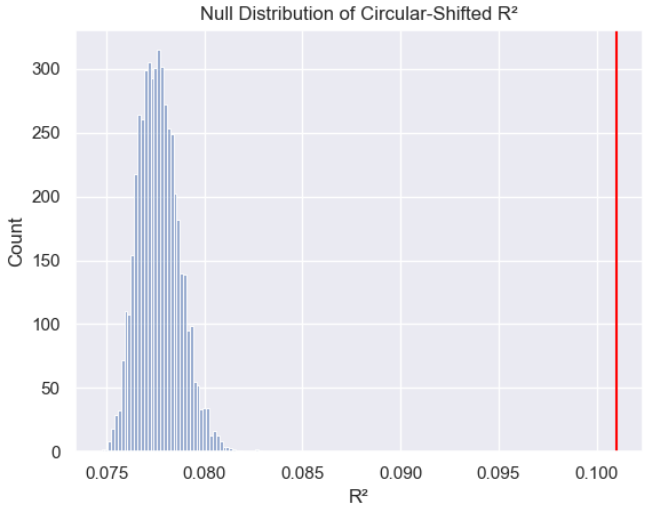


Figure 5. Summaries of how the true and null distributions compare. The null distribution is constructed from circular-shifted, whole-corpus means, while the red line indicates the grand-mean of all videos' R^2 values.

Additionally, I probed for differences between the two set's cross-modal predictability. To do so, I grouped null R^2 values by their target-predictor pair and the video they originated from. With this grouping, I constructed 5,000 null grand-means for each target-predictor pair within either videoset, following the same scheme as used for *Figure 5*. I then compared this distribution with each true videoset-level grand-mean to calculate significance. This revealed only two differences between the collections: in TEDTalks, embeddings are predictive of instantaneous speed (though not in the stand-up set), while face measures are predictive of instantaneous speed in the stand-up set (and not in the TEDTalk set).

