A blueprint for using deepfakes in sociolinguistic matched-guise experiments

Nathan Young[1]*, David Britain*[2] *and Adrian Leemann*[3]

[1]Department of Linguistics and Scandinavian Studies, University of Oslo, Norway
[1]Centre for Research on Bilingualism, Stockholm University, Sweden
[2]Department of English, University of Bern, Switzerland
[3]Center for the Study of Language and Society, University of Bern, Switzerland
n.j.young@iln.uio.no, david.britain@unibe.ch, adrian.leemann@unibe.ch

Abstract

Matched-guise paradigms, which are used extensively in speaker and accent evaluation studies, have long been hampered by empirical holes. We offer a solution by incorporating *deepfake* technology, which greatly reduces the number of potential confounds. We constructed a sociophonetic experiment whereby high-rising terminal (a.k.a. "uptalk") – and the lack thereof – was superimposed onto a deepfaked "beautiful" and "less beautiful" female guise. The resulting four guises were incorporated into a 2x2-factor between-subjects experiment tested on female evaluators. Each evaluator assessed their respective guise against a list of prescribed attributes and offered free-form comments. Results align with studies on high-rising terminal as well as intuitions concerning conventional beauty, which validates the technique and motivates its wider adoption.
Index Terms: sociophonetics, matched guise, methodological innovations

1. Introduction

Our contribution is the use of *deepfake* technology for accent and speaker-evaluation studies. A deepfake is a video that has been digitally manipulated to portray someone as doing or saying something that was not actually done or said (detailed further in Sec. 2.3). The incorporation of deepfake technology into experimental sociolinguistics relates to two overlapping research strands. "Strand 1" seeks to describe and understand ideologies about accents and accent features. "Strand 2" seeks to describe and understand how social and contextual cues influence evaluations of linguistic variables. As we describe below, both strands of research have relied on experimental techniques that introduce too many confounds.

1.1. Strand 1: Evaluation studies of accents

A classic example of this sort of experiment will have one guise with working-class speech and another guise with middle-class speech; evaluators will then assess them in a between or within-subjects fashion [1]. The experimental paradigm is referred to as the *matched guise* technique [2]. Here, a sole actor produces several guises, i.e. accents. In this sense, the designation "matched" refers to the fact that each guise matches back to the speaker. These guises are played to listeners, sometimes along with "distractor voices", all of which are rated on an attribute, usually with a Likert scale (e.g., aesthetics, prestige, employability etc. [3, 4, 5, 6]). The benefit of this design is relative control of acoustic features such as f_0, voice quality, or speaking rate, which – if left uncontrolled – may bias listener responses [7]. The pitfall, of course, is that (a) actors need to be very skilled impersonators and (b) listeners may pick up on the fact that one speaker produced most of the stimuli.

To avoid the complications and confounds of matched guises, researchers introduced so-called "verbal guises" [8], where each guise is produced by different native speaker(s) of each variety or accent. This circumvents the imitation and speaker-identification problem, yet eliminates the control of acoustic features, introducing other potential confounds. For example, a male speaker of accent X may inadvertently have a low f_0 while a male speaker of accent Y has a high f_0. If one wished to test "pleasantness" ratings between accent X and Y, the experiment would lack sufficient control because other research shows that lower f_0 in males is perceived as more pleasant [7]. There is little recourse for the high number of potential confounds that different speakers introduce.

1.2. Strand 2: Social cues and linguistic evaluation

A classic example of this sort of experiment will play the same audio sample for two evaluator groups while showing a picture of a black person to one group and a picture of a white person to the other group [9]. This is a more recent strand of research that – with the help of the matched-guise technique – investigates the extent to which language attitudes/perception are biased by various social and contextual cues. Niedzielski [10], who pioneered this approach, found that subjects identified vowels as being more Canadian-like or Michigan-like when their experiment answer-sheet had 'Canada' respectively 'Michigan' written at the top – even though they were being exposed to the very same vowel. Similarly, Hay and Drager [11] found that hearers were more likely to identify a vowel as Australian when they had seen a toy koala in the room and as a New Zealand vowel with a toy kiwi. Squires [12, pp. 230–231] showed that such cues also shaped perception of morphosyntactic non-standardness; subjects shown a picture of both a "low" and "high"-status speaker were more likely to assume that non-standard utterances were uttered by low status-appearing speakers. Others have shown that age [13], gender [14], race [15] and persona [16] all can shape linguistic perception. The stimuli in this research, however, have been textual or static images or physical objects (e.g. toys), all of which are physically disconnected from the speech signal that subjects have evaluated.

