Automatic Classification of Regular vs. Irregular Phonation Types

Tamás Bőhm, Zoltán Both, Géza Németh

Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics,
Magyar Tudósok krt. 2., 1117 Budapest, Hungary
{bohm, bothzoli, nemeth}@tmit.bme.hu

Abstract. Irregular phonation (also called creaky voice, glottalization and laryngealization) may have various communicative functions in speech. Thus the automatic classification of phonation type into regular and irregular can have a number of applications in speech technology. In this paper, we propose such a classifier that extracts six acoustic cues from vowels and then labels them as regular or irregular by means of a support vector machine. We integrated cues from earlier phonation type classifiers and improved their performance in five out of the six cases. The classifier with the improved cue set produced a 98.85% hit rate and a 3.47% false alarm rate on a subset of the TIMIT corpus.

Key words: Irregular phonation, creaky voice, glottalization, laryngealization, phonation type, voice quality, support vector machine

1 Introduction

Voiced speech can be characterized by the regular vibration of the vocal folds, resulting in a quasi-periodic speech waveform (i.e. the length, the amplitude and the shape of adjacent periods show only slight differences). However, sometimes the vocal folds vibrate irregularly and thus successive cycles exhibit abrupt, substantial changes in their length, amplitude or shape (Fig. 1). In this paper, the term irregular phonation is used to describe regions of speech that display "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" [1]. This latter case refers to periodic vibration well below the speaker's normal fundamental frequency range. Phonation types, in which the irregularity arises from additive noise (e.g. breathiness) are not considered in this study.

It is likely that irregular phonation plays a role in various aspects of speech communication. It has been shown to be a cue to segmental contrasts and to prosodic structure in several languages [2]. Further, its occurrence seems to be characteristic to certain speakers [3] and to certain emotional states [4]. An automatic method for

classifying phonation into regular and irregular can help the analysis of these communicative roles and can also be useful in improving speech technologies (e.g. speech recognition, emotional state classification and speaker identification).

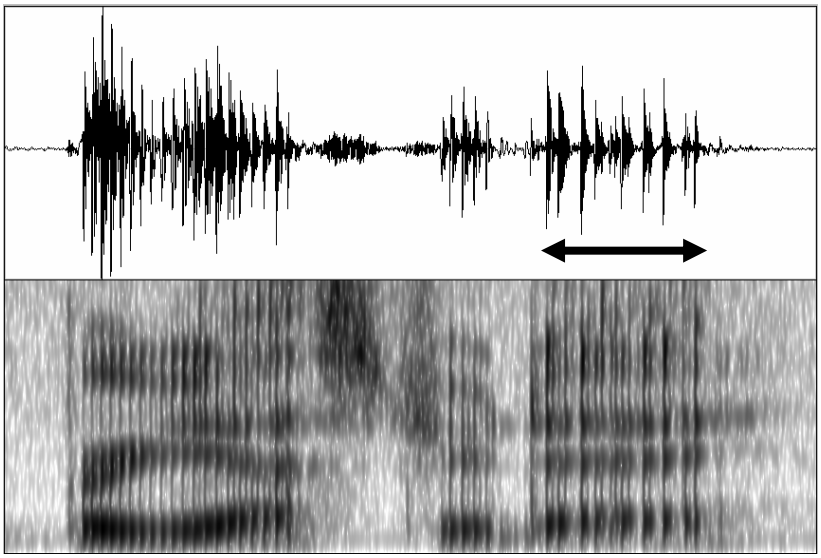

Fig. 1. Waveform (top) and spectrogram (bottom) of a speech signal exhibiting irregular phonation (denoted by the arrow).