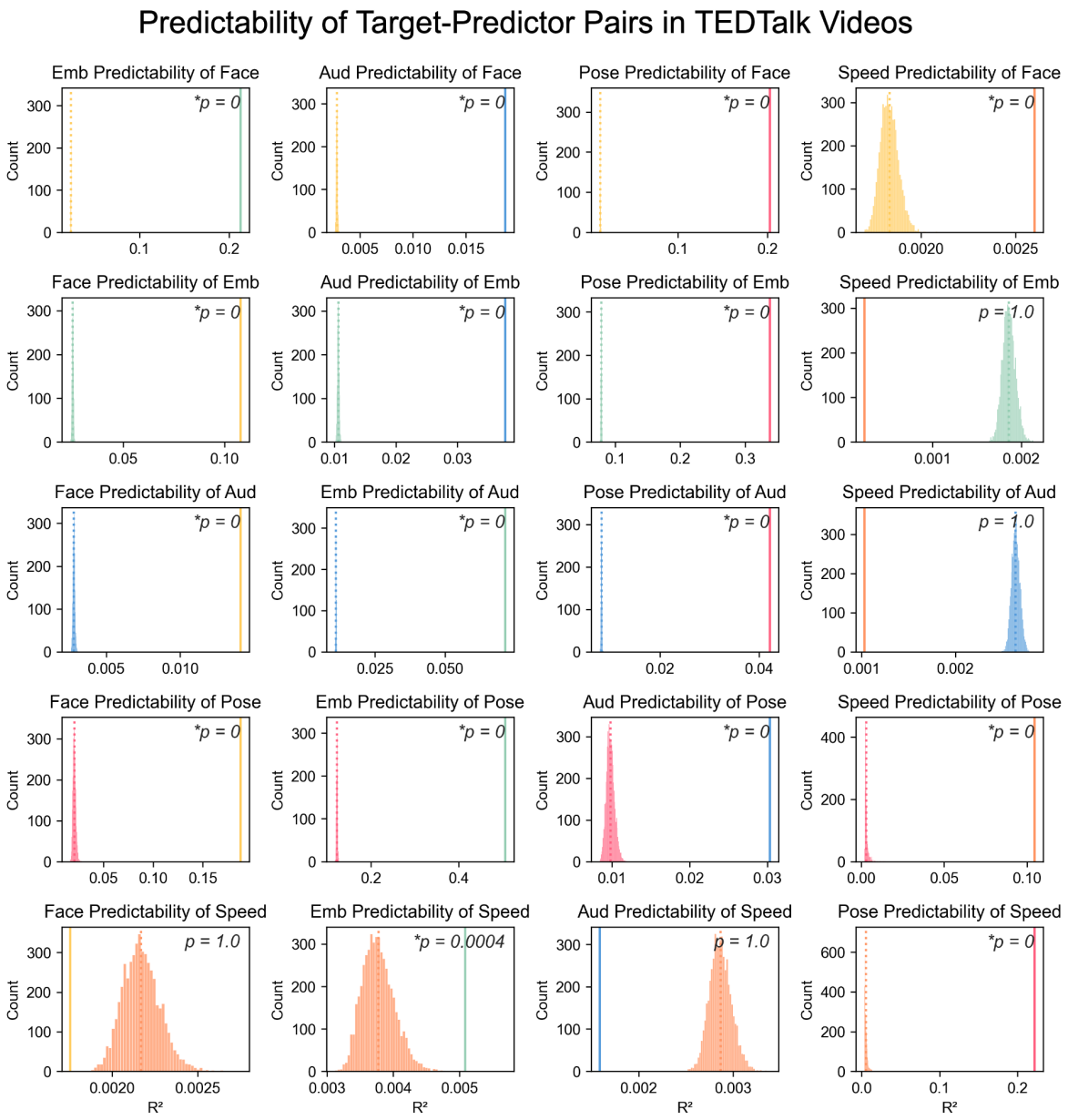


Figure 6. Visualizations of significance tests for different target-predictor pairs in the TEDTalk data. Each element is colored with respect to the modality it represents (yellow for face; green for embeddings; blue for audio features; red for pose; orange for speed). Each graph depicts the null distribution of R^2 values (histogram bars, colored by target modality), the mean of that distribution (dashed line), and the true mean of that target-predictor pair across all videos within the set (solid line, colored by predictor modality).

