Automatic Detection of Irregular Phonation in Continuous Speech

Srikanth Vishnubhotla and Carol Espy-Wilson

Institute for Systems Research and Dept. of Electrical & Computer Engineering,
University of Maryland, College Park, MD, USA 20742
{srikanth, espy}@glue.umd.edu

10.21437/Interspeech.2006-178

Abstract

Voice quality is one of the most important source characteristics of a speaker's speech production process, and creakiness is one of the variations of voice quality. This paper describes the development of an algorithm to automatically detect irregular phonation, including creakiness and other variations, in continuous running speech. The algorithm is an extension of the Aperiodicity, Periodicity and Pitch (APP) Detector. The algorithm has been run on 485 files of the TIMIT database, which contained 677 instances of irregular phonation. The test set comprised of 97 speakers, of which 57 were male and 40 were female. The algorithm has been found to give an accuracy of 86.7 % on average, with performance being almost the same for both male and female speakers. Automatic detection of irregular phonation should help characterize speakers for speaker identification applications.

Index Terms: irregular phonation, creakiness, voice quality, speaker recognition, speaker characterization

1. Introduction

Speakers have a certain characteristic quality to their voice, which is a consequence of their style of phonation, and thus, their individual source properties. This kind of voice quality is perceived by listeners as the breathiness, creakiness, harshness etc., of the voice. While some speakers exhibit a particular voice quality throughout their speech, some others show either a continuous or an abrupt change in their voice quality. The variety in voice quality is a well-studied phenomenon, and researchers from various disciplines like speech processing, voice pathology, phonetics, linguistics and music have examined the various aspects of phonation. [1,2,3,4].

In this study, we focus on one of the variations of voice quality, namely irregular phonation. We define this particular category to comprise of sounds that various researchers from different disciplines have called creak, vocal fry, diplophonia, diplophonic double pulsing [1], glottalization [2], laryngealization [3], pulse register phonation [1], and glottal squeak [2]. We rely on the work in [1] that has shown that most of these phenomena are all perceptually similar to each other, and can thus be classified together.

The most common example of irregular phonation is the creak. A perceptual definition for creak is [4]: "the acoustic result of a creak is a series of irregularly spaced vocal pulses that give the auditory impression of a rapid series of taps, like a stick being run along a railing". Figure 1 is a spectrogram showing three instances of creakiness in an utterance, at time intervals as indicated in the caption. The rather high, irregular spacing between consecutive vocal pulses – a characteristic of creak – is seen as vertical striations at those locations. We define our task as the "automatic detection of all sounds that fall in this *perceptual category* of irregular phonation" and will use the terms creak and irregular phonation interchangeably here.

The task is to automatically detect any of these varieties of irregular phonation. This is especially important in present-day speaker identification systems, since speaker identification relies on the reliable extraction of both the source and vocal tract features [5]. The work presented here is particularly useful in characterizing the voice quality (or equivalently, the source information) of a speaker, by distinguishing it from the modal or breathy kinds of speech. In addition to speaker identification, this work can contribute to the task of language identification, as it can aid to identify a variety of languages that exploit creakiness and breathiness to articulate certain sounds [4]. Automatic identification of irregular phonation can also aid in the study of phonetics and voice pathology.

We have developed an algorithm that can process an input speech file and automatically identify regions of the signal where the speaker exhibits irregular phonation. The algorithm is an extension of the Aperiodicity, Periodicity and Pitch (APP) Detector [6], a system that processes a speech file on a frame-by-frame basis to give a spectro-temporal profile indicating the amount of aperiodicity and periodicity in different frequencies with time. Currently, the APP Detector does not distinguish between aperiodicity due to turbulence from that due to irregular phonation. We modified the APP Detector to separate these different classes of aperiodicity.

