Modelling of Amplitude Modulated Vocal Fry Glottal Area Waveforms Using an Analysis-by-Synthesis Approach

Vinod Devaraj [1,2,*] and Philipp Aichinger [1]

1   Department of Otorhinolaryngology, Division of Phoniatrics-Logopedics, Medical University of Vienna, 1090 Vienna, Austria; philipp.aichinger@meduniwien.ac.at
2   Signal Processing and Speech Communication Laboratory, University of Technology, 8010 Graz, Austria
*   Correspondence: vinod.devaraj@meduniwien.ac.at

Abstract: The characterization of voice quality is important for the diagnosis of a voice disorder. Vocal fry is a voice quality which is traditionally characterized by a low frequency and a long closed phase of the glottis. However, we also observed amplitude modulated vocal fry glottal area waveforms (GAWs) without long closed phases (positive group) which we modelled using an analysis-by-synthesis approach. Natural and synthetic GAWs are modelled. The negative group consists of euphonic, i.e., normophonic GAWs. The analysis-by-synthesis approach fits two modelled GAWs for each of the input GAW. One modelled GAW is modulated to replicate the amplitude and frequency modulations of the input GAW and the other modelled GAW is unmodulated. The modelling errors of the two modelled GAWs are determined to classify the GAWs into the positive and the negative groups using a simple support vector machine (SVM) classifier with a linear kernel. The modelling errors of all vocal fry GAWs obtained using the modulating model are smaller than the modelling errors obtained using the unmodulated model. Using the two modelling errors as predictors for classification, no false positives or false negatives are obtained. To further distinguish the subtypes of amplitude modulated vocal fry GAWs, the entropy of the modulator's power spectral density and the modulator-to-carrier frequency ratio are obtained.

Keywords: voice quality; glottal area waveforms; vocal fry; irregular phonation

1. Introduction

Vocal fry is a voice quality which is synonymously referred to as creaky voice, pulse register, glottal fry or creak [1–3]. In addition, the term vocal fry is used to designate a subtype of creaky voice [4]. The remaining subtypes are multipulsed voice, aperiodic voice, nonconstricted creak and tense/pressed voice. Vocal fry is mainly characterized by a low fundamental frequency which gives an auditory impression of "a stick being run along a railing", "popping of corn" or "cooking of food on a pan" [1,2,5]. Other characteristics include shimmer, jitter, and damping of pulses. Furthermore, subglottal air pressure and air flow were found to be smaller in vocal fry than in modal registers [2]. However, fundamental frequency is one of the main factors which distinguishes vocal fry from modal and harsh voice [6]. In our work, we identify vocal fry based on the impulsivity of voice samples, i.e., the auditory attribute associated with the separate perception of the glottal cycles. It was shown in the past that the peak prominence observed in loudness versus time curves may reflect the temporal auditory segregation of the glottal cycles in vocal fry [7].

Vocal fry has been identified in healthy and dysphonic speakers. In particular, vocal fry was reported to be one of the salient voice characteristics of male patients with contact granuloma [8]. Hence, vocal fry may be an indicator for the presence of a pathology which makes the study of characteristics of vocal fry clinically important. Although it can be a sign of a voice disorders, Hollien et al. reported vocal fry voice quality in nonpathological voice [9,10]. These studies investigated the fluctuation of pulse rate of vocal fry samples,

and evidence against classifying vocal fry as pathological is provided. The results were further supported by an investigation of vocal fry with 104 first year graduate students [11]. Of the students in the study group, 86% had a nonpathological voice and the remaining 14% were reported to use vocal fry along with a pathological voice. However, 16% of the students with nonpathological voice quality also used vocal fry.

For the objective detection of vocal fry and creak, these following acoustic features have been used in the past. The presence of vocal fry segments in speech utterances was detected based on the autocorrelation properties of the audio signals [3]. The insertion and detection rates obtained were 13% and 74%, respectively. In [12], audio features like inter-frame periodicity, interpulse similarity, peak fall and peak rise, H2-H1, i.e., the difference in amplitudes of the first two harmonics, F0 contours in each frame and peak prominence were used for vocal fry detection. An aperiodicity, periodicity and pitch (APP) detector, which uses a dip profile of the average magnitude difference function (AMDF) at lags, was used to distinguish between irregular phonation and modal phonation [13]. A Fourier spectrum analysis approach of the audio signals was also proposed for distinguishing vocal fry segments from diplophonic voice [14]. The ratio of the numbers of harmonics present in consecutive frames was determined to detect transitions of modal to vocal fry regimes, or vice versa. In total, 81% of the voice segments were correctly detected as vocal fry. The distinct characteristics of the creaky and modal regions were obtained using so called epoch parameters [15]. Epochs are negative to positive zero crossing instances of zero frequency filtered audio signals. The used epoch parameters were the number of epochs in a frame, the time instants of epochs, and the time interval between successive epochs. Based on the variance of these epoch parameters, a neural network classifier was developed to identify creaky regions. This study was done using voice samples of several speaker groups. The F1 scores were analyzed to validate the detector. The maximum F1 score was found to be a little over 0.8 for female Finnish speakers when compared to the other groups. Convolutional neural networks (CNN) based on the detection of the creaky voice from emergency calls was also proposed, where 32 mel-scale log filter-bank outputs were used as input features [16]. The data consist of emergency call recordings and a conversational speech corpus with 30 recordings each. Though these methods detect vocal fry or creaky segments, they do not allow a detailed study of voice production.

Previous studies have tried to interlink vocal fold vibration patterns with voice quality types. Several studies have investigated the vibration patterns of vocal fry, which include the number of opening and closing phases of the vocal folds in a single modulator cycle, and the duration of the closed phase. The authors of [17] and [18] found single- and double-pulsed patterns, respectively, using high-speed videos, where the closed phase is longer than the open phase. More evidence for the existence of multiple pulses in a single cycle was reported in [1,2,5]. Furthermore, multiple pulses in a single cycle without a long closed phase were observed using electroglottograms which reflect translaryngeal electrical resistance that is proportional to the contact area of the vocal folds [19]. In this paper, we observe vocal fry GAWs with multiple single-peak pulses in a single cycle. A GAW is the time series of the glottal area, i.e., the projected area of the space between the vocal folds, which vibrate during phonation. In the open phase, the magnitude of the GAW monotonically increases and decreases while the vocal folds are opening and closing, respectively. The magnitude becomes zero when the folds are closed completely, i.e., during the closed phase. The pulses are amplitude modulated and no long closed phase is observed. Throughout the paper, we refer to single-peak pulses. We model the GAWs using an analysis-by-synthesis approach. A modified version of the analysis-by-synthesis approach implemented in [20] for modelling diplophonic GAWs is used. Here, we use two supporting points per pulse of GAW instead of one while modelling the input GAWs. This helps in improving the estimation of the upper and lower envelopes,—discussed in Section 2.4. Furthermore, instead of two fundamental frequencies, one fundamental frequency is used for modelling GAWs.

