State-Space Approach to Structural Representation of Perturbed Pitch Period Sequences in Voice Signals

Gabriel A. Alzamendi, Gastón Schlotthauer, and María E. Torres, *Entre Ríos, Argentina*

Summary: Objectives. The aim of this study was to propose a state space-based approach to model perturbed pitch period sequences (PPSs), extracted from real sustained vowels, combining the principal features of disturbed real PPSs with structural analysis of stochastic time series and state space methods.

Methods. The PPSs were obtained from a database composed of 53 healthy subjects. State space models were developed taking into account different structures and complexity levels. PPS features such as trend, cycle, and irregular structures were considered. Model parameters were calculated using optimization procedures. For each PPS, state estimates were obtained combining the developed models and diffuse initialization with filtering and smoothing methods. Statistical tests were applied to objectively evaluate the performance of this method.

Results. Statistical tests demonstrated that the proposed approach correctly represented more than the 75% of the database with a significance value of 0.05. In the analysis, structural estimates suitably characterized the dynamics of the PPSs. Trend estimates proved to properly represent slow long-term dynamics, whereas cycle estimates captured short-term autoregressive dependencies.

Conclusions. The present study demonstrated that the proposed approach is suitable for representing and analyzing real perturbed PPSs, also allowing to extract further information related to the phonation process.

Key Words: Perturbed pitch periods–Stochastic pitch model–Jitter–Structural time-series analysis–State-space models.

INTRODUCTION

Precise period perturbation assessment is one of the most difficult tasks in speech pathology and voice therapy.[1] This is because the perturbations arise in an unpredictable fashion and are usually concealed in the speech records. Specialists have not yet reached a full agreement on the nature and origin of these irregularities.[2] Indeed, it can be shown that irregularities arise even in a nonpathologic stable voice.[3,4] In the last few decades, this situation has drawn great attention from researchers and clinicians in the speech community. They have found that perturbations arise as a result of the combination of neurologic, biomechanical, aerodynamic, and acoustic sources throughout the speech production system.[5] Additionally, it has been argued that perturbations behave noticeably different in pathologic and nonpathologic voices.[6,7] In this work, we propose a method for analyzing and modeling a pitch period sequence (PPS), consisting of successive pitch periods extracted from a voice signal, which explicitly considers fluctuations and instantaneous perturbations.

Real PPSs are generally composed by identifiable structures presenting short- and long-term behavior.[8] Short-term structures carry information related to period perturbations, consisting of random cycle-to-cycle variations denominated jitter.[3,9] In objective voice analysis, there are diverse acoustical parameters conceived to quantify jitter, which can be classified as absolute measures (eg, perturbation factor or directional perturbation factor) or fundamental frequency-related measures (eg, jitter factor, jitter ratio, or coefficient of variation).[6] However, these objective features are highly sensitive to slow long-term components, normally associated with prosody information or intonation. Consequently, relative average perturbation features have been defined (eg, average absolute perturbation or period perturbation quotient).[6] Nevertheless, these parameters are inadequate where long-term components are strong. Therefore alternative methods are required for robust perturbation assessment. To solve this problem, we propose a state-space approach to PPS structural analysis, and we show that it allows to optimally separate jitter and long-term components.

Recent advances in PPS characterization have found great applicability in modern technology. It has been demonstrated that PPSs carry information belonging to the speaker itself (eg, identity, gender, or mood)[10–12] and related to the physiological condition of the speech production system (eg, vocal folds dynamic).[13] Therefore, theoretical models, considering jitter and long-term fluctuations, were developed and successfully used in several applications, eg, to enhance the naturalness of artificial voices in synthesis methods,[14] to synthesize expressive voices in human-computer interfaces,[11] to verify speakers in security systems,[10] and to simulate pathologic voices under controlled conditions.[15] Moreover, these models provide a theoretical framework to understand period perturbations.[5]

Nowadays, there is a great number of PPS models available in speech literature. The simplest one consists of a sequence of constant fundamental periods, not allowing the representation of aperiodic signals. On the other hand, versatile perturbed PPS models have been developed using simple stochastic laws. A straightforward strategy for jitter generation involves

Accepted for publication November 20, 2014.

From the Laboratorio de Señales y Dinámicas no Lineales, Facultad de Ingeniería de la Universidad Nacional de Entre Ríos and Consejo Nacional de Investigaciones Científicas y Técnicas, Entre Ríos, Argentina.