This paper gives a description of the algorithm and its performance. Section 2 of this paper discusses the APP Detector in brief, to lay the background for the algorithm. Section 3 discusses the algorithm, and the features of irregular phonation

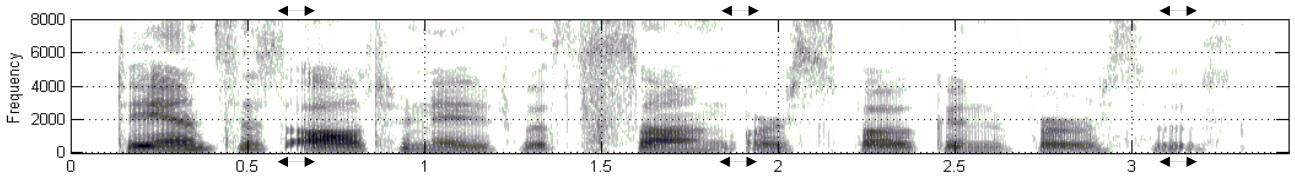

Figure 1: *Spectrogram of a sample speech file containing creakiness in three different location (marked by arrows): t=0.6 to 0.7 sec, t=1.8 to 1.9 sec and t=3.1 to 3.2 sec. The x-axis shows time in seconds.*

that have been considered for design. Section 4 describes the performance of the algorithm, and section 5 concludes the paper, outlining the future work briefly.

2. The Aperiodicity, Periodicity and Pitch (APP) Detector

The APP Detector [6] estimates the proportion of periodic and aperiodic energy in a speech signal, and the pitch period of the periodic component, on a frame-by-frame basis. In brief, the first block in the APP Detector is an auditory gamma-tone filter-bank that splits the channels into 60 frequency bands. The outputs of the higher frequency channels are then smoothed using the Hilbert transform to extract the envelope information and remove the finer structure. The next stage of the APP Detector incorporates silence detection by thresholding, followed by the use of the Average Magnitude Difference Function (AMDF) to identify the amount of periodicity or aperiodicity in the signal.

For each frame, a windowed portion of the signal, centered at the frame center, is used to compute the AMDF. The AMDF $\gamma_n[k]$ of a signal $x[n]$ is defined as

$$\gamma_n[k] = \sum_{m=-\infty}^{\infty} \left| x[n+m]w[m] - x[n+m-k]w[m-k] \right|$$

where $w[n]$ represents a rectangular window centered at n and having a width as specified. When the speech signal is periodic, this function will contain dips at values of k equal to a multiple of the pitch period. Figure 2 contrasts a sample AMDF function from a single channel of a strongly periodic signal with that of a strongly aperiodic signal. The vertical lines superimposed on the figure represent the strength of the dips. The noteworthy features are that (i) the AMDF shows strong dips at lags equivalent to the pitch period and its multiples in the case of a strongly periodic signal, (ii) the dips are weaker and very randomly distributed in the strongly aperiodic case.

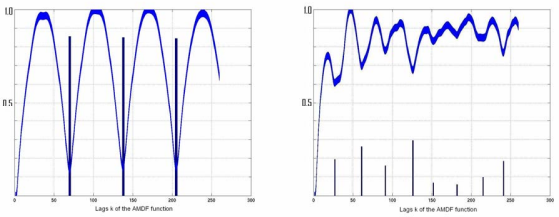

Figure 2: *AMDF and Dip Behavior for a periodic signal (left) and aperiodic signal (right)*

Decision of periodicity and aperiodicity of the frame is made by summarizing the trend across all channels. For a periodic frame, it is expected that all non-silent channels will exhibit a similar trend in the AMDF dips. Thus, when the dips are summed across all channels, the dips will cluster tightly together at lags equaling the pitch period and its integer multiples, and give a significant strength of dips. If this is the case for a particular frame, then that frame is classified as being periodic. For an aperiodic frame, the dips, when summarized, will display a random behavior as a consequence of the individual channel behavior. The summary measure of periodic and aperiodic content is obtained by multiplying the frame per/aper decision by its energy and then adding it across channels. Thus, for each frame, a decision of per/aper is made for the frame, its individual channel-wise per/aper profile is produced, and the amount of aperiodic and periodic energy is also obtained.

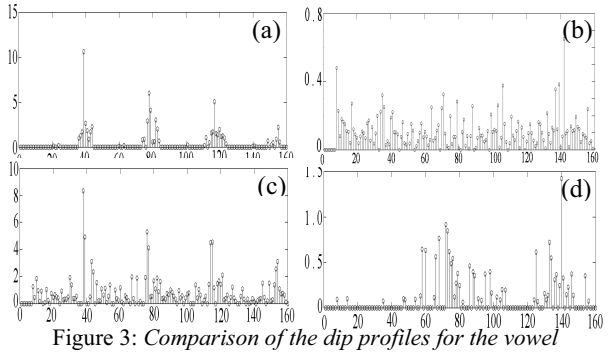

Figure 3: *Comparison of the dip profiles for the vowel frame (a), fricative frame (b), voiced fricative frame (c) and creaky voicing frame (d).The x-axis represents lags, and the y-axis represents strength of dips*

