Rocha *et al. EURASIP Journal on Advances in Signal Processing* (2021) 2021:78
https://doi.org/10.1186/s13634-021-00746-2

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Voice production model based on phonation biophysics

Raissa Bezerra Rocha[1], Wamberto José Lira de Queiroz[2] and Marcelo Sampaio de Alencar[3*]

*Correspondence:
malencar@iecom.org.br
[3]Institute for Advanced Studies in Communications, Campina Grande-PB, Brazil
Full list of author information is available at the end of the article

Abstract

This paper presents a proposal to a source-filter theory of voice production, more precisely related to voiced sounds. It is a proposal of a model to generate signal using linear and time-invariant systems and takes into account the phonation biophysics and the cyclostationary characteristics of the voice signal, related to the vibrational behavior of the vocal cords. The model suggests that the oscillation frequency of the vocal cords is a function of its mass and length, but controlled by the longitudinal tension applied to them. The mathematical description of the model of glottal excitation is presented, along with a mathematical closed expression for the power spectral density of the signal that excites the glottis. The voice signal, whose parameters can be adjusted for detection and classification of glottis pathologies, is also present. As a result, the output of each block diagram that represents the proposed model is analysed, including a power spectral density comparison between emulated voice, original voice, and classic source-filter model. The Log Spectral Distortion is computed, providing values below 1.40 dB, indicating an acceptable distortion for all cases.

Keywords: Voice production model, Transmission of cyclostationary information, Glottal pulse of Liljencrants-Fant, Power spectral density of voice signal

1 Introduction

The observation of the voice production mechanism begun in the eighteenth century, when it was stipulated that the vocal fold vibration was produced by air vibration.

In 1950, Husson proposed that the vocal fold vibration is a consequence of individual nervous impulses, generated at a rate given by fundamental frequency, sent to vocal muscles, resulting in an air force exhaled through the vocal cords.

Currently, the most accepted theory for the description of train of glottal pulses was proposed by Helmholtz and Muller, improved by van den Berg [1], in 1958, and Titze [2], in 1980, and is referred as Aerodynamic-Myoelastic Theory. According to this theory, the movement of opening and closing of the vocal folds is related to mechanical properties of the muscle that constitutes the vocal folds and to aerodynamic forces that are distributed along the larynx during the phonation.

During a discourse the rate of vocal fold vibration continuously changes due to the intonation of sentences. The question "Are you happy?" presents a growing intonation,

and the sentence "I am happy" has a decreasing intonation. The differences in intonation is justified by variation of oscillation of the vocal folds.

The pattern of vocal fold vibration is related to its length, mass, and tension, as presented in Formula 1. These parameters are associated to the sex and age of speaker. For men, for example, the vocal fold length is between 17 to 24 mm, while the feminine vocal fold is from 13 to 17 mm. For children, this length is smaller, from 6 to 8 mm. Usually, these length vary from 3 to 4 mm [3, 4].

$$F_o = \frac{1}{L_m}\sqrt{\frac{\sigma_c}{\rho}}, \tag{1}$$

in which L_m represents the length of membrane in vibration on the vocal folds, σ_c the longitudinal tension and ρ the volumetric mass of tissue.

A particular vocal cord, with a certain mass and length, has its vibration pattern increased by the elongation and tension of the cords, which reduces its mass and increases its elasticity. In this case, the mass and tension are more important than the length, in the determination of the vibration rate of the vocal cords, since the elongation reduces the mass and increases the tension, causing an increase in the oscillation frequency.

So in order to accomplish a more faithful acoustic analysis of the voice generation, this article has the objective of presenting a new mathematical model for the voice formation, which, differently from other works models in the literature, includes the cyclostationary behavior of the vocal cords.

The model is based on one of the most widespread theories in the voice production, the source-filter theory, proposed by Fant in 1970 [5]. The voice formation process is proposed based on a linear and time invariant system and considers that the cyclostationary movement of the vocal cords comes from changes in the vocal fold oscillation frequency, controlled by an electric signal measured at the vocal cords, referred as the control signal. In this case, the modelling is based on the fact that the oscillation frequency variation is proportional to the longitudinal tension applied to the vocal cords.

This model of voice production has the purpose of being utilized for the construction or improvement of systems which use speech processing, and mainly in the acoustic analysis for detection and classification of pathology in the glottis, by means of alterations in its vibratory pattern.

As the consonant phonemes are not generated by the vibration of the vocal cords, they are not suitable for detection and classification of cord pathologies. Thus, the proposed model is focused on the production of vowel phonemes.

The model is based on the assumption that the glottal flux results from an excitation produced by a cyclostationary impulse train generator. Since a linear process does not generate new frequencies and does not amplify the frequency range, it is assumed in model that variations in the fundamental frequency are present in the voice waveform and are represented by a signal which is obtained by means of crossing points the zero of the voice signal.

To achieve the proposed objective, the mathematical representation of a cyclostationary impulse generator is initially developed, which characterizes the variation of the fundamental frequency along a speech, governed by the control signal.

Then, the parameters of the probability distribution that models the control signal are estimated by means of a curve adjustment to the Unilateral Gamma and Rayleigh

probability density functions. A mathematical expression is proposed, to describe the power spectral density (PSD) of the considered glottal pulse, as well as expressions which characterizing the behavior of the voice signal in time and frequency domains.

Besides this introductory section, this article is organized in four more sections. Section 4 describes the new voice production model, with emphasis on the mathematical modeling of a cyclostationary impulse generator, and also presents a new mathematical expression for the analysis of the glottal pulse in the frequency domain and a probability density function estimate of the vibration frequency of the vocal cords. Section 5 presents results of the model's performance, with an analysis of the voice signal in the time and frequency domains, as well as a comparison of the classic source-filter voice generation models. Finally Section 6 presents the conclusions and future works suggestions.

2 Methods/experimental

The methods are:

1 The objective of this work is to present a new mathematical model that emulates the power spectral density (PSD) of the voice signal.
2 The model was tested with 6 human voices, that is, 3 male and 3 female.

3 Main contributions

The voice production models in the literature consider that a train of periodic impulses excites the glottis. Articles [6–8] show a continuous study to develop a voice production model. The authors are motivated to develop a new model, since the existing models in literature model the voice production from the separate analysis of the voice production steps, making it incomplete.

In [6], a model is presented, called probabilistic acoustic tube (PAT), in which the pitch, vocal tract, and energy are modeled together. In [7], an improvement of the model described in [6] is presented, in which the effect of breathing and glottal variation are incorporated. However, the models do not consider the cyclostationary vibration of the vocal cords.

In [8] an evolution of the PAT model is presented, including the AM/FM (amplitude modulation/frequency modulation) effect. The authors emphasize that the amplitude and frequency variations inside a voice frame are important and cannot be neglected, like what happens in many works that consider the voice perfectly stationary. The article proposes an adaptation of Bayesian Spectrum Estimation to rebuild the voice signal spectrum. However, it presents a high computational cost and does not present a closed expression for the PSD in function of the parameters obtained in the voice signal.

In this work, a new voice production model is presented, with the objective of providing a new expression for the voice signal PSD, to emulate the healthy voice signal PSD, such as the ones presented in the results, as well as voices with pathology. To this end, the cyclostationary movement of the vocal cords is considered, which in other works is neglected, and the voice PSD is proposed from probability distribution functions obtained in the voice waveform.

In comparison with the other works in the literature, this article presents the following contributions:

- It considers the cyclostationary movement of the vocal cords, making the mathematical modeling more faithful to the voice production process.

- It utilizes a cyclostationary sequence of impulses, with an average period given by the fundamental period, to excite the filter that models the vocal tract.
- It accomplishes the mathematical modelling of the vocal cords from the pulse position modulation (PPM), which variation of the impulse position is in function of and electric signal, $M(t)$, measured in the vocal cords.
- Considers a linear estimator for the PPM signal phase deviation, in a way that inside an appropriate band of modulating signal amplitude variation, the PPM signal is approximated by a wide-sense stationary processes (WSS).
- Proposes a connection between the movement of the vocal cords with the zero crossing points of the voice waveform.
- The mathematical model proposed is useful for sonorous and non-sonorous phonemes.
- It presents a closed expression for the voice signal PSD in function of the probability distribution of the modulating signal $M(t)$, which governs the vocal cord movement.
- A useful model to emulate the voice PSD to be used in applications which identify and classify the healthy and pathological voice, among other applications.

4 New model of voice production

Acoustic analysis is an area which attracts researchers in an increasingly manner, by representing an important tool for studying of applications in which the voice signal is present. Systems such as voice segmentation, voice coding, identification, and classification of pathologies and voice disturbances emulation, among others can be developed or improved based on acoustic analysis.

Particularly in the case of pathology emulation, research results found in the literature aim to obtain methods which can discriminate between the pathological voices and the healthy ones. In this case, the acoustic analysis can be combined with techniques that accomplish the direct observation of the vocal cords, with the objective of obtaining indicators that can identify disturbances in the voice.

Several pathologies of the voice can be detected by means of the observation of the vocal cords, which are one of the main tissues that involve the voice production. However, to make the identification of the vocal cord disturbances viable, the technique that makes it possible to analyze the acoustics must be the most faithful possible in its representation, during the phonation process.

In this context, the mathematical modelling that represents the behavior of the vocal cords during phonation, as described in this article, is a powerful acoustic analysis technique. From this method is possible to generate a voice signal in the time domain and estimate the voice power spectral density, by means of mathematical expressions which parameters can be adjusted to the emulation of the healthy and pathological voice.

In the voice production process, a sub-glottal pressure causes the separation of the vocal cords, and due to the Bernoulli effect, which explains a reduction in the supra-glottal pressure against the internal sides of each vocal fold, which comes together again and the air travels through the glottis at a higher speed. The opening and closing cycle of the glottis repeats, generating a train of pulses which feeds the vocal tract. This whole process is only possible because the vocal cords are elastic [4].

Particularly in the case of the vowel phonemes, this procedure causes the vibration of the vocal cords. On average, the vocal cords vibrate at each period $T_o = 1/F_o\ s$, or, in

other words, the vocal cords vibrate at a rate given by the fundamental frequency F_o [9].

However, the vibration frequency of the vocal cords is constantly changing while different patterns of intonation in the sentences are pronounced. Thus, a certain frequency F that is produced by a certain speaker has its value altered all the time during the speech. In this case, along the duration of a locution, the fundamental frequency is obtained for a brief moment, with frequencies larger and smaller than the average frequency being obtained.