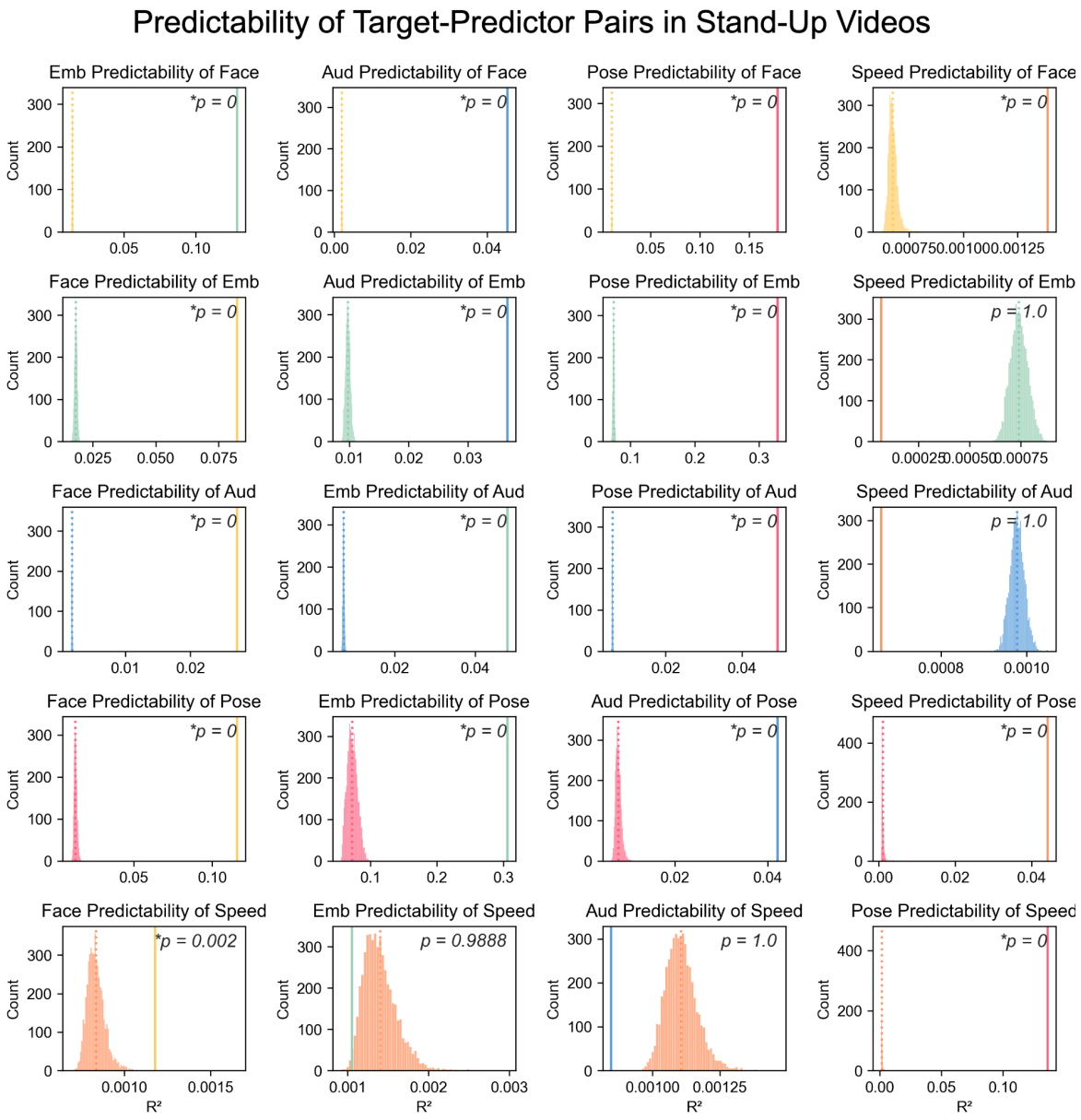


Figure 7. Visualizations of significance tests for different target-predictor pairs in the stand-up data. Each element is colored with respect to the modality it represents (yellow for face; green for embeddings; blue for audio features; red for pose; orange for speed). Each graph depicts the null distribution of R^2 values (histogram bars, colored by target modality), the mean of that distribution (dashed line), and the true mean of that target-predictor pair across all videos within the set (solid line, colored by predictor modality).

4. Discussion

4.1 Redundancy in Communication

These results validate greater cross-modal redundancy in TEDTalks than in stand up videos, and the separability of the two support the hypothesis that levels of redundancy can indicate differences in context.

This work may cohere with findings from studies which probe for the impact of expectations on speaker decisions-making. For instance, Galati & Brennan (2010) found that speakers, when relating a story to naive and familiarized addressees, augment which events they mention, the count of words they use, and the amount of detail they convey. Lockridge & Brennan (2002) similarly found that speakers emphasize unexpected elements when relating a story to an addressee who has less information than the speaker. These results may suggest that such tailoring is not constrained to utterances, but is present throughout the spectrum of modalities.

Additionally, these findings can also be interpreted with the same information theoretic scheme laid out at this paper's start. If performers have equal access to the same communicative behaviors, this would imply that TEDTalks are making use of a smaller subspace of those potential signals and combinations. This high-level hypothesis appears out-of-step with intuition. While it may seem that TEDTalks must communicate more "information" given their educational tilt, this intuitive sense of the term is separate from the theoretic one I have used throughout. TEDTalks introduce ideas which are *new* to the viewer, but each video's contents are bound tightly to their central arguments. Stand-ups, meanwhile, can vacillate between family dynamics, food, sex, drugs, and culture, thus covering a greater span of semantic space. Recall that unimodal redundancy is inverse to discriminability: more referents mean fewer phantoms.

TEDTalks, despite making more precise arguments and using more specialized concepts, rely on a smaller subset of those referents. This interpretation presents redundancy as dependent on the semantics present in performances, but redundancy may also serve a more relational, syntactic function.

