DOI: 10.61782/fa.2023.0426

forum acusticum 2023

Auditory Perception of Impulsiveness and Tonality in Vocal Fry

Vinod Devaraj[1] **Imme Roesner** [1] **Florian Wendt** [1]
Jean Schoentgen [2] **Philipp Aichinger** [1*]

[1] Division of Phoniatrics and Logopedics, Department of Otorhinolaryngology, Medical University of Vienna, 1090 Vienna, Austria

[2] Department of Bio-Mechatronics, Faculty of Applied Sciences, Universite Libre de Bruxelles, 1050 Brussels, Belgium

ABSTRACT

In this study, we examined the relationship between the parameters of voice production and the perceptual aspects of vocal fry, which is a voice quality present in both healthy and disordered voices. Two perceptual experiments were conducted to investigate the impact of the fundamental frequency, open quotient, and glottal area pulse skewness on the perception of vocal fry in synthetic vowels. Thirteen listeners participated in the experiments and rated binary fry (yes/no) and attributes such as impulsiveness, tonality, and naturalness on 7-point Likert scales. The results indicate that a low fundamental frequency primarily triggers the perception of vocal fry, although the open quotient also plays a role, and narrower glottal area pulses slightly increase the probability of perceived fry. The perceived tonality is inversely related to perceived impulsiveness.

Keywords: voice quality, psychoacoustics, glottal area waveforms, vocal fry

1. INTRODUCTION

Vocal fry is a vocal quality that can be heard in both healthy and disordered voices [1-3]. It is used to mark the end of sentences, turn-taking, and social interactions [4]. In some languages, like Jalapa Mazatec, vocal fry is phonemic, indicating that the presence or absence of creaky voice can change the meaning of a sentence [5]. In disordered voices, vocal fry can be a sign of a contact granuloma or muscle tension dysphonia [6]. Therefore, assessing vocal fry is essential in the clinical care of disordered voices for treatment and monitoring of treatment effect.

Vocal fry is defined differently in terms of voice production and auditory perception, creating a gap between these two levels of description. On the voice production level, Laver noted that fry and creak are often used interchangeably [7]. He observed that fry is characterized by a low vocal fold vibration frequency, small vibrating portion of the vocal folds, low airflow rate, small subglottal air pressure, slackness of the vocal folds, and large pulse-to-pulse frequency variation. Furthermore, it is possible to contrast vocal fry with falsetto, which is characterized by a high fundamental frequency Fo and a large open quotient Qo [8, 9]. Keating et al. use "creak" to refer to various voice types, including fry [10]. The so-called prototypical creak is characterized by low vibration frequency, strong jitter, and a constricted glottis. The latter is characterized by a small peak glottal opening, a long closed phase, and a low flow rate. According to Keating, fry shares most properties with prototypical creak but has smaller jitter. They also list double and triple pulsing and pressed voice as types of creak. Gerrat and Kreiman note that vocal fry is used to describe voices with low vibration frequencies involving large variations or double pulsing irrespective of pitch [11]. Imazumi and Gauffin distinguish between creak and fry based on their position within a vowel. They stated that creak is a form of fry which occurs at the end of a vowel [12]. In terms of auditory perception, vocal fry may be evocative of the sound of a stick being run along a railing or the popping of corn, with the individual acoustic pulses being temporally segregated [13, 14]. Running a stick along a railing produces quasi-periodic pulses, whereas the popping of corn results in random pulses. For the ear to

[*] ***Corresponding author***: *philipp.achinger@meduniwien.ac.at*

perceive individual glottal pulses, there must be "pauses" between them, which means that there are sufficiently long intervals between subsequent excitation peaks. In particular, a decay of 42-44 dB SPL between pulses was reported to be the threshold above which the ear can temporally segregate individual glottal pulses [15].