2. Guise construction

We conducted a 2x2-factor experiment whereby we altered a single linguistic variable – high-rising terminal (HRT) or the lack thereof (low-falling terminal, LFT) – and a single social variable – conventional beauty ("beautiful") or less of it ("less beautiful"). We built a female guise and elicited evaluations from female subjects to control for gender-related extraneity.

Figure 1: *Four experimental guises were constructed using deepfake technology for the visual element and using f_0 synthesis for the audio element: (1) “beautiful” with low-falling terminal H*L-L% (LFT), (2) “less beautiful” with LFT, (3) “beautiful” with high-rising terminal L*H-H% (HRT), (4) “less-beautiful” with HRT.*

2.1. Deciding on variables: HRT and “beauty”

HRT is popularly known as “uptalk” and was first researched in New Zealand [17] and Australian English [18]. It is also commonly associated with the “Valley Girl” persona in Southern California [19, 20]. It has since spread to the U.K., and we have sought to investigate it in Southern British English. A key reason why we selected this specific feature for this experiment is (a) because of its relative popularity in the sociolinguistic literature, which gives us references to use for validation, and (b) suprasegmentals are easier to superimpose onto videos than segmentals. On point a, the literature has tied HRT use *in production* to politeness [19, 17], a rhetoric strategy for engaging listeners [21], a floor-holding technique [22], and a way to introduce new information [22]. Popular ideologies associate it with “ditziness” [23], and research on evaluations partly substantiates this. HRT correlates with evaluations of *attentive* [18] (but also *inattentive* [20]), *excited* [20], *expressive* [18], *friendly* [18], *happy* [20], *less certain* [24] (but also *more certain* [20]), less *confident* [18, 24] (but also *more confident* [20]), less *forceful* [18], *unintelligent* [20], *unprofessional* [18], and *youthful* [18, 20]. On point b, HRT is particularly suitable for video guises because manipulating the mouth is not required. In contrast, a variable like /θ/, which has [f] and [θ] as variants in the U.K., would require additional visual manipulation.

Conventional beauty was selected [25, 26, 27] because, while indeed also being a sensitive topic of inquiry [28], it is less sensitive than superimposing gender or race onto a guise by means of deepfake technology (see, e.g., [29]). Echoing recent discussions on the ethics of deepfakes, we call for a measured approach when deploying our technique in research [29, 30].

2.2. Constructing HRT and LFT guises

Figure 1 offers a visual aid for the four guises we constructed. A 74-second excerpt was extracted from a video of a young female YouTube “influencer” talking about makeup. We ensured that (a) she, herself, was a user of HRT, (b) one exemplar of HRT was present in the excerpt, and (c) one exemplar of LFT was present in the excerpt. We then used those HRT and LFT contours as respective templates for simulating and applying them to the remainder of her intonational phrases (henceforth, “IP”). We also ensured that the excerpt did not include *listing* [31, 20].

The soundtrack was extracted into a wav file using *MediaHuman Audio Converter*, transcribed within ELAN [32], and phonetically time-aligned with the *Montreal Forced Aligner* [33]. The phonetically time-aligned file allowed us to demarcate IP’s – defined as continuous speech located before a pause of $\geq 70ms$ [34, p. 340].

