Perception of Breathy Phonation in Gujarati

MAX NELSON*, KELLY BERKSON*, SAMEER UD DOWLA KHAN**, CHRISTINA ESPOSITO***
**Indiana University, Bloomington; **Reed College; ***Macalester College*

ABSTRACT

The correlates of breathiness are similar across consonants and vowels, raising a question about whether breathy consonant/breathy vowel contrasts are confusable in languages with both, e.g. Gujarati. We investigate the perception of phonemically breathy Cs and Vs in Gujarati via three tasks: free-sort, AX discrimination, and picture-matching identification. Results from six native listeners indicate that breathiness is indeed confusable: participants reliably identify the presence of breathiness if the acoustic correlates thereof are strong enough, but cannot reliably assign it to the appropriate segment (consonant or vowel), rendering it difficult to distinguish CʱV from CV̤.

1. Introduction

Phonation refers to production of sound via vibration of the vocal folds. Different types of phonation are achieved by adjusting the manner of vocal fold vibration; breathy voice, for example, is produced with increased airflow as compared with modal/plain voicing, resulting in increased turbulence/noise in the signal (Bhaskararao & Vuppala 2014, Gordon & Ladefoged 2001, Ladefoged & Maddieson 1996).

In both consonants and vowels, breathy voice is associated with increases in spectral balance and spectral slope, as well as increases in measures of noise (Berkson 2012, Dutta 2007, Esposito 2006, Huffman 1987, Khan 2012, among others). In terms of localization, the acoustic correlates of breathy voiced consonants are housed primarily in the the following vowel (Berkson 2012, Esposito & Khan 2012). A question is therefore raised as to how CʱV and CV̤ sequences differ from one another. Esposito and Khan (2012) investigated this via acoustic analyses of White Hmong and Gujarati, two languages that contrast breathy consonants and breathy vowels. In both languages, the timing and degree of acoustic difference were found to pattern differently in consonants and vowels: breathy consonants are characterized by a short period of intense breathiness at the onset of the vowel followed by decreasing breathiness, while breathy vowels showed stable (Gujarati) or increasing (White Hmong) breathiness throughout the vowel.

Perceptually, we know that Gujarati speakers can reliably distinguish breathy from modal vowels in Gujarati[1] stimuli (Bickley 1982, Fischer-Jørgensen 1967). But can they leverage the differences in timing and degree of breathiness in CʱV vs. CV̤ sequences in order to reliably distinguish the two? This is the question addressed herein.

2. Methods

This study includes three tasks (free sort, AX discrimination, and picture-matching identification) to investigate the perception of CV, CʱV, and CV̤ sequences by native Gujarati listeners. Tasks

[1] Breathy vowels in Gujarati vary across dialect and register. They may be produced as a disyllabic [əhV] sequence in careful speech (Cardona & Suthar 2003, Khan 2012). Also, some dialects may not have breathy vowels (p.c., Gujarati informants).

were ordered, rather than randomized across participants, because the design of the ID task imposed three experimenter-defined categories on participants. To minimize potential vowel-context or gender effects, stimuli consisted of a well-known minimal triplet (breathy vowel: ba̤ɾ ‘outside’, breathy consonant: bʱaɾ ‘burden’, modal: baɾ ‘twelve’) produced by four female native speakers in their 20s.

	Gujarati	IPA	Gloss
Breathy Vowel	બહાર	ba̤ɾ	‘outside’
Breathy Consonant	ભાર	bʱaɾ	‘burden’
Modal	બાર	baɾ	‘twelve’

Table (1) Stimuli List

Stimuli were extracted from running speech recorded in Khan (2012) and zero-crossed to maximize naturalness. Two repetitions of each member of the triplet was used, for a total of 24 tokens (3 items X 2 repetitions X 4 speakers). Participants included six native Gujarati listeners, five males in their mid-20s and one 52 yr old female.

3. Free Sort Task and Results

The free sort task, which followed the auditory free classification methodology of Clopper (2008), investigated whether listeners independently proposed three target categories ([baɾ], [bʱaɾ], and [ba̤ɾ]) when presented with a screen containing 24 numbered icons arranged in columns (Fig. 1a) and asked to categorize them by dragging them to the right and placing them in groups (see sample outcome in Fig. 1b). Icons corresponded randomly to one of the 24 audio stimuli, and played when clicked.