In this study, we examine the relationship between parameters of voice production and perceived vocal fry. We use the definition of fry by Keating et al., which requires low vibration frequencies, low jitter, and a constricted glottis. The voice production parameters examined include the vibration frequency Fo, the open quotient Qo, and the pulse skewness Qs. We asked listeners to rate perceived impulsiveness, tonality and naturalness using 7-point Likert scales. Impulsiveness reflects the extent to which individual glottal pulses segregate temporally. At one end of the scale, the individual glottal pulses are distinct and countable, while on the other end, they merge into a continuous percept. Tonality refers to the perceived strength of pitch, and naturalness reflects how natural, i.e., human, stimuli sound.

2. MATERIALS AND METHODS

Synthesized vowels are used as test stimuli. The process of synthesis is composed of three distinct stages. During the first stage (Stage I), a phase-delayed overlapping sinusoids (PDOS) model is utilized to generate the glottal area waveform. In the second stage (Stage II), the generated glottal area waveform is integrated into the Rothenberg model, resulting in the glottal flow rate being determined by a differential equation. The third and final stage (Stage III) involves filtering the glottal flow rate through a concatenation of 25 2nd order filters that simulate the resonances of the vocal tract. Subsequently, a numerical derivative of the volume velocity is applied to simulate the radiation of the acoustic pressure at the lips. Finally, a temporal amplitude envelope is imposed, which simulates the attack, decay, sustain, and release.

Two experiments were conducted in this study. In the first experiment, the effect of vocal frequency Fo, open quotient Qo, and pulse skewness Qs on the perception of vocal fry in vowels with a homogeneous voice quality over time is investigated. The studied parameters of voice production are constant throughout a stimulus. We hypothesize that the perception of impulsiveness and the likelihood of perceiving binary fry increase, while the perception of tonality decreases when there is a decrease of vocal frequency Fo or open quotient Qo, and an increase in pulse skewness Qs.

In the second experiment, we investigate whether the perception of fry is influenced by short-term context. The stimuli used are vowels with high-low-high changes in frequency and open quotient. The study hypothesizes that increasing the variation in glottal parameters leads to a greater likelihood of perceiving fry. This means that when the differences in frequency (ΔFo) and open quotient (ΔQo) increase, the stimuli are more likely to be perceived as vocal fry and impulsive, and less likely to be perceived as tonal.

In total, 13 listeners (age ranging from 24 to 66 years, seven women and six men) participated. Five were speech therapists, five were speech scientists, and three were medical doctors. That diversity made it possible to observe differences in behavior between individuals of diverse backgrounds. The listeners were instructed to assess the tonality and impulsiveness of the stimuli and evaluate the presence or absence of vocal fry. Additionally, the perceived naturalness of the synthetic sounds was monitored by asking listeners to rate the naturalness of the stimuli.

For each of the attributes, including tonality, impulsiveness, and naturalness, a 7-point Likert scale was presented to participants via a graphical user interface (GUI). The scales ranged from -3 (representing atonal, non-impulsive, and artificial) to +3 (representing tonal, impulsive, and natural). Additionally, participants were asked to indicate whether or not they perceived the presence of fry. To familiarize participants with the range of voice qualities covered by the evaluation scales and to increase the reliability of the ratings, eight stimuli obtained by combining the most extreme parameter values were presented for anchoring purposes at the beginning of each experiment [17]. Following the anchoring phase, the main corpus was rated by participants. Then, participants were asked to rate a second set of stimuli randomly selected from the main corpus for testing intra-rater reliability. The stimuli were presented in random order within three blocks (anchoring, main corpus, repetitions), which was consistent for all participants. Participants were permitted to listen to each stimulus as many times as they wished

3. RESULTS

This section reports the results of the two listening experiments. Listener reliability and predictions of binary fry, tonality, impulsiveness, and naturalness by means of regression analyses are presented, as well as receiver operating characteristic (ROC) analyses.

