ISCA Archive
http://www.isca-speech.org/archive

Speech Prosody 2012
Sixth International Conference
Shanghai, China
May 22-25, 2012

Automatic detection of voice creak

Philippe Martin

CLILLAC-ARP, EA 3967, UFR Linguistique
Université Paris Diderot Sorbonne Paris Cité
philippe.martin@linguist.jussieu.fr

Abstract

The analysis of large spontaneous speech corpora reveals that creaky mode appears more frequently than expected, especially for young female speakers. Creaky mode usually creates fundamental frequency measurement errors and creaky voice segments must be often identified manually beforehand to avoid erroneous reading of F0 in large speech databases.

Various approaches have been proposed to identify creaky segments with diplophonic and vocal fry automatically, based on autocorrelation, AMDF, HMM, pitch markers, etc. The approach proposed here is based on narrow band Fourier spectrum analysis, operating not on a single frame but on the evaluation of sudden changes in the harmonic distribution of consecutive frames. The implemented algorithm simulates the visual detection of creak from spectrographic display where so-called sub harmonics appear on short voice segments.

Index Terms: creaky voice, diplophonia, vocal fry, fundamental frequency, spontaneous speech.

1. Introduction

The analysis of large spontaneous speech corpora reveals that creaky mode is used more frequently than expected, especially by young urban female speakers [6]. As automatic or semi-automatic prosodic analysis requires reliable acoustic parameters measurements such as fundamental frequency, it is necessary to identify the creaky speech segments to invalidate beforehand erroneous pitch tracking values that would be given for those segments.

Various approaches have been proposed to identify creaky segments automatically [1], based on autocorrelation [3], AMDF [2], HMM [5], pitch markers [4], etc., but all involve some drawbacks, due to the complexity and variability of the speech signal and the presence of various noise sources inherent to spontaneous speech recordings.

The approach proposed here is based on narrow band Fourier spectrum analysis, operating not on a single frame but on the evaluation of sudden changes in the harmonic pattern of consecutive frames. The implemented algorithm simulates the visual detection of creak from spectrographic display where so-called sub harmonics appear on short voice segments.

2. Voice creak

Perceived voice creak may correspond to different characteristics in the speech signal. It may be due to large period to period irregularities in the signal, with a jitter values well over 10 % (Fig. 1), by a very low fundamental frequency (e.g. below 50 Hz) (Fig. 2) or by paired pulsing (diplophonia) (Fig. 3).

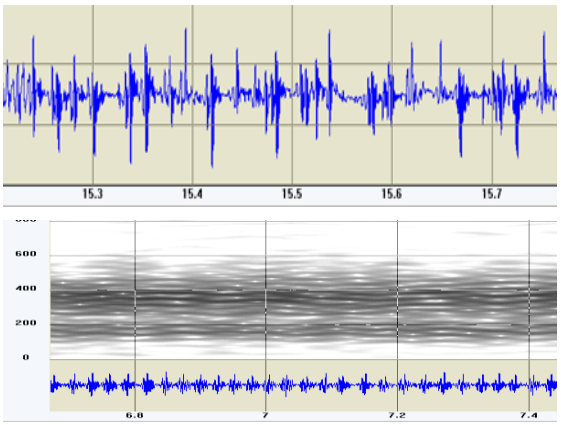

Figure 1 *Creaky segment with consecutive pitch periods large irregularities (jitter) (vocal fry)*

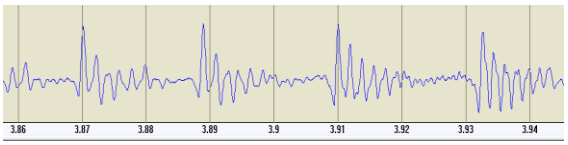

Figure 2 *Creaky segment with very long pitch periods (above 20 ms)*

The type considered more specifically here is diplophonia, characterized by paired pulsing [9]: consecutive pitch periods may differ by more than 10% in value, whereas alternate pitch periods may differ by less than 5% (Fig. 3). This is different from vocal fry, where such similarity between alternate periods is not found (Fig. 1).