In this study, GAWs are used to distinguish between vocal fry and normal voice quality. The remainder of this article is structured as follows. In Section 2, the extraction of GAWs, the different types of observed amplitude modulated GAWs, and the modelling of GAWs using analysis-by-synthesis are described. The results of automatic classification of amplitude modulated vocal fry GAWs and euphonic GAWs are reported in Section 3. Section 4 discusses the performance, limitations and the possible future works of this study. Finally, Section 5 concludes the paper.

2. Materials and Methods

2.1. Data

High-speed videos (HSVs) of the vocal folds with a frame rate of 4000 frames per second are obtained by inserting a laryngeal endoscopic camera (HRES ENDOCAM 5562) through the mouth into the pharynx [21]. Subjects are instructed to produce sustained phonation of vowels for a duration of two seconds while the vocal fold vibrations are filmed. Audio files are recorded simultaneously using a head worn microphone. The audio files are synchronized with the captured video files. The voice samples are annotated by three expert annotators with regard to presence or absence of vocal fry. Based on the annotations, the corresponding video frames of the HSVs are cropped and used for GAW extraction. The GAWs are extracted using a seeded region growing segmentation algorithm [22]. Seven GAWs annotated as vocal fry contain amplitude modulations without long phases. They are used as a positive group. Eight euphonic i.e., normophonic GAWs are used as a negative group.

The vocal fry GAWs are further distinguished into two categories based on the cyclicity of the GAWs by visual inspection. Three out of seven amplitude modulated GAWs are observed to be acyclically modulated. These GAWs are termed as Acyclically Amplitude Modulated Pulse train (AAMP). The remaining four vocal fry GAWs are observed to have a cyclic modulator. These cyclic GAWs are further subclassified based on the number of pulses in each modulator cycle, one of which is characterized by double-pulsing, i.e., alternation of pulse magnitudes, timing, and/or shapes. This type is termed as Cyclically Amplitude Modulated Pulse train-2 (CAMP = 2). The other cyclically modulated type is characterized by more than two pulses per cycle. This type is termed as Cyclically Amplitude Modulated Pulse train = 4 (CAMP = 4). Along with the natural GAWs, 1200 synthetic GAWs are used. Of the synthetic GAWs, 300 belong to the euphonic group and the remaining 900 are vocal fry GAWs, with each of the vocal fry type having 300 GAWs. Synthetic GAWs are generated using parameter distributions obtained from estimated parameters of the natural GAWs.

Regularity and symmetry of vocal fold vibration is analyzed using phonovibrograms (PVGs). A PVG displays the deflections of the vocal folds from the glottal axis represented as the color intensities [23]. The brighter the color intensity, the farther away are the vocal folds from the glottal axis. Figure 1 shows PVGs of the GAWs shown in Figure 2, i.e., of the amplitude modulated vocal fry and euphonic types of vocal fold vibration. A negligible anterior–posterior asymmetry of vocal folds is observed in the PVG belonging to CAMP = 4 in Figure 1a. Here, each modulator cycle contains four pulses where the borders of one of the modulator cycles is marked by two vertical green lines. The vibration pattern is described as an increase in the amplitude of the vocal fold vibration for four consecutive glottal pulses followed by a sudden decrease in amplitude. In contrast, the anterior–posterior phase difference is observed in PVGs of CAMP = 2, AAMP, and the euphonic type. In Figure 1b, the anterior–posterior phase difference in every second pulse of the vocal folds results in a period-doubled GAW, as shown in Figure 2b.
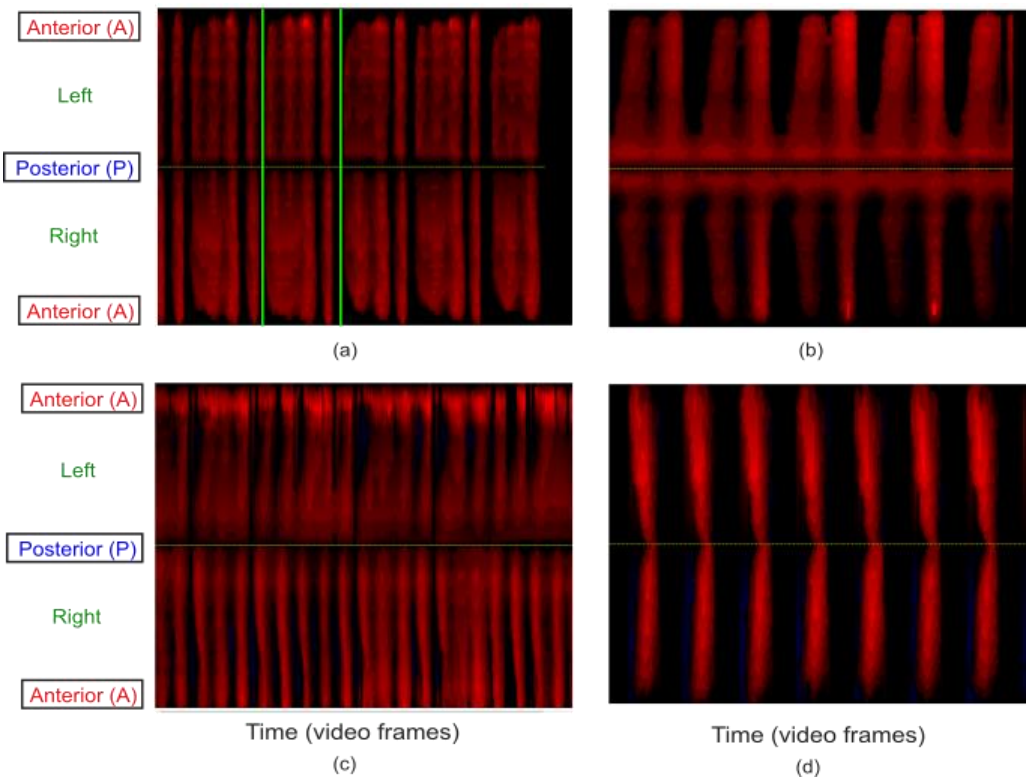
Figure 1. Phonovibrogram (PVG) representation of vocal fold kinematics of the GAWs shown in Figure 2: (**a**) CAMP = 4, (**b**) CAMP = 2, (**c**) AAMP and (**d**) euphonic type.
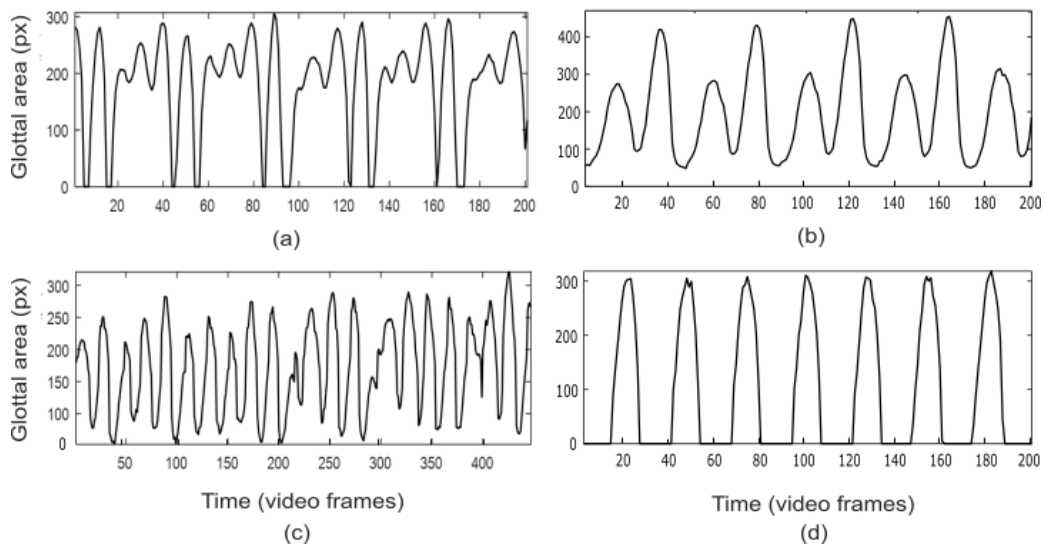
Figure 2. Glottal area waveforms (GAWs) extracted by segmentation of the high-speed videos (HSVs): (**a–c**) vocal fry and (**d**) euphonic voice. (**a**) illustrates an example of a CAMP = 4 type GAW, having four pulses in a modulator cycle; (**b**) illustrates an example of CAMP = 2 type, where each modulator cycle contains two pulses; and (**c**) belongs to an example of the Acyclically Amplitude Modulated Pulse train (AAMP) type, which is acyclically modulated. The euphonic GAW has smaller amplitude modulations than the vocal fry GAWs.

