STOCHASTIC MODELS OF PITCH JITTER AND AMPLITUDE SHIMMER FOR VOICE MODIFICATION

Dima Ruinskiy and Yizhar Lavner

1 Department of Computer Science, Tel-Hai Academic College, Israel
2 Israeli Development Center, Intel Corporation, Haifa, Israel

ABSTRACT

We present a voice modification algorithm for transforming a modal voice to a hoarse voice. The algorithm is based on modifying the jitter and shimmer of the voice, which are long known to be connected to hoarseness. A pitch-synchronous linear prediction analysis-resynthesis process is used to increase jitter and shimmer in modal voices, where the data for the jitter and the shimmer is obtained from a stochastic model based on naturally hoarse voices. A formal evaluation of the algorithm conducted with multiple voices and listeners showed that the produced voices are perceived as close to natural hoarseness, suggesting the applicability of the algorithm to various voice modification and voice synthesis tasks.

Index Terms— Hoarseness, voice modification, voice effects, jitter, shimmer.

1. INTRODUCTION

Voice modification is the task of altering the acoustic characteristics of the voice, while retaining the phonetic characteristics. That is, producing a voice that sounds different, but carries the same speech information, and is still perceived as natural. Examples include time-scale modification [1], pitch-scale modification [2], voice morphing [3], voice personality transformation [4], or manipulations of various voice qualities, such as hoarseness, creakiness, breathiness, and others [5]. In recent years, along the advancement of speech synthesizers, there is an increased interest in voice modification systems, which enable the ability to produce more natural, expressive, and personalized synthetic voices [6]. It is also a subject of interest in the professional music industry, [5, 7] and in voice transformation systems [8].

Among the various vocal quality types, a group of perceptually and acoustically similar voices, which is commonly referred to as rough or hoarse voices, has been the focus of extensive study (see for example [6-14]). Many of these studies aim to find the acoustic or the perceptive correlates of the hoarse or rough voices ([9, 11-14]). For example, it was found that hoarseness is related to measures of jitter, shimmer and noise levels (e.g., [12, 13]). Other studies confirmed that the perceived roughness is linked with high levels of jitter, but also found that this perception is affected by the relation between the jitter and the fundamental frequency [14].

Despite the multitude of studies on hoarse and rough voices, few of them attempt to synthesize such voices or to convert modal voice into hoarse voice [5, 6, 8]. In most such studies it was found that the naturalness of the effect is often dependent on the input voice.

In this study, we present an algorithm for transforming a modal voice to a hoarse voice, aimed at the professional music industry. The goal of the algorithm is to introduce hoarseness in a singing or speaking voice, while retaining its naturalness. Additionally, the hoarseness model produced by the algorithm may be used to enhance voice synthesis systems, by introducing different vocal qualities to the synthesized voices.

The transformation is achieved by modeling two key parameters, whose relation to hoarseness has been established in previous studies – pitch jitter (e.g., [15, 16]) and amplitude shimmer (e.g., [17]). Jitter refers to the short-term (cycle-to-cycle) perturbation in the fundamental frequency, while shimmer refers to the short-term perturbation in amplitude of the voice [18].

Our algorithm constructs a model based on statistical analysis of natural hoarse voices, and utilizes it to modify the jitter and shimmer properties of a modal voice, using a pitch-synchronous linear prediction analysis-resynthesis mechanism.

Double-blind listening tests conducted with nine listeners, using several different voices showed that the algorithm is capable of producing hoarse voices that are perceived as close to natural, suggesting the applicability of the algorithm to various voice modification and voice synthesis tasks.

2. PITCH AND JITTER IN SPEECH

Voiced speech is produced by the quasi-periodic vibrations of the vocal folds, modulated by the vocal tract. The speech waveform, therefore, has a locally pseudo-periodic nature: for the signal denoted by \( x(t) \), at any point \( t_0 \) there exists a pseudo-period \( T \), such that \( x(t) \approx x(t + T) \) for each \( t \in [t_0, t_0 + T) \). The fundamental period \( T \) for which the above holds is called the local pitch period of the signal, with the
corresponding fundamental frequency denoted as pitch frequency.

In regular speech, typical values of the pitch period are between 2.5 ms and 16 ms. The pitch period varies over time, depending on the utterance, intonation, emotional state of the speaker, and other factors. The long term pitch variations are often intentional, and are due to changes in the intonation or transitions between phonemes. The short-term changes are caused by the time-variant characteristics of the vocal system, and are usually not controllable: small, random fluctuations exist between consecutive glottal cycles even in sustained vowels. These small variations are commonly referred to as jitter, and exist naturally in every voice.