Boundary tones have differing numbers of unstressed syllables following the nuclear stress accent, which hindered the straightforward deployment of the LFT and HRT templates. The literature presents two options for the H% portion of HRT: (a) a steady rise or (b) a rise followed by a plateau (see, e.g., figures on p. 154 in [35]). It is unclear whether this distinction is phonotactic, typological, or a combination of the two. It is therefore also unclear what rules govern the timing of the plateau in, among other constraints, brief versus extended syllabic ecologies. We therefore experimented with what sounded subjectively natural to us while adhering as closely as possible to the naturally-occurring HRT and LFT templates:

HRT We used a template of L*H-H% by constructing a valley at 234 Hz midway into the ultimate nuclear pitch accent of the IP. Then, if no unstressed vowels followed the ultimate stress, we built a rise that continued linearly up to 347 Hz at the boundary. If one unstressed vowel followed the ultimate nuclear pitch accent, we also had the rise continue linearly to 347 Hz at the boundary. If, however, two or more unstressed vowels followed the ultimate nuclear pitch accent, we had the rise continue linearly to 347 Hz at the midpoint of the second unstressed vowel.

LFT We used a template of H*L-L% that started with whichever level the frequency naturally occurred at. We then constructed a valley at 188 Hz midway into the ultimate nuclear pitch accent. We then constructed a subsequent plateau.

The 74-second recording had 28 separate IPs, and the above templates were manually deployed onto duplicates of each IP to make one “HRT version” and one “LFT version”. All IPs were adjusted using the *Manipulate* function in *Praat* [36]. The two template-IPs that were not adjusted were also run through

the Manipulate function in order to render synthetic every IP in both guises. Some of the synthetically-produced HRT tones had a robotic tinniness that we alleviated with the program *TwistedWave*. The final LFT guise had LFT contours on all 28 of the IPs, and the final HRT guise had HRT contours on 21 of the 28 IPs and LFT contours on the remaining seven IPs.

2.3. Constructing "beautiful" and "less beautiful" guises

We harnessed what is referred to as "cheapfake" [37] technology to superimpose a "beautiful" and "less-beautiful" face onto our 74-second clip. We used the web application *deepfakesweb.com* and outline below why we used cheapfakes instead of more sophisticated deepfake technology. We also discuss how we selected faces for superimposition.

Cheapfake technology A "cheapfake" is the product of an out-of-the-box application that produces low-quality deepfakes [37]. We offer our own definition of "deepfake": technology that triangulates one face in a video, triangulates a second face in another video, and then superimposes the graphical surface of the second face onto the graphical axes of the first video, thereby rendering the illusion that the second face is moving as the first. This technology relies on machine-learning, and its computational expense increases exponentially for every additional axis added to the triangulation. This means that the most advanced techniques were unavailable to our team – especially since we needed to trial-and-error a large number of faces.

Face selection We conjectured a priori – based on emic knowledge of the U.K. – that our selected influencer was somewhat above-average in terms of conventional standards of beauty there. We therefore sought faces that were highly above average and highly below average according to those same conventional standards. This involved mining YouTube, collecting various faces, and incorporating them into our cheapfakes on a trial-and-error basis. Some faces fit quite realistically onto our selected influencer, and others did not. We therefore made our selection of the "beautiful" and "less-beautiful" faces based solely on *which of each was the most realistic*.

The final faces looked like *blends* of the base and the veil face, which made the substrate identities unidentifiable. To make the videos even more realistic, we then sent them to a visual effects specialist who doctored them. Since the resulting "less-beautiful" guise did not appear sufficiently below-average in beauty, we also had the editor design a blemish mask. Figure 1 on page 2 contains screenshots of the final two video guises.

3. Validation – analysis

3.1. Silent video guise validation

Procedure Because we are incorporating two innovations into the same study, we first validated the cheapfake component by testing it without sound. We created a survey that randomized the same silent 20-second excerpt from the "beautiful" and "less-beautiful" guise. We asked 155 female participants to rank the video on a five-point Likert scale according to the question *"Based on your knowledge of the real world, how likely could this person work in a job that requires looks and beauty?"*.

This specific formulation was designed to externalize the assessment from the participant, which we hoped would achieve two goals. First, we intended to minimize social desirability bias, which would be quite high for something as taboo as asking participants to rate a stranger on their beauty. Second, we wanted to distinguish conventional beauty from personal attraction. In other words, one might find certain "niche types" physically attractive despite knowing that the greater population would not necessarily agree.

