Resonator-based Creaky Voice Detection

Thomas Drugman[1]*, John Kane*[2]*, Christer Gobl*[2]

[1]TCTS Lab - University of Mons, Belgium
[2]Phonetics and Speech Laboratory,
School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, Ireland
thomas.drugman@umons.ac.be, kanejo@tcd.ie, cegobl@tcd.ie

Abstract

Creaky voice is used by speakers for a variety of interactive, expressive and stylistic reasons. As a result the accurate detection of creaky regions in speech can yield important information not captured within the propositional content of spoken utterances. Hence, we describe a new method for automatically detecting creaky regions following the observation that secondary peaks occur in the linear prediction residual signal. The proposed approach was shown to significantly outperform the state-of-the-art in an objective evaluation on a range of speech databases.
Index Terms: Voice quality, glottal source, creak, vocal fry

1. Introduction

Creak (as well as other voice quality labels such as: glottal fry, vocal fry and laryngealisation) refers to a voice quality stemming mainly from a distinctive, non-modal laryngeal articulation. Creak typically involves ventricular incursion [1] which occurs when the ventricular folds press down and slightly cover the *true* vocal folds, resulting in an increased mass which produces a lower frequency of vibration and can result in vibration also occurring above the level of the glottis. Creak further involves strong adductive vocal fold tension, weak longitudinal tension and low levels of subglottal pressure [2]. These settings can be combined with those used for producing modal voice to get the compound voice quality *creaky voice*.

This mode of phonation produces dramatically different acoustic characteristics than those resulting from modal phonation. The most striking features include: extremely long glottal pulse duration and as a consequence, little or no superposition of formant oscillations between adjacent glottal pulses, very long glottal closed phase [3] and the presence of secondary (and at times even tertiary) excitations [4]. These secondary excitations are likely due to ventricular incursion, and, hence, may be produced slightly above the level of the glottis.

As a result of these characteristics many standard analysis methods (including F_0 tracking, and spectral analysis) are often unsuitable. For instance, the very low F_0 values in creak (where pulses can occasionally be as long as 100 ms [4]) may be below the lower limits of many F_0 algorithms. As the standard frame length for various analysis methods is typically no longer than 32 ms, and as at least two glottal periods are required for periodicity information (resulting in a minimum F_0 value of 62.5 Hz), analysis of creak segments may not provide any meaningful information.

Consequently, many speech technology applications (e.g., statistical text-to-speech synthesis) tend to discard creak segments following spurious acoustic values. However, creak is commonly produced in speech for a variety of interactive, expressive and stylistic reasons. It is so often produced in utterance final position in Finnish that listeners preferred synthetic speech containing creak [5]. The use of creak has been shown to be an important conversational tool in a range of languages, particularly in regard to turn-taking [6] and hesitations [7]. It is also thought to allow insights into speaker's affective and expressive states [8, 9].

In order to exploit this source of information one requires the ability to automatically detect creaky regions in speech signals. To the best of our knowledge, only two approaches have been proposed in the literature for such a purpose [10, 11], though several further methods exist for detecting the broader class, irregular phonation. These two techniques are described in Section 3.1 and although they are useful for analysing speech containing creak, in our experience they can at times lead to excessive false positives.

In the present work, we describe a Resonator-based Creaky Voice Detection (RCVD) method which emphasises the presence of secondary peaks in the linear prediction (LP) residual signal. We provide a comprehensive description of the method (Section 2) which is then compared to the state-of-the-art in an objective evaluation using annotated speech data from a number of speech databases (Section 3).

2. Proposed technique

The proposed technique, called Resonator-based Creaky Voice Detection (RCVD), arises from the observation that creaky voice production results in secondary peaks in the LP-residual signal. These extra peaks can occur due to secondary laryngeal excitations and also from sharp discontinuities at glottal opening, following a long glottal closed phase.

The aim is therefore to determine the significance of these secondary excitation pulses in order to detect creaky regions in speech. The key idea behind our method is that by passing the LP residual signal through a resonator, secondary peaks will perturb its output and cause the appearance of a greater amount of harmonics. To illustrate this, Figs. 1 and 2 display the resonator output (described below) for 'normal', voiced phonation and for creaky voice, respectively. In the first case, it is noticed that the residual excitation exhibits major peaks only at the Glottal Closure Instants (GCIs, [12]). As a result, perturbations between two major excitation peaks are relatively weak and the oscillating signal at the output of the resonator will contain a small amount of harmonics. On the other hand, for creaky voice, secondary pulses significantly re-excite the resonator between two consecutive GCIs, leading to perturbations in its output which will be reflected by a greater richness of harmonics.