In modal voices, the jitter typically varies between 0.1% and 1% of the pitch period [19], whereas for hoarse voices it is typically higher [11]. In our empirical tests, the strength of the jitter was found to be up to 5% in hoarse voices. An example is shown in Figure 1, which compares the pitch contours of the same male singer, singing once in a modal voice, and once in a hoarse voice, which exhibits noticeably stronger jitter.

In order to introduce hoarseness in modal voices, we analyzed the jitter patterns occurring naturally in hoarse voices. A schematic description of the analysis framework is shown in Figure 2. After segmentation of the signal to voiced/unvoiced sections, we obtain the local pitch period using a two stage detector, based on the real cepstrum (for initial detection) and the short-time normalized cross-correlation (for refinement). After the pitch contour is obtained, the average local pitch and the long-term variations are cancelled out, using a high-pass filter, leaving only the short-term variations, namely the jitter.

The jitter derivative (first order difference function) is also computed, as shown in the example in Figure 3. Next, we extract statistics on the strength of the jitter (relative to the instantaneous pitch), and on the possible durations of the jitter trend - the number of consecutive pitch cycles without sign alternations of the jitter derivative. Our analysis has shown that these trends tend to be very short, typically only a single cycle, and occasionally 2-4 cycles, as demonstrated in Figure 3 (bottom plot). We calculate the probabilities of each possible trend duration (1, 2, 3 and 4+ cycles), and for each such case construct a jitter bank, which stores the possible jitter strength values. The jitter banks are used to model the jitter in the synthesis stage, where an artificial jitter contour is constructed for synthesizing hoarseness in a given input voice (see Section 4).

In our approach, jitter is simulated by stretching or shortening each pitch cycle according to a jitter factor associated with that cycle. A vector of such jitter factors denoted as jitter contour. In this vector, $J_i$ is the local jitter factor of each pitch cycle $i$, whose length is $P_i$ samples. Positive values of $J_i$ indicate stretching, whereas negative values indicate shortening. Thus, the synthesized pitch cycle is $P_i + J_i$ samples long (Figure 4).

In order to introduce hoarseness in modal voices, we
analyzed the jitter patterns occurring naturally in hoarse voices. A schematic description of the analysis framework is shown in Figure 2. After segmentation of the signal to voiced/unvoiced sections, we obtain the local pitch period using a two stage detector, based on the real cepstrum (for initial detection) and the short-time normalized cross-correlation (for refinement). After the pitch contour is obtained, the average local pitch and the long-term variations are cancelled out, using a high-pass filter, leaving only the short-term variations, namely the jitter.

The jitter derivative (first order difference function) is also computed, as shown in the example in Figure 3. Next, we extract statistics on the strength of the jitter (relative to the instantaneous pitch), and on the possible durations of the jitter trend - the number of consecutive pitch cycles without sign alternations of the jitter derivative. Our analysis has shown that these trends tend to be very short, typically only a single cycle, and occasionally 2-4 cycles, as demonstrated in Figure 3 (bottom plot). We calculate the probabilities of each possible trend duration (1, 2, 3 and 4+ cycles), and for each such case construct a jitter bank, which stores the possible jitter strength values. The jitter banks are used to model the jitter in the synthesis stage, where an artificial jitter contour is constructed for synthesizing hoarseness in a given input voice (see Section 4).

In our approach, jitter is simulated by stretching or shortening each pitch cycle according to a jitter factor associated with that cycle. A vector of such jitter factors denoted as jitter contour. In this vector, $J_i$ is the local jitter factor of each pitch cycle $i$, whose length is $P_i$ samples. Positive values of $J_i$ indicate stretching, whereas negative values indicate shortening. Thus, the synthesized pitch cycle is $P_i + J_i$ samples long (Figure 4).

In order to introduce hoarseness in modal voices, we
analyzed the jitter patterns occurring naturally in hoarse voices. A schematic description of the analysis framework is shown in Figure 2. After segmentation of the signal to voiced/unvoiced sections, we obtain the local pitch period using a two stage detector, based on the real cepstrum (for initial detection) and the short-time normalized cross-correlation (for refinement). After the pitch contour is obtained, the average local pitch and the long-term variations are cancelled out, using a high-pass filter, leaving only the short-term variations, namely the jitter.

