Improved automatic detection of creak

John Kane[a,*], Thomas Drugman[b], Christer Gobl[a]

[a] *Phonetics and Speech Laboratory,*
School of Linguistic, Speech and Communication Sciences,
Trinity College Dublin, Ireland
[b] *TCTS Lab, University of Mons, Belgium*

Abstract

This paper describes a new algorithm for automatically detecting creak in speech signals. Detection is made by utilising two new acoustic parameters which are designed to characterise creaky excitations following previous evidence in the literature combined with new insights from observations in the current work. In particular the new method focuses on features in the Linear Prediction (LP) residual signal including the presence of secondary peaks as well as prominent impulse-like excitation peaks. These parameters are used as input features to a decision tree classifier for identifying creaky regions. The algorithm was evaluated on a range of read and conversational speech databases and was shown to clearly outperform the state-of-the-art. Further experiments involving degradations of the speech signal demonstrated robustness to both white and babble noise, providing better results than the state-of-the-art down to at least 20 dB signal to noise ratio.

Keywords:
Creak, Creaky voice, vocal fry, glottal source, glottal closure instant

1. Introduction

Several recent studies in the literature have been devoted to the development of improved methods for modelling aspects of the glottal source and excitation characteristics in speech (e.g., Drugman et al., 2009; Degottex

[*] Corresponding author. Tel. +353 1 896 1348
Email addresses: kanejo@tcd.ie (John Kane), thomas.drugman@umons.ac.be (Thomas Drugman), cegobl@tcd.ie (Christer Gobl)

et al., 2011; Kane and Gobl, 2011). Many of these aspects contribute significantly to the perception of voice quality. In this study we focus on one particular voice quality sometimes referred to as creak. Several voice quality labels such as glottal fry, vocal fry, laryngealisation or creaky voice are often used in the literature and are in the vicinity of each other in terms of their physiological and acoustic characteristics. For the present work we subsume these labels into one voice quality class, *creak*, which will be defined solely on the basis of the auditory criterion: 'a rough quality with the sensation of additional impulses' (as is done in Ishi et al., 2008a). This approach is justified as previous studies have demonstrated some of these voice quality variations to be perceptually similar (Gerratt and Kreiman, 2001). Laver (1980), however, makes the distinction between creak and creaky voice, stating that creaky voice is a compound of creak and modal voice[1]. We do not make this distinction in the present work and there is evidence to suggest that such a distinction is not utilised by speakers for any linguistic or paralinguistic contrast (Ishi et al., 2008a; Laver, 1994).

Although the voice quality class we are considering here is defined solely by the auditory criterion, it can be instructive to consider the physiological as well as the acoustic features typically associated with creak.

In Laver's review of the physiology (Laver, 1980), he states that creaky voice typically involves low levels of longitudinal vocal fold tension (probably the main physiological parameter utilised for pitch variation) and high levels of adductive tension (i.e. the muscular tension involved in bringing the vocal folds together). This is combined with low levels of subglottal pressure from the lungs (Laver, 1980). It has been claimed, however, that so-called 'irregular phonation' can sometimes involve low levels of adductive tension with incomplete vocal folds closure (Slifka, 2006).

Some further insights into the physiology involved in creak production were highlighted in Edmondson and Esling (2006) including the presence of *ventricular incursion*. Ventricular incursion is when the ventricular folds push down and cover the *true* vocal folds, causing an increased mass and, as a result, lowers the frequency of vibration (Moisik and Esling, 2011). This ventricular incursion can also result in secondary vibrations occurring above the glottis.

[1] We interpret modal voice, following Laver (1980), as the case of periodic vocal fold vibration, with full glottal closure and no audible frication.

Many of the resulting acoustic characteristics of creak are clearly distinct from modal voice. One of these features is the very long glottal pulse duration (where pulses can occasionally be as long as 100 ms, see Blomgren et al., 1998; Hollien and Wendahl, 1968). Such findings are corroborated by the results of psychoacoustic experiments carried out in Titze (1994), which demonstrated that human listeners begin to perceive individual pulses from around 70 Hz. Another acoustic feature often reported is the presence of secondary excitations, shown in electroglottographic (EGG) and speech pressure signals (Blomgren et al., 1998) as well as in voice source signals estimated by glottal inverse filtering (Gobl and Ní Chasaide, 1992). These secondary excitations may perhaps be explained by the occurrence of ventricular incursion, mentioned above. A further observation is that there is little or no superposition of formant oscillations between adjacent glottal pulses (Ishi et al., 2008a). One can frequently observe that oscillations from the vocal tract resonances have almost completely decayed before the start of the next pulse.

These distinctive acoustic characteristics can cause problems for standard speech analysis methods (including f_0 tracking and spectral analysis). The very low f_0 values and, at times, irregular temporal patterning may not be properly handled by standard f_0 tracking algorithms. Standard frame lengths (usually no longer than 32 ms) may be too short to capture two glottal pulses and, hence, will be unsuitable for obtaining strong periodicity information. Commonly used f_0 trackers tend to either output spurious values in creaky regions or consider creaky regions to be unvoiced. As a result of this, creaky regions will be poorly modelled in most speech technology applications. This problem was highlighted in a previous study (Silen et al., 2009) which involved the development of a speech synthesis system for Finnish (a language in which creak frequently occurs).

However, creak is commonly produced for a range of interactive, expressive and stylistic reasons. It has previously been studied in relation hesitations (Carlson et al., 2006) and turn-taking (Ogden, 2001), as well in the context of various forms of expression and emotion (Yanushevskaya et al., 2005; Gobl and Ní Chasaide, 2003; Ishi et al., 2008b). It has also been shown to frequently occur at phrase boundaries and utterance final position in American English (Surana and Slifka, 2006b). If low subglottal pressure is a necessary condition for creak, then it is not surprising that it frequently occurs in phrase/utterance/turn-final position, where the speaker will have less air available than at the start of the utterance. Creak has also recently received attention from popular science articles following the study in Wolk and Abdelli-Beruh (2012) which demonstrated that two thirds of the young female American English speakers analysed displayed creak at the end of read sentences. The authors state that continuous use of creak is likely to be more prevalent in more sociable, conversational speech settings (however this was not formally investigated).

The study of creak (and indeed voice quality in general) has been hindered because of the lack of suitable automatic detection algorithms and, as a result, most applied studies on creak tend to either rely on qualitative interpretation or use small amounts of data.

In terms of specific applications robust detection could be used to segment creaky regions in corpora used for text-to-speech synthesis which would facilitate the use of more appropriate acoustic modelling and, hence, better rendering of these regions (Drugman et al., 2012a). This would be particularly important for expressive or conversational speech synthesis as the use of creak (and indeed other aspects of voice quality) are known to play a critical role in spoken interaction (Campbell and Mokhtari, 2003), e.g., with turn-taking (Ogden, 2001). Also, as creak is known to be frequently produced during hesitations (Carlson et al., 2006) its detection could also be used to identify hesitations which could, in turn, be used for distinguishing speaking styles or, for instance, providing feedback on presentation skills. The robust automatic detection of creak would be beneficial for sociological studies (e.g., Wolk and Abdelli-Beruh, 2012) and studies on tonal patterns (Yu and Lam, 2011) in terms of allowing quantitative analysis on larger volumes of data. Furthermore, as studies have shown listeners to be sensitive to creak in terms of recognition of the speaker's identity (Böhm and Shattuck-Hufnagel, 2007), the detection of creak can be exploited for improving speaker recognition systems (Espy-Wilson et al., 2006; Elliot, 2002).

Motivated by this, the present work describes a new algorithm for automatically detecting creak through the use of two new acoustic parameters which describe aspects of the LP-residual signal. The approach builds on previous work by the same authors (Drugman et al., 2012b). This initial study involved the use of a single parameter which was evaluated on a rather small set of read text-to-speech (TTS) synthesis data. The current study involves the inclusion of a further acoustic parameter as well as the incorporation of the features into a classifier for detecting creak. A much larger evaluation is carried out here on a wide range of speech data, covering a variety of speakers, gender, languages, recording conditions and speaking styles. Furthermore, additional robustness experiments are conducted examining the effect of different noise types and levels on the performance of the different methods. Also, this manuscript provides a more thorough survey and analysis of the literature in relation to the production, acoustics and usage of creak. Finally, the present work includes more formal statistical comparisons of the evaluated methods.

The paper is organised as follows. First we present a review of the state-of-the-art (Section 2) and then move to a complete description of the proposed algorithm (Section 3). In Section 4 we describe the speech data used in our experimental setup (outlined in Section 5). We then give a presentation of the results on both 'clean' (Section 6) and degraded (Section 7) speech. Next we discuss the findings and finally provide a conclusion (Section 8).

2. State-of-the-art

Although a considerable amount of research has been carried out investigating the acoustic characteristics of creak there is a clear lack of algorithms for detecting it automatically. Some studies describe automatic detection of 'irregular phonation' (see e.g., Böhm et al., 2010; Surana and Slifka, 2006a; Vishnubhotla and Espy-Wilson, 2006), a class within which creak is contained. For instance in Böhm et al. (2010) the authors derive six acoustic parameters and use them as input to a support vector machine (SVM) based classification system. Their method involved using acoustic measurements from previous studies (i.e. Surana and Slifka, 2006a; Ishi et al., 2008a). In Surana and Slifka (2006a) the authors propose the use of acoustic parameters including normalised root mean squared amplitude and smoothed-energy-difference amplitude measures. However, misdetections apparently occur in low f_0 regions.

In the present study we include two creak detection algorithms from the literature (Ishi et al., 2008a; Vishnubhotla and Espy-Wilson, 2006) for comparison with the proposed algorithm. They are now described in detail.

