Two Pragmatic Functions of Breathy Voice in American English Conversation

Nigel G. Ward[1], *Ambika Kirkland*[2], *Marcin Włodarczak*[3], *Éva Székely*[2]

[1]University of Texas at El Paso, [2]KTH Royal Institute of Technology, [3]Stockholm University

nigelward@acm.org, kirkland@kth.se, wlodarczak@ling.su.se, szekely@kth.se

Abstract

Although the paralinguistic and phonological significance of breathy voice is well known, its pragmatic roles have been little studied. We report a systematic exploration of the pragmatic functions of breathy voice in American English, using a small corpus of casual conversations, using the Cepstral Peak Prominence Smoothed measure as an indicator of breathy voice, and using a common workflow to find prosodic constructions and identify their meanings. We found two prosodic constructions involving breathy voice. The first involves a short region of breathy voice in the midst of a region of low pitch, functioning to mark self-directed speech. The second involves breathy voice over several seconds, combined with a moment of wider pitch range leading to a high pitch over about a second, functioning to mark an attempt to establish common ground. These interpretations were confirmed by a perception experiment.

Index Terms: CPPS, voice quality, self-directed speech, common ground, grounding, explaining, prosodic constructions

1. Introduction

Previous work on breathy voice has focused on cases where breathiness by itself is discriminative or informative, but has rarely considered how it may interact with other prosodic features. Previous work has also mostly considered its roles in paralinguistics and in phonemic contrasts, with comparatively little study of its role in pragmatic functions. This paper is a case study of breathy voice in American English, focusing on pragmatic functions and examining how breathy voice participates in larger prosodic constructions.

2. Related Work

Our focus is on the pragmatic functions of breathy voice, but it is worth first noting some of its paralinguistic roles. Breathy voice is involved in the expression of emotions and related states, such as intimacy and positive emotions [1, 2, 3, 4]. Breathy voice is also involved in marking various aspects of speaker identity and relation to the interlocutor [5, 6, 7]. Concomitantly, breathy voice is often seen in the production of reported speech, in which a speaker acts out some third party's reported utterance. Breathiness is also common in laughter, laughed speech, and sighs [8, 9] and possibly at phrase starts [10].

Turning now to its pragmatic functions, various roles of breathy voice have been identified for various languages. The only systematic bottom-up study appears to be Ishi *et al.*'s study of Japanese [11], which identified 14 pragmatic functions, of which one of the least frequent, "confidential talking, talking-to-oneself, diffidence" is perhaps related to one of the functions for English we report below. Breathiness has been related to aspects of turn taking, backchanneling and filler production in French, Spanish, English, Slovak, and Arabic [12, 13, 14, 15]. In Spanish, Slovak, and English it also helps differentiate among functions of affirmative cue words, such as backchanneling, providing feedback or communicating agreement [15]. Breathy voice also tends to be characteristic of non-prominent syllables in German [16]. Our own previous work on American English identified two roles relevant to the functions reported below: a role for breathy voice in marking "interpolated, meta- and high-priority utterances" [17], and a role in indicating dissatisfaction [18], although follow-up work found that breathy voice over about 3 seconds, with a short half-second excursion to non-breathy voice, was a better indicator of the latter (J. Avila, personal communication). Also relevant to our discussion below is the involvement of breathy voice in various kinds of questions in several languages: confirmation requests in Lachixío [19], polar questions in Ikaan [20], and rhetorical questions in German [21].

3. Analysis Method

While languages may occasionally convey meanings by variation in one prosodic feature, they more commonly use multiple prosodic features, often in specific temporal configurations. Following Ogden [22], we refer to these as prosodic constructions. Previous work on prosodic constructions has generally not considered breathy voice. We therefore ask, how does breathy voice commonly co-occur with other prosodic features? and, What meanings does it bear in such combinations?

Because prosodic constructions in dialog frequently occur superimposed, identifying them with unaided perception is difficult. We accordingly used a previously developed toolset and workflow [23, 17] that uses Principal Component Analysis (PCA) to automatically identify commonly co-occurring configurations of prosodic features. These are then candidates for being prosodic constructions, whose meanings can be sought by listening to examples where the configurations are strongly present.

