The Journal of the Acoustical Society of America

MAY 16 2025

How does a credible voice sound?

Jochen Steffens ; Patrick Blättermann; Maximilian Sattler; Kevin Tang

Check for updates

J. Acoust. Soc. Am. 157, 3780–3792 (2025)
https://doi.org/10.1121/10.0036646

View Online

Export Citation

JASA ARTICLE

How does a credible voice sound?

Jochen Steffens,[1,a)] Patrick Blättermann,[1] Maximilian Sattler,[1] and Kevin Tang[2,3,b)]

[1]*Institute of Sound and Vibration Engineering (ISAVE), Hochschule Düsseldorf, Düsseldorf 40476, Germany*

[2]*Department of English Language and Linguistics, Institute of English and American Studies, Faculty of Arts and Humanities, Heinrich Heine University Düsseldorf, Düsseldorf 40225, Germany*

[3]*Department of Linguistics, University of Florida, Gainesville, Florida 32611-5454, USA*

ABSTRACT:
Credibility is a crucial social factor that influences people's perception and decision-making. This study explores the acoustic attributes that define credible speech compared to neutral and ironic speech. A custom-built German corpus was developed, containing speech samples recorded from amateurs to enhance ecological validity. The study extracted a broad set of audio features from these samples, employing recursive feature elimination to identify the most influential attributes. These were then analyzed using a machine-learning-supported multinomial logistic regression model. The results indicated significant differences in the acoustic features associated with credible speech compared to neutral and ironic speech. Key findings include the role of a higher energy level (1st mel-frequency cepstral coefficient) in credible compared to neutral and ironic speech and a higher speaking rate in both credible and ironic compared to neutral speech. Also, irony is characterized by more high-frequency content (mean spectral centroid) compared to credible or neutral speech. Gender differences in spoken irony involve a greater influence of speaking rate in women's speech, while high pitch plays a more significant role in men's speech. This research thus contributes to the understanding of how credibility is conveyed through speech and offers insights for applications in communication, media, and artificial intelligence. The study also highlights the methodological advancements made by incorporating a diverse range of acoustic features and employing a robust machine-learning framework.

https://doi.org/10.1121/10.0036646

(Received 16 August 2024; revised 20 April 2025; accepted 22 April 2025; published online 16 May 2025)

[Editor: Jody Kreiman] Pages: 3780–3792

I. INTRODUCTION

Effective communication relies on the perception of credibility, an important semantic dimension that influences the way we interpret and respond to information (Hovland and Weiss, 1951). The concept of credibility generally refers to the extent to which a message as well as its source is perceived as trustworthy and reliable by its recipients (Hovland *et al.*, 1953). It plays a significant role in various domains such as politics (von Hohenberg and Guess, 2023), media (Metzger *et al.*, 2003), and education (da Rosa and Otero, 2018) as it influences people's decision-making process. This study therefore examines the acoustic attributes that distinguish credible speech from neutral and ironic speech of male and female speakers.

Most contemporary studies of credibility examine dimensions of source, message, and media credibility (Metzger *et al.*, 2003). Source credibility refers to the evaluations made by a perceiver about how believable the communicator is. Message credibility focuses on how the characteristics of the message influence perceptions of its believability, whether regarding the source or the message itself; in this respect, there are overlaps in source and message credibility. Media credibility refers to the trustworthiness and reliability of the medium or platform through which the information is delivered (e.g., newspapers, TV channels, social media).

Research concerning message and media credibility witnessed a resurgence in popularity in recent years, due to recurring discussions in the cultural mainstream concerning social media and the veracity of news online, which warranted the adaptation and modification of known concepts and scales to the heavily changing cultural communication landscape (Appelman and Sundar, 2016; Chung *et al.*, 2012; Metzger *et al.*, 2003).

Yet, how can one tell if a speaker is credible or not? Both vocal and non-vocal properties of a speaker have been found to influence the perception of credibility. Studies have suggested positive effects of prolonged eye contact (Beebe, 1974), appropriate clothing (O'Neal and Lapitsky, 1991), or general non-vocal behavior and movement (Burgoon *et al.*, 1990). Compared to non-verbal cues, vocal cues have been suggested to already contain a great deal of information to discern the credibility of a speaker. Bond and DePaulo (2006) conducted a large scale meta-analysis of research on the accuracy of deception judgments, synthesizing results from 206 documents and 24 483 judges. They revealed that

a)Email: jochen.steffens@hs-duesseldorf.de
b)Email: kevin.tang@hhu.de

the accuracy of identifying deception is significantly higher when relying solely on audio recordings compared to video-only recordings. The current study, consequently, centers on vocal indicators of credibility.

Given that the characterization of credibility remains an open question, not surprisingly, studies that investigate the vocal properties of credibility have used a range of related terminologies, such as trust(worthiness) (Belin *et al.*, 2017; Chen *et al.*, 2020; De Meo, 2012; De Meo *et al.*, 2011), truthfulness (Chen *et al.*, 2020; Kirchhübel and Howard, 2013; Syed *et al.*, 2019), and certainty and honesty (Goupil *et al.*, 2021). Furthermore, to better examine the acoustic correlates of credibility, researchers would select a baseline "control" condition and a contrasting dimension, such as deception, especially when comparing to trust (Kirchhübel and Howard, 2013; Patel *et al.*, 2023).

Instead of deception, however, the current study selected irony being an appropriate contrasting dimension. An ironic message is a deliberately false statement that includes counterattitudinal (a facetious display of an attitude) information with the intention for the recipient to recognize its falsehood (Averbeck, 2010; Averbeck and Hample, 2008; Kreuz and Link, 2002). While both ironic and deceptive messages contain the element of falsehood, the intention to be detected by the listener is deliberate in ironic messages. Therefore, one could expect irony to provide a starker contrast to credibility than deception does.

How exactly does irony connect with credibility? Averbeck (2010) discussed the relationship between irony and credibility under the framework of language expectancy theory, which addresses expectations of language patterns, particularly examining how various aspects of a message can conform to or deviate from expectations concerning appropriate communication (Burgoon, 1995).

Meeting or exceeding expectations is favorable for the source, leading to higher ratings of persuasiveness and credibility of the source (Burgoon, 1995); conversely, a negative violation of expectations would result in lower ratings of persuasiveness and source credibility (Hamilton *et al.*, 1990). A high-credibility source is expected to satisfy expectations, while this is not expected for a low-credibility source. Ironic messages are intentionally counterattitudinal, and they attempt to maintain and not change an attitude and its corresponding behavioral intentions. According to language expectancy theory, when someone communicates a message that contradicts their actual attitude, the perceiver is inclined to align with the intended outcome of the message. However, the sender typically does not want the recipient to take the ironic statement literally (Sperber and Wilson, 1981).

What are the specific acoustic cues of credibility and similar and contrasting dimensions? One of the few studies that focused on the credibility of speech in German was by Schröder *et al.* (2017). They examined a set of acoustic cues and their influence on the credibility of speech in German using an analysis-by-synthesis approach. Synthetic utterances with manipulated acoustic cues were judged by participants in terms of credibility. They found that an increase in breathiness or an insertion of a tremolo (trembling of the voice) increases credibility, while an insertion of a pause or a raise in pitch decreases it. In other languages and related semantic dimensions, additional acoustic cues have been found to correlate with credibility, such as intensity, pitch variance, and speaking rate (e.g., Belin *et al.*, 2017, Chen *et al.*, 2020, De Meo, 2012, Hartwig and Bond, 2011, Zuckerman *et al.*, 1981).

In irony, prominent features of its acoustical signature were shown to be higher pitch levels or a wider pitch range (Bryant and Tree, 2002; González Fuente *et al.*, 2016; Scharrer and Christmann, 2011), a reduction in speech rate (Bryant, 2010), or a prolonged vowel and syllable duration (Adachi, 1996; Anolli *et al.*, 2000; González Fuente *et al.*, 2016; Laval and Bert-Erboul, 2005; Scharrer and Christmann, 2011), as well as unusual or exaggerated stress on particular words (Kochetkova *et al.*, 2021). Contrasting research by Rockwell (2000) and Leykum (2021), however, has observed ironic speech to also be associated with lower pitch. This example illustrates that the findings about the acoustics of credibility and irony are mixed and sometimes contradictory, which motivates the present study using novel, more robust statistical methods.

Also, regarding speaker gender, an inconsistent effect on credibility was observed. Some studies have been reported that the gender of the speaker does not play a strong role in credibility and trust. De Meo (2012) examined the segmental and suprasegmental acoustic details of native and non-native speech in Italian and found that gender did not affect credibility judgements. Syed *et al.* (2019) created and examined a gender-balanced database of speech samples representing individuals with low and high public trust and found that, for most acoustic features, the degree of public trust of a speaker did not differ by their gender with respect to prosody and voice quality. The few features that did show a gender difference were those that define the lower-frequency end of the speech spectra; concretely, individuals who have a perceptually deeper voice tend to have a higher trustworthiness. Chen *et al.* (2020), however, have revealed a gender effect, with female English speakers being trusted more than male speakers.

