Modeling the Creaky Excitation for Parametric Speech Synthesis

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

In order to produce natural sounding output, corpus-based speech synthesis systems need to be able to properly model the acoustic variability in the corpus. Creaky voice is a voice quality frequently produced in many languages, in both read and conversational speech settings. However, the creaky excitation displays different acoustic characteristics than modal excitations and is, hence, not suitably modelled by standard vocoders. This study presents an analysis of the creaky excitation which is used to derive an extension of the Deterministic plus Stochastic Model of the residual signal. This proposed model is designed to appropriately model creaky voice and is integrated into a vocoder for parametric speech synthesis. Copy-synthesis versions of short speech segments containing creaky voice were used in a subjective listening test which revealed clearly better rendering of the voice quality than a standard vocoder.
Index Terms: Voice quality, speech synthesis, creak, vocal fry

1. Introduction

For statistical parametric synthesis to produce natural sounding speech output there needs to be proper acoustic modeling of the different speech sounds produced in the corpus. Recently there have been some significant gains in the naturalness of HMM-based statistical speech synthesis (HTS), particularly from improved excitation modeling in vocoders (see e.g., [1, 2, 3]).

Speakers, however, frequently adopt different phonation modes, often in conversational settings but also in read speech for text-to-speech (TTS) corpora. These different phonation modes may produce acoustic characteristics that are not properly modeled by existing vocoders. In this study we focus on one particular voice quality, namely creaky voice, which arises from a particular non-modal phonation. Creaky voice displays distinctive acoustic characteristics, such as: extremely long glottal pulse duration (and as a consequence, little or no superposition of formant oscillations between adjacent glottal pulses), long glottal closed period, and the presence of secondary excitations (see, for example, [4]). In our experience traditional vocoders are not properly suited to these characteristics.

In many languages creaky voice is frequently produced even in read speech. In Finnish, for instance, speakers frequently produce creaky voice in sentence final position [5]. This is also true for other languages including many dialects of American English (see e.g., the BDL sentences from the ARCTIC database [6]). Other languages like some North Caucasian languages and Quiavin Zapotec utilise creaky voice quality for phonetic contrast [7]. In conversational speech creaky voice is particularly prevalent, with studies demonstrating associations with turn-taking [8], hesitations [9] and various forms of expression [10]. The development of a vocoder which can provide a natural rendering of creaky voice is, hence, clearly desired for standard TTS where creaky voice exists in the corpus or for the development of conversational or expressive speech synthesis.

Some studies have involved synthesis of creaky speech, mainly using formant synthesis. For instance in [11] the authors use the KLATT88a formant synthesiser to produce creaky voice by reducing F_0 controlling the DI-diplophonia parameter, as well as modifying a number of other parameters. In [9] the authors use the KTH formant synthesiser to generate creaky voice for the purpose of studying the perception of hesitations. They modeled creaky voice by modifying glottal parameters in time and amplitude for every second pulse. To the best of our knowledge the only study on modeling creaky voice in HTS was presented in [5] and arose from the need to provide natural rendering of creaky voice in Finnish. Their method focuses on providing robust f_0 estimation and suitable voicing decision in creaky regions. They do not, however, focus on modeling the characteristics of the creaky excitation which may play a significant role in producing the correct timbre.

In the present study we focus on modelling the creaky excitation as an extension of the Deterministic and Stochastic Model (DSM, [1]) synthesis system. We begin by introducing the speech data used in this study (Section 2). The framework adopted for analyzing creaky excitation is then described in Section 3. The proposed vocoder is described in Section 4 and is put to a subjective evaluation (Section 5).

2. Speech Data

The speech data used in the present study comes from two separate speech synthesis databases. The first is the recorded speech of a male American speaker (the BDL voice in [6]) and the second is a Finnish male speaker (the MV database used in [12]). Both databases were downsampled at 16 kHz. We carefully selected 100 sentences from each database which included creaky regions. These creaky regions were then manually annotated by following a similar annotation procedure as was used in [10]. Qualitative analysis of the creaky excitations revealed similar patterns across the two speakers with the presence of sharp secondary excitation peaks. Note, however, that a range of excitation settings can give rise to the perception of creak and can involve more irregular patterns than those observed in the two speakers in the present study.

