American Psychological Association

Technology, Mind & Society
2021 Conference Proceedings

© 2021 The Author(s)
ISBN:

http://dx.doi.org/10.1037/xxxxxxx

A Measurement-based Foundation for AI Applied to the Audio of Interviews

Isaac Thompson[1] , Marc Cubrich[1], Madiha Qasim[1], Yizhen Egyn Zhu[2], Kristen Hassenkamp[1], Farshad Koohifar[1], Nick Koenig[1], and Nikki Dudley[1]

[1]Modern Hire

[2]North Carolina State University

The ability to express oneself verbally is essential to success in employment interviews and in a wide array of occupations. Yet, the exploration and measurement of audio features in video interviews remains largely unexplored. The current work reviews previous research of auditory features and relates them to pre-employment interviews, creating an operationalization of verbal expression, with both a definition and behavioral anchors. Across two pilot studies, we demonstrate the reliability of a pairwise comparison approach to the measurement of the construct and provide preliminary validity evidence. The construct and associated measurement techniques represent a measurement-driven approach that can create the foundation for the implementation of large-scale deep learning models that will automate the measurement of acoustic features as they relate to employment screening.

Keywords: 5-7 Verbal Expression, Artificial Intelligence, Machine Learning, Audio, Employment Interviews, Measurement, Employee Selection

Organizations are increasingly using advanced tools and technology to capture candidate responses to assessment questions in the hiring process in both audio and video formats. While artificial intelligence, machine learning, and deep learning have become increasingly integrated into the hiring process, the exploration of verbal expression and audio features in video interviews using these approaches remains unexplored. The ability to express oneself verbally is essential to the interview and many professions yet no standard of measurement exists. If verbal expression can be reliably measured and then that measurement replicated by computers, the resulting AI tools can aid in hiring decisions as part of the interview process or as coaching tools for virtual interactions such as conference calls.

Verbal expression is defined in disparate ways across different streams of research. Although these definitions cover a broad range of information regarding verbal expression and allude to its importance in many professions, an integration of the wide range of professional perspectives is needed to advance research and ultimately development of automated machine learning tools to measure it at scale. The current state of verbal expression research relies heavily on subjective definitions, and the subjectivity of this process poses a risk for decision making and measurement. Without a rigorous and consistent definition, any attempted measurement of the construct is likely contaminated by unrelated constructs, such as cultural similarity between speaker and listener, likability, and perceived organizational fit of the speaker (Dipboye, 1994). Further, taking a construct-driven approach to verbal expression can help ground such approaches in theory and machine-extracted features that are readily interpretable. To help guide the research and measurement of verbal expression, a definition is developed, and measurement method piloted.

Verbal Expression in Employment Interviews

Employment interviews are a common method through which employers evaluate candidates (Downs, 1969). Hiring managers draw conclusions about traits an interviewee may or may not have, such as their ability to work with others or desire to work. Oftentimes, these traits are assessed from the conversational component in interviews (Hollandsworth et al., 1979) and thus candidates who are better at expressing themselves usually have a marked advantage over other applicants that may be equally qualified but unable to express their proficiency as well (Prazak, 1969). On the flip side, verbal expression, called oral expression by O*NET, is considered a *critical* skill for 90% of occupations

This article was published [to be completed by publisher].
(insert author ORCiD or remove if not applicable)
The authors acknowledge the following people:
This research was supported in part by [funder agency and grant number]. This study was registered with . The authors have the following conflict of interest to disclose:

Correspondence concerning this article should be addressed to Name, Affiliation, City, State, Zip code, Country. Email:

(O*NET, 2020). Regardless of whether verbal expression is measured and then controlled for in statistical analyses to get a better sense of true skill or is itself considered a true expression of job relevant traits, the measurement of verbal expression in pre-employment assessment is highly applicable to current hiring and selection systems.