The proposed voice production model is composed of five parts, as presented in Fig. 1: pulse generator, glottal pulse, gain, vocal tract, and radiation.

Differently from other works presented in the literature, the new voice formation model is based on the biophysics of phonation and has the characteristic of modeling the glottal flux taking into consideration the cyclostationary vibratory movement of the vocal cords.

In this case, the variation of the fundamental frequency as the speech occurs, as well as the gain parameter, related to the pressure of the air coming from the diaphragm, added to the glottal flux, are modeled with the purpose of obtaining a voice generation model which allows the relation of parameters with biomedical data and which has the possibility of adjusting parameters to emulate voice pathologies.

The way the voice is generated, the vibration frequency is determined by the elasticity, mass, and especially the longitudinal tension applied to the vocal cords. In a secondary manner, it is affected by the vertical tension obtained by the elevation lowering of the larynx, as well as the variation of the sub-glottal pressure.

In the proposed model, the source of excitation is controlled by an electric signal (control signal) $M(t)$ that can be measured at the vocal cords, whose amplitude controls the liberation of glottal pulses. This signal commands the vibration mechanism of the vocal cords, and mathematically, the relation between the variation of the vibration frequency and tension can be expressed by

$$\Delta\omega = \omega_o \alpha_1 T(t), \tag{2}$$

in which $\Delta\omega$ is the frequency variation of the impulses in the impulse sequence, $T(t)$ is the mechanical tension at the vocal cords, and α_1 is a sensibility constant of the process and $\omega_o = 2\pi F_o$.

The mechanical tension $T(t)$ can be written directly proportional to signal $M(t)$, so that

$$T(t) = \frac{1}{\alpha_2} M(t), \tag{3}$$

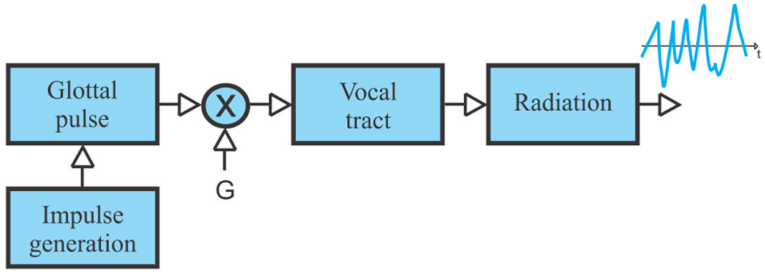


Fig. 1 Block diagram of the new voice generation model

in which $M(t)$ is an electric signal that can be measured at the vocal cords and α_2 is a sensibility constant of the process.

Thus

$$\Delta\omega = \omega_o \frac{\alpha_1}{\alpha_2} M(t). \tag{4}$$

Thus it is possible to define

$$\beta = \frac{\alpha_1}{\alpha_2}, \tag{5}$$

in which β represents the relation between the sensibility constants of the process.

The analysis of the accomplished production of the voice signals, using the source-filter model as a prototype, consists of defining, mainly, a mathematical model for the source excitation. Following, each step of the new model of voice production is described.

4.0.1 Cyclostationary impulse generation

In the process of producing human voiced sounds, a glottal pulse $E(t)$, originated in the lungs, is forced through the opening between the vocal cords, the glottis. In this process, when the vocal cords are under a larger tension, they vibrate more, contributing to generate the high-pitched speech sounds. When the cords are under a smaller tension, their vibration is smaller, contributing to the low pitched sounds.

This process of voice generation can be modeled as the passage of an impulse train $C(t)$ through a time invariant linear system with impulse response equal to the glottal pulse $E(t)$.

The impulse train $C(t)$ can be interpreted as the output of a cyclostationary impulse generator, since it consists of a sequence of impulses in time, whose position is controlled by a cyclostationary signal $M(t)$ which works as a modulating signal.

When the random signal $M(t)$ is not present, it is possible to consider that the vocal cords are under an average tension F_o, and thus, the signal $C(t)$ has equally spaced impulses for a duration T_o and can be written in terms of a trigonometric Fourier series,

$$C(t) = a_0 + \sum_{k=1}^{\infty} a_k \cos(k\omega_o t) + \sum_{l=1}^{\infty} b_l \ \sin(l\omega_o t), \tag{6}$$

in which

$$a_0 = \frac{1}{T_o} \int_{-\frac{T_o}{2}}^{\frac{T_o}{2}} \delta(t) dt = \frac{1}{T_o}. \tag{7}$$

$$a_k = \frac{2}{T_o} \int_{-\frac{T_o}{2}}^{\frac{T_o}{2}} \delta(t) \cos(k\omega_o t) dt = \frac{2}{T_o}. \tag{8}$$

And,

$$b_l = \frac{2}{T_o} \int_{-\frac{T_o}{2}}^{\frac{T_o}{2}} \delta(t) \ \sin(l\omega_o t)\, dt = 0, \tag{9}$$

so that

$$C(t) = \frac{1}{T_o} + \frac{2}{T_o} \sum_{k=1}^{\infty} \cos(k\omega_o t). \tag{10}$$

In the case in which the signal $M(t)$ is present, for the cyclostationary signal $C(t)$, the intervals between the occurrences of the impulses are controlled by the integral of $M(t)$, permitting $C(t)$ to be written as

$$C(t) = \frac{1}{T_o} + \frac{2}{T_o}\sum_{k=1}^{\infty}\cos\left(k\left(\omega_o t + \omega_o\beta\int_{-\infty}^{t} M(\tau)d\tau + \phi_o\right)\right), \tag{11}$$

in which

1 The average frequency ω_o corresponds to the period T_o of the non-modulated train of impulses.
2 The phase ϕ_o is directly proportional to the initial position δ_o of the impulses, $\phi_o = \omega_o\delta_o$, and is uniformly distributed in an interval of length 2π.
3 The signal $M(t)$ is an electric signal measured at the vocal cords and has a dimension of mV.
4 The parameter β can be seen as a sensibility constant for the modulation process of the glottis and has a dimension of V^{-1}.

For this model, the variation of the vibration frequency of the vocal cords is also directly proportional to the signal $M(t)$, that is

$$\Delta\omega = \omega_o\beta M(t). \tag{12}$$

and this variation occurs in the interval

$$0 \leq \Delta\omega \leq \omega_o\beta M_{max}, \tag{13}$$

in which M_{max} is the maximum amplitude of the random signal $M(t)$. Thus, the frequency deviation is such that

$$0 \leq \beta\omega_o M(t) \leq \omega_m, \tag{14}$$

in which ω_m is the maximum frequency of the vocal cord oscillation and the amplitude of $M(t)$ varies in the interval

$$0 \leq M(t) \leq \frac{\omega_m}{\beta\omega_o}. \tag{15}$$

4.0.2 Spectral analysis of the signal C(t)

Before initializing the spectral analysis of the random signal $C(t)$ it is appropriate to consider

$$\phi(t) = \omega_o\beta\int_{-\infty}^{t} M(\tau)d\tau. \tag{16}$$

and rewrite $C(t)$ as

$$C(t) = \frac{1}{T_o} + \frac{2}{T_o}\sum_{k=1}^{\infty}\cos\left(k\left(\omega_o t + \phi(t) + \phi_o\right)\right). \tag{17}$$

Since ϕ_o is a uniformly distributed random variable in an interval of length 2π, then the expected value of $C(t)$ is a constant $\frac{1}{T_o}$. Thus, if it is possible to write the autocorrelation of $C(t)$ in terms of only the difference between the instants of observation t and $t+\tau$ then it is possible to affirm that $C(t)$ is a wide-sense stationary process (WSS). The autocorrelation of $C(t)$ can be written as

$$\begin{aligned} R_C(\tau) = E\Bigg[&\left(\frac{1}{T_o} + \frac{2}{T_o}\sum_{k=1}^{\infty}\cos\left(k(\omega_o t + \phi(t) + \phi_o)\right)\right) \\ &\cdot\left(\frac{1}{T_o} + \frac{2}{T_o}\sum_{l=1}^{\infty}\cos\left(l\left(\omega_o(t+\tau) + \phi(t+\tau) + \phi_o\right)\right)\right)\Bigg]. \end{aligned} \tag{18}$$

After realizing the product of the terms, applying the expected value and using the fact that the expected value of the cosine of a random variable uniformly distributed in a interval of length 2π is null, one can write

$$R_C(\tau) = \frac{1}{T_o^2} + \frac{2}{T_o^2}\sum_{k=1}^{\infty} E\left[\cos\left(k\left(\omega_o\tau + (\phi(t+\tau) - \phi(t))\right)\right)\right]. \tag{19}$$

According to [10], the random process $\phi(t+\tau)$ can be approximated by a mean square error linear estimator, so

$$\phi(t+\tau) \approx \phi(t) + \tau\phi'(t). \tag{20}$$

Thus, $R_C(\tau)$ can be rewritten as

$$R_C(\tau) = \frac{1}{T_o^2} + \frac{2}{T_o^2}\sum_{k=1}^{\infty} E\left[\cos\left(k\left(\omega_o\tau + \tau\phi'(t)\right)\right)\right]. \tag{21}$$

Applying the Euler's formula for cosine function, $R_C(\tau)$ can be rewritten as

$$R_C(\tau) = \frac{1}{T_o^2} + \frac{1}{T_o^2}\sum_{k=1}^{\infty} E\left[e^{jk(\omega_o\tau+\tau\phi'(t))}\right] + \frac{1}{T_o^2}\sum_{k=1}^{\infty} E\left[e^{-jk(\omega_o\tau+\tau\phi'(t))}\right]. \tag{22}$$

At this point of the development, it is important to remember that $\phi(t)$ is an instantaneous phase defined by Formula (16), so its derivative is a random instant frequency, represented by $\omega(t)$. So $R_C(t)$ can be written as

$$R_C(\tau) = \frac{1}{T_o^2} + \frac{1}{T_o^2}\sum_{k=1}^{\infty} e^{jk\omega_o\tau} E\left[e^{jk\tau\omega(t)}\right] + \frac{1}{T_o^2}\sum_{k=1}^{\infty} e^{-jk\omega_o\tau} E\left[e^{-jk\tau\omega(t)}\right]. \tag{23}$$

Both of the expected values in this expression correspond, respectively, to the characteristic function of $\omega(t)$ and its complex conjugate sampled at $k\tau$. If this function is denoted $\varphi_{\omega(t)}(\nu)$, then $R_C(\tau)$ can be written as

$$R_C(\tau) = \frac{1}{T_o^2} + \frac{1}{T_o^2}\sum_{k=1}^{\infty} e^{jk\omega_o\tau}\varphi_{\omega(t)}(k\tau) + \frac{1}{T_o^2}\sum_{k=1}^{\infty} e^{-jk\omega_o\tau}\varphi^*_{\omega(t)}(k\tau). \tag{24}$$

From this expression, and the approximation $\phi(t+\tau) \approx \phi(t) + \tau\phi'(t)$, it is possible to write $R_C(\tau)$ as a function of τ and say that $C(t)$ is approximately wide-sense stationary.