2.1. Ishi's method for detection of vocal fry/creak

Recently an algorithm was presented for the automatic detection of vocal fry/creak (Ishi et al., 2008a) which builds on previous work by the same authors (Ishi, 2004; Ishi et al., 2005). This algorithm involves detecting candidate regions in a power contour measured from a bandlimited speech signal. Then a combination of autocorrelation and cross-correlation methods are used to discriminate creak from ‘normal’ voiced speech and unvoiced/silence regions, respectively. The full details of the algorithm are as follows.

The algorithm operates on the speech signal, which has been bandlimited to 100 - 1500 Hz. A ‘very short-term’ power contour is measured, with a frame length of 4 ms and shift of 2 ms, in order to highlight the amplitude variation within individual pulses (see Figure 1 panel b). Peaks are then detected in this contour and Power Peak (PwP) parameters are derived for each peak based on the previous (PwP-rising) and following (PwP-falling) 5 frames (i.e. 10 ms) in the contour. The maximum power difference in each direction is used as the PwP value and a threshold is applied to this parameter to determine whether the peak can be used as a creak candidate location. In addition to this, it has been suggested by the author (personal communication) that peaks more than 20 dB below the maximum power peak (for each utterance) can also be discarded.

Given these peak locations (creak candidates) a check is performed against a frame-synchronised periodicity strength measure in order to discriminate creaky regions from ‘normal’ voiced regions. The Intra-Frame Periodicity (IFP, see Figure 1, panel (c)) contour is calculated with:

$$\text{IFP} = \min\left\{\frac{N}{N-\tau}\cdot \text{autoCorr}(\tau);\quad \tau = j\cdot\tau_0;\quad j = 1,2,...\right\} \tag{1}$$

where N is the frame length (set to 32 ms), τ is the autocorrelation lag, autoCorr is the normalised autocorrelation function, and τ_0 is the lag of the strongest autocorrelation peak. Note also that the search space for τ is limited to 15 ms and that the factor $\frac{N}{N-\tau}$ is used to compensate for the decrease in amplitude with increasing τ in the autocorrelation function. ‘Normal’ voiced regions are expected to have an IFP value close to 1, while creaky (and non-voiced) regions are likely to show a value closer to 0. This is based on the observation that creaky regions can display irregular temporal patterns and also, that even in cases where creaky regions display a reasonable amount of periodicity, the very long pulses will mean that the frame is not sufficiently long to capture strong periodicity information. Finally, IFP values are set to 0 unless three successive frames are found to be above a given threshold.

Next an Inter-Pulse Similarity (IPS, see Figure 1, panel (d)) measure is calculated with:

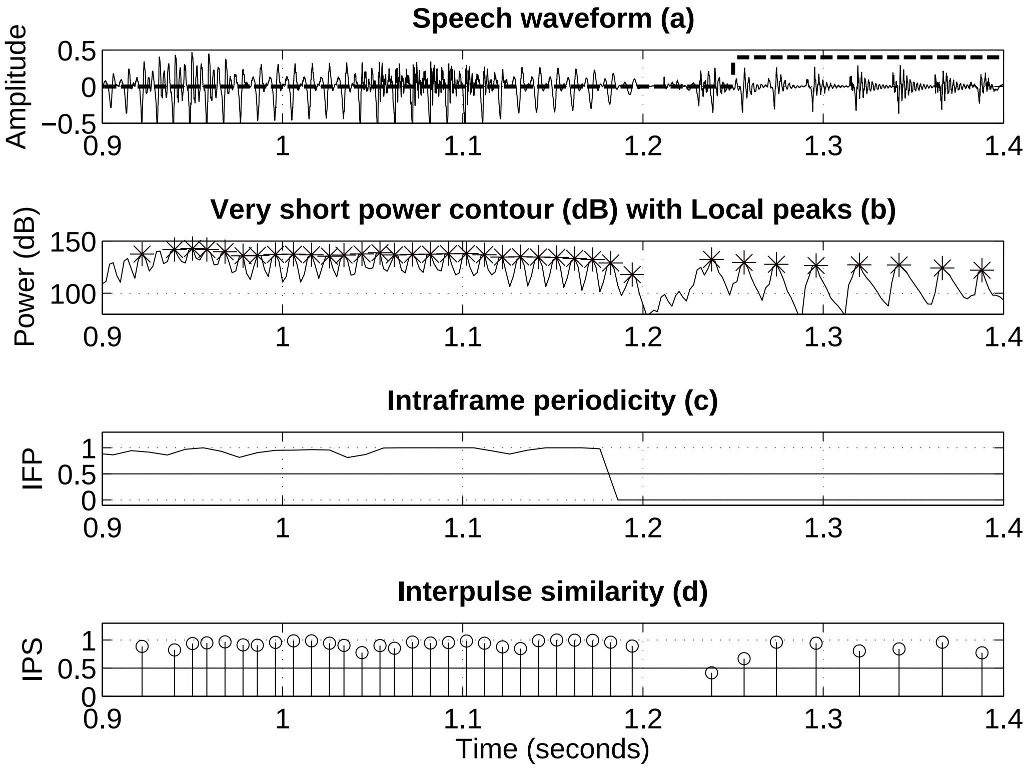

Figure 1: *Illustration of creak detection using the method proposed in Ishi et al. (2008a). The speech waveform, with binary creak decision (dashed line), is shown with the creaky region beginning from around 1.22 seconds (panel a), along with the very short term power contour and detected local peaks (panel b), the Intraframe Periodicity (IFP) contour (panel c) and the Interpulse Similarity (IPS) values (panel d). Horizontal lines for IFP and IPS are given to illustrate the thresholds used.*

$$\mathrm{IPS} = \max\left\{\mathrm{CCorr}(F_{\tau_1}, F_{\tau_2}); \quad \tau_1 - \tau_2 < T_{max}\right\} \tag{2}$$

where CCorr is the cross-correlation function, F_{τ_1} and F_{τ_2} are the frames centred on successive candidate peak locations, and T_{max} is the maximum allowed distance between adjacent peaks, and is set to 100 ms. Each frame is selected as the range of 5 ms around the peak location. Adjacent creak pulses are expected to display a reasonably high similarity (as the vocal tract setting is not expected to have significantly changed) and, hence, IPS values are expected to be high (e.g., above 0.5). Non-speech and unvoiced regions, on the other hand, are expected to display low levels of similarity and, hence, and IPS close to 0. If adjacent pulses are too far apart, i.e. above T_{max}, IPS values are also set to 0.

We use the optimal thresholds suggested in the original publication (Ishi et al., 2008a), i.e. $\{\text{PwP} \geq 7 \text{ dB} \ \& \ \text{IFP} \leq 0.5 \ \& \ \text{IPS} \geq 0.5\}$ for a peak to be considered to be creaky (however, different thresholds have also been applied in a separate applied study, Ishi et al., 2008b). The binary creak decision is then made by merging regions between detected creak peaks (see Figure 1, panel (a)). This method is given the label **Ishi Orig.** throughout this paper.

In the original publication (Ishi et al., 2008a) the authors report an upper bound detection rate of 74 % with a false alarm rate of 10 %, using thresholds optimised on the same dataset. Note that frame level results were not reported.

2.2. Extension of the Aperiodicity, Periodicity and Pitch (APP) detector

This method has been proposed in Vishnubhotla and Espy-Wilson (2006) for the automatic detection of 'irregular phonation' (a term they say they use interchangeably with creak). The authors interpret this label also to include sounds referred to as vocal fry, diplophonia, glottalisation, laryngealisation, pulse register phonation and glottal squeak (Vishnubhotla and Espy-Wilson, 2006).

As part of the algorithm they make use of the Aperiodicity, Periodicity and Pitch (APP) detector (originally presented in Deshmukh et al., 2005). This involves applying a gamma-tone filterbank to decompose the speech signal into 60 frequency bands. An Average Magnitude Difference Function (AMDF) is calculated on the separated frequency bands (smoothed by the use of the Hilbert envelope) to determine aperiodicity/periodicity in the signal. The AMDF function, $\gamma_n(k)$, for each outputted signal is calculated with:

$$\gamma_n(k) = \sum_{m=-\infty}^{\infty} |x(n+m)w(m) - x(n+m-k)w(m-k)| \tag{3}$$

where $w(m)$ is a rectangular window centered on n and given a specified width. When the signal is periodic the AMDF function will display pronounced 'dips' when k (the lags of the function) is equal to integer multiples of the fundamental period. In the implementation of the algorithm we obtained, the frame length was set to 25 ms, with a shift of 2.5 ms.

Next, ‘irregular phonation’ is differentiated from aperiodic frames, breathy vowels and voiced fricatives using the so-called dip profile of their AMDF in various frequency bands. The dip profiles for ‘irregular phonation’ display distinctive clustering characteristics to both regular phonation and speech with turbulent excitations. Identification of ‘irregular phonation’ is on the basis of detection of this characteristic.

Finally, the problem of false positives in some stops is addressed by calculating the spectral slope. This is done by fitting a regression line to the amplitude spectrum from 2 to 4 kHz. A threshold of -0.05 is empirically used to differentiate ‘irregular phonation’ from these stops. In our evaluation, we used the original implementation of the algorithm kindly shared by the authors. The method is given the label **Vishnu**.

3. Proposed method

The current section presents a description of the proposed algorithm for automatically detecting creaky regions in a speech signal. The method is based on further development of the algorithm described in Drugman et al. (2012b). First an analysis of the excitation characteristics of creak is carried out which highlights the main acoustic features which our algorithm is designed to detect. Then follows a description of the two components of the algorithm, both of which are used as input features to a binary classifier (see Section 3.4). Note that the two components attempt to describe different aspects of the LP-residual signal in creaky regions. These two aspects may, at times, both be present. At other times, however, just one is displayed.

Figure 2 provides a summary of the proposed creak detection method, including feature extraction, classification and post-processing stages.