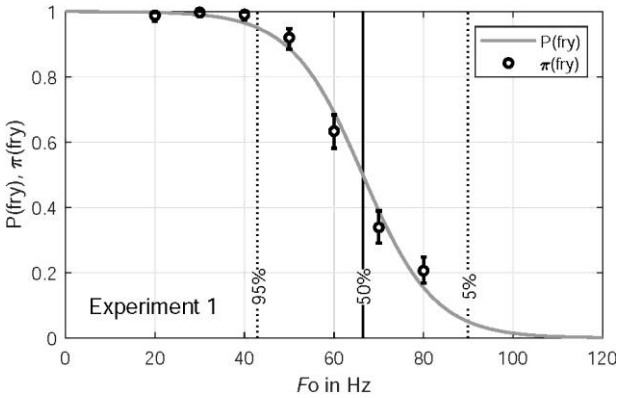
3.1 *Perceived Binary Fry Predicted from Glottal Parameters*

Logistic regression is used to model the likelihood of perceived binary fry for the two experiments, and the z-normalized coefficient estimates and p-values for the all-listeners binomial logistic regression model are shown in Table 1. In Experiment 1, i.e., without short-term context, both frequency Fo and open quotient Qo are statistically significant predictors. A decrease in Fo and Qo corresponds to an increase in the probability of perceiving binary fry. However, the coefficient values suggest that the impact of Qo is relatively small compared to that of Fo. Furthermore, the AUC (not shown) obtained with Fo as the sole predictor is 0.907, which only increases slightly to 0.91 when Qo is added as a second predictor.

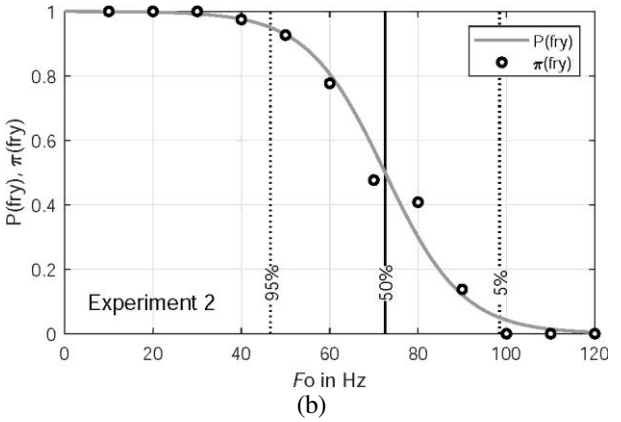
Table 1: All-listener z-normalized coefficient estimates of binomial logistic regression models designed to predict binary fry labels from glottal parameters.

	Coefficient		p values	
	Exp. 1	Exp. 2	Exp. 1	Exp. 2
Const.	1.88	0.70	<0.001	<0.001
Fo	−3.51	−2.16	<0.001	<0.001
Δ Fo	n.a.	0.41	n.a.	0.062
Qo	−0.20	−0.67	0.013	<0.001
ΔQo	n.a.	−0.08	n.a.	0.484
Qs	0.05	n.a.	0.531	n.a.

Figure 1 shows for Experiments 1 and 2 proportion π(fry) and probability P(fry) for a stimulus being labeled as fry as a function of Fo. The predicted values for P(fry) are based on binomial logistic regression, and vertical lines are included to indicate the thresholds of P(fry) at 5%, 50%, and 95%. For Experiment 2, i.e., with varying short-term context, the predicted 50% threshold of 72 Hz is higher than the corresponding threshold for Experiment 1 at 66 Hz. One possible explanation for this difference is that the anchoring approach varied between the two experiments. Specifically, Experiment 2 used a larger maximum Fo (120 Hz) than Experiment 1 (80 Hz) to more effectively sample the transition between fry and non-fry stimuli.



(a)



(b)

Figure 1: The two plots show how the perception of binary fry by all listeners change in relation to the fundamental frequency (Fo). The black circles (π(fry)) indicate the proportion of stimuli that were rated as fry. In Experiment 1, corresponding 95% binomial confidence intervals are provided. In Experiment 2, π(fry) is grouped in 10 Hz increments. The gray lines (P(fry)) represent the predicted probabilities of a fry label occurring based on binomial logistic regression models as a function of Fo. The vertical lines indicate the 95%, 50%, and 5% probability thresholds for the corresponding probabilities.