We also asked participants *"Do you have any other observations that you would like to share?"* in order to provide an opportunity for them to identify the video as a fake if needed.

Participant selection To control for social extraneity, we prescreened participants to ensure they came from as uniform a social profile as possible. We recruited through our own personal networks and through the services *Prolific* and *Call for Participants*. Prescreening criteria were: (a) female, (b) ages 18–29, (c) lives in a dialectally-leveled region of England (London, Southwest, Southeast, East), (d) has an average household income of £50,000 or more, (e) has English as a first language, (f) and spent the majority of life in England. Eighty participants viewed the beautiful guise; 75 participants viewed the less-beautiful guise. We also asked *"Did you recognize this person?"*, and all participants answered "no".

3.2. Video and sound guise validation

Procedure After validating the beauty differential between the two silent guises, we sought to validate the audio-visual combination by building a survey that included the full 74-second guise complete with audio and video. We deployed 193 *new* participants to evaluate one of the four guises in a between-subjects paradigm, *viz.* only one participant saw one randomly-selected guise. The evaluations were made on a Likert scale according to the prescribed attributes listed in Table 2. We then compared these evaluations to popular ideologies surrounding HRT and ideologies concerning conventional beauty in order to see whether they were in concordance or not. As we detail in Section 4.2, we found that they were. We also asked participants to provide open-ended comments about the actual video in the hope that those who could identify it as a "deepfake" would.

Participant selection Participants were recruited the same way as in Section 3.1. To ensure that our participants were attentive to the video-watching task, we asked two questions on the video's content. Three participants failed the attention check, leaving us with 190 participants in total (beautiful HRT: $n = 53$, beautiful LFT: $n = 45$, less-beautiful HRT: $n = 47$, less-beautiful HRT: $n = 45$).

Attributes tested We asked participants to rate the person in the video according to a list of attributes that we selected from other projects [38, 39, 20]. We list them in Table 1, column 1. A five-point Likert scale was provided under each attribute with 5 for *"very"* and 1 for *"not at all"*. The attributes that showed a statistically-significant difference between beautiful and less-beautiful are indicated with the diagnostic from a Wilcoxon rank sum test in column 2. The attributes that showed a statistically-significant difference between HRT and LFT are indicated with a diagnostic in column 3.

4. Validation – results

4.1. Silent video guises

The mean beauty assessment for the "beautiful" guise was 4.5 out of 5[1], and the mean beauty assessment for the "less-beautiful" guise was 3.7 out of 5, which is why we use the language "less beautiful" instead of "not beautiful". We modeled

[1]Controversy circulates on the use of means for ordinal data. We believe, however, that the granularity of a mean makes it more heuristically accessible than a median. We take, however, a more conservative position on our diagnostics by using the non-parametric Wilcoxan rank sum test. We refer the reader to [40] and [41] for further discussion.

Table 1: *Attribute evaluations of video-with-audio guises*

attributes	HRT vs. LFT	beautiful vs. less beautiful	interpretation and mean Likert scores
ambitious			
annoying			
articulate		$p < 0.05$	beautiful more (3.9) than less beautiful (3.7)
complex			
confident		$p < 0.05$	beautiful more (4.3) than less beautiful (4.1)
down to earth			
educated	$p = 0.00$		LFT more (3.5) than HRT (3.2)
formal	$p < 0.05$		LFT more (2.2) than HRT (2.0)
generous			
hardworking			
independent			
intelligent	$p < 0.05$		LFT more (3.4) than HRT (3.2)
kind			
mature			
outgoing			
polite			
reliable			
tough			
trendy		$p = 0.00$	beautiful more (4.2) than less beautiful (3.6)

the assessments in a Wilcoxon rank sum test, and the difference was statistically significant with a diagnostic of $p < 0.0001$. As it pertains to the question *"Do you have any other observations that you would like to share?"*, three participants identified the less-beautiful guise as a deepfake, and no participants wrote anything in about the beautiful guise.