According to 25, the power spectral density of $C(t)$ can be obtained calculating the Fourier Transform of $R_C(\tau)$.

$$S_C(\nu) = \int_{-\infty}^{\infty} R_C(\tau)\, e^{-j\omega\tau} d\tau \tag{25}$$

Considering the definitions of characteristic function and continuous time Fourier transform, the PSD of $C(t)$ can be written as

$$\begin{aligned} S_C(\nu) = {} & \frac{1}{T_o^2}\sum_{k=1}^{\infty}\int_{-\infty}^{\infty} f_\Omega(\omega)\int_{-\infty}^{\infty} e^{-j(\nu-k\omega_o-k\omega)\tau} d\tau d\omega \\ & + \frac{1}{T_o^2}\sum_{k=1}^{\infty}\int_{-\infty}^{\infty} f_\Omega(\omega)\int_{-\infty}^{\infty} e^{-j(\nu+k\omega_o+k\omega)\tau} d\tau d\omega \\ & + \frac{2\pi}{T_o^2}\delta(\nu). \end{aligned} \tag{26}$$

Considering the fact, from the Dirac delta function theory, which

$$\delta\left(t-t_o\right)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-j\omega(t-t_o)}d\omega, \tag{27}$$

the PSD $S_C(\nu)$ can be rewritten as

$$\begin{aligned}S_C(\nu)=&\frac{2\pi}{T_o^2}\sum_{k=1}^{\infty}\int_{-\infty}^{\infty}f_\Omega(\omega)\delta\left(\nu-k\omega_o-k\omega\right)d\omega\\&+\frac{2\pi}{T_o^2}\sum_{k=-\infty}^{-1}\int_{-\infty}^{\infty}f_\Omega(\omega)\delta\left(\nu-k\omega_o-k\omega\right)d\omega\\&+\frac{2\pi}{T_o^2}\delta(\nu).\end{aligned} \tag{28}$$

Applying the filtering and scaling in time properties of the Dirac delta function, the expression $S_C(\nu)$ can be rewritten as

$$S_C(\nu)=\frac{2\pi}{T_o^2}\sum_{k=1}^{\infty}\frac{1}{|k|}f_\Omega\left(\frac{\nu}{k}-\omega_o\right)+\frac{2\pi}{T_o^2}\sum_{k=-\infty}^{-1}\frac{1}{|k|}f_\Omega\left(\frac{\nu}{k}-\omega_o\right)+\frac{2\pi}{T_o^2}\delta(\nu), \tag{29}$$

which can be further rewritten as

$$S_C(\omega)=\frac{2\pi}{T_o^2}\delta(\omega)+\frac{2\pi}{T_o^2}\sum_{\substack{k=-\infty\\k\neq0}}^{\infty}\frac{1}{|k|}f_\Omega\left(\frac{\omega}{k}-\omega_o\right). \tag{30}$$

Using the fact that

$$\phi(t)=\omega_o\beta\int_{-\infty}^{t}M(\tau)d\tau \tag{31}$$

and that $\Omega(t)=\frac{d}{dt}\phi(t)$ is given by

$$\Omega(t)=\omega_o\beta M(t) \tag{32}$$

then

$$f_{\Omega(t)}(\omega)=\frac{1}{\omega_o\beta}f_{M(t)}\left(\frac{\omega}{\omega_o\beta}\right). \tag{33}$$

Substituting this result in (30), $S_C(\omega)$ can be rewritten as

$$S_C(\omega)=\frac{2\pi}{T_o^2}\delta(\omega)+\frac{2\pi}{\beta\omega_oT_o^2}\sum_{\substack{k=-\infty\\k\neq0}}^{\infty}\frac{1}{|k|}f_{M(t)}\left(\frac{\omega}{k\beta\omega_o}-\frac{1}{\beta}\right). \tag{34}$$

4.0.3 Probability distribution of M(t)

For $M(t)$, a continuous time WSS random process with probability distribution equal to distribution of zero crossings of the voice signal is proposed.

This proposal is based on the voice production model, described by a linear time-invariant filter, in which there is no generation of new frequencies when the glottal flux transverses the vocal tract. In this case, the variation of tension and the vocal cord oscillation frequency is directly related to the zero crossings obtained in the voice waveform.

The fundamental frequency consists of an average frequency reached by each orator. However, in a speech, the rate of variation of the vocal cords can be larger or smaller than the fundamental frequency. In light of this, two probability distributions are proposed to model the amplitude variation of the control signal $M(t)$: unilateral gamma and Rayleigh.

1 Unilateral gamma distribution
The unilateral gamma distribution can be characterized by the PDF (probability density function)

$$f_X(x) = \frac{1}{\Gamma(k_x)\theta^{k_x}} x^{k_x-1} e^{-\frac{x}{\theta}} u(x), \tag{35}$$

in which k_x and θ represent, respectively, the format and scale parameter, and $u(x)$ the unitary degree function.

2 Rayleigh Distribution
The Rayleigh distribution can be characterized by the PDF

$$f_X(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{(2\sigma^2)}} u(x) \text{ para } x \geq 0, \tag{36}$$

in which σ is the scale parameter for the Rayleigh PDF.

4.1 Glottal pulse model

In the literature it is possible to find some mathematical models to represent the glottal pulse analytically. Although there are many different quantities of parameters between the models, all of them have similarities in their characteristics, like representing the glottal pulse always positive or null, considering the glottal pulse almost periodic and as continuous function in time.

Besides that, the functions which represent the glottal flux are differentiable in time, except in the opening and closing instants of the glottis. Rosenberg's glottal model [11], Fant's [12], Liljencrants-Fant's (LF) [13] and Klatt's [14], are some of the glottal flux models found in the literature.

The LF model represents the derivative of the glottal flux and is divided in two segments. The first comprehends the opening process of the vocal cords. This segment initializes at the instant t_o, when the vocal cords are closed, until the instant t_e, when the glottis returns to its initial state, after opening, whose derivative assumes its maximum negative value, $-E_e$ [15]. Mathematically, this segment can be written a

$$E(t) = E_o e^{\alpha t} \sin(\omega_g t), t_o \leq t \leq t_e, \tag{37}$$

in which ω_g is the increase rate of the amplitude, determined by α, and E_o is a scale factor to reach an area.

The second segment of the glottal pulse consists of an exponential function which models the phase of return from the main excitation to the total closure phase [16]. This segment starts at the instant t_e and ends at the instant t_c, whose duration is also T_b. Mathematically, this segment can be described by

$$E(t) = \frac{-E_e}{\epsilon T_a} \left(e^{-\epsilon(t-t_e)} - e^{-\epsilon T_b} \right), t_e \leq t \leq t_c, \tag{38}$$

in which ϵ is a constant decaying to the phase recovery of the exponential.

The main parameter for the second segment is T_a, which represents the efficiency to return the phase. For the Liljencrants-Fant model, the glottal flux, $U(t)$, is given by

$$\begin{aligned} U(t) &= \frac{E_o e^{\alpha t}\sin\left(\omega_g t - \arctan\frac{\omega_g}{\alpha}\right)}{\sqrt{\alpha^2 + \omega_g}} + \frac{E_o\omega_g}{\alpha^2 + \omega_g}, \\ &\text{for } t_o \le t \le t_e. \\ &= \frac{E_e}{\epsilon^2 T_a}\left[e^{-\epsilon(t-t_e)} + \epsilon e^{-\epsilon T_b}\left(t - \left(t_c + \frac{1}{\epsilon}\right)\right)\right], \\ &\text{for } t_e \le t \le t_c. \end{aligned} \tag{39}$$

In this article, the frequency domain representation of the glottal pulse derivative is given by the Expression 40

$$\begin{aligned} E(\omega) &= \frac{\sqrt{\alpha_e^2 + \beta_e^2\omega^2}e^{-j\left(\omega t_e - \arctan\left(\frac{\beta_e}{\alpha_e}\omega\right)\right)}}{(\omega + j\alpha)^2 - \omega_g^2} + \frac{E_e}{\epsilon T_a}(t_c - t_e)e^{\frac{-j\omega}{2}(t_c+t_e)} \\ &\quad - \frac{\sqrt{\alpha_o^2 + \beta_o^2\omega^2}e^{-j\left(\omega t_o - \arctan\left(\frac{\beta_o}{\alpha_o}\omega\right)\right)}}{(\omega + j\alpha)^2 - \omega_g^2} \\ &\quad \cdot\left[e^{-\epsilon T_b}\mathrm{Sa}\left(\frac{(t_c - t_e)}{2}\omega\right) - \frac{1}{2}e^{-\frac{\epsilon}{2}(t_c-t_e)}\mathrm{Sa}\left(j\frac{(t_c - t_e)}{2}(\epsilon + j\omega)\right)\right], \end{aligned} \tag{40}$$

in which Sa(x) is the sample function, that is $\sin(x)/x$.

4.2 Vocal tract

Vocal tract is the space between the vocal folds and the lips. In the process of voice generation, the glottal flux is the entrance to the vocal tract, whose muscles cause the movement of the articulators, which, in turn, change the shape of the vocal tract, causing the production of different sounds.

Compared to the movement of the vocal folds, the vocal tract changes shape relatively slowly. The minimum time interval necessary for nerves and muscles to vary the articulations which participate in the speech formation, correspond to the duration of a phone. This duration is of the order of 50 ms, which represents the emission of 20 phones per second [17, 18].

During the voice production, the vocal tract is excited by a generator of pulses produced by the vocal folds for the formation of sonorous sounds, and, for the case of non-sonorous sounds, by turbulent air passing through the constrictions of the vocal tract. The vocal tract acts as a resonant filter, whose different configurations of articulators define the distinct formative frequencies, which have the objective of molding the frequency spectrum of the sound which propagates through its cavities. In general, for the generation of phones, three to five formatives are necessary.

