DETECTION OF IRREGULAR PHONATION IN SPEECH

Srikanth Vishnubhotla & Carol Y Espy-Wilson

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

ABSTRACT

The problem addressed here is that of detecting irregular phonation during conversational speech. While most published work tackles this problem only by focusing on the voiced regions of speech, we focus on detecting irregular phonation without assuming prior knowledge of voiced regions. In addition, we improve the pitch estimation accuracy of a current pitch tracking algorithm in regions of irregular phonation, where most pitch trackers fail to perform well. The algorithm has been tested on the TIMIT and NIST 98 databases. The detection rate for the TIMIT database is 91.8% (17.42% false detections). The detection rate for the NIST 98 database is 91.5% (12.8% false detections). The pitch detection accuracy increased from 95.4% to 98.3% for the TIMIT database, and from 94.8% to 97.4% for the NIST 98 database.

Keywords: irregular phonation, creakiness, voice quality, speaker characterization, pitch estimation

1. INTRODUCTION

In the speech production process, while the vocal tract determines the articulated phoneme, it is the mechanism of production of the glottal pulses that determines the quality and perceptual attributes of the speech signal [1,2]. Speakers have a certain characteristic quality to their voice, which is a consequence of their style of phonation and source properties. This kind of voice quality is perceived by listeners as breathiness, creakiness etc. In this study, we focus on the class of voicing that exhibits irregular phonation, including creak, vocal fry, diplophonia, diplophonic double pulsing [3], glottalization [4], laryngealization [2], pulse register phonation [3], and glottal squeak [4]. The main characteristic of these various forms of phonation is that the vocal folds do not vibrate as they would for a modal case. The difference can occur due to several possible reasons: asymmetry in physical properties of the vocal folds, voice pathology, behavioral tendencies etc.

The task is to process spontaneous speech and identify regions where the speaker exhibits irregular voicing of any kind, and estimate the pitch period in such regions. Since a common manifestation of irregular phonation is the creak, we use the terms creakiness and irregular phonation interchangeably here. This work has several contributions. While [5] and [6] address only clean speech, this work gives results for telephone speech, which is harder due to channel distortion, noise and non-speech (laughter, coughing). [7] investigates telephone speech, but restricts analysis to only the voiced regions.

Automatic detection of creakiness has several potential applications: characterization of speakers for speaker identification [8,9], identification of languages exploiting creakiness to articulate certain sounds [10], ASR by identifying word boundaries and turn-taking during speech, non-intrusive diagnosis of voice pathology etc.

The proposed algorithm is an extension of a pitch detection algorithm called the Aperiodicity, Periodicity and Pitch (APP) Detector [11]. We modified the APP Detector to distinguish between aperiodicity due to turbulence from that due to irregular phonation, and corrected the pitch estimation of the algorithm during irregular phonation.

2. ANALYSIS OF THE APP DETECTOR

The APP detector is a time domain algorithm that gives a spectro-temporal profile of the amount of voiced and aperiodic energy in a signal, and a pitch estimate for the identified voiced regions [11]. The output of the APP detector for a sample speech signal is given in Fig. 3.

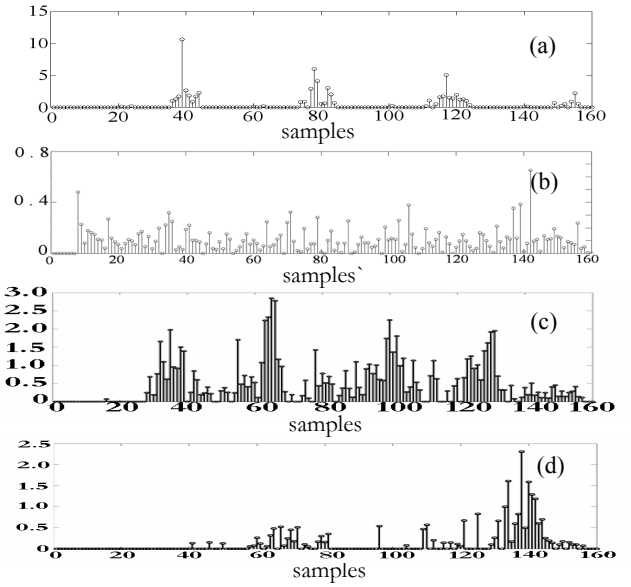
In brief, the speech signal is split into frequency channels by a filter-bank. The Average Magnitude Difference Function (AMDF) is calculated for each channel, to look for periodic structure. When the speech signal is periodic, the AMDF will contain strong dips at lags equal to multiples of the pitch period. When the signal is aperiodic, the dips are weaker and randomly distributed. Decision of