4.2. Video and sound guises

As we flag in Section 3.2, we conducted a separate series of Wilcoxon rank sum tests between HRT and LFT guises and a separate series of Wilcoxon rank sum tests between beautiful and less-beautiful guises. Of the 19 attributes tested, six rendered significant results, three for HRT and three for beauty, shown in Table 1. For HRT vs. LFT, the attributes *formal*, *educated*, *intelligent* are favored when LFT is present. For beauty vs. less beauty, the attributes *articulate*, *confident*, *trendy* are favored when more beauty is present.

5. Discussion

5.1. Silent video guises

Section 4 demonstrates a clear distinction in beauty evaluations for the silent deepfake guises, which implies success. However, three participants were able to identify the "less beautiful" guise as fake. This indicates that the experiment did not fully live up to its confound-free goal. However, we would argue that this confound is preferable over the sorts of confounds that arise from using entirely different individuals. What we mean is that we can somewhat contain the fallout of this error by removing those three participants. There will always remain the chance that others may also have silently noticed the video to be fake, but this, in our view, can still be averaged out by those who have not noticed it if the sample size is great enough. On the other hand, the elusive confounds introduced by entirely different speakers can never be contained or averaged out.

5.2. Video and sound guises

Of the 19 attributes tested, six rendered significant results – three for HRT and three for beauty. These trend in the direction that one would expect for beauty and HRT, which we believe further validates our technique.

Educated The LFT guise was assessed as more *educated* than the HRT guise with a mean score of 3.5 versus 3.2, respectively. Note that the topic of makeup is colloquial, so we expected a penalty for all guises. Nonetheless, the application of HRT appeared to impart even less educatedness, which echos earlier findings of HRT as unprofessional [18] and unintelligent [20].

Formal The LFT guise was assessed as more *formal* than the HRT guise with a mean score of 2.2 versus 2.0, respectively. Again, the topic is colloquial, which would penalize formality across the board. Nonetheless, the application of HRT appeared to impart it with even more informality, which resembles earlier findings that HRT is evaluated as unprofessional [18].

Intelligent The LFT guise was assessed as more *intelligent* than the HRT guise with a mean score of 3.4 versus 3.2, respectively. Intelligence is collinear with education, so we view this result as somewhat redundant, but again also in line with earlier findings of HRT as unintelligent [20].

Articulate The beautiful guise was assessed as more *articulate* than the less-beautiful guise with a mean score of 3.9 versus 3.7, respectively. This was in line with our expectation that status-related attributes would be bolstered by the beautiful guise and attenuated by the less-beautiful guise.

Confident The beautiful guise was assessed as more *confident* than the less-beautiful guise with a mean score of 4.3 versus 4.1, respectively. As we indicated for *articulate*, this is in line with our expectation that status-related attributes would be bolstered by the beautiful guise and attenuated by the less-beautiful guise.

Trendy The beautiful guise was assessed to be more *trendy* than the less-beautiful guise with a mean score of 4.2 versus 3.6, respectively. As we indicated for *articulate* and *confident*, this was also in line with our expectation that status-related attributes would be bolstered by the beautiful guise and attenuated by the less-beautiful guise. Important also is the *degree* of difference between the two guises: conventional beauty seems to result in an especially significant bump in trendiness, something that we intuitively would expect.

6. Conclusion

We have offerred a blueprint for successfully harnessing deepfake technology to build fully-controlled experimental guises for the investigation of language attitudes and perceptions. Our experiment contains four guises that – by means of cutting-edge technology in video and audio manipulation – are identical, save for two interacting variables: (a) HRT or the lack thereof and (b) "beautiful" or "less beautiful". Participant evaluations of these four respective guises are in agreement with earlier literature on HRT, and they are in agreement with intuitions concerning conventional beauty, which further validates our approach. Three participants spotted the "less beautiful" guise as fake, which draws attention to the importance of giving participants the option to comment on the guise. This will then allow researchers to (at least partially) identify and contain the confound, which will bolster the reliability of their study.