In this article, the characteristic frequencies of the vocal tract are estimated by the linear predictive coding (LPC) method, which represents future samples by the linear combination of precedent samples, besides determining the fundamental frequency, spectrum, formatives, among other parameters [19–21].

In the LPC, the transference function of the vocal tract is given by

$$H(z) = \frac{1}{1 - \sum_{h=1}^{p} a_h z^{-h}}, \tag{41}$$

which consists of a all-pole filter, with all poles in a unitary radius circle, such as $z = e^{-j\omega}$, as

$$H(\omega) = \frac{1}{1 - \sum_{h=1}^{p} a_h e^{-jh\omega}}. \tag{42}$$

4.3 Power spectral density of voice signal

The prototype for the voice generation is based on the source-filter model, whose main difference is the modelling of the cyclostationary vibration of the vocal cords. The model presents the voice as a signal produced from a invariant-time linear system, in a interval in which the voice can be considered cyclostationary, typically from 16 to 32 ms, being possible to estimate the behavior of the voice signal in the frequency domain.

The proposed source-filter model considers the voice generation in a independent steps base, which are excitation model, vocal tract and radiation.

In the new voice production model, the source of excitation takes into consideration the cyclostationary movement of the vocal cords, based on its physical parameters, such as tension, mass and length. Since for a certain orator, the mass of the vocal cords is fixed and the length varies moderately, the vocal cords vibration is strongly related to the longitudinal tension applied to them.

In this context, the frequency of oscillation of the vocal cords is considered directly proportional to the tension to which they are submitted, controlled by a signal which characteristics are present in the voice signal waveform.

The control signal is given in the time domain and its period is inversely proportional to the longitudinal tension applied to the vocal cords. The new voice production model considers then that the voice signal is a result of the convolution between the signal resulting from the cyclostationary impulse generator controlled by the tension, glottal pulse, response to the impulse of the vocal tract and radiation at the lips, as illustrated in Fig. 1 and mathematically presented by

$$V(t) = GC(t) * E(t) * H(t) * L(t), \tag{43}$$

in which $V(t)$ represents the output signal, $C(t)$ the impulse train controlled by the tension signal, $E(t)$ the glottal pulse, $H(t)$ the impulse response of the vocal tract, G a positive gain related to the power of the air that comes from the diaphragm and $L(t)$ the effect of the radiation.

The effect of the radiation at the lips and nostrils is jointly represented by a high pass filter approximated by a first order derivative in the time domain, meaning that the derivative of the glottal flux is the excitation for the vocal tract. The radiation step amplifies the high frequencies with an average gain of 6 dB per octave and mathematically is given by

$$L(\omega) = 1 - \alpha e^{-j\omega}, \tag{44}$$

in which α is the lips/nostrils radiation coefficient which, usually, assumes values between 0.95 and 0.99 so that the zeros stay located inside the unitary circle in the z plain.

The proposed model assumes that each of the subsystems for voice generation is a invariant-time linear filter. In this scenario, at each step of the generation process, the resulting power spectral density is given by multiplying the input signal by the square module of the filter frequency response [22, 23].

This way, the power spectral density of the voice given by the new production model is given by Expression 45, in which $S_c(\omega)$ represents the PSD of the impulse train which excites the vocal folds, $E(\omega)$ is the frequency response of the glottal pulse model, G is the gain associated to the air power and $H(\omega)$ and (ω) are the selectivity in frequency of the vocal tract and effect of the radiation, respectively.

$$S_V(\omega) = G^2 S_c(\omega)|E(\omega)|^2|H(\omega)|^2|L(\omega)|^2. \tag{45}$$

The purpose of the developed voice production model is the possibility of, based on its mathematical expressions, accomplish adjustments of the parameters for detection and classification of vocal cord pathologies.

Since the vibratory behavior of the vocal cords is altered when facing pathologies, the parameters for the obtained expressions, such as probability density function, sensibility constant, and representation of the train of cyclostationary impulses in time, can be adjusted in order to adapt to a healthy voice, as well as to the characteristics of each pathology.

5 Results and discussion

In order to analyze the performance of the new voice production model, six locutions were randomly selected, from a male speaker and from a female speaker.

The locutions come from voice databases recorded by speakers from the interior of São Paulo state. The sentences were recorded at a rate of 22.05 ksamples/s and quantized with 16 bits per sample, for the male speaker, and 32 bits female speaker. The locutions are on average 3 s long and were recorded with the minimum amount of noise possible. All the processing is accomplished in the interval in which the voice signal is considered stationary, or in other words, at each 20 ms, with partitioning utilizing the Hamming window.

5.1 Cyclostationary control signal

The voice signal waveform is the result of the entire speech formation process. Basically, speech is formed from a cyclostationary excitation, for the sonorous signals, or a broad spectrum noise, similar to white noise, for the non-sonorous sounds.

The sonorous signals have in their waveform a cyclostationarity provided by the type of excitation in their formation process. Besides that, the waveform possesses short duration segments, delimited by zero crossings.

The new voice production model assumes that the vibration frequency of the vocal cords is directly proportional to the tension applied to them, stimulated by a control signal which represents the tension signal measured at the vocal cords.

For the model, the signal which controls the cyclostationary movement of the vocal cords is present in the voice signal waveform and is represented by the zero crossings. In this context, the signal which governs the opening and closing activity of the vocal cords is obtained by means of fragmentation of the voice waveform in each zero crossing, which is done by the passage of the signal through a two-level quantizer, according to Expression 46, in which each voice signal sample is associated to a specific level, depending on if it assumes a higher or lower value than the threshold.

$$\text{sgn}(s(n)) = \begin{cases} 1, & s(n) \geq 0, \\ -1, & s(n) < 0. \end{cases} \tag{46}$$

The process of quantization results in the representation of the voice signal by means of a matrix formed by regions constituted by sequences of 1s and −1s, whose transitions consist of the zero crossings of the voice signal.

The speech signal is represented by Expression 47, in which the matrix M_N expresses the quantity of samples contained in each short duration segment delimited by the instants of intersections with zero, estimated by amount of 1s or −1s, contained in each sequence.

$$M_N = [m_1, m_2, m_3, \ldots, m_i, \ldots, m_N], \text{ for } i = 1, \ldots, N, \tag{47}$$

in which m_i represents the duration parameter which describes the amount of samples in the i-th segment between zero crossings and N consists of the quantity of zero crossings.

From the matrix M_N it is possible to find the matrix T_N which represents the period or interval of time between each adjacent zero crossing. This way, the matrix T_N is obtained by the multiplication of the matrix M_N by the sampling period, T_s, and given by

$$T_N = [T_1, T_2, T_3, \ldots, T_i, \ldots, T_N], \text{ for } i = 1, \ldots, N, \tag{48}$$

in which T_i represents the period of session i.

5.2 Analysis of the probability distribution of the control signal

With the purpose of establishing the spectral representation of the voice signal, it is necessary to characterize the control signal based on the estimate of its probability density function, to describe the behavior in the frequency domain of the cyclostationary impulse generator using Expression 30.

During a locution, the vocal cords have a higher probability of oscillating at a rate given by the fundamental frequency, which consists of a specific frequency for each speaker, determined by the length, mass, and especially the tension applied to the vocal cords. However, during a speech, the vocal folds can hit a rate of vibration that is higher or lower than the rate established by the fundamental frequency, with a higher probability of values higher than it.

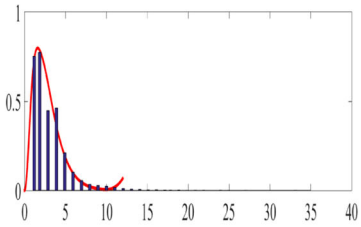
The control signal, which represents the oscillation movement of the vocal cords, is cyclostationary and has its probability distribution specified by a peak which represents the probability of variation in the fundamental frequency. Besides that, the distribution of the control signal presents higher probability values for higher fundamental frequency values.

In this case, the control signal possesses such a behavior that its probability distribution function can be adjusted to unilateral gamma and Rayleigh probability distributions, which were chosen for presenting similar behavior.

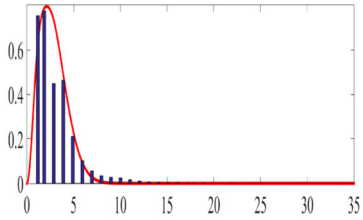
Figures 2, 3, and 4 present the histograms of the zero intersection point variables for the test locutions.

5.3 Cyclostationary impulse generation

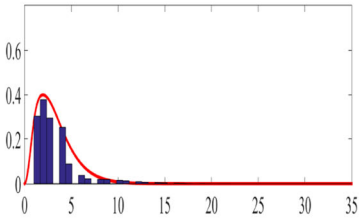
The voice generation model is based on the physics of phonation considering the cyclostationary of the voice signal, caused by the oscillation of the vocal cords.



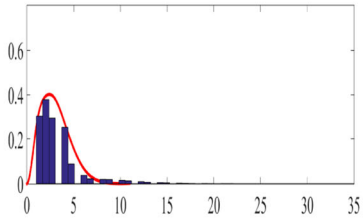
(a) Unilateral Gamma PDF, male speaker 1, voice sample for parameters $K_x = 2.7$ and $\theta = 1.0$.



(b) Rayleigh PDF, male speaker 1, voice sample for parameters $\sigma = 2.2$.

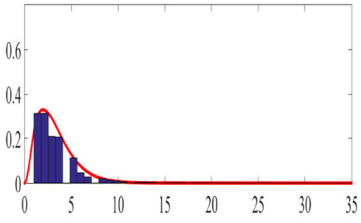


(c) Unilateral Gamma PDF, male speaker

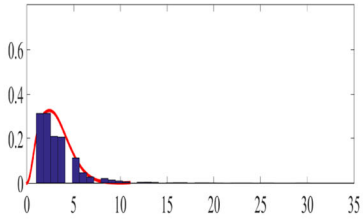


(d) Rayleigh PDF, male speaker 2, voice

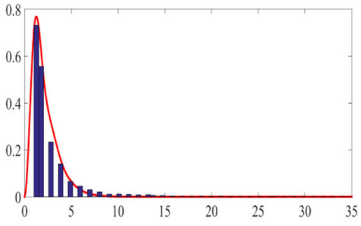
Fig. 2 PDF curves of Rayleigh and unilateral gamma distributions superimposed to histograms generated from voice samples of male speakers



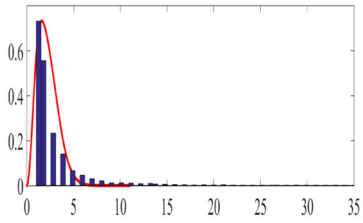
(a) Unilateral Gamma PDF, male speaker 3, voice sample for parameters $k_x = 2.0$ and $\theta = 1.0$.