3. Analysis of Creaky Excitation

This section summarizes our framework for the analysis of the creaky residual excitation and leads to our proposed model,

which will be integrated into a vocoder (Section 4). The general analysis workflow which we adopted is displayed in Figure 1. From a database of a given speaker producing creaky voice, it aims at estimating the data-driven components of the proposed model. The process consists of four consecutive steps: detecting the creaky segments of speech, estimating the Glottal Closure Instant (GCI) positions and the pitch contour in creaky voice, determining the instants of the secondary pulses and finally estimating the data-driven waveforms used in the proposed model of the creaky excitation. These four operations are detailed in the following sections.

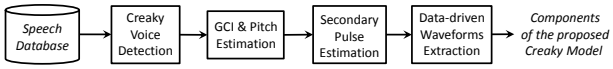

Figure 1: *Workflow for the analysis of creaky excitation leading to the estimation of the components of the proposed model.*

3.1. Creak Voice Detection

In the first analysis step, the creaky segments are detected from continuous speech. Some algorithms have recently been developed [13] for the automatic detection of these segments with relative success (sensitivity between 65 and 77%, specificity between 97.5 and 99.5%). Nonetheless, the manual creaky annotations are considered in the following in order to remove the influence of possible errors made by the creaky voice detector in the process.

3.2. GCI and Pitch Estimation

Glottal Closure Instants (GCIs, [14]) refer to the moments of significant excitation that occur at the level of the vocal folds during voiced speech. Locating GCIs is a necessary first step in pitch synchronous speech processing, in particular when there is processing of the main excitation region in the LP-residual signal. The performance of GCI detection on modal speech has reached a certain level of maturity [14]. However, for voice qualities like creaky voice, which display dramatically different glottal closing characteristics, most GCI algorithms fail to provide usable levels of performance. A recent study looked at optimising GCI detection on non-modal voice qualities [15]. The study found that by taking the SEDREAMS algorithm (described in [14]) and applying a post-processing method to it can make the performance suitable for analysis of creaky regions.

Considering GCIs estimated by the SEDREAMS algorithm in relation to the derivative electroglottographic (dEGG) signal (bottom panel of Figure 2) there is clearly appropriate GCI detection up to around 1.2 seconds. For the following creaky region, although the algorithm appears to output 'correct' GCIs, there are also clearly a large number of false alarms. In order to remove these false alarms one can make use of the output of a resonator applied the residual signal, derived from LPC analysis and subsequent inverse filtering. The resonator is designed with a centre frequency of $F_{0,mean}$ and with a bandwidth set to 150 Hz. This outputs a waveform (shown in the top panel of Figure 2) which displays strong negative peaks in the region of the GCI. The post-processing step involves getting the average of the negative resonator peak value at the previous and following estimated GCI values. This average is then multiplied by a weight, w_{pp}, and if the result is greater than the negative peak corresponding to the present GCI then this GCI is removed. In a previous study the optimal w_{pp} was found to be 0.4 [15] and is used here. Note that this method works sufficiently well for creaky regions which display a moderate level of irregularity.

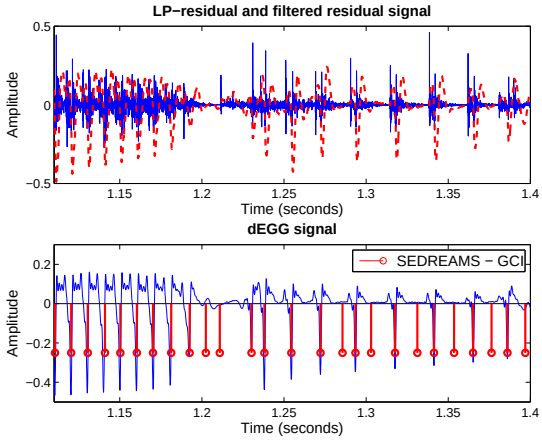

Figure 2: *LP-residual signal (solid line) and the output of a resonator (dashed line) applied to this signal (top panel) and dEGG signal with GCIs as estimated by the SEDREAMS method (bottom panel). Creaky region begins from around 1.2 seconds.*

3.3. Secondary Peak Estimation

As can be seen in the top panel of Figure 2, one striking acoustic feature of creaky speech is the presence of secondary pulses in the excitation signal [4]. These extra peaks can occur due to secondary laryngeal excitations, butalso from sharp discontinuities at glottal opening, following a long glottal closed period. Through our analysis of speakers BDL and MV, we noticed these secondary excitation peaks to be mainly linked to the glottal opening instant. This can also be seen from the inspection of Figure 2 with parallel dEGG recordings. Therefore we use in the following the term *open period* to refer to the timespan between the secondary pulse and its consecutive GCI, although the reader should be aware of the possible limitations of this terminology.