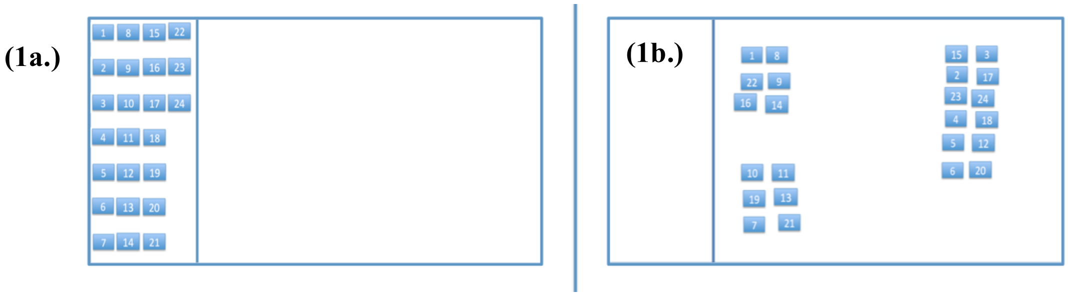

Figure (1) Free sort task set-up (a) and sample outcome (b)

To avoid any experimenter-imposed biases, participants had absolute freedom over how to categorize the items and how many categories to propose.

Different approaches to categorization were utilized, and so a purely descriptive report of the outcomes is most informative here. Participants (a) and (b) attempted to pair all stimuli by both token and speaker, resulting in 12 groups of two (see Figure 2a).[2] Participant (a) was highly accurate in grouping stimuli this way, while Participant (b) was less so. Participants (c) and (d) formed two unique groups, a response pattern illustrated in Fig. (2b). For both, one of the groups represents a well-defined [bʱaɾ] category while the other combined [baɾ] and [ba̤ɾ]. This is of particular interest in light of a note in Fischer-Jørgensen (1967) mentioning that a modal vowel can serve as an acceptable realization of a breathy vowel but the reverse is not true. Participant (e) created three groups which may have been intended to represent the three categories of stimuli: each consisted of a majority of one type of stimuli, but all groups were mixed and contained at least one member of each of the three stimuli types. Even here, however, the most consistently grouped stimuli were breathy consonants—perhaps indicating that these are the least confusable type of stimuli. The responses of participant (f) appeared random, highlighting the problems that can arise in a task with so few guidelines.

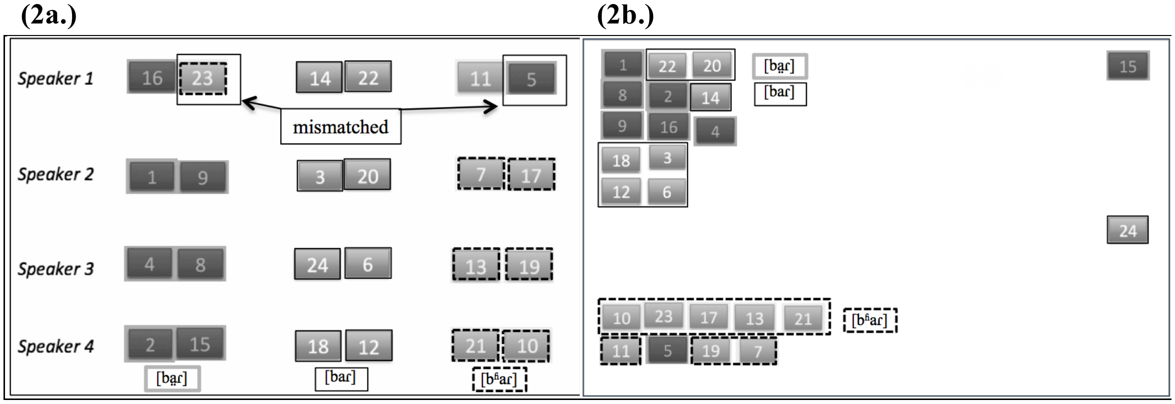

Figure (2) Recoded outcomes illustrating two response patterns. Two participants grouped by speaker and token-type with relative accuracy (2a), and two created two large groups (2b)—one consisting of breathy consonant items, and one grouping plain tokens and breathy vowel tokens together.

Overall, responses to the free sort task can be divided in three categories: those pairing by token and speaker (a and b), those separating breathy consonants from all other tokens (c and d), and those following less interpretable orders (e and f). Participants (a) and (b) matched tokens from the same speaker with great accuracy, suggesting that they can leverage speaker-specific acoustic information, while the results from (c) and (d) suggest overlap in the modal and breathy vowel categories.