A number of phonation type classifiers employing a wide spectrum of acoustic cues and decision algorithms are described in the literature. The method of Surana and Slifka [1,5] calculates four acoustic cues (e.g. fundamental frequency and normalized amplitude) and then applies a support vector machine (SVM) classification scheme to decide whether a phone was produced with regular or irregular phonation. Ishi et al. [6] proposed three cues based on the peaks of the very-short-term power of the speech signal (e.g. cues related to the rate of power change around the peaks and the periodicity between the peaks). The final decision is based on thresholds for the values of the three cues. Vishnubhotla and Espy-Wilson [7] employed cues derived from the AMDF (average magnitude difference function) dip profile, as well as zero-crossing rate, spectral slope and pitch detection confidence (autocorrelation peak value). Yoon et al. [8] also used the peak value of the autocorrelation function and measured open quotient, too (by means of the amplitude difference between the first and second harmonics). Kiessling et al. [9] used five acoustic cues extracted from the cepstrally smoothed spectrum and classified phonation by means of a Gaussian phone component recognizer. The other approach presented in the same paper employs an artificial neural network (ANN) to inverse filter the speech signal, and then another ANN to classify the source signals into regular, irregular and unvoiced.

In this paper, we propose a phonation type classifier that can categorize vowels either as regular or as irregular with high accuracy. Our method integrates cues from two earlier systems: all the four cues from [1] (fundamental frequency, normalized

RMS amplitude, smoothed-energy-difference amplitude and shift-difference amplitude) and two cues from [6] (power peak rising and falling degrees, and intraframe periodicity). These six cues were reimplemented based on their description in the literature and most of them were improved by algorithmic refinements and by exploring their parameter space. The performance of each cue was measured by the increase in the area under the receiver operating characteristic (ROC) curve [10]. These cues are inputted into a support vector machine in order to classify the vowel as regular or irregular.

2 Speech Data Set

This work was carried out using the subset of the TIMIT corpus in which occurrences of irregular phonation in vowels were hand-labelled by Surana and Slifka [1,5]. This subset included recordings of 151 speakers (both males and females) uttering 10 sentences each (114 speakers in the train set and another 37 speakers in the test set). Among the vowels labeled, they found 1751 produced with irregular phonation and 10876 with regular phonation.

3 Acoustic Cues

For the extraction of five of the acoustic cues, the input speech signal is spliced into 30 ms frames with a 5 ms step. These five cues are expressed as a summary statistic (e.g. mean or minimum) over the values calculated for each frame in the input. The remaining one cue (power peaks) processes the entire input vowel in one run. For each cue, we first present its original calculation method (as described in the corresponding paper) and then we explain our improvements (if there was any).

3.1 Fundamental Frequency (F0)

One can expect that for irregularly phonated speech, a pitch detector extracts an F_0 value that is lower than that for regular voiced speech, or detects it as unvoiced (denoted by F_0=0 Hz). Surana and Slifka [1,5] used an autocorrelation-based pitch detector to calculate this acoustic cue. Before computing the autocorrelation function, the speech signal is Hamming-windowed and inverse filtered (by means of a 12th order LPC filter) and the residual signal is low-pass filtered with a 1 kHz cutoff. The peaks of the normalized autocorrelation function of the low-passed residual are used to estimate F_0:

- If there are no autocorrelation peaks higher than the voicing threshold (0.46) in the lag interval corresponding to 70-400 Hz, then the frame is considered unvoiced and F_0 is set to 0.
- If there is only one peak fulfilling the above criterion, then the reciprocal of the peak's lag is returned as the F_0.

- If there are more than one such peaks and their lags are integer multiples of each other, then the second one is chosen as the one corresponding to F_0 ("second peak rule").
- If there is no such regularity among the multiple peaks, then the highest peak is chosen.

For a given vowel, the F0 cue value is the minimum of the F_0's of the frames comprising that vowel.

By some changes in the algorithm, we could increase the area under the ROC curve of the cue from 0.87 to 0.93 (Fig. 2.a). These changes included the removal of the second peak rule as it usually resulted in halving errors, the use of an unbiased autocorrelation function (whose envelope does not decrease to zero at large lags [11]). Further, the voicing threshold was changed to 0.35 based on a systematic evaluation of a number of values in the range of 0.25-0.5.