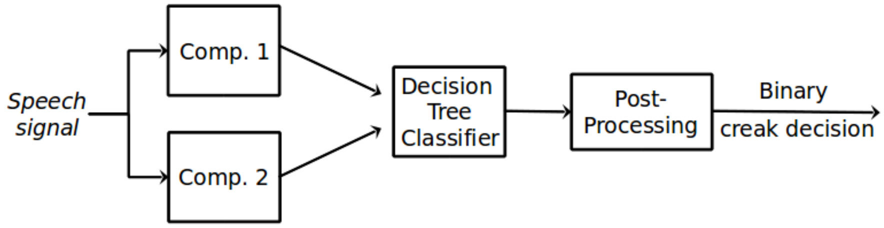

Figure 2: *Block diagram of the proposed creak detection method.*

3.1. Excitation characteristics

Following extensive qualitative analysis of the Linear Prediction (LP) residual signal (obtained by LPC analysis and subsequent inverse filtering) we observed some distinctive characteristics which appeared to be closely associated with creaky regions. The first observation was the presence of secondary (or even tertiary) peaks proceeding the main excitation peak which corresponds to the glottal closure instant (GCI, Drugman et al., 2012c). This is illustrated in Figure 3, where strong residual peaks can be observed before the residual peak which corresponds to the GCI (as shown by the derivative of the EGG signal). Although secondary excitations and double-pulsing have frequently been reported in the literature (e.g., Blomgren et al., 1998; Gobl and Ní Chasaide, 1992) these secondary residual peaks frequently did not appear to correspond to secondary laryngeal excitations (which would show up in the EGG signal). Instead, we hypothesise that often these peaks correspond to an abrupt glottal opening following a long closed phase. Some preliminary comparison with glottal source signals, estimated by inverse filtering, and EGG signals appeared to support this. Nevertheless, strong secondary laryngeal excitations did at times cause secondary peaks to occur in the LP-residual signal. Following these observations Component 1 of the algorithm is designed to exploit these secondary peaks occurring in the LP-residual signal.

Further analysis of the LP-residual of creaky speech regions revealed that although the above trend is very prevalent, secondary LP-residual peaks may sometimes be absent. Considering the LP-residual signal in Figure 4 one can observe strong impulse-like peaks with no secondary peaks, even when the EGG derivative is displaying small secondary peaks.

Component 2 of the algorithm is, hence, designed to capture this specific feature combined with the knowledge that creaky voicing produces considerably longer glottal pulse duration (Titze and Sundberg, 1992; Blomgren et al., 1998).

3.2. Component 1: Detection of secondary excitation peaks

The block diagram of Component 1 of the proposed algorithm is shown in Figure 5. This method is designed to exploit the secondary peaks in the residual excitation signal. The residual signal is estimated following LPC analysis. The aim of this analysis is to cancel the spectral contribution of both the vocal tract filter and the glottal source signal, thereby rendering the spectrum of the LP-residual essentially flat. However, this residual signal

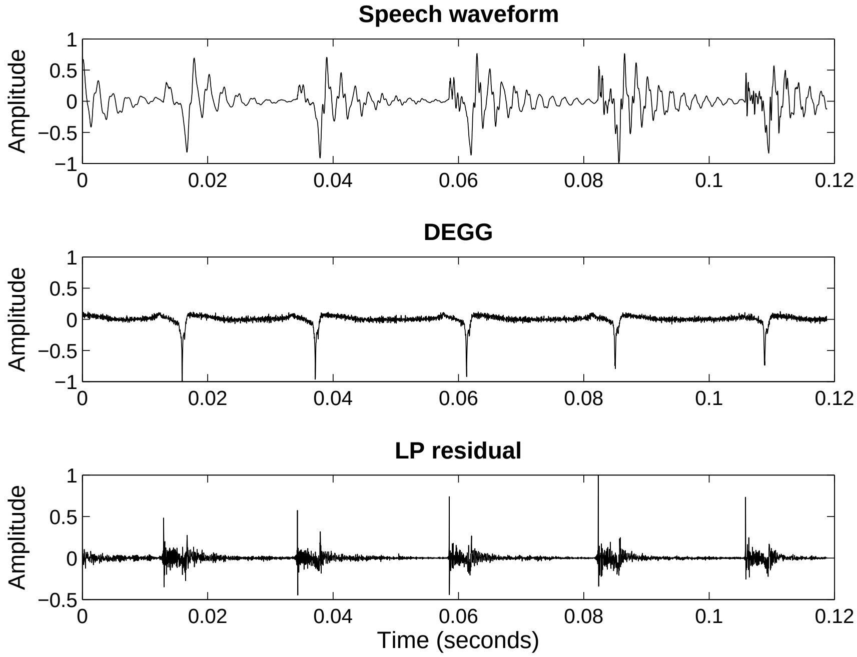

Figure 3: *Speech waveform (top panel), EGG derivative (middle panel) and LP-residual (bottom panel) signals from a creaky region in an utterance produced by a male speaker.*

exhibits important phase properties of the excitation source, including the presence of secondary peaks. The key idea of Component 1 is that when applying a resonator to the LP-residual, secondary peaks will perturb its output and produce a greater amount of harmonics. The full details of this method are as follows.

The LP-residual is obtained by LPC analysis of order $(Fs/1000) + 2$ [2], where Fs is the sampling frequency. Two separate resonators are then applied to the residual signal for different purposes. One is used for getting a more robust estimate of the f_0 contour in creaky regions while the other is to highlight the presence of secondary residual peaks. Both resonators

[2] This LPC order is chosen in order to obtain a spectral envelope which is not biased to harmonics and is frequently used in the literature. Although it is widely known that LPC analysis becomes biased towards harmonics for high f_0 values (Villavicencio et al., 2006; Kay, 1988), creaky regions display a low f_0 and here we have found LPC analysis to be suitable.

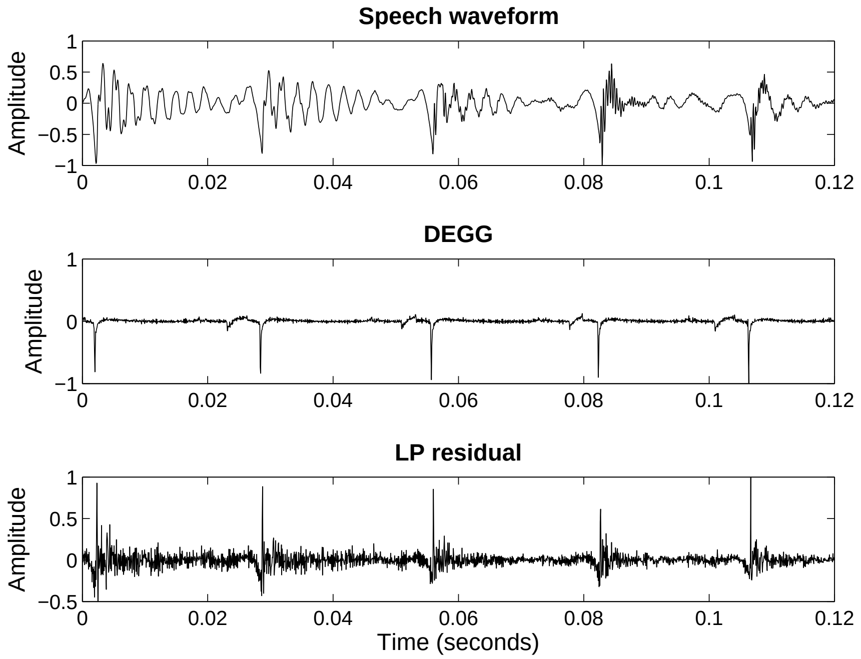

Figure 4: *Speech waveform (top panel), EGG derivative (middle panel) and LP-residual (bottom panel) signals from a creaky region in an utterance produced by a male speaker.*

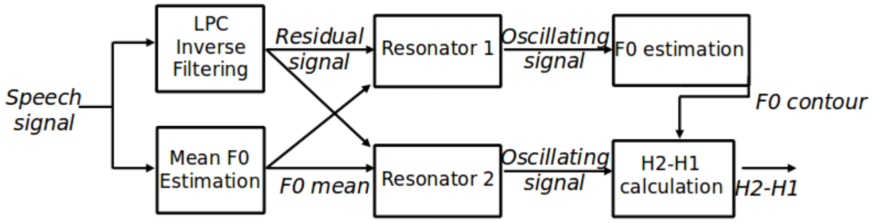

Figure 5: *Workflow of Component 1 of the proposed technique. The H2-H1 parameter is derived from the response of two resonators excited by the LP-residual signal. Both resonators are centred on* $f_{0,mean}$ *but use different damping factors. For details see text.*

use a centre frequency of $f_{0,mean}$, the mean f_0 of the speaker. This value is here estimated using the Summation of Residual Harmonics (SRH) method (Drugman and Alwan, 2011), although the choice of algorithm for this is not critical. The z-transform $H(z)$ of these resonators, which are characterised by two complex conjugate poles of modulus ρ and of angle $\pm\frac{2\pi f_{0,mean}}{F_s}$, can be written as:

$$H(z) = \frac{1}{1 - 2\rho\cos(\frac{2\pi f_{0,mean}}{F_s})z^{-1} + \rho^2 z^{-2}} \tag{4}$$

Resonator 1 is used for its ability to estimate the f_0 contour, even in creaky regions. It uses poles with a modulus $\rho = 0.8$, which provides a compromise between avoiding ambiguity with octave jumps and capturing the spread of f_0 values from $f_{0,mean}$ (as will inevitably happen when there are creaky regions). To estimate f_0 a 50 ms-long Hanning window is applied to the resonator output (note that a longer than usual window is required due to the longer glottal pulses present in creaky regions). From this, the corrected autocorrelation function $r'(\tau)$ is calculated:

$$r'(\tau) = \frac{N}{N-\tau} \cdot \text{autoCorr}(\tau) \tag{5}$$

where N is the window length (in samples) and τ is the number of autocorrelation lags. A correction of $\frac{N}{N-\tau}$ is applied to compensate for the decreasing properties of the autocorrelation function with increasing τ (as is used in Ishi et al., 2008a). The local fundamental period is then considered as the position of the maximum value in $r'(\tau)$ above the peak centred on $\tau = 0$.