The workflow of the proposed approach is shown in Fig. 3. First the residual signal is obtained by Linear Preditive Coding

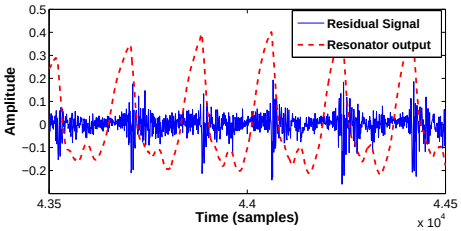

Figure 1: *Illustration of the output of a resonator excited by the residual signal of a* ***voiced segment of regular phonation****.*

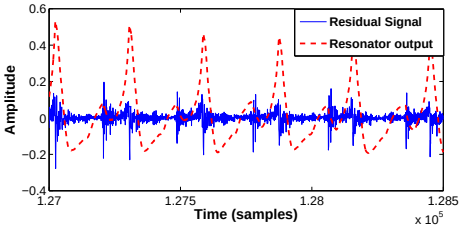

Figure 2: *Illustration of the output of a resonator excited by the residual signal of* ***creaky voice****.*

(LPC) inverse filtering, using an order of $F_s/1000 + 2$. This residual excitation is the input of two resonators used for different purposes. Both resonators are set using complex-conjugate poles. One is used for estimating the F_0 contour and the other for measuring the significance of secondary pulses (as shown in Figs. 1 and 2). Both resonators are centred on the mean fundamental frequency ($F_{0,mean}$) but use different bandwidths. In this work, $F_{0,mean}$ is estimated via the Summation of Residual Harmonics (SRH, [13]) algorithm available in the GLOAT toolbox, although this choice is not critical.

For estimating the F_0 contour the bandwidth of Resonator 1 was set to 1100 Hz as it gives a reasonable compromise between avoiding ambiguity with octave jumps (bandwidth too high) and capturing the spread of F_0 values from the $F_{0,mean}$ often found in creaky parts (which might not be achieved correctly if the bandwidth is too low). To estimate the local F_0, a 50 ms-long Hanning window is applied to the resonator output and the corrected autocorrelation function $r'(\tau)$ is calculated:

$$r'(\tau) = \frac{N}{N-\tau} \cdot \text{autoCorr}(\tau), \qquad (1)$$

where N is the window length (in samples) and τ is the number of autocorrelation lags. As in [8], the correction $\frac{N}{N-\tau}$ compensates the decreasing properties of autocorrelation functions as τ increases. The local glottal period is then determined by the position of the maximum in $r'(\tau)$ after removing the peak centred on $\tau = 0$.

For highlighting secondary excitations, a more pronounced resonating character is needed and, hence, the bandwidth of Resonator 2 is set to 150 Hz. To measure the importance of secondary pulses, the amplitude difference (in dB) between the two first harmonics ($H2 - H1$) is computed on the spectrum of the autocorrelation function, as it allows to enhance harmonic peaks. Note that $H2 - H1$ is then filtered by a 100 ms-long moving average filter to lessen the impact of outlier values.

Fig. 4 gives an example of creaky voice detection using the proposed method. It can be noted that $H2 - H1$ in creaky regions clearly emerges from its values in regular phonation. Detected creaky parts are then simply obtained by applying a threshold to $H2 - H1$ (setting this threshold is discussed in Section 3.3). As shown in Fig. 4, the $H2 - H1$ contour clearly increases above 0 dB in the creaky region.

In addition, a module of post-processing has been appended to the workflow presented in Fig. 3 to remove possible detections in silent and unvoiced parts. For this, silent regions are detected based on the signal energy and unvoiced segments relying on the zero-crossing rate. For both, 20 ms-long windows are considered and thresholds have been carefully set so that no actual creaky regions are removed.

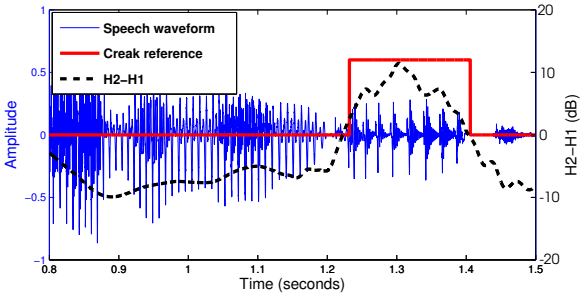

Figure 4: *Illustration of $H2 - H1$ contour for detecting creaky regions.*

3. Experiments

3.1. Comparison algorithms

In order to evaluate the performance of the proposed method, we used the algorithms described respectively in [10] and [11] as a comparison. To the best of our knowledge, these are the only two methods available in the literature for creaky voice detection.