At a psychological level, increased redundancy may allow performers to narrow the focus of viewers and brace unintuitive lines of reasoning. As a unimodal example, Jaeger (2010) found that speakers employ the complementizer "that" after using a verb whose subordinate clause is poorly predicted by the main clause. Redundancy, manifested in the strictly unnecessary word "that", clarifies the structure of a speaker's message. Holler et al. (2018)'s finding that gesture-accompanied questions receive shorter gaps in conversation also suggests that signals from one channel can modulate expectations about another. This study's findings show that certain target-predictor pairs are significant in one context but not another. Embeddings' ability to predict speed in TEDTalks and not in stand-up videos could suggest that profiles of coherence could also be informative. Exploring this possibility, however, demands behavioral measures from viewers of the videos, rather than theoretical explanations of the results. The present study cannot answer questions about whether or how humans use these relationships, but establishing their existence may help direct future investigations.

Additionally, the potential contribution of methodological errors may push these results in the opposite direction. Comedy Central videos retained a greater proportion of featurized frames, and they had the advantage of consistent editing techniques, camera angles, lighting, and venues. Despite these advantages and regularities, comedic performances were less suited for cross-modal prediction.

Comedy's reduced redundancy also follows from folk refrains for the form: violate expectations. Though, this same imperative should be reflected in TEDTalks; viewers are likely to be bored by predictable arguments and concepts. However, this analysis searched for structural redundancies, which may only recover lower-level patterns of coherence. This is to say that, while TEDTalks may describe less predictable ideas, they evade this study's operationalization.

4.2 Limitations

This study has attempted to draw in a wide spectrum of features from human, social interaction. Part and parcel of this, though, is understanding that the data itself should be taken as their own results. In this section, I begin from the broadest challenges and work down to technical questions and discrepancies.

Given the importance of context outlined in the introduction, it makes sense to levy the same critique against this work. Whenever measures have been extracted, this study has inherently divorced them from their context. While I have aimed to restore some of that context by incorporating other features, each measure confers assumptions about what meaningful contributes to a calculation of human redundancy. By operationalizing pose as key points, for instance, I suggest that the performers' particulars — their body shape, appearance, and mode of dress — are less relevant to the questions spurring this paper. A study which can safely exclude the role of such assumptions of importance must work with the data in its raw form.

Another broad challenge is the role of audience members in moderating the behaviors and decision making of the performers. At an even lower level, this study has struggled throughout to adequately separate the contributions of audience members and performers. For instance, audience laughter overlaps with speaker utterance, and they cannot be removed from

the audio stream (or their measures). This same challenge also appeared throughout pose and facial expression estimates: their inclusion reduced the number of usable frames, given my heuristics for data cleaning. While this study has proposed embracing context, I have excised it at many junctures. For instance, while both video sets contain solo performances geared toward gaining publicity, this study does not actively seek to address the role of the audience, whose reactions serve as an important component of the performers' decision making. Ultimately, the gold standard of data for investigating this claim of redundancy would likely involve recruiting participants to, in an isolated room, deliver a stand-up routine and educational presentation. This design, however, still fails to heed the importance of context in psychological research. Despite the promise of these models, ecological validity and measurability are still at odds.

Moving from performance-level confounds, each modality carries its own complications. In the visual world, this study's most pervasive challenge was the role of camera angles and video editing. Py-Feat and PARE have an unfortunate relationship: wide-shots improve the amount of data relevant to pose estimates, but they reduce the quality of face data. To compensate, I decided to remove leg articulation from the set of collected features, but this also strays further from the performance's ground-truth. Additionally, while the composition of training data is better tended to today, these models are still subject to biases along dimensions of settings, race, and technical details, even when inferring camera specifications to predict depth from 2D video. Suffice to say, they were not trained for the explicit purpose of analyzing TEDTalks and Stand-up.

Just as in broader machine learning literature, accurately capturing semantics poses the greatest challenge. These embeddings were recovered from what is essentially a kind of large language model "pug". While this already brings the quality of these semantic representations

into question, they also suffer from a broader challenge for natural language processing. Words can be interpreted in myriad ways, and while these models have made great strides, they still rely heavily on established meanings. The term "grandma", for instance, likely has a strong conceptual loading. Grandmothers are often invoked as prototypical examples of maternal care and domesticity. However, engaging performances often seek to upend common associations. In comedy, a performer may reference numerous stories from their grandmother's youth. While an audience member, during the course of the performance, may come to associate the word "grandma" with obscenities and substance use, large language models are unlikely to incorporate the modulated meaning. It is also worth noting that these models are trained on text.