periodicity or aperiodicity of a frame is made by summing the dips across all channels to get a dip profile for that frame. The dips for a periodic frame will cluster tightly together, and for an aperiodic frame, will display random behavior (see Fig. 1). The algorithm calls a frame periodic if it can find strong, tight clusters in the dip profile; channels contributing to those clusters are judged periodic. The dip with maximum strength is judged as the pitch period.

Fig. 1 also shows the dip profiles for a breathy frame and creaky frame. In the case of the breathy vowel, the dips show a mixed behavior – there is clustering of some dips (due to the periodic voiced source dominating at the low frequencies), and there is also some randomness of some dips (noise due to the leakage of air through the glottis, dominating at the higher frequencies). The case of the creaky frame is midway between that of the periodic frame and the aperiodic frame. It lacks a tightly packed dip structure, lacks randomness because of the voicing source and has wider, farther placed clusters due to low pitch.

This loose clustering can be traced back to the AMDF of the creaky voicing. Owing to the fact that the pitch shows some jitter or irregularity during irregular phonation, the AMDF does not have exact alignment of the signal and its delayed version. Thus, the dips are weak, and not in alignment with their counterparts from other channels. The clusters are much broader than would be for the periodic frames and thus, the APP detector calls these frames aperiodic. A simple approach of increasing the allowed cluster width will cause aperiodic frames, especially those occurring during co-articulation, to also be counted as periodic. Further, this will allow the frame to be called periodic, and does not provide provision for identifying creakiness. A different approach needs to be taken – using additional acoustic cues, and identifying characteristic traits of the dip profile.

Figure 1: Dip-profiles for a frame of (a) periodic vowel (b) fricative (c) breathy vowel (d) creaky vowel

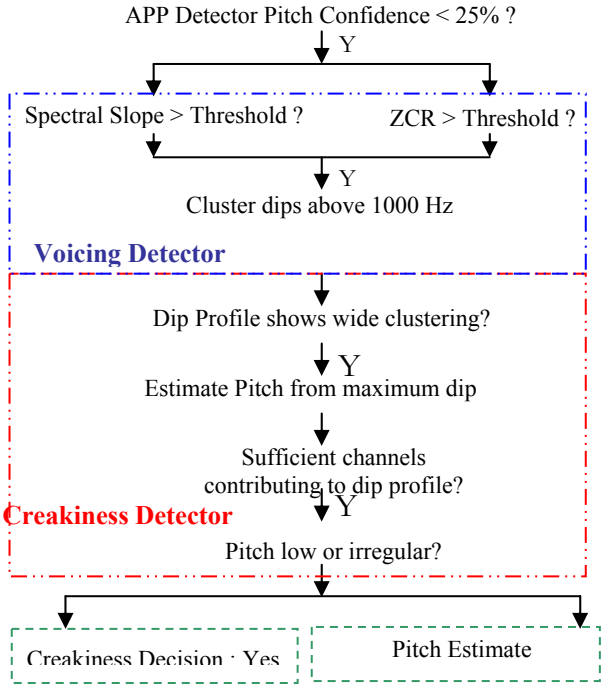


3. IRREGULAR PHONATION DETECTOR

In regions of creakiness, the pitch confidence [11] is seen to be typically less than 25% of the maximum possible (equal to the number of channels). Thus, a frame is labeled a suspect frame if the pitch confidence falls below that, and is processed using the Irregular Phonation Detector. Suspect frames include frames with irregular phonation, but also those with breathy voicing and obstruents. Thus, the first step in this detector is to separate voiced from unvoiced speech. Next, the dip profiles are analyzed for creakiness. Fig. 2 shows a flowchart for the algorithm.