Exploring the Facets of Verbal Expression

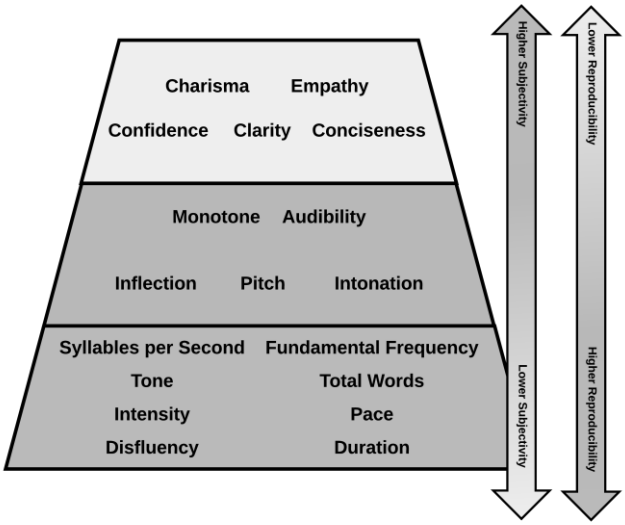
One of the demarcations of the diverse streams of research regarding verbal expression is the subjective vs. objective measures of auditory aspects. Objective measures are most often studied under the research of prosody, which examines the acoustic properties of speech, such as tone, pace, frequency, etc. Prosodic research is the study of those cues independently of the words being spoken (Wagner & Watson, 2010), and can focus on the concrete or physical definitions of the acoustic properties (e.g., pitch, duration, intensity, etc.) or the abstract evaluation of speech as a whole and how it influences listener interpretation of a message (Fujisaki, 1997; Ladd & Cutler, 1983). For example, prosody used when conveying information can alter both the interpretation and reception of a message based on variables such as tone, intonation, speaker, and other signals in the environment (Wilson & Wharton, 2006). Alternatively, and perhaps more subjectively, cognitive psychology and the employment interview literature, defines verbal expression as the ability to express oneself (Ericsson, 2002), clearly and convincingly (Huffcutt, 2011).

Reviewing the different aspects that could be measured in speech, Figure 1 is presented. Given the many ways in which verbal expression and audio features can be defined, conceptualized, and measured, the constructs and features represented in Figure 1 are not intended to be exhaustive. Rather, the figure provides illustrative examples of more subjective, psychological construct-driven approaches to conceptualizing verbal expression (e.g., charisma, empathy, etc.), and machine-extraction approaches that seek to capture and quantify aspects of sound waves. Holistically, the features represented at the bottom are less subjective and easier to extract mathematically with computers, whereas the higher-level features are more subjective and require more human discernment to measure or train machine learning algorithms on.

The constructs at the bottom represent objective, concrete, and quantifiable features of audio data. These features can be easily extracted using machines and reproduced with high reliability and fidelity. For example, fundamental frequency and intensity can be directly extracted from the sound wave derived from speech. Features at the lowest level of the hierarchy are the most objective and reproducible.

The mid-level constructs can also be extracted, but may require additional calculation, engineering, or programming. Compared to the low-level features, these constructs may not be considered purely objective properties. For example, although pitch may be quantified using fundamental frequency, it is considered a subjective psychoacoustical attribute of sound. The term "psychoacoustics" references insights from the cognitive psychology literature which suggest that personal expectations, prejudices, and predispositions have effects on listeners' evaluations and comparisons of audio. In other words, hearing is not a purely mechanical phenomenon, but an experience that is subject to sensation and perception. Different combinations of low-level objective features may be used to represent or serve as high-fidelity proxies for these mid-level constructs, but they remain subject to human judgment.

Figure 1. ***Audio Features and Constructs of Verbal Expression*.**



At the top of the hierarchy, these constructs represent high-level acoustic attributes of the voice. These constructs are the most difficult to extract and replicate and are likely to require many human ratings to train deep learning models. Conceptually, these constructs are likely to represent complex combinations of low- and mid-level features. For example, speech that is characterized as charismatic is likely to be represented by a complex combination of features such as pitch, inflection, and pauses. Further, the definition may vary across contexts and people. Research has suggested that human raters are reliable in their subjective judgments of verbal expression qualities such as charisma and empathy (Blanch-Hartigan, 2011; D'Errico et al., 2013; Jokisch et al., 2018), but the complex nature of these constructs makes them difficult to reproduce using machine learning. All these features must be considered to lay the foundation for the creation of job relevant and reliable measurement of verbal expression.