In contrast to the aforementioned studies that examine traditional acoustic parameters, only a handful of studies have examined acoustic feature sets such as eGeMAPS (extended Geneva Minimalistic Acoustic Parameter Set) (Eyben *et al.*, 2016), which are often used as standard features in speech classification tasks. Specifically, mel-frequency cepstral coefficients (MFCCs), a common feature type in acoustic feature sets, were underexamined. MFCCs are widely used in natural language processing models concerning spoken speech, such as speech recognition (phone/word) and speaker recognition. They are derived through a series of transformations that include taking the logarithm of the power spectrum and applying a discrete cosine transform (Lerch, 2012). This process results in coefficients that are not directly related to intuitive physical properties of the sound; however, they can be generally interpreted as a compact representation of the spectral content as perceived by the human ear (Lerch, 2012).

Syed *et al.* (2019) examined the eGeMAPS feature sets, which include MFCCs of speech samples representing individuals with low and high public trust. Santos *et al.* (2016) focused on whether MFCCs can be used as part of a domain-agnostic approach for opinion prediction and found that MFCCs were important indicators for predicting persuasiveness (which presumably relates to credibility) in debates (Brilman and Scherer, 2015). Given how widely used MFCCs are in voice classification tasks, a likely reason why MFCCs are not often examined is their aforementioned complex representation of the frequency spectrum and thus their poorer interpretability. A recent study by Tracey *et al.* (2023) has begun to reconcile this shortcoming by correlating MFCCs with more interpretable speech biomarkers, such as the high-to-low frequency energy ratio in the case of MFCC 2.

Beyond the choice of acoustic features, studies that examined MFCCs employed the usual approach in speech emotion recognition, which is different from traditional phonetic studies. Traditional phonetic studies typically focus on examining the effect of a small set of acoustic features statistically using analysis of variance (ANOVA) or regressions, while more recent approaches have begun to examine a large set of acoustic features using machine learning (Sonderegger and Soskuthy, 2024; Tavakoli *et al.*, 2025; Tomaschek *et al.*, 2018). In speech emotion recognition studies, usually, a set of standard descriptive audio features is extracted from a database of audio files first. The feature extraction is most commonly performed using tools such as PRAAT, OPENSMILE (Eyben *et al.*, 2010), or LIBROSA, which are used to extract the most prominent prosodic and spectral descriptors, like MFCCs, speaking rate, pitch contours, and other notable spectral components. Features are reduced and selected using statistical procedures, such as principal component analysis (PCA) or recursive feature elimination, which aim to find the most influential features as a means of simplifying the data. The resulting features are then analyzed using statistical modeling, such as multinomial logistic regressions, or other, more complex, models, such as support vector machines, to define the exact influence and relation of the features to the respective emotional classes. Studies following this approach have achieved average classification accuracies of around 80% (El Ayadi *et al.*, 2011; Semwal *et al.*, 2017). Apart from emotions, these techniques have also been successfully used to classify more abstract semantic dimensions, such as intent (Gu *et al.*, 2017), sincerity, or deception (Schuller *et al.*, 2016); however, no such efforts have been made in the field of credibility research so far.

In the current study, we bring together these empirical and methodological concerns to ask two questions regarding the vocal cues of credibility. First, what are the acoustic characteristics of credible speech (RQ1) and ironic speech (RQ2) compared to neutral speech (RQ3)?[1] Second, do speakers of different genders differ in their vocal cues in credible and ironic speech compared to neutral speech? The current study also fills a methodological gap. We approach the investigation of these questions differently from those that examine traditional acoustic parameters in two ways. First, we include a broader set of features, including MFCCs. Second, we employ a non-linear machine learning method to select relevant features for better interpretability and to increase the power to predict unseen data. This improves the generalization of our findings by optimizing the trade-off between the accuracy of predictions and the number of features in the model. Speech is inherently multidimensional and can vary across time (e.g., formant trajectories), as evidenced by acoustic feature sets such as the ComParE feature set (Schuller *et al.*, 2013), which consists of 6373 static acoustic features. When faced with such high dimensionality, traditional statistical methods, such as ANOVA and regression, suffer from the issue of collinearity (Tomaschek *et al.*, 2018), which is when predictor variables in a model are highly correlated, estimates of features may become unstable, leading to the wrong conclusions. One way of addressing collinearity is to employ machine learning techniques. We build on recent phonetic studies that have adopted this approach, such as Howell *et al.* (2017) and Villarreal *et al.* (2020), which extract hundreds of acoustic measurements and analyze them using feature selection algorithms and machine learning models, including random forest and support vector machines. This approach guards against the challenges caused by the so-called "researcher degrees of freedom" in phonetics (Roettger, 2019) as to what acoustic parameters to extract and how and intentionally or unintentionally hunting for significant *p* values. It therefore reduces the chance of researchers overlooking potentially important acoustic features and of misattributing the effects under investigation to other acoustic features. Furthermore, unlike traditional statistical methods, the use of machine learning approaches typically necessitates the validation and stabilization of model performances. This reduces the risk of overfitting the model to a specific dataset and increases the generalizability of the findings. We follow the standard methodology found in speech emotion recognition and aim to strike a balance between model accuracy and interpretability (e.g., Nfissi *et al.*, 2024). We created a corpus of elicited credible, ironic, and neutral speech and carried out an analysis with a feature set that is broad enough to ensure a decent classification accuracy while also choosing features that are understandable and interpretable in an attempt to approach a better understanding of the "sound of credibility."

II. METHOD

A. Speech corpus and participants

Standard German, also known as *Hochsprache*, *Schriftsprache*, and *Standardsprache*, is the target language of the current study. Standard German is an official language in Germany, Austria, Switzerland, Luxembourg, and

Liechtenstein, spoken by approximately 95×10^6 speakers (Gordon, 2005). This variety of German is the supraregional variety codified in grammars, such as the Duden Grammar (Duden Editorial Team, 2005), and it serves as the "official" form of the language. Written German usually adheres to the spelling and grammatical standards established for this particular variety (Fagan, 2009a,b). Many existing speech corpora elicit deceptive speech in a laboratory setting. Some studies, such as Hirschberg *et al.* (2005) and Kirchhübel and Howard (2013), follow an interview scheme, where participants are asked to lie in parts of an interview, while other studies, such as Huang *et al.* (2019), Levitan *et al.* (2015), Schuller *et al.* (2016), and Zhang *et al.* (2022), make use of interactive games. Please see De Luca *et al.* (2024) for a discussion of recent deception experiments and corpora. These existing corpora are limited in their language of choice, namely only English and Mandarin Chinese. These corpora were not suitable for this study, since they do not cover the German language and they do not examine irony. As the acoustic speech analysis of the chosen semantic dimensions required a highly specific collection of recordings that was not covered by existing corpora or archival recordings, a custom corpus had to be created.

Previous corpora in the field of speech emotion recognition, such as the Berlin Database of Emotional Speech (Burkhardt *et al.*, 2005), have used professional actors to perform the respective emotional dimensions. While this is a popular approach, it is often criticized for the fact that it can lead to inauthentic, exaggerated representations of the desired contexts (Iriondo *et al.*, 2007). We therefore recruited amateurs for the recordings of the corpus instead, to ensure a variety of everyday-life-like voices leading to more ecologically valid results. Regarding the paradigm of the experiment, we decided not to use interviews or social games as previous speech corpora on deception had done. Rather, we opted to use a reading task following Schröder *et al.* (2017). We acknowledge that, while this approach establishes more control, it does not fully capture truly natural speech. It remains a challenge to elicit speech naturally in a laboratory setting for semantic dimensions such as deception; please see Enos (2009) for a list of criteria.

B. Material

Existing corpora often use either pseudo-linguistic utterances (e.g., Banse and Scherer, 1996) or random neutral sentences (e.g., Burkhardt *et al.*, 2005) for their recordings, having the actors record the same phrases in all emotional dimensions. Both are popular approaches that ensure maximum acoustic comparability by completely separating the prosodic from the lexical content. However, authentically acting out a specific semantic dimension while only using completely unrelated, random, or even ungrammatical phrases is a rather complex and abstract task that would presumably be even more challenging for non-actors. Therefore, a meaning-guided elicitation method was applied as it has successfully been employed in previous studies featuring amateur-based speech corpora (e.g., Niebuhr, 2010). In this method, the desired semantic dimensions are elicited from speakers by having them recite specifically designed passages, representing familiar situations in which one would commonly speak in the desired tone.

Other scenarios, such as being a news reporter (De Meo, 2012) or being accused of committing a crime (Kirchhübel and Howard, 2013), often require specialized knowledge or training to perform convincingly, such as acting skills. To avoid introducing variability stemming from participants' acting abilities, we intentionally opted for a scenario that would be universally relatable and familiar to laypersons and a setting in which one would commonly want to come across as highly credible: the job interview. Furthermore, the job interview scenario enables a practical application of our research. Credibility in job interviews is a socially relevant topic, and insights from this study may contribute to a better understanding of how speech features influence perceived trustworthiness and professionalism in such high-stakes settings. Finally, existing experimental paradigms such as using a social game have not been tested to elicit more than one dimension, while the job interview scenario allows us to elicit both credible and ironic speech by means of changing the script. Therefore, for the `credibility` condition and the `irony` condition, two monologues from a job interview situation were chosen.

