Acoustic cues for the classification of regular and irregular phonation

Kushan Surana and Janet Slifka

Speech Communication Group, Research Laboratory of Electronics
Massachusetts Institute of Technology, Cambridge, MA, USA
slifka@speech.mit.edu

Abstract

Irregular phonation serves an important communicative function. It can be a cue to linguistic contrasts, and often serves as a marker for word and utterance boundaries. Automatic methods for classification and detection of regions of irregular phonation can be used to improve analyses of occurrences of irregular phonation and support technologies such as speech recognition and synthesis. This study proposes a set of acoustic cues from both the temporal and frequency domains — fundamental frequency, normalized RMS amplitude, smoothed-energy-difference amplitude and shift-difference amplitude— for separation of regions of regular and irregular phonation. Tokens from the TIMIT database are classified using support vector machines, trained on 114 different speakers and tested with 37 different speakers. Both genders are well represented in the data set and the tokens occur in various contexts within the utterance. In the test set, 292 of 320 irregular tokens (recognition rate of 91.25%), and 4105 of 4320 regular tokens (recognition rate of 95.02%) are correctly identified. [Work supported by NIH/NIDCD # DC02978.]

Index Terms: irregular phonation, nonmodal phonation, speech recognition, acoustic analysis, speaker variation

1. Introduction

Irregular phonation is used to convey both linguistic and non-linguistic information, where the specific type of information associated with an occurrence of irregular phonation depends on the language and the context. Irregular phonation may serve as a cue to speech segmentation (for example, [1-4]), and is lexically contrastive in some languages (for example, [4]). Automatic methods for detection and analysis of regions of irregular phonation not only support studies of the communicative role of such phonation patterns (in terms of database annotation and acoustic analysis), but also aid the development of technologies for speech recognition and synthesis capable of processing the range of natural phonation variation.

Additionally, irregular phonation can result from some voice disorders, and a reliable, accurate, and non-invasive automatic system for detection and monitoring of vocal fold abnormalities is one in a range of necessary tools for pathological speech assessment [5].

One approach to an automatic system for irregular phonation detection has been to expand a phone-based automatic speech recognition platform to include phones that are realized with irregular phonation [6]. Five acoustic cues derived from cepstral coefficients are used to train on three speakers and test on a fourth. The recognition rate for frames of irregular phonation is 67% with a false alarm rate of 7%.

Autocorrelation-based cues obtained from the residual signal as derived from inverse-filtering with linear prediction coefficients are used in [7] for speaker-dependent classification of normal, aspirated, and creaky phonation. For frames extracted from 404 utterance-final syllables for a single female speaker, a decision-tree paradigm resulted in a deletion rate of 13.7% and a substitution rate of 7.9% for irregular phonation, or a recognition rate of 78.4%.

In this paper, we propose a set of four acoustic cues, based in both the time-domain and the frequency-domain, which are designed to separate regions of regular phonation from regions of irregular phonation in a speaker-independent and context-independent manner for a large number of speakers. The cues are used in an automatic classification scheme and a discussion of the classification failures is provided.

2. Irregular phonation

Normal, voiced speech is characterized by quasi-regular vibration of the vocal folds. Although the vocal folds oscillate regularly when variables such as transglottal pressure, vocal fold tension, and vocal fold adduction are in particular ranges, irregularities in vocal fold vibration are observed for certain combinations of the values of the control variables. These irregularities in vocal fold vibration are more pronounced than the small cycle-to-cycle variations associated with the quasi-periodic quality of regular phonation. The terms *"modal"* and *"periodic"* are often used interchangeably with *"regular"* phonation. Similarly, *"nonmodal"* and *"aperiodic"* are often used to denote *"irregular"* phonation. However, nonmodal phonation includes irregular, aperiodic phonation as well as some forms of regular, periodic phonation such as breathy voice. Regions with very low frequency, periodic glottal pulses are also not typical of the normal range of phonation for a given speaker and are classified as irregular in this study, in spite of being periodic. In this study, irregular phonation is defined as:

> "A region of phonation is an example of irregular phonation if the speech waveform displays either an unusual difference in time or amplitude over adjacent pitch periods that exceeds the small-scale jitter and shimmer differences, or an unusually wide-spacing of the glottal pulses compared to their spacing in the local environment, indicating an anomaly with respect to the usual, quasi-periodic behavior of the vocal folds."

