DYNAMICS OF VOICE QUALITY OVER THE COURSE OF THE ENGLISH UTTERANCE

Elizabeth Bird[1], Marc Garellek[2]

[1]University of California San Diego Department of Bioengineering, [2]University of California San Diego Department of Linguistics
embird@ucsd.edu, mgarellek@ucsd.edu

ABSTRACT

Ends of utterances are often creaky in American English, but the changes in voice quality across the utterance are still poorly understood. In this study, we use electroglottographic contact quotient to track voice quality changes over the course of the English utterance. 10 speakers were recorded using both audio and electroglottography while reading sentences containing mostly sonorants. Contact quotient was measured from the beginning of the utterance to the onset of phrase-final creak, providing a continuous "contact" track, akin to a pitch track. Results indicate some consistent patterns: beginnings of utterances are usually characterised by a rise in contact; ends of utterances, which are usually creaky, are characterised by either a drop in contact (suggesting that phrase-final creak may be unconstricted in quality) or by a constant degree of contact.

Keywords: voice quality, phonation, electroglottography, contact quotient, phrasing

1. INTRODUCTION

In American English, the ends of utterances and utterance-internal phrases are frequently characterised as being 'creaky'; that is, they have irregular voicing. This phenomenon, which we refer to as *phrase-final creak*, has been well documented in American English [3,6,9,12]. Although phrase-final creak (indeed, all creaky voice, [8]) is typically thought to involve increased vocal fold constriction, this is not always the case. For instance, Slifka [13] found that speakers sometimes produced phrase-final creak with increased airflow and glottal opening, suggesting an unconstricted creak [8]. Still, more recent work suggests that speakers of American English generally produce phrase-final creak with increased constriction, as indexed acoustically by lower H1*–H2* (difference in amplitude between first and second harmonics) [4].

In contrast to the ends of phrases and utterances, it is still largely unclear how voice quality varies earlier in the utterance. There is some evidence to suggest that the very beginning of phrases tends to be breathier than later parts [5], and that phrase-medial words with post-lexical prominence (i.e., a pitch accent) are produced with more constriction [2]. Still, the dynamics of voice quality over the course of the American English utterance have yet to be detailed in any systematic way: How much cross-speaker variability is there in voice quality changes over the course of the utterance? Are there consistent patterns in voice quality as a function of phrasal position, similar to declination of f0 [10]?

In this study, we address these questions by measuring and modelling changes in voice quality over the English utterance. Voice quality is analysed primarily using contact quotient (CQ) measured from the electroglottographic (EGG) signal, allowing a continuous measuring of voice quality independent of the segmental variation in speech due to changes between consonants and vowels.

Below we outline a method of modelling the CQ track, comparable to that of a pitch track. This should enable researchers to analyse dynamic changes in voice quality over different portions of speech, with implications for understanding both laryngeal dynamics and coordination between laryngeal and supralaryngeal articulation. Then, by relating dynamic changes in quality to changes in f0, we determine the extent to which voice quality is dependent on or co-varies with pitch.

2. METHOD

Speaker audio and EGG waveforms were recorded during the reading of sentences constructed of modally voiced sonorants and vowels. From the EGG signal we calculated the contact quotient (CQ) of these tokens, which was then analysed to determine its behaviour at the beginning, middle and end of the sentences. Details of this procedure are outlined below.

2.1. Stimuli

Stimuli sentences contained mostly modally voiced sonorants and vowels, without glottalization or aspiration noise. We avoided creating sentences with obstruents and non-modal voice so as to be able to determine a baseline of voice quality for modally

voiced sounds. We avoided non-modal voicing using the following criteria: (1) consonants were mostly limited to nasals, liquids, and glides; (2) sounds that would lead to constriction or spreading of the vocal folds, including glottal stops and aspirated stops, were excluded; (3) /t/ and /d/ were included only if they were likely to be produced as taps; and (4) word-initial stressed vowels were avoided because they are likely to have glottalization [5]. Sentences were constructed in sets to contain similar sounds to produce sentences of 7 to 11 syllables in length: e.g., "We will win a lottery" [wi wɪl wɪn ə lɑɾəɹi] and "We will win a yearly lottery" [wi wɪl wɪn ə jiɹli lɑɾəɹi] were both included.