For Resonator 2, poles with a modulus $\rho = 0.97$ are applied in order to produce a more pronounced resonating character necessary to highlight the presence of secondary residual peaks. Again a 50 ms Hanning window is applied to the output of Resonator 2, before calculating the spectrum of the autocorrelation function (which enhances the harmonic peaks). Then using the local f_0 value, obtained using Resonator 1, the difference between the second and first harmonics (H2-H1) in dB is measured. The parameter contour is then smoothed using a 100 ms-long moving average filter in order to lessen the impact of outlying values. An example of the contour is shown for a speech utterance in Figure 6 where it is clear that H2-H1 reaches much higher values in the annotated creak region at the end of the speech segment, and for which applying a threshold (of around 0 dB in this case) would lead to good detection results.

An illustration of the steps involved in the workflow (depicted in Figure 5) is given in Figure 7 for both a segment of speech involving modal phonation (on the left) and creak (on the right). The speech signal is displayed in

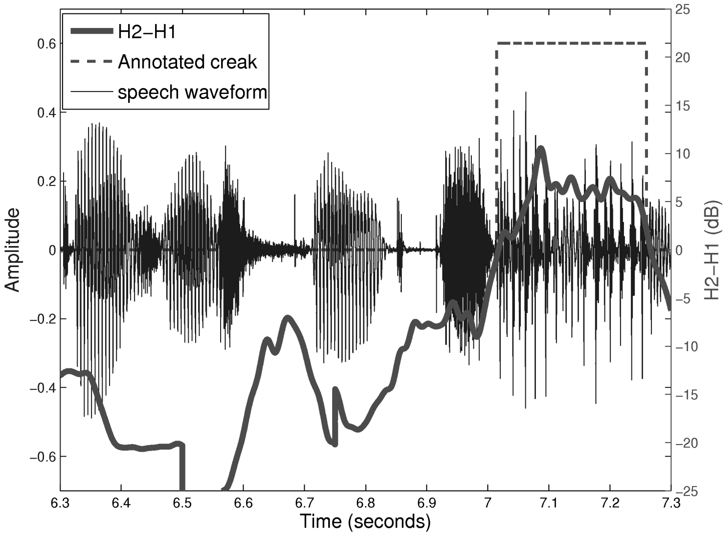

Figure 6: *Illustration of the H2-H1 contour (thick line) for a speech segment with annotated creak region (dashed line). The left y-axis shows the scale for the amplitude of the speech signal, whereas the right y-axis shows the scale for the H2-H1 contour in dB.*

the top row plots. In the middle row, the residual signal and the output of Resonator 2 are represented.

In the case of modal phonation, it can be noted that the residual signal exhibits a regular structure with major peaks only at the GCI positions. As a result, perturbations between two major excitation peaks are relatively weak and the oscillating signal outputted by Resonator 2 will only contain a small amount of harmonics. This is reflected in its amplitude spectrum (in dB) in the bottom row of Figure 7 where the level at f_0 is much higher than for the harmonic at $2 \cdot f_0$. The difference H2-H1 in such a modal phonation then reaches low negative values (-15 dB in the case of Figure 7).

On the other hand, for creak, secondary pulses significantly re-excite the resonator between two consecutive GCIs, leading to perturbations in its output. This effect is seen in the corresponding amplitude spectrum which displays a greater richness of harmonics. More specifically, the level at the harmonic $2 \cdot f_0$ is much higher compared to the modal phonation case, and can even overtake the level at f_0. As a consequence, the presence of secondary peaks in the excitation of creak is reflected by a higher harmonicity in the output of Resonator 2. This leads therefore to higher values of H2-H1 (9.22 dB for the creaky example of Figure 7) compared to what is obtained for modal phonation.

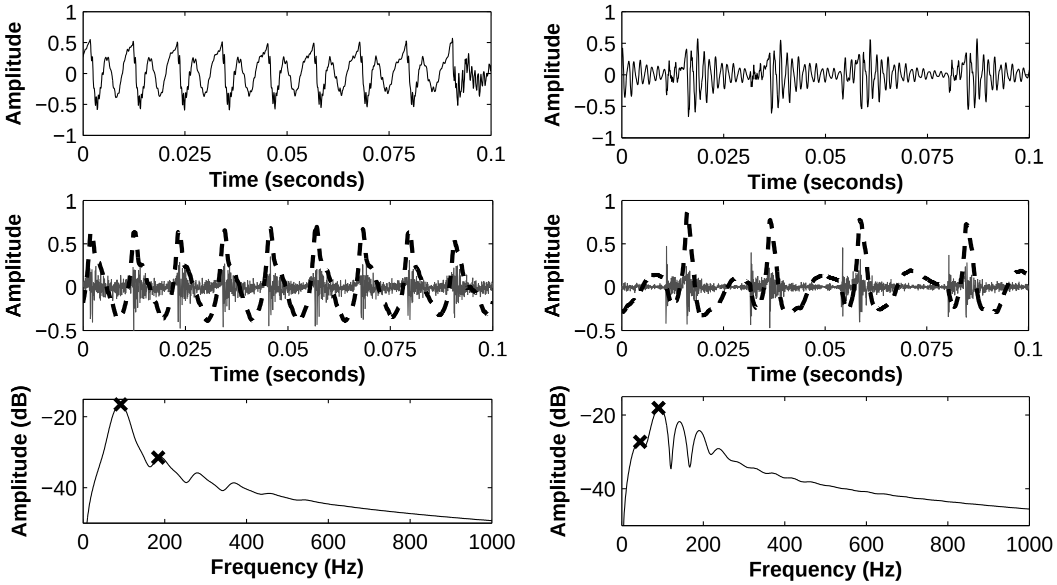

Figure 7: *Example of modal phonation (left column) and creak (right column) uttered by the same speaker, with the speech waveform (top row), the LPC residual signal (middle row, solid line) together with the output of Resonator 2 (in dashed line), and the amplitude spectrum (in dB) of a frame of the output of Resonator 2 where the values for f_0 and $2 \cdot f_0$ are indicated by crosses (bottom row).*

3.3. Component 2: Residual peak prominence

Component 2 of the algorithm is designed to detect creaky speech regions where there are prominent residual peaks (as shown in Figure 4). The prominent residual peaks may stem from the sharp vocal fold closure resulting from high levels of adductive tension (Laver, 1980). This is combined with the knowledge that creaky regions contain very long glottal pulses (Blomgren et al., 1998; Gobl and Ní Chasaide, 1992). The very long glottal pulses may be due to ventricular incursion (Edmondson and Esling, 2006) where the ventricular folds push down on the *true* vocal folds, causing an increased mass which vibrates at a lower frequency. Creaky regions can, at times, display irregular temporal patterning which can render frequency domain methods unsuitable. It follows that this second component of the algorithm does not rely on signal information to do with periodicity but instead looks to characterise individual pulses in the time domain. The method is carried out as follows.

Initially we measured residual peak prominence directly from the LP-residual signal, but further examination revealed this approach to be rather sensitive to additive noise. Instead, the output of Resonator 1 (see Figure 5) was used. The setting of the pole modulus to $\rho = 0.8$ is suitable for highlighting the prominence of the main residual peaks without being overly biased towards secondary peaks (if present).

The method operates on a fixed, non-overlapping frame basis using a rectangular window and with a frame-length of 30 ms. This roughly corresponds to two periods at 70 Hz. In this method correct polarity of the speech signal is assumed (this can be determined automatically for example using the method described in Drugman and Dutoit, 2011) and the output of the resonator will display strong negative peaks. We invert this signal so that it displays strong positive peaks, corresponding to positive peaks in the LP-residual.

For each frame the absolute maximum peak in the resonator output is identified and the frame is then shifted to be centred on this peak. Considering Figure 8 one can observe for a creaky region (right panel) the prominent peak amplitude in the centre of the frame. For a modal region (left panel), however, peaks from neighbouring glottal pulses are captured within the frame length.

By measuring the amplitude difference between the maximum peak (in the centre of the frame) and the next strongest peak one can obtain a parameter value which differentiates modal and creaky regions. In order to avoid selecting values in the vicinity of the main centre peak, the search for the next strongest peak is made outside a distance of 6 ms of the centre of the frame. This corresponds to 40 % of half the frame length which ensures that there is sufficient space for peaks to occur from neighbouring glottal pulses. A value is thus obtained for each frame producing the outputted parameter contour. This contour was then filtered with a 3-point median filter to remove misdetections due to transients in the signal.

