FACTORS FOR IMPROVING CLASSIFICATION OF CREAKY VOICE

Theresa Kettelberger
Department of Computer Science
University of Rochester

ABSTRACT

Very few algorithms exist to identify vocal fry, or voice creak, in speech. Despite this, creaky voice (CV) is an important paralinguistic feature across languages and its detection would benefit many systems. Existing algorithms are reliant on periodicity of the pulses of CV despite the fact that CV is often aperiodic. Newer algorithms propose alternate measures, but these are generally limited in the environments where they are effective. This paper investigates possible heuristics for CV, several taken from these newer methods and some new ones, and integrates the most effective into a support vector machine. While the support vector machine does not render very good results, the results offer opportunity for further study.

1. INTRODUCTION

"Creaky voice" or "CV" is a descriptor for an unusual vocal effect which is popularly known for its use by the Kardashians and other female users of so-called "vapid" speech. This type of phonation, which sounds low and croak-like, is also referred to as "vocal fry", "laryngealization", "glottalization", and "the pulse register". Its unique sound is caused by compressed, slack vocal folds which result in extra glottal pulses which may be of a very large period or irregular [1].

CV occurs in far more registers than "vapid speech". In English, it indicates prosodic and emotional information. In many Native North American languages, voice creates phoneme-level contrasts. In tonal languages, certain tones may be indicated by the presence of creaky voice, such as the third tone in Mandarin. This is because CV sometimes is used as a stand-in for actually lowering ones voice pitch, because CV sounds very low. Relatedly, CV creates problems for pitch estimation algorithms. Interest in detecting CV lies in its value to speech recognition, emotion detection, intonation categorization, improved pitch detection, and other systems.

In the face of these interests, detecting CV is a nontrivial problem. In the literature, CV is a term that encompasses several more specific types of vocal glottalization, all of which have slightly variant acoustic properties [2]. CV is also particularly difficult to detect in male speech and speech with a low fundamental frequency, or f0, without generating a large number of false positives. Finally, CV is not a homogenous effect. Even within consistently presenting CV, there are two separately occurring features: high-energy pulses and low-energy suppressed areas between. Algorithms must accommodate both of these to detect whole segments of voice creak.

For a long time, the detection of creaky voice relied on autocorrelation to detect low-frequency periodicity which might be the successive glottal pulses resulting from CV [4]. The flaw of this method is that it is far more likely to detect CV where glottal pulses fall periodically or close to that. While this is true of some CV, many CV pulses are irregular.

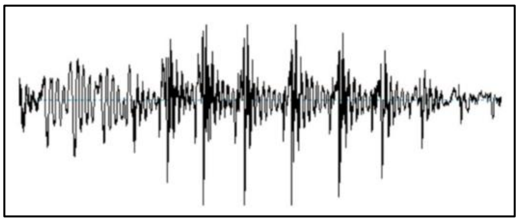

Figure 1. An example of a creaky voiced vowel in a female speaker, including irregular pulses and amplitude

Improved algorithms exist, although these all have some limitations in the types of creak detected as well as the environments in which they are detectable.

One algorithm classifies segments based on measures of aperiodicity, periodicity, and "very short term" power peak detection [3]. This algorithm was designed for a specific subset of CV and may struggle with the others. It also used data sets that were mostly female speakers and scores more poorly on male speakers than its reported overall accuracy [1].

Another algorithm applies a resonator to speech to detect glottal pulses. This algorithm claims accuracy but only detects pulses, not entire segments of CV [1].

The only algorithm that takes advantage of the spectral domain is Martin which uses sudden changes in the number of harmonics. However, this algorithm cannot detect CV after silences and unvoiced segments, which is an extremely large number of environments in which to fail [4].

The most state of the art and well-used heuristic is Ishis's measure, the ratio between the first and second harmonics [5]. However, this algorithm was not well-tested on male speech and presents itself as a primary, but not exclusive, measure

The goal of the experiment in this paper was to quantify another measure which might make these somewhat effective algorithms more robust to different environments and speakers. Particularly, I wanted to investigate measures that might be agnostic to the average f0 range of the speaker. I investigate sound levels and a variation on spectral flux as parameters with mixed results.