(b) Rayleigh PDF, male speaker 3, voice sample for parameters $\sigma = 2.7$.

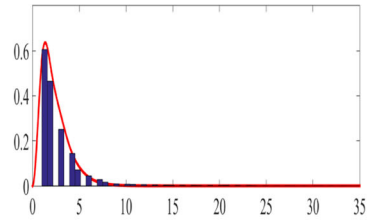


(c) Unilateral Gamma PDF, female speaker 1, voice sample for parameters $K_x = 2.0$ and $\theta = 1.0$.

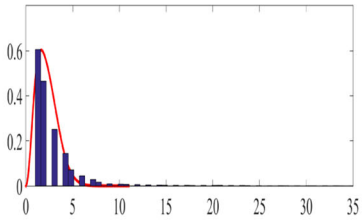


(d) Rayleigh PDF, female speaker 1, voice sample for parameters $\sigma = 2.2$.

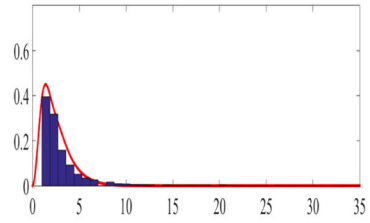
Fig. 3 PDF curves of Rayleigh and unilateral gamma distributions superimposed to histograms generated from voice samples of male and female speakers



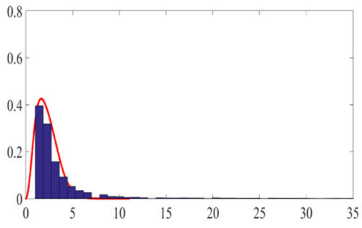
(a) Unilateral PDF, female speaker 2, voice sample for parameters $k_x = 2.3$ and $\theta = 1.0$.



(b) Rayleigh PDF, female speaker 2, voice sample for parameters $\sigma = 2.6$.



(c) Unilateral Gamma PDF, female speaker 3, voice sample for parameters $k_x = 2.0$ and $\theta = 1.0$.



(d) Rayleigh PDF, female speaker 3, voice sample for parameters $\sigma = 2.6$.

Fig. 4 PDF curves of Rayleigh and unilateral gamma distributions superimposed to histograms generated from voice samples of female speakers

In this scenario, the vocal cord prototype for voice generation is given by the generation of a train of cyclostationary impulses, which excite the vocal cords providing a cyclostationary glottal flux.

The impulse generator has as a result a sequence of impulses whose spacing is cyclostationary and established by the control signal which stimulates the tension at the vocal cords occasioning its oscillation. Figure 5 illustrates the output signal of the generator, $C(t)$, formed from each element of the matrix T_N, which represents a measure of spacing between the impulses.

The signal $C(t)$ can be seen as a PMM scheme used to transmit to the glottis the longitudinal tension information, in which the spacing or period between the impulses in time is inversely proportional to the tension, and consequently, the oscillation frequency of the vocal cords. Mathematically, this relation is given by Formula 12, in which β is a proportionality constant between the tension signal and the signal which represents the vibration

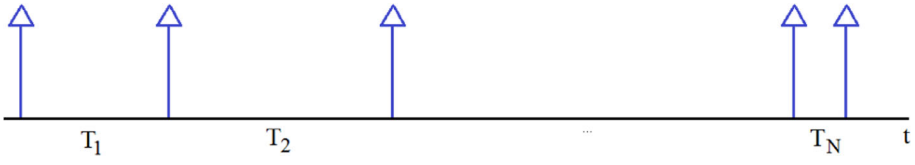


Fig. 5 Example of output from the impulse train generator, $C(t)$, with cyclostationary spacing obtained from the matrix T_N

frequency of the vocal cords. The sensibility constant utilized for better adjustment to the obtained results was $\beta = 0.1\ V^{-1}$.

Figures 6, 7, and 8 present the power spectral densities, $S_C(f)$, of the cyclostationary impulse train for each of the test locutions. The simulations were obtained with the gain values shown in Table 1.

5.4 Temporal and spectral analysis after glottis

For the new voice production model, the vocal cords are excited by a train of cyclostationary impulses, more adequately characterizing the generation of the voice signal. The vocal cords are modeled by means of the derivative of the glottal pulse of Liljencrants-Fant, whose impulse response given by Expressions 37 e 38.

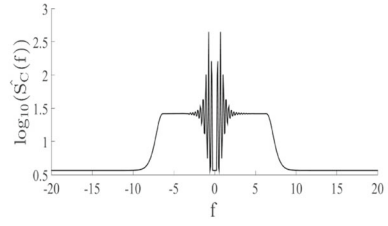
In the time domain, the glottal flux is represented by a sequence of glottal pulses with cyclostationary spacing between adjacent pulses, resulting from the convolution between the impulse response of the vocal cords and the cyclostationary impulse train. Mathematically, the glottal flux $Y(t)$, illustrated in Fig. 9, can be written as

$$Y(t) = C(t) * E(t). \tag{49}$$

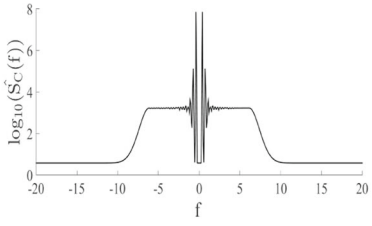
Figure 10 illustrates the comparison between the estimated Fourier transform of Liljencrants-Fant glottal pulse derivative and the Fourier transform obtained with Expression 40, proposed in this article, in which it is possible to observe the concordance with other works in literature [24–27].

Since the production model assumes that the voice is generated by means of a time invariant linear system, the PSD, after the vocal cords, can be written as

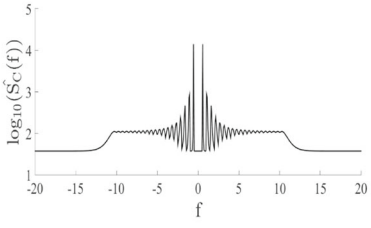
$$S_Y(\omega) = |E(\omega)|^2 S_C(\omega), \tag{50}$$



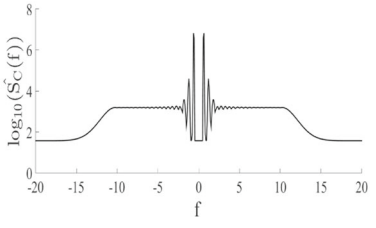
(a) $M(t)$ characterized by the unilateral Gamma PDF, for male speaker 1.



(b) $M(t)$ characterized by the Rayleigh PDF, for male speaker 1.

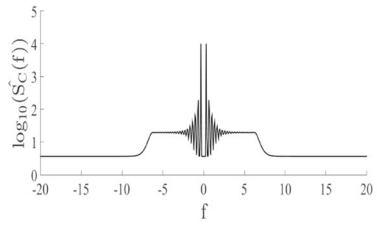


(c) $M(t)$ characterized by the unilateral Gamma PDF, for male speaker 2.

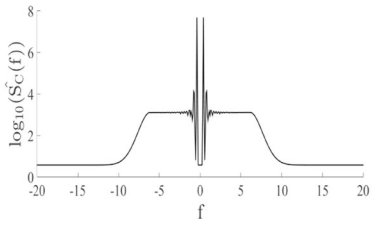


(d) $M(t)$ characterized by the Rayleigh PDF, for male speaker 2.

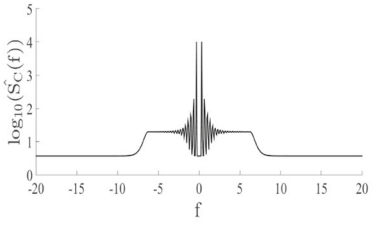
Fig. 6 PSD estimates, $\log_{10}(\hat{S}_C(f))$ (db), as a function of f ($\times\pi$ rad/s), of an impulse train modulated in position by a random signal $M(t)$ characterized by two PDFs for male speakers



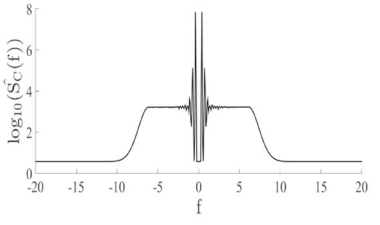
(a) $M(t)$ characterized by the unilateral Gamma PDF, for male speaker 3.



(b) $M(t)$ characterized by the Rayleigh PDF, for male speaker 3.

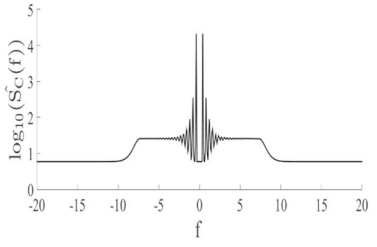


(c) $M(t)$ characterized by the unilateral Gamma PDF, for female speaker 1.

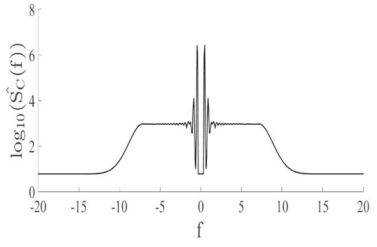


(d) $M(t)$ characterized by the Rayleigh PDF, for female speaker 1.

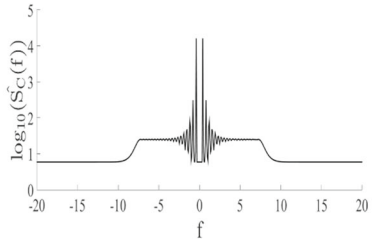
Fig. 7 PSD estimates, $\log_{10}(\hat{S}_C(f))$ (db), as a function of f ($\times\pi$ rad/s), of an impulse train modulated in position by a random signal $M(t)$ characterized by two PDFs for male and female speakers



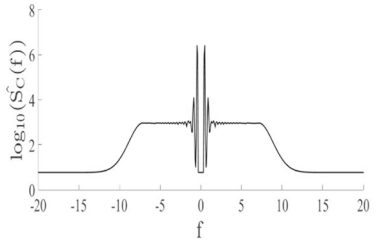
(a) $M(t)$ characterized by the unilateral Gamma PDF, for female speaker 2.



(b) $M(t)$ characterized by the Rayleigh PDF, for female speaker 2.