3.4. Classification using the two parameters

In order to detect creaky regions we used the two new parameters as input features to a binary decision tree classifier (Breiman et al., 1984). The separation of the two classes is done using a top-down approach where both

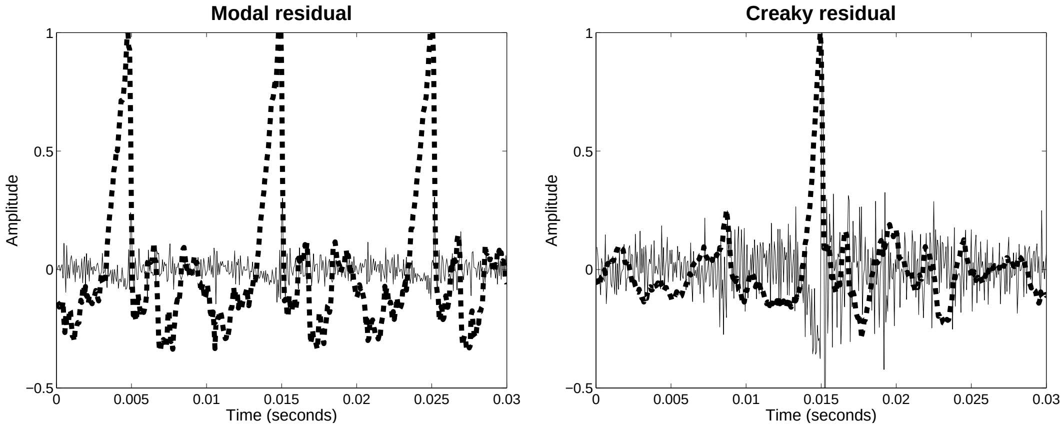

Figure 8: *A 30 ms peak centred residual frame (thin line), with superimposed Resonator 1 output (thick dashed line) for a modal (left panel) and a creaky (right panel) utterance segment. The residual peak prominence value for the modal segment is very close to 0, while the value for the creaky segment is over 0.75. Both LP-residual and resonator output signals are normalised in amplitude for clarity.*

classes are initially placed at the root node and then a series of binary questions are asked (to do with the input features) and for each question a new child node is created. This creates the decision tree, the ends of which are leaf nodes. The commonly used Gini's Diversity Index (GDI) was used for the splitting criterion and splits are selected in order to reduce the GDI criterion. The splitting was stopped if the current node was *pure* (i.e. contained only observations of a single class) or if the branch node contains fewer than 10 observations.

Decision trees were developed on a training set using the extracted parameter values and the binary creak annotation labels (see Section 5.1). When inputting a training example to the trained classifier the output is the posterior probability, P_1, of the example corresponding to class 1 (i.e. creaky) and the posterior probability, P_0, of it corresponding to the class 0 (i.e. non-creak). The standard binary decision is typically set to 1 if $P_1 > 0.5$, and otherwise 0. In the training phase the decision tree classifier is optimised for minimising the error rate. However, for skewed datasets which contain a given class to be detected which displays sparse occurrence (e.g., creak or laughter) this error criterion is not suitable. To address this a further processing was carried out during the training phase. This involved making the decision strategy such that if $P_1 > \alpha$ then the output is set to 1, otherwise it is set to zero. α is varied in the range [0, 1] and the setting which produced the higher F1 score on the training set is subsequently used for the decision strategy in testing. Note that the F1 score is a more suitable criterion for a dataset of the type used in this study (see Section 5.2).

In order to illustrate the characteristics of the two parameters learnt by the training, a decision tree was trained using all the data described below in Section 4. This exemplary decision tree was purposely set to have a low complexity in order to clearly highlight the questions asked at each node. x1 refers to Comp. 1 and x2 refers to Comp. 2. The first few nodes of the decision tree are listed:

```
1  if x1<0.362994 then node 5 elseif x1>=0.362994 then node 2
2  if x2<0.000172 then node 3 elseif x2>=0.000172 then node 4
3  class ‘non-creak’
4  class ‘creak’
5  if x2<0.430789 then node 7 elseif x2>=0.430789 then node 6
6  class ‘creak’
7  ....
```

It can be seen that if x1/Comp. 1 is greater than 0.36 and if x2/Comp. 2 is above close to 0 then a positive creak decision is made. Also, if x1/Comp. 1 is less than 0.36 but x2/Comp. 2 is greater than 0.43 a positive creak decision is arrived at. The learning displayed by the classifier here clearly shows intuitive decision making considering the descriptions of the parameters given in Section 3.2 and 3.3.

In this study four methods involved the use of the classification approach described above. The first component (labelled **Comp. 1**) and the second component (labelled **Comp. 2**) of the proposed method were used separately and in combination (labelled **Comp. 1 & 2**). We also included the four parameters derived using the method by Ishi et al. (2008a) and described in Section 2.1, i.e. PwP-rising, PwP-falling, IFP and IPS (this method was labelled **Ishi Opt**).

3.5. Post-processing

Some final post-processing can then be carried out on the binary decision of the proposed (or other) methods. To help remove misdetections in unvoiced and non-speech areas, zero-crossings are measured on 20 ms frames.

Areas with a zero crossing rate (ZCR, i.e. number of zero-crossings per ms) of more than 5, were considered to be unvoiced or silent parts and therefore excluded as potential creak areas. Note that the use of energy contours was deemed unsuitable particularly because conversational speech data can display widely varying energy values. This is, of course, a rather basic method for determining unvoiced regions and could be substituted with a more sophisticated method (see for example Ghosh et al., 2011).

Finally, overly short detected creak regions were removed and nearby adjacent creak regions were merged. We used a minimum creak length of 30 ms which corresponds roughly to two periods at 70 Hz. The assumption here is that at least two pulses are required for the perception of creak and, again, that the perception of individual glottal pulses starts around 70 Hz (Titze, 1994). The binary creak decision vectors used in this study were sampled every 10 ms, so the removal of short regions and merging of close regions was done in the same operation through the use of a 5-point (i.e. 50 ms) median filter applied to the binary decision. For instance this would remove one or two positive creak samples (i.e. ones) surrounded by negative decision samples (i.e. zeros). On the other hand, if a zero had two positive decision samples on either side, the median filter would merge the two regions and convert the zero to a one.

4. Speech material

In order to provide a thorough evaluation of detection performance we decided to use a wide range of speech databases which cover gender, language, read/conversational speech and recording condition variations. All speech data were downsampled to 16 kHz. A summary of the speech data used in the evaluation in the present study is given in Table 1.

4.1. Text-To-Speech databases

To evaluate performance in *ideal* recording conditions we selected 100 sentences containing creaky regions from three Text-To-Speech (TTS) databases. An American male speaker (BDL) was selected from the ARCTIC database (Kominek and Black, 2004). We also used utterances from a Finnish male speaker (MV, as was used in Vainio, 2001) and from a Finnish female speaker (HS, as was used in Silen et al., 2009).

4.2. Spontal corpus

Next we wished to include conversational speech data recorded in high quality conditions. The Spontal corpus (described in Edlund et al., 2010) contains audio, video and motion capture data from a large number of dialogues lasting at least 30 minutes, carried out in a recording studio. The dialogues were in Swedish and participants were encouraged to talk about whatever topic they wished.

We selected the audio data from the microphone channel of one male (label: 09-13-02) and one female speaker (label: 09-03-01) who were deemed to produce frequent creaky utterances. Audio was captured through the use of two microphones per speaker: a Brüel & Kjær 4003 omni-directional gooseneck at 1m distance, and a head-mounted Beyerdynamic Opus 54 cardioid which was used to obtain optimal recording quality. In the current study only the audio channels from the head-mounted microphone were used. The original sampling rate was 48 kHz.

4.3. American conversational data

We selected audio streams from two male and two female American English speakers who engaged in conversations on the topic of food. The speech data were part of a larger number of conversational recordings, additional to the data used in Yuasa (2010). For each speaker the audio was of approximately 10 minutes in duration and was captured using a headset resting around the speaker's neck with the microphone pointing to the mouth. Conversations were carried out in a booth of a media center, a relatively quiet but relaxing environment suited for natural conversations.

4.4. Japanese conversational data

We also included the audio recordings of two female Japanese speakers, previously described in Magnuson (2011). Both speakers spoke a Japanese dialect spoken in Western Japan and engaged in a 30 minute conversation. The speakers were shown some short animated films before starting the conversation. For the conversation itself they were encouraged to talk about any topic they wished but to refer to the short film at some stage during the conversation. Audio was recorded on AKG C420 III PP MicroMic headset microphones wired through a BeachTek DXA-2S pre-amp connected to the video camera (Sony DCR-TRV38 Mini DV camera). WAV files were extracted from the video into separate channels.

Table 1: *Summary of the speech data used for evaluating creak detection performance.*

Database	ID	Gender	Country	Conditions	Speech	Duration
TTS	US-M	Male	USA	Studio	Read	100 Sentences
	Fin-F	Female	Finland	Studio	Read	100 Sentences
	Fin-M	Male	Finland	Studio	Read	100 Sentences
Spontal	F	Female	Sweden	Studio	Conversation	30+ Minutes
	M	Male	Sweden	Studio	Conversation	30+ Minutes
US	F1	Female	USA	Quiet room	Conversation	10+ Minutes
	M1	Female	USA	Quiet room	Conversation	10+ Minutes
	M2	Male	USA	Quiet room	Conversation	10+ Minutes
	F2	Male	USA	Quiet room	Conversation	10+ Minutes
Japan	F1	Female	Japan	Quiet room	Conversation	30+ Minutes
	F2	Female	Japan	Quiet room	Conversation	30+ Minutes

5. Experimental protocol

For the experiments conducted as part of this study we detected creak regions using the proposed algorithm (Section 3) as well as Ishi's algorithm (Section 2.1) and the algorithm by Vishnubhotla (Section 2.2). The experimental setup is now described in full.

5.1. Human annotation

Unfortunately there is no obvious way of obtaining an automatic reference for creaky regions in speech. Furthermore, the only relatively large database of speech data labelled for creak was the database used in Ishi et al. (2008a) and was carried out by a single person. As a result, in order to have a reference to evaluate detection performance we needed to carry out human annotations of the speech data. At the same time we wanted to evaluate performance of the algorithms on a large set of data covering a range of different speaking styles, languages, recording conditions etc. As the manual annotation of this volume of data is both tedious and time-consuming, annotation was shared between the first two authors. Both annotators strictly followed the annotation procedure outlined in Ishi et al. (2008a). Ultimately the binary decision on the presence of creak was based on the auditory criterion “a rough quality with the additional sensation of repeating impulses”.

However, the annotation was also guided through the use of spectrograms and f_0 contours. Wideband spectrograms typically display vertical striations (Ogden, 2009) and f_0 contours can frequently display spurious values or disappear (i.e. are considered unvoiced) and, hence, these displays were used to help guide the annotation.

Furthermore, manual voice activity segmentation was carried out for the speech data containing conversational speech. The exception to this was in the Spontal corpus where automatic voice activity detection was carried out with the algorithm proposed in Heldner et al. (2011). Finally, the long audio signals from conversational data were split into 5 second segments for analysis.