2. DATA AND ANNOTATIONS

I performed my experiment with data that I collected myself. I did this for several reasons. Firstly, many preexisting datasets were behind a paywall. Secondly, many of these datasets only included 1-3 speakers which were almost all female. Finally, these datasets were not pre-annotated for creaky voice segments, and I would have been annotating myself regardless. I decided that I could build a more useful dataset than existed previously.

I recorded about an hour and a half of speech from 8 different speakers, 4 male and 4 female, using a ZOOM recorder with two microphones. I interviewed speakers late at night because tired speakers are more likely to produce CV than average speakers. I then annotated the data to use as training and test data.

The annotation schema I used was very simple. In the software Praat, I used audio as well as visual cues from the wave form and labeled segments of creaky voice as 1 and modal (normal) voice as 0 [6]. Each segment was an average of 0.9 seconds in length. Modal segments could include anything from silence to vowels to unvoiced segments. The split was about 60% modal and 40% creak.

These annotations were fed through a Python script with wave files. The wave files were then split into separate files based on the annotations for use as training data and test data.

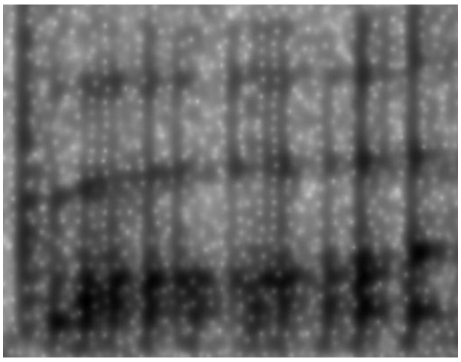

Figure 2. A creaky spectrum: Dark indicates more energy

3. MEASURES OF CREAK

3.1 Score

Score was implemented with the intent to imitate the way humans identify creak through the spectrum by positively scoring items with large fluctuations in the power per spectral frame. CV undulates between very energized pieces (the glottal pulse) and very damped sections. The goal of score is to capture this factor. This method felt intuitive and is a very simple measure.

The score has two components: change in spectral power and change in harmonic composition. Good candidates for creak should have a high change in spectral power, indicative of the irregular amplitude that accompanies CV as well as the way that the spectrum switches between high energy glottal pulses and low energy intermediary sections.

Good candidates will preferably also have a low change in harmonic composition. This avoids erroneously marking pieces of speech that change spectral power as creak when they are caused by something else, such as changes in the phoneme.

I measure the difference in spectral power by simply summing each window of the spectrogram across all frequencies. This measure was best when I put an upper bound on the frequencies included. An upper bound at 3500 Hz excluded high frequency background noise, as well as some of the high frequency noise caused by aspirated plosives and stridents such as /s/. The upper frequency noise caused by these phonemes is very high energy and highly variant and may appear as false positives for creak because the noise varies highly from frame to frame. This does not cut off all of this noise, but stops it from having a disproportionate effect on score measurements.

$$\Delta\text{Power} = \sum S_t - \sum S_{t-1}$$

My measure of change in harmonic structure is not based on measuring harmonics through autocorrelation or other more sophisticated measures. Not all of my training data has harmonic structure, and attempting to measure multiple autocorrelation peaks in segments without a fundamental frequency led to many errors. Instead, I approached it as more of a frequency profile than a harmonic one.

I found the highest energy k frequencies in each window of the spectrum. The index of the frequency bins of these are then stored and sorted. The difference between these sorted lists becomes the change in frequency/harmonics. The k can vary, although I found best results around 3-7 points. While this measurement is not equivalent to real measures of the harmonic structure, it has practical value. For voiced segments, which are the segments we are focused on, these maximum energy frequencies should be located around the formants of the phoneme.

$$\Delta\text{Frequency} = \sum^{k}(i_{k(t)} - i_{k(t-1)})$$

The score is a measure of the ratio of change in spectral power and change in harmonic structure. It is helpful to apply a weight, α, to the change in harmonic structure. Weighting it higher produces improved accuracy overall. We want to lower this weight, we only need it to have a small effect. My optimal α was 0.25.

$$\text{score} = \Delta\text{Power}/\alpha\Delta\text{Frequency}$$

This measure functioned best with a very high time resolution. I used a Gaussian window (the same as the default on my annotation software) of length 128. Raising the size of this window means that creak may not be captured as glottal pulses may lie entirely within windows and their high power will appear less significant buffered by surrounding lower power glottal suppression.