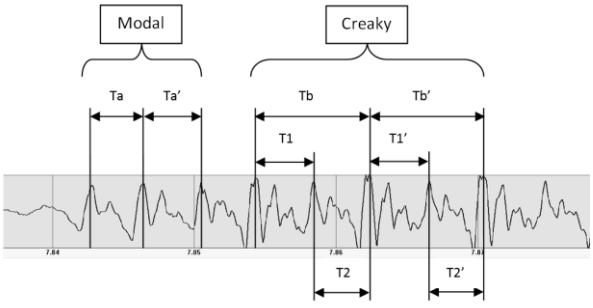

Figure 3. *An example of voice creak with paired pulsing(diplophonia).*

Fig. 3 illustrates a case of paired pulsing appearing after few laryngeal periods in modal phonation. In modal phonation, consecutive pitch periods Ta and Ta' differ only by few %. In diplophonia creaky mode, consecutive periods T1 and T2 are less similar than T1-T1' and T2 T2', so that Tb and Tb'

appear (wrongly) as the laryngeal period. This type is thus different from the other modes perceived as creak, irregular (Fig. 1) or with a low F0 (Fig. 2).

Since the consecutive periods are relatively close in value the limited frequency resolution of Fourier analysis of a creaky segment merges the two frequencies 1/T1 and 1/T2 in the spectrographic display. Furthermore, the Tb Tb' sequence (with Tb=T1+T2, Tb'=T1'+T2), due to the newly created time pattern regularity, will result in a new harmonic component (often called sub harmonic) with a frequency = 1/Tb (Fig. 4).

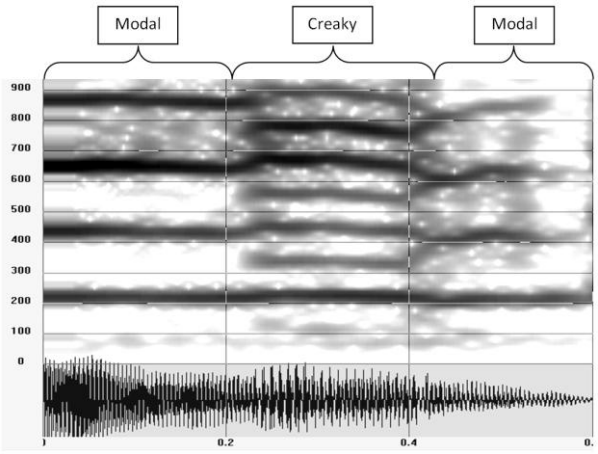

Figure 4. *Spectrogram of a modal-creaky-modal speech segment sequence.*

Fig. 4 shows a spectrogram of a modal-creaky-modal segment sequence. The diplophonic section displays the sub harmonics (with a very weak intensity first component) witch disappear when the modal phonation is restored. Fig. 4 shows as well that the first subharmonic of a creaky segment is very weak or absent. This is the consequence of the alternate pulsing in creaky mode, the time domain pattern created having low amplitude for its main component (as shown Fig. 3). Vowels like [u] in French sometimes exhibit the same property with a fundamental 6 dB to 12 dB lower than the second harmonic.

3. A spectrogram based algorithm

One of the reasons explaining why autocorrelation of pitch markers methods do not always give satisfactory results may stem from the fact that the identification of creak (any type) relies on one signal frame, generally of 30 or 40 ms duration, whereas successful visual inspection of a spectrogram spans on the whole creaky segment, i.e. generally on 200 to 500 ms or more, as well as the surrounding speech segments produced in modal phonation.

In order to automatically identify creaky segment, the proposed algorithm detects a sudden variation of the number of harmonics detected when a transition from a modal segment to a diplophonic or vocal fry segment occurs. It attempts to simulate an operator visual inspection of a narrow band spectrogram by considering large voice segments at once to establish the creaky character of a speech segment instead of just one frame.