To be able to compare the acoustical features of the `credibility` condition, a `neutral` condition was established as well as one `irony` condition that can be considered the opposite of `credibility`, or at least provide a stark contrast. For the `neutral` passage, a specifically created German translation of the "Rainbow Passage" (Fairbanks, 1960) was chosen: a short public domain passage that is commonly used in a wide range of linguistic and speech assessment contexts as a standardized stimulus due to its neutral tone and its quality of being approximately phonetically balanced, meaning that it contains an accurate representation of all the phonemes of the English language. The `irony` passage used the same job interview monologue from the `credibility` condition, re-phrased in a more exaggerated, ironic, presumptuous tone, expecting the obvious contrast to the other version and the choice of phrasing to successfully elicit the right speaking tone.[2] As writing accurately phonetically balanced passages is a challenging and highly specified linguistic task, the large language model "ChatGPT" (OpenAI, 2023) was used to write all the monologues and to translate the "Rainbow Passage" in a way that maintains its phonetic balance.[3] The resulting ~300-word passages were then converted to a phonetic transcription using the "G2P" tool off the BAS web service package (Reichel, 2012, 2014) to check whether phonetic balance was actually achieved. The individual phone percentages of the resulting transcriptions were calculated, which were then compared to the phone statistics [the BAStat resource; see Schiel (2010)] based on a large variety of German conversational speech corpora. Using this as a reference, parts of the individual passages were manually rewritten to further refine the general phoneme balance as well as the balance among the passages.[4]

1. Validation of material

To validate the meaning-guided elicitation method applied in our study, to test whether speakers intend to speak the respective texts in the intended credible or ironic manner, and to check for potential cross-correlations to other related semantic dimensions, we conducted an online experiment with $N = 28$ Standard German-speaking participants (39.3% female, 60.7% male, mean age 47.4 years) recruited via the platform clickworker.com. According to our sample requirement, only native speakers aged 18–80 years from Germany, Austria, and Switzerland were admitted to the study.

In the course of the experiment, participants rated the two texts (presented in random order) using the following instruction: "Please read the text above first. Then imagine that you have to speak this text yourself. How would you do this?" The list of adjectives was derived from the General Music Branding Inventory (Lepa *et al.*, 2020), a self-developed and validated inventory to measure the semantic expression in musical and acoustical stimuli, and included the following adjectives: relaxed, happy, authentic, ironic, sarcastic, serious, credible, anxious, honest, joyful, dreamy, annoyed, friendly, stressed, bored, curious, dominant, passionate, surprised, intelligent, reliable, and likeable. The used scale was binary and included only a Yes and No option.

Results indicate that the credible text was rated as such, with an agreement of 89.3%. However, also other semantic properties achieved high agreement: for example: serious, 92.9%; honest, 89.3%; friendly, 85.7%; reliable, 82.1%; likeable, 82.1%. In contrast, only 3.6% of the participants reported that they would speak the credible text in an ironic manner. Results further confirmed that also the ironic text was rated as intended, with an agreement of 71.4%. Here, also other semantic attributes were elicited: for instance, joyful, 71.4%; happy, 64.3%; friendly, 60.7%.

The descriptive results already indicate that certain semantic properties might be inherently correlated in the specific context of a job interview. To further test this assumption, we conducted a Varimax-rotated PCA to explore the semantic dimensions underlying the presented adjectives. A scree plot hinted to a well-interpretable three-factor solution, whereby the first factor can be interpreted as credibility, with credible loading highest on this factor. Moreover, factor 2, including the highest loading of joyful and happy, might be interpreted as emotional valence, the first dimension of the Circumplex model by Russell (1980). Finally, factor 3, including "annoyed," "surprised," and (negative) dreaminess, might be associated with emotional arousal, the second dimension of the Circumplex model.

To validate whether listeners perceive the recordings obtained in the current study to reflect credible versus ironic speech, we conducted an online experiment in which 183 listeners (mean age 44.6 years, 78 female, 105 male), each rated 20 short utterances. Those utterances stem from a pool of 520 utterances produced by the 65 speakers (i.e., 4 credible and 4 ironic utterances per speaker) from the current production study. Participants rated audio excerpts on a 1–6 scale for the adjectives ironic and credible. As expected, the finding suggests that raters assign a higher credible rating for the credible speech than the ironic speech and a lower ironic rating for the credible speech than the ironic speech. The full data analyses of these two validation experiments are available in the OSF repository at http://doi.org/10.17605/OSF.IO/MTCUH.

C. Recordings and data collection

The recordings were conducted at a recording studio at the Institute of Sound and Vibration Engineering at Hochschule Düsseldorf. Speakers were recruited through word of mouth and mailing lists. A small set of additional recordings was conducted in the private homes of an author's friends and family members, all in all adding up to 65 participants (36.9% female, 61.5% male, 1.5% diverse), primarily consisting of students with a mean age of 27.2 years [standard deviation (SD) = 8.2 years]. The participants self-identified as speakers of Standard German. Thirteen participants (20%) reported on acting experience (whereas 80% did not), in particular related to school or amateur theatre groups. Furthermore, 14 participants (21.5%) reported speech impediments, primarily referring to stuttering and sigmatism during childhood. These impediments, however, were self-reported to not have any implications for their speech in the present. Based on this and on listening to the acoustic recordings, we refrained from excluding participants from the analysis.

The studio recordings were conducted using a t.bone (Burgebrach, Germany) condenser microphone, while the home recordings were done with a mobile setup using a Røde (Sydney, Australia) NT1-A condenser microphone, both ensuring a decently clean recording quality.[5]

D. Procedure

Participants were first given a quick briefing, informing them that they were about to be asked to read aloud three different short passages. They were told that the first one would serve as a calibration for the study and should just be read freely without too much thought (neutral passage). Participants were informed that, for the other passages, they would be asked to make themselves familiar with their content and their intention first by reading them silently and then, once they felt ready, to recite them in a manner that they would deem accurate for the situation displayed. They were also told that before the recording of each new passage, they would get a short, more detailed introduction of the respective situation. Participants were told that in case they misread a word or sentence, they should repeat the whole sentence in its entirety.

For the `credibility` passage, they were told that the situation should be imagined as a real job interview and that they would be a serious, ambitious candidate who wanted to appear professional to get the job. For the `irony` passage, they were told that the situation is still a job interview, but they are now a candidate who is not exactly serious about the situation, does not care much, and is more unprofessional. In all cases, they were also told that the passages were already written in a tone that should guide them and to recite the monologues in ways that felt natural to them in the provided contexts and tones. The `neutral` passage was always presented first, and the other two passages were presented in random order. After the recordings were finished, participants filled out a questionnaire concerning their demographic data, potential influences like recording and acting experience, speech impediments, and the Big Five Inventory (BFI-10) personality assessment (John *et al.*, 1991; Rammstedt *et al.*, 2014) (not the focus of this study).

E. Data processing and analysis

The speech samples were manually trimmed to remove silence at the beginnings and ends of the recording sessions, whereas natural speaking pauses were kept, and if repetitions were found, keeping only the last instance without speech errors or disfluencies. The recordings were processed by a slight denoising filter removing ambient and preamp noise as well as by some individual cleanup by removing particularly loud vocal pops or other unwanted noises using the audio cleanup tool iZotope RX 10 (iZotope, Inc., 2024) and rendered as normalized 44.1 kHz/16 bit PCM files. Finally, recordings were divided into 5-s-long chunks to increase statistical power. This is a common segmentation approach in speech classification tasks, such as stuttering event detection (e.g., Bayerl *et al.*, 2022 and Lea *et al.*, 2021) and speech emotion recognition (e.g., Kim and Provost, 2016 and Zhang *et al.*, 2020), where speech is modeled at a fixed length chunk level by dividing the original signal of arbitrary length into short segments with a predefined size. For an overview of more advanced segmentation approaches, see Lin and Busso (2023) and Barrett *et al.* (2024). As one central goal of this work was to analyze and interpret which acoustical features would be associated with the different semantic dimensions, a custom feature set tailored to the specific needs of this study was extracted using a combination of algorithms from the "LibROSA" PYTHON library v.0.10.2 (McFee *et al.*, 2023) as well as custom PYTHON scripts. The chosen features were partly some of the most commonly used ones for audio classification tasks like mel frequency cepstral coefficients (MFCCs), spectral centroid, or zero crossing rate, while others were more specifically tailored to the corpus based on expectations and initial observations made during the recording process, like pitch variance and speaking rate. The final feature set consisted of four main categories (pitch, MFCC, spectral, speed/percussive), containing a total of 28 main features (see Table I). Where applicable, feature values were aggregated using their minimum (min), maximum (max), SD, and mean values, adding up to a final total of 106 individual feature variables in the dataset. Also, all resulting feature variables were z-standardized across speakers and conditions. Thus, the final dataset contained 3441 observations over 106 feature variables for the three semantic categories and used for all further analyses. From this dataset, gender-split sub-datasets were derived, with 2173 observations in the male set and 1,328 observations in the female set.[6]

TABLE I. Acoustic features extracted from the recorded speech corpus.