Address correspondence and reprint requests to Gastón Schlotthauer, Laboratorio de Señales y Dinámicas no Lineales, Facultad de Ingeniería, Universidad Nacional de Entre Ríos, cc 47, Suc 3 (3100) Paraná, Entre Ríos, Argentina. E-mail: gschlotthauer@conicet.gov.ar

Journal of Voice, Vol. 29, No. 6, pp. 682-692
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2014.11.007

a constant fundamental period perturbed by random noise. This approach has been applied in expressive speech synthesis for neutral voice transformation, where fundamental period and random noise depend on different emotions.[16] Moreover, Gaussian mixture models were applied to emotional speech synthesis, where different emotions were characterized considering long-term components as multimodal processes.[17] Recently, we developed a strategy to synthesize both normal and pathologic perturbed sustained vowels, where PPSs were obtained from a stochastic model based on jitter factor. This method proved to be useful for testing algorithms for fundamental frequency estimation[15] and for voice synthesis with high perceptual quality.[18]

All the methods mentioned previously assume that PPSs are independent and identically distributed stationary stochastic processes. Nevertheless, examination of real sequences demonstrates that these hypotheses are unrealistic and, as a consequence, previous methods are not able to suitable represent a real PPS. Schoentgen[19] has summarized the principal features of PPSs extracted from real normal voices. For the present work, some of those were considered:

- PPS presents Gaussian probability distribution;
- adjacent periods in a PPS are correlated where correlation degree varies with voice signals;
- there are structures that reinforce period correlation (microtremors);
- jitter size is small (0.1–1% relative to fundamental period);
- jitter appears to be a genuine stochastic phenomenon;
- meaningful statistics of jitter can be obtained from sustained vowel waveforms.

Considering the previously listed features, it is clear that more versatile models, able to represent complex structures, are required. The first attempt to understand period correlation made use of time series analysis methods based on autoregressive (AR) or AR moving average models.[20] These methods allowed to represent the existing strong correlation in both normal and pathologic voices, where model order depended on the analyzed voices.[12,21] Later, Ruinskiy and Lavner[14] proposed a jitter bank-based approach to characterize the relative amplitude and correlation in the PPSs, suitable for naturally hoarse voice synthesis. Despite the ability to represent correlation information, the previously mentioned methods assume that PPSs are stationary signals. Although it has been shown that short-term jitter is indeed a stationary process, PPSs are not necessarily stationary signals.[5]

Stochastic difference equations have demonstrated to be useful for modeling complicated random dynamics. Therefore, these methods provide the required theoretical framework for perturbed PPSs representation. Using this method, a jitter model able to represent aperiodic vocal folds oscillations was proposed,[19] and it was used to analyze the influence of glottal and external factors in period perturbations. Moreover, this model was applied in hoarse voice synthesis, showing that hoarseness strongly depends on jitter dynamics.[22,23] Additionally, artificial voices synthesized by this method were used to evaluate the effects of experience and training of voice pathologists in the correct identification of periods in perturbed sustained vowels, under controlled conditions of jitter[24] and additive noise.[25]

Other strategies for PPS modeling have been published in speech literature, eg, strategies based on biomechanical models,[13,26] spectral information of perturbation signals,[27] and nonlinear or chaotic signal processing.[28] Although most of previously mentioned methods have been successfully applied in voice synthesis tasks and theoretical perturbation modeling, only few of them can be applied to real PPS analysis. As far as the authors know, none of these methods incorporates the principal PPS features pointed out by Schoentgen[19] into the objective analysis of real voice signals. Therefore, in this article, we propose a state space-based approach to analyze and model PPSs extracted from real sustained vowels. State-space methods (SSMs) allow combining the PPS features with model-based, stochastic, time-series analysis. Within this framework, real PPSs are analyzed, and stochastic trend and cycle structures possessing a straightforward relationship with PPS features are estimated. In addition, the performance of this method is evaluated through statistical tests.

MATERIALS AND METHODS

PPS samples

In this section, the required procedure to obtain the PPSs and the principal materials used throughout this article are presented.

Voice database

In this work, the database (DB) developed by the Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Lab[29] is used, which includes sustained vowels /a/ of healthy individuals and patients with a wide range of voice disorders—eg, organic, neurologic, traumatic, and psychogenic. Voice samples are accompanied by detailed medical information gathered from tests and professional opinions. Only voices from healthy participants were considered. The participants were 21 males and 32 females, 38.81 ± 8.49 years and 34.16 ± 7.87 years, respectively. All samples in the DB were collected in a controlled environment and the duration of each signal was 3 seconds. The sampling rate and quantization level were 50 kHz and 16 bit, respectively.

PPS calculation

Voice samples were processed to extract PPSs with *Praat* software, developed by Boersma and Weenink[30] of the Institute of Phonetic Sciences, University of Amsterdam. *Praat* is one of the most widely used software in objective voice perturbation analysis. It applies a short-term analysis procedure, where pitch periods are obtained by waveform-matching methods. The technique applies autocorrelation analysis to estimate the location of fixed points in the glottal cycle, called pitch marks, where two consecutive waveforms look maximally similar. Therefore, pitch periods are calculated as the difference between consecutive pitch marks and the PPS $\{P\_1,P\_1,\ldots,P\_N\}$ is obtained, with N the number of elements in the sequence. In Figure 1, 20 milliseconds of a typical sustained vowel /a/