The jitter derivative (first order difference function) is also computed, as shown in the example in Figure 3. Next, we extract statistics on the strength of the jitter (relative to the instantaneous pitch), and on the possible durations of the jitter trend - the number of consecutive pitch cycles without sign alternations of the jitter derivative. Our analysis has shown that these trends tend to be very short, typically only a single cycle, and occasionally 2-4 cycles, as demonstrated in Figure 3 (bottom plot). We calculate the probabilities of each possible trend duration (1, 2, 3 and 4+ cycles), and for each such case construct a jitter bank, which stores the possible jitter strength values. The jitter banks are used to model the jitter in the synthesis stage, where an artificial jitter contour is constructed for synthesizing hoarseness in a given input voice (see Section 4).

In our approach, jitter is simulated by stretching or shortening each pitch cycle according to a jitter factor associated with that cycle. A vector of such jitter factors denoted as jitter contour. In this vector, $J_i$ is the local jitter factor of each pitch cycle $i$, whose length is $P_i$ samples. Positive values of $J_i$ indicate stretching, whereas negative values indicate shortening. Thus, the synthesized pitch cycle is $P_i + J_i$ samples long (Figure 4).
The jitter is modeled as a stochastic variable with short-term memory, constructed point-by-point. The value of the jitter at any point may depend on (up to 4) past samples, as follows: at each step, the jitter trend (positive or negative) and its duration (number of cycles between 1 and 4) are chosen randomly, based on the probabilities calculated in the analysis stage. The strength of the jitter for each of the cycles is selected from the suitable jitter bank (see Section 3).

In preliminary experiments this model was shown to yield hoarse voices that sound more natural than the simpler Gaussian memory-less model, where the value of the jitter at each step is simply taken at random from the jitter bank, according to a Gaussian distribution. The naive approach of a fixed sign-alternating jitter was even less successful, leading to a voice that was perceived as "metallic" or "buzzy".

The generated jitter contours are normalized, so as to contain only information about the trend and the relative jitter strength. The actual amount of jitter for each pitch cycle is determined as a combination of several parameters:
- The length of the pitch cycle (instantaneous pitch)
- The local value of the normalized jitter contour $J_i$.
- The maximum relative jitter $J_{max} \in (0,1)$. $J_{max}$ is user-defined, to control the strength of the hoarseness effect.
- The short-time energy of the pitch cycle $E_i$ (computed relative to the minimum energy $E_{min}$ and the maximum energy $E_{max}$ measured across the signal). The naturalness of the produced voice increases when jitter and energy are correlated, as was demonstrated in preliminary tests.

The local jitter factor $J_i$ is determined by the following formula:

$$J_i = \left[ J_{max} \cdot J_0 \cdot C_i \cdot R_i \right],$$

where

$$C_i = 0.5 - \frac{E_i - E_{min}}{E_{max} - E_{min}} \cdot 0.5$$

(2)

is a scaling factor, ranging from 0.5 to 1, depending on the local energy. $J_i$ is rounded to the nearest integer because its value must be determined in samples.

5. THE HOARSENESS SYNTHESIS ALGORITHM

The synthesis algorithm consists of the following three steps:
- Pitch-synchronous analysis, using Linear Prediction (LPC) inverse filtering.
- Introduction of jitter by modifying the LPC residual of individual pitch cycles.
- Reconstruction of the voice using the LPC synthesis filter.

A schematic block diagram of the process is shown in Figure 5. The LPC residual, denoted as $e[n]$, is computed using LPC inverse filtering, where

$$e[n] = x[n] - \sum_{i=1}^{M} a_i x[n-i],$$

(3)

$M = 12$ is the filter order and the coefficients $a_i$ are obtained by the Levinson-Durbin algorithm applied separately to each pitch cycle.

The residual signal $e[n]$ contains mostly the information related to the glottal pulse and the fundamental frequency, since the information related to the formants is eliminated by the inverse filtering. Unvoiced segments are unchanged by the algorithm, since they do not exhibit pitch, and therefore, jitter is not applicable. Indeed, since vocal folds vibrations do not contribute to unvoiced speech, hoarseness and other voice modalities are irrelevant.

The residual of each pitch cycle, obtained through the inverse filtering, is either stretched or compressed by a few samples, according to the desired jitter factor (see Section 4), as follows: if the $i$-th pitch cycle is $P_i$ samples long, and the jitter factor is $J_i$ samples ($J_i$ may be negative), the residual is resampled by a factor of $(P_i + J_i)/P_i$. The resampling is performed using a poly-phase FIR filter implementation, with a Kaiser window to minimize distortions.

Finally, the signal is reconstructed cycle-by-cycle from the resampled residual signal $\hat{e}[n]$, using the LPC synthesis filter:

$$y[n] = \hat{e}[n] + \sum_{i=1}^{M} a_i \hat{e}[n-i],$$

At this step, shimmer may be applied to the signal, as described in the following section.

