Two Pragmatic Functions of Breathy Voice in American English Conversation

Nigel G. Ward¹, Ambika Kirkland², Marcin Włodarczak³, Éva Székely²

¹University of Texas at El Paso, ²KTH Royal Institute of Technology, ³Stockholm 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 the role of breathy voice 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

While 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 [5, 6]. 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 [7, 8].

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 [9], 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 [10, 11, 12, 13]. In Spanish, Slovak, and English it also helps differentiate among functions of afﬁrmative cue words, such as backchanneling, providing feedback or communicating agreement [13]. Our own previous work on American English [14] noted a role for breathy voice in marking “interpolated, meta- and high-priority utterances,” and in indicating dissatisfaction [15], 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, (J. Avila, personal communication). Breathy voice also tends to be characteristic of non-prominent syllables in German [16]. It also is involved in various kinds of questions in several languages: conﬁrmation requests in Lachixío [17], polar questions in Ikaan [18], and rhetorical questions in German [19].

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 [20], 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 [21, 14] that uses Principal Component Analysis (PCA) to automatically separate out 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 [14], but modiﬁed to include a feature representing breathy voice, sampled at {-1200, -700, -350, -250, -150, -50, 50, 150, 250, 350, 700, 1200} milliseconds offset from the point of interest. Speciﬁcally, we estimated breathiness using smoothed Cepstral Peak Prominence (CPPS) [22]. 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 [23, 24], and to a lesser extent in phonetics [25], 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 [26]. While CPPS, like NHR and other measures, does not highly correlate with perceptions of breathiness [27], it seems to be the best simple measure of dysphonia in pathological voices [23, 24], and to a lesser extent in phonetics [25], 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 [26]. While CPPS, like NHR and other measures, does not highly correlate with perceptions of breathiness [27], it seems to be the best simple measure [28]. 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 [29]. We stress that CPPS is not robust. For example, for our data, the lowest CPPS values were commonly at times of noisy exhalation. While most work using CPPS sidesteps the robust-
ness issue by data curation — using data elicited under controlled conditions or carefully annotated and segmented — here we instead rely on a big data approach and on the use of CPPS in concert with other features.

We applied the modified recipe (code available at http://www.cs.utep.edu/nigel/breathy/), including computation of CPPS and the usual 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 indicated 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 meanings. Each dimension involved two poles, when its values were negative and when they were positive, each potentially representing a meaningful construction. 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 dimensions and both poles of each we were eventually able to find a short description that seemed to validly describe most of the samples examined.

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 pitched and creaky speech. In the dialog, this extract comes after the speaker has segued from a response to a question into a dissertation on 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 as marking “self-directed speech” [30], 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]

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, in a sense, 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 these meanings. 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 different 6 conversations, using times with the lowest (respectively, highest) scores on the dimension of interest, after excluding candidate clips that began at the dialog onset or contained no speech, only laughter or noises. We thus 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 with 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, as judged by chi-square tests. We chose 0.05 as the significance threshold, and, having three sets of controls in each case, 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 hypothesized (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. 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

As seen in Table 1, all hypotheses were supported. All results are \( \chi^2 \) (3, N=99), as data for one subject was lost. “Self-directed” was chosen more often for Dim5neg stimuli than for Dim5pos, \( \chi^2 = 38.04, p < .001 \), Dim10neg, \( \chi^2 = 56.24, p < .001 \), and Dim10pos \( \chi^2 = 6.38, p < .05 \). “Other-directed” was chosen more often for Dim5pos than for Dim5neg, \( \chi^2 = 37.98, p < .001 \), or Dim10pos, \( \chi^2 = 17.30, p < .001 \), but not for Dim10neg, \( \chi^2 = 4.14 \). “Working towards common ground” was chosen more often for Dim10neg than for Dim10pos, \( \chi^2 = 21.05, p < .001 \), or Dim5neg, \( \chi^2 = 23.19, p < .001 \), but not significantly more than for Dim5pos, \( \chi^2 = 2.18 \). “Explanation” was chosen more often for Dim10pos than for Dim10neg, \( \chi^2 = 48.24, p < .001 \), or Dim5pos, \( \chi^2 = 48.11, p < .001 \), but not significantly more than Dim5neg, \( \chi^2 = 0.02 \).

#### 6.4. Discussion

We obtained statistical significance for all hypotheses: subjects judged the stimuli exemplifying these patterns to have the hypothesized meanings more often than the controls. While the detailed statistics reveal some nuances, overall, lumping together all controls for each case, each of the four hypothesis was clearly supported (p < .002, \( \chi^2 \)).

The effect sizes are modes, but these likely understate the true strength of connection between these two 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 the subjects never chose the “other-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 [31], compounded 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 or 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 con-

---

**Figure 2: Summaries of the Identified Constructions**

<table>
<thead>
<tr>
<th>The Self-Directed Construction (dim5neg)</th>
<th>The Grounding Construction (dim10neg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>form: 3 sec region with: low pitch</td>
<td>form: breathy voice over about 3 sec</td>
</tr>
<tr>
<td>slightly wide pitch range</td>
<td>with a moment of creaky voice and wide pitch</td>
</tr>
<tr>
<td>creaky voice</td>
<td>rising to high pitch</td>
</tr>
<tr>
<td>breathiness over about 1 sec within that region</td>
<td>function: mark the utterance as self-directed</td>
</tr>
<tr>
<td>function: indicate an attempt at grounding</td>
<td>function: indicate an attempt at grounding</td>
</tr>
</tbody>
</table>
such contexts calls out for further investigation. The exact relationship between creaky voice and breathy voice in that phonation types are not really discrete categories, but the We know that CPPS and creak measures do correlate [13], and distinct phenomena, in these constructions they occur together.

While creaky voice and breathy voice are often thought of as coordinating joint action [35].

conclusions and reaching decisions, for the sake of effectively communicating will need this, to be able to distinctively produce both self talk, intended to unobtrusively give status up-

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 [14], 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 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 is 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


<table>
<thead>
<tr>
<th>Dimension</th>
<th>self-directed</th>
<th>other-directed</th>
<th>grounding</th>
<th>explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 neg</td>
<td>16.9%**</td>
<td>3.8%**</td>
<td>7.1%**</td>
<td>55.8%</td>
</tr>
<tr>
<td>5 pos</td>
<td>8.4%**</td>
<td>10.4%**</td>
<td>15.5%</td>
<td>41.7%**</td>
</tr>
<tr>
<td>10 neg</td>
<td>6.6%**</td>
<td>7.8%*</td>
<td>13.2%</td>
<td>41.3%**</td>
</tr>
<tr>
<td>10 pos</td>
<td>13.1%*</td>
<td>5.6%**</td>
<td>7.3%**</td>
<td>56.0%</td>
</tr>
</tbody>
</table>

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.

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. Based on the existence of the “self-directed” construction, we conjecture that breathiness is relevant to the commercially important problem of detecting self-directed speech versus device-directed speech [32].

Our findings also suggest a need for speech synthesizers be able to produce utterances with varying degrees of breathiness in varying positions, an ability which exists for HMM-based TTS [33] but it is yet to be added to state-of-the-art neural synthesizers. We conjecture that robots, for example [34], to effectively communicate will need this, to be able to distinctively produce both self talk, 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 [35].

8. Open Questions and Conjectures

While creaky voice and breathy voice are often thought of as distinct phenomena, in these constructions they occur together. We know that CPPS and creak measures do correlate [13], and that phonation types are not really 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 [36, 37], 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 [14], 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 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 is 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