3.2 Normalized RMS Amplitude (NRMS)

In irregularly phonated speech, there are generally fewer glottal pulses in a given time frame than in regularly phonated speech. This can be captured by the RMS amplitude (intensity), normalized by the RMS amplitude of a longer interval. Surana and Slifka [1] divided the RMS intensity of each frame by the intensity calculated over the entire sentence and then took the mean of these values in order to compute the NRMS cue for the vowel.

If the intensity level changes along the sentence, then normalizing with the sentence RMS can become misleading. Thus instead of the entire sentence, we used the vowel and its local environment for normalization. In the range examined (25-500 ms, with 25 ms steps), an environment of 50 ms (25 ms before and 25 ms after) led to a small increase in the area under the ROC curve (0.84 to 0.86; Fig. 2.b).

3.3 Smoothed-Energy Difference Amplitude (SED)

This cue attempts to characterize the rapid energy transitions in irregular phonation that are due to the wider spacing of the glottal pulses. The energy curve (calculated as the mean magnitude between 300 and 1500 Hz of the FFT in 16 ms windows, stepped by 1 ms) is smoothed with a 16 ms and separately with a 6 ms window. The difference of the two smoothed energy functions is usually near zero for regular phonation, while in case of irregular phonation, it has several peaks. In the original description, the SED cue is the maximum of the difference function [1]. (The first and last 8 ms of the difference function is not taken into account because artifacts due to the different window sizes can appear in these regions.)

Instead of taking the maximum, we calculate the absolute maximum. Further, our tests examining the effect of various window size combinations revealed that using a 2 ms and a 4 ms smoothing window can improve the separation of regular and irregular tokens. After these changes, the area under the ROC curve increased by 0.08 (from 0.74 to 0.82; Fig. 2.c).