Figure 3 shows an overview of the analysis and synthesis combined with parameter distribution fitting. First, the analyzer models the natural GAW $y(t)$ (input) using an analysis-by-synthesis approach, which is described in Section 2.2. A Fourier synthesizer is used to obtain the modelled GAWs $\hat{y}(t)$. The modelling error E is obtained from the root

mean square error between the input GAW and the modelled GAW $\hat{y}(t)$. The modelled GAW is refined iteratively to obtain a minimum modelling error. Parameters of the input GAW ψ_i are also estimated during the modelling process. Distributions of estimated parameters of the input GAW are fitted as described in Section 2.3 for the purpose of generating the synthetic corpus, which is larger than the corpus of the natural GAWs. Section 2.4 explains the Fourier synthesizer in detail. The synthetic GAWs $y_s(t)$ are again input into the analyzer. The modelling errors obtained for the natural and synthetic GAWs are used for the classification of the positive and the negative group. In Section 2.5, the features and classification are explained.
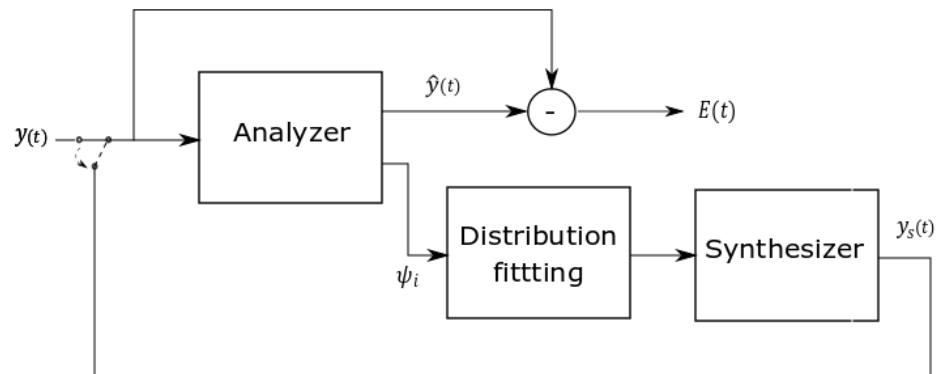
Figure 3. Overview of the analyzer and the synthesizer used for modelling and synthesizing GAWs. The analyzer models and estimates parameters of input GAWs, which is described in Section 2.2. The synthesizer described in Section 2.4 generates synthetic GAWs $y_s(t)$ is based on the distributions of the estimated parameters ψ_i that are fitted as described in Section 2.3.

2.2. Analyzer

Figure 4 shows the block diagram of the analysis-by-synthesis approach used in the analyzer to model input GAWs. A similar approach was proposed in the past to model diplophonic GAWs [20] where two simultaneous fundamental frequency (f_o) tracks were used. In this study, instead of two f_o tracks, only one f_o track is estimated. First, the f_o track of the input GAW is extracted by a hidden Markov model (HMM) combined with repetitive execution of a Viterbi algorithm. An unmodulated quasi unit pulse train u_t is generated by an oscillator driven by the extracted f_o track. The pulse locations of this pulse train approximate the time instants of the maxima of the input GAW. An additional pulse train for indicating the locations of the minima of the input GAW is obtained by phase shifting the train's phase by $180°$. The instantaneous phase and amplitude of the maxima and minima are extracted from the quasi unit pulse train, which is explained in detail in Section 2.5.
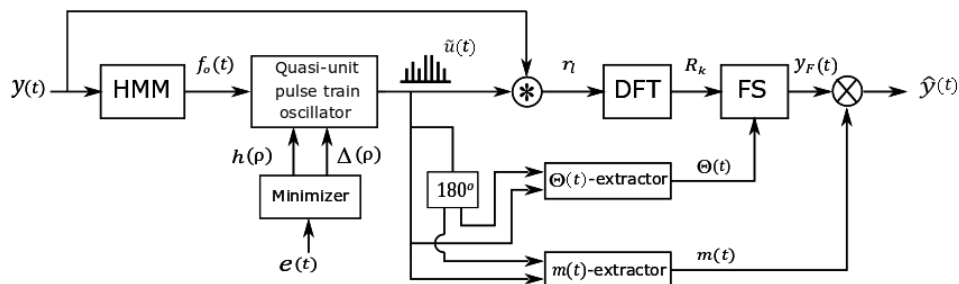
Figure 4. The modified version of the analyzer used for modelling GAWs, modified from [20].