(c) $M(t)$ characterized by the unilateral Gamma PDF, for female speaker 3.



(d) $M(t)$ characterized by the Rayleigh PDF, for female speaker 3.

Fig. 8 PSD estimates, $\log_{10}(\hat{S}_C(f))$ (db), as a function of f ($\times\pi$ rad/s), an impulse train modulated in position by a random signal $M(t)$ characterized by two PDFs for female speakers

Table 1 Gain values $|G|^2$ for each of the test locutions

Locution	Gain $(\times 10^{-9})$
Locution (male voice)	3
Locution (female voice)	7

in which $E(\omega)$ represents the Fourier transform of the glottal pulse $E(t)$. Since $E(t)$ was considered a impulse response of a time invariant linear system, then $E(\omega)$ represents the response in frequency of this system. With this development it is possible to affirm that, at the glottis output, the spectrum of the observed signals in a time window that justifies the stationarity in a wide-sense can be adjusted to the spectrum $S_Y(\omega)$.

5.5 Spectral analysis of vocal tract

After the passage through the glottis, the glottal flux represents the input of the vocal tract, which has the function of filtering from a transference function determined by the position of the articulators in the moment of the phonation of each phoneme.

As mentioned, the estimation of the frequency selectivity magnitude spectrum of the vocal tract is obtained by means of an LPC analysis. Since the model treats the voice generation as a linear and time invariant model, the frequency response for the vocal tract is given by $|H(\omega)|^2$, in which $H(\omega)$ consists of the frequency response obtained by the LPC representation.

Figure 11 present the magnitude squared of the frequency response, $|H(f)|^2$, for the voice production model for each of the test locutions.

5.6 Final spectral and temporal analysis of voice signal

The temporal and spectral representation of the voice signal is a primordial acoustic analysis for applications which include voice signal processing. The model presented in this paper proposes that the voice is produced by a linear and time invariant system, for intervals in which it is considered stationary in the wide-sense.

Figure 12 illustrates an example of a segment of the voice signal in which it is possible to observe the glottal pulses with cyclostationary spacing, determined by the tension signal which controls the movement of the vocal cords, modified by the vocal tract and the radiation of the lips and nostrils.

The sequence of Figs. 13, 14, and 15 presents the comparison between the power spectral densities obtained by the simulation of locutions, by the new voice generation model and by the classic source-filter model.

From the observation of the figures, it is possible to notice that the PSD provided by the new voice generation model adjusts well to the frequency behavior of the test locutions.

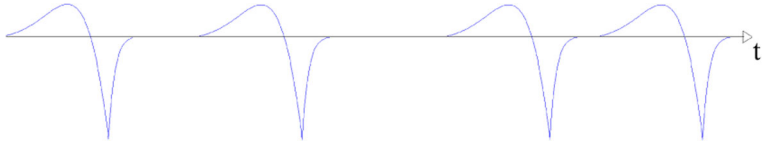


Fig. 9 Example of glottal flux obtained after the convolution of the cyclostationary impulse train with the derivative of the impulse response of the Liljencrants-Fant glottal model

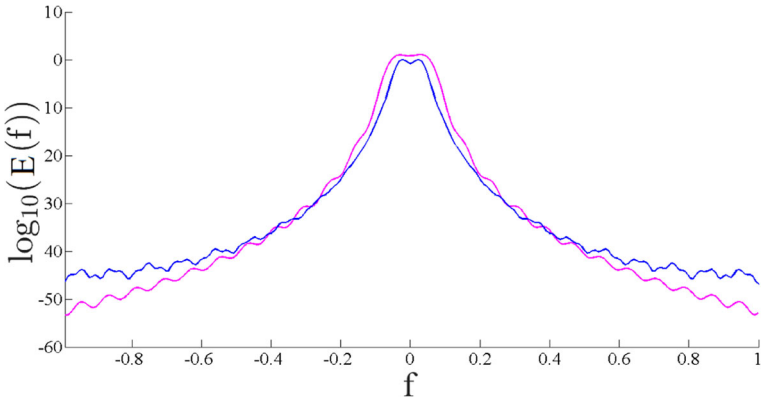
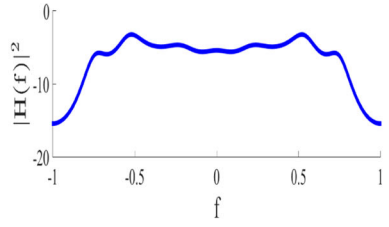
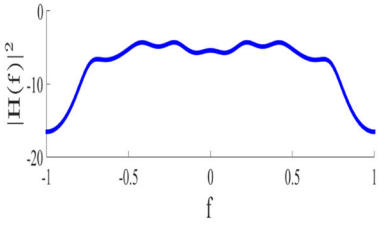


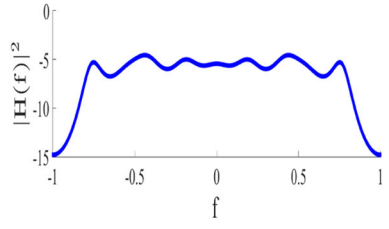
Fig. 10 Log of the estimated Fourier transform in dB, as a function of the frequency f ($\times\pi$ rad/s), obtained by the Welch method (blue) versus the one obtained with the proposed expression (lilac) for the derivative of the impulse response of the Liljencrants-Fant glottal model



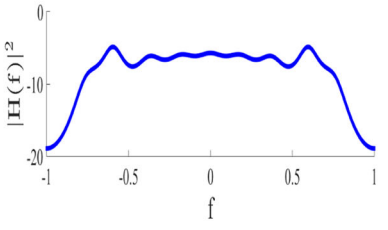
(a) Estimates of the vocal tract for male speaker 1.



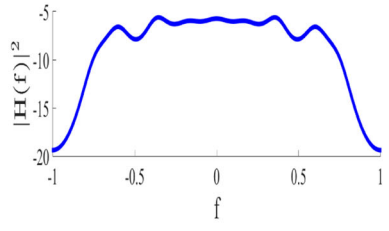
(b) Estimates of the vocal tract for male speaker 2.



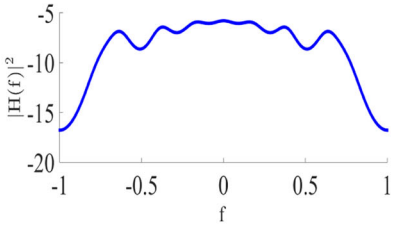
(c) Estimates of the vocal tract for male speaker 3.



(d) Estimates of the vocal tract for female speaker 1.



(e) Estimates of the vocal tract for female speaker 2.



(f) Estimates of the vocal tract for female speaker 3.

Fig. 11 Estimation of the vocal tract for male and female speakers

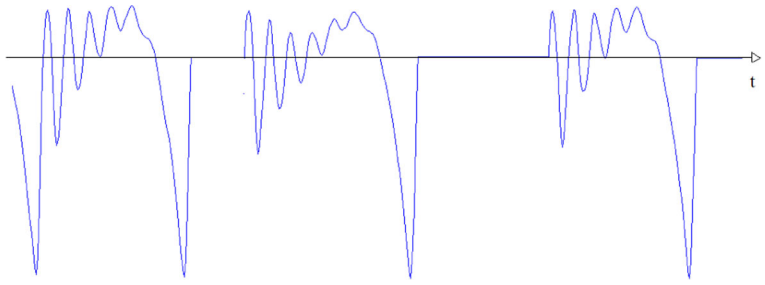


Fig. 12 Example of the signal in the time domain obtained by the proposed voice generation model

In the filtering process, the PSD of the cyclostationary impulse train is filtered and its bandwidth is equal to voices bandwidth. Its influence causes oscillations in glottal flux's PSD, or in other words, after the passage of the impulse train through the glottis, which do not exist in the classic source-filter model. The oscillations are derived from the cyclostationary movement of the vocal cords, providing a better adjustment when compared with voice signal PSD.

The classic voice production model, by the source-filter theory, on which many works in the literature are based on, was proposed by Fant in 1970 [5]. According to Fant, the representation of the voice formation is accomplished by means of the convolution between a glottal pulse model, the response to the impulse of the vocal tract and the effect of the radiation at the lips and nostrils.

In the time domain, Fant proposes the representation of the voice by

$$V_c(t) = E(t) * H(t) * L(t). \tag{51}$$

and in the frequency domain

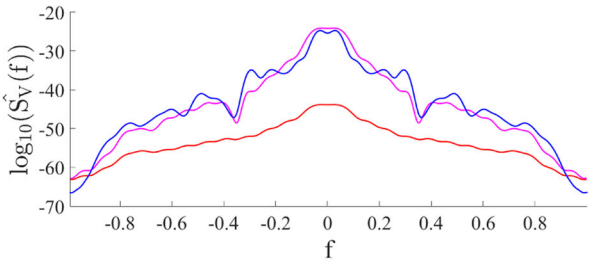
$$V_c(\omega) = E(\omega)H(\omega)L(\omega). \tag{52}$$

The results allow affirming that the spectrum of the tested signal in a stationary time window in the broad sense can be adjusted to the spectrum $S_V(\omega)$, with the use of two probability distribution functions. However, the unilateral gamma distribution, in general, has shown better adjustment to the PSD of the tested locutions.

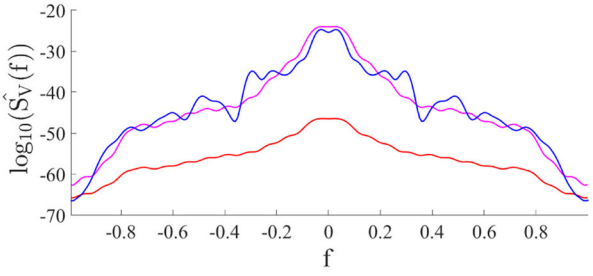
Besides that, the accomplished modeling for the voice formation based, on the physics of phonation, permits a larger applicability. One of them is that the estimation of the oscillation of the vocal folds, by means of a cyclostationary impulse generator, can also be used in the differentiation between sonorous and non-sonorous phonemes, since these two types of phonemes are distinguished by the zero intersection rate.

With the proposed model, it is also possible to accomplish a spectral estimation for pathological analyses. The vocal cords behave irregularly towards a pathology, and the present voice formation proposal is a robust method for the detection of pathologies, and a promising technique for the classification of pathologies, since it models the behavior of the vocal folds.