Figure 3 gives an illustration of the dip summary (or profile) for the cases of a single frame of a vowel, an unvoiced fricative a voiced fricative and a creaky vowel. The vowel, being a periodic signal, shows strong dip clusters at multiples of pitch period. The unvoiced fricative displays a random distribution of small dips. In the case of the voiced fricative, the dips show a mixed behavior – there is clustering of some dips, which is due to the voiced (glottal) source, and there is also some randomness in distribution of some other dips, which is due to the turbulence of the supra-glottal source. The fourth dip profile belongs to a creaky frame. Because of the irregular phonation which causes the pitch to change over consecutive frames, the AMDF does not show a regular behavior. This is due to misalignment between the windowed signal and its delayed version, which causes the dips to scatter from their cluster. The clusters are thus not as well defined as in the periodic vowel, and yet not as randomly distributed as will be due to a turbulent source. It is this kind of dip profile that we target to identify from others. It is worthy of interest that the number of clusters in the frame have halved from the periodic case, as is expected due to the fall in pitch occurring in a creak.

A noteworthy fact is that the APP Detector is indeed capturing the source information from the signal. An illustrative example is the comparison of the APP Detector output for the case of a vowel utterance by the same speaker in a modal and a creaky voice. Figure 4 shows the spectro-temporal per/aper profile – a visual display of periodic and aperiodic channels for consecutive frames for these two cases, as obtained by the APP Detector. It is seen that the creaky vowel displays aperiodic energy during the majority of the vowel. This is due to the fact that creakiness arises due to irregular vocal fold vibration, which is not periodic in nature. Figures 4 (c)-(e) show the time waveform, the corresponding glottal waveform obtained by an inverse filtering technique [7] and the corresponding AMDF obtained by the APP Detector for a frame of the modal and creaky vowels. The most striking point here is that the AMDF closely captures the information in the glottal waveform – this is evident in the relative position of the peak of the glottal waveform and dips of the AMDF in the case of the modal vowel. This correspondence is not so exact for the creaky

vowel, owing to the fact that the pitch shows some jitter and thus, the AMDF does not have exact alignment of the signal and its delayed version to give dips at exactly the pitch period. This is what causes the AMDF to have dips that are weak and do not cluster properly, and are not in alignment with their counterparts from other channels. One may therefore conclude that the APP Detector is indeed capturing the source information and is a very suitable candidate to use to identify irregular phonation.

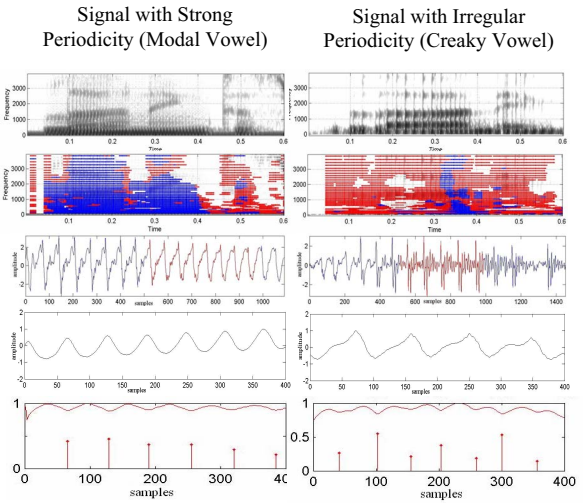

Figure 4: *Modal vowel versus creaky vowel: spectrogram (a), aperiodicity/periodicity spectro-temporal profile (b), section of the time waveform between t=0.1 to 0.2 sec for the modal vowel and t=0.28 to 0.38 for the creaky vowel (c), glottal waveform obtained by inverse filtering (d) and AMDF corresponding to the above glottal waveform, for one of the channels.*

3. Automatic Detection of Irregular Phonation

The decision process of the APP Detector for frames with irregular phonation can be traced to their dip profile. Separating irregular frames from periodic frames is an easy task – the APP Detector shows periodicity for all frames that have modal phonation – thus, such frames can be eliminated from further analysis by conditioning on the periodicity measure.