A Fourier synthesizer (FS) uses the extracted instantaneous phase $\Theta(t)$ and the Fourier coefficients R_k obtained from pulse shapes r_l to model the input GAW. The pulse shapes

are obtained by cross-correlating the quasi unit pulse with each block of the input GAW of length 32 ms obtained using a Hanning window with a 50% overlap. These single-peak pulse shapes are then modelled using Chen's model [24]. Chen's model estimates the pulse parameters which are used for generating synthetic GAWs. The pulse shapes are transformed to the Fourier coefficient R_k using the discrete Fourier transform (DFT). The synthesized GAW $y_F(t)$ is obtained by the Fourier synthesizer using the Fourier coefficients of the pulse shapes and the extracted instantaneous phase. The synthesized GAW $y_F(t)$ is multiplied with an amplitude modulator $m(t)$ to obtain a modelled GAW,

$$\hat{y}(t) = y_F(t) \cdot m(t) \tag{1}$$

The modelling error E is the level of the root mean squared difference between the input GAW $y(t)$ and the modelled GAW $\hat{y}(t)$ i.e.,

$$E(t) = 20 \cdot log_{10} \left(\frac{\sqrt{\overline{(y(t) - \hat{y}(t))^2}}}{\sqrt{\overline{y^2(t)}}} \right) \quad [dB] \tag{2}$$

$\hat{y}(t)$ obtained using the unmodulated quasi unit pulse train u_t is the output of a nonmodulating model where the frequency and amplitude modulation present in the original GAWs are not modelled. Random modulations of the individual pulses of the input GAW are modelled by minimizing the error E via modulating the quasi unit pulse trains, which approximates a nonlinear system in a linear way. Pulse heights and time instants of the unit pulse trains are modulated on a pulse-to-pulse time scale. Additionally, the modelled GAWs are clipped at the baseline to minimize the modelling error E. The amplitude modulation vector $h(\rho)$ and the pulse time modulation vector $\Delta(\rho)$ of the quasi unit pulse train are estimated iteratively by minimizing the time domain modelling error $E(t)$. ρ is the pulse index.

2.3. Distribution Fitting

The parameters of the pulse shape of the estimated input GAW are used for generating synthetic GAWs. These parameters are fitted to Gaussian distributions independently of each other and random numbers are drawn from the distribution to obtain original synthesis parameters. A is a parameter matrix containing N rows and P columns. N is the number of rows of a single GAW type, and P is the number of frames in each GAW, which may be different for each GAW. First, the mean μ_A and the standard deviation σ_A of the matrix is taken over all frames for each GAW, i.e.,

$$\mu_A = \frac{\sum_{i=1}^{P} A_i}{P} \tag{3}$$

$$\sigma_A = \sqrt{\frac{\sum_{i=1}^{P} (A_i - \mu_A)^2}{P}} \tag{4}$$

These two measures describe the distribution of a parameters observed in a GAW. To reflect inter-GAW variation of the parameter distribution, means and standard deviations of μ_A and σ_A are also obtained.

$$\mu_{\mu_A} = \frac{\sum_{i=1}^{N} \mu_{A,i}}{N} \tag{5}$$

$$\sigma_{\mu_A} = \sqrt{\frac{\sum_{i=1}^{N} \left(\mu_{A,i} - \mu_{\mu_A} \right)^2}{N}} \tag{6}$$

$$\mu_{\sigma_A} = \frac{\sum_{i=1}^{N} \sigma_{A,i}}{N} \tag{7}$$

$$\sigma_{\sigma_A} = \sqrt{\frac{\sum_{i=1}^{N}(\sigma_{A,i} - \mu_{\sigma_A})^2}{N}} \tag{8}$$

In particular, two Gaussian distributions are estimated indicating the variation of the means and standard deviations of a GAW type's parameter over time. From each of these two Gaussian distributions $\mathcal{N}(\mu, \sigma)$ with mean μ and standard deviation σ, M random numbers are drawn.

$$\mu_{A_s,M} \sim \mathcal{N}\left(\mu_{\mu_A}, \sigma_{\mu_A}\right) \tag{9}$$

$$\sigma_{A_s,M} \sim \mathcal{N}\left(\mu_{\sigma_A}, \sigma_{\sigma_A}\right) \tag{10}$$

These M random pairs of μ_{A_s} and σ_{A_s} indicate the means and standard deviations of the parameters for synthesizing M GAWs. {\displaystyle {\mathcal {N}}(\mu,\sigma ^{2})}Using each pair of μ_{A_s} and σ_{A_s}, a Gaussian distribution is obtained. From this distribution, R numbers are drawn which indicate the values of a parameter for R frames.

$$A_{s,M} \sim \mathcal{N}(\mu_{A_s}, \sigma_{A_s}) \tag{11}$$

All parameters of the pulse shape are generated using the same process, but different distribution parameters. The parameters are input to the synthesizer for generating synthetic GAWs.

2.4. Fourier Synthesizer

This subsection explains the Fourier synthesizer. The Fourier synthesizer generates $y_F(t)$ using the estimated locations and amplitudes of the extrema of the input GAW and the Fourier coefficients obtained by transforming the pulse model. The location of two quasi unit pulse trains approximate the maxima and minima of the individual pulses of the input GAW. They are the supporting points at which the instantaneous phase is equal to odd and even integer multiples of π, i.e., (π, 3 π, 5 π... and 2 π, 4 π, 6 π ...), respectively. These supporting points are sorted and interpolated to obtain $\Theta(t)$. Instantaneous phase $\Theta(t)$ and the Fourier coefficients R_k obtained by transforming prototype pulses are provided to the Fourier synthesizer to synthesize the GAWs.