In this study, the secondary excitation peaks are simply located by looking for the greatest discontinuity between two consecutive GCIs. The regions around GCIs (with a tolerance of $\pm$ 1 ms) are not considered in order to avoid detection in the close vicinity of the GCI discontinuity. Figure 3 shows the histograms of the durations of the resulting F_0 and open periods for the creaky segments produced by speaker BDL. It is observed that the F_0 has very low values with a large variability ranging from about 45 Hz to 110 Hz. Interestingly, the histogram for the open period durations displays one prominent narrow peak. This suggests that in creaky voice, even when F_0 is varying, the open period remains relatively constant. The same pattern was also found for speaker MV.

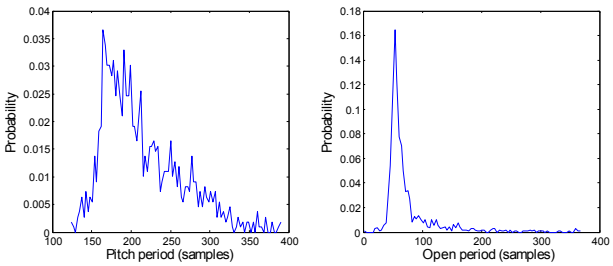

Figure 3: *Distribution, for the creaky regions of speaker BDL, of the duration (in samples) of the F_0 period (left panel) and of the open period (right panel).*

3.4. Extracting the Data-driven Waveforms of the Proposed Model

The approach we adopted for modeling the creaky excitation is an extension of the Deterministic plus Stochastic Model (DSM) proposed in [1]. However, the original DSM models the excitation of modal speech while the model we propose in this paper integrates the presence of secondary pulses in the creaky excitation. As in DSM, the excitation signal is assumed to consist of two components acting in two separate spectral bands delimited by a cut-off frequency F_m (sometimes referred to as the maximum voiced frequency): the deterministic and stochastic components modeling the low and high frequencies [1], respectively. However, the estimation of the data-driven waveforms for these two components is now split into two parts: the open and the closed period.

In other words, the open and closed periods of the residual excitation are extracted from the analysis corpus and isolated in two separate datasets. For each dataset, data-driven waveforms are estimated in the same way as for DSM [1]. These latter waveforms are the first eigenvector obtained by Principal Component Analysis (PCA) for the deterministic component, and the energy envelope for the stochastic component. These components (for two glottal cycles centered on a GCI) are shown in Figure 4 for speaker BDL. Two clear discontinuities are observed in the deterministic component (left panel) at the GCI and secondary peak locations. This latter waveform exhibits some residual phase information of the glottal formant in the open period (as was the case for DSM [1]) while it mainly consists of zero values in the closed period. The energy envelope (right panel of Figure 4) gives some indication of how the turbulence noise in the excitation are temporally distributed in creaky voice.

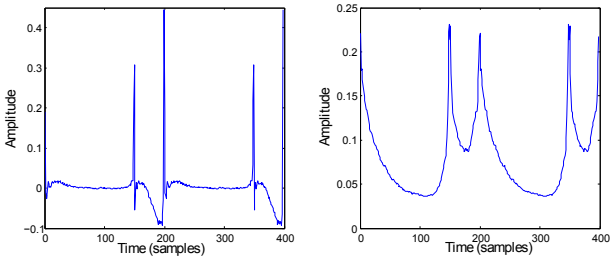

Figure 4: *Two glottal cycles with the data-driven components of the proposed model of the creaky excitation for BDL: the first eigenvector for the deterministic component (left panel) and the energy envelope for the stochastic component (right panel).*

4. The Proposed Vocoder

The vocoder incorporating the proposed model for the reconstruction of the creaky excitation is presented in Figure 5. The deterministic component $r_d(t)$ of the residual signal is obtained from the first eigenvector calculated in both the open and closed periods, as extracted in Section 3.4. Given the observation in Section 3.3 that the open period duration has a very sharp distribution, it is taken to be constant for a given speaker. According to our analysis, it is fixed to a value of 3.75 ms for BDL and 5 ms for MV. For this reason, the conversion towards the target pitch is achieved by only resampling the closed period. The final $r_d(t)$ component is obtained by retaining its low-pass content below the frequency F_m (fixed to 4 kHz in this work).