3.1.1. Ishi's method for creaky voice detection [10]

The algorithm described in [10] involves processing the speech signal band-limited to a frequency range of 100-1500 Hz. Part of the analysis is carried out on a standard frame-synchronised basis, and the other part is glottal pulse synchronised following the measurement of peaks in a 'very short-term' power contour, which are used to mark the glottal pulses. Glottal pulses displaying strong Power Peaks (PwP) in the 'very short-term' power contour are considered creak candidates. The frame-synchronised part involves estimation of an Intraframe Periodicity (IFP) strength contour which is used for differentiating 'normal voiced' and creaky voiced regions. This is done by considering multiples of the strongest peak in the normalised autocorrelation function taken on 32 ms frames. Next, an interpulse similarity (IPS) measure is used to differentiate unvoiced and creaky regions. This is done by calculating the normalised cross-correlation function for regions around adjacent creak candidate glottal pulses. The IPS parameter is the maximum cross-correlation value when adjacent pulses are below 100 ms apart.

PwP values above or equal to 7 dB are considered candidates and IFP $\leq$ 0.5 & IPS $\geq$ 0.5 are the necessary conditions for the glottal pulse to be consider 'creak'. Adjacent creak pulses, below 100 ms apart, are merged to construct the creak region.

3.1.2. Extension of the Aperiodicity, Periodicity and Pitch (APP) detector [11]

This method has been proposed in [11] for the automatic detection of irregular phonation, including sounds referred to as creak, vocal fry, diplophonia, glottalization, laryngealization, pulse register phonation and glottal squeak [11]. In a first

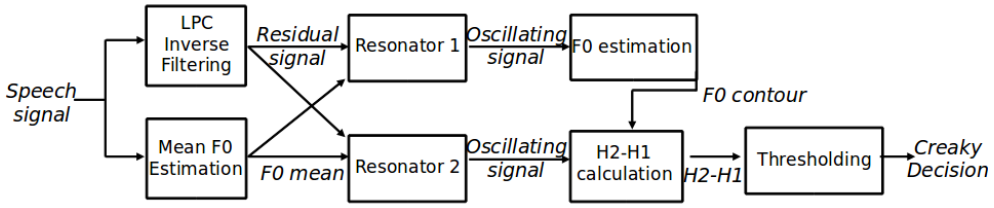

Figure 3: *Workflow of the proposed technique. The creaky voice decision stems from the harmonic structure produced when Resonator 2 is excited by the LP-residual. Resonator 1 is used to estimate the F_0 contour, including creaky regions. Both resonators are centred on $F_{0,mean}$ but have different bandwidths.*

step, the algorithm separates irregular frames from periodic frames using the periodicity measure of the APP detector. In a second step, irregular phonation is differentiated from aperiodic frames, breathy vowels and voiced fricatives using their 'dip profile' on the Average Magnitude Difference Function (AMDF) in various frequency bands. Finally, the potential confusion with some stops is addressed by calculating the spectral slope. In our evaluation, we used the original implementation kindly shared by the authors of this method.

3.2. Experimental protocol

The characteristics of creak can differ considerably across speakers, so in order to properly evaluate the detection algorithms we used different speech databases. The first three contain instances of creak and include text-to-speech corpora of a Finnish male speaker (MV, as used in [14]), a Finnish female speaker (HS, as used in [5]) and an American English male speaker (ARCTIC-BDL [15]). We also included two further databases which do not have creaky segments; a Scottish male speaker (ARCTIC-AWB [15]) and an American female speaker (ARCTIC-SLT [15]). 50 sentences from each database were used in the evaluation (250 sentences in total). All speech data were downsampled to a sampling frequency of 16 kHz.

Human annotation of creaky regions was required to evaluate the performance of detection. The first two authors carried out this annotation using a similar approach to that described in [10]. To determine the creaky regions an auditory criterion was used: *a rough quality with the additional sensation of repeating impulses*. However, inspection of waveforms, spectrograms and F_0 contours was used to help guide the annotation. Furthermore, sentences for which a possible ambiguity about the annotation remained were not considered for the evaluation.

The performance of the methods is evaluated at both the frame and the event levels. For frames, three metrics are employed: the True Positive Rate (TPR, also called recall), the False Positive Rate (FPR), and the F1 score. TPR is the proportion of actual creaky frames that are retrieved. FPR is the proportion of actual non-creaky frames that are erroneously detected as creaky. The F1 score is a single measure (bound between 0 and 1) computed using precision and recall. The better the technique, the higher the TPR and F1, and the lower the FPR. At the event-based level, the metrics used are the number of hits, misses and false alarms.