To distinguish the irregular phonation frames from aperiodic frames, the amount of aperiodicity in the highest frequencies is seen. It is expected that fricatives will have high amount of aperiodicity in the higher frequencies (above 3000 Hz) due to turbulence, while irregular phonation frames do not have such high-frequency aperiodicity. Thus, it is conditioned that the algorithm eliminates frames showing aperiodicity in all channels above 3000 Hz. The next step is to characterize the dip profile that is unique to non-modal phonation. This is done by finding all the local maxima in the dip profile, and eliminating all those maxima that lie too close to each other. This is done in order to ensure that the corresponding pitch estimate remains below 150 Hz, as expected for irregular phonation. The local maxima are then defined as cluster centers, and loose clusters are formed around these centers. A score is made of how many of the channels have their dips lying within these clusters defined by the cluster center, and this is called the channel confidence. The cluster center is then recalculated, using only those channels which have their dips lying within the cluster. Redefining the clusters, the channel confidences are calculated again. The channel confidence is then summarized across frequency, and if at least 50% of the channels show confidence 1, the frame is declared to have the characteristic dip profile.

We observed that by this stage, the majority of the frames detected had irregular phonation, but we were also detecting some breathy vowels and voiced fricatives. This can be explained by the fact that similar to the irregular phonation, the breathy vowels and voiced fricatives have some underlying periodic energy because of voicing. To eliminate the breathy sounds and voiced fricatives, we exploited the fact that the voicing in these two cases will often be in frequencies 0 – 1000 Hz, while irregular phonation is expected to show its characteristic irregularity at all frequencies. Thus, the dip profiles are obtained by summarizing across only those channels that span frequencies above 1000 Hz. We further reduce the possibility of capturing turbulence by conditioning the ratio of number of dips present within the cluster, to the number of non-zero dips, to exceed a threshold of 40%. For frames with irregular phonation, the aperiodicity is not random Therefore the number of non-zero dips lying inside the cluster will be higher compared to the cases of breathiness and voiced fricative, which show random aperiodicity and hence will have a large number of dips outside the cluster.

Our algorithm also showed some confusion with some stops, though this was the case with only certain speakers and not for all stops for that speaker. In order to handle this issue, we made use of the spectral slope of the frame by taking a central portion of each frame and calculating the slope of the spectrum between frequencies 2000 to 4000 Hz by fitting a line using Minimum Mean-Square Error (MMSE) criterion. The spectral slope for stops is expected to be very low due to their flatter spectrum. Thus, a threshold of -0.05 is used to distinguish between the two. However, we also speculate another reason for the detection of stops in our algorithm, and this will be discussed in the results section.

4. Experiment & Results

We have no available standard database that can be used for describing voice quality and testing the performance of the algorithm. We therefore made use of the TIMIT database that has been studied and hand-marked for irregular phonation at MIT, to test our algorithm. Since there is no clearly defined reference that marks phonation, we have called all those locations marked in the reference, as well as those identified by our algorithm but missed in the reference (confirmed by visual inspection by the first author), as the total number of instances of irregular phonation. False alarms have first been investigated by looking at the transcriptions for obvious cases, and by visual inspection of files in cases that are unclear or ambiguous. Quantitatively, both the accuracy and false alarm rate are identified in terms of the percentage of total number of instances that are irregular phonation.

We made use of the "test" subset of the TIMIT database. The number of files processed was 485, and it included 57 male and 40 female speakers from 5 dialect regions (dr1 through dr5). The total actual number of instances of irregular phonation was 677. Table 1 shows our results. A sample output of the Irregular Phonation Detector is shown in figure 5. The input is the speech file shown in figure 1. It may be seen that all three creaky regions have been identified by our algorithm.

Table 1: Performance of the algorithm for detection of irregular phonation

	Total # of instances	# of instances identified	Percentage identified
Male + Female	677	587	86.7 %
Female	312	276	88.5 %
Male	365	311	85.2 %

The percentage of instances identified, 87%, seems an encouraging figure, considering the fact that identification is made during continuous running speech and in the presence of various confusing elements. Both female and male samples are handled equally well by the algorithm, which confirms that irregular phonation possesses acoustic features that do not depend on gender [3]. We investigated the cases of the missed instances, and found that we were missing the irregular phonation in one of two cases. The first is when the location of irregular phonation is at the very beginning of the file – this is because the AMDF can be computed only after a few initial frames. This will be remedied by refining our conditions for boundaries. The next case is when the pitch falls so low that the analysis window used cannot capture even one full cycle – when that happens, the AMDF structure cannot capture the characteristic dip profile. A solution for this is to adaptively change the analysis window size.