We conclude that the next generation of linguistic experimentation will benefit immensely from deepfake technology, especially if (a) societal taboos are heeded concerning what social categories can be superimposed, (b) careful trial-and-error is deployed to ensure a good build, and (c) survey questions are devised that try to identify problems surrounding realism.

7. Acknowledgements

We would like to thank Darren for his video-editing smarts and Aidan Coveney, Hannah Hedegard, David Hornsby, Matthew Hunt, and Rodney Jones for their help recruiting participants.

8. References

[1] E. Levon, D. Sharma, D. J. Watt, A. Cardoso, and Y. Ye, "Accent Bias and Perceptions of Professional Competence in England," *Journal of English Linguistics*, vol. 49, no. 4, pp. 355–388, 2021.

[2] W. E. Lambert, R. C. Hodgson, R. C. Gardner, and S. Fillenbaum, "Evaluational reactions to spoken languages," *The Journal of Abnormal and Social Psychology*, vol. 60, no. 1, p. 44, 1960.

[3] K. T. Strongman and J. Woosley, "Stereotyped reactions to regional accents," *British Journal of Social and Clinical Psychology*, vol. 6, no. 3, pp. 164–167, 1967.

[4] J. A. Dixon, B. Mahoney, and R. Cocks, "Accents of guilt? Effects of regional accent, race, and crime type on attributions of guilt," *Journal of Language and Social Psychology*, vol. 21, no. 2, pp. 162–168, 2002.

[5] R. Mai and S. Hoffmann, "Four positive effects of a salesperson's regional dialect in services selling," *Journal of Service Research*, vol. 14, no. 4, pp. 460–474, 2011.

[6] H. Giles, "Evaluative reactions to accents," *Educational review*, vol. 22, no. 3, pp. 211–227, 1970.

[7] E. Pustka, M. A. Pöchtrager, A. N. Lenz, J. Fanta-Jende Kamerhuber, N. Klingler, H. Leykum, and J. Rennison, Eds., *How to measure a pleasant voice*, vol. 22. Institut für Romanistik, Universität Wien, 2020.

[8] R. L. Cooper and J. A. Fishman, "A Study of Language Attitudes," *Bilingual Review / La Revista Bilingüe*, vol. 4, no. 1/2, pp. 7–34, 1977.

[9] M. Babel, "Evidence for phonetic and social selectivity in spontaneous phonetic imitation," *Journal of Phonetics*, vol. 40, no. 1, pp. 177–189, 2012.

[10] N. Niedzielski, "The effect of social information on the perception of sociolinguistic variables," *Journal of language and social psychology*, vol. 18, no. 1, pp. 62–85, 1999.

[11] J. Hay and K. Drager, "Stuffed toys and speech perception," *Linguistics*, vol. 48, pp. 865–892, 2010.

[12] L. Squires, "It don't go both ways: Limited bidirectionality in sociolinguistic perception," *Journal of Sociolinguistics*, vol. 17, no. 2, pp. 200–237, 2013.

[13] C. Koops, E. Gentry, and A. Pantos, "The effect of perceived speaker age on the perception of PiN and PeN vowels in Houston, Texas," *University of Pennsylvania Working Papers in Linguistics*, vol. 14, no. 2, p. 12, 2008.

[14] E. A. Strand, "Uncovering the role of gender stereotypes in speech perception," *Journal of language and social psychology*, vol. 18, no. 1, pp. 86–100, 1999.

[15] L. S. Casasanto, "Does social information influence sentence processing?" in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 30, no. 30, 2008.

[16] A. D'Onofrio, "Persona-based information shapes linguistic perception: Valley Girls and California vowels," *Journal of Sociolinguistics*, vol. 19, no. 2, pp. 241–256, 2015.

[17] D. Britain, "Linguistic change in intonation: The use of high rising terminals in New Zealand English," *Language Variation and Change*, vol. 4, no. 1, pp. 77–104, 1992.

[18] G. R. Guy and J. Vonwiller, "The meaning of an intonation in Australian English," *Australian Journal of Linguistics*, vol. 4, no. 1, pp. 1–17, 1984.