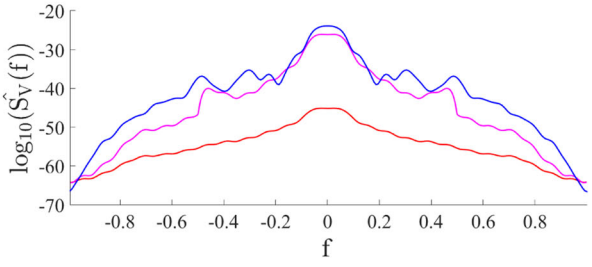
Observing the figures, it is possible to notice that the new voice production model provides a good estimate for the PSD of the voice in comparison to Fant's model. The voice signal is better modeled when it is considered to be generated from a linear and time



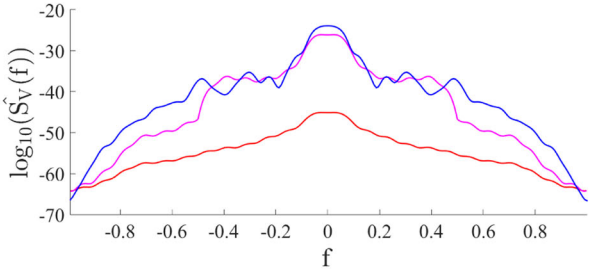
(a) PSDs comparison for male speaker 1 utilizing the unilateral Gamma probability distribution.



(b) PSDs comparison for male speaker 1 utilizing the Rayleigh probability distribution.

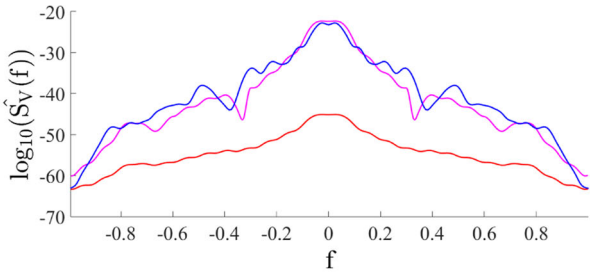


(c) PSDs comparison for male speaker 2 utilizing the unilateral Gamma probability distribution.

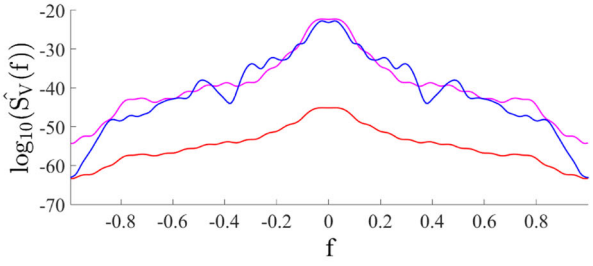


(d) PSDs comparison for male speaker 2 utilizing the Rayleigh probability distribution.

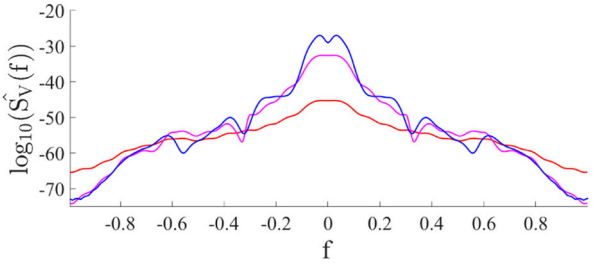
Fig. 13 Power spectral densities curves, as a function of the frequency $f(\times\pi\, rad/s)$, for male speakers: new voice generation model (lilac), classic source-filter (red) and estimated by voice locution (blue)



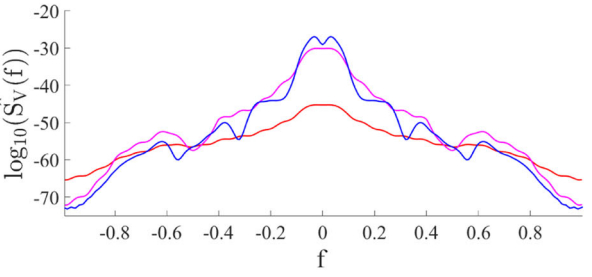
(a) PSDs comparison for male speaker 3 utilizing the unilateral Gamma probability distribution.



(b) PSDs comparison for male speaker 3 utilizing the Rayleigh probability distribution.

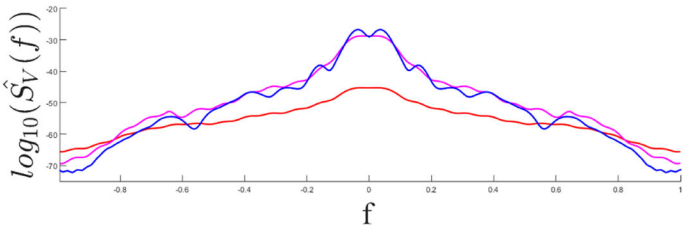


(c) PSDs comparison for female speaker 1 utilizing the unilateral Gamma probability distribution.

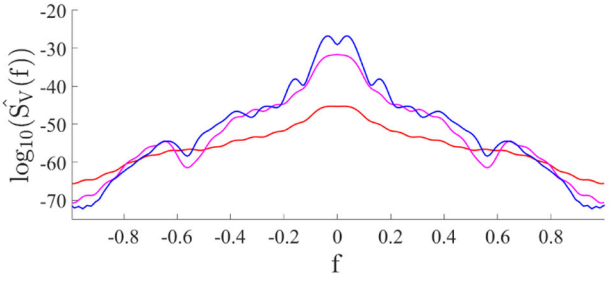


(d) PSDs comparison for female speaker 1 utilizing the Rayleigh probability distribution.

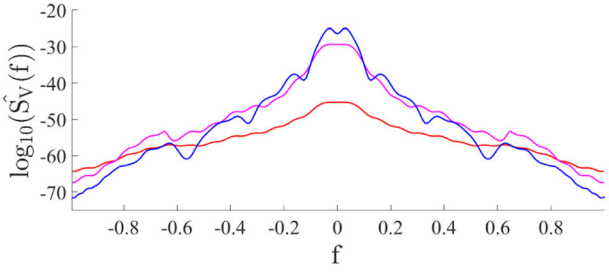
Fig. 14 Power spectral densities curves, as a function of the frequency $f(\times\pi\, rad/s)$, for male and female speakers: new voice generation model (lilac), classic source-filter (red) and estimated by voice locution (blue)



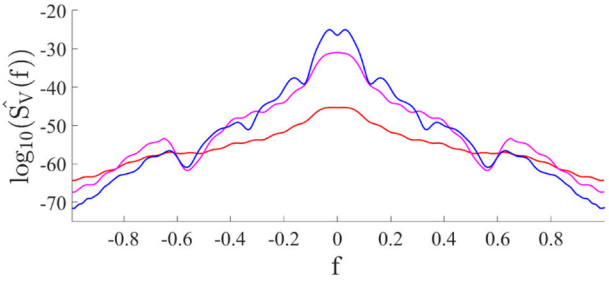
(a) comparison for female speaker 2 utilizing the unilateral Gamma probability distribution.



(b) PSDs comparison for female speaker 2 utilizing the Rayleigh probability distribution.



(c) PSDs comparison for female speaker 3 utilizing the unilateral Gamma probability distribution.



(d) PSDs comparison for female speaker 3 utilizing the Rayleigh probability distribution.

Fig. 15 Power spectral densities curves, as a function of the frequency $f(\times\pi rad/s)$, for female speakers: new voice generation model (lilac), classic source-filter (red) and estimated by voice locution (blue)

invariant system. Besides that, the proposed model considers the cyclostationary movement of the vocal cords in the phonation of sonorous sounds, providing a more faithful representation of the oscillations of the voice spectrum.

5.7 Model evaluation using distortion measures

In order to evaluate the proposed model of voice production, it is considered in this article the spectral distortion metric log spectral distortion (LSD), defined as [28]

$$LSD = \sqrt{\frac{1}{B}\int_0^B \left[10\log S(\omega) - 10\log\hat{S}(\omega)\right]^2 d\omega}, \tag{53}$$

in which B is the voice signal bandwidth, $S(\omega)$ is the original voice signal PSD and $\hat{S}(\omega)$ is the emulated voice signal PSD (represented in this article by $S_V(\omega)$).

In this evaluation, 76 locutions were considered, 32 from female speakers and 44 from male speakers. The speakers are of different ages and locutions are 3 kHz bandwidth.

Figure 16 presents LSD metric values to female and malle speakers, for unilateral Gamma and Rayleigh probability distributions. It is possible to observe that the LSD values are below 2 dB, limit at which signals have acceptable distortion.

On average, the female speakers provide 1.30 and 1.33 with standard deviation 0.22 and 0.25, for unilateral Gamma and Rayleigh probability distributions, respectively. On the other hand, on average, the male speakers provide 1.39 and 1.34 with standard deviation 0.21 and 0.19, for unilateral gamma and Rayleigh probability distributions, respectively.

6 Conclusions and future work

This article presents a new voice synthesis method based on the source-filter theory. The objective is the development of a theory that is more faithful to the biophysics of phonation, with the intent of enabling the detection and classification of pathologies in the vocal cords from its specific characteristics.

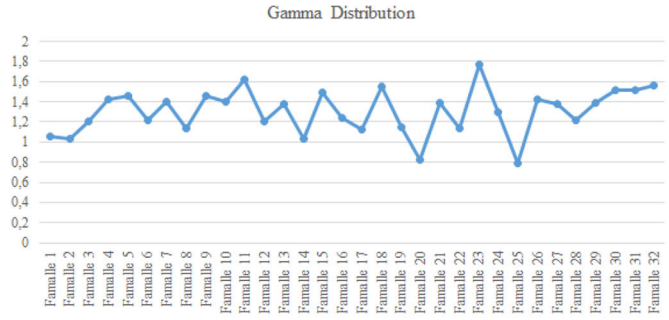
Due to its appropriateness to identification of pathologies, the model is directed to the generation of sonorous sounds, since in its process of synthesis is included the behavior of the vocal cords.

The new generation model proposes that the vocalic sounds are formed from a cyclostationary oscillation of the vocal cords, represented by an average oscillation frequency, which is the fundamental frequency, besides frequencies above and under it. This movement is proportional to the mass, length of vocal folds, and especially, a longitudinal tension signal.