The false detection rate, however, seems higher than the error rate. We had a false detection rate of 20%, and these were mostly due to stops. Confusions were more for male speakers than female speakers, and this is attributed to the low pitch of male speakers, which may often cause the dips to scatter from clusters and thus manifest the sound as being similar to irregular phonation. Of all these false triggers, 25% were due to voiced fricative /sh/. About 35% of the false detections were due to stops being called irregular, while the remaining 40% of the false triggers were due to stops wherein the vowel preceding the stop was identified as creaky. Though we have currently included the detections in the latter category as false detections, studies have shown [2,4] that there do exist cases of stops in both American English and other languages, where both voiced and unvoiced stops may be accompanied by irregular phonation. Further, it is possible that the end of vowels preceding such stops may also exhibit irregular phonation. Thus, it may actually be the fact that we are capturing such instances of irregular phonation too, in which case our false triggering rate will actually be considerably low. In addition, we have also observed that we are identifying what we suspect are some glottal squeaks (4 instances). However, due to the lack of a standard reference, we are not able to confirm our speculations at this juncture, and have added this count to false triggers.

5. Conclusion & Future Directions

We have discussed our algorithm for the automatic detection of irregular phonation in continuous speech. The algorithm is seen to show good recognition rate of 87% on continuous running speech. We show a false-triggering rate of 20%, but are also speculative that our false alarm rate is lower than the figures presented. We plan to continue our work by using a standard reference database to verify the performance of our algorithm, and improving the recognition rates by tackling the two problems discussed in the paper. We eventually plan to work towards the development of a creaky voice quality parameter that can be used for speaker recognition applications.

6. Acknowledgements

This research was supported by NSF grant # BCS-0519256. We thank Kushan Surana and Dr. Janet Slifka of MIT for providing us with the hand-marked transcription of occurrence of irregular phonation in the TIMIT database.

7. References

[1] Bruce R. Gerratt and Jody Kreiman (2001). "Toward a taxonomy of nonmodal phonation", Journal of Phonetics, 29, 365-381

[2] Redi, L. & Shattuck-Hufnagel, S. (2001). "Variation in the realization of glottalization in normal speakers", Journal of Phonetics, 29, 407-429.

[3] Klatt, D.H., and Klatt, L.C. (1989). "Analysis, Synthesis and Perception of Voice Quality Variations among female and male talkers", J. Acoustic. Soc. Am. 82 (2), February 1990, 820-857

[4] Matthew Gordon & Peter Ladefoged (2001). "Phonation types: a cross-linguistic overview", Journal of Phonetics, 29, 383-406.

[5] Carol Y. Espy-Wilson, Sandeep Manocha & Srikanth Vishnubhotla (2006), "A New Set of Features for Text-Independent Speaker Identification", ICSLP (Interspeech) 2006, Pittsburgh, PA, USA

[6] Om Deshmukh, Carol Y. Espy-Wilson, Ariel Salomon, and Jawahar Singh (2005). "Use of Temporal Information: Detection of Periodicity, Aperiodicity, and Pitch in Speech", IEEE Transactions on Speech And Audio Processing, 13(5), September 2005, 776-786

[7] Moore, E. and Clements, M. (2004), "Algorithm for Automatic Glottal Waveform Estimation Without the Reliance on Precise Glottal Information", Proc. IEEE-ICASSP, 2004 (Vol. 1), 101-104.

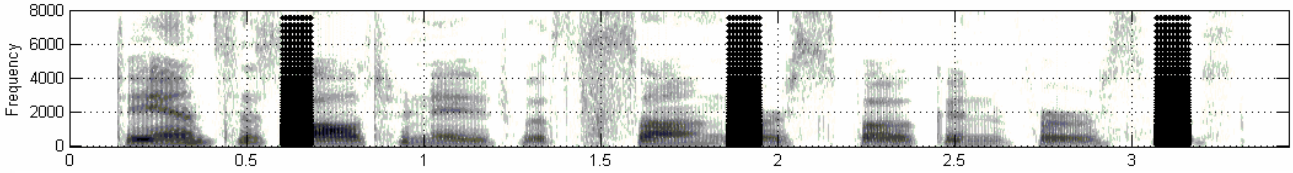

Figure 5: *Output of the Irregular Phonation Detector – the dark areas represent the locations irregular phonation has been detected. The x-axis shows time in seconds.*