3. Data Set

Both regular and irregular tokens were extracted from a subset of the TIMIT corpus produced by speakers from the dialect regions "Northern" (dr1) and "New England" (dr2). The speech material is divided into training and testing subsets. The data set consists of utterances from 151 different speakers, with 114 different speakers in the train set and another 37 different speakers in the test set.

Irregular tokens were hand-labeled and extracted by analyzing the waveform in both the temporal and frequency domains to find regions which corresponded to the stated definition of irregular phonation. Tokens were confirmed by listening. The set of regular phonation tokens consists of all the vowels in dialect regions 1 and 2 labeled as \iy\, \ey\, \ae\, \aa\, \aw\, \ay\, \ao\, \oy\, \ow\, \uw\, \ux\, \er\, and \axr\ which do not contain an instance of irregular phonation. Table 1 shows the breakdown of regular and irregular tokens according to gender. Tokens occur in various contexts within the utterance (i.e. phrase-final, utterance-final, phrase-initial, etc.).

Table 1. *Gender breakdown of the number of regular and irregular tokens.*

	Total	Male	Female
Regular	8055	5458	2597
Irregular	1279	735	544

4. Acoustic Cues

A set of acoustic cues that can provide a rational basis for separating regular and irregular phonation was chosen that consists of fundamental frequency (F0), normalized root-mean-square (NRMS) amplitude, smoothed-energy-difference (SED) amplitude and shift-difference (SD) amplitude [8]. Table 2 summarizes the rationale behind choosing these cues, and highlights the expected range for the cue values for regular and irregular tokens.

Table 2: *Brief cue descriptions with expected ranges for cue values shown in bold within parentheses.*

Cue	Regular tokens	Irregular tokens
F0	Quasi-periodic signal with F0 in the range of ~72 to 266 Hz **(higher)**	Aperiodic signal either lacking an F0 or with an unusually low F0 **(lower)**
NRMS	Mid-range NRMS due to regular spacing of glottal pulses **(higher)**	Low NRMS from irregular and wide spacing of glottal pulses **(lower)**
SED	Lack of rapid energy transitions **(lower)**	Rapid energy transitions due to widely spaced glottal pulses **(higher)**
SD	Repeatable waveform structure **(lower)**	Lack of repeatable waveform structure **(higher)**

4.1. Fundamental Frequency

This study estimates F0 in a conservative manner, that is, unless F0 can be confidently estimated, a zero-value is output for the F0 estimate. The estimator is based on the peaks in the filtered-error-signal-autocorrelation sequence (FEAS) to minimize formant interaction. The steps for the calculation of the F0 estimate from the autocorrelation sequence are itemized below:

- If no peaks > 0.46 in FEAS, then the F0 estimate is 0.
- If only one peak is > 0.46 in FEAS, then the associated index is estimated as the fundamental period.
- If more than one peak is > 0.46 in FEAS, then a test is conducted to determine if all the peak indices are proportional to each other within a threshold of 0.02. If true, then the second peak index is the fundamental period.
- If all the above-mentioned criteria fail, the maximum peak above the threshold value is selected and its index determined as the fundamental period.

4.2. Normalized Root Mean Square Amplitude

Root-mean-square (RMS) amplitude is a common tool used in signal processing to estimate the average amplitude of a signal. The RMS amplitude of a token is normalized by the RMS amplitude of the entire speech signal from which the token is extracted to account for interspeaker variation in signal amplitude. This normalization assumes that the speaker uses the same "speaking level" over the course of the utterance. The mathematical formulation to compute this cue is,

$$A_{RMS} = \frac{(\frac{1}{L}\sum_{n=1}^{L-1} s[n]^2)^{0.5}}{(\frac{1}{N}\sum_{n=1}^{N-1} S[n]^2)^{0.5}}$$

where s[n] is the token; S[n] is the entire speech signal; N is the length of the speech signal in samples; and L is the length of the token in samples.