This investigation followed the analysis recipe of [17], modified to include a feature representing breathy voice. Specifically, we estimated breathiness using smoothed Cepstral Peak Prominence (CPPS) [24]. The measure corresponds to the amplitude of the first rahmonic relative to the regression line over the log power cepstrum of a signal, effectively quantifying the strength of its harmonic structure. CPPS has been used extensively as a measure of dysphonia in pathological voices [25, 26], and to a lesser extent in phonetics [27], but has so far seen little use in corpus-based work, despite its advantages over traditional perturbation measures, such as jitter or shimmer, when calculated on running speech [28]. While CPPS, like the standard noise-to-harmonics ratio and other measures, does not highly correlate with perceptions of breathiness [29], Panfili's work suggests that it is the best simple proxy [30]. For this study we reimplemented CPPS in Matlab, and we have made this code publicly available, in the latest version of the Midlevel Prosodic Features Toolkit [31].

We stress that CPPS is not robust. For example, for our

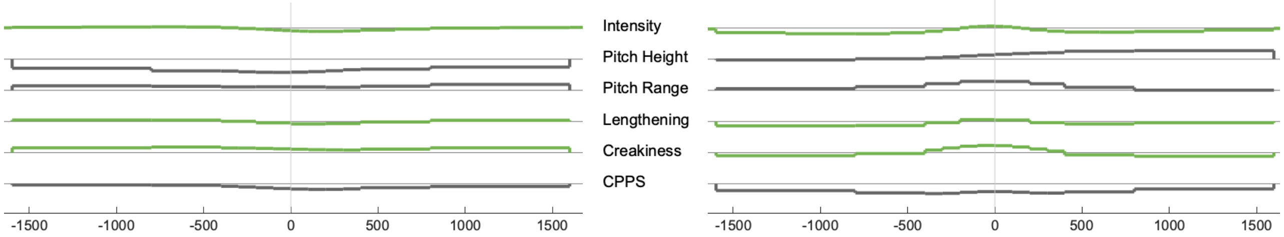

Figure 1: *Weights of the prosodic feature. Left: Dimension 5, Right: Dimension 10. In both diagrams the weights are inverted to show the negative-side pattern. Times are in milliseconds.*

data, the lowest CPPS values were commonly at times of noisy exhalation. While most work using CPPS sidesteps the robustness issue by data curation — using data elicited under controlled conditions or carefully annotated and segmented — here we instead exploit a big data approach and the use of CPPS in concert with other features. Wanting to capture any patterns in the way breathiness varies over time, the CPPS was sampled at {–1200, –600, –350, –250, –150, –50, 50, 150, 250, 350, 600, 1200} milliseconds offset from the point of interest, and these were used as features for the analysis. (Subsequently we repeated the analysis using as features the average CPPS values over 12 windows: $\{-1600 \sim -800, -800 \sim -400, -400 \sim -300, -300 \sim -200, -200 \sim -100, -100 \sim 0\}$, and symmetrically on the right side of the timeline. The results were qualitatively the same.)

We applied the modified recipe (code available at https://www.cs.utep.edu/nigel/breathy/), including computation of CPPS and other prosodic features and then PCA, to 80 minutes of American English conversations between friends and classmates. The result was many dimensions (factors, components) of common variation, of which we examined the 10 that explained most of the variance. Most of these did not significantly involve CPPS: that is, the weights for the CPPS feature indicating nothing more than the expected correlation with voice activity, but for two of the dimensions CPPS was involved, and for these we looked closer.

In particular, we looked for associated meanings. Each dimension involved two poles, when its values were negative and when they were positive, each potentially representing a meaningful construction. Following the method of [17], we accordingly examined a sampling of each, listening to audio at times when the value of a dimension was low and times when it was high, looking for generally present dialog activities or pragmatic functions. For both poles of both dimensions we were eventually able to find a short description that seemed valid for most of the samples examined, as detailed in the next section.