Category	Features
Pitch	f_0 min, max, variance, mean
MFCC	MFCC 1-13 min, max, SD, mean
Spectral	Zero crossing rate, spectral centroid, spectral flatness, spectral roll-off, spectral bandwidth, spectral contrast (in 6 bands) min, max, SD, mean
Speed/percussive	Speaking rate, onset rate, onset strength, min, max, SD, mean

III. RESULTS

A. Feature reduction

First, a feature reduction method was applied to the total dataset to determine the smallest amount of relevant features that still provide a decent classification accuracy. To that end, a recursive feature elimination (RFE) algorithm from the "caret" package (Kuhn, 2008) in R (R Core Team, 2023) was used. The algorithm recursively eliminates features of a given feature set and target, based on a cross-validated statistical model: in our case, a random forest model. We chose a fivefold cross-validation with ten repetitions as parameters for the RFE. The RFE resulted in an optimal total of 85 features for the general dataset, which reported an accuracy of 68.6%.

Additionally, we used a knee-point detection algorithm (Satopaa *et al.*, 2011) to find the best trade-off between classification accuracy and amount of features in the RFE results for the general dataset. We selected a sensitivity value of 2 as parameter for the knee-point detection algorithm, which leads to a more conservative approach with detection that tends towards higher classification accuracy. The knee-point within the RFE results included 25 features and obtained a classification accuracy of 65.6%, features that were thus identified as most appropriate for the classification task.

B. Model selection

For statistical modeling and further evaluation of feature importance, a multinomial logistic regression from the NNET package (Venables and Ripley, 2002) in R (R Core Team, 2023) was used for each of the three datasets with the reduced set of input variables, resulting in three statistical models. Each of the three models combines multiple logistic regressions to determine the influence of the input variables on the output classes. The multiple output classes are examined in their relation to a common reference class, in this case relating the `credibility` class and the `irony` class to the `neutral` reference class. The multinomial logistic regression models were assessed by examining the performances in terms of goodness-of-fit, predictive ability, stability and generality. The in-sample McFadden's pseudo-R^2 (McFadden, 1973) was calculated as the measurement for goodness-of-fit, with values of 0.2–0.4 representing a good fit, unlike standard R^2. A repeated fivefold cross-validation with 20 repetitions was conducted for all the models to evaluate the predictive abilities and to stabilize the metrics of each evaluation. Based on the predictions within each fold of the cross-validation cycles, confusion matrices were determined to analyze the accuracies of the models in predicting unseen data. The matrices were then averaged over the validations and repetitions, leading to one matrix for each model. The out-of-sample-based McFadden's pseudo-R^2 was calculated for each cross-validation to further assess the predictive power and potential overfitting of the full models on the datasets according to

$$R_M^2 = 1 - \frac{L_{est.out}}{L_{null}}, \tag{1}$$

where $L_{est.out}$ denotes the log-likelihood as the sum of the logarithm of the probabilities that each predicted observation takes on its observed value within the cross-validations and L_{null} denotes the log-likelihood of the corresponding model with only intercept. The out-of-sample-based pseudo-R^2 values were averaged over the repetitions of each cross-validation. The coefficients of the evaluated models were then examined to find the most important features for all three datasets. The features for each class of the general dataset are provided in Table II, ranked by their absolute z-scored coefficient values, which are calculated in relation to the `neutral` reference class.

TABLE II. Final multinomial logistic regression model (general dataset) predicting `credibility` (CRD) and `irony` (IRO) compared to the `neutral` condition. All 25 predictors are in descending order of importance of the CRD model based on the absolute value.

Feature	CRD	IRO
`spectral_centroid_mean`	−1.213	−1.210
`mfcc3_mean`	−1.041	−1.419
`mfcc2_mean`	−0.959	−2.067
`mfcc2_stdev`	0.873	0.664
`speaking_rate`	0.421	0.880
`onset_strength_mean`	0.396	0.315
`mfcc3_stdev`	−0.374	−0.274
`mfcc8_mean`	−0.362	−0.781
`mfcc12_mean`	−0.313	−0.678
`(Intercept)`	0.292	0.111
`zero_crossing_rate_mean`	0.291	0.203
`mfcc4_stdev`	0.262	0.496
`spectral_roll-off_mean`	0.261	0.262
`pitch_mean`	−0.261	−1.077
`mfcc1_mean`	−0.231	−0.905
`spectral_bandwidth_stdev`	−0.230	−0.132
`spectral_flatness_min`	0.203	0.418
`mfcc9_mean`	0.190	−0.067
`mfcc5_min`	−0.158	−0.577
`mfcc1_max`	0.148	−0.225
`mfcc12_stdev`	−0.140	0.146
`spectral_contrast_6_mean`	−0.136	−0.041
`mfcc3_min`	0.135	0.306
`spectral_bandwidth_min`	0.112	−0.067
`mfcc7_mean`	−0.111	−0.324
`mfcc1_min`	0.013	−0.144

C. Model interpretation

The final model, as presented in Table II, achieves an averaged, out-of-sample-based McFadden's pseudo-$R^2 = 0.262$ and a mean accuracy of 62.6%, suggesting a good fit. It becomes obvious that the MFCCs are the feature group with the highest predictive power (number of included features, $N = 15$), followed by (other) spectral cues (e.g., the mean spectral centroid, $N = 7$), speed and percussive features (e.g., speaking rate, $N = 2$), and pitch ($N = 1$).

To further enhance the interpretability of the model, we computed the z-transformed top 25 features averaged across all recordings for the three conditions (general dataset), as shown in Fig. 1. It can be observed that, for instance, the mean values of the MFCC 1 (`mfcc_mean`) are highest under the `credibility` condition, followed by the `neutral` and `irony` conditions. Since the MFCC 1 is an indicator of the overall energy in an audio signal (Tracey *et al.*, 2023), it can be assumed that credible speech is, on average, louder than neutral and ironic speech, whereby the latter achieves the lowest energy values. Similar findings can be observed for the minimum and maximum overall energy (`mfcc1_min`, `mfcc1_max`), whereby credible speech reaches similar, above-average values and whereby ironic speech again obtains lower values.

As for the role of higher-order MFCCs, results show substantial differences regarding, for example, the mean values of the MFCCs 2 and 3 (`mfcc2_mean`, `mfcc3_mean`). As stated by Tracey *et al.* (2023), the MFCC 2 can be interpreted as a weighted ratio of low- to high-frequency energy and the MFCC 3 can be regarded a measure of the weighted ratio of low-mid and high-mid spectral components. Accordingly, neutral speech is shown to have highest relative low-frequency content, whereas under both the `credibility` and `irony` conditions, speakers might tend to increase the pitch and general spectral energy of their speech towards higher-frequency ranges, leading to lower MFCC 2 and 3 values. This is in line with the more straightforward interpretation of the averaged spectral centroid (`spectral_centroid_mean`) and the averaged spectral roll-off (`spectral_roll-off_mean`). The spectral centroid can be described as the "gravity center" of an audio signal's spectral energy, and similarly, the spectral roll-off is the cut-off frequency of a signal below which the dominant spectral energy is located (Lerch, 2012). Both the mean spectral centroid and the mean spectral roll-off achieve