The ROC curves in Figure 2 indicate that when using Fo as the predictor, the model fitted to the data of Experiment 1 outperforms the model fitted to the data of Experiment 2. Despite this, both models show similar cutoff thresholds around 70 Hz, which approximate the 50% thresholds displayed in Figure 1.

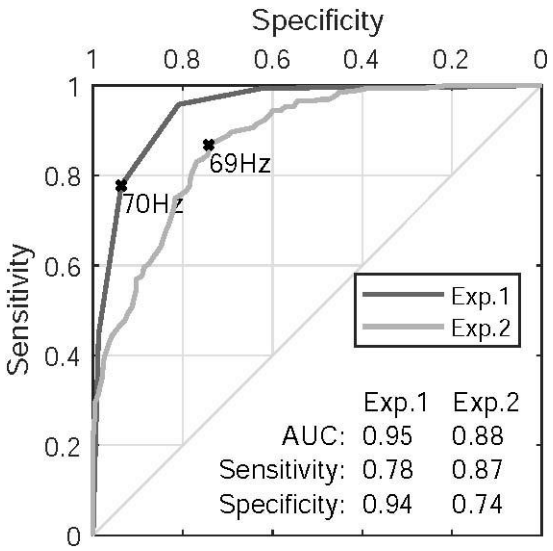


Figure 2: All-listener ROC curves regarding the prediction of perceived binary fry by **Fo** after exclusion of Listeners 4 and 5. The corresponding cutoff thresholds are 69 Hz and 70 Hz.

3.2 *Impulsiveness, Tonality, and Naturalness Predicted from Glottal Parameters*

Table 2 lists all-listener weights of the linear regression models regarding perceived impulsiveness, tonality, and naturalness as functions of Fo, Qo and Qs (Experiment 1) and Fo, ΔFo, Qo, and ΔQo (Experiment 2). The results indicate that the parameters Fo, ΔFo, and Qo have a significant impact on perceived impulsiveness, and their effect directions are consistent with those observed in binary fry (refer to Table 1). Additionally, the context parameter ΔFo, which showed only marginal significance in the binary fry ratings ($p = 0.06$), is found to be significant in the 7-point Likert scale ratings of impulsiveness

Table 2: All listeners' z-normalized coefficients estimates of linear regression models of perceived impulsiveness, tonality and naturalness. (gray highlights: $p < 0.05$)

	Impulsiveness		Tonality		Naturalness	
	Exp.1	Exp.2	Exp.1	Exp.2	Exp.1	Exp.2
Const.	0.48	0.31	−0.30	0.34	−0.21	0.32
Fo	−1.75	−1.21	1.63	1.03	1.27	0.64
ΔFo	n.a.	0.22	n.a.	−0.34	n.a.	−0.25
Qo	−0.09	−0.24	0.32	0.21	0.18	−0.04
ΔQo	n.a.	−0.03	n.a.	−0.03	n.a.	−0.02
Qs	0.01	n.a.	−0.06	n.a.	0.03	n.a.

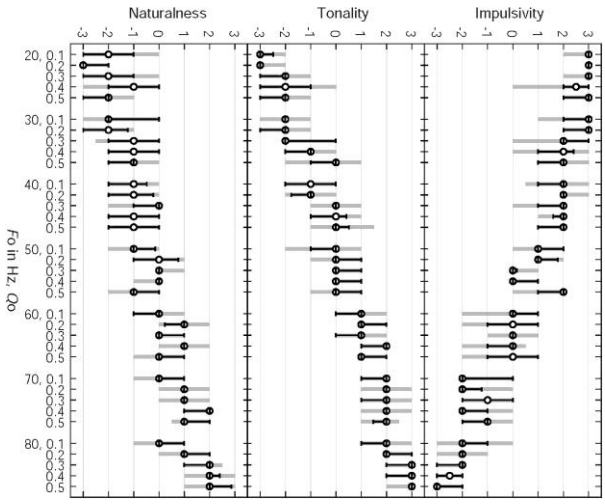


Figure 3: Factored median scores with 95% confidence intervals (black) and interquartile ranges (gray) of perceived impulsiveness, tonality, and naturalness as functions of Fo and Qo for Experiment 1.