Four techniques are compared in our experiments: Ishi's method as described in Section 3.1.1, the APP method described in Section 3.1.2, the proposed RCVD approach, and RCVD with the Post-Processing (PP) removing silent and unvoiced regions (hereafter called RCVD+PP).

3.3. Results

As a reminder, the proposed RCVD approach makes use of a threshold applied to $H2 - H1$ to detect creaky regions. The influence of this threshold on the F1 score is presented in Fig. 5. Clearly, a threshold of 0 dB gives consistent performance for all databases, therefore this setting was used throughout the evaluation. However, to fully explore the optimal setting of this threshold, further analysis on larger amounts of data is required.

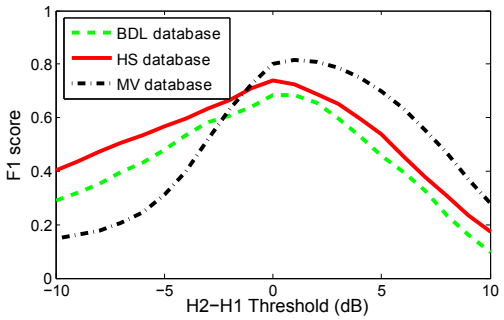

Figure 5: *Impact of the threshold on the creaky voice detection.*

The results in terms of the frame level evaluation metrics are presented for the speech databases containing creak in Fig. 6. Interestingly the RCVD method clearly outperforms the two comparison algorithms, leading to a significant increase in both TPR and F1 score, across all databases. Notably for the MV dataset, Ishis method displayed a surprisingly high level of false positives. We further examined these false positives and found that in a large number of instances, the F_0 value of the detected segment was very low (speaker MV has a generally low F_0 frequently around 80 Hz), but not low enough for the individual glottal pulses to be perceived. In these cases the IFP (Section 3.1.1) contour fell below the threshold, leading to creak being incorrectly detected. In contrast, this was not the case for the HS database where Ishi's algorithm and RCVD+PP provide very low FPR values. Overall, the RCVD+PP technique produced the lowest level of false positives and the APP-based detection exhibited the lowest F1 score, mainly due to the low level of true positives. This can be explained by the fact that this method detects very short segments of irregular phonation. And even though these detections are in the middle of a creaky part, most of it is missed by the algorithm.

Results in terms of event detection are displayed in Table 1. Considering voices containing creak (BLD, HS and MV), it can be seen that APP and Ishi's method produced a high number of false alarms (except for the female speaker with Ishi's method), for the reasons explained above. Although the false alarms for APP correspond to very short detected segments, this was also true for its hits and our attempts to apply a post-processing to remove its FAs while keeping a similar hit rate failed. It is also worth mentioning that the original implementation of the APP-

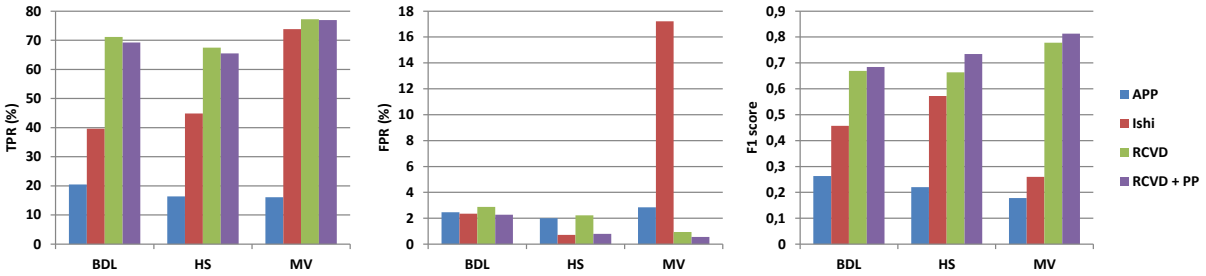

Figure 6: *TPR (left), FPR (middle) and F1 scores (right) for the creak detection algorithms on the three databases containing creak.*

Table 1: *Summary of event level evaluation metrics for the four algorithms across the 5 databases. Best performance is highlighted in bold.*

Database	Metric	APP	Ishi	RCVD	RCVD + PP
BDL	Misses	22	8	**3**	**3**
	FAs	215	93	37	**27**
	Hits	47	61	**66**	**66**
HS	Misses	59	**12**	**12**	15
	FAs	301	38	89	**37**
	Hits	46	**93**	**93**	90
MV	Misses	44	**2**	7	7
	FAs	353	920	49	**14**
	Hits	31	**73**	68	68
AWB	FAs	371	76	33	**12**
SLT	FAs	122	6	19	**4**

based technique was designed to detect 'irregular phonation' (and not only creaky voice), which can also partially explain its high proportion of FAs.