Rating Approach

In order to obtain machine predictions of verbal expression, a model structure needs to be specified and trained on a set of data labelled by humans, referred to as the training set. This training set is then used to estimate model parameters (e.g., regression coefficients) as a means to replicate these human labels. Building machine and deep learning models often involves amassing large amounts of reliably labeled data, and given the time

and resources needed to rate this amount of data, researchers often resort to hiring a large number of untrained workers whose proposed labels are averaged to reduce variance. As such, it is critical to consider the potential benefits of different rating approaches used to obtain the human labels.

Several methods have been proposed for reducing bias and overcoming calibration problems in human labeling. Pairwise ratings, in particular, offer a way to reduce bias and also demonstrate other benefits for raters. Compared to rating scale methods, ordinal methods (e.g., pairwise comparison) require workers to rank patterns as opposed to rating them (Chen et al., 2016). For ordinal methods, a pair of patterns A and B (e.g., two video interviews) is presented and the rater is asked whether value A < or > value B on the designated property (Chen et al., 2016).

Comparison-based ratings have several advantages over rating scale methods. First, comparison-based ratings naturally identify the strength of each example in the context of relational distance from the other examples (Junior et al., 2018). Further they do not require raters to establish the baseline or scales in their ratings (e.g., "what does a score of 3 on verbal expression mean?"). This approach avoids previously annotated samples from biasing future scores (Junior et al., 2018). Finally, research has provided evidence that ordinal rankings are easier than rating scale methods for untrained workers, and that workers are more engaged and less easily bored when they had to make comparisons rather than rating single items (Chen et al., 2016).

When a large amount of labeled data is required using an outsourced workforce, it appears that comparison-based ratings offer some promise above traditional ratings of single items. For these reasons, we describe a pilot study that investigates the use of the pairwise rating approach to differentiate verbal expression. Compared to rating scale methods that assign a fixed value, pairwise approaches can be used to map properties onto a continuum which more fully captures the complexity of verbal expression. Considering the potential benefits of pairwise comparison, the first pilot study investigates the comparability of the pairwise rating approach with the traditional BARS rating scale format. Building on these findings, the second pilot study uses the pairwise approach to test if meaningful differences of verbal expression can be discerned from interview responses, and whether these ratings are associated with job-relevant criteria.

Method

Operationalizing Verbal Expression

To arrive at an operational definition of verbal expression, multidisciplinary reviews of the relevant literature from cognitive psychology, linguistics, machine learning, and employment interviews were reviewed. Following the review, the definition and behaviorally based anchors were developed using an iterative process using a set of subject matter experts (n = 8). The iterative process followed was similar to developing Likert-type items. Subject matter experts (SMEs) read summaries of the literature, then provided definitions, and then tweaked others' definitions. Consensus was sought multiple times, through voting and discussions, and resulted in sometimes the expansion of scope and then the narrowing of it, as ideas were refined. For example, one of the decision points was around paralinguistic cues (e.g., facial expressions, hand gestures). While paralinguistics cues can impact the reception and effectiveness of an acoustic message, since they are not acoustic in their nature they were excluded from our conceptualization.

The end of the processes resulted in a definition and cues that were agreed upon by all SMEs. Verbal expression was defined as *the use of auditory features in the voice to increase understandability and engage listeners*. In addition to defining verbal expression, behaviorally based anchors for effective and ineffective verbal expression were delineated and are displayed in Table 1. These anchors include pitch, tone, speed, volume, understandability, pauses, and general cues.

