

On the use of Creak in Hungarian Spontaneous Speech

John Kane¹, Kinga Pápay², László Hunyadi², Christer Gobl¹

¹ Centre for Language and Communication Studies, Trinity College, Dublin

² Department of General and Applied Linguistics, University of Debrecen

kanejo@tcd.ie, kinga.papay@gmail.com, hunyadi@llab2.arts.klte.hu, cegobl@tcd.ie

ABSTRACT

Creak, also referred to as vocal fry or glottal fry, is a perceptually distinctive vocal occurrence that is frequently produced in running speech. The presence of creak causes problems in speech processing particularly with f_0 and spectral analysis. However, as creak frequently occurs in natural spoken communication and as it often carries important segmental and paralinguistic information we believe that applying appropriate processing methods for the study of creak in spontaneous conversation is warranted. This preliminary study demonstrates trends in the use of creak by 9 Hungarian speakers talking in both formal and informal settings. Creaky segments are identified and the distributions of these segments were examined in relation to frequency of occurrence in formal versus informal settings. The study shows trends of increased frequency of creak in informal settings and an apparent convergence in speaking style on the part of the interviewer with interviewees who produced creak frequently.

Keywords: Voice quality, creaky voice, glottal fry, interactive speech, phonetic convergence

1. INTRODUCTION

The voice quality or colouring of a person's voice can allow insights into their affective state and can be used for certain communicative functions in spoken discourse. One frequently produced voice quality is creak (also referred to also glottal fry or vocal fry). Perceptually creak has been given impressionistic descriptions, such as sounding like: the "popping of corn" or a "rapid series of taps, like a stick being run along a railing" [1].

Laver [2] described the physiology involved in producing creak as having strong adductive glottal tension and very low longitudinal tension, as well as low subglottal pressure [3]. During creaky phonation the thyroarytenoid muscles can shorten resulting in a thickening of the vocal folds [4]. This also results in the vocal fold margins becoming more flaccid while at the same time being tightly adducted [3]. This vocal gesture produces the 'bubbling' of subglottal air between the folds [5]. The resulting

speech signal is typically characterised by a very low f_0 , normally lower than 80 Hz and ranging as low as 10 Hz. Interestingly, male and female speakers tend not to produce significantly different f_0 when using creaky phonation [6, 3]. This may imply that overall longitudinal vocal fold length is less of a factor at determining vibration rate in creaky speech [3].

Creaky speech signals can sometimes be periodic but are frequently irregularly spaced and can also display multi-pulsed patterns [7]. These multi-pulsed patterns, involving multiple excitations within a single glottal cycle, can take the form of both doublets and triplets [3]. They have also been reported as having very low open quotients, meaning a long closed phase in each pulse length [7].

Processing of speech containing creak is frequently hampered by these acoustic characteristics. Pitch detection algorithms are typically designed with in-built assumptions about likely f_0 ranges. These assumptions, however, frequently do not hold for analysis of creak and, hence, algorithms produce spurious values. Furthermore, standard windowing operations, e.g., those used for obtaining spectral and cepstral measurements, typically involve using window lengths of no longer than 32 ms. As two pitch periods are required for periodic/harmonic information, and as creaky pulse lengths are frequently longer than 16 ms (i.e. $f_0 < 62.5$ Hz), standard analysis strategies are clearly unsuitable.

As a result of the above analysis difficulties, researchers involved in speech processing often disregard or avoid speech segments containing creak. However, by doing this researchers fail to extract what can be rich information pointing to the speaker's mood (in recognition systems) and also fail to incorporate creak segments in synthesis systems, missing the potential for improved naturalness. In fact a recent study demonstrated improved naturalness in statistical parametric speech synthesis by appropriately modelling creak segments [8].

Creak often carries important segmental as well as paralinguistic information [9]. The study carried out in [9] involved analysis of a short Swedish utterance and resynthesis using a formant synthesiser involving the setting of source parameters to produce

different voice qualities. Perception tests indicated that creaky voice was associated with low activation emotions, both positive and negative.

There are relatively few studies reporting on the use of creaky voice in spontaneous, interactive speech. It has been, however, investigated in the case of turn yielding functions in Finnish spoken conversations [10].