The Zero Crossing Rate (ZCR) is used to eliminate fricatives, laughter and coughing. While the modal voiced regions have a low ZCR than do fricatives, the case is different for creaky voicing. The ZCR of creaky frames is high and comparable to that of the fricatives because of high-frequency damping of the weak glottal pulses. Also, laughter and coughing have significant randomness at lower frequencies. In order to eliminate (1) the ZCR due to high-frequency damping in creakiness, and (2)

Figure 2: Flowchart of the algorithm for the Irregular Phonation Detector



low frequency noise of non-speech, only the low-frequency spectrum of the signal is used to compare ZCR.

In spontaneous speech, due to co-articulation or inaccurate articulation, certain consonants like stops show a phonation pattern very similar to that of a creaky vowel [4]. The spectral slope is a useful parameter to eliminate such cases, since most obstruents have a low frequency spectrum that either rises (fricatives) or remains flat (stops), while creaky voicing shows a falling slope [1,2]. The range of frequencies over which the tilt is calculated is restricted to (100 – 2000) Hz, and a threshold of -16dB / octave gives good separation.

Because of low frequency voicing energy, most breathy vowels and voiced fricatives also pass the voiced / unvoiced detector test. In order to separate these frames from creaky frames, we use the dip profile that only sums dips from channels above 1000 Hz. The motivation is that the voicing information in the former cases is predominantly at the lower frequencies, typically below 1000 Hz, while the creaky frames show their voicing energy even at high frequencies as evidenced by their characteristic vertical striations in spectrograms. Once the voicing information is removed from their dip profiles, the breathy and voiced fricative frames look similar to the aperiodic frames and can be easily separated.

The next step is to characterize the dip-profile that is unique to irregular phonation. This is done by finding all the local maxima in the dip-profile and forming loose clusters around them. If the ratio of in-cluster dips, to total number of non-zero dips, falls below a threshold of 40%, the frame is judged aperiodic. Otherwise, a score is made of the channels contributing to the clusters. The cluster center is then recalculated, using only those channels that have been identified to contribute. The cluster with the maximum dip is then judged as the main cluster, and its maximum is declared as the pitch period.

Using the new cluster centers, the channel contributions of all channels are recalculated to check for consistency. If the number of channels exceeds a threshold, then the pitch estimate is used to make a final decision about creakiness. As irregular phonation is accompanied by either a significant fall or irregular behavior in the pitch, the current pitch is compared with the previous pitch estimates to check for such behavior. In the presence of such activity, the frame is judged as creaky, the irregular phonation pitch estimate replaces the old estimate, and the channel energies are added to get the new pitch confidence.

4. DATABASE & RESULTS

4.1. Reference Software & Database Used

Due to non-availability of a standard database for ground truth, creaky regions were marked by hand in two standard speech databases – TIMIT and NIST 98. The TIMIT database is clean speech sampled at 16 kHz, with clearly uttered phonemes and little non-speech. The database was hand-transcribed for irregular phonation by a graduate student at MIT [6] and checked by the first author. The speaker set included 65 male and 45 female speakers from eight dialect regions (dr1 through dr8). The algorithm was also tested on conversational speech from telephone data sampled at 8 kHz, namely the NIST 98 database, which was hand-transcribed by the first author. Here, the speech is subject to channel distortion that corrupts the amplitudes of harmonics up to 300 Hz and contains background noise. Further, effects like laughter, coughing, etc are also seen.

For the pitch estimation accuracy, the reference pitch was obtained using the ESPS Wavesurfer software using the autocorrelation method and recommended specifications.

4.2. Results

4.2.1. Performance of the Irregular Phonation Detector

The first set of results demonstrates the detection performance and false alarm rate of the Irregular Phonation Detector. The frame rate is 2.5 msec and the window size is 20 msec. Fig. 3 shows a spectrogram of a creaky speech segment, and its spectro-temporal profile, with the creaky region identified in green. It is seen that the creaky region is clearly identified. The following table gives the detection rate of the algorithm:

	Total # creaky instances	# creaky instances identified	Percentage Identified
TIMIT Female	584	543	93.0%
TIMIT Male	816	742	90.9%
NIST Female	100	94	94.0%
NIST Male	100	89	89.0%

Table 1: Detection rate of the algorithm on the TIMIT & NIST 98 databases

The average detection rate is 91.8% for TIMIT and 91.5% for NIST 98 datasets. The algorithm performs well with the NIST 98 database in spite of channel effects. Errors were more for male speakers than female speakers, which might be because creakiness detection is conditioned on the pitch falling below a certain threshold, which is a harder threshold for males since they have a low modal pitch to begin with. The performance is comparable to or exceeds those reported in earlier experiments [5,6,7,9].