Table 1. ***Verbal Expression Definition and Behavioral Scale Anchors***

Definition: Use of auditory features in the voice to increase understandability and engage listeners.	
Behaviorally Based Anchors	
Effective Indicators	Ineffective Indicators
Pitch: Proper modulation of voice and pitch to increase the reception of communication; Use of varying pitch range and intonation throughout speech.	**Pitch**: Uses little or no intonation in speech; Speech is monotone; The use of modulation and pitch detracts from meaning or understandability.
Tone: Uses tone that is appropriate for the message; Uses tone that matches the conveyed attitudes or feelings.	**Tone**: Uses tone that is inappropriate for the message; Comes across as bored or uninterested; Tone suggests inappropriate emotion for the situation.
Speed: Uses appropriate speed, pacing, and cadence to ensure understandability.	**Speed**: Speaks too fast, slow, or with too many pauses which diminishes understandability.
Volume/Understandability: Uses appropriate volume (not yelling or muted) when speaking; Clear pronunciation and articulation of words aiding understandability.	**Volume/Understandability:** Uses volume that is too loud or too soft, making it hard to listen to; Unclear or muted speech; Speech includes disruptive speaking patterns (e.g., filler words, unnatural pauses, grammatical errors).
Pauses: No unnecessary breaks in speech; effectively uses pauses to ensure understanding; Speech is not disfluent.	**Pauses**: Engages in the use of vocal pauses that detract from the message (stumbling, unnatural pauses); Several unnatural pauses in speech.
General cues: Use of cues make listeners want to listen; Engages listener with audio cues; Uses emotions and verbal cues to draw in the listener.	**General cues**: Uses cues that diminish understandability and reception of the message; Generally, makes it difficult to listen to; Listener must work hard to understand and follow the speech.

Pilot One: BARS vs. Pairwise

To test the measurement properties of the construct, both pairwise comparisons (comparing audio a and b and selecting the one higher in the ability) and BARS (Behaviorally Anchored Rating Scales) were rated on a small set (n = 10) of audio responses to a work styles pre-employment screening questions by four raters. Work styles are different than interviews as each candidate is essentially roleplaying a response to a structured assessment question. The candidate responses were gathered by a pre-employment screening for a real-world call center. The 10 audios were sampled to create a stratified sample of predicted verbal expression based on a previously created machine-learning model. Each rater was exposed to 45 combinations of audio clips and asked to select the audio higher in verbal expression according to the provided definition and anchors. The 180 comparisons (4 raters x 45 rated pairs) were used to compute Elo scores for each of the audios. Elo rating was originally developed to match chess players based on their previous matches and record but has been applied to a variety of contexts. After a comparison, or match, is made, points are exchanged based on the predicted probability of winning the match (Clark et al., 2018). We used a k-factor of 25 which represents the maximum amount an audio could move in ranking based on a single match. Correlational results showed that on this small set of data, BARS were essentially comparable to Elo scores derived from pairwise comparisons (r = .93).

Pilot Two: Pairwise on Interviews

A second pilot was conducted to test if meaningful differences of verbal expression could be decerned from interview responses. A small sample of interviews (n = 12) were rated with every possible pairwise comparison by four trained subject matter experts. With 64 total comparisons per rater, 256 comparisons were completed. Elo scores were then computed for the 12 interview responses. Consistency measurement of rater with self (Clark et al., 2018), a slightly more conservative rating of internal consistency than ICC was high, ranging from .83 to .92 on the four raters. Between rater correlations were .81 to .96 showing reliability between raters. Lastly, while not significant due to the small sample size, the derived Elo score related moderately to externally gathered supervisor ratings of communication effectiveness (r = .34). Table 2 provides a summary of the results from the second pilot study.

Table 2. ***Intercorrelations Between Raters, Elo Score, and Supervisor Rated Communication.***

Variable	1	2	3	4	5
1. Rater 1	(.83)				
2. Rater 2	.75**	(.92)			
3. Rater 3	.84**	.53*	(.83)		
4. Rater 4	.82**	.60*	.81**	(.87)	
5. Average Elo score	.96**	.81**	.89**	.90**	
6. Supervisor Rated Communication	.32	.35	.33	.22	.34

Note. Reliabilities reported on the diagonal represent consistency of measurement of the rater with their self; **Correlation is significant at the .01 level (2-tailed); *Correlation is significant at the .05 level (2-tailed).

Discussion and Future Direction

As virtual interviews and other virtual interactions increase in popularity, the need for assessing verbal expression grows. Specifically, having a concrete definition of verbal expression and a reliable method to assess it provides a novel trait that may someday be used in employee development and selection. Specifically, at scale, a large pool of human ratings based on a reliable measure of verbal expression could be used to train a machine to automate and scale its measurement. Secondly, while it may be technologically possible, the measurement must show both validity and a lack of group differences in real world use cases. These following issues are discussed below.