3.2 SNR and RMS

The literature leads us to expect a lower signal-noise-ratio for almost every type of voice creak [2], because of irregular f0 increasing the amount of noise, and glottal pressure stifling some of the sound between pulses. However, SNR returned no statistically significant or even interesting looking results no matter what parameters and noise were used for its calculation. While I do believe it is possible to use this measure, a quality measurement of SNR customized to creak was outside of the scope and limited timeline of this project.

While I intended to use RMS to detect silences that should not be tested or trained on, RMS became my most useful measure. RMS was low for any voiceless segments of speech, as expected, but rather than serving to differentiate between voiced segments and voiceless segments (segments with no f0, where the vocal chords do not vibrate), RMS was similarly low between voiceless segments and creaky voiced segments. Because creaky segments must be voiced – you cannot produce sounds with slack vocal folds if you are not using your vocal folds – this is a very good heuristic to tell the difference between creak and modal voiced segments. This is particularly intriguing because this method is agnostic to the distance between glottal pulses, theoretically neutralizing the difficulties with male and low f0 speech.

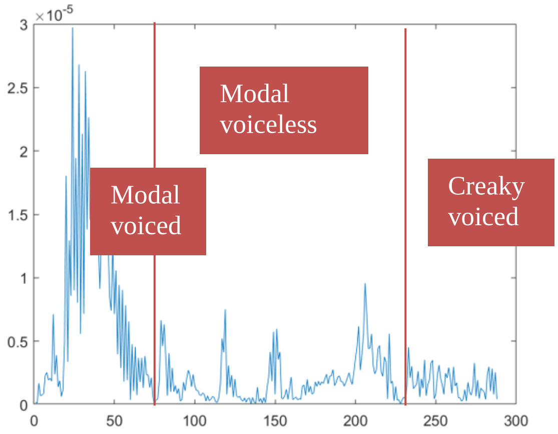

Figure 3. A graph of squared RMS across a short utterance, labelled for phonemes and phonation.

4. LEARNING

I used RMS and my own score parameter to train a simple linear support vector machine to classify creaky and noncreaky segments. The machine was trained on 50,000 examples. This is within the recommended range for the SVM that I used [7]. The input was ordered pairs of <score, RMS> for each spectral frame. Labels were 1 or 0 for creaky or modal generalized from my earlier annotations.

The training data was 30% creak and 70% modal voice. This imbalance is difficult to avoid because modal voice is much more common.

This SVM used an adaptive learning rate for fastest convergence. It used a shrinking heuristic, I trained and tested it several times on different samples of my data to make sure that my results were consistent, especially given my small amount of data. Another positive effect of my high time resolution in my spectrogram was that it stretched my data farther, enabling a more accurate model and lessening the chance of overfitting.

5. RESULTS AND DISCUSSION

My results had very low accuracy, and I will discuss the reasons for this in the rest of the paper.

The predicted labels on my test data, which had a similar imbalance to my training data, were an average of 56% accurate. When accounting for the balance of the data, the score was 54%, meaning we only managed to do about 4% better than random guessing would have been.

This was a discouraging result which caused me to investigate the possibility of overfitting. I tested my implementation on my training data, and while the results were somewhat higher – 57.3% - they were by no means good.

Of the incorrect predictions, 96% were false negatives for CV. This means my model is significantly skewed towards modal voice, which makes sense and is probably at least partially a result of my data skew.

In the following section, I discuss these negative results, the challenges that caused them, and what can be learned from this flawed implementation.

6. CHALLENGES

The main challenge in using a SVM for this problem is the noisiness of the data. This problem is two-fold, firstly because the data may be noisy in the same way any data might be, but secondly because of the extreme diversity of the measurements even within the same annotated segment. Figure 2 shows the variation that can occur in measurements like this. There are many peaks and valleys even in the relatively sound RMS measurement. Peaks and valleys in both score and RMS measurements are considered equally representative while training an SVM.