The amplitude of the modelled GAW is modulated by multiplying the carrier $y_F(t)$ by an estimated modulator $m(t)$. The modulator is obtained as a superposition of the estimated upper envelope $m_u(t)$ and the lower envelope $m_l(t)$, i.e.,

$$m(t) = \left(\frac{1 + y_F(t)}{2}\right) \cdot m_u(t) - \left(\frac{-1 + y_F(t)}{2}\right) \cdot m_l(t) \tag{12}$$

The envelopes are estimated by interpolating the pulse heights of the modulated unit pulse trains using shape preserving piecewise cubic interpolation. The carrier $y_F(t)$ is obtained as the negative cosine of the estimated instantaneous phase $\Theta(t)$. Hence, at times when $y_F(t)$ is +1 and -1, the modulator is set equal to the upper and lower envelope respectively. At times in between, (12) enables smooth transition.

To obtain the signals $y_F(t)$ used in the generation of the synthesized corpus, the same synthesizer is used, but instead of using as input the parameters estimated from an input GAW, parameters drawn from the distributions explained in Section 2.3 are used.

2.5. Features and Classification

Modelling errors are used as features to distinguish between the positive and the negative group. E_{unmod} is the modelling error obtained using an unmodulated quasi unit pulse $u(t)$ train and E_{mod} is the improved modelling error obtained using a modulated pulse train $\tilde{u}(t)$. An SVM classifier with a linear kernel is used for classifying the positive (vocal fry GAWs) and the negative (euphonic GAWs) groups with the two modelling errors as predictors.

The subtypes of AMS-type vocal fry GAWs are further distinguished based on the cyclicity of their modulators. The modulators for CAMP = 2 and CAMP = 4 are cyclic, whereas the modulators are acyclic in AAMP type GAWs. As a result, the magnitude spectra of the acyclic modulators are more homogeneous across frequencies than the magnitude spectra of the cyclic modulators. The homogeneity of a magnitude spectrum is reflected by Shannon's Entropy

$$H = -\frac{\sum_{f=1}^{N} S(f) \, log_2 S(f)}{log_2 N_f} \tag{13}$$

where $S(f) = \frac{G(f)}{\sum_f G(f)}$, is the probability distribution of the modulators' power spectrum $G(f) = |X(f)|^2$, $X(f)$ is the discrete Fourier transform of the modulator and N_f is the total number of discrete frequencies f. To distinguish between CAMP = 2 and CAMP = 4 GAWs, the modulator-to-carrier frequency ratio $r_f = f_m / f_c$, where the modulator frequency f_m is the frequency corresponding to the maximal magnitude of the modulator's spectrum and f_c is the carrier frequency estimated using the time instances of the quasi unit pulse train. A pairwise statistical comparison is conducted to check the variability of the two features, i.e., the entropy of the modulator spectrum and the r_f ratio, in terms of p-values between the different subtypes of vocal fry GAWs.

3. Results

Figure 5 shows waveforms involved in the modelling of an example of an amplitude modulated natural vocal fry GAW using the nonmodulating model (top plot) and the modulating model (bottom plot). The input GAW is an amplitude modulated signal with four pulses within a modulator cycle (CAMP = 4). The modulating model has a modelling error of −11.55 dB for the example GAW, which is larger than the modelling error obtained using the nonmodulating model (−2.15 dB). A smaller modelling error indicates a better fit to the input GAW.
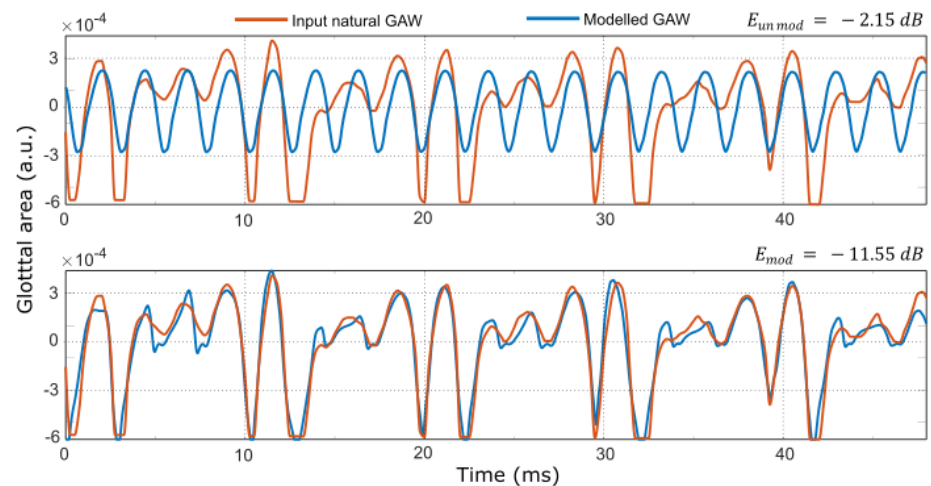
Figure 5. Modelling of an example of an amplitude modulated vocal fry GAW using the nonmodulating model (top plot) and the modulating model (bottom plot).

The modulating model is observed to fit the input GAW better than the nonmodulating model. In particular, the nonmodulating model fails to track the jitter and shimmer (instantaneous frequency and amplitude modulations) of the individual pulses of the input GAW. As a result, the individual extrema of the estimated unit pulse trains (not shown here) are equally high, which results in constant upper and lower envelopes. Hence, the modulator estimated using these constant envelopes has negligible fluctuation. On the other hand, in the modulating model, the quasi unit pulse trains track the instantaneous

frequency and amplitude modulation of the individual pulses of the input GAW. The modulator estimated using these quasi unit pulse trains fluctuates in accordance with the input GAW, which results in a modelling error smaller than the error achieved when using the nonmodulating model.