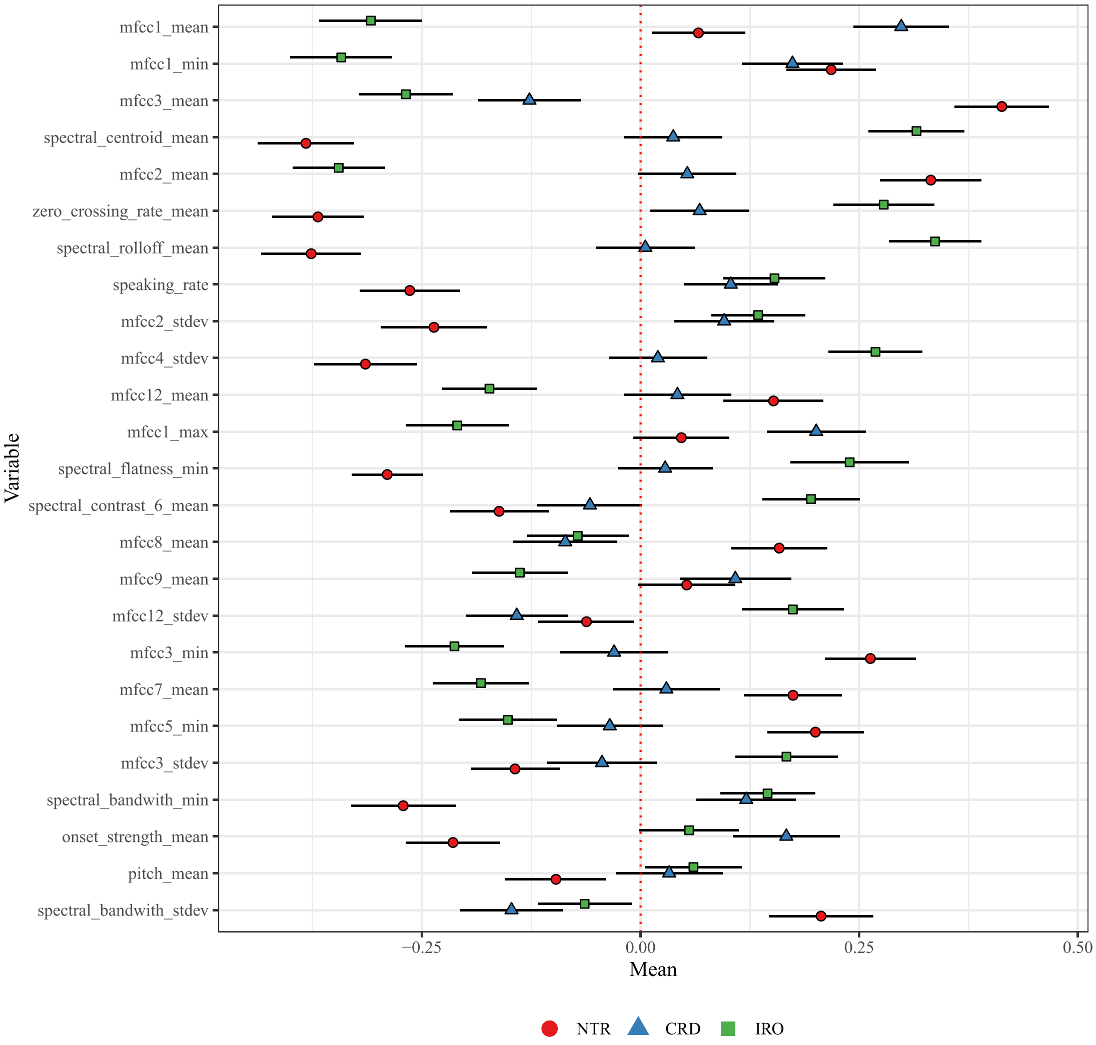

FIG. 1. Mean values of the top 25 features (z-scores) across recordings under the three conditions `credibility` (CRD), `irony` (IRO), and `neutral` (NTR) (general dataset).

highest values for ironic speech, followed by credible and neutral speech. This finding suggests that, when speakers try to convey irony, their speech contains more high-frequency components than in credible and neutral speech. This might be due to a generally higher pitch in terms of the fundamental frequency (f_0) and first formants (f_1 and f_2), but also due to a potential overemphasis of high-frequency consonants. Regarding the role of pitch, the averaged pitch (`pitch_mean`, i.e., f_0) indeed indicates higher values for ironic and credible compared to neutral speech, but no substantial difference across ironic and credible speech. The hypothesis of an overemphasis of high-frequency consonants, in contrast, might be supported by similar differences in the averaged zero-crossing rate (`zero_crossing_rate_mean`), indicating rapid changes in amplitude as occurring in high-frequency transient components, such as unvoiced consonants (Bachu *et al.*, 2010).

Finally, beyond timbral and pitch-related features, the speaking rate was selected by the RFE algorithm as one of the top 10 features. Here, our findings suggest that when speakers aim to convey irony or credibility, they tend to speak faster than under the `neutral` condition. However, no major difference can be observed between credible and ironic speech.

D. Gender and individual differences

We did not perform a gender-specific RFE in the first place, since it led to a decrease in model accuracy (potentially due to a loss to statistical power and small differences in feature importance across genders). However, to clarify potential gender differences regarding the effect size and direction of the features selected for the overall dataset, we computed the mean values of the top 25 features across conditions separately for female and male speakers (see the supplementary material). For this analysis, gender-diverse speakers were removed from both subsets of the data, leaving only female speakers or male speakers. It can be observed that, for most of the features, the effects described in Sec. III C are identical or highly similar across genders. However, in few cases, slight differences could be detected: for instance, in the case of the speaking rate, the SDs of

MFCC 2 and MFCC 3, the mean values of MFCC 12, and the SD of the spectral bandwidth. While an interpretation of the MFCC 12 is not straight forward, the gender-specific results indicate that women speak slightly faster under the irony than under the credibility condition, whereas no differences can be observed for men. In contrast, men tend to speak with a higher pitch under the irony than under the credibility condition, whereas women do not show such a difference. To assess potential individual differences in the effect of the speech style on a specific acoustic feature, we additionally built a linear mixed-effects model for each of the 25 most relevant features included within the RFE results, with the respective feature as dependent variable, semantic dimensions as fixed effect, and the ID of individuals as a grouping variable, i.e., random effect, while allowing random slopes for the fixed effect. We then calculated proportions of random slope variances (Goldstein *et al.*, 2002) as contributions to the total variance of the target variable to identify the magnitude of individual variability in the effect. Results show comparatively small proportions of variance, with means of 8.8% (min = 2.2%, max = 22.8%, SD = 5.5%) for random slopes of credibility and 14.8% (min = 3.8%, max = 33%, SD = 7.6%) for random slopes of the irony dimension over all 25 features, indicating that individual differences explain only a small fraction of the total variability in the acoustic features. The directions of these effects, however, are not consistent between individuals and tend to be equally distributed, with an average of 49% (min = 40%, max = 55.4%, SD = 4.9%) of the individuals who show a positive effect of credible speech and an average of 49.6% (min = 43.1%, max = 60%, SD = 4.2%) who show a positive effect of ironic speech on the acoustic features. Both the relatively small differences in the effect of speech style on specific acoustic features between individuals and the equally distributed directions of effect support the need for a set of acoustic features and their aggregate relationships to the target for the prediction of semantic dimensions.

IV. CONCLUSION

Under the framework of language expectancy theory, credibility is influenced by the degree to which expectations are met. Irony, in contrast, often involves a counterattitudinal statement that is not meant to be taken literally, with the speaker intending for the recipient to recognize its nonliteral nature. Speakers convey these meanings through various cues, including acoustic features. This study aimed to explore the acoustic features that differentiate credible speech from neutral and ironic speech. Previous studies have emphasized the role of traditional acoustic features like pitch, intensity, and speech rate in determining speech credibility. This study extended these findings by incorporating a broader set of features, including MFCCs, and employing methods from machine learning to select relevant features for better interpretability and to increase the power to predict unseen data to improve the generalization of our findings. Moreover, while past research has often relied on professional actors, potentially leading to exaggerated and inauthentic speech styles, this study used amateur speakers to ensure more ecologically valid results. The present work can be expanded to enhance the ecological validity in a number of ways, such as eliciting spontaneous speech that is not scripted, bringing in an interlocutor as the receiver of the speech (e.g., an experimenter or another participant acting as the hiring manager), and incorporating a reward component (e.g., additional financial compensation for how highly the speech is rated to be credible).

The study has undercovered various acoustic characteristics of credible speech (RQ1) and ironic speech (RQ2) compared from neutral speech. Concerning the acoustic characteristics of credible speech (RQ1), the list of the 25 most important features suggests that mel-frequency cepstral coefficients (MFCCs), particularly the MFCC 1, showed a strong predictive power for credible speech, suggesting that credible speech generally tends to be of greater intensity compared to neutral and ironic speech. This finding aligns with previous research that associates higher energy levels in speech with perceptions of credibility and trustworthiness (Chen *et al.*, 2020). One interpretation is that speech with greater intensity conveys a higher level of confidence (Jiang and Pell, 2017), a semantic dimension that is related to credibility.

Concerning the acoustic characteristics of ironic speech (RQ2), the findings suggest that it is characterized by a higher spectral centroid and roll-off, indicating a greater presence of high-frequency components. This observation is in line with previous findings by Bryant (2010) and González Fuente *et al.* (2016), who observed similar effects for English and French. The increase in high-frequency content compared to neutral speech might be due to the use of a generally higher pitch and overemphasis of certain consonants conveying a sarcastic or mocking tone. This can be verified in further research by examining the acoustics of individual phones or by conducting an ablation study with the classification models. These acoustic markers are essential for recognizing irony in speech, which often relies on subtle vocal cues to convey a meaning that contrasts with the literal words spoken.

Furthermore, compared to neutral speech (RQ1 and RQ2), our findings suggest that both credible and ironic speech tend to be faster. The finding of credible speech being faster supports previous research by Zuckerman *et al.* (1981) and Hartwig and Bond (2011), while the finding of ironic speech tends to be faster contradicts the results of many studies that observed ironic speech to be slower (or longer) (Adachi, 1996; Anolli *et al.*, 2000, 2002; Bryant, 2010; González Fuente *et al.*, 2016; Laval and Bert-Erboul, 2005; Milosky and Ford, 1997; Rockwell, 2000; Scharrer and Christmann, 2011). We, however, cannot offer a clear explanation for the contrasting findings.

In addition, this study examined gender-specific differences in the acoustic features associated with credibility (RQ3). It is noteworthy that most features showed similar effects across genders. These findings might be explained by the fact that speakers mainly consisted of students in their 20's, of which many might have politically liberal attitudes, and thus no longer conformed to the antiquated expectation that women should talk differently from men. However, some differences were noted. For instance, women tended to speak slightly faster under the irony condition compared to the credibility condition, whereas men showed no such difference. Conversely, men spoke with a higher pitch under the irony condition, a difference not observed in women. These findings suggest that gender-specific vocal traits can influence the perception and production of credible and ironic speech.