In its implementation, the algorithm proceeds with the analysis performed on time segments separated by 80 ms. For each of these segments, the number of spectral peaks is evaluated in a frequency range of 60 to 800 Hz using an analysis time window of 64 ms in order to obtain a sufficient frequency resolution in the resulting frequency spectrum and select reliable peaks. A positive difference above a predefined threshold indicates a modal-creak transition, a negative difference a creak-modal transition (again for the diplophonic and vocal fry varieties of creak).

Instead of the difference of detected harmonics, for each transition a creak index is calculated from the ratio of detected harmonics, theoretically equal to 2 for modal-creak transitions, and to 0.5 for creak-modal transitions. Vocal fry segments will also be detected with this index since the drop in F0 will lead to the presence of a larger number of harmonics in the same frequency range considered by the algorithm.

The validity of the creak index has been first evaluated by visual inspection on recordings sampled from the Rhapsodie project [8]. 177 segments have been identified perceptually as creaky and reported automatically in a data base included in the WinPitch speech analysis program [9]. Clicking on each entry of this table in the left window (Fig. 5) allows for an easy inspection of the creaky part and a visual analysis of the validity of the creak index given by the algorithm, leading to an adequate set of parameters to evaluate creaky indexes (frequency range for harmonics identification, temporal spacing of harmonics selection windows, number of spectra in selection windows).

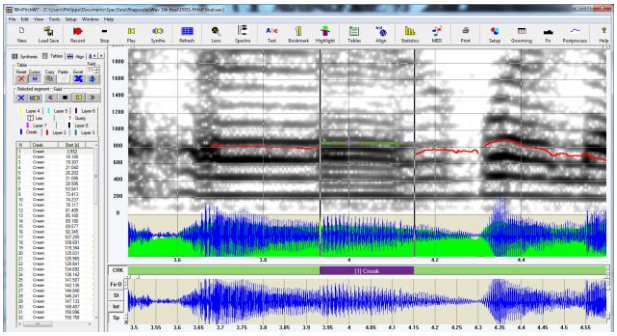

Figure 5. *Example of detection of a creaky segment (female voice)*

4. Implementation

The algorithm was implemented in the speech analysis software WinPitch package using the same routines implemented to select harmonics for the spectral comb evaluation of fundamental frequency. The Fourier transform operates on a 64 ms window giving a 15 Hz frequency resolution. Harmonic peak frequencies are obtained by parabolic interpolation.

To avoid perturbation effects in the selection of harmonics due to a change in the formant structure inside a vowel, the frequency range retained to select harmonics has been set to 60 Hz-800 Hz (this upper limit is user adjustable). Furthermore, to avoid confusion between vowel harmonics and noise peaks, an adjustable intensity threshold is also used to reject false harmonics. Likewise, a simple voicing zero crossing threshold is used to reject from the creak index calculation segments possibly unvoiced.

Intervals between consecutive windows for harmonic selections were set to 80 ms. In order to average the harmonic detection errors due to the position of the Hann(ing) window

on the signal, 10 successive measures spaced by 4 ms are taken for each window, A creaky index is computed for each window shifted position and is equal to the average number of harmonics detected in the right window divided by the average number of harmonics detected in the left window, the right window being 50 ms ahead in time from the left window, and each window corresponding to 10 measures of the number of detected harmonics taken every 4 ms. A creaky transition modal-creaky threshold has been set to 1.5, when the number of right harmonics exceeds the number of left harmonics by a factor of 1.5 (Fig. 6).

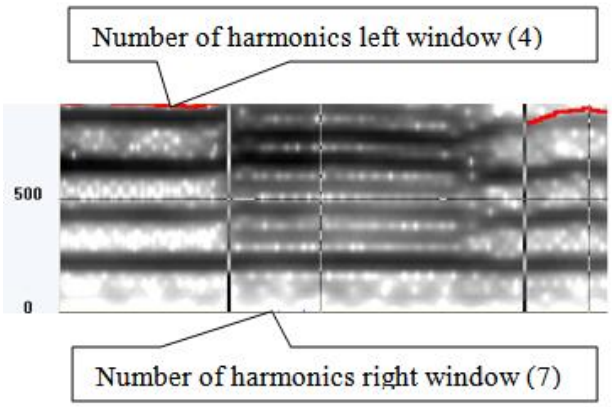

Figure 6. *The creaky index (here 7/4) is computed from the ratio of the number of harmonics detected in a 60 Hz-800 Hz range on two consecutive time windows spaced by 80 ms.*

5. Evaluation and limitations

A total of 177 tokens of visually identified creaky (i.e. diplophonic, vocal fry and low F0) and overlapping segments were identified by the algorithm according to Table 1.