Figure 6 shows waveforms involved in the modelling of an example of a euphonic GAW using the nonmodulating model (top plot) and the modulating model (bottom plot). The modulation is negligible as compared to what is observed in vocal fry GAW in Figure 5. Thus, the GAWs modelled by the two models are similar which makes the difference between the modelling errors smaller compared to the difference between the modelling errors obtained for the GAW shown in Figure 5.
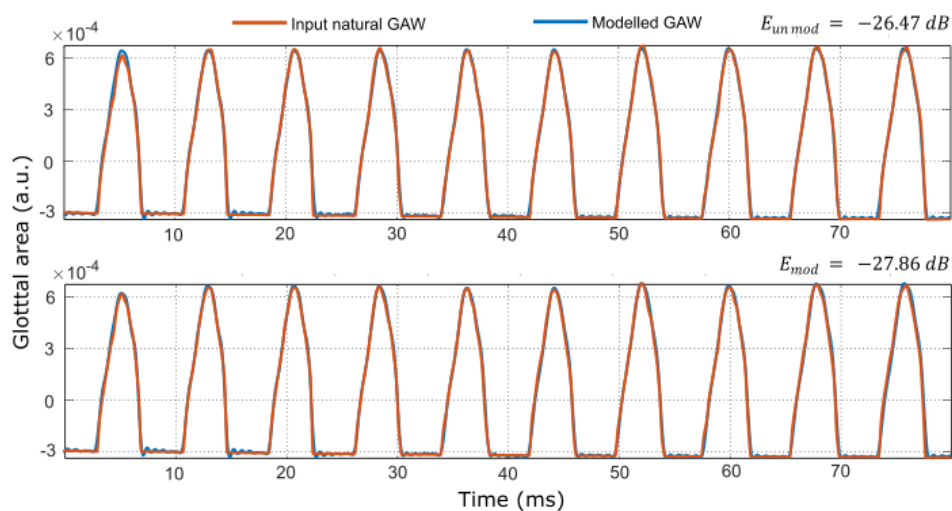
Figure 6. Modelling of an example of a natural euphonic GAW using the nonmodulating model (top plot) and the modulating model (bottom plot).

E_{mod} vs. E_{unmod} are the two features used for classifying amplitude modulated vocal fry GAWs (positive group) and euphonic (negative group). Figure 7 shows the scatter plots of the two modelling errors plotted against each other for natural (a) and synthetic (b) data. For vocal fry GAWs, the use of the modulating model results in modelling errors smaller than the modelling errors achieved when using the nonmodulating model. The line of equality (LoE) is dashed. Modelling errors of the euphonic GAWs are well separated from the modelling errors of vocal fry GAWs in the feature space. An SVM classifier with a linear kernel acheives a 5-fold cross validated accuracy of 100% for natural and synthetic GAWs. The performance of the classification is reported using sentivities, specificities and accuracies with 95% confidence intervals (CI) in Table 1. For both natural and synthetic GAWs, no false positives or false negatives are observed.

Table 1. Sensitivities, specificities and accuracies of classification between vocal fry and euphonic GAWs with 95% confidence interval (CI).

GAW	Sensitivity		Specificity		Accuracy	
	Value	95% CI	Value	95% CI	Value	95% CI
Natural	100%	59.04% to 100.00%	100%	63.06% to 100.00%	100%	78.20% to 100.00%
Synthetic	100%	99.59% to 100.00%	100%	98.78% to 100.00%	100%	99.69% to 100.00%
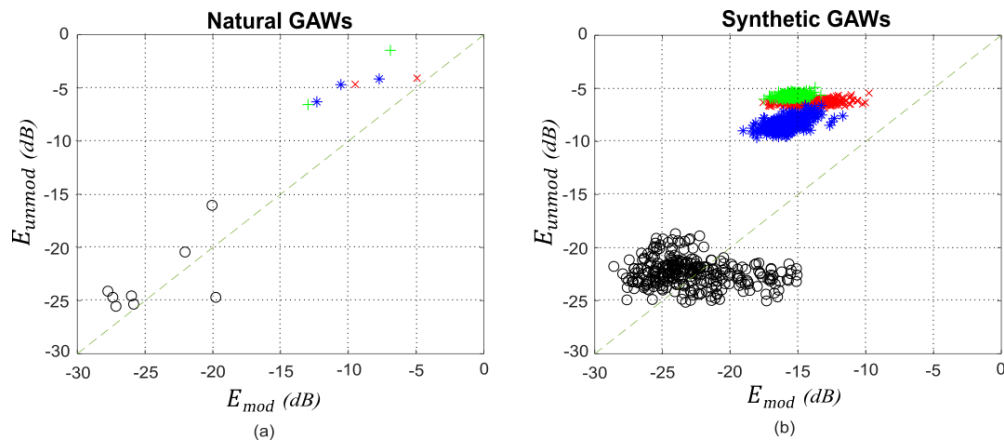
Figure 7. Scatter plots of the modelling errrors vocal fry GAWs and other types of GAWs: (**a**) natural and (**b**) synthetic GAWs.

The subtypes of amplitude modulated vocal fry GAWs are distinguished using the entropy of the modulator's power spectral density (PSD) and the modulator-to-carrier frequency ratio r_f. Figure 8 shows box plots of the entropy and r_f values of natural and synthetic vocal fry GAW subtypes. The entropy of the modulator's PSD is used to distinguish between cyclically and acyclically modulated vocal fry GAWs. The entropy of the modulators' PSD is larger for modulator with a homogeneous spectrum than for a modulator with inhomogeneous spectrum. This results in larger modulator spectrum entropies for AAMP than for CAMP. An Anova test is performed to check the variance of entropy between the subtypes of vocal fry GAWs. The p-values obtained between CAMP (CAMP = 2 and CAMP = 4) and AAMP are less than 0.01, which indicates significant differences between the group means of CAMP and AAMP. For the two cyclically modulated groups, the obtained p-value is 0.2526. Therefore, a distinction is possible between CAMP and AAMP type GAWs based on the entropies of the modulator spectrum. To make a distinction between the two subtypes of cyclically modulated vocal fry GAWs, the r_f ratio is used. The r_f ratio for double pulsing (CAMP = 2) is approximately 0.5 and for quadruple pulsing (CAMP = 4) is approximately 0.25. In double pulsing, a modulator cycle contains two pulses, whereas in quadruple pulsing, a modulator cycle contains four pulses. The Anova test for the r_f ratio resulted in a p-value of less than 0.05, which indicates a significant difference between the cyclically modulated vocal fry GAWs.
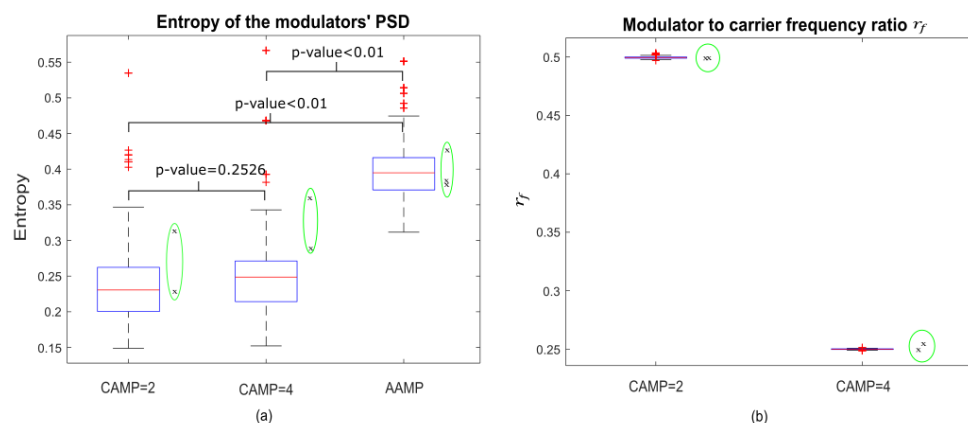
Figure 8. Features used for distinguishing the subtypes of synthetic vocal fry GAWs: (**a**) entropy of the modulators' power spectral density (PSD) for distinguishing cyclically vs. acyclically modulated vocal fry and (**b**) modulator-to-carrier frequency ratio r_f to distinguish CAMP = 2 and CAMP = 4 type vocal fry GAWs. The features of natural GAWs are represented using"x" encircled in green.