The study's methodological strengths include the use of a large set of acoustic features and a robust statistical modeling approach. Our *post hoc* analyses of individual differences for each of the 25 most relevant acoustic features suggest that the acoustic correlates of credibility and irony can be better understood by considering the aggregate relationships of the acoustic features, e.g., through the use of non-linear feature selection algorithm. The inclusion of MFCCs and other spectral features provides a nuanced understanding of the acoustic properties that differentiate credible speech. The use of a custom corpus with amateur speakers enhances the ecological validity of the findings. However, there are several limitations to consider. The study did not account for visual cues, which have been found to play a significant role in speech perception, particularly for irony (Kochetkova *et al.*, 2022). Additionally, the dataset was not balanced in terms of gender, which could introduce biases in the results. Future research should aim to include a more balanced sample and consider multimodal data to provide a more holistic understanding of speech credibility.

Future research should investigate the dynamic profiles of acoustic features, such as changes in pitch and intensity over time (Goupil *et al.*, 2021), which could provide deeper insights into the temporal aspects of credible speech. By examining a broader set of languages, future research can begin to evaluate the impact of cultural and linguistic differences on the perception of speech credibility, as these factors can influence vocal characteristics and listener interpretations (Andrist *et al.*, 2015; Castillo, 2011).

The findings of this study have practical implications for various fields, including communication training, forensic linguistics, and automated speech analysis. Understanding the acoustic features that contribute to speech credibility can help in training individuals to improve their persuasive communication skills, for instance as part of witness preparation training with criminal defendants (Boccaccini *et al.*, 2005). In forensic settings, these insights can assist in evaluating the reliability of spoken testimonies (Chapman, 1993; Gojkovich *et al.*, 2019). Moreover, automated systems for speech analysis can be enhanced by incorporating the identified features to better assess speech credibility in real-time applications, as demonstrated in emotion analysis by Pfister and Robinson (2011). Overall, this study contributes to the growing body of knowledge on speech credibility by highlighting the importance of specific acoustic features and offering new avenues for research and application in various domains.

SUPPLEMENTARY MATERIAL

See the supplementary material for the mean values of the top 25 features (*z*-scores) across recordings under the three conditions credibility (CRD), irony (IRO), and neutral (NTR) for the female dataset (SuppPub1.pdf) and the male dataset (SuppPub2.pdf).

ACKNOWLEDGMENTS

We thank Nele Mastracchio for help with typesetting the paper. J.S. and K.T. are the senior and co-corresponding authors. We follow the CRediT taxonomy (https://credit.niso.org/). J.S. contributed Conceptualization, Formal Analysis, Methodology, Project Administration, Resources, Supervision, and Writing—Original Draft Preparation. P.B. contributed Formal Analysis, Methodology, Software, Validation, Visualization, and Writing—Original Draft Preparation. M.S. contributed Data Curation, Investigation, Methodology, Formal Analysis, Software, Visualization, and Writing—Original Draft preparation. K.T. contributed Conceptualization, Formal Analysis, Methodology, Supervision, and Writing—Original Draft Preparation.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Ethics Approval

The experiment was conducted as part of the author's (M.S.) Bachelors' thesis. M.S.'s institution did not require formal ethics procedures for this type of non-invasive survey experiment as part of the thesis. However, we always followed standard ethics procedures. As part of the informed consent, we stated to the participants that their participation in the experiment was completely voluntary, they were given a chance to read the instructions before deciding to participate, if they had any questions (during or after the experiment), contact details of the experimenters were provided, and they could withdraw from the experiment at any stage of the experiment without any explanation. The participants were informed that the data cannot be copied and used for other purposes, besides research. After reading the consent form, the participants were asked to give explicit consent that they would like to take part in the experiment.

DATA AVAILABILITY

The data that support the findings of this study are available within the article and its supplementary material and are openly available in the Open Science Framework (OSF) repository at https://doi.org/10.17605/OSF.IO/MTCUH.

[1] Neutral speech is speech that has a neutral tone and lacks a particular paralinguistic dimension, such as credibility and irony.

[2] A comparison of the two passages can be found in the Open Science Framework (OSF) repository.

[3] The prompts used in ChatGPT can be found in the OSF repository (see Data Availability).

[4] All materials and phoneme distribution are available open access at https://doi.org/10.17605/OSF.IO/XZD7H.

[5] Any miniscule differences in sound should be insignificant since the acoustically conveyed credibility should presumably not depend heavily on recording quality.

[6] The dataset is available open-access at https://doi.org/10.17605/OSF.IO/XZD7H. The codes for data analyses are available open-access at https://doi.org/10.17605/OSF.IO/EG5K8.

Adachi, T. (**1996**). "Sarcasm in Japanese," Stud. Lang. **20**(1), 1–36.

Andrist, S., Ziadee, M., Boukaram, H., Mutlu, B., and Sakr, M. (**2015**). "Effects of culture on the credibility of robot speech: A comparison between English and Arabic," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI '15*, Portland, OR (Association for Computing Machinery, New York), pp. 157–164.

Anolli, L., Ciceri, R., and Infantino, M. G. (**2000**). "Irony as a game of implicitness: Acoustic profiles of ironic communication," J. Psycholinguistic Res. **29**(3), 275–311.

Anolli, L., Ciceri, R., and Infantino, M. G. (**2002**). "From 'blame by praise' to 'praise by blame': Analysis of vocal patterns in ironic communication," Int. J. Psychol. **37**(5), 266–276.

Appelman, A., and Sundar, S. S. (**2016**). "Measuring message credibility: Construction and validation of an exclusive scale," Journalism Mass Commun. Q. **93**(1), 59–79.

Averbeck, J. M. (**2010**). "Irony and language expectancy theory: Evaluations of expectancy violation outcomes," Commun. Stud. **61**(3), 356–372.

Averbeck, J. M., and Hample, D. (**2008**). "Ironic message production: How and why we produce ironic messages," Commun. Monogr. **75**(4), 396–410.

Bachu, R., Kopparthi, S., Adapa, B., and Barkana, B. (**2010**). "Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy," in *Advanced Techniques in Computing Sciences and Software Engineering*, edited by K. Elleithy (Springer Netherlands, Dordrecht, Netherlands), pp. 279–282.

Banse, R., and Scherer, K. (**1996**). "Acoustic profiles in vocal emotion expression," J. Pers. Soc. Psychol. **70**, 614–636.

Barrett, L., Tang, K., and Howell, P. (**2024**). "Comparison of performance of automatic recognizers for stutters in speech trained with event or interval markers," Front. Psychol. **15**, 1155285.

Bayerl, S. P., von Gudenberg, A. W., Hönig, F., Nöth, E., and Riedhammer, K. (**2022**). "KSoF: The Kassel State of Fluency dataset—A therapy centered dataset of stuttering," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022*, June 20–25, Marseille, France, edited by N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis (European Language Resources Association, Paris, France), pp. 1780–1787.

Beebe, S. A. (**1974**). "Eye contact: A nonverbal determinant of speaker credibility," Speech Teach. **23**(1), 21–25.

Belin, P., Boehme, B., and McAleer, P. (**2017**). "The sound of trustworthiness: Acoustic based modulation of perceived voice personality," PLoS One **12**(10), e0185651.

Boccaccini, M. T., Gordon, T., and Brodsky, S. L. (**2005**). "Witness preparation training with real and simulated criminal defendants," Behav. Sci. Law **23**(5), 659–687.

Bond, C. F. J., and DePaulo, B. M. (**2006**). "Accuracy of deception judgments," Pers. Soc. Psychol. Rev. **10**(3), 214–234.

Brilman, M., and Scherer, S. (**2015**). "A multimodal predictive model of successful debaters or how I learned to sway votes," in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, MM '15*, October 26–30, Brisbane, Australia, edited by X. Zhou, A. F. Smeaton, Q. Tian, D. C. A. Bulterman, H. T. Shen, K. Mayer-Patel, and S. Yan (Association for Computing Machinery, New York), pp. 149–158.

Bryant, G. A. (**2010**). "Prosodic contrasts in ironic speech," Discourse Processes **47**(7), 545–566.

Bryant, G. A., and Tree, J. E. F. (**2002**). "Recognizing verbal irony in spontaneous speech," Metaphor Symbol **17**(2), 99–119.

Burgoon, J. K., Birk, T., and Pfau, M. (**1990**). "Nonverbal behaviors, persuasion, and credibility," Human Commun. Res. **17**(1), 140–169.

Burgoon, M. (**1995**). "Language expectancy theory: Elaboration, explication, and extension," in *Communication and Social Influence Processes*, edited by C. R. Berger and M. Burgoon (Michigan State University Press, East Lansing, MI), pp. 29–52.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (**2005**). "A database of German emotional speech," in *Proceedings of Interspeech 2005*, Lisbon, Portugal (International Speech and Communication Association, Stockholm, Sweden), pp. 1517–1520.

Castillo, P. A. (**2011**). "Cultural and cross-cultural factors in judgments of credibility," Ph.D. thesis, Charles Sturt University, Wagga Wagga, NSW, Australia.

Chapman, V. V. (**1993**). "The effects of language style on the credibility of legal testimony," Ph.D. dissertation, Indiana University, Bloomington, IN.

Chen, X. L., Levitan, S. I., Levine, M., Mandic, M., and Hirschberg, J. (**2020**). "Acoustic-prosodic and lexical cues to deception and trust: Deciphering how people detect lies," Trans. Assoc. Comput. Linguistics **8**, 199–214.