The ratings of tonality have been found to have an inverse correlation with perceived impulsivity. The parameters Fo, Qo, and ΔFo have a significant impact on perceived tonality, but in the opposite direction to perceived impulsivity. Specifically, a decrease in Fo leads to a significant decrease in perceived tonality which aligns with previous findings for frequencies larger than 120 Hz. Decreasing Qo also decreases tonality, but to a lesser extent than Fo. Tonality and ΔFo are inversely related, possibly due to the influence of short-term context on listener perception.

In Experiment 1, the naturalness of the stimuli is observed to increase significantly with the increase in frequency Fo and open quotient Qo. This reflects the difficulty of synthesizing vocal fry that sounds completely natural (see Figure 3).

Figure 4 illustrates the ROC curves that depict the prediction of binary fry based on perceived impulsiveness ratings. The aim of the figure is to demonstrate that perceived impulsiveness can be considered as an indicator of perceived binary fry. The curves are comparable for both experiments, with substantial AUC values of 0.96 and 0.91 for Experiment 1 and 2, respectively. Therefore, perceived impulsiveness can be deemed as a reliable measure of perceived binary fry. Experiment 1 has an optimal cutoff threshold of 0, which corresponds to the midpoint of the 7-point Likert scale. On the other hand, Experiment 2 has a

higher optimal threshold of 1, leading to a smaller model sensitivity (true positive rate). Similar to the ROCs presented in Figure 2, the AUC is higher for Experiment 1 than for Experiment 2. Finally, the AUC values for both experiments increased when the individual impulsiveness ratings of listeners were substituted with majority votes, resulting in AUC values of 0.99 and 0.97 for Experiments 1 and 2, respectively (not shown here).

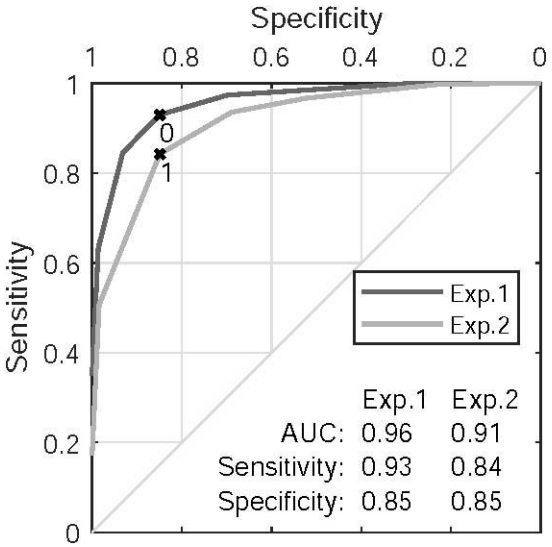


Figure 4: All-listeners' ROC curves and cutoff thresholds regarding the prediction of binary fry from perceived impulsiveness.

4. DISCUSSION AND CONCLUSION

We report two listening experiments that examine the perception of vocal fry in regular vocal fold oscillations. In the first experiment, glottal frequency (Fo), open quotient (Qo), and glottal area pulse skewness (Qs) were examined. The results indicated that Fo was the main factor affecting the perception of vocal fry, with phonation classified as fry when Fo was below a listener-individual threshold ranging from 40 Hz to 80 Hz. Qo also had a minor effect on fry perception, but Qs did not have any measurable impact. The second experiment examined the influence of short-term context on fry perception and found that in the context Fo had a small but significant effect on perceived impulsiveness, tonality, as well as naturalness. We further observed that perceived impulsiveness can serve as a proxy of perceived binary fry. Notably, perceived impulsiveness was able to predict perceived binary fry with only a small error. The main observation regarding naturalness was that as the vibration frequency increased, there was a noticeable improvement in perceived naturalness. This implies that synthesizing vocal fry with frequencies as low as 20 Hz is challenging.