4. The Self-Directed Construction

Dimension 5, as seen in Figure 1 (left), involved breathy voice on the negative side; that is, at times in a dialog when the value of Dimension 5 is low, the value of CPPS tends to be momentarily low and the breathiness high for a second or so. In addition there tends to be about 3 seconds of low pitch and wide pitch range, with a tendency to creaky voice. We refer to this configuration as the dim5neg configuration, as it corresponds to the negative pole of Dimension 5. Examining examples, the functions seemed to involve a related set of stances. These included musings not intended to affect the hearer nor to evoke a reaction. Consider for example (audio available at http:www.cs.utep.edu/nigel/breathy/).

1. need to think about, so I can write it up in the document, to where we're gonna give to the taggers [utep-social00@9:26]

Here a second or so of breathier speech, marked with the dotted underline, occurs in the midst of several seconds of low-pitch and creaky speech. In the dialog, this extract comes after the speaker has segued from responding to a question to reviewing his list of project workitems. Around the word *document* his prosody gives the (weak) impression that he is disengaging from his interlocutor to instead visualize the next task he has to work on.

Other examples included statements based on the speaker's own direct perception, knowledge, or imaginings, as in

2. and like, I could just imagine being part of something like that [utep-social04@5:43]

where the speaker is wrapping up discussion of his dream job. Overall, we call this family of functions "marking self-directed speech" [32], as summarized in Figure 2 (left).

While secondary to our focus here, the form and function of the other pole of this dimension are also worth mentioning. Its weights are of course the exact opposites of dim5neg, and thus this pattern involves a second or so of non-breathy voice, typically modal or even harmonic, in the midst of some 3 seconds of high pitch and narrow pitch range. Across many examples, this pattern, for short "dim5pos," occurred with a family of related pragmatic functions, notably inviting the interlocutor to see the point in an argument or the point of a joke, so we summarize this in general as marking "other-directed" speech.

5. The Seeking Common Ground Construction

As seen in Figure 1 (right), there is another pattern involving breathy voice. In this, crucially, the breathiness is much longer, lasting typically around at least 3 seconds (the limit of the span of analysis used here). This pattern also involves a moment of wide pitch and creaky voice, followed by a rise to a region of high pitch of about a second.

This pattern commonly occurs with shared laughter. It also often occurs when the speakers are trying to establish the context as a preliminary to a deeper discussion, as in

3. okay, so it's been a little while since you've been programming⇑, or // no↑, it, um, it's actually been really quick; I've only had a summer off. [utep-social04@1:00]

The Self-Directed Construction (dim5neg)		
form:	3 sec region with:	low pitch slightly wide pitch range creaky voice
	breathiness over about 1 sec within that region	
function: mark the utterance as self-directed		

The Grounding Construction (dim10neg)	
form:	breathy voice over about 3 sec with a moment of creaky voice and wide pitch rising to high pitch
function: indicate an attempt at grounding	

Figure 2: *Summaries of the Identified Constructions*

where // marks a speaker change, ⇑ the pitch rise central to the dialog segment best matching this pattern, and ↑ other pitch rises. Occurrences of this pattern often might be characterized as uptalk. It often includes contributions by both speakers; thus it can operate as a joint construction. This pattern also supports the functions of trying to reach agreement on the nature of something or what to call it, as in

4. They're searching for content? they're searching for↑ ...// yeah⇑, content, um↑ // a feeling↑ [utep-social00@9:12]

Overall, we called this family of functions "working towards common ground," as summarized in Figure 2 (right).

Although again secondary, the opposite pattern is also interesting. In form it involves high CPPS, thus modal or harmonic voice, over a few seconds, ending lower in pitch. In terms of function, this seems to be "speaking to inform," as when the speaker was explaining something from a stance of expert knowledge. In direct contrast to examples from the negative pole, the focus is on providing true information, regardless of whether this is understandable by or engaging to the listener. We labeled this "explaining."