Chung, C. J., Nam, Y., and Stefanone, M. A. (**2012**). "Exploring online news credibility: The relative influence of traditional and technological factors," J. Comput-Mediated Commun. **17**(2), 171–186.

da Rosa, C. W., and Otero, J. (**2018**). "Influence of source credibility on students' noticing and assessing comprehension obstacles in science texts," Int. J. Sci. Educ. **40**(13), 1653–1668.

De Luca, A., Clark, A., and Dellwo, V. (**2024**). "Numberlie: A game-based experiment to understand the acoustics of deception and truthfulness," in *Proceedings of Interspeech 2024*, Kos, Greece (International Speech and Communication Association, Stockholm, Sweden), pp. 3659–3663.

De Meo, A. (**2012**). "How credible is a non-native speaker? prosody and surroundings," in *Methodological Perspectives on Second Language Prosody: Papers from ML2P 2012 (CLEUP)*, edited by B. M. Grazia and S. Antonio (CLEUP, Padua, Italy), pp. 3–9.

De Meo, A., Vitale, M., Pettorino, M., and Martin, P. (**2011**). "Acoustic-perceptual credibility correlates of news reading by native and Chinese speakers of Italian," in *17th International Congress of Phonetic Sciences, ICPhS 2011*, August 17–21, Hong Kong, China (International Phonetic Association, London, UK), pp. 1366–1369.

Duden Editorial Team (**2005**). *Duden: Die Grammatik (Duden: The Grammar)*, 7th ed. (Duden Publishing House, Mannheim, Germany).

El Ayadi, M., Kamel, M. S., and Karray, F. (**2011**). "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognit. **44**(3), 572–587.

Enos, F. (**2009**). "Detecting deception in speech," Ph.D. thesis, Columbia University, New York, NY.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and Truong, K. P. (**2016**). "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," IEEE Trans. Affective Comput. **7**(2), 190–202.

Eyben, F., Wöllmer, M., and Schuller, B. (**2010**). "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia, MM '10*, Firenze, Italy (Association for Computing Machinery, New York), pp. 1459–1462.

Fagan, S. M. B. (**2009a**). *Introduction, 1–3* (Cambridge University Press, Cambridge, UK).

Fagan, S. M. B. (**2009b**). *Regional Variation, 214–243* (Cambridge University Press, Cambridge, UK).

Fairbanks, G. (**1960**). *Voice and Articulation Drillbook* (Harper & Brothers, New York), pp. 124–139.

Gojkovich, K. L., Reitz, N. L., Monstrola, V. N., Campbell, M. M., and Rivardo, M. G. (**2019**). "Effects of dress and speech style on credibility of co-witness and misinformation acceptance," N. Am. J. Psychol. **21**, 707–721.

Goldstein, H., Brown, W., and Rasbash, J. (**2002**). "Partitioning variation in multilevel models," Understanding Stat. **1**(4), 223–231.

González Fuente, S., Prieto Vives, P., and Noveck, I. A. (**2016**). "A fine-grained analysis of the acoustic cues involved in verbal irony recognition in French," in *Speech Prosody 2016*, edited by J. Barnes, A. Brugos, S. Shattuck-Hufnagel, and N. Veilleux (International Speech Communication Association, Stockholm, Sweden), pp. 902–906.

Gordon, R. G., Jr. (**2005**). *Ethnologue: Languages of the World*, 15th ed. (SIL International, Dallas, TX).

Goupil, L., Ponsot, E., Richardson, D., Reyes, G., and Aucouturier, J.-J. (**2021**). "Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature," Nat. Commun. **12**(1), 861.

Gu, Y., Li, X., Chen, S., Zhang, J., and Marsic, I. (**2017**). "Speech intention classification with multimodal deep learning," in *Advances in Artificial Intelligence*, edited by M. Mouhoub and P. Langlais (Springer International Publishing, Cham, Switzerland), pp. 260–271.

Hamilton, M. A., Hunter, J. E., and Burgoon, M. (**1990**). "An empirical test of an axiomatic model of the relationship between language intensity and persuasion," J. Lang. Soc. Psychol. **9**(4), 235–255.

Hartwig, M., and Bond, C. F., Jr. (**2011**). "Why do lie-catchers fail? A lens model metaanalysis of human lie judgments," Psychol. Bull. **137**(4), 643–659.

Hirschberg, J., Benus, S., Brenier, J. M., Enos, F., Friedman, S., Gilman, S., Girand, C., Graciarena, M., Kathol, A., Michaelis, L., Pellom, B. L., Shriberg, E., and Stolcke, A. (**2005**). "Distinguishing deceptive from non-deceptive speech," in *Proceedings of Interspeech 2005*, Lisbon, Portugal (International Speech Communication Association, Stockholm, Sweden), pp. 1833–1836.

Hovland, C. I., Janis, I. L., and Kelley, H. H. (**1953**). *Communication and Persuasion; Psychological Studies of Opinion Change* (Yale University Press, New Haven, CT).

Hovland, C. I., and Weiss, W. (**1951**). "The influence of source credibility on communication effectiveness," Public Opin. Q. **15**(4), 635–650.

Howell, J., Rooth, M., and Wagner, M. (**2017**). "Acoustic classification of focus: On the web and in the lab," Lab. Phonol. **8**(1), 16.

Huang, C.-H., Chou, H.-C., Wu, Y.-T., Lee, C.-C., and Liu, Y.-W. (**2019**). "Acoustic indicators of deception in Mandarin daily conversations recorded from an interactive game," in *Proceedings of Interspeech 2019*, Graz, Austria (International Speech Communication Association, Stockholm, Sweden), pp. 1731–1735.

Iriondo, I., Planet, S., Claudi Socoró, J., Alías, F., Monzo, C., and Martíınez, E. (**2007**). "Expressive speech corpus validation by mapping subjective perception to automatic classification based on prosody and voice quality," in *Proceedings of the 16th ICPhS (ICPhS XVI)*, Saarbrücken, Germany (International Phonetic Association, London, UK), pp. 2125–2128.

iZotope, Inc. (**2024**). "iZotope RX 10, [software]," https://docs.izotope.com/rx10/en/index.html (Last viewed July 19, 2023).

Jiang, X., and Pell, M. D. (**2017**). "The sound of confidence and doubt," Speech Commun. **88**, 106–126.

John, O. P., Donahue, E. M., and Kentle, R. L. (**1991**). *The Big Five Inventory—Versions 4a and 5* (Institute of Personality and Social Research, University of California, Berkeley, Berkeley, CA).

Kim, Y., and Provost, E. M. (**2016**). "Emotion spotting: Discovering regions of evidence in audio-visual emotion expressions," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI '16*, Tokyo, Japan (Association for Computing Machinery, New York), pp. 92–99.

Kirchhübel, C., and Howard, D. M. (**2013**). "Detecting suspicious behaviour using speech: Acoustic correlates of deceptive speech—An exploratory investigation," Appl. Ergon. **44**(5), 694–702.

Kochetkova, U., Evdokimova, V., Skrelin, P., German, R., and Novoselova, D. (**2022**). "Interplay of visual and acoustic cues of irony perception: A case study of actor's speech," in *Artificial Intelligence and Natural Language*, edited by V. Malykh and A. Filchenkov (Springer Nature, Cham, Switzerland), pp. 82–94.

Kochetkova, U., Skrelin, P., Evdokimova, V., and Novoselova, D. (**2021**). "The speech corpus for studying phonetic properties of irony," in *Language, Music and Gesture: Informational Crossroads: LMGIC 2021*, edited by T. Chernigovskaya, P. Eismont, and T. Petrova (Springer Singapore, Singapore), pp. 203–214.

Kreuz, R. J., and Link, K. E. (**2002**). "Asymmetries in the use of verbal irony," J. Lang. Soc. Psychol. **21**(2), 127–143.

Kuhn, M. (**2008**). "Building predictive models in R using the caret package," J. Stat. Soft. **28**(5), 1–26.

Laval, V., and Bert-Erboul, A. (**2005**). "French-speaking children's understanding of sarcasm," J. Speech Lang. Hear. Res. **48**(3), 610–620.

Lea, C., Mitra, V., Joshi, A., Kajarekar, S., and Bigham, J. P. (**2021**). "Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada (IEEE, New York), pp. 6798–6802.

Lepa, S., Herzog, M., Steffens, J., Schoenrock, A., and Egermann, H. (**2020**). "A computationa402l model for predicting perceived musical expression in branding scenarios," J. New Music Res. **49**, 387–402.

Lerch, A. (**2012**). *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics* (Wiley, Hoboken, NJ).

Levitan, S. I., An, G., Wang, M., Mendels, G., Hirschberg, J., Levine, M., and Rosenberg, A. (**2015**). "Cross-cultural production and detection of deception from speech," in *Proceedings of the 2015 ACM Workshop on Multimodal Deception Detection, WMDD '15*, Seattle, WA (Association for Computing Machinery, New York), pp. 1–8.

Leykum, H. (**2021**). "Voice quality in verbal irony: Electroglottographic analyses of ironic utterances in standard Austrian German," in *Proceedings of Interspeech 2021*, Brno, Czech Republic, edited by H. Hěrmanský, H. Cěernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlicek (International Speech Communication Association, Stockholm, Sweden), pp. 991–995.