[19] M. K. Ching, "The question intonation in assertions," *American Speech*, pp. 95–107, 1982.

[20] J. C. Tyler, "Expanding and mapping the indexical field: Rising pitch, the uptalk stereotype, and perceptual variation," *Journal of English Linguistics*, vol. 43, no. 4, pp. 284–310, 2015.

[21] B. M. Horvath, *Variation in Australian English: The Sociolects of Sydney*. Cambridge: Cambridge University Press, 1985.

[22] E. Levon, "Gender, interaction and intonational variation: The discourse functions of High Rising Terminals in London," *Journal of Sociolinguistics*, vol. 20, no. 2, pp. 133–163, 2016.

[23] M. Woo, "Is valley girl speak, like, on the rise?" *National Geographic*, vol. 7, 2013.

[24] V. Shokeir, "Evidence for the stable use of uptalk in South Ontario English," *University of Pennsylvania Working Papers in Linguistics*, vol. 14, no. 2, p. 4, 2008.

[25] K. Dion, E. Berscheid, and E. Walster, "What is beautiful is good." *Journal of personality and social psychology*, vol. 24, no. 3, p. 285, 1972.

[26] J. H. Langlois, L. Kalakanis, A. J. Rubenstein, A. Larson, M. Hallam, and M. Smoot, "Maxims or myths of beauty? A meta-analytic and theoretical review," *Psychological bulletin*, vol. 126, no. 3, p. 390, 2000.

[27] S. Kanazawa, S. Hu, and A. Larere, "Why do very unattractive workers earn so much?" *Economics & Human Biology*, vol. 29, pp. 189–197, 2018.

[28] J. H. Langlois, J. M. Ritter, L. A. Roggman, and L. S. Vaughn, "Facial diversity and infant preferences for attractive faces," *Developmental Psychology*, vol. 27, no. 1, p. 79, 1991.

[29] R. Iyengar, "A controversial photo editing app slammed for AI-enabled 'blackface' feature," *CNN Business*, vol. September 24, 2020. [Online]. Available: https://edition.cnn.com/2020/09/23/tech/gradient-app-ai-blackface/index.html

[30] E. Meskys, J. Kalpokiene, P. Jurcys, and A. Liaudanskas, "Regulating deep fakes: legal and ethical considerations," *Journal of Intellectual Property Law & Practice*, vol. 15, no. 1, pp. 24–31, 2020.

[31] P. J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *The Journal of the Acoustical Society of America*, vol. 90, no. 6, pp. 2956–2970, 1991.

[32] H. Sloetjes and P. Wittenburg, "Annotation by category – ELAN and ISO DCR," in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

[33] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proceedings of Interspeech*, 2017, pp. 498–502.

[34] E. Thomas and P. Carter, "Prosodic rhythm and African American English," *English World-Wide*, vol. 27, pp. 331–355, 2006.

[35] A. Arvaniti and M. Atkins, "Uptalk in Southern British English," in *Proceedings of Speech Prosody, 2016*, J. Barnes, A. Brugos, S. Shattuck-Hufnagel, and N. Veilleux, Eds., 2016, pp. 153–157.

[36] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer software], Version 6.0.36," 2017. [Online]. Available: http://www.praat.org/

[37] B. Paris and J. Donovan, "Deepfakes and cheap fakes," *Data & Society's Media Manipulation research initiative*, 2019.

[38] G. M. White, "Conceptual universals in interpersonal language," *American Anthropologist*, vol. 82, no. 4, pp. 759–781, 1980.

[39] K. Campbell-Kibler, "Accent, (ING), and the social logic of listener perceptions," *American Speech*, vol. 82, no. 1, pp. 32–64, 2007.

[40] G. Norman, "Likert scales, levels of measurement and the "laws" of statistics," *Advances in Health Sciences Education*, vol. 15, no. 5, pp. 625–632, 2010.

[41] G. M. Sullivan and A. R. Artino Jr, "Analyzing and interpreting data from Likert-type scales," *Journal of graduate medical education*, vol. 5, no. 4, pp. 541–542, 2013.