4. Discrimination Task and Results

The discrimination task aimed to determine the accuracy with which participants can distinguish pairs of target words. In one sense, it is the task that most directly addresses the issue of perceiving the difference between CV, CʱV, and CV̤ sequences, because while in other perceptual tasks the participant may categorize stimuli in some way and then compare categories, a

[2] All icons were identical when the participant arranged them, and pairs of icons were distributed randomly throughout the screen, but in Figure (2) icons have been rearranged and re-colored for clarity.

discrimination task encourages participants to compare the stimuli directly (Key 2012). Items were presented in a classic AX task. In the trials, participants heard two of the 24 stimuli in succession and indicated whether the two were 'same' or 'different'. No trial included two words from the same speaker, so there were 54 unique AX pairings. All pairings were played in both orders, for a total of 108 randomly ordered trials. The three categories of stimuli included modal [baɾ] (henceforth referred to as M), breathy consonant [bʱaɾ] (henceforth C), and breathy vowel [ba̤ɾ] (henceforth V). In some trials the two items were the same, and in some they were different. "Same" trials were of three types (MM, CC, and VV), as were "different" trials (MC, MV, and VC).

Given our main question, the crucial trials are CV (contrasting breathy C [bʱaɾ] and breathy V [ba̤ɾ]), where participant responses can reveal whether the two types of stimuli are reliably distinguished. Overall accuracy by trial-type is presented in Fig. (3).

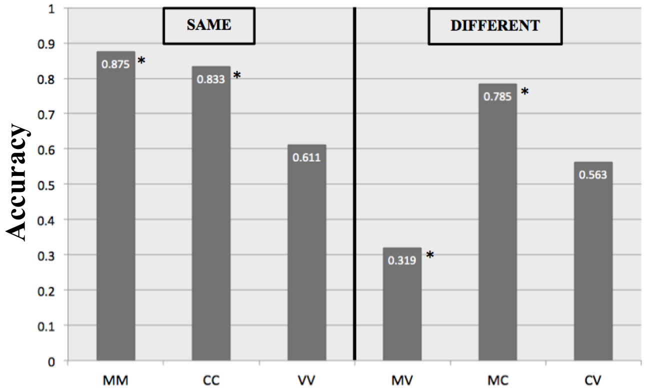

Figure (3) Mean accuracy in AX discrimination. * = responses significantly different from chance.

To confirm that a contrast is perceptually salient, participants must discriminate stimuli at a rate significantly above chance (in a task with two possible answers, this is 0.5). Chi-squared tests compared the accuracy of each trial type to chance. Participants performed significantly above chance in MM and CC trials (p < .0001), and reliably differentiated these two types of tokens as evidenced by their above-chance performance in MC trials (p < .0001). However, they were not above chance in the target CV trials, those differentiating [bʱaɾ] and [ba̤ɾ] (p = .1136). Breathy V [ba̤ɾ] stimuli were problematic for listeners in general. In VV trials, participants correctly identified two breathy vowel stimuli as being "the same" with just 61.1% accuracy. This is not above chance (p = .0593). In MV trials ([baɾ] vs. [ba̤ɾ]), their average accuracy of 31.9% was significantly below chance (p < .0001). Rather than correctly distinguishing [baɾ] and [ba̤ɾ] as different, participants actively considered them to be the same. Plainly stated, listeners did not reliably consider two breathy vowel stimuli to be the same, and yet actively considered a breathy vowel and a fully modal stimulus to be the same. Recall Fischer-Jørgensen 1967's comment that a modal vowel can "pass" as a breathy vowel, but not vice versa. The results partially support this, as a modal vowel can pass as a breathy vowel to the extent that a "same" response was preferred in MV trials. What remains confusing, and begs further investigation, is the finding that breathy vowels themselves are not sufficiently alike so as to trigger an above-chance "same" response.

5. Identification Task and Results

The identification (ID) task sought to determine overlap between categorization of the target words. Unlike the previous two tasks, the ID task allowed participants to determine whether a stimulus was an acceptable member of an experimenter-defined category. Consider: [baɾ] might be an acceptable realization of /ba̤ɾ/ 'outside' to a participant at least sometimes, but they may still fail to group them together in a free sort task where they can play the tokens repeatedly and deliberate about how to group them, and they may recognize auditory differences between [baɾ] and [ba̤ɾ] in a discrimination task. In an ID task, however, they may indicate that [baɾ] can correspond to the meaning 'outside'.