2.2. Participants

Ten native speakers of Californian English were recruited from the undergraduate student body at UC San Diego (age mean and standard deviation of 20 ± 1 years, 7 women and 3 men). Speakers all had normal speech on initial interview and consenting, and received course credit for their participation. Exclusion criteria included having primary language other than English, a known speech pathology, or being a speaker of another variety of English.

2.3. Task

Speakers were visually presented with one sentence at a time, and were instructed to read each individual sentence fluently, without emphasis on a particular word (that is, with broad focus). They were instructed to take a breath before beginning to read each sentence. In one session, speakers said 252 tokens over 10 minutes. Speakers were instructed to avoid list intonation and to take a break if the manner in which they were reading the stimuli changed mid-task.

2.4. Audio and EGG measurements

Audio and electroglottographic (EGG) waveforms were simultaneously recorded in Audacity, digitized to the computer's sound card at a sampling rate of 44.1 kHz and stored as a 16-bit wav file. Audio was recorded with a Shure SM10A microphone placed 10 cm from speakers' mouths in a sound booth. The EGG signal was recorded using a two-channel electroglottograph (Model EG2, Glottal Enterprises), and was filtered with a 20 Hz high-pass filter before being stored.

Waveforms were annotated using Praat [1]. Also annotated was the presence of a glottal stop, full oral stops (e.g. pronouncing a [t] instead of tap), presence of pauses or perceived utterance-medial phrase boundaries, and atypical declarative-statement intonation. Sentences were excluded if the speaker was disfluent: i.e. if they used long pauses, misread the sentence, or had large pitch disruptions.

EggWorks [15] was used to calculate CQ for each pulse of the EGG signal, giving a continuous CQ contour. CQ is measured as the ratio of contact phase duration to the total duration of the EGG pulse. The contact phase boundaries were determined using the hybrid method [7].

VoiceSauce [14] was used to obtain acoustic measurements such as f0 and cepstral peak prominence (CPP), a harmonics-to-noise ratio measure. Measures of f0 were calculated to relate changes in voice quality to those of pitch. CPP was used as a method of estimating the onset of phrase-final creak, as we discuss below.

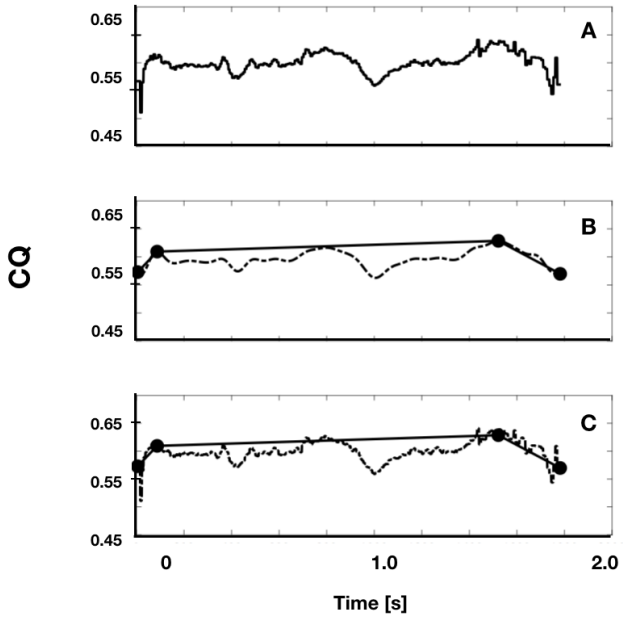
3. ANALYIS OF THE CONTACT QUOTIENT SIGNAL