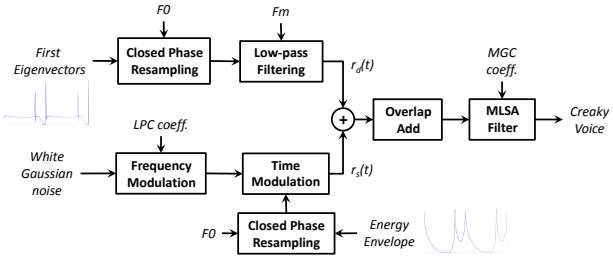

Figure 5: *Vocoder incorporating the proposed model of the creaky excitation. Its inputs are the target F_0 and the Mel-Generalized Cepstral (MGC) coefficients. All other data are precomputed from the speaker-dependent analysis step.*

The stochastic component $r_s(t)$ of the residual signal consists of a white Gaussian noise modulated both in the time and frequency domains. As in DSM, the spectral envelope of the noise is captured at analysis time by LPC modeling [1]. The time dispersion of the noisy component in the excitation is taken into account by making use of the energy envelope extracted in Section 3.4. Here again, the open period is considered to have a fixed constant duration and the pitch transposition is performed by resampling only the closed period.

The final residual signal is obtained by adding its $r_d(t)$ and $r_s(t)$ components and by overlapping and adding the resulting GCI-synchronous excitation frames. In a last step, this excitation signal is the input of the Mel-Log Spectrum Approximation (MLSA) filter, controlled by the Mel-Generalized Cepstral (MGC) coefficients to get the creaky voice. Note that in non-creaky segments of speech, the vocoder makes use of the standard DSM which is known to be suited for modeling modal speech.

5. Subjective Evaluation

The subjective evaluation aims at quantifying the perceptual improvement brought by the integration of the proposed model of the creaky excitation. For this, the vocoder described in Section 4 is compared to a standard method: the DSM vocoder. Indeed standard techniques (such as the traditional DSM) have been designed to model the excitation in modal voiced speech. Nonetheless, they exhibit some drawbacks when producing non-modal effects such as creaky voice.

Twenty-two people, all in the area of speech research, participated in the subjective evaluation. Participants were either native English speakers or had English as a second language and none of them spoke Finnish. Two experiments were carried out: an ABX and a CMOS test. In both tests we focused on the copy-synthesis of short segments of speech (typically 2 second-long) containing creak. For this, the two male speakers BDL (US English) and MV (Finnish) were considered. Each test consisted of 20 stimuli to be scored (10 per speaker).

In the first test (ABX), participants were given the original segment of speech as a reference (X) and were asked which one of versions A or B is the closest. A and B were the segment X vocoded either by the traditional or proposed technique (randomly shuffled). Participants had also the possibility to say that A and B were equivalent. Results of the ABX test are displayed in Figure 6, and show consistent improvement for BDL and MV. About 60% of the responses indicated a preference for the proposed model, while around 15% of the responses found the version vocoded by DSM to be closer to the original. Finally, about 20-25% of responses did not favour either method.

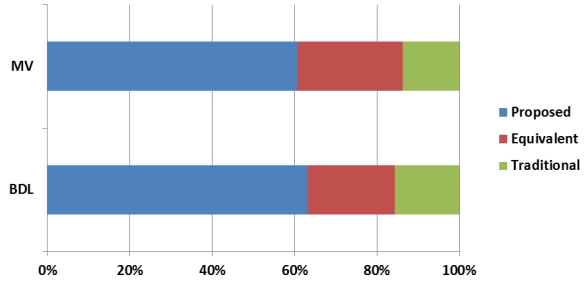

Figure 6: *Results of the ABX test for speakers BDL and MV.*

The second test was a Comparative Mean Opinion Score (CMOS) test which aimed at assessing the participant preference by providing pairwise stimuli vocoded by the traditional and proposed method. For each segment of speech considered, participants were asked to listen to both versions (randomly shuffled) and to attribute a score according to their overall preference using the 7-point gradual CMOS scale.