The false instances rate was 12.8% for the NIST 98 database, and 17.4% for the TIMIT database. Of the false triggers, 35% were due to voiced fricatives, 40% due to stops, while the remaining 25% were due to stops with a creaky vowel preceding the stop. Though we have currently included these latter two cases as false detections, studies [4,10] discuss that there do exist cases of stops in both American English and other languages, where both voiced stops may be accompanied by irregular phonation.

4.2.2. Improvement in Pitch Estimation Accuracy

This section compares the pitch tracking of the modified APP detector with the ESPS pitch tracker. The frame rate of both pitch tracks is the same, namely 2.5 msec. Only regions where the reference pitch detector has non-zero pitch estimates have been considered. It may be noted that the ESPS pitch tracker may be erroneous, and the reported results are therefore representative. In voiced regions, if the pitch values of the APP detector and the reference vary by more than 15 Hz or 10% of the current pitch, then the APP detector is declared to be in error. The pitch comparison tests were run on 20,000 frames of the NIST database and 8,000 frames of the TIMIT database.

The pitch detection accuracy improved from 95.4% to 98.3% for the NIST database and from 94.8% to 97.4% for the TIMIT database, due to pitch correction in creaky regions. The increase in false pitch estimation (reporting pitch for unvoiced frames) was from 6.8% to 7.2% for the NIST database and 3.2% to 3.4% for the TIMIT database. Thus, we have an overall improvement of 3% at the expense of 0.3% false pitch estimates.

5. CONCLUSIONS & FUTURE WORK

We have discussed an algorithm for detection of irregular phonation in speech. The algorithm shows good recognition rate and pitch correction on both clean & telephone speech. We plan to improve our results by looking for other acoustic correlates that are robust to co-articulation and channel distortions. We also plan to develop an algorithm to non-intrusively diagnose voice disorders by arriving at the vocal fold activity using signal processing strategies.

6. REFERENCES

[1] Stevens, K.N. 1998. *Acoustic Phonetics*, Cambridge, Massachusetts: MIT Press.

[2] Klatt, D., Klatt, L. 1990. Analysis, synthesis and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.* 87, 820-857.

[3] Gerratt, B.R., Kreiman, J. 2001. Toward a taxonomy of nonmodal phonation. *J. Phonetics*, 29, 365-381

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

[5] Ishi, C. T., Ishiguro, H., Hagita, N. 2005. Proposal of acoustic measures for automatic detection of vocal fry. *Proc. Eurospeech 2005*, 481-484

[6] Surana K., Slifka, J. 2006. Acoustic cues for the classification of regular and irregular phonation. *Proc. 9th ICSLP,* Pittsburgh.

[7] Yoon, T. Cole, J., Hasegawa-Johnson, M. Shih, C. (2005) Detecting non-modal phonation in telephone speech. UIUC ms.

[8] Espy-Wilson, C.Y. , Manocha, S., Vishnubhotla, S. 2006. A new set of features for text-independent speaker identification. *Proc. 9th ICSLP,* Pittsburgh.

[9] Vishnubhotla, S., Espy-Wilson, C.Y. 2006. Automatic detection of irregular phonation in continuous speech. *Proc. 9th ICSLP,* Pittsburgh.

[10] Gordon, M. Ladefoged, P. 2001. Phonation types: a cross-linguistic overview. *J. Phonetics*, 29, 383-406

[11] Deshmukh, O., Espy-Wilson, C. Y., Salomon, A., Singh, J. 2005. Use of Temporal Information: Detection of Periodicity, Aperiodicity, and Pitch in Speech", *IEEE Trans. Speech & Audio Proc*, 13(5), 776-786.

Figure 3: (top) a sample speech signal, (middle) its spectrogram, (bottom) output of the Irregular Phonation Detector superposed on the APP detector. Green shows region where irregular phonation was detected. Blue shows regions where the signal is periodic, while red shows the aperiodic regions. For the latter two panes, the y axis shows frequency in decimal scale.