With the help of a native speaker, pictures representing the three target words were selected. Participants heard an audio stimulus, saw an image representing one of the target words, and indicated whether the two matched. Like the discrimination task, there were "same" and "different" trials: three types of "same" trials, wherein the presented audio and image match, and six types of "different" trials, where they do not.

Average accuracy rates appear in Table (2), where an asterisk indicates a result significantly above or below chance. The trend here is similar to the discrimination task.

		Audio		
		[baɾ]	[ba̤ɾ]	[bʱaɾ]
Image	'twelve'	97.5*	70*	5*
	'outside'	65	62.5	42.5
	'burden'	17.5*	22.5*	70*

Table (2) Percentage "same" response in ID task. Shaded cells are those where the audio and picture matched (correct answer: "same"). In unshaded cells, the audio and picture did not match (correct answer: "different"). Thus, greater accuracy is indicated by **high values in shaded boxes** and **low values in unshaded boxes**. Asterisks indicate response rates that differ significantly from chance (0.5).

Listeners accurately identified that the picture and audio stimulus matched in [baɾ] 'twelve' and [bʱaɾ] 'burden' trials, but—like the AX task—did not perform significantly above chance in identifying that the audio [ba̤ɾ] matched the image for 'outside'. In fact, participants performed at chance whenever the image for [ba̤ɾ] 'outside' was used.

Given the audio stimulus [ba̤ɾ] and the breathy consonant image 'burden', listeners identified the mismatch between image and word with above-chance accuracy (*p* = .0114). This indicates that participants identify [bʱaɾ] as an acceptable realization of /ba̤ɾ/ 'outside'. However, they do not do the inverse: [ba̤ɾ] is not an acceptable realization of /bʱaɾ/ 'burden'. We hypothesize that because /bʱaɾ/ contains a breathy consonant, only an utterance with sufficiently salient breathiness can pass as a realization of this item.

The complex relationship between modal and breathy vowels is also evident in these data. In both trials with a modal audio stimulus and breathy vowel image, and trials with breathy vowel audio and breathy vowel image, listeners were at chance. They did not know whether the audio and image matched, but rather were guessing. Furthermore, they correctly indicated that the [ba̤ɾ] stimuli did not correspond with the image for [baɾ] 'twelve' only 30% of the time: the remaining 70% of the time, they (incorrectly) indicated that the audio and image matched. This is significantly below chance (p = .0114), meaning once again that they weren't guessing; rather, they actively indicated that the breathy vowel audio stimuli corresponded with the definition of the modal word.

6. Discussion

The primary question driving this research revolves around how well native listeners are able to distinguish between CʱV and CV̤ sequences in Gujarati, sequences known to have similar acoustic cues but with differences in degrees and timing (Esposito & Khan 2012). A related issue proposed by Fischer-Jørgensen (1967) was also addressed: the variation that allows for a fully modal production to be accepted as a target breathy vowel.

Two results are important to highlight: (1) the inability of listeners to discriminate between [ba̤ɾ] and [bʱaɾ] significantly above chance; and (2), the inability of listeners to reliably identify that [bʱaɾ] does not correspond with the image for 'outside' (/ba̤ɾ/). Both suggest that CʱV and CV̤ sequences are not reliably differentiated by listeners.

The discrimination task most directly addressed the salience of the difference between any two categories. Ideally, presentation of two audio stimuli in immediate succession causes participants to compare the acoustic properties of the stimuli without categorizing them (Key 2012), and the findings presented here for listeners of Gujarati indeed strongly suggest that the acoustic differences between CʱV and CV̤ sequences are not sufficiently robust. When presented with a trial pairing a breathy Cʱ and a breathy V̤ stimulus, listeners responded with "different" just 56% of the time, meaning that they were at chance: they could not accurately distinguish the two stimuli as being different.

Similarly, in the identification task, participants did not reliably indicate the mismatch between a [bʱaɾ] audio stimulus and a [ba̤ɾ] 'outside' image. Again, they performed at chance. In the inverse type of trial, however, in which the [ba̤ɾ] audio was paired with the [bʱaɾ] 'burden' image, participants reliably indicated the mismatch. In other words, participants were willing to identify [CʱV] as a realization of /CV̤/, but not willing to identify [CV̤] as a realization of /CʱV/. This raises the possibility that there is ambiguity in the robust breathiness associated with consonants: listeners recognize the breathiness in [bʱaɾ] stimuli, but are willing to assign it to either the consonant or the vowel. Thus, it is deemed acceptable as a realization of either /bʱaɾ/ or /ba̤ɾ/. This too supports the hypothesis that the two types of breathy stimuli are not well distinguished.