The clinical long-term goals consist of three aspects. Firstly, enhancing the comprehension of the connection between voice production and perception can serve as a basis for refining clinical intervention. Secondly, acquiring knowledge on the characteristics of the auditory benchmarks of listeners is crucial for the advancement of voice quality evaluation. Lastly, distinguishing between pathological and non-pathological fry may also be a clinical long-term perspective of this research.

5. ACKNOWLEDGMENTS

This work was supported by the Austrian Science Fund (FWF): KLI 722-B30.

6. DISCLAIMER

An extended version of this paper has been published elsewhere [18].

7. REFERENCES

[1] Chen, Y.; Robb, M.P.; Gilbert, H.R. Electroglottographic evaluation of gender and vowel effects during modal and vocal fry phonation. J. Speech Lang. Hear. Res. 2002, 45, 821–829. https://doi.org/10.1044/1092-4388(2002/066).

[2] Michel, J.F.; Hollien, H. Vocal fry as a phonational register. J. Speech Hear. Res. 1968, 11, 600–604.

[3] Hollien, H. On vocal registers. J. Phon. 1974, 2, 125–143.

[4] Wolk, L.; Abdelli-Beruh, N.B.; Slavin, D. Habitual use of vocal fry in young adult female speakers. J. Voice 2012, 26, e111–e116.

[5] Ashby, M.; Maidment, J. Introducing Phonetic Science; Cambridge University Press: Cambridge, UK, 2005.

[6] Patel, R.; Liu, L.; Galatsanos, N.; Bless, D.M. Differential vibratory characteristics of adductor spasmodic dysphonia and muscle tension dysphonia on high-speed digital imaging. Ann. Otol. Rhinol. Laryngol. 2011, 120, 21–32.

[7] Laver, J. The Phonetic Description of Voice Quality; Cambridge University Press: Cambridge, UK, 2009; 200p.

[8] Henrich, N.; d'Alessandro, C.; Doval, B.; Castellengo, M. Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency. *J. Acoust. Soc. Am.* **2005**, *117*, 1417–1430. https://doi.org/10.1121/1.1850031.

[9] Herbst, C.T.; Ternström, S.; Švec, J. Investigation of four distinct glottal configurations in classical singing—A pilot study. *J. Acoust. Soc. Am.* **2009**, *125*, EL104–EL109. https://doi.org/10.1121/1.3057860.

[10] Keating, P.A.; Garellek, M.; Kreiman, J. Acoustic properties of different kinds of creaky voice. In Proceedings of the 18th International Congress of Phonetic Sciences, ICPhS, Glasgow, UK 10-14 August, 2015; pp. 2–7.

[11] Gerratt, B.R.; Kreiman, J. Toward a taxonomy of nonmodal phonation. *J. Phon.* **2001**, *29*, 365–381.

[12] Imaizumi, S.; Gauffin, J. Acoustical and perceptual characteristics of pathological voices: Rough, creak, fry, and diplophonia. *Ann. Bull. RILP* **1991**, *25*, 109–119.

[13] Catford, J.C. In honour of Daniel Jones: Papers contributed on the occasion of his eightieth birthday.1964.

[14] Klatt, D.H.; Klatt, L.C. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.* **1990**, *87*, 820–857.

[15] Coleman, R.F. Decay Characteristics of Vocal Fry. *Folia Phoniatr. Logop.* **1963**, *15*, 256–263. https://doi.org/10.1159/000262970.

[16] Blomgren, M.; Chen, Y.; Ng, M.L.; Gilbert, H.R. Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. *J. Acoust. Soc. Am.* **1998**, *103*, 2649–2658. https://doi.org/10.1121/1.422785.

[17] Chan, K.M.K.; Yiu, E.M.-L. The effect of anchors and training on the reliability of perceptual voice evaluation. *J. Speech Lang. Hear. Res.* **2002**, *45*, 111–126. https://doi.org/10.1044/1092-4388(2002/009).

[18] Devaraj, V., Roesner, I., Wendt, F., Schoentgen, J., and Aichinger, P. Auditory perception of impulsiveness and tonality in vocal fry. *Appl. Sci. Basel*, **2023**, 13(7), p. 4186, doi: 10.3390/app13074186.