Orthographic representations of language are not the same as their verbal forms. While the corpuses which train LLMs are often broad and diverse, they still capture text-specific patterns and structures. The semantic meaning of spoken words are often modulated by their tone, for instance. The featurization pipeline I have assembled, however, divorces the acoustic qualities from the words: audio is transcribed, and transcriptions are embedded. Additionally, there is the question of whether the content of TEDTalks is more accurately represented in these semantic embeddings. TEDTalks, and educational presentations more broadly, often make use of "standard" English. Meanwhile, stand-up performances often rely on less dominant strands of language, incorporating community-specific terms and constructions. By refraining from a representation of semantics which updates over time, the quality of these embeddings may meaningfully differ.

As a clarification, the analysis I performed has no inherent coherence to human psychology. Though it seems clear that humans make use of and integrate the contents of multiple channels simultaneously, there is no reason to suspect that those operations mirror the

operation of a generalized linear model. This analysis offers an information theoretic measure of redundancy, not a human one. As I emphasized previously, the level of interpretive redundancy is strongly dependent on the interpretive scheme. Without knowledge of that scheme (which is to say, the sum of human psychology), interpretive redundancy is out of reach. Instead, this study has analyzed the presence of structural redundancy, and it has made assumptions of the salient structures, their contents, and their dynamics. Kelly et al. (2004), for instance, found that gestures which precede and "match" the content of speech (i.e. gesturing up along the wall of a drinking glass while referring to its height) reduces the surprise of that speech in ERP measures. These "matching" relationships are unavailable to the analysis I used, as there are currently no semantic embeddings for poses. A comedian, while relating some narrative which barrels though semantic space, may need a wider array of gestures to accompany the wider array of topics. With an increased count of states and no way to relate them, the low-level features of these modalities cannot predict each other, reducing this measure of redundancy.

At a less realistic level, the gold standard for a form of *analysis* might be a fully unified, attention-equipped model. Much in the same way that signals in one channel may modulate the meaning of when combined with another channel (e.g. binaural perception, where the delay between ears modulates the location inference without modulating either ear's contents), a unified model should attend to all features, and it should use low-level estimates from one modality to inform the inferences of other modalities.

Lastly, to turn to the central question of the paper — whether performers in educational contexts are in fact more redundant — this study must also contend with the fact that humor is not exclusive to comedy, and arguments are not exclusive to education. While the topics of the videosets are distinct, performers' ability to educate and entertain have no simple measure.

4.3 Future Directions

If redundancy truly is a salient feature in human communication, additional research could search for the pattern of redundancies. Though these analyses incorporate the role of synchrony across modalities, it does not attempt to contend with the role of temporality. In comedy, for instance, "callbacks" often involve performers making references to prior topics or stories. Do these callbacks also mirror the multimodal signals of their first introduction? Is this profile of similarities consistent throughout between modalities? Might, say the pose be reduced while abandoning gesture.

Additionally, research looking to advance this strand, seemingly, will always be able to find new features to operationalize and incorporate. One such feature would be hand pose. A simple "thumbs up", for instance, has its own semantic loadings, yet this study introduces no measures to describe them. Additionally, future work could investigate redundancy across the various levels of interpretation. Low-level information would consist of raw data, such as pixels, and individual samples from the accompanying audio file. Climbing up a kind of representation hierarchy, abstractions may include the features collected here, and further refinements could try to group these collections of "multimodal states" and map their associations.

Such an analysis could, for example, make use of machine learning models, which have become the method-of-choice for contending with noisy, high-quantity data. These investigations could also investigate the impact of which modalities are available to the model. Such an approach, however, would also run into the classic "big data" problem, where more data seems to improve model performance in some cases while sacrificing generalizability in others.

Lastly, any work searching for regularities in social behaviors should consider the role of culture. These performances are often geared toward international audiences, but they often

retain a distinctly Western — and at times explicitly American — orientation. Broader accounts of redundancy in communication should attend to other styles of performance. Though, these efforts may also encounter difficulties in obtaining clean, naturalistic data. This study, constrained to English-speaking, solo performances produced by well-funded organizations, is no basis for sweeping generalizations.