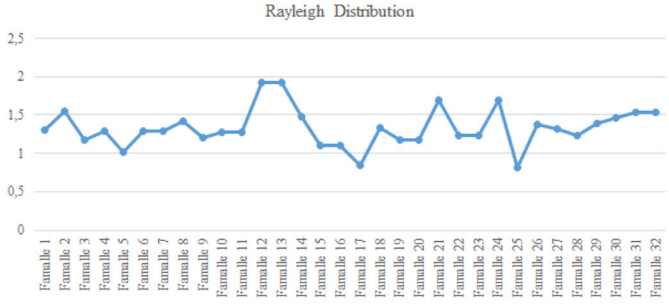
For a certain speaker, the new model considers a determined mass and vocal cord length and proposes that the tension signal that rules the cyclostationary movement of the vocal cords is directly proportional to the oscillation frequency and is present in the voice signal waveform, at each zero intersection point.

In this scenario, the mathematical analysis of the vocal cord excitation process is accomplished from its representation by a train of cyclostationary impulses resulting from a generator.

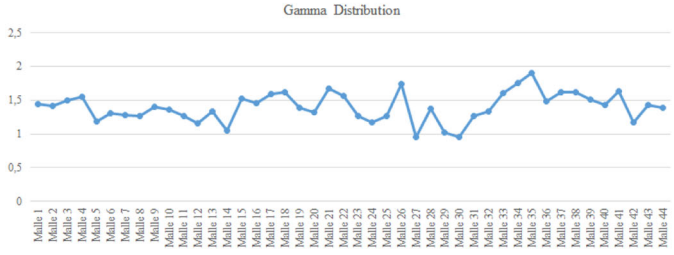
The model accomplishes the synthesis considering that the source-filter system is constituted from linear and time invariant subsystems, for time intervals in which the voice signal is considered stationary. In this case, in time domain, the flux after the glottis is



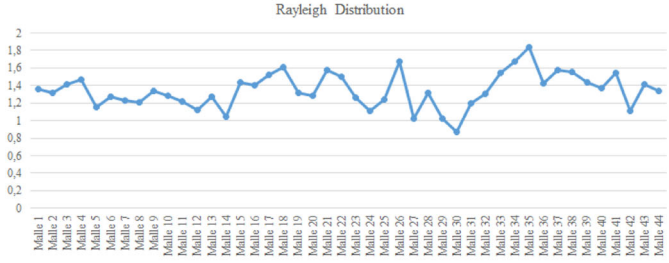
(a) LSD results for 32 female speakers utilizing the unilateral Gamma probability distribution.



(b) LSD results for 32 female speakers utilizing the Rayleigh probability distribution.



(c) LSD results for 44 male speakers utilizing the unilateral Gamma probability distribution.



(d) LSD results for 44 male speakers utilizing the Rayleigh probability distribution.

40

Fig. 16 LSD results for 76 speakers

described by the convolution between the glottal pulse model and the cyclostationary impulse train.

In this article, the response of the glottis was modeled by the derivative of the Liljencrants-Fant glottal pulse model. For the formation of the voice signal, the glottal flux is convoluted with the response of the impulse of the vocal tract, and the radiation of the lips and nostrils. Besides that, a gain is considered for the representation of the air power that comes from the diaphragm.

A new mathematical formulation is presented for the voice power spectral density, in function of the PSD of the cyclostationary impulse generator, which is given in function of the tension signal probability distribution. A representation of the tension signal's probability distribution is proposed, and consequently, a representation of the vocal cord oscillation frequency, by means of unilateral Gamma and Rayleigh distributions, is also proposed

To evaluate the performance of the proposed voice synthesis model, six locutions were selected at random, one with a female voice and other with a male voice. The output results for each voice generation subsystem are presented.

Observing the results, it is noticeable that the voice generation is better modeled by a linear time invariant synthesis, with the inclusion of the cyclostationary movement of the vocal cords, which describes more faithfully the oscillations in the power spectrum.

Besides that, spectral distortion evaluations with 76 speakers were performed and the results indicate that the distortion obtained is acceptable for all, that is, below 1.40 dB.

The developed voice generation model is promising for the characterization of pathological voices, because it models the voice by means of parameters related to the physics of phonation.

The detection and classification of larynx disturbances is intended, based on the estimation of parameters like the probability distribution of the signal which controls the movement of the vocal cords, the mass, length, and average oscillation frequency, since it is known that the presence of pathologies, like edemas, nodules, polyps, and cysts, causes modifications in the vocal cords, like the increase of mass, causing irregular vibration. Besides that, it can lead to incorrect closing of the glottis, causing noisy components, and significant modifications in the sonorous sounds.

Abbreviations

PSD: Power spectral density; PAT: Probabilistic acoustic tube; AM: Amplitude modulation; FM: Frequency modulation; PPM: Pulse position modulation; WSS: Wide-sense stationary processes; PDF: Probability density function; LF: Liljencrants-Fant's; LPC: Linear predictive coding; LSD: Log spectral distortion

Acknowledgements

The authors would like to express their thanks to the Federal University of Campina Grande, the Federal University of Sergipe, and the Institute for Advanced Studies in Communications.

Authors' contributions

Raissa Bezerra Rocha participated in the design of the study, performs literature review, and implemented the simulations and tests, besides writing the manuscript.
Wamberto José Lira de Queiroz participated in the design of the study, performs literature review, and helped to draft and review the manuscript.
Marcelo Sampaio de Alencar conceived of the study and participated in its design and coordination and helped to draft and review the manuscript.
All authors read and approved the final manuscript.

Funding

Not applicable

Availability of data and materials

Not applicable

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Federal University of Sergipe, São Cristóvão-SE, Brazil. [2]Federal University of Campina Grande, Campina Grande-PB, Brazil. [3]Institute for Advanced Studies in Communications, Campina Grande-PB, Brazil.

Received: 3 February 2021 Accepted: 15 June 2021
Published online: 08 September 2021

References

1. J. Van den Berg, Myoelastic-aerodynamic theory of voice production. J. Speech Hear. Res. **1**, 227–244 (1958)
2. I. R. Titze, Comments on the myoelastic-aerodynamic theory of phonation. J. Acoust. Soc. Am. **23**, 495–510 (1980)
3. T. B. Patel, H. A. Patil, in *The 9th International Symposium on Chinese Spoken Language Processing*, Novel Approach for Estimating Length of the Vocal Folds using Fujisaki Model (IEEE, Singapore, 2014), pp. 308–312. https://doi.org/10.1109/ISCSLP.2014.6936673
4. L. J. Raphael, G. J. Borden, K. S. Harris, *Speech Science Primer, Sixth edition*. (LWW, 2011)
5. G. Fant, *Acoustic Theory of Speech Production*. (The Hague, Paris, 1970)
6. Z. Ou, Y. Zhang, in *International Conference on Artificial Intelligence and Statistics*, Probabilistic Acoustic Tube: a probabilistic generative model of speech for speech analysis/synthesis, (2012)
7. Y. Zhang, Z. Ou, M. Hasegawa-Johnson, in *2014 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2014*, Improvement of Probabilistic Acoustic Tube Model for Speech Decomposition (Institute of Electrical and Electronics Engineers Inc., Florence, 2014), pp. 7929–7933
8. Y. Zhang, Z. Ou, M. Hasegawa-Johnson, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Incorporating AM-FM Effect in Voiced Speech for Probabilistic Acoustic Tube Model, (2015)
9. R. B. Rocha, V. V. Freire, F. Madeiro, M. S. Alencar, Sistema de Segmentação de Fala Baseado na Observação do Pitch. Revista de Tecnologia da Informação e Comunicação. **4**(1), 6 (2014)
10. M. S. de Alencar, *Communications Systems*. (EUA, New York, 2005)
11. G. Degottex, Glottal Source and Vocal-Tract Separation. Estimation of Glottal Parameters, Voice Transformation and Synthesis using a Glottal Model. Tese de doutorado, Université Paris (2010)
12. G. Fant, Vocal-Source Analysis – A Progress Report. TL-QPSR. **20**, 31–53 (1979)
13. G. Fant, J. Liljencrants, Q. Lin, A Four Parameter Model of Glottal Flow. TL-QPSR. **26**, 1–13 (1985)
14. D. Klatt, L. Klatt, Analysis, synthesis, and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Am. **87**(2), 820–857 (1990). https://doi.org/10.1121/1.398894
15. S. de O. Dias, Estimation of the Glottal Pulse from Speech or Singing Voice, Master's thesis, School of Engineering of the University of Porto (2012)
16. C. Gobl, The Voice Source in Speech Communications. Tese de doutorado, Vetenskap Och Konst (2003)
17. E. D. S. Paranaguá, Segmentação Automática do Sinal de Voz Para Sistemas de Conversão Texto-Fala. Tese de Doutorado. Universidade Federal do Rio de Janeiro (Março 2012)
18. D. O'Shauqhnessy, Modern Methods of Speech Synthesis. IEEE Circ. Syst. Mag. **7**(3), 6–23 (2007)
19. R. F. B. Sotero, Novas Abordagens para Codificação de Voz e Reconhecimento Automático de Locutor Projetadas Via Mascaramento Pleno em Frequência por Oitava, Dissertação de mestrado, Universidade Federal de Pernambuco (2009)
20. E. L. F. da Silva, Estimativas de Comportamento Vocálico de Locutores e um Novo Sistema de Separação Silábica. Dissertação de mestrado, Universidade Federal de Pernambuco (2012)
21. L. R. Rabiner, B. Juang, *Fundamentals on Speech Recognition,*. (Prentice Hall, Engliewood Cliffs, 1996)
22. M. S. de Alencar, *Probabilidade e processos estocásticos*. (Érica, Português, 2009), p. 288
23. B. P. Lathi, *Modern Digital and Analog Communication Systems,*. (Oxford University Press, New York, 2009)
24. G. Fant, The LF-model Revisited. Transformations and Frequency Domain Analysis. Q. Prog. Status Rep. (STL-QPSR). **36**(2-3), 119–156 (1995)
25. G. Fant, K. Gustafson, LF-Frequency Domain Analysis. Q. Prog. Status Rep. (STL-QPSR). **37**(2), 135–138 (1996)
26. B. Doval, C. d'Alessandro, N. Henrich, The Spectrum of Glottal Flow Models. Acta Acustica U. Acustica. **92**(1), 1–21 (2006)
27. J. Kane, M. Kane, C. Gobl, in *INTERSPEECH*, A Spectral LF Model Based Approach to Voice Source Parameterisation, (2010)
28. R. P. Ramachandran, R. Mammone, Modern Methods of Speech Processing. Springer Science + Business Media, LLC (1995)

Publisher's Note