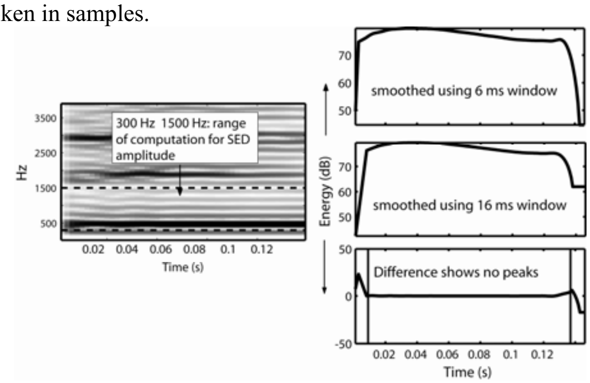

Figure 1. *Illustration of smoothed-energy-difference amplitude calculation for a regular token.*

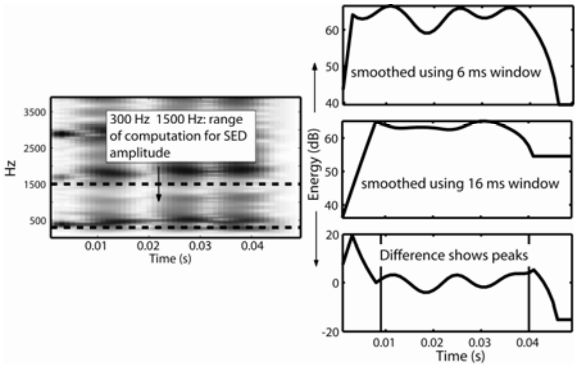

Figure 2. *Illustration of smoothed-energy-difference amplitude calculation for an irregular token.*

4.3. Smoothed-energy-difference amplitude

The smoothed-energy-difference (SED) amplitude is found by first computing the 512-point Fast Fourier Transform for each frame in a given token, and deriving the average energy in the 300 to 1500 Hz band (in dB). This frequency-averaged amplitude is smoothed using window sizes of 6 ms and 16 ms.

The difference between the smoothed averaged energy values for the two smoothing window lengths is called the smoothed-energy-difference (SED) waveform. Since the energy in regular phonation is smoothly varying, few peaks are expected in its SED waveform (Figure 1). On the other hand, the SED waveform should show a more jagged structure for irregular phonation (Figure 2). Inadvertent peaks might be produced at the beginning and end of the SED waveform due to filtering artifacts from the different window lengths. In order to avoid these artifacts, max(window size)/2 + 1 samples from the beginning and end of the waveform are excluded from analysis. The SED cue is set to the largest peak in the SED waveform.

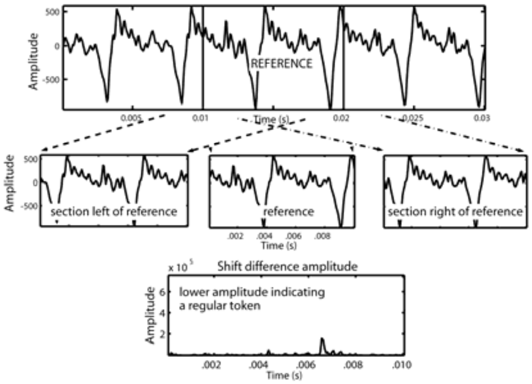

Figure 3. *Illustration of shift-difference amplitude calculation for a regular token.*

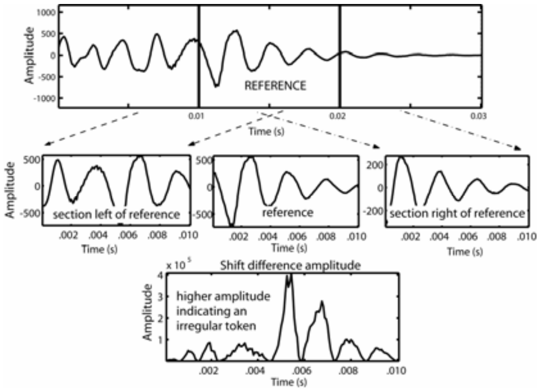

Figure 4. *Illustration of shift-difference amplitude calculation for an irregular token.*

4.4. Shift-difference amplitude

The shift-difference amplitude cue is largely based on work by Kochanski *et. al.* [9] with minor modifications. It is a measure of aperiodicity and the authors used it to detect prominence in speech. This aperiodicity measure uses 10 ms from the middle of the token as a reference. After windowing by a Gaussian with 20 ms standard deviation, sections of the token are compared within 2 ms to 10 ms from the time of the reference, in increments of the sampling rate. The value of the cue is equal to the minimum difference between shifted sections after normalizing by the reference section. The cue should result in a minimum for periodic tokens.