Type	Diplophonic	Vocal fry	F0 < 50 Hz	Overlapping
Observed	95	27	9	43
Detected	74	22	0	35
%	**87 %**	**81 %**	**0 %**	**81 %**

TABLE 1 – *Percentages of correct automatic detection of visually identified diplophonic and vocal fry segments*

Table 1 gives the percentages of correct detection of segments identified visually on narrow band spectrograms as diplophonic, vocal fry, with F0 < 50 Hz or speech overlapping. These values were obtained from the analysis of 5 rather low recording quality recordings (mp3 compression, background noise, low recording amplitude,…) from the Rhapsodie corpus (D001, D003, D004, D005 and D008). The speakers voice characteristics vary considerably from one recording to the next, D001 contains many cases of diplophonia, while D005 has as many diplophonia than vocal fry occurrences. The algorithm detected a large number of speech overlapping in the D004 recording.

A detailed examination of detection errors reveals that most wrong creak identifications were due either to the presence of unvoiced segments before or after the diplophonic or vocal fry section, or the presence of a high intensity noise, both conditions leading to a wrong selection and thus a erroneous measure of the number of harmonics and the creak index.

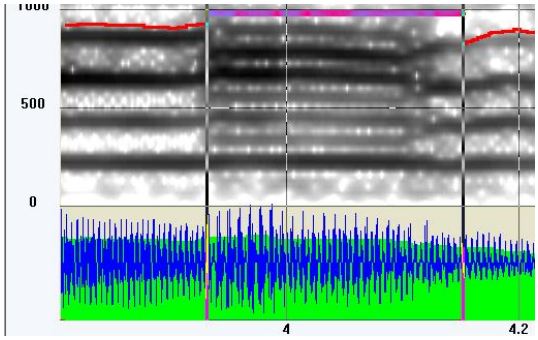

Figure 7. *A clear example of a modal to creaky transition, with a detected creaky index of 7/4 (female voice).*

Some sources of errors are illustrated in Fig. 8 to Fig. 10. They pertain to a rise or fall in fundamental frequency which changes the count of harmonics in the given frequency range (60 Hz to 800 Hz). Fig. 8 and Fig. 9 show examples of a somewhat low creaky index due to a rapid change of fundamental frequency, involving the disappearance of harmonics close to the upper limit of the spectrum frequency range.

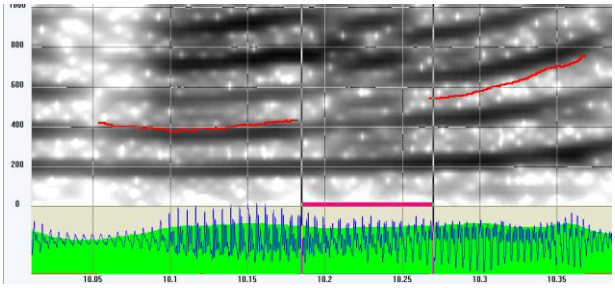

Figure 8. *An example where the creak index = 7/5. The upward variation of harmonics in the modal-creaky transition zone gives a rather low modal/creaky ratio (female voice).*

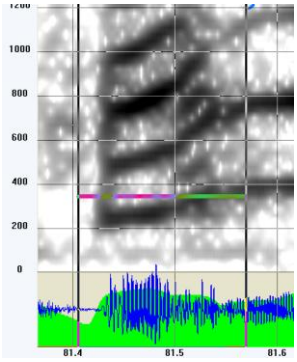

Figure 9. *An example where F0 changes rapidly leading to an index equal to 7/5 as the modal segment harmonics leave the frequency range used for the numbering of harmonics.*

Figure 10 is an example of an erratic detection of harmonics in the zone perceptually identified as creaky. This error is due to the high level of noise in the segment, which prevents the harmonics to be properly identified.