6. Experiment

While the meanings of these constructions were generally evident to us, we wanted to see whether others would also perceive them. We hypothesized that native speakers would frequently judge stimuli that exemplified these constructions as having the claimed meanings.

6.1. Stimuli and Statistical Test

As stimuli, for each construction we used timepoints identified by the model as among the lowest (respectively, highest) on the dimension of interest. To provide subjects with context, we selected 10-second clips such that the minimal (respectively, maximal) timepoint occurred 3 seconds before the end. Clips were in stereo, including audio from both speakers. We chose clips from 6 different conversations, using times with the lowest (respectively, highest) values on the dimension of interest, after excluding candidate clips that began at the dialog onset or contained no speech, only laughter or noises. We selected 2 such clips each for both the negative and positive poles for both dimensions of interest for each of the 6 conversations, for a total of 47 stimuli (not 48 because one clip was accidentally duplicated).

As controls, we used the clips selected to exemplify the other constructions. Thus, for example, the connection between dim5neg and self-directed speech was assessed by the frequency by which subjects chose the "self-directed" descriptor more frequently for the 12 clips exemplifying that pattern than for the 35 clips selected to exemplify dim5pos, dim10neg, and dim10pos. The controls were used in two ways: lumped and split into the three sets. In both cases we judged significance using chi-square tests with $p < 0.05$ as the threshold. When testing against the split controls, we used a Bonferroni correction factor of 3.

6.2. Procedure and Subjects

Subjects were presented the audio clips in randomized order and asked to "Listen to the audio, and focusing on the last few seconds, select one or more descriptors that best describe what is happening in the conversation." 13 descriptors were provided, with checkboxes. These included one each for the hypotheses (self-directed, other-directed, working toward common ground, and explanation), four for other attributes commonly observed with the four patterns (low involvement, high involvement, agreement, and elaboration), and five pure distractors (disagreement, confusion, anger, amusement, and hesitation). Participants performed the experiment online on their own computers, using the Cognition.run platform to listen to each stimulus and then record their judgments.

One hundred native speakers of English were recruited using Prolific.com, but data for one subject was lost and so N=99. The reported dialects were British (82%), Irish (7%) Canadian (7%), and American (4%). 64% identified as female. The median age was 34.

6.3. Results

We obtained statistical significance for all hypotheses: subjects judged the stimuli exemplifying these patterns to have the hypothesized meanings more often than the lumped controls. Indeed, the support was strong for each of the four (p < .002, χ^2).

Analysis using the split controls however revealed some nuances, as seen in Table 1. "Self-directed" was chosen significantly more often for Dim5neg stimuli than for Dim5pos, χ^2 = 38.04, Dim10neg, χ^2 = 56.24, and Dim10pos χ^2 = 6.38. "Other-directed" was chosen significantly more often for Dim5pos than for Dim5neg, χ^2 = 37.98, or Dim10pos, χ^2 = 17.30, but not for Dim10neg, χ^2 = 4.14. "Working towards common ground" was chosen significantly more often for Dim10neg than for Dim10pos, χ^2 = 21.05, or Dim5neg, χ^2 = 23.19, but not for Dim5pos, χ^2 = 2.18. "Explanation" was chosen significantly more often for Dim10pos than for Dim10neg, χ^2 = 48.24, or Dim5pos, χ^2 = 48.11, but not more than Dim5neg, χ^2 = 0.002.

6.4. Discussion

The effect sizes are modest, but these likely understate the true strength of connection between these prosodic forms and the meanings, for several reasons. First, our descriptors were short phrases and single words, some of whose intended meanings were likely not familiar to many of the crowdworkers. We noted, for example, that 28 of them never chose the "other-

	percent of stimuli rated as			
	self-directed	other-directed	grounding	explanation
Dimension 5 neg	**16.9%**	3.8%**	7.1%**	55.8%
Dimension 5 pos	8.4%**	**10.4%**	15.5%	41.7%**
Dimension 10 neg	6.6%**	7.8%	**13.2%**	41.3%**
Dimension 10 pos	13.1%*	5.6%**	7.3%**	**56.0%**