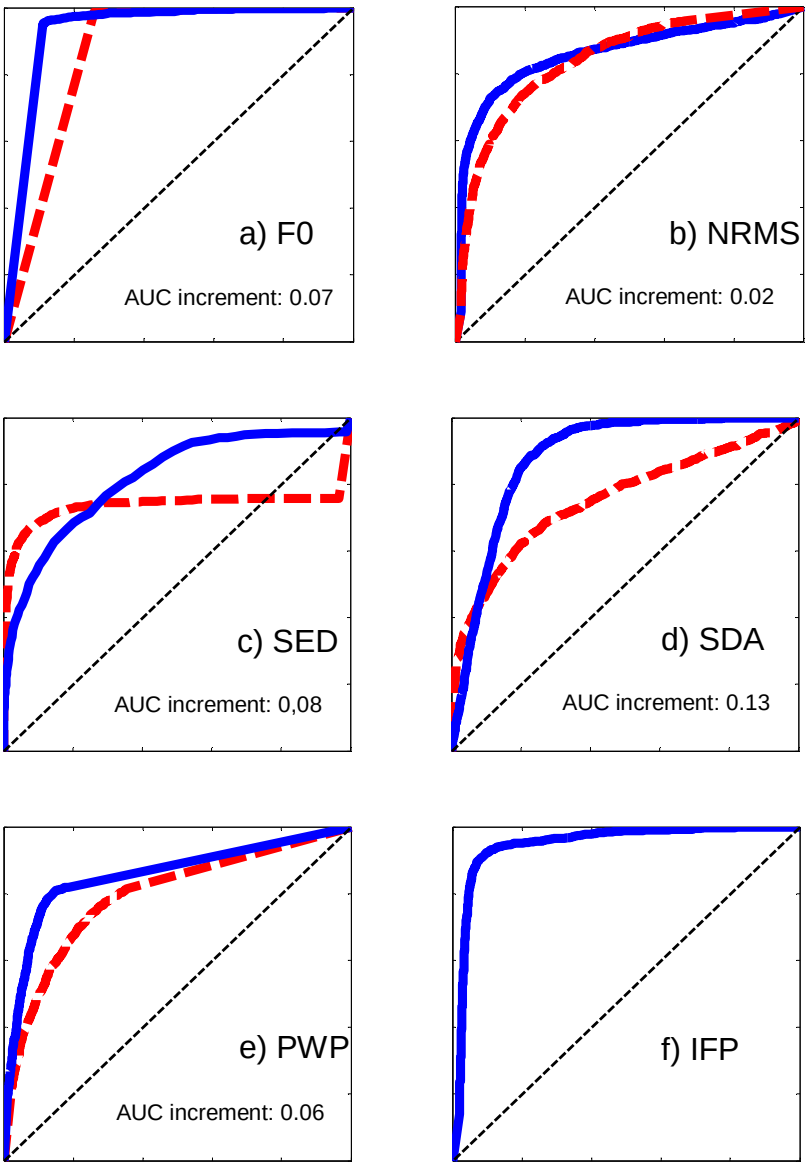

Fig. 2. Receiver operating characteristic (ROC) curves of the six acoustic cues: a) fundamental frequency, b) normalized RMS amplitude, c) smoothed-energy difference amplitude, d) shift-difference amplitude, e) power peak rising and falling degree, f) intraframe periodicity. The red dashed lines correspond to the original versions of the cues, while the blue continuous lines represent the improved versions (there were no improvements for the IFP cue). Both axes range from 0 to 1 on all panes. The increment in the area under the curves is also shown.

3.4 Shift-Difference Amplitude (SDA)

This cue estimates the irregularity of the glottal pulses. After preprocessing, two 10 ms windows are employed: one of them is initially shifted 2 ms to the left, while the other one is shifted 2 ms to the right of the center of the 30 ms frame [1]. In each step, both of the windows are shifted one sample away from the frame center and the squared difference signal of the two windowed waveforms is calculated. Then, the minimum of the difference signals is computed in each time point over all the window shifts. The resulting 10 ms signal is normalized by the square of the middle 10 ms of the frame and then averaged over time. The SDA cue for the vowel is the average of the SDA values obtained for each frame.

The shift-difference amplitude cue in [1] is based on Kochanski's "aperiodicity measure" [12]. According to our tests, applying the aperiodicity measure (with slight changes) results in higher separation between regular and irregular vowels than the SDA cue in [1] (the area under the ROC curve is 0.88, compared to 0.75 with the method described in [1]; Fig 2.d).

To calculate the "aperiodicity measure", we add 5 ms of silence to the beginning and to the end of the 30 ms frame. Then, as described in [12], we define two 30 ms windows: one at the beginning and the other one at the end of the zero-padded signal. As we shift both windows sample-by-sample towards the center, we calculate their squared difference function in each step. After down-sampling each difference function to 125 Hz (so that there are only 5 values remaining for each 40 ms long signal), the minimum value is taken at each time point. After dropping the first and the last value, these minimums are averaged to obtain the aperiodicity measure for the frame.

3.5 Power Peak Rising and Falling Degrees (PWP)

Like the smoothed-energy difference amplitude, this cue also attempts to capture the rapid energy transitions in irregular phonation. As described by Ishi et al. [6], the "very-short-term power contour" is calculated, with a window size of 4 ms and a step size of 2 ms (note that this cue is not calculated using the 30 ms framing employed by all the other cues). In this function, irregular glottal pulses usually appear as peaks with a long, steep rise and fall. Thus first the peaks are detected and then a measure of the rate of power rise and fall (before and after the peak) is computed. A point in the very-short-term power contour is considered to be a power peak, if it has a value of at least 2 dB higher than both the point 3 samples before and the point 3 samples later. The degree of power rising before the peak is obtained as the maximum power difference between the peak and the 5 preceding values. The power falling was estimated similarly, with the 5 following values. When adapted to our phonation type classification framework, the PWP cue is calculated as the average of the maximum power rising and the maximum power falling value in the vowel.