We calculated the percentage of time speaking which was annotated as being creak for the 11 speakers used in the evaluation and the average was 6.69 % with a range of 3.55 - 10.53 %.

5.2. Evaluation metrics

To assess the performance of the algorithms we calculated evaluation metrics at both the event level and the frame level. For the event level we used the metrics: hit (i.e. some part of a reference creak region was correctly detected), miss (i.e. for a reference creak region no positive detection was made) and false alarm (i.e. within a detected creak region there was no reference creak).

On the frame level we use the standard metrics True Positive Rate (TPR, also known as recall):

$$TPR = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \tag{6}$$

and False Positive Rate (FPR):

$$FPR = \frac{\text{False positives}}{\text{False positives} + \text{True negatives}} \tag{7}$$

We also use the F1 score which combines true positives, false positives and false negatives into one single metric. This metric is particularly useful when analysing skewed datasets where the feature being considered has a rather sparse occurrence (e.g., for laughter detection, Scherer et al., 2009), and is therefore well suited for assessing the performance of creak detection techniques. The metric is bound between 0 and 1, with 1 indicating perfect detection:

$$F1 = \frac{2 \cdot \text{True positives}}{2 \cdot \text{True positives} + \text{false positives} + \text{false negatives}} \in [0, 1] \tag{8}$$

5.3. Experiments on clean speech

In order to evaluate the detection performance of the various methods, analysis was carried out on the speech databases described in Section 4. For the methods using the decision tree classifier (i.e. Comp. 1, Comp. 2, Comp. 1 & 2 and Ishi Opt.), a leave one speaker out design was used whereby the speech data of a given speaker was held out for testing and the remainder of the speech data was used for training the classifier and optimising the decision strategy (see Section 3.4). The procedure was repeated for each speaker. For the methods Vishnu. (Vishnubhotla and Espy-Wilson, 2006) and Ishi Orig. (Ishi et al., 2008a) the same settings as were described in the original publications were used for all speakers.

There were three main aims of the experiments on clean speech:

1. A preliminary assessment of the suitability of each of the classification methods for detecting creaky regions.
2. To examine the effect of the post-processing (described in Section 3.5) on the methods, with analysis of the entire speech dataset.
3. To provide a thorough presentation of the performance of the difference methods using both event and frame level metrics.

5.4. Robustness to additive noise

In addition to examining detection performance on a range of speech databases we also carried out experiments to analyse the robustness of the algorithms to additive noise. For this we only used the TTS databases which were recorded in high-quality conditions.

Degraded conditions were simulated by adding noise to the original speech waveform at various signal-to-noise ratio (SNR) settings. Both white noise and babble noise (also known as cocktail party noise) were considered. The noise signals were taken from the Noisex-92 database (Varga and Steeneken, 1993), and were added so as to control the overall SNR without silence removal. All parameters from the different methods were extracted from speech with the various types and levels of noise added. Then a similar leave one speaker out approach, as was used described in Section 5.3, was applied.

Here, however, optimisation of the classification was done on a 'clean' training set, whereas testing was carried out on a test set with noise added. The F1 score was calculated for each test set (i.e. each speaker) for the various noise types/levels. Note that for these robustness testing experiments the zero-crossing rate (ZCR) feature used as part of the post-processing was omitted. This is because additive noise is likely to severely affect the rate of zero-crossings and, hence, this will hinder the assessment of the detection methods themselves.

6. Results on 'clean' data

6.1. Preliminary results on TTS database

An illustration of the performance of the six detection methods on the TTS databases is shown in Figure 9. The F1 scores for Comp 1. and Comp 2. are comparatively high, and the synergic effect of their combination is clearly apparent with an important improvement over the individual components for each of the three speakers. The prevalence of the two measures can be observed here and although Comp. 1 is more prevalent, Comp. 2 still provides better detection than the other comparison methods for two of the three speakers. Furthermore, as it is clear that their combination improves the detection performance, only the combination of the two (i.e. Comp 1 & 2) will be considered for the remainder of the study.

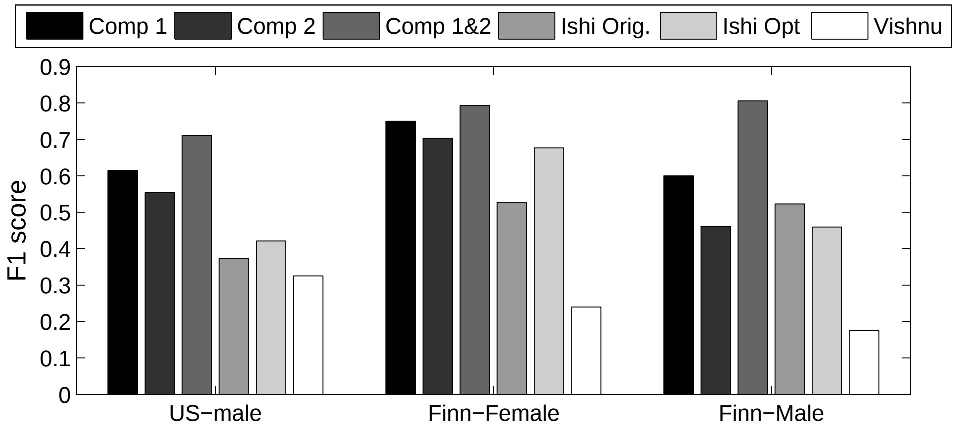

Figure 9: *F1 scores for the six different detection methods on the TTS database.*

Ishi's algorithm shows strong performance on the Finnish female speaker, with an improvement over the original algorithm when the parameters are used as inputs to the decision tree classifier. A small improvement is also seen for the American male speaker and a slight reduction for the Finnish male.

The algorithm described in Vishnubhotla and Espy-Wilson (2006) (labelled Vishnu.) displayed relatively low performance compared to the other algorithms, with a consistently low F1 score. This was found to be largely due to an high number of false alarms. In the original study (Vishnubhotla and Espy-Wilson, 2006) the authors use the terms 'irregular phonation' and creak interchangeably, but upon examination of the false alarms we found that a broader class of 'irregular phonation' types were detected, many of which did not match the auditory criterion used in this study and in Ishi et al. (2008a). As a result we decided to exclude this algorithm from the remainder of the analysis as the grounds for comparison seemed tenuous.

6.2. Effect of post-processing

The effect of the post-processing step on the event level metrics, summed across all speakers, on the three detection methods is shown in Table 2. Although an increase in misses can be observed, there is, nevertheless, a substantial reduction (in the region of 50 %) in false alarms for the three methods. It is clear that the use of the decision tree classifier with the parameters from Ishi et al. (2008a) as input features (i.e. Ishi Opt.) causes both an increase in hits but also a substantial increase in false alarms. The post-processing step considerably reduces the number of false alarms, however Ishi Opt. still displays more false alarms than the other two methods. As will be seen in Section 6.3 a large proportion of these false alarms come from the Finnish male speaker in the TTS database, where Ishi Opt. frequently produces misdetections in low pitch, non-creak, voiced segments.

A slight improvement of F1 score has been noticed for all techniques. The F1 score for Comp 1 & 2 has been improved on average by 0.015 ± 0.007 (standard deviation), for Ishi paper by 0.001 ± 0.011, and for Ishi Opt by 0.019 ± 0.016. There is a rather minor effect of the post-processing on F1, although there is a considerable improvement at the event level. This can be explained by the fact that the majority of the false alarms shown in the event level results were short in duration and, hence, their removal did not contribute strongly to the resulting F1 scores.

Table 2: *Hits, misses and false alarms totalled across all speakers for the four detection methods. Results are shown without and with additional post-processing.*

	WITHOUT POST-PROCESSING			WITH POST-PROCESSING		
Method	**Hits**	**misses**	**false alarms**	**Hits**	**misses**	**false alarms**
Comp 1&2	2320	426	2039	2221	525	1009
Ishi Orig.	1808	938	2311	1617	1129	1206
Ishi Opt.	2264	482	7142	2086	660	3561

As the post-processing brought an improvement to the three methods (particularly in terms of event level metrics) it will be used in with these three methods (i.e. Comp 1 & 2, Ishi Orig. and Ishi Opt.) for remainder of the study.

6.3. Detailed survey of detection performance

The frame level results for the three detection methods are presented in Figure 10, with the event level results shown in Table 3. Note that the F1 score, shown in the right column of Figure 10, gives the clearest impression of the performance of the detection methods in a single measure and, hence, will receive most attention.

Considering Figure 10 one can observe that the proposed detection method (Comp 1 & 2) produced higher F1 scores, for every speaker, than the two comparison methods (this is with the exception of Female 2 from the US database where Comp 1 & 2 and Ishi Opt. produced the same F1 score of 0.49). This is due to a comparatively high true positive rate (TPR) and low false positive rate (FPR). However, for female 1 in the US database there is relatively high FPR for Comp 1 & 2. We investigated this and found that a large proportion of the false alarms contained noises from the speaker's mouth colliding with the microphone or other background noises, mostly occurring as the person was speaking. Additional features for detecting such noises would help reduce the number of false alarms of this kind.