The present study examines the use of creak by Hungarian speakers in one-to-one interview situations with both formal and informal settings. Creak segments are automatically identified using a slight adaptation of the algorithm described in [11]. Identified segments are then analysed in relation to the formality of the setting. Furthermore, trends in the speaking style used by the interviewer with the various interviewees is examined, particularly in relation to the use of creak.

2. DETECTING CREAKY SEGMENTS

2.1. Speech data

The speech data are part of the HuComTech Multimodal Database which was recorded and is being annotated by the Hungarian HuComTech (Human-Computer Interaction Technologies) Research Group of Debrecen [12]. The database building is part of a comprehensive project which carries out multidisciplinary research on spontaneous multimodal human-human and human-computer interaction.

From the database we selected conversations of 8 participants (4 male, 4 female). The same interviewer (female) was present for all participants. The participants had three short tasks during the recording, two of which were involved in the current study. The first was participation in a formal dialogue which involved answering a set of questions of a simulated job interview. Participants used the formal form of ‘you’ (in Hungarian there are two different forms of ‘you’; one is formal - *ön*, the other is informal - *te*). The second task included in this study was participation in an informal dialogue. Although the interviewer asked questions, the conversations were allowed to deviate quite a bit from the question topic. The interviewer behaved as an equal partner, told her own stories as well and used the informal form of ‘you’. Both settings were starkly contrasted in terms of formality. Around 30 minutes of material was recorded per speaker (10 minutes formal and 20 minutes informal).

The recordings were made in the pre-equipped studio of the Institute of English-American Studies of the University of Debrecen. Audio recordings were made using a Shure 16 A cardioid microphone

(one per speaker). Audio was digitised at a sampling frequency of 44.1 kHz and with 16 bit quantization, and then downsampled to 16 kHz. Post-processing involved manually chopping the conversation waveforms to just include the sections relating to the specific speaker and no overlapping speech was kept (for the present study). In this study the speaker numbering was identical to that used in the database, to allow comparison in future studies. ‘M’ or ‘F’ was attached to the speaker number to mark the speaker as male or female.

2.2. Automatic detection

Parameters to model acoustic characteristics of creak were described in [11] and these parameters are employed in the current study. A power contour of the speech signal using a very short window size of 4 ms (with the shift being 2 ms) is measured. Creak segments typically display rapid fluctuations in this contour and a power peaks (*PwP*) parameter is used to analyse the peaks in the contour and identify potential creak candidates.

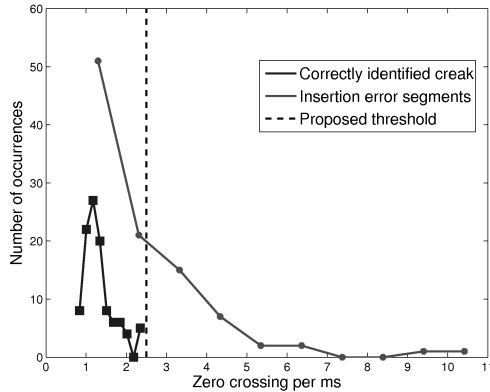
A parameter is used to measure Intra-Frame Periodicity (IFP). A standard window size (i.e. 32 ms) is used. Using the normalised autocorrelation function of the windowed signal of non-creaky, voiced speech will produce strong repeating peaks. However, as creak segments are predominantly either irregularly spaced or have pitch periods longer than half the window length, peaks in the normalised autocorrelation function will be much weaker.

The final parameter, *IPS*, measures the similarity of successive candidate pulses using a cross-correlation function. The maximum peak in the cross-correlation function is likely to be high when comparing two creak pulses, but will be low for unvoiced segments. All of the analysis is done on speech signals filtered below 100 Hz and above 1500 Hz.

The thresholds for identifying creak segments suggested in [11] are also used here (i.e. $PwP \geq 7$ dB, $IFP \leq 0.5$, $IPS \geq 0.5$). However, from our experience of using this method the insertion errors are mainly speech-less or voiceless segments, which is consistent with the results in [11]. To deal with this we propose adding a simple zero-crossing rate parameter (*zeroXrate*) as further parameter to the set ($zeroXrate = \frac{\text{Number of zero crossings}}{\text{Time (ms)}}$). This measurement is taken from the unfiltered speech signal.