According to a grid search we performed, it is more advantageous to find the peaks based on the 4th value (instead of the 3rd) before and after the point in question. Further, the power rising and falling features separate the two phonation types better,

if they are computed as the maximum power difference in a ±4 sample environment of the peak. The ROC curve of the original PWP cue has an area of 0.79, while the area under the curve corresponding to the improved version is 0.85 (Fig 2.e).

3.6 Intraframe Periodicity (IFP)

As the shift-difference amplitude, IFP also tries to capture the repeatable waveform structure of regular phonation and the reduction in this repeatability in irregular phonation. For each frame, the unbiased autocorrelation function is calculated and the first prominent peak with a positive lag is found [6]. Then the autocorrelation values at integer multiples of this peak lag are obtained and their minimum is selected as the measure of intraframe periodicity. For white noise, this value is near zero, while for a perfectly periodic signal, it is one. The IFP cue of the vowel is the maximum of IFP values of the individual frames. This cue provides excellent separation between regular and irregular vowels (the area under ROC curve is 0.95; Fig 2.f) and thus we did not attempt to improve it.

4 Classification

Using the six described cues (their improved versions, where available) as features, the classification was carried out by a support vector machine (SVM) with a radial basis function (RBF) kernel [13]. The SVM was implemented using the publicly available OSU SVM toolbox (http://sourceforge.net/projects/svm).

Before training, two parameters of the SVM needed to be set: C, the cost of incorrect classifications, and γ, a property of the Gaussian kernel. We ran a grid search on a wide range of values for the two parameters. A subset of the training set, containing 1380 regular and 1380 irregular vowels, was used for the grid search. For each parameter combination, we performed both a 3-fold and a 10-fold cross-validation (e.g. for the 3-fold cross-validation, we trained the SVM three times, always leaving out one third of the set that we subsequently used for testing the performance of the classifier) and calculated the average of the hit rates and false alarm rates. The two averages were then combined by an equal weight to obtain the accuracies for each parameter setting. The results are shown on Fig. 3, with the highest accuracy achieved with C=0.0313 and γ=0.0313.

Using the above parameter values, the phonation type classifier was trained with an equal number of regular and irregular tokens. This training set contained all the 1403 irregular vowels and 1403 regular vowels randomly selected from the 8196 available.

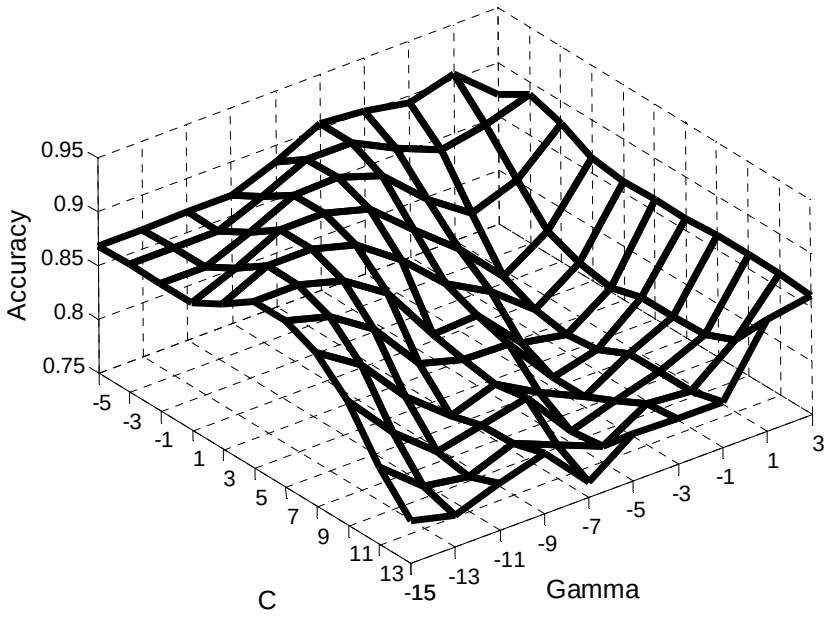

Fig. 3. Results of the grid search over an array of SVM training parameter values. For each *C* and γ, the average accuracy calculated from the 3-fold and 10-fold crossvalidations are shown. The two parameters are expressed as powers of 2.