Table 1: *Frequency of descriptors chosen for each pole of each dimension. Percentages in bold are for the functions that were hypothesized to be associated with each prosodic pattern. * indicates that the difference in response frequency to the hypothesized function in the column was statistically significant by a chi-square test, $p < .05.$, and ** $p < .001$. All significant differences were in the predicted directions.*

directed" descriptor. Second, the stimuli were not vetted. Although CPPS is informative on average, as when used to detect patterns across an entire corpus, it is susceptible to local phonetic content [33], compounded here by our use of point samples rather than smoothed values, and thus some of the stimuli may not actually have been perceptually breathy or non-breathy. Also, our controls were not well chosen, not being representative of any typical or neutral prosody. Third, some of the crowdworker judgements may not have been meaningful. Typically some fraction of workers rush through the task without paying much attention, introducing random noise, but our procedure lacked attention checks and quality checks to detect this. Fourth, although all stimuli exemplified the prosodic construction of interest, they also included lead-in content, other superimposed prosodic constructions, and specific lexical content, any of whose meanings may have been more salient to our crowdworkers than the meaning element of interest.

We also note that the mappings from prosody to meaning are here, as always, not one-to-one. For example, subjects often chose the descriptor "amusement" for the examples of dim10neg. This likely reflects the fact that laughter, which is often breathy, often involves amusement, but also suggests that dim10neg can bear more than one meaning. Subjects also commonly chose "explanation" for the examples of dim5neg. This likely reflects that fact that several speakers in the corpus talked at length about how they accomplished recent programming assignments or what they needed to work on next, and the fact of being an explanation was likely more salient to our judges than the fact that these were mostly self-directed disquisitions, and also suggests that the activity of explaining can be prosodically marked in more than one way. Clearly the mappings between prosodic patterns and pragmatic functions are complex.

7. Likely Practical Significance

Systems that aim to recognize aspects of speakers' state or intent from their speech might benefit by including features such as CPPS to represent breathiness. In particular, the existence of the "self-directed" construction suggests that breathiness is relevant to the commercially important problem of detecting self-directed speech versus device-directed speech [34].

Our findings also indicate the need for speech synthesizers to be able to produce utterances with varying degrees of breathiness in varying positions, an ability which exists for HMM-based TTS [35] but it is yet to be added to state-of-the-art neural synthesizers. For example, effective spoken-language interaction with robots [36] will likely need this: to be able to distinctively produce both self talk [37], intended to unobtrusively give status updates to the user, and user-directed talk, needing immediate attention; and also able to clearly distinguish between the activity of establishing common ground and the activities of drawing conclusions and reaching decisions, for the sake of effectively coordinating joint action [38].

8. Open Questions and Conjectures

While creaky voice and breathy voice are sometimes thought of as distinct phenomena, in these constructions they occur together. We know that CPPS also correlates with creaky voice [15], and that phonation types are not discrete categories, but the exact relationship between creaky voice and breathy voice in such contexts calls out for further investigation.

We have provided only circumstantial evidence that breathy voice causally contributes to these perceived meanings. In future work we plan to build a high quality speech synthesizer where the degree of breathiness can be directly manipulated, and use it to create stimuli for controlled experiments.

Many of the questions in these dialogs involved the dim10neg pattern, often with very salient breathy voice. While studies of the prosody of English questions and related phenomena have mostly focused on pitch contours, slope and peak shapes [39, 40], we conjecture that, at least in dialog, breathy voice is an important cue to questionhood and question type.

Pervasive breathy voice is considered to be not a benign individual characteristic, but something requiring medical consultation and possible treatment. Our findings suggest an explanation beyond intelligibility and aesthetics: we conjecture that proper control of breathy voice is important for communicative effectiveness.

Our earlier attempt at a comprehensive description of the pragmatic functions of prosody in American English [17], was marred by the lack of a feature for breathy voice. Adding CPPS to the feature inventory enabled us to discover more: how prosody serves two very important dialog functions: marking self-directed speech and marking grounding attempts. We conjecture that consideration of breathy voice will be important for a full understanding of the prosody of other languages also.