5. Conclusion

This study has aimed to explore the role of cross-modal redundancies establishing performance context, and I hypothesized that educational contexts, demanding more specificity, would demonstrate greater redundancy. Emphasizing early information theory, I operationalized redundancy as predictability and captured a wide array of features to investigate this claim. Using a video set of 24 TEDTalk videos and 16 episodes from *Comedy Central Presents Stand-up* performances, I extracted measures to represent speakers' facial action units, pose, speed, acoustic profile, and the meaning of their spoken language. This effort made use of multiple machine learning models, which allowed me to extract large amounts of data with high temporal resolution. Ultimately, I show that, with the features collected, it is possible to separate the performances of these two forms using their cross-modal predictions. I also show that the quality of these predictions are above chance, emphasizing the role of synchrony in multimodal redundancy. I lastly offer interpretations of these results and suggest that relationships between modalities could offer a coherent foundation for further studies.

References

Albiero, V., Chen, X., Yin, X., Pang, G., & Hassner, T. (2020). img2pose: Face Alignment and Detection via 6DoF, Face Pose Estimation. CoRR, abs/2012.07791. Retrieved from https://arxiv.org/abs/2012.07791

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-accurate speech transcription of long-form audio. In arXiv [cs.SD]. arXiv. http://arxiv.org/abs/2303.00747

Cer, D., Yang, Y., Kong, S.-Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). Universal Sentence Encoder. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1803.11175

Chen, S., Liu, Y., Gao, X., & Han, Z. (2018). MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices. ArXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/1804.07573

Cheong, J. H., Jolly, E., Xie, T., Byrne, S., Kenney, M., & Chang, L. J. (2023). Py-Feat: Python Facial Expression Analysis Toolbox. ArXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2104.03509

Clark, E. A., Kessinger, J., Duncan, S. E., Bell, M. A., Lahne, J., Gallagher, D. L., & O'Keefe, S. F. (2020). The Facial Action Coding System for characterization of human affective response to consumer product-based stimuli: A systematic review. Frontiers in Psychology, 11, 920.

Cook, N. D., Carvalho, G. B., & Damasio, A. (2014). From membrane excitability to metazoan psychology. *Trends in Neurosciences*, 37(12), 698–705.

Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R., & Keltner, D. (2019). The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. Nature human behaviour, 3(4), 369–382. https://doi.org/10.1038/s41562-019-0533-6

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies. IEEE Multimedia, 19(3), 34–41.

D'Mello, S. K., & Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. User Modeling and User-Adapted Interaction, 20(2), 147–187. https://doi.org/10.1007/s11257-010-9074-4

Donato, G., Bartlett, M. S., Hager, J. C., Ekman, P., & Sejnowski, T. J. (1999). Classifying Facial Actions. IEEE transactions on pattern analysis and machine intelligence, 21(10), 974. https://doi.org/10.1109/34.799905

Dursun, P., Emül, M., & Gençöz, F. (2010). A review of the literature on emotional facial expression and its nature. *Yeni Symposium*, 48(3), 207–215.

Ekman P. & Friesen W. V. (1978). Facial action coding system: manual. *Consulting Psychologists Press*.

Fausti, S. A., Erickson, D. A., Frey, R. H., Rappaport, B. Z., & Schechter, M. A. (1981). The effects of noise upon human hearing sensitivity from 8000 to 20 000 Hz. The Journal of the Acoustical Society of America, 69(5), 1343–1347.

Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, *62*(1), 35–51.

Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple Contrastive Learning of Sentence Embeddings. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2104.08821

de Gelder B. (2009). Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 364(1535), 3475–3484. https://doi.org/10.1098/rstb.2009.0190

Gleick, J. (2011). *The information: A history, a theory, a flood*. Vintage Books.

Gielen, S. C., Schmidt, R. A., & Van den Heuvel, P. J. (1983). On the nature of intersensory facilitation of reaction time. *Perception & Psychophysics*, 34(2), 161–168.