Overall, RCVD+PP achieved the best results. The advantage of applying the post-processing (PP) is clearly noted: except for three extra missed events on the HS dataset, there were no additional missed detections, however there was a considerable reduction of false alarms. This was also reflected in Fig. 6: at the expense of a minor reduction of TPR, applying PP led to a clear improvement in the FPR and F1 scores. As shown at the end of Table 1, the same conclusions hold for voices with no creak (AWB and SLT). RCVD+PP is clearly the most successful method with only a small number of false alarms. Strikingly, APP generated a prohibitively high number of false alarms as did Ishi's method on AWB (male speaker).

4. Conclusion

This paper presents a Resonator-based Creaky Voice Detection (RCVD) technique, which focuses on the characteristics of the secondary excitations typically occurring in creaky speech. This technique is aimed at characterising the secondary excitation pulses typically occurring in creaky speech, through the use of a resonator and subsequent harmonic measurement. Compared to the state-of-the-art, the RCVD method consistently produced better results for the five voices tested with particularly notable reductions in false alarms (especially for male speakers) and missed detections.

5. Acknowledgements

The authors would like to thank Hanna Silén and Martti Vainio for kindly providing the HS and MV databases and Srikanth Vishnubhotla and Carol Espy-Wilson for kindly sharing the APP method. The first author is supported by the Walloon Region (Grant WIST 3 COMPTOUX # 1017071). The second and third authors are supported by the Science Foundation Ireland, Grant 07/CE/I1142 (Centre for Next Generation Localisation, `www.cngl.ie`) and Grant 09/IN.1/I2631 (FASTNET).

6. References

[1] Edmondson, J.A., Esling, J.H.: The valves of the throat and their functioning in tone, vocal register and stress: laryngoscopic case studies, Phonology 23(2), pp. 157-191, 2006.

[2] Laver, J. "The Phonetic Description of Voice Quality", Cambridge University Press, 1980.

[3] Gobl, C., Ní Chasaide, A., "Acoustic characteristics of voice quality", Speech Communication, 11, pp. 481-490, 1992.

[4] Blomgren, M., Chen, Y., Ng, M., Gilbert, H. "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers", J. Acoust. Soc. Am., 103(5), pp. 2649-2658, 1998.

[5] Silén, H., Helander, E., Nurminen, J., Gabbouj, M., "Parameterization of vocal fry in HMM-based speech synthesis", in Proc. of Interspeech, pp. 1775-1778, 2009.

[6] Ogden, R., "Turn transition, creak and glottal stop in Finnish talk-in-interaction", Journal of the Internation Phonetic Association, 31 (1), pp. 139-152, 2001.

[7] Carlson, R., Gustafson, K., Strangert, E., "Prosodic Cues for Hesitation," in Proceedings from Fonetik 2006 , pp. 2124, 2006.

[8] Ishi, C., Ishiguro, H., Hagita, N., "Using Prosodic and Voice Quality Features for Paralinguistic Information Extraction," in Proc. of Speech Prosody 2006.

[9] Kane, J., Pápay, K., Hunyadi, L., Gobl, C., "On the use of creak in Hungarian spontaneous speech," in Proc. of ICPhS, pp. 1014-1017, 2011.

[10] Ishi, C., Sakakibara, K., Ishiguro, H., and Hagita, N., "A method for automatic detection of vocal fry", IEEE Transactions on Audio, Speech and Language Processing, 16 (1), pp. 47-56, 2008

[11] Vishnubhotla, S., Espy-Wilson, C., "Automatic detection of irregular phonation in continuous speech", Proceedings of Interspeech, Pittsburgh, pp. 949-952, 2006.

[12] Drugman, T., Thomas, M., Gudnason, J., Naylor, P. and Dutoit, T. "Detection of Glottal Closure Instants from Speech Signals: a Quantitative Review", IEEE Transactions on Audio, Speech and Language Processing, 20 (3) pp. 994-1006, 2012.

[13] Drugman, T., Alwan, A., "Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics", Proc. Interspeech, Florence, Italy, pp. 1973-1976, 2011.

[14] Vainio, M., "Artificial neural network based prosody models for Finnish text-to-speech synthesis," Ph.D. dissertation, University of Helsinki, Finland, 2001.

[15] [Online], "CMU ARCTIC speech synthesis databases", *http://festvox.org/cmu arctic/*.