Using Deep Learning to Replicate Human Expertise

Machine learning and deep learning models have been widely adopted to make predictions to replicate human ratings on essays and have generally reached inter-rater reliabilities of over 0.7 (Rodriguez et al., 2019; Taghipour & Ng, 2016). However, there has been little research on machine replicating human ratings of verbal expression, or any voice-related constructs. One example of an effort to use voice to replicate human ratings is the ChaLearn competition (Ponce-López et al., 2016), in which participants attempted to parse out how the acoustic channel of videos contribute to apparent personality ratings. This conceptual model and measurement framework for verbal expression can help fill this gap. It not only extends the literature of replicating voice-related ratings, but also guides practices in predicting voice-related constructs. The conceptual information and pilots discussed above help to establish a foundation for this future research. Human ratings of verbal expression can, at scale, be used to train deep learning models that will be able to automate the extraction of verbal expression and use it in selection or development context.

Practical Considerations: Validity

Second, criterion-related validity must be established if verbal expression, as captured by humans or ultimately replicated by a machine, empirically links to on-the-job measures of performance or other organizationally relevant outcomes. The second pilot study hints at such promise with empirical relationships to on-the-job supervisor ratings but a larger data set is needed to fully explore

that relationship. There is the hope that acoustic features such as verbal expression offer an avenue for incremental validity beyond that of self-report or approaches that focus on the content of interviews, as verbal expression is a unique skill set independent of what is said. If there is incremental validity how much does verbal expression offer?

Practical Considerations: Fairness

Finally, if voice-related constructs are shown to demonstrate meaningful validity in an employment context, researchers must apply and analyze the results of the deep learning models to further our understanding of both group differences between populations in respect to verbal expression and the unintended impact those differences may have on hiring. Further, the mere perception of fairness from those who are evaluated by such technologies may matter as much as the reality of fairness. For these reasons, tools based on this technology may be better suited to provide developmental feedback rather than making hiring decisions on such information. These risks are currently theoretical, however, and need to be established through further research so interviewers can be aware of and work to mitigate any potential bias that may occur both by any machine learning programs that are developed and through hiring via other interview platforms that are currently popular

Conclusion

Pairing the proliferation of virtual human interactions with the rapidly increasing power of AI, there will soon come a time where technology will give us feedback about our interactions with each other. To maximize the potential positive application of this technology, namely job relevance, and to limit potential biases, a novel measurement-based construct-focused approach is proposed, developed, and piloted. This framework may assist in the development of robust data sets that will be used to train machine learning and deep learning algorithms to score virtual interviews, adding the unique variance of "how" a candidate responds rather than only "what" the candidate said. While the measurement foundation may provide a more legally sound approach being more job relevant than purely black box approach, more questions exist than answers. The next natural stage of this research is for deep learning methods to be trained on the raw audio to replicate these types of human ratings on constructs like verbal expression and then subjecting those predictions of human judgment to the rigors of selection science, namely that they show empirical evidence of fairness and job relevance. Ultimately, how we humans express ourselves will be quantifiable and it is in the best interest of society and research to proactively investigate, discuss, and guide this development.

References

Blanch-Hartigan, D. (2011). Measuring providers' verbal and nonverbal emotion recognition ability: Reliability and validity of the Patient Emotion Cue Test (PECT). *Patient education and counseling*, *82*(3), 370-376.

Butcher, M. (2019, April 23). *The robot-recruiter is coming - VCV's AI will read your face in a job interview.* https://techcrunch.com/2019/04/23/the-robot-recruiter-is-coming-vcvs-ai-will-read-your-face-in-a-job-interview/

Chen, B., Escalera, S., Guyon, I., Ponce-López, V., Shah, N., & Simón, M. O. (2016, October). Overcoming calibration problems in pattern labeling with pairwise ratings: application to personality traits. *European Conference on Computer Vision, Amsterdam, Netherlands, 9915.* doi:10.1007/978-3-319-49409-8_33

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM*. doi:10.1145/2939672.2939785

Clark, A. P., Howard, K. L., Woods, A. T., Penton-Voak, I. S., & Neumann, C. (2018). Why rate when you could compare? Using the "EloChoice" package to assess pairwise comparisons of perceived physical strength. *PloS one*, *13*(1), e0190393.