![Figure 5](image)

Figure 5. A block diagram of the process of introducing jitter to a modal voice.

6. SHIMMER

Shimmer is defined as the fluctuation of amplitudes of consecutive pitch cycles of voiced speech. Following previous studies, such as [17], which established a relation between the amount of shimmer and the hoarseness of the voice, we attempted to use shimmer in combination with jitter, in order to enhance the hoarse effect. This was done by multiplying each pitch cycle by a special type of window function, with varying peak amplitudes:

$$\tilde{x}_i[n] = x[n] \cdot (1 + S_i \cdot w[n]),$$

(5)
where \( x[n] \) is the pitch cycle (after application of the jitter procedure), \( S, \in (-1,1) \) is the designated shimmer factor and \( w[n] \) is a Hamming window. This window function achieves maximum shimmer near the middle of the pitch cycle, while ensuring continuity at the boundaries of consecutive cycles. The shimmer factor \( S \) varies over time, similarly to the jitter factor \( J \) (see Section 4), and is bounded by the maximum shimmer \( S_{max} \). It was experimentally found that using values of \( S_{max} < 0.2 \) can enhance the effect achieved by the jitter, whereas higher values can lead to noticeable distortions and degradation of the voice.

7. RESULTS

The evaluation of the algorithm has been carried out using a subjective listening test, in which listeners had to judge the level of hoarseness and naturalness in several voices, sampled and digitized using a sampling frequency of 44.1 KHz. Voices of four different singers, two male, and two female, with different pitches and timbres were used. The average duration of the recordings was 17 seconds. All the original voices were modal or very close to modal. For each voice, the algorithm was applied to construct three hoarse samples – with light jitter/shimmer (3-4%), with slightly stronger jitter/shimmer (5-6%), and with extremely strong jitter/shimmer (20%). The jitter was modeled according to jitter banks extracted from voices of two hoarse speakers, male and female, each used for the corresponding gender.

Nine listeners participated in a subjective evaluation test, all of them without any hearing loss. Five samples of each of the four speakers’ voices were used: the original (clear) voice, the three hoarse samples constructed by the algorithm, and a distracter constructed using the naïve sign-alternating approach known to yield metallic sound.

For each voice, the listeners were presented with the five samples in a double-blind setting (neither the listeners nor the testers knew the order of the samples). The listeners were asked to grade each of the voices according to two criteria: hoarseness and naturalness, on a free scale between 0 and 100 (using sliders). For the evaluation, the listeners could play each sample any number of times, in any order, before assigning the grades.

The results are shown on Figure 6 for each of the voices separately. Each column indicates the average of the scores assigned by the listeners. All listeners perceived the samples generated by algorithm as noticeably hoarse, when compared to the original modal voices. However, the naturalness of the voices decreased as the hoarseness increased, as evident by the grades given to the version with 20% jitter. The “metallic” sound also received significantly lower marks in naturalness, even though its hoarseness marks were high.

It is also evident from the results that the success of the algorithm can depend on the original voice, with M1/F1 receiving significantly lower marks in naturalness compared to M2/F2.

![Figure 6. The hoarseness and naturalness grades for each of the four voices.](image)

8. CONCLUSION

We presented an algorithm for transforming a modal voice into a hoarse voice, by controlling two parameters of the speech waveform, whose correlation with hoarseness has been previously established – jitter and shimmer. Our experimental results showed that enhancing these two parameters leads to a voice that is perceived as hoarse. The results demonstrate a few limitations of the algorithm, namely, the decrease of the perceived naturalness as the hoarseness increases, and the large variation in naturalness between different voices. However, in all cases tested, it was possible to synthesize a voice that sounds noticeably hoarse, while retaining most of its naturalness. Thus, the algorithm seems to be a promising direction in hoarseness emulation for voice synthesis and voice modification applications.

Additional improvement may be achieved by further study of the properties of jitter and shimmer in hoarse voices, in order to obtain methods that can better preserve the naturalness of the modified voice. Another approach is combining the presented algorithm with other techniques. For example, in order to achieve natural hoarseness in high-pitched voices, it may be desirable to slightly lower the fundamental frequency, using a pitch-scale modification algorithm.
9. ACKNOWLEDGMENT

We would like to thank Mr. Meir Shaashua and Mr. Itai Neoran, both of Waves Audio, for valuable ideas and discussions. We also wish to thank Mr. Yefim Yakir for technical assistance. This study was partially supported by Waves Audio.

REFERENCES