4. Discussion

In this study, the detection of vocal fry based on characterizing the variation observed in the vibration patterns of the vocal folds is proposed. The results of the classification accuracies suggest that the proposed model enables the distinction of normal and vocal fry GAWs. The obtained classification accuracies of detection were found to be competitive with the accuracies reported for past detection techniques proposed in [3,12–14]. In these past techniques, the detection rates of vocal fry were less than 90% and the models were based mainly on audio signals, whereas in our study, the high-speed videos of vocal folds enable a more detailed study of voice production. This may help in studying the cause–effect relationship between voice production and perception, and aid the diagnosis of voice disorders.

Bailly et al. investigated the ventricular fold dynamics in human phonation [25,26]. A correlation between the vibration of the ventricular and vocal folds was demonstrated using time derivative of EGG signals (DEEG) [25]. A periodic contact between the ventricular folds was observed for every second glottal cycle which resulted in a period-doubling pattern. In [26], ventricular motion was studied using synchronized high-speed cinematographic and audio recordings. The recorded audio sequences consist of glides with transition between different voice qualities. In most of vocal fry sequences, the ventricular folds were observed to follow a slow nonoscillatory motion with a partial ventricular contact, or a total contact. However, in our study, no significant ventricular motion was observed in the high-speed videos. The audio sequences were also recorded for sustained phonation of vowels without any glides. The cyclic modulation of the vibration patterns of vocal fry which we observed using high-speed videos were only due to distinct vocal fold vibration behavior. For CAMP = 2 and CAMP = 4 type GAWs, in each modulator cycle, the peak area of the glottis increased after every subsequent opening of the vocal folds. It is suggested that the myoelastic aerodynamic cause of this phenomenon is to be studied in the future.

A central aspect of our study is the analysis of noise contained in the GAWs. The GAWs extracted from the videos contain modulation noise (i.e., jitter and shimmer) and quantization noise. We modelled the modulation noise on a pulse-to-pulse time scale by modulating the quasi unit pulse train that the model uses. The additive quantization noise arises from the finite pixel resolution of the videos and is part of the modelling error levels E. In particular, depending on the relative size of the glottis in the video, a GAW is quantized to only a few hundred available values, i.e., a discrete number of pixels. We focused in this study on modelling modulation noise instead of additive quantization noise, because (i) the modulations and their properties were used to directly describe the voice types that we report, and (ii) the amount of modulation noise in voices labeled as vocal fry exceeded the amount of additive quantization noise by an order of a magnitude.

Added values of the presented approach include the following. First, analysis-by-synthesis combined with parameter distribution fitting provides a means of data augmentation. That is particularly relevant because not much natural data is available yet. The selection of the synthesis parameter distributions that we fitted to create the synthetic, i.e., augmented corpus appears to be reasonable since the scatter plots of modelling errors as well as the boxplots of the entropy of the modulator's spectrum and the modulator-to-carrier frequency ratio reflect comparable distributions for natural and synthetic data. Second, we opted for a detailed signal modelling based approach for classification instead of using a black box approach. This has the advantages that the classification appears to be better explainable, and that qualitative knowledge regarding voice production kinematics is obtained.

The limitations of our proposal and suggestions for future work include the following. First, all of the vocal fold vibrations patterns that we found in the voices which were labelled as vocal fry are auditorily perceived as pulsatile, i.e., individual glottal cycles appear to be audible due to temporal segregation. However, it is not yet clear whether listeners may reliably distinguish the subtypes that we found, or whether they may be

trained to do so. Second, although our model gives a classification with no true negatives or false negatives, it only distinguishes between vocal fry and euphonic GAWs. For future studies, we suggest the addition other types of dysphonic voices as negative data, and the combination of analyses with audio waveforms. Third, to further test and improve our waveform model, more positive natural data will be needed. With more natural data, the performance estimation of the classification could be made more confident. In particular, the lower limit of the 95% CI of classification accuracy for the natural data could be improved. Finally, instead of modelling the temporal transitions between the voice qualities, intervals of homogeneous voice qualities were preselected. Thus, models of the temporal transitions of the voice qualities related to vocal fry may be proposed in the future.

5. Conclusions

This paper investigated different types of amplitude modulated vocal fry GAWs. They were modelled using an analysis-by-synthesis approach and distinguished automatically from euphonic GAWs based on their modelling errors. Traditionally, vocal fry GAWs are characterized by a long closed phase. However, we also observed amplitude modulated vocal fry GAWs without a long closed phase, and analyzed them in a detailed way. These GAWs have been termed as AAMP, CAMP = 2 and CAMP = 4 based on their cyclicity and the number of pulses in a single modulator cycle. Modulated and unmodulated GAWs were modelled for vocal fry and euphonic GAWs. The rational for using a nonmodulating model is to obtain larger modelling errors for amplitude modulated GAWs than for nonmodulated GAWs. Modelling errors of the natural and synthetic vocal fry GAWs are observed to be well separated from the euphonic GAWs in the feature space. These modelling errors are used as predictors for classifying the vocal fry and euphonic GAWs. For the natural and synthetic GAWs, no false positives or false negatives were obtained for classification between vocal fry and euphonic GAWs.