D'Errico, F., Signorello, R., Demolin, D., & Poggi, I. (2013). The perception of charisma from voice. A cross-cultural study. In *Proc. humaine association conference on affective computing and intelligent interaction*, Geneva, Switzerland (pp. 552-557). http://dx.doi.org/10.1109/ACII.2013.97

Dipboye, R. L. (1994). Structured and unstructured selection interviews: Beyond the job-fit model. *Research in Personnel and Human Resources Management*, *12*, 79-123.

Doughty, H., Damen, D., & Mayol-Cuevas, W. (2018). Who's Better? Who's Best? Pairwise Deep Ranking for Skill Determination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6057-6066).

Ericsson, K. A. (2002). Towards a procedure for eliciting verbal expression of non-verbal experience without reactivity: Interpreting the verbal overshadowing effect within the theoretical framework for protocol analysis. *Applied Cognitive Psychology*, *16*(8), 981-987.

Ericsson, K. A. (2002). Towards a procedure for eliciting verbal expression of non-verbal experience without reactivity: Interpreting the verbal overshadowing effect within the theoretical framework for protocol analysis. *Applied Cognitive Psychology*, *16*(8), 981-987.

Fujisaki, H. (1997). Prosody, models, and spontaneous speech. In *Computing prosody* (pp. 27-42). Springer, New York, NY.

Harwell, D. (2019, October 25). *A face-scanning algorithm increasingly decides whether you deserve the job*. https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job

Hollandsworth Jr, J. G., Kazelskis, R., Stevens, J., & Dressel, M. E. (1979). Relative contributions of verbal, articulative, and nonverbal communication to employment decisions in the job interview setting. *Personnel Psychology*, *32*(2), 359-367.

Huffcutt, A. I. (2011). An empirical review of the employment interview construct literature. *International Journal of Selection and Assessment*, *19*(1), 62-81.

Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology*, *83*(2), 179-189.

Jokisch, O., Iaroshenko, V., Maruschke, M., & Ding, H. (2018). Influence of age, gender and sample duration on the charisma assessment of German speakers. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2018*, 224-231.

Joo, J., Steen, F. F., & Zhu, S. C. (2015). Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *Proceedings of the IEEE international conference on computer vision* (pp. 3712-3720).

Junior, J. C. S. J., Jacques, C., Güçlütürk, Y., Pérez, M., Güçlü, U., Andujar, C., ... & van Gerven, M. A. (2018). First impressions: A survey on computer vision-based apparent personality trait analysis. *arXiv preprint arXiv:1804.08046*.

Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology*, *100*(1), 128-161.

Ladd, D. R., & Cutler, A. (1983). Introduction. Models and measurements in the study of prosody. In *Prosody: Models and measurements* (pp. 1-10). Springer, Berlin, Heidelberg.

National Center for O*NET Development. O*NET OnLine Help: Data Collection Information. *O*NET OnLine*. Retrieved October 12, 2020, from https://www.onetonline.org/help/online/data

Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, ... Escalera, S. (2016). Chalearn lap 2016: First round challenge on first impressions-dataset and results. *European Conference on Computer Vision*, *9915*, 400-418.

Prazak, J. A. (1969). Learning job-seeking interview skills. *Behavioral counseling: Cases and techniques*, 414-428.

Reilly, N. P., Bocketti, S. P., Maser, S. A., & Wennet, C. L. (2006). Benchmarks affect perceptions of prior disability in a structured interview. *Journal of Business and Psychology*, *20*(4), 489-500.

Rodriguez, P. U., Jafari, A., & Ormerod, C. M. (2019). Language models and automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1882-1891).

Taghipour, L., & Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, *25*(7-9), 905-945.

Wilson, D., & Wharton, T. (2006). Relevance and prosody. *Journal of Pragmatics*, *38*(10), 1559-1579.

Wolgast, S., Björklund, F., & Bäckström, M. (2018). Applicant Ethnicity Affects Which Questions Are Asked in a Job Interview. *Journal of Personnel Psychology*.

Yao, T., Mei, T., & Rui, Y. (2016). Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 982-990).