Acknowledgments This research was supported by the Swedish Research Council projects *Perception of speaker stance – using spontaneous speech synthesis to explore the contribution of prosody, context and speaker* (VR-2020-02396) and *Prosodic functions of voice quality dynamics* (VR-2019-02932), and by the Riksbankens Jubileumsfond project *CAPTivating – Comparative Analysis of Public speaking with Text-to-speech* (P20-0298).

9. References

[1] K. Hammerschmidt and U. Jürgens, "Acoustical correlates of affective prosody," *Journal of voice*, vol. 21, no. 5, pp. 531–540, 2007.

[2] C. Gussenhoven, "Foundations of intonational meaning: Anatomical and physiological factors," *Topics in Cognitive Science*, vol. 8, pp. 425–434, 2016.

[3] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[4] A. Anikin, "A moan of pleasure should be breathy: the effect of voice quality on the meaning of human nonverbal vocalizations," *Phonetica*, vol. 77, no. 5, pp. 327–349, 2020.

[5] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *the Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[6] R. J. Podesva and P. Callier, "Voice quality and identity," in *Annual Review of Applied Linguistics, 35*. Cambridge University Press, 2015, pp. 173–194.

[7] Y. Li, N. Campbell, and J. Tao, "Voice quality: Not only about "you" but also about "your interlocutor"," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4739–4743.

[8] J. Trouvain, "Phonetic aspects of "speech-laughs"," in *Oralité et Gestualité: Actes du colloque ORAGE, Aix-en-Provence. Paris: L'Harmattan*, 2001, pp. 634–639.

[9] K. P. Truong, G. J. Westerhof, F. de Jong, and D. Heylen, "An annotation scheme for sighs in spontaneous dialogue," in *Interspeech*, 2014, pp. 228–232.

[10] E. Bird and M. Garellek, "Dynamics of voice quality over the course of the english utterance," in *Proceedings of the 19th International Congress of Phonetic Sciences*, 2019, pp. 2406–2410.

[11] C. T. Ishi, H. Ishiguro, and N. Hagita, "Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–12, 2010.

[12] C. Smith, "Marking the boundary: utterance-final prosody in french questions and statements," in *International Congress of Phonetic Sciences*, 1999.

[13] N. Ward and Y. Al Bayyari, "A prosodic feature that invites backchannels in Egyptian Arabic," in *Perspectives on Arabic Linguistics XX*, M. Mughazy, Ed. John Benjamins, 2007, pp. 186–206.

[14] E. Székely, J. Mendelson, and J. Gustafson, "Synthesising uncertainty: The interplay of vocal effort and hesitation disfluencies." in *Interspeech*, 2017, pp. 804–808.

[15] M. Heldner, M. Wlodarczak, Š. Beňuš, and A. Gravano, "Voice quality as a turn-taking cue," in *Interspeech*, 2019, pp. 4165–4169.

[16] B. Ludusan, P. Wagner, and M. Włodarczak, "Cue interaction in the perception of prosodic prominence: the role of voice quality," in *Interspeech*, 2021, pp. 1006 – 1010.

[17] N. G. Ward, *Prosodic Pattterns in English Conversation*. Cambridge University Press, 2019.

[18] N. G. Ward, J. E. Avila, and A. M. Alarcon, "Towards continuous estimation of dissatisfaction in spoken dialog," in *SigDial*, 2021.

[19] M. A. Sicoli, "Voice registers," in *The handbook of discourse analysis*, D. Tannen, H. E. Hamilton, and D. Schiffrin, Eds. Wiley, 2015, pp. 105–126.

[20] S. Salffner, "West African languages enrich the frequency code: Multi-functional pitch and multi-dimensional prosody in Ikaan polar questions," *Laboratory Phonology*, vol. 8, no. 1, pp. 1–44, 2017.

[21] B. Braun, N. Dehé, J. Neitsch, D. Wochner, and K. Zahner, "The prosody of rhetorical and information-seeking questions in German," *Language and Speech*, vol. 62, no. 4, pp. 779–807, 2019.