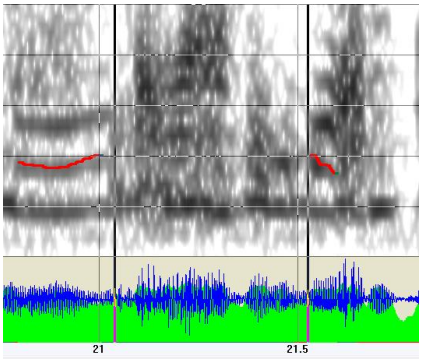

Figure 10. *An example where the creak index (15/8) indicates a creak, but harmonic detection in the creaky zone is erratic due to noise in the signal (male voice).*

Fig. 11 gives an erratic creak index due to the presence of an unvoiced segment (in this case a silence) before the diplophonic segment, preventing the identification of an harmonic pattern change.

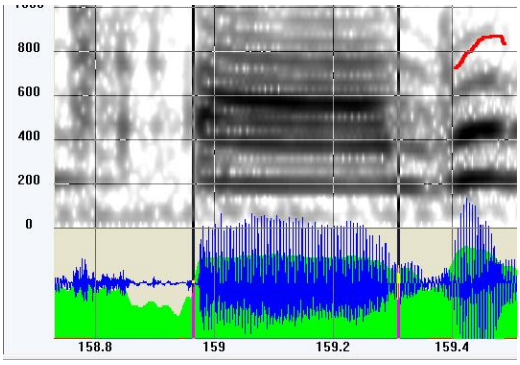

Figure 11. *The diplophonic segment is not detected due to the unvoiced segment that precedes (a silence in this example). The index is then 7/0.*

6. Conclusions

More and more research projects on sentence intonation use spontaneous speech recordings which show for some speakers numerous creaky segments. As those creaky speech segments give numerous erroneous fundamental frequency values, whatever the pitch tracking method used, it is important to identify beforehand creaky segments in the speech continuum.

The method of creaky segments detection presented here relies on the identification of a sudden change of the number of harmonics in a given time frame and in a given frequency range. As it actually mimics visual identification of diplophonic and vocal fry creak on a narrow band spectrogram, its implementation is very simple and gives satisfactory results despite some limitations due to adverse recording conditions or fast F0 change in the signal.

A future version of the algorithm will incorporate a test to better differentiate vocal fry from low frequency creak, as well as the integration in the index of both modal-creak and creak-modal transitions.

7. References

[1] Hollien, H., Michel, J., and Doherty, E.T. (1973). A Method for Analyzing Vocal Jitter in Sustained Phonation, J. Phon. 1, 85-91.

[2] Vishnubhotla, S. and Espy-Wilson, C. Y. (2006). Automatic detection of irregular phonation in continuous speech, In *INTERSPEECH-2006*, paper 1893-Tue1CaP.12.

[3] Ishi, C. Toshinori, Ishiguro, H. and Hagita, N. (2005). Proposal of acoustic measures for automatic detection of vocal fry, In *INTERSPEECH-2005*, 481-484.

[4] Hagmuller, M. and Kubin, G. (2006) Poincaré pitch marks, Speech Communication 48 (2006) 1650–1665.

[5] Lugger, M., Stimm, F. and Yang, B. (2008). Extracting voice quality contours using discrete hidden Markov models, In *Speech Prosody 2008*, 29-32.

[6] Yuasa, I. P. (2010) Creaky Voice: A New Feminine Voice Quality for Young Urban-Oriented Upwardly Mobile American Women? American Speech Fall 2010 85(3) 315-337.

[7] McKinney N. P. (1965). *Laryngeal Frequency Analysis for Linguistic Research*, Ann Arbor, University of Michigan Communication Sciences Laboratory, VII.

[8] Rhapsodie (2011) Reference prosody corpus on spoken French, http://rhapsodie.risc.cnrs.fr/en/index.html.

[9] WinPitch (2011) www.winpitch.com