Fig. 1 shows the distribution of *zeroXrate* values for creak segments and insertion errors. The speech data for this were the conversations of two male speakers both with the female interviewer. Identified creak segments were annotated by the first author as

Figure 1: Distributions of *zeroXrate* values for crack segments (N = 100) and insertion errors (N = 100).



either being crack or non-crack. From these two sets 100 segments were chosen randomly and analysed using the *zeroXrate* parameter. It can be seen in Fig. 1 that by setting a threshold of *zeroXrate* = 2.5, 31 % of insertion errors are removed with no correctly identified samples removed.

This approach was used to identify crack segments in the speech data used in the present study. However, as there were still a small number of remaining insertion errors, all identified crack segments were exported as sound segments. The first author then listened to the sound files and segments deemed not to contain crack were excluded. It should be noted that the perceptual criteria for crack in the current study is consistent with that in [11] (i.e. a “rough quality with an additional sensation of repeating impulses”).

2.3. Analysing crack segments

For the first experiments we looked at the number of crack segments per conversation. A simple crack rate value was used ($\text{Crack Rate} = \frac{\text{Number of crack segments}}{\text{Time (seconds)}}$). We examined whether participants used crack more in informal or formal settings. We then investigated whether the interviewer used crack more for participants who produced crack frequently and whether any convergence in terms of this voice quality was more apparent in formal or informal settings.

3. RESULTS

Voice quality analysis for each speaker revealed higher frequency of crack in informal settings, compared with formal settings, for every speaker (see Fig. 2). For speakers 55F and 78F the increase was only marginal but for all other speakers the increase was substantial (53 % increase in the lowest other case).

Interviewer crack rates were then analysed for each of the participants in both setting types. The scatterplot in Fig. 3 shows a trend towards the interviewer using a higher frequency of crack for speakers who ‘creaked’ more frequently in informal settings. Correlation scores (Pearson R) were substantially higher for informal settings (R = 0.87) compared with formal settings (R = 0.77). However, as the informal setting always came after the formal setting we carried out a further experiment to determine whether the above findings were indeed as a result of setting type and not solely as a result of general increase in frequency with time.

Figure 2: Crack rate for each speaker in formal and informal settings

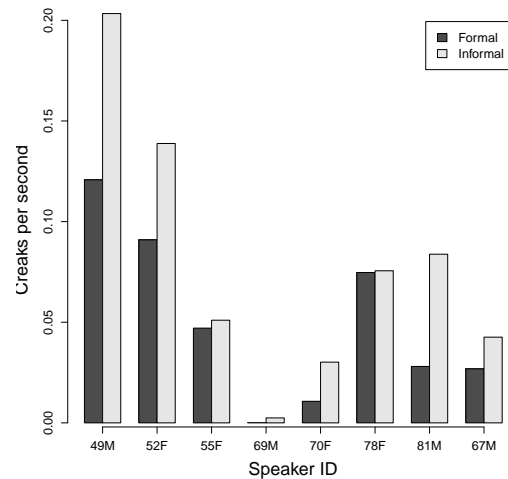
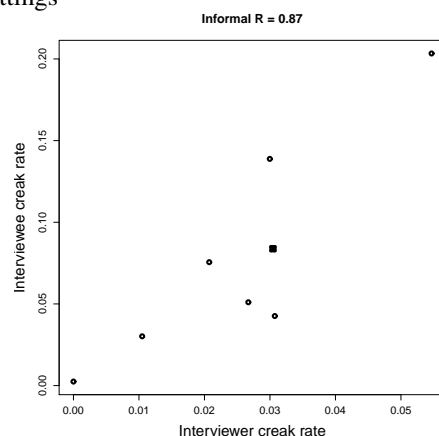


Figure 3: Scatterplot of crack rate for the interviewer compared with the interviewee in informal settings



The test involved dividing each conversation into equal time lengths of one third the total length of the conversation (henceforth referred to as tertiles). The crack rate was measured for each of the tertiles. Then correlation values (Pearson R) were

Table 1: Correlation values (Pearson R) comparing creak rates in successive tertiles with the time index of the tertile.