Continuous CQ tracks (**Fig 1A**) from each sentence were analysed from onset to the beginning of phrase-final creak (present in more than 95% of all speaker sentences) to facilitate description and summary of their contour shape. We excluded CQ measures after the onset of phrase-final creak, because on visual inspection, the CQ values were frequently spurious, owing to the high occurrence of multiple, lower amplitude, pulsing. The onset of phrase final creak was determined using a signal-to-noise threshold determined from the audio waveform. This threshold was identified by finding the time point at which CPP dropped below 1.1 times its mean value for a specific utterance. Additionally, a research assistant marked the onset of creak by visual inspection of the f0 track, based on the point at which the f0 became irregular up until the end of the sentence. Local f0 irregularities that were short and did not last till the end of the sentence were ignored. We found large convergence between the visual inspection method and the automatic detection method, suggesting that our criteria using CPP worked well for estimating the onset of phrase-final creak. Moreover, any extreme CQ values (> 0.8 or < 0.2) were removed.

The CQ track was then smoothed and low-pass filtered (with a cut off frequency 10 Hz; **Fig 1B**). The low-pass filter had its local maxima detected using the built-in *findpeaks* function in MATLAB [11]. Using these peaks, three slopes in the filtered CQ track were defined. These included: (1) the initial slope, defined as the slope from the first CQ point to the first maximum; (2) the middle slope, defined as the first to the final CQ maximum; and (3) the final slope, defined as the slope from the

final CQ maximum to the last CQ value before onset of phrase-final creak (**Fig 1C**). These slopes were chosen because they connect prominent landmarks seen in all tokens on visual inspection. Time was not normalized to preserve fidelity of the slopes.

Figure 1: Overview of CQ track analysis method. The unedited CQ track (**1A)** is smoothed and low-pass filtered **(1B)**. From this smoothed waveform, the points of interest are selected and the slopes calculated. The calculated contours are then fit to the original unedited waveform **(1C)**.



4. RESULTS

For all speakers, CQ tracks have an increasing initial contour, with a slope value range of +1.36 to + 0.30s^{-1}. A rise in CQ for this initial slope indicates a tendency towards more constriction at the onset of the utterance. The middle CQ contour is generally flat, ranging from -0.03 to +0.03s^{-1}. The middle contour is positive for 7 of the speakers and 0 for one of the speakers. The final CQ contour decreases for all speakers, with a range of -0.33 to -0.07s^{-1}, indicating a tendency towards *less* constriction before the onset of phrase-final creak.

In general, CQ contours (**Figs 2** and **3**) of the speakers (n = 10, 7 females) all show an increase in CQ from utterance onset to first CQ peak, and show a relatively flat CQ between the first and final peaks. (Below, we indicate that the CQ peaks tend to coincide with f0 peaks, suggesting they relate to post-lexical prominence.) The final contours of different speakers (from the last peak in the CQ track to the final point before phrase-final creak) can have a very rapid (**Fig 2A**) or very slight (**Fig 2B**) decrease or increase from the final CQ maximum to the final point of the utterance before creak onset. These contour values ranging from negative to slightly positive do not appear to be due to distinct patterns. Instead, when plotted as a histogram, they are part of a continuous distribution (**Fig 3**).

Figure 2: Representative CQ contours over the time course of an utterance of an individual token from a speaker with a rapidly-decreasing final slope (**2A**), and from a speaker with a steady final slope (**2B**).