This general trend was also supported in the event level results (Table 3) where Comp 1 & 2 gave a higher number of hits and a lower number of misses than Ishi Orig. for every speaker. This was often combined with a lower number of false alarms. Compared with Ishi Opt., Comp 1 & 2 had a more similar level of hits and misses, however it generally led to a much

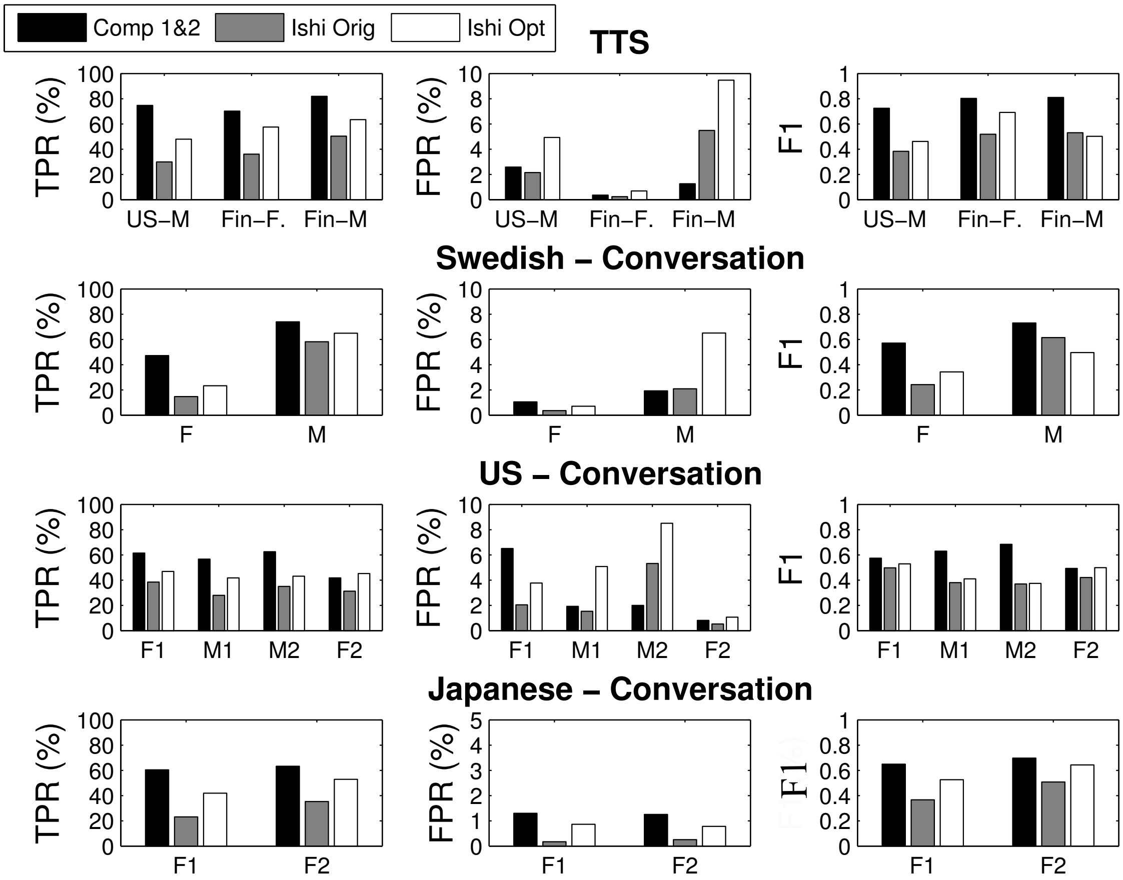

Figure 10: *Frame level metrics (i.e. TPR, FPR and F1) for the three detection methods shown for each speaker in the four databases. M is used for male and F for female.*

lower number of false alarms.

The F1 score for Ishi Opt was higher than Ishi Orig for almost every speaker (the exception being the Finnish male speaker from the TTS database). This was due to an increased TPR over Ishi Orig, however this was also coupled with an increased FPR for every speaker.

For Ishi Opt. there was generally a higher FPR and number of false alarms for male speakers. This was particularly true for the Finnish male speaker in the TTS database. This speaker had the lowest pitch of all the speakers with a mean f_0 typically around 80 Hz. We investigated the false alarms and

Table 3: *Event level results (i.e. hits, misses and false alarms) for the three creak detection methods for each speaker in the four databases.*

		COMP 1& 2			ISHI ORIG			ISHI OPT		
Database	**Speaker**	**hits**	**misses**	**FAs**	**hits**	**misses**	**FAs**	**hits**	**misses**	**FAs**
TTS	US-M	153	12	42	111	54	97	145	20	312
	Finn-F	211	51	7	171	91	13	220	42	90
	Finn-M	193	20	53	173	40	477	195	18	1098
Swedish	F	378	142	192	174	346	72	247	273	225
	M	237	33	103	221	49	157	240	30	652
US	F1	192	37	212	154	75	84	178	51	233
	M1	75	12	53	55	32	60	73	14	257
	M2	89	16	35	64	41	182	71	34	294
	F2	37	17	38	31	23	18	39	15	68
Japanese	F1	413	132	167	274	271	28	422	123	250
	F2	243	53	107	189	107	18	256	40	82

found them to be frequently due to a lack of distinction between low pitch, voiced, non-creaky segments and true creaky segments. The IFP parameter, which is utilised in these methods for differentiating normal voiced regions from creaky regions, frequently produced very low values in these low pitch regions which led to these false alarms. Further features to help disambiguate these two classes would certainly improve the detection performance when using the parameters from Ishi et al. (2008a) for detecting creak.

In order to investigate whether the F1 scores for the proposed detection method were significantly higher than those from the two comparison methods, we carried out a one-way ANOVA with F1 score treated as the dependent variable and detection method as the independent variable. This revealed that the detection method had a significant effect on the F1 score [$F_{(2,30)} = 15.002, p < 0.001$], and subsequent pairwise comparisons carried out using Tukey's Honestly Significant Difference (HSD) test revealed that the Comp 1 & 2 method gave significantly higher F1 scores than both Ishi Orig. ($p < 0.001$) and Ishi Opt. ($p < 0.01$).

We also wanted to investigate whether gender had a significant effect on the F1 scores for each of the methods. To do this we carried out a two-way ANOVA with F1 score as the dependent variable and with detection method and gender as the independent variables. Although lower mean F1 scores were observed for females with the Comp 1 & 2 method, and lower means for males in the two Ishi methods, the two-way ANOVA revealed no significant effect of gender [$F_{(1,27)} = 0.057, p = 0.81$]. However, repeating the same test with FPR as the dependent variable we found that gender did have a significant effect [$F_{(1,27)} = 25.078, p < 0.001$] and pairwise comparisons (using Tukey's HSD) revealed that for Comp 1& 2 and Ishi Opt. FPR was significantly higher (p $<$ 0.05) for males.

7. Results on degraded data

The effect of additive noise on the detection performance of the three methods is illustrated in Figure 11. It can be observed that in the white noise condition the Comp 1 & 2 method achieves the highest F1 score at all levels of signal to noise ratio (SNR). Ishi Opt. achieves a slightly higher F1 than Ishi Orig. down to an SNR of 10 dB. For the babble noise condition again the Comp 1 & 2 method attains a higher F1 than the two comparison methods down to an SNR of 20 dB. However, at SNR of 10 dB the F1 for Comp 1 & 2 falls slightly below that of Ishi Opt. All methods deteriorate severely at 0 dB SNR.

The findings here are encouraging for the proposed method as they suggest Comp 1 & 2 can provide superior detection of creak even in conditions with moderately high levels of noise. Babble noise is shown to have a stronger negative effect on the performance of the three methods compared to white noise. This is likely due to the more pronounced low frequency characteristic of babble compared to white noise which more severely affects the parameters used in the three methods at low SNR levels.

8. Discussion and conclusion

This paper presented a new method for automatically detecting creak in speech signals by exploiting characteristics of the LP-residual signal, namely the presence of secondary peaks and long glottal pulse lengths with prominent impulse-like excitation peaks. Resonators were applied to the LP-residual and two parameters were derived from the characteristics of the resonator output. These parameters were then used as input features to a decision

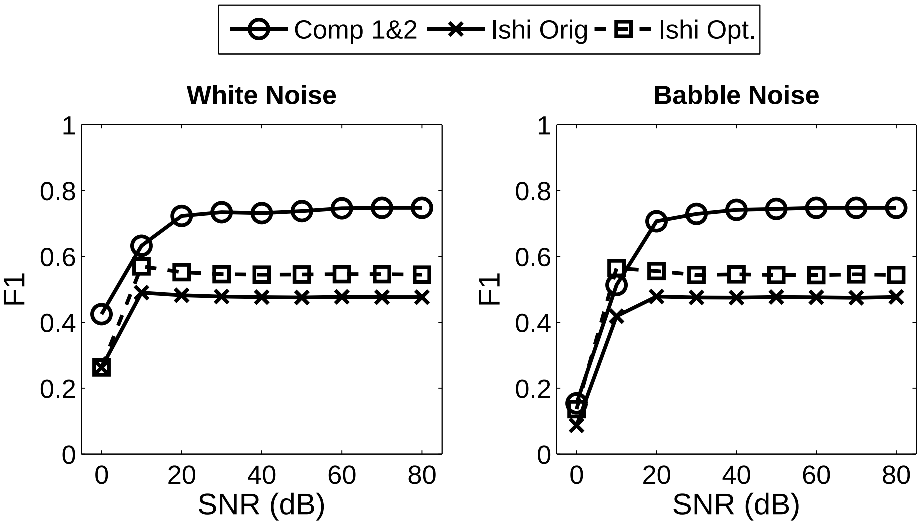

Figure 11: *Effect of white noise (left panel) and babble noise (right panel) on the F1 score (averaged across the three speakers in the TTS database) achieved by the three creak detection methods.*

tree classifier. The resulting detection performance was shown to significantly outperform existing creak detection methods on a large range of speech data covering different speakers, gender, languages, recording conditions and speaking styles (i.e. read vs conversational speech). These findings build on the initial promising results reported in Drugman et al. (2012b).

Furthermore, the new method demonstrated robustness to white noise, with the highest performance across all SNR levels, and to babble noise, with improved detection over the comparison methods down to 20 dB SNR.