SpeakerID	R (Formal)	R (Informal)
49M	-0.87	-0.86
52F	0.97	-0.50
55F	-0.24	1
69M	0	0.86
70F	-0.50	0
78F	-0.96	0
81M	0.86	0.98
67M	-0.86	-0.86

taken comparing creak rates to the tertile sequence number for formal and informal settings separately (see Table 1). Only for speaker 81M was there successive positive correlation for both formal and informal settings, suggesting that time may have been a factor for the increase in the frequency of creak productions. For the remaining speakers no clear trends of frequency of creak with time were observed and, hence, we do not believe that time was the deciding factor in the findings displayed in Figs 2-3.

Furthermore, as the observations for speaker 81M are represented as a square in Fig. 3 it can be seen that it is not heavily influencing the trends. This also suggests that there is no evidence of convergence over the course of the interaction (a phenomenon also requiring a relationship with time)

4. DISCUSSION

In a previous study involving analysis and synthesis of creaky voice [9] perception tests suggested a relationship mainly with low activation emotions. In the current study there is evidence to suggest that creaky utterances are more frequently produced in informal/relaxed settings. This appears to be reasonably inline with the findings in [9]. However, the interviewer informally described the conversations considered as formal to be more tedious and boring compared with the informal conversations. The results here, hence, suggest that of the lower activation emotions (i.e. boredom and relaxedness) creaky utterances are more commonly used in relaxed settings. There is also evidence here of an apparent convergence in the use of creaky voice quality on the part of the interviewer with the use of creak by the interviewees and that this is shown more clearly to be the case in informal settings. However, the results do not show evidence of speakers adapting over the course of time. Many studies on convergence demonstrate adaption in speaking styles over time in cooperation tasks, but not in spontaneous

non-directed conversation.

Voice quality has rarely received attention from studies on phonetic convergence, however the results here point to a convergence between speakers in the use of creaky phonation. As these findings are based on the speaking tendencies of only one speaker (i.e. the interviewer) it would be worthwhile extending this analysis to more speakers to examine whether the trends here can be generalised.

5. FUTURE WORK

A possible extension of this study is to examine the location of creak segments in relation to different tiers of annotation (e.g., syntax, discourse function etc). Future work would also include extending the set of voice qualities used in the analysis. Furthermore, as the database contains multi-modal recordings it would also be possible to carry out examinations comparing acoustic features with features from the video recordings (e.g., head movements, gaze).

6. REFERENCES

- [1] Catford J. 1964. "Phonation types: the classification of some laryngeal components of speech production", Blackwell.
- [2] Laver, J. 1980. "The Phonetic Description of Voice Quality", Cambridge University Press.
- [3] Blomgren, M., Chen, Y., Ng, M., and Gilbert, H. 1998. "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers", *J. Acoust. Soc. Am.*, 103(5), pp.2649-2658.
- [4] Allen, E., and Hollien, H. 1973. "A laminagraphic study of pulse (vocal fry) phonation", *Folia Phoniat*, 25, pp.241-250.
- [5] Zemlin, W. 1988. *Speech and hearing science: Anatomy and physiology*, Prentice-Hall, NJ.
- [6] McGlone, R. 1967. "Airflow during vocal fry phonation", *J Speech Hear Res.*, 10, pp.299-304.
- [7] Gobl, C., and Ní Chasaide, A. 1992. "Acoustic characteristics of voice quality", *Speech Commun.*, 11, pp.481-490
- [8] Silén, H., Helander, E., Nurminen, J., and Gabbouj, M. 2009. "Parameterization of vocal fry in HMM-based speech synthesis", *Proc. of Interspeech*, pp. 1775-1778.
- [9] Gobl, C., and Ní Chasaide, A. 2003. "The role of voice quality in communicating emotion, mood and attitude", *Speech Commun.*, 40, pp. 189-212.
- [10] Ogden, R., 2001. "Turn transition, creak and glottal stop in Finnish talk-in-interaction", *J. Int. Phonetic Association*, 31 (1), pp. 139-152.
- [11] Ishi, C. 2008. "A method for automatic detection of vocal fry", *IEEE Trans. on Audio, Speech and Language Processing*, 16 (1).
- [12] Papay, K. 2011. "Designing a Hungarian Multi-modal Database - Speech Recording and Annotation." In: *Proc. of COST 2102 International Training School Caserta, Italy*. Springer, Heidelberg, pp. 403-411.