Even if trained to recognize high scores and low RMS as a good measure of creak, this will still lead to a great many false negatives because only a few frames per annotated segment may really demonstrate those qualities.

While overall, creaky segments have a higher score than non-creaky segments, there are many points in between these where the score is quite low because between glottal pulses, the spectral power may be decently consistent. It would be preferable to smooth these curves in a way that ignores valleys between very large peaks, but too much smoothing may ignore the fine-grained details necessary to detect CV in the first place.

I believe that a better use for the score parameter would be to measure for larger segments the density of peaks in the score. High density would indicate creaky voice. For this purpose, the change in frequency would not need to be included in the measurement.

Another difficulty is the relative nature of RMS. RMS measurements vary significantly between recordings and even within recordings. While I tried to normalize volume based on the longer recordings which sourced my shorter training and testing data, there is not a single number that divides modal and creaky RMS measurements. Some modal voiced segments, even within a single recording, may be quieter or not as well-voiced as others. Furthermore, normalizing a segment which contains only creaky voice may artificially raise its RMS to modal levels.

Finally, due to variant speaker pitch range and vocal habits, the optimal window size for any type of glottal pulse analysis is heavily variant. For modal segments, some of my speakers' normal glottal pulses happen at the same rate as creaky pulses for my speakers with a higher f0. Tailoring my window size to the higher-pitched speakers gave false positives on lower pitched speakers on spectral measurements (although not RMS). Tailoring my window size to lower pitched speakers resulted in ignoring higher pitched speakers creaky voice. On top of this, many of my male speakers used a lot of pressed or tensed voice, a similar effect to creaky voice that is not low-pitched or irregular, but still causes unusually strong glottal pulses with suppression in between. This increases the problem of false positives. A better system could train individually on a single speaker or recording environment, or would first take the average pitch range of the speaker into account.

7. CONCLUSION

While the SVM was not particularly successful, we did discover the usefulness of RMS as a measure to differentiate between CV and modal voiced segments. This was not exceptionally helpful for this project, because my data was not annotated for phoneme. However, many speech data sets are annotated by phoneme. In future experiments, given this information, we can attempt to improve popular existing methods' accuracy on male voices by testing only segments labelled as voiced phonemes and then testing the RMS in addition to other heuristics, such as Ishi's H1-H2. This is an interesting avenue for further study.

I do not think we should entirely throw out the idea of a trained classifier or SVM. Algorithms designed for this purpose continually show drastically better performance on the datasets which they were developed for [1], which is not a surprising result. One reason that this is such a significant effect is that optimal window size and spectrogram resolution, along with other parameters for wave-based analysis, vary based on individual speaker's voice and pitch range. A partially trained agent which can be completely trained on the beginning of a long recording and annotate what follows afterwards would in and of itself be helpful within intonation-detection and similar algorithms, solving some of the most salient problems for the application of CV detection.

I would love to try many techniques, including sending the whole spectrum to a classifier, separately training on men and women, and finding more sophisticated measures for score and noise levels. Unfortunately, as the semester ends, this project cannot include those further investigations which I look forward to pursuing on my own.

8. REFERENCES

[1] Drugman, Thomas. Kane, John. Gobl, Christer. Resonator-based Creaky Voice Detection. (2012). INTERSPEECH.

[2] Acoustic properties of different kinds of creaky voice (2015). 18th International Congress of Phonetic Sciences, Glasgow, Scotland.

[3] Ishi, Carlos T. Sakakibara, Ken-Ichi, Ishiguro, Hiroshi. (2011). A Method for Automatic Detection of Vocal Fry. IEEE Transactions on Audio, Speech, and Language Processing.

[4] Martin, Phillipe. Automatic detection of voice creak. (2012). CLILLAC-ARP, EA 3967, UFR Linguistique.

[5] Ishi, Carlos T. Analysis of Autocorrelation-based Parameters for Creaky Voice Detection. JUST/CREST.

[6] Paul Boersma and D Weenik. Praat: a system for doing phonetics by computer. Report of the institute of phonetic sciences of the University of Amsterdam. Amsterdam: University of Amsterdam, 1996.

[7] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.