Recall that listeners performed at chance when provided with a matched breathy V audio and image pair in the ID task, and when presented with two breathy V stimuli in the AX discrimination task. This raises the possibility that the breathiness associated with vowels is variable in a way that consonant breathiness is not: listeners do not reliably identify the breathy nature of the breathy vowel stimulus, and are therefore unwilling to consider it a realization of something that should have salient breathiness, namely /bʱ/.

The story, then, is that this confusion runs in only one direction: [bʱaɾ] can be mistaken for /ba̤ɾ/, but the reverse is not true. Furthermore, participants have a tendency to allow [ba̤ɾ] stimuli to serve as acceptable realizations of /baɾ/. This may shed light on the reason [ba̤ɾ] is so rarely mistaken for /bʱaɾ/: the breathiness in [ba̤ɾ] is subtle enough to pass for a fully modal /baɾ/, and not robust enough to pass for breathy consonant /bʱaɾ/.

The idea that a modal sequence [CV] may serve as an acceptable realization of a breathy vowel /CV̤/ but that a breathy vowel [CV̤] is not an acceptable realization of modal /CV/ was initially proposed by Fischer-Jørgensen (1967), and the results of this experiment support this claim. The free sort results, for instance, indicate that there is an increased probability of overlap between [ba̤ɾ] and [baɾ], which some speakers put together in a single group, while the tendency for [bʱaɾ] to remain distinct. This was true across the different response patterns exhibited by our participants. In the discrimination task, participants performed at a rate significantly below chance in trials involving a modal stimuli and breathy vowel, indicating that participants reliably consider breathy vowel stimuli and modal stimuli to represent the same word. They're not guessing; they're actively calling the two "the same". Under the hypothesis that [baɾ] can serve as a realization of /ba̤ɾ/, participants may reliably hear the difference between stimuli of each type yet consider them two acceptable variants of the same word.

Participants also performed significantly below chance when identifying that a breathy vowel stimulus, [ba̤ɾ], does not correspond with the picture for the fully modal 'twelve', and they could not accurately identify that [ba̤ɾ] corresponds with picture for the breathy V 'outside.' Participants were not inaccurate at correctly identifying breathy vowels, in other words; they were accurate at misidentifying them as modal stimuli. This indicates that the degree of breathiness in [ba̤ɾ] stimuli was not sufficiently salient.

For the results of the ID task to be consistent with the hypothesis that /CV̤/ can be realized as [CV], but /CV/ cannot be realized as [CV̤], participants should identify modal [baɾ] stimuli as corresponding with the images representing both /baɾ/ 'twelve' and /ba̤ɾ/ 'outside'. The results align with this expectation. One would also expect salient breathiness in the stimulus to prevent participants from incorrectly identifying a stimulus as modal, and this too is borne out in both the discrimination tasks and the ID task. Listeners identified that audio stimuli of fully modal [baɾ] and breathy consonant [bʱaɾ] were different 79% of the time, and noted the mismatch between a breathy consonant [bʱaɾ] audio and the picture for modal /baɾ/ 'twelve' 95% of the time.

A potential explanation for the trends seen here is that the differences between CʱV and CV̤ sequences are not robust enough to be perceptually salient, because breathy vowels are inadequately cued. We propose that breathiness functions no differently from other continuous variables that are perceived categorically. For example, the perception of VOT in English is not completely categorical; there is a window of ambiguity in which an alveolar stop can be perceived as either a /t/ or a /d/ (Eimas & Corbit 1973, among others). The perception of breathiness can be thought of similarly but with a suite of continuous variables representing spectral tilt, spectral balance, and noise. If the strength of the acoustic cues for breathy vowels lies near the perceptual threshold between breathiness and modality but those for breathy consonants do not, the breathiness of CʱV stimuli should be easily identifiable while that of CV̤ stimuli should be more ambiguous. Consistent with the data, listeners who are sensitive to cues in degree of breathiness in CʱV sequences but not to cues in its timing would be able to correctly identify a sequence as breathy but not be able to reliably categorize it as either CʱV or CV̤. Similarly, if cues to the degree of breathiness of a CV̤ sequence are insufficient to determine with certainty that the stimulus is breathy, then that sequence would be incorrectly categorized as CV. A schematization of this proposal appears in Figure (4), in which the vowel after Cʱ is represented with intense breathiness at first before a gradual decrease, and V̤ is represented with more moderate, increasing breathiness. The breathiness associated with Cʱ falls outside the zone of ambiguity, while the breathiness of V̤ does not.