For each possible shift, between 2 ms and 10 ms to the left and right, $d_\delta[n] = (s[n+\delta/2] - s[n-\delta/2])^2$ is computed where $s[n]$ is the reference at time n. The reference is multiplied by itself to give $P[n] = s[n]^2$, a measure of the power in the reference. Both $d_\delta[n]$ and P[n] are convolved with 20 ms standard deviation Gaussians to yield $\tilde{d}_\delta[n]$ and $\tilde{P}[n]$. $\hat{d}[n] = \min_\delta\{\tilde{d}_\delta[n]\}$ is the minimum difference over all the shifts δ. In order to normalize the output, the shift-difference amplitude cue is $(\hat{d}[n]/\tilde{P}[n])^{0.5}$.

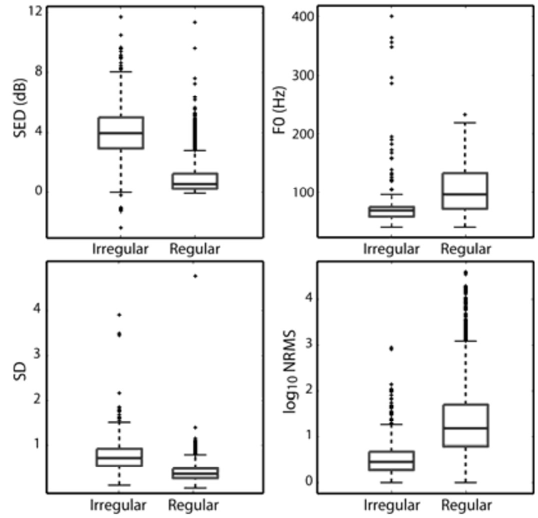

Figure 5. *Illustration of distributions of cue values for the four acoustic cues in this study.*

5. RESULTS

5.1. Cue separation

For each cue, two-sample t-tests between irregular and regular tokens yielded a p-value << 0.001 (df 9332). Box plots of the distributions are given in Figure 5. The middle line of the box is the median, and the upper and lower lines are the upper and lower quartile values. While the cues are statistically separable, the ability of the set of cues to correctly classify tokens was evaluated using support vectors machines (SVMs).

5.2. Token classification

SVMs are learning machines for pattern classification and regression tasks based on statistical learning theory [10]. Given a set of training vectors $\{x_i\}_{i=1}^{l}$, and the corresponding class labels $\{y_i\}_{i=1}^{l}$ such that

$$y_i \in \{-1,+1\},\ \ x_i \in \Re^n$$

SVMs select a set of support vectors $\{x_i^{SV}\}_{i=1}^{SV}$ that is a subset of the training set $\{x_i\}_{i=1}^{l}$ and find an optimal decision function $f(x) = sign(\sum_{i=1}^{N_{SV}} y_i \alpha_i K(x_i^{SV}, x) - b)$ where K is an *a priori* chosen kernel function. The weights α_i, the set of support vectors $\{x_i^{SV}\}_{i=1}^{N_{SV}}$ and the bias term b are found from the training data using quadratic optimization methods. A Gaussian kernel is used to classify regular and irregular phonation. For the Gaussian kernel, $K(x_i, x) = \exp(-\gamma \mid x_i - x \mid^2)$. The experiment was carried out using the OSU SVMs Toolbox (http://www.ece.osu.edu/˜maj/osu svm/). The trade off between the false negative and false positive rates was evaluated for every possible threshold. The number of irregular tokens in the training set is 959. The number of regular tokens used for training was increased from 959→1500→2500→3500. Classification of irregular tokens improved as the number of regular training tokens increased, but improvement decreased after 2500 samples. The classification rate of regular tokens does not change as the number of regular training tokens increased from 2500 tokens to 3500 tokens. Therefore, 2500 regular tokens and 959 irregular tokens were used for training the SVM. The test set consists of 4320 regular tokens and 320 irregular tokens. The unequal size of the test set should not affect the performance of the SVM. It is merely an artifact of having an unequal number of regular and irregular tokens and is somewhat representative of the occurrences of regular and irregular phonation in normal speech. Using a threshold of 0, a recognition rate of 91.25% is obtained for irregular phonation with a false alarm rate of 4.98%, while regular phonation is classified with a recognition rate of 95.02% and a false alarm rate of 9.75%.