Results of the CMOS test are presented in Figure 7 for the two speakers BDL and MV and provide a comparison between the proposed model and the traditional vocoder according to the CMOS 7-point scale. It is seen that the proposed technique is almost never perceived as *much worse* or *worse* than the traditional DSM method. About 10% of opinions considered our model to be *slightly worse* while around 35% found both versions to have an equivalent quality. For the remaining 55%, the proposed technique was preferred over the traditional approach. Interestingly, our proposed model was perceived to be *better* or *much better* in 20% of cases for BDL and 30% for MV. The averaged CMOS scores with their 95% confidence intervals are respectively of 0.71 ± 0.14 for BDL and 0.88 ± 0.16 for MV, confirming the better creaky rendering using the proposed model.

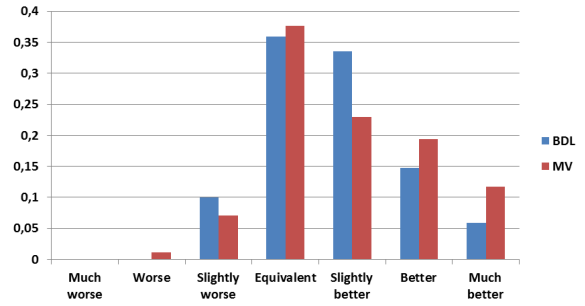

Figure 7: *Results of the CMOS test comparing the proposed model with the traditional one.*

6. Conclusion

This paper addressed the modeling of the residual signal in creaky voice. The first step focused on the analysis of the creaky excitation. This led to the development of specific tools for the detection of GCIs, for F_0 estimation as well as for the determination of the secondary excitation peaks. Some interesting observations were also made. Despite the pitch variability, it turned out that the open period kept a constant duration in creaky voice. Besides, the data-driven waveforms of the proposed model were extracted, showing the evolution of its deterministic component (whose closed period mostly consists of zero values) and of the energy envelope of the noise. In a second step, a vocoder incorporating the proposed model of the creaky excitation was built. It was compared to the traditional DSM vocoder through a subjective evaluation made of both an ABX and a CMOS test. The ABX results showed that the proposed model was found to be closer to the original speech signal in 60% of cases, against 15% for the traditional approach. Finally, the CMOS test emphasized the improvements brought by the proposed vocoder in which 20 to 30% of scores was perceived to be better or much better than the traditional approach, while the opposite occurred only in less than 1% of preference scores.

7. Acknowledgements

The authors would like to Martti Vainio for kindly providing us the MV databases. 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).

8. References

[1] Drugman, T, Dutoit, T., "The Deterministic plus Stochastic Model of the Residual Signal and its Applications", IEEE Trans. on Audio, Speech and Language Processing, 20(3), pp. 968-981, 2012.

[2] Cabral, J., Renals, S., Yamagishi, J., Richmond, K., "HMM-based speech synthesiser using the LF model of the glottal source", in Proc. of ICASSP, pp. 4704-4707, 2011.

[3] Raitio, T., Suni, A., Pulakka, H., Vainio, M., Alku, P., "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis", in Proc. of ICASSP, pp. 4564-4567, 2011.

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

[7] Moisik, S., Esling, J., "The 'whole larynx' approach to laryngeal features", in Proc. of ICPhS, pp. 1406-1409, 2011.

[8] 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.

[9] Carlson, R., Gustafson, K., Strangert, E., "Prosodic Cues for Hesitation," in Proc. of Fonetik 2006 , pp. 2124, 2006.

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

[11] Gobl, C, Ní Chasaide, A., "The role of voice quality in communicating emotion, mood and attitude," Speech Communication, pp. 189-212, 2003.

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

[13] Drugman, T., Kane, J. and Gobl, C. "Resonator-based Creaky Voice Detection", Accepted for Interspeech, 2012.

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

[15] Kane, J., Gobl, C.: "Evaluation of glottal closure instant detection in a range of voice qualities", submitted to Speech Communication.