Author Contributions: V.D.: conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, and writing—original draft preparation. P.A.: conceptualization, methodology, validation; formal analysis, investigation, resources, data curation, writing—review and editing, visualization, supervision, project administration, and funding acquisition. Both the authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Austrian Science Fund (FWF), grant number KLI722-B30.

Institutional Review Board Statement: Ethical review and approval were waived by ethics committee of Medical University of Vienna. The approval number is 1473/2016.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: P. Aichinger, I. Roesner, M. Leonhard, D. Denk-Linnert and B.Schneider-Stickler, "A database of laryngeal high-speed videos with simultaneous high-quality audio recordings of pathological and nonpathological voices", *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, vol.10, pp. 767–770, 2016.

Acknowledgments: This work was supported by the Austrian Science Fund (FWF): KLI 722-B30. The authors would like to thank the University Hospital Erlangen for providing the segmentation tool.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Whitehead, R.L.; Metz, D.E.; Whitehead, B.H. Vibratory patterns of the vocal folds during pulse register phonation. *J. Acoust. Soc. Am.* **1984**, *75*, 1293–1297. [CrossRef] [PubMed]
2. Blomgren, M.; Chen, Y.; Gilbert, H.R. Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. *J. Acoust. Soc. Am.* **1998**, *103*, 2649–2658. [CrossRef] [PubMed]
3. Ishi, C.T.; Ishiguro, H.; Hagita, N. Proposal of acoustic measures for automatic detection of vocal fry. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portuga, 4–8 September 2005.
4. Keating, P.A.; Garellek, M.; Kreiman, J. Acoustic properties of different kinds of creaky voice. In Proceedings of the International Congress of Phonetic Sciences, Glasgow, UK, 10–14 August 2015.
5. Degottex, G. Glottal Source and Vocal-Tract Separation. Ph.D. Thesis, Université Pierre et Marie Curie-Paris VI, Paris, France, 2010.
6. Laver, J. The phonetic description of voice quality. *Camb. Stud. Linguist. Lond.* **1980**, *31*, 1–186.
7. Aichinger, P.; Roesner, I.; Schoentgen, J. Auditory sensation of impulsivity and tonality in vocal fry. *J. Acoust. Soc. Am.* **2019**, *145*, 1909. [CrossRef]
8. Ylitalo, R.; Hammarberg, B. Voice characteristics, effects of voice therapy, and long-term follow-up of contact granuloma patients. *J. Voice* **2000**, *14*, 557–566. [CrossRef]
9. Hollien, H.; Moore, P.; Wendahl, R.W.; Michel, J.F. On the nature of vocal fry. *J. Speech Hear. Res.* **1966**, *9*, 245–247. [CrossRef] [PubMed]
10. Hollien, H.; Wendahl, R. Perceptual study of vocal fry. *J. Acoust. Soc. Am.* **1968**, *43*, 506–509. [CrossRef] [PubMed]
11. Gottliebson, R.O.; Lee, L.; Weinrich, B.; Sanders, J. Voice problems of future speech-language pathologists. *J. Voice* **2007**, *21*, 699–704. [CrossRef] [PubMed]
12. Drugman, T.; Kane, J.; Gobl, C. Data-driven detection and analysis of the patterns of creaky voice. *Comput. Speech Lang.* **2014**, *28*, 1233–1253. [CrossRef]
13. Vishnubhotla, S.; Espy-Wilson, C.Y. Automatic detection of irregular phonation in continuous speech. In Proceedings of the Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006.
14. Martin, P. Automatic detection of voice creak. In Proceedings of the Speech Prosody, Sixth International Conference, Shanghai, China, 22–25 May 2012.
15. Narendra, N.P.; Rao, K.S. Automatic detection of creaky voice using epoch parameters. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
16. Lauri, T.; Tanel, A.; Werner, S. Recognition of Creaky Voice from Emergency Calls. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 1990–1994.
17. Hollien, H.; Girard, G.T.; Coleman, R.F. Vocal fold vibratory patterns of pulse register phonation. *Folia Phoniatr. Logop.* **1977**, *29*, 200–205. [CrossRef] [PubMed]
18. Moore, P.; von Leden, H. Dynamic variations of the vibratory pattern in the normal larynx. *Folia Phoniatr. Logop.* **1958**, *10*, 205–238. [CrossRef] [PubMed]
19. Childers, D.G.; Lee, C.K. Vocal quality factors: Analysis, synthesis, and perception. *J. Acoust. Soc. Am.* **1991**, *90*, 2394–2410. [CrossRef] [PubMed]
20. Aichinger, P.; Pernkopf, F. Synthesis and Analysis-by-Synthesis of Modulated Diplophonic Glottal Area Waveforms. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 914–916. [CrossRef]
21. Aichinger, P.; Roesner, I.; Leonhard, M.; Denk-Linnert, D.; Schneider-Stickler, B. A database of laryngeal high-speed videos with simultaneous high-quality audio recordings of pathological and non-pathological voices. In Proceedings of the InternationalConference on Language Resources and Evaluation (LREC), Portorož, Slovenia, 23–28 May 2016; Volume 10, pp. 767–770.
22. *Glottis Analysis Tools (GAT-2018), Computer Program: Version 5*; Department of Phoniatrics and Pediatric Audiology, University Hospital Erlangen: Erlangen, Germany.
23. Voigt, D.; Döllinger, M.; Braunschweig, T.; Yang, A.; Eysholdt, U.; Lohscheller, J. Classification of functional voice disorders based on phonovibrograms. *Artif. Intell. Med.* **2010**, *49*, 51–59. [CrossRef] [PubMed]
24. Chen, G.; Shue, T.L.; Kreiman, J.; Alwan, A. Estimating the voice source noise. In Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association, Portland, OR, USA, 9–13 September 2012.
25. Bailly, L.; Bernardoni, N.H.; Müller, F.; Rohlfs, A.K.; Hess, M. Ventricular-fold dynamics in human phonation. *J. Speech Lang. Hear. Res.* **2014**, *57*, 1219–1242. [CrossRef] [PubMed]
26. Bailly, L.; Nathalie, H.; Xavier, P. Vocal fold and ventricular fold vibration in period-doubling phonation: Physiological description and aerodynamic modeling. *J. Acoust. Soc. Am.* **2010**, *127*, 3212–3222. [CrossRef] [PubMed]