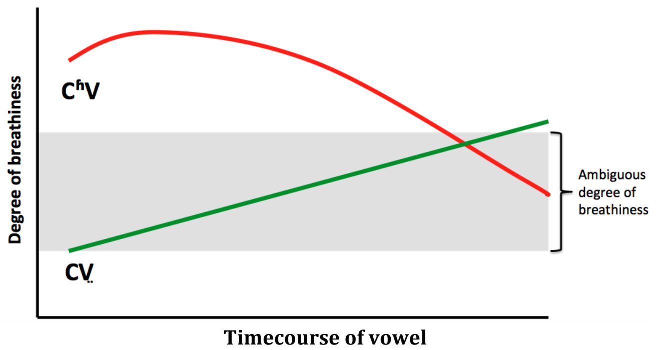

Figure (4) Schematization of degree of breathiness across timecourse of vowel

In the scenario proposed by this explanation, listeners are sensitive to the presence or absence of breathiness. Significant indication of breathiness may be sufficient cause for excluding a stimulus from being modal, but insufficient cause for determining if the breathiness is associated with the consonant or vowel. The results of the present study align with this interpretation, and strongly suggest that it merits further investigation.

6. Conclusion

This study investigated the perception of C^{ɦ}V, CV̤, and CV sequences by native listeners of Gujarati. Participants reliably recognize the presence of breathiness in C^{ɦ}V sequences, but are not necessarily capable of determining whether that breathiness is associated with the vowel or consonant. They do not reliably recognize the presence of breathiness in CV̤ sequences, however, often indicating them to be the same as or an acceptable realization of a modal CV sequence. The overarching trend, then, is that C^{ɦ}V can be perceived as either C^{ɦ}V or CV̤, and CV̤ is often indistinguishable from CV. While ongoing work will further explore the specifics of these trends, it is evident from this study that there is a problem in differentiating C^{ɦ}V and CV̤ sequences as well as an overlap in either the categorization or perception of CV̤ and CV.

References

Bhaskararao, Peri. & Vuppala, Anil Kumar. 2014. Automatic detection of breathy vowels in Gujarati speech. *International Journal of Speech Technology*, 17(01): 75-82.

Berkson, Kelly Harper. 2012. Phonation types in Marathi: an acoustic investigation. Ph.D. diss.

Bickley, Corine. 1982. Acoustic analysis and perception of breathy vowels. *Speech Communication Group Working Papers,* Research Lab of Electronics, MIT, 73-93

Clopper, Cynthia G. 2008. Auditory free classification: Methods and Analysis. *Behavior Research Methods,* 40(2): 575-581.

Dutta, Indranil. 2007. Four-way stop contrasts in Hindi: An acoustic study of voicing, fundamental frequency and spectral tilt. Ph.D. diss.

Eimas, Peter, & Corbit, John. 1973. Selective adaptation of linguistic feature detectors. *Cognitive Psychology,* 4.1: 99-109.

Esposito, Christina M. 2006. The effects of linguistic experience on the perception of phonation. Ph.D. diss.

Esposito, Christina M. & Khan, Sameer D. 2012. Contrastive breathiness across consonants and vowels: a comparative study of Gujarati and White Hmong. *JIPA,* 42(2): 123-143.

Fischer-Jørgensen, Eli. 1967. Phonetic analysis of breathy (murmured) vowels in Gujarati. *Indian Linguistics*, 28: 71-139.

Gordon, Matthew. Ladefoged, Peter. 2001. Phonation types: a cross-linguistic overview. *JPhon,* 29: 383-406

Huffman, Marie K. 1987. Measures of phonation in Hmong. *JASA*, 81: 495–504.

Key, Michael. 2012. Phonological and phonetic biases in speech perception. Ph.D. diss.

Khan, Sameer D. 2012. The phonetics of contrastive phonation in Gujarati. *JPhon,* 40: 780-95.

Ladefoged, Peter, & Maddieson, Ian. 1996. *The sounds of the world's languages*. Blackwell.