Harrigan, J. A., Rosenthal, R., & Scherer, K. R. (2005). The New Handbook of Methods in Nonverbal Behavior Research. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&db=e000xna&AN=142080&site=ehost-live&scope=site&authtype=ip,shib&custid=dartcol&group=main

Holler, J., Kendrick, K. H., & Levinson, S. C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic Bulletin & Review*, 25(5), 1900–1908.

img2pose: Face Alignment and Detection via 6DoF, Face Pose Estimation | Papers With Code (n.d.). Retrieved April 18, 2023, from https://paperswithcode.com/paper/img2pose-face-alignment-and-detection-via

Jaeger, T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2), 201–211. https://doi.org/10.3758/BF03212378

Kathirvel, P., Sabarimalai Manikandan, M., Senthilkumar, S., & Soman, K. P. (2011). Noise robust zerocrossing rate computation for audio signal classification. *3rd International*

Conference on Trends in Information Sciences & Computing (TISC2011). https://doi.org/10.1109/tisc.2011.6169086

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, *89*(1), 253–260.

Kocabas, M., Huang, C.-H. P., Hilliges, O., & Black, M. J. (2021). PARE: Part Attention Regressor for 3D Human Body Estimation. In arXiv [cs.CV]. arXiv. http://arxiv.org/abs/2104.08527

Kuhlmeier, V. A., Troje, N. F., & Lee, V. (2010). Young infants detect the direction of biological motion in point-light displays. *Infancy: The Official Journal of the International Society on Infant Studies*, 15(1), 83–93.

Lockridge, C. B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic bulletin & review*, *9*(3), 550-557.

Middlebrooks, J. C., & Green, D. M. (1991). Sound localization by human listeners. *Annual review of psychology*, *42*, 135–159. https://doi.org/10.1146/annurev.ps.42.020191.001031

Miller, J. (1982). Divided attention: evidence for coactivation with redundant signals. Co*gnitive Psychology*, 14(2), 247–279.

Oviatt, S. (1997). Mulitmodal interactive maps: Designing for human performance. Human-Computer Interaction, 12(1-2). https://doi.org/10.1080/07370024.1997.9667241

Paul Ekman Group LLC. (2020, January 30). Facial action coding system. Retrieved April 19, 2023, from https://www.paulekman.com/facial-action-coding-system/

Politis, I., Brewster, S.A., & Pollick, F.E. (2014). Evaluating multimodal driver displays under varying situational urgency. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. In arXiv [eess.AS]. arXiv. http://arxiv.org/abs/2212.04356

Ritchie, D. (1986). Shannon and weaver. Communication Research, 13(2), 278–298.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. https://doi.org/10.48550/ARXIV.1908.10084

Rubi, T.L., & Stephens, D.W. (2016). Does multimodality per se improve receiver performance? An explicit comparison of multimodal versus unimodal complex signals in a learned signal following task. *Behavioral Ecology and Sociobiology, 70*, 409 - 416.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. MIT Press.

STS benchmark (n.d.). Retrieved April 22, 2023, from https://paperswithcode.com/sota/semantic-textual-similarity-on-sts-benchmark

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423.

Vergin, R., O'Shaughnessy, D., & Gupta, V. (2002). Compensated mel frequency cepstrum coefficients. 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA. https://doi.org/10.1109/icassp.1996.541097

Wallbott, H. G. (1998). Bodily expression of emotion. European Journal of Social Psychology, 28(6), 879–896.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2002.10957

Wasson, D., & Donaldson, R. (1975). Speech amplitude and zero crossings for automated identification of human speakers. IEEE Transactions on Acoustics, Speech, and Signal Processing, 23(4), 390–392.

Xu, H., Zhang, X., & Jia, L. (2012, May). The extraction and simulation of Mel frequency cepstrum speech parameters. 2012 International Conference on Systems and Informatics (ICSAI2012). 2012 International Conference on Systems and Informatics (ICSAI), Yantai, China. https://doi.org/10.1109/icsai.2012.6223385

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). A survey of large language models. https://doi.org/10.48550/ARXIV.2303.18223

Zhao, X., Yu, Z., Wu, M., & Li, L. (2022). Compressing sentence representation for semantic retrieval via Homomorphic Projective Distillation. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2203.07687