[22] R. Ogden, "Prosodic constructions in making complaints," in *Prosody in Interaction*, D. Barth-Weingarten, E. Reber, and M. Selting, Eds. Benjamins, 2010, pp. 81–103.

[23] N. G. Ward, "Automatic discovery of simply-composable prosodic elements," in *Speech Prosody*, 2014, pp. 915–919.

[24] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *Journal of Speech Language and Hearing Research*, vol. 39, no. 2, 1996.

[25] Y. D. Heman-Ackah, D. D. Michael, and G. S. Goding, "The relationship between cepstral peak prominence and selected parameters of dysphonia," *Journal of Voice*, vol. 16, no. 1, pp. 20–27, 2002.

[26] Y. D. Heman-Ackah, R. J. Heuer, D. D. Michael, R. Ostrowski, M. Horman, M. M. Baroody, J. Hillenbrand, and R. T. Sataloff, "Cepstral peak prominence: a more reliable measure of dysphonia," *Annals of Otology, Rhinology & Laryngology*, vol. 112, no. 4, pp. 324–33, 2003.

[27] Y. Maryn and D. Weenink, "Objective dysphonia measures in the program Praat: smoothed cepstral peak prominence and acoustic voice quality index," *Journal of Voice*, vol. 29, no. 1, pp. 35–43, 2015.

[28] C. Moers, B. Möbius, F. Rosanowski, E. Nöth, U. Eysholdt, and T. Haderlein, "Vowel- and text-based cepstral analysis of chronic hoarseness," *Journal of Voice*, vol. 26, no. 4, pp. 416–424, 2012.

[29] A. Chanclu, I. B. Amor, C. Gendrot, E. Ferragne, and J.-F. Bonastre, "Automatic classification of phonation types in spontaneous speech: towards a new workflow for the characterization of speakers' voice quality," in *Interspeech*, 2021, pp. 1015–1018.

[30] L. Panfili, "Cross-linguistic acoustic characteristics of phonation: A machine learning approach," Ph.D. dissertation, University of Washington, 2018.

[31] N. G. Ward, "Midlevel prosodic features toolkit (2016-2021)," 2021, https://github.com/nigelg ward/midlevel.

[32] K. Mády and U. D. Reichel, "How to distinguish between self-and other-directed wh-questions?" in *Proc. Phonetik und Phonologie im deutschsprachigen Raum*. Munich University, 2016.

[33] M. Sampaio, M. L. Vaz Masson, M. F. de Paula Soares, J. E. Bohlender, and M. Brockmann-Bauser, "Effects of fundamental frequency, vocal intensity, sample duration, and vowel context in cepstral and spectral measures of dysphonic voices," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 5, pp. 1326–1339, 2020.

[34] C.-W. Huang, R. Maas, S. H. Mallidi, and B. Hoffmeister, "A study for improving device-directed speech detection toward frictionless human-machine interaction," in *Interspeech*, 2019, pp. 3342–3346.

[35] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise," *Computer Speech & Language*, vol. 28, no. 2, pp. 648–664, 2014.

[36] M. Marge, C. Espy-Wilson, N. G. Ward *et al.*, "Spoken language interaction with robots: Research issues and recommendations," *Computer Speech and Language*, in press, 2021.

[37] A. Geraci, A. D'Amico, A. Pipitone, V. Seidita, and A. Chella, "Automation inner speech as an anthropomorphic feature affecting human trust: Current issues and future directions," *Frontiers in Robotics and AI*, vol. 8, p. 66, 2021.

[38] H. H. Clark, *Using Language*. Cambridge University Press, 1996.

[39] N. Hedberg and J. M. Sosa, "A unified account of the meaning of English questions with non-canonical intonation," in *International Seminar on Prosodic Interfaces, Jawaharlal Nehru University, November*, 2011, pp. 25–27.

[40] A. Ritchart and A. Arvaniti, "The form and use of uptalk in Southern Californian English," in *Speech Prosody*, 2014, pp. 20–23.