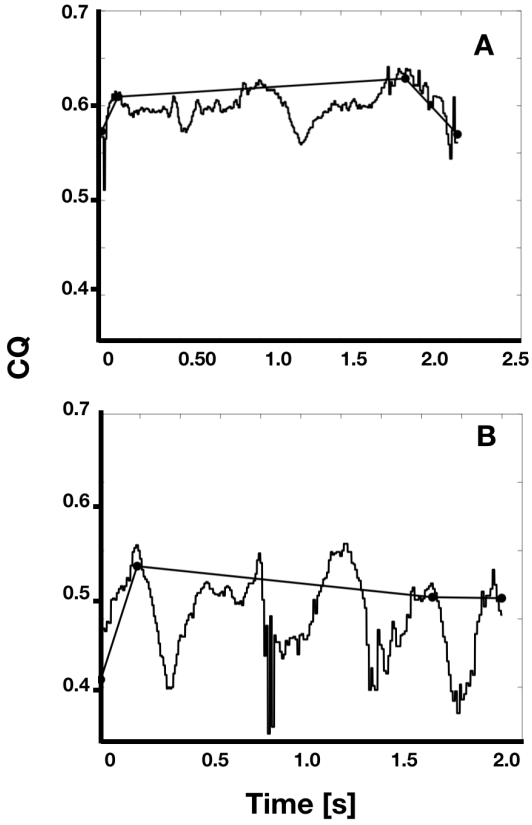
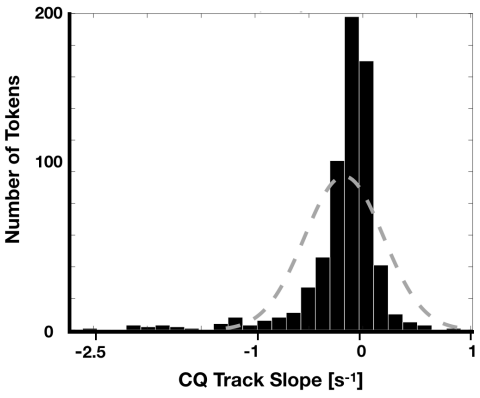
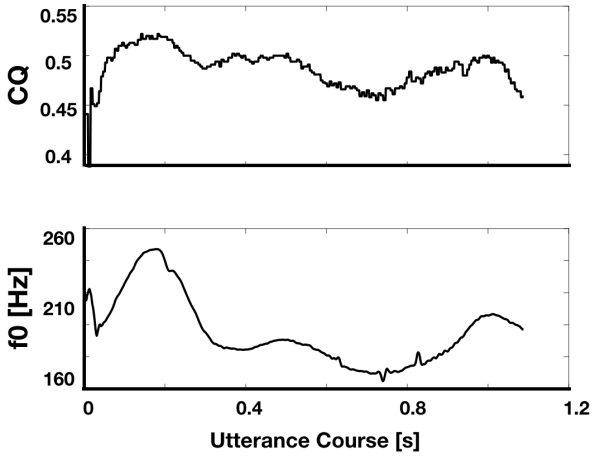


Figure 3: Distribution of final CQ slopes. The distribution appears negatively skewed but also contains near zero and positive final slopes.



The correlation coefficient between mean CQ and f0 over all tokens for a given speaker varied from r = -0.12 to +0.99, with 6 speakers showing positive correlations. The peaks of the CQ and f0 contours tend to occur at the same time (**Fig 4**).

Figure 4: CQ and f0 tracks for a representative token. The peaks in the CQ track (including the initial and final peaks) coincide with f0 peaks, suggesting that higher CQ accompanies post-lexical prominence.



5. DISCUSSION AND CONCLUSIONS

The contact quotient (CQ) tracks of 10 speakers of Californian English demonstrate similar patterns across all utterance segments. All speakers exhibit increasing CQ contours at the beginning of utterances (here operationalized as the utterance onset to the first peak in CQ), indicating a tendency towards increased glottal constriction. The CQ slope also had a larger range of values at utterance onset. Together, this implies that speakers use a breathier voice quality at utterance onset compared with later in the utterance [5], and suggests that voice quality tends to be more variable at the utterance onset before stabilizing in the middle of the utterance.

Both patterns also show a relatively flat, minimally variable, CQ contour over the middle of the utterance compared to beginnings and ends. This less-variable middle contour suggests that speakers may have a target voice quality for the majority of the utterance. It should be noted that the maxima in CQ over an utterance generally coincided with f0 peaks (i.e., with high pitch accents). This suggests that prominent syllables are also associated with increased constriction [2]. Given that the utterance-medial slope is fairly level, there may be a target voice quality associated with prominent syllables, such that there is little change from one prominence peak to another.