Lin, W.-C., and Busso, C. (**2023**). "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," IEEE Trans. Affective Comput. **14**(2), 1215–1227.

McFadden, D. (**1973**). "Conditional logit analysis of qualitative choice behaviour," in *Frontiers in Econometrics*, edited by P. Zarembka (Academic Press, New York), pp. 105–142.

McFee, B., McVicar, M., Faronbi, D., Roman, I., Gover, M., Balke, S., Seyfarth, S., Malek, A., Raffel, C., Lostanlen, V., van Niekirk, B., Lee, D., Cwitkowitz, F., Zalkow, F., Nieto, O., Ellis, D., Mason, J., Lee, K., Steers, B., Halvachs, E., Thomé, C., Robert-Stöter, F., Bittner, R., Wei, Z., Weiss, A., Battenberg, E., Choi, K., Yamamoto, R., Carr, C., Metsai, A., Sullivan, S., Friesch, P., Krishnakumar, A., Hidaka, S., Kowalik, S., Keller, F., Mazur, D., Chabot-Leclerc, A., Hawthorne, C., Ramaprasad, C., Keum, M., Gomez, J., Monroe, W., Morozov, V. A., Eliasi, K., nullmightybofo, Biberstein, P., Sergin, N. D., Hennequin, R., Naktinis, R., beantowel, Kim, T., Åsen, J. P., Lim, J., Malins, A., Hereñú, D., van der Struijk, S., Nickel, L., Wu, J., Wang, Z., Gates, T., Vollrath, M., Sarroff, A., Xiao-Ming, Porter, A., Kranzler, S., Voodoohop, Gangi, M. D., Jinoz, H., Guerrero, C., Mazhar, A., toddrme2178, Baratz, Z., Kostin, A., Zhuang, X., Lo, C. T., Campr, P., Semeniuc, E., Biswal, M., Moura, S., Brossier, P., Lee, H., and Pimenta, W. (**2023**). "librosa/librosa: 0.10.0.post2," Zenodo. https://zenodo.org/records/7746972.

Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D. R., and Mccann, R. M. (**2003**). "Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment," Ann. Int. Commun. Assoc. **27**(1), 293–335.

Milosky, L. M., and Ford, J. A. (**1997**). "The role of prosody in children's inferences of ironic intent," Discourse Processes **23**(1), 47–61.

Nfissi, A., Bouachir, W., Bouguila, N., and Mishara, B. (**2024**). "Unveiling hidden factors: Explainable AI for feature boosting in speech emotion recognition," Appl. Intell. **54**(11), 7046–7069.

Niebuhr, O. (**2010**). "On the phonetics of intensifying emphasis in German," Phonetica **67**(3), 170–198.

O'Neal, G. S., and Lapitsky, M. (**1991**). "Effects of clothing as nonverbal communication on credibility of the message source," Clothing Textiles Res. J. **9**, 28–34.

OpenAI (**2023**). "ChatGPT (Version GPT-3.5) [software]," https://www.openai.com/ (Last viewed July 19, 2023).

Patel, B., Zhang, Z., McGettigan, C., and Belyk, M. (**2023**). "Speech with pauses sounds deceptive to listeners with and without hearing impairment," J. Speech Lang. Hear. Res. **66**(10), 3735–3744.

Pfister, T., and Robinson, P. (**2011**). "Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis," IEEE Trans. Affective Comput. **2**(2), 66–78.

Rammstedt, B., Kemper, C. J., Klein, M. C., Beierlein, C., and Kovaleva, A. (**2014**). "Big five inventory (BFI-10)," Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS), https://zis.gesis.org/DoiId/zis76.

R Core Team (**2023**). "R: A language and environment for statistical computing" (R Foundation for Statistical Computing, Vienna, Austria), https://www.R-project.org/ (Last viewed July 19, 2023).

Reichel, U. D. (**2012**). "PermA and Balloon: Tools for string alignment and text processing," in *Interspeech 2012, 13th Annual Conference of the International Speech Communication Association*, Portland, OR (International Speech Communication Association, Stockholm, Sweden), pp. 1874–1877.

Reichel, U. D. (**2014**). "Language-independent grapheme-phoneme conversion and word stress assignment as a web service," in *Elektronische Sprachverarbeitung 2014*, edited by R. Hoffmann (TUDpress, Dresden, Germany), pp. 42–49.

Rockwell, P. (**2000**). "Lower, slower, louder: Vocal cues of sarcasm," J. Psycholinguistic Res. **29**(5), 483–495.

Roettger, T. B. (**2019**). "Researcher degrees of freedom in phonetic research," Lab. Phonol. **10**(1), 1–27.

Russell, J. A. (**1980**). "A circumplex model of affect," J. Pers. Soc. Psychol. **39**(6), 1161–1178.

Santos, P. B., Beinborn, L., and Gurevych, I. (**2016**). "A domain-agnostic approach for opinion prediction on speech," in *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, Osaka, Japan, edited by M. Nissim, V. Patti, and B. Plank (The COLING 2016 Organizing Committee, Osaka, Japan), pp. 163–172.

Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (**2011**). "Finding a 'kneedle' in a haystack: Detecting knee points in system behavior," in *2011 31st International Conference on Distributed Computing Systems Workshops*, Minneapolis, MN (IEEE, New York), pp. 166–171.

Scharrer, L., and Christmann, U. (**2011**). "Voice modulations in German ironic speech," Lang. Speech **54**(4), 435–465.

Schiel, F. (**2010**). "BAStat: New statistical resources at the Bavarian archive for speech signals," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, edited by N. Calzolari (European Language Resources Association, Paris, France), pp. 1069–1076.

Schröder, A., Stone, S., and Birkholz, P. (**2017**). "The sound of deception—What makes a speaker credible?," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, August 20–24, Stockholm, Sweden, edited by F. Lacerda (International Speech Communication Association, Stockholm, Sweden), pp. 1467–1471.

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., Elkins, A., Zhang, Y., Coutinho, E., and Evanini, K. (**2016**). "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *Proceedings of Interspeech 2016*, San Francisco, CA (International Speech Communication Association, Stockholm, Sweden), pp. 2001–2005.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., and Kim, S. (**2013**). "The Interspeech 2013 Computational Paralinguistics Challenge: Social signals, Conflict, Emotion, Autism," in *Proceedings of Interspeech 2013* (International Speech Communication Association, Stockholm, Sweden), pp. 148–152.

Semwal, N., Kumar, A., and Narayanan, S. (**2017**). "Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models," in *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, New Delhi, India (IEEE, New York), pp. 1–6.

Sonderegger, M., and Soskuthy, M. (**2024**). "Advancements of phonetics in the 21st century: Quantitative data analysis," OSF Preprints, https://sciety.org/articles/activity/10.31234/osf.io/mc6a9.

Sperber, D., and Wilson, D. (**1981**). "Irony and the use-mention distinction," in *Radical Pragmatics*, edited by P. Cole (Academic Press, New York), pp. 295–317.

Syed, M. S. S., Stolar, M., Pirogova, E., and Lech, M. (**2019**). "Speech acoustic features characterising individuals with high and low public trust," in *2019 13th International Conference on Signal Processing and Communication Systems (ICSPCS)*, Gold Coast, QLD, Australia (IEEE, New York), pp. 1–9.

Tavakoli, S., Matteo, B., Pigoli, D., Chodroff, E., Coleman, J., Gubian, M., Renwick, M. E., and Sonderegger, M. (**2025**). "Statistics in phonetics," Annu. Rev. Stat. Its Appl. **12**, 133–156.

Tomaschek, F., Hendrix, P., and Baayen, R. H. (**2018**). "Strategies for addressing collinearity in multivariate linguistic data," J. Phon. **71**, 249–267.

Tracey, B., Volfson, D., Glass, J., Haulcy, R., Kostrzebski, M., Adams, J., Kangarloo, T., Brodtmann, A., Dorsey, E. R., and Vogel, A. (**2023**). "Towards interpretable speech biomarkers: Exploring MFCCs," Sci. Rep. **13**(1), 22787.

Venables, W. N., and Ripley, B. D. (**2002**). *Modern Applied Statistics with S*, 4th ed. (Springer, New York).

Villarreal, D., Clark, L., Hay, J., and Watson, K. (**2020**). "From categories to gradience: Auto-coding sociophonetic variation with random forests," Lab. Phonol. **11**(1), 6.

von Hohenberg, B. C., and Guess, A. M. (**2023**). "When do sources persuade? The effect of source credibility on opinion change," J. Exp. Polit. Sci. **10**(3), 328–342.

Zhang, S., Chen, A., Guo, W., Cui, Y., Zhao, X., and Liu, L. (**2020**). "Learning deep binaural representations with deep convolutional neural networks for spontaneous speech emotion recognition," IEEE Access **8**, 23496–23505.

Zhang, Z., McGettigan, C., and Belyk, M. (**2022**). "Speech timing cues reveal deceptive speech in social deduction board games," PLoS One **17**(2), e0263852.

Zuckerman, M., DePaulo, B. M., and Rosenthal, R. (**1981**). "Verbal and nonverbal communication of deception," in *Advances in Experimental Social Psychology*, edited by L. Berkowitz (Academic Press, New York), pp. 1–59.