5 Evaluation

The subset of the TIMIT test set that was annotated by phonation type was used to evaluate our classifier. It contained 348 irregular and 2680 regular tokens.

The proposed system achieved a 98.85% hit rate (correct recognition of irregular phonation) and a 3.47% false alarm rate (incorrectly classifying regular phonation as irregular). These results can be compared to the results of Surana and Slifka [1], as they used the same training and test sets, and their phonation type classifier was designed with the same assumptions (the input is a vowel and the output is a binary regular/irregular decision). They reported a hit rate of 91.25% and a false alarm rate of 4.98%.

6 Conclusions

We proposed a phonation type classifier that aims to distinguish irregularly phonated vowels from regularly phonated ones. Our system achieved a 7.60% higher hit rate and a 1.51% lower false alarm rate than a comparable previously published system.

The improvements are due to using a wider range of acoustic cues (integrating cues employed earlier in different systems) and to refining these cues either in terms of the calculation algorithm itself or in terms of the parameters of the algorithm.

The high hit rate (98.85%) and reasonably low false alarm rate (3.47%) are likely to be sufficient for most applications. We have to note however that this classifier has only been tested on vowels. In many cases, this may not limit practical application (e.g. for contributing to the automatic recognition of the emotional state or the identity of the speaker). But in other cases, it would be more useful to have a classifier capable of working on both vowels and consonants, and without phone-level segmentation. Future work should address these issues.

References

1. Surana, S., Slifka, J.: Acoustic cues for the classification of regular and irregular phonation. In: Interspeech 2006, 693–696 (2006)
2. Slifka, J.: Irregular phonation and its preferred role as cue to silence in phonological systems. XVIth International Congress of Phonetic Sciences, 229–232 (2007)
3. Henton, C.G., Bladon, A.: Creak as a sociophonetic marker. In: Hyman, L.M., Li, C.N. (eds.) Language, speech and mind: Studies in honour of Victoria A. Fromkin, pp. 3–29, Routledge, London (1987)
4. Gobl, C., Ní Chasaide, A.: The role of voice quality in communicating emotion, mood and attitude. Speech Communication 40, 189–212 (2003)
5. Surana, K.: Classification of vocal fold vibration as regular or irregular in normal voiced speech. MEng thesis, MIT (2006)
6. Ishi, C.T., Sakakibara, K.-I., Ishiguro, H., Hagita, N.: A method for automatic detection of vocal fry. IEEE Tr. on Audio, Speech and Language Proc. 16(1), 47–56 (2008)
7. Vishnubhotla, S., Espy-Wilson, C.: Detection of irregular phonation in speech. XVIth International Congress of Phonetic Sciences, 2053–2056 (2007)
8. Yoon, T.-J., Zhuang, X., Cole, J., Hasegawa-Johnson, M.: Voice quality dependent speech recognition. International Symposium on Linguistic Patterns in Spontaneous Speech (2006)
9. Kiessling, A., Kompe, R., Niemann, H., Nöth, E., Batliner, A.: Voice source state as a source of information in speech recognition: Detection of laryngealizations. In: Rubio-Ayuso, Lopez-Soler (eds.) Speech Recognition and Coding: New advances and trends, pp. 329–332, Springer, Heidelberg (1995)
10. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters 27, 86–874 (2006)
11. Boersma, P.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. IPS Proceedings 17, 97–100 (1993)
12. Kochanski, G., Grabe, E., Coleman, J., Rosner, B.: Loudness predicts prominence: Fundamental frequency lends little. JASA 118(2), 1038–1054 (2005)
13. Bennett, K.P., Campbell, C.: Support vector machines: Hype or hallelujah? SIGKDD Explorations 2(2), 1–13 (2000)