The CQ behaviour at the ends of the utterances is is interesting because it not consistently negative, and instead can be negative or near zero and slightly negative or positive. The fact that the majority of speakers do not show an increase in CQ does not necessarily preclude an increase in constriction during phrase-final creak; this discrepancy may instead be due to our analysis of the CQ track. In this study, the CQ track ended at the onset of phrase-final creak, which may have led to the removal of constricted phrase-final creak in many tokens. However, this was done because of spurious CQ values during creak, which often showed multiply-pulsed voice [8], whose weaker pulses do not figure in to the CQ values. With more accurate detection of CQ during creak, future work will look at how CQ contours change during creak.

Speakers who have a negative sloping CQ contour and noticeable decrease in CQ between the final prominence peak and onset of phrase final creak could be producing less constricted phrase-final creak, rather than prototypical creak with constriction [8,13]. The middle flat CQ contour, coupled with the decreasing final CQ contour, together shows a pattern similar to that of declination in f0. From the similarity in prominence peaks between the f0 and CQ tracks, as well as the decrease in CQ that many speakers show at the ends of utterances, it might be expected for f0 to be positively correlated with CQ. But the correlation between f0 and CQ was speaker dependent, ranging from positive to negative values.

To conclude, in this study we quantified how CQ varies over the English utterance. We find that utterances usually begin with a rise in the CQ track, consistent with a breathier voice quality at utterance onsets. The flat CQ contour over the middle of the utterance between prominence peaks suggests that many speakers do not vary much in voice quality in that position. However, many speakers produce a a drop in constriction towards the end of the utterance, consistent with a less constricted voice quality leading up to the onset of phrase-final creak.

6. ACKNOWLEDGEMENTS

We thank Viktoriya Dorokhina for help with recording of participants, as well as Emily Huo, Min Kyung Michelle Lee, Nicole Oberman, Eileen Prieto, and Erin Wu for their help with file annotation. We thank members of the UCSD Phonetics-Phonology group for helpful feedback. An earlier version of this work was presented at the Fall 2017 meeting of the Acoustical Society of America.

7. REFERENCES

[1] Boersma, P. and Weenink, D. Praat. 2016. Retrieved from www.praat.org

[2] Campbell, N., Beckman, M. 1997. Stress, prominence, and spectral tilt. *Intonation: theory, models, and applications*, 67–70.

[3] Crowhurst, M. J. 2018. The joint influence of vowel duration and creak on the perception of internal phrase boundaries. *J. Acoust. Soc. Am*. 143, EL147–EL153.

[4] Davidson, L. *To appear*. The effects of pitch, gender, and prosodic context on the identification of creaky voice. *Phonetica*. https://doi.org/10.1159/000490948.

[5] Garellek, M. 2014. Voice quality strengthening and glottalization. *J. Phon.* 45, 106–113.

[6] Garellek, M. 2015. Perception of glottalization and phrase-final creak. *J. Acoust. Soc. Am*. 137, 822–831.

[7] Howard, D.M., 1995. Variation of Electrolaryngographically Derived Closed Quotient for Trained and Untrained Adult Female Singers. *Log Phon Vocol 9*, 163-72.

[8] Keating P., Garellek M., Kreiman, J. 2015. Acoustic properties of different kinds of creaky voice. *Proc. 18th ICPhS* Glasgow.

[9] Kreiman, J. 1982. Perception of sentence and paragraph boundaries in natural conversation. *J. Phon.* 10, 163–175.

[10] Ladd, D.R., 1984. Declination: a review and some hypotheses. *Phonol. Yearbook* 1, 53–74.

[11] MATLAB_R2018b. 2018. Natick, Massachusetts: The Mathworks Inc.

[12] Redi, L. Shattuck-Hufnagel, S. 2001. Variation in the realization of glottalization in normal speakers. *J. Phon.* 29, 407–429.

[13] Slifka, J. Some physiological correlates to regular and irregular phonation at the end of an utterance. *J. Voice* 20, 171–186.

[14] Shue, Y.L. VoiceSauce. 2016. Retrieved from http://www.phonetics.ucla.edu/voicesauce/

[15] Tehrani, H. EGG Works. 2012. Retrieved from http://www.appsobabble.com/login.aspx?ReturnUrl=functions/EGGWorks.aspx