6. Discussion

The proposed set of cues provides good separation of the regular and irregular tokens (as evidenced by the cue distributions and the greater than 90% classification rates), but is not completely successful. In this section, we discuss the cases where tokens were mis-classified and present possible strategies for accounting for these cases in future work.

For irregular tokens, the F0 cue is unexpectedly high for tokens with widely-spaced glottal pulses and evidence of relatively strong oscillations between pulses, where the periodicity of these oscillations may be detected as F0. Regular tokens show an unexpectedly low F0 in cases where the amplitude of the waveform decreases across the token, leading to peaks in the autocorrelation function which do not exceed the given threshold.

Similarly, widely-spaced pulses with strong oscillations between pulses in irregular tokens can lead to a higher than normal NRMS value, and low or decreasing signal amplitude across a regular token can lead to unexpectedly low NRMS.

The SED and SD cues tend to fall outside of expected ranges for irregular tokens when the token contains only one or two glottal pulses. The SED cue for regular tokens tends to show inappropriately high values for some tokens from male speakers where the F0 is low enough such that the shorter window (6 ms) may not always capture at least one glottal pulse, leading to peaks in the SED waveform. The SD cue for regular tokens tends to show unexpectedly high values for regular tokens when the signal amplitude changes across the token.

Overall, two recurring reasons for cues outside of expected ranges are: irregular tokens with widely-spaced glottal pulses showing strong oscillations between pulses or with only one or two glottal pulses, and (2) regular tokens in which the signal amplitude changes across the token or with a low F0. While, in general, in each of these cases, the values of the other cues are sufficient to offset the unexpectedly out-of-range cue, there is room to improve these cues with possible reduction of between-pulse oscillations or detection of envelope changes.

7. Conclusions

A set of four acoustic cues -- F0, normalized RMS amplitude, smoothed-energy-difference amplitude and shift-difference amplitude – have been proposed for separation of regular and irregular tokens of phonation. In general, cue distributions are widely separated statistically and classify tokens with accuracy rates greater than 90%. The results support the stated aim of classification of tokens from a relatively large set of speakers—114 different speakers for training and 37 different speakers for testing – and confirm the ability of the cues to separate tokens in spite of the high inter-speaker variation of irregular phonation. In addition, both male and female speakers are well represented in the data set and the regular and irregular tokens used for training and testing occur in various contexts (i.e. utterance-initial, phrase-final, utterance-final etc.).

8. Acknowledgements

This work was supported by NIH/NIDCD grant # DC02978.

9. References

[1] Kreiman, J. (1982) “Perception of sentence and paragraph boundaries in natural conversation,” *Journal of Phonetics*, 10, 163-175.

[2] Huber, D. (1992) “Perception of aperiodic speech signals,” In: *Proceedings of the 2nd International Conference on Spoken Language Processing*, Alberta: Banff, 503-505.

[3] Pierrehumbert, J. (1995) “Prosodic effects on glottal allophones,” In: *Vocal fold physiology: voice quality control*, (O. Fujimura and M. Hirana, eds.) San Diego: Singular Publishing Group, 39-60.

[4] Gordon, A. and Ladefoged, P. (2001) “Phonation types: a cross-linguistic overview,” *Journal of Phonetics*, 29, 383-406.

[5] Dibazar, A.A., Narayanan, S. and Berger, M. (2002) “Feature analysis for automatic detection of pathological speech,” In: *Proc. Engineering, Medicine, and Biology Symposium*, 1, 182-183.

[6] Kiessling, A., Kompe, R., Niemann, H., Nöth, E., and Batliner, A. (1995) “Voice source state as a source of information in speech recognition: Detection of laryngealizations,” In: *Speech Recognition and Coding – New Advances and Trends*, (Rubio-Ayuso and Lopez-Soler, eds.), New York: Springer Verlag, 329-332.

[7] Ishi, T.C. (2004) “Analysis of autocorrelation-based parameters for creaky voice detection,” in *ISCA Archive*, presented at Speech Prosody 2004, Nara, Japan, 23-26.

[8] K. Surana, (2006) “Classification of vocal fold vibration as regular or irregular in normal, voiced speech,” M. Eng. Thesis, Massachussetts Institute of Technology, Cambridge, MA.

[9] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, (2005) “Loudness predicts prominence: Fundamental frequency lends little,” *Journal of Acoustical Society of America*, 118(2), 1038–1054.

[10] V. Vapnik, (1995) *The nature of statistical learning theory*, New York: Springer Verlag.