The inclusion of the parameters derived using the methods described in Ishi et al. (2008a) in a decision tree classifier which was optimised on training sets brought some improvement to the overall detection performance compared to the original algorithm, with the original threshold settings. However, despite this improvement there was still an increase in the level of false alarms. The inclusion of further features, in such a classifier, which could help disambiguate non-creaky voiced segments and creaky segments would certainly bring a further improvement to the detection performance. In fact a direction of our future work is to determine the mutual and complementary information in the various features relevant to creaky voice detection and include these features in a single classification system to help better identify creaky voice regions. We also intend to investigate the existence of various possible patterns of creaky production.

For other future work we intend to investigate the extent to which contextual factors (phoneme, word stress, position in the sentence, prosodic context etc) can be used to predict locations of creak. We also intend to study and model f_0 patterns proceeding creaky regions. Both these studies will be used to contribute to our initial efforts in incorporating creak in parametric speech synthesis (Drugman et al., 2012a) and may facilitate transformation of voice characteristics. Finally, we hope that the new algorithm can be used to help quantitatively study the use of creak on larger volumes of data and we also hope to investigate its potential in applications like speaker identification.

9. Acknowledgements

The first 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). The second author is supported by the Walloon Region, Belgium (Grant WIST 3 COMPTOUX # 1017071) The authors would like to thank Jens Edlund for providing us with the Spontal corpus, Ikuko Patricia Yuasa for providing us with the US dataset and Thomas Magnuson for sharing the Japanese data with us. The authors would also like to thank Srikanth Vishnubhotla and Carol Espy-Wilson for kindly sharing the APP method.

References

Blomgren, M., Chen, Y., Ng, M., Gilbert, H., 1998. Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. Journal of the Acoustical Society of America 103 (5), 2649–2658.

Böhm, T., Both, Z., Németh, G., 2010. Automatic classification of regular vs. irregular phonation types. Advances in nonlinear speech processing, 43–50.

Böhm, T., Shattuck-Hufnagel, S., 2007. Listeners recognize speakers' habitual utterance-final voice quality. Proceedings of ParaLing07, 29–34.

Breiman, L., Stone, C. J., Olshen, R. A., Friedman, J. H., 1984. Classification and regression trees. Wadsworth Inc.

Campbell, N., Mokhtari, P., 2003. Voice quality: The 4th prosodic dimension. Proceedings of ICPhS, Barcelona, Spain, 2417–2420.

Carlson, R., Gustafson, K., Strangert, E., 2006. Cues for hesitation in speech synthesis. Proceedings of Interspeech, Pittsburgh, USA, 1300–1303.

Degottex, G., Roebel, A., Rodet, X., 2011. Phase minimization for glottal model estimation. IEEE Transactions on Audio, Speech, and Language Processing, 19 (5), 1080–1090.

Deshmukh, D., Espy-Wilson, C., Salomon, A., Singh, J., 2005. Use of temporal information: Detection of periodicity, aperiodicity and pitch in speech. IEEE Transactions on Audio Speech and Language processing 13 (5), 776–786.

Drugman, T., Alwan, A., 2011. Joint robust voicing detection and pitch estimation based on residual harmonics. Proceedings of Interspeech, Florence, Italy, 1973–1976.

Drugman, T., Bozkurt, B., Dutoit, T., 2009. Complex cepstrum-based decomposition of speech for glottal source estimation. Proceedings of Interspeech, Brighton, UK, 116–119.

Drugman, T., Dutoit, T., 2011. Oscillating statistical moments for speech polarity detection. Proceedings of Non-Linear Speech Processing Workshop (NOLISP11), Las Palmas, Gran Canaria, Spain, 48–54.

Drugman, T., Kane, J., Gobl, C., 2012a. Modeling the creaky excitation for parametric speech synthesis. Proceedings of Interspeech, Portland, Oregon, USA.

Drugman, T., Kane, J., Gobl, C., 2012b. Resonator-based creaky voice detection. Proceedings of Interspeech, Portland, Oregon, USA.

Drugman, T., Thomas, M., Gudnason, J., Naylor, P., Dutoit, T., 2012c. Detection of glottal closure instants from speech signals: a quantitative review. IEEE Transactions on Audio Speech and Language processing 20 (3), 994–1006.

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., House, D., 2010. Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. Proceedings of LREC, Malta, 2992–2995.

Edmondson, J., Esling, J., 2006. The valves of the throat and their functioning in tone, vocal register and stress: laryngoscopic case studies. Phonology 23 (2), 157.

Elliot, J. R., 2002. The application of a Bayesian approach to auditory analysis in forensic speaker identification. Proceedings of the 9th Australian International Conference on Speech Science and Technology, 315–320.

Espy-Wilson, C., Manocha, S., Vishnubhotla, S., 2006. A new set of features for text-independent speaker identification. Proceedings of Interspeech (IC-SLP), Pittsburgh, Pennsylvania, USA, 1475–1478.

Gerratt, B. R., Kreiman, J., 2001. Toward a taxonomy of nonmodal phonation. Journal of Phonetics 29, 365–381.

Ghosh, P., Tsiartas, A., Narayanan, S., 2011. Robust voice activity detection using long-term signal variability. IEEE Transactions on Audio, Speech, and Language Processing, 19 (3), 600–613.

Gobl, C., Ní Chasaide, A., 1992. Acoustic characteristics of voice quality. Speech Communication 11, 481–490.

Gobl, C., Ní Chasaide, A., 2003. The role of voice quality in communicating emotion, mood and attitude. Speech Communication 40, 189–212.

Heldner, M., Edlund, J., Hjalmarsson, A., Laskowski, K., 2011. Very short utterances and timing in turn-taking. Proceedings of Interspeech, Florence, Italy, 2837–2840.

Hollien, H., Wendahl, R. W., 1968. Perceptual study of vocal fry. Journal of the Acoustical Society of America 47 (3), 506–509.

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

Ishi, C. T., 2004. Analysis of autocorrelation-based parameters for creaky voice detection. Proceedings of Speech Prosody, Nara, Japan, 643–646.

Ishi, C. T., Ishiguro, H., Hagita, N., 2005. Proposal of acoustic measures for automatic detection of vocal fry. Proceedings of Interspeech, Lisbon, Portugal, 481–484.

Ishi, C. T., Ishiguro, H., Hagita, N., 2008b. Automatic extraction of paralinguistic information using prosodic features related to f0, duration and voice quality. Speech communication 50 (6), 531–543.

Kane, J., Gobl, C., 2011. Identifying regions of non-modal phonation using features of the wavelet transform. Proceedings of Interspeech, Florence, Italy, 177–180.

Kay, S., 1988. Modern spectral estimation: theory and application. Prentice Hall Englewood Cliffs, NJ.

Kominek, J., Black, A., 2004. The CMU ARCTIC speech synthesis databases. ISCA speech synthesis workshop, Pittsburgh, PA, 223–224. URL `http://festvox.org/cmuarctic/`

Laver, J., 1980. The Phonetic Description of Voice Quality. Cambridge University Press.

Laver, J., 1994. Principles of Phonetics. Cambridge University Press.

Magnuson, T., 2011. Realizations of /r/ in Japanese talk-in-interaction. Proceedings of ICPhS, Hong Kong, 1306–1309.

Moisik, S., Esling, J., 2011. The 'whole' larynx approach to laryngeal features. Proceedings of ICPhS, Hong Kong, 1406–1409.

Ogden, R., 2001. Turn transition, creak and glottal stop in finnish talk-in-interaction. Journal of the International Phonetic Association 31 (1), 139–152.

Ogden, R., 2009. The larynx, voicing and voice quality. In: An introduction to English phonetics. pp. 40–55.

Scherer, S., Schwenker, F., Campbell, N., Palm, G., 2009. Multimodal laughter detection in natural discourses. Human centered robot systems, Cognitive Systems Monographs 6, 111–120.

Silen, H., Helander, E., Nurminen, J., Gabbouj, M., 2009. Parameterization of vocal fry in HMM based speech synthesis. Proceedings of Interspeech 2009, Brighton, UK, 1775–1778.

Slifka, J., 2006. Some physiological correlates to regular and irregular phonation at the end of an utterance. Journal of Voice 20 (2), 171–186.

Surana, K., Slifka, J., 2006a. Acoustic cues for the classification of regular and irregular phonation. Proceedings of Interspeech (ICSLP), Pittsburgh, Pennsylvania, USA, 693–699.

Surana, K., Slifka, J., 2006b. Is irregular phonation a reliable cue towards the segmentation of continuous speech in American English. Proceedings of Speech Prosody, Dresden, Germany, Paper 177.

Titze, I., 1994. Vocal registers. In: Principles of Voice Production. Englewood Cliffs, MJ: Prentice Hall, pp. 252–259.

Titze, I., Sundberg, J., 1992. Vocal intensity in speakers and singers. Journal of the Acoustical Society of America 91 (5), 2936–2946.

Vainio, M., 2001. Artificial neural network based prosody models for finnish text-to-speech synthesis. Ph.D. thesis, University of Helsinki, Finland.

Varga, A., Steeneken, H., 1993. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication 12 (3), 247–251.
URL `http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html`

Villavicencio, F., Robel, A., Rodet, X., 2006. Improving lpc spectral envelope extraction of voiced speech by true-envelope estimation. Proceedings of ICASSP, Toulouse, France.

Vishnubhotla, S., Espy-Wilson, C., 2006. Automatic detection of irregular phonation in continuous speech. Proceedings of Interspeech, Pittsburgh, USA, 949–952.

Wolk, L., Abdelli-Beruh, N., 2012. Habitual use of vocal fry in young adult female speakers. Journal of Voice 26 (3), 111–116.

Yanushevskaya, I., Gobl, C., Ní Chasaide, A., 2005. Voice quality and f0 cues for affect expression. Proceedings of Interspeech, Lisbon, Portugal, 1849–1852.

Yu, K. M., Lam, H. W., 2011. The role of creaky voice in Cantonese tonal perception. Proceedings of ICPhS, Hong Kong, 2240–2243.

Yuasa, I. K., 2010. Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile american women? American Speech 85 (3), 315–337.