

ANALYSIS BY SYNTHESIS OF FM MODULATION AND ASPIRATION NOISE COMPONENTS IN PATHOLOGICAL VOICES

Brian Gabelman and Abeer Alwan

Department of Electrical Engineering, UCLA
Los Angeles 90095, USA
Gabelman@ucla.edu, alwan@icssl.ucla.edu

ABSTRACT

FM and source noise characteristics of pathological voices are analyzed and modeled using precision interpolating pitch tracking. Detailed tracking data allows segregation of pitch variations into low frequency (tremor) and high frequency pitch variation (HFPV) time series. Tremor data is used to resample the original voice into a quasi-constant pitch signal, which results in a more accurate source noise estimate using the noise analysis algorithm described by de Krom [1]. Gaussian distributions are used for both source HFPV and aspiration noise models. Combined analysis parameters are used to drive a formant synthesizer, resulting in improved perceived fidelity.

1. INTRODUCTION

As part of our on-going effort to accurately model and synthesize pathological voices, this study focuses on two of the aperiodic components of pathological voices: FM modulation of both high frequency (HFPV) and low frequency (tremor), and aspiration noise. Analysis of pathological voice samples is performed to extract formants, source flow derivative, LF source model [2], pitch, tremor, HFPV, and NSR (ratio of aperiodic "noise" to periodic "signal"). Using the source - vocal tract filter model, synthesis is performed using the parameters determined in the analysis. Aperiodic components are simulated using FM modulation and the addition of spectrally-shaped noise to the source.

Measurement of the aspiration noise component of a voice is performed using a modification of [1]. A limitation of that approach is the tradeoff between the inaccuracies due to the windowing effects of a short time sample and the instabilities of the voice in a long time sample. The current approach overcomes one limitation of longer sample times by removing errors in calculated NSR due to variation in pitch during the sample. The original voice is re-sampled to remove low frequency pitch variation before noise analysis is performed, thus permitting the use of a longer time sample. Changes in measured NSR of as much as 12 dB are observed with pitch stabilization, matching perceived noise levels more closely.

Previous investigations have used a variety of measures of HFPV, usually some type of average. The current approach refines HFPV analysis and synthesis by modeling the variation in the pitch period as a Gaussian distribution.

2. ANALYSIS OF PATHOLOGICAL VOICES

All steps of data analysis are summarized in Fig. 1

2.1 Data collection

Voice samples were collected by Dr. B. Gerratt at UCLA. A B&K condenser microphone with a flat response to 20 kHz is used to record 1 second samples of patients vocalizing the sustained vowel /a/. The signal is low-pass filtered with a FIR filter and decimated to 10 kHz. Thirty-one voices are randomly selected representing a range of disorders, age, and gender.

2.2 Formant analysis

Formant extraction is done using LPC analysis. A group of selected pulses are interactively analyzed [3], generating average formant values across the open and closed phase of selected pulses. The operator checks results for reasonable spectral tilt, expected formant frequencies, bandwidths, and number of formants for /a/.

2.3 Inverse filtering

The user selects specific single pulses and estimates the length of the glottal closed phase. The covariance method of LPC is applied to the targeted window, and a new vocal tract model, inverse filter, and flow derivative waveform are generated. The user has the ability to vary pole locations and observe the flow derivative for improvement (formant ripple reduction).

2.4 LF model fitting

The estimated source flow derivative waveform is fitted to a simplified LF model [2]. The fit is performed by first identifying key features of the pulse (time of zero crossing, minima, time of minima, and decay half-point) to determine a first approximation for LF parameters. From this estimate, least squares minimization is used across the model's four degrees of freedom, determining a best fit. Fig. 2 illustrates the LF model used in this study.

2.5 Pitch tracking

Time domain pitch tracking is carried out pulse by pulse, so that short duration variations (HFPV) may be captured. Pitch tracking allows four variations of maxima/minima detection for cycle-marking: original pulse, source, smoothed derivative, or smoothed differentiated source. In difficult cases the user may manually mark features on as many pulses as necessary to reestablish track lock. Interpolation is used to determine pitch periods to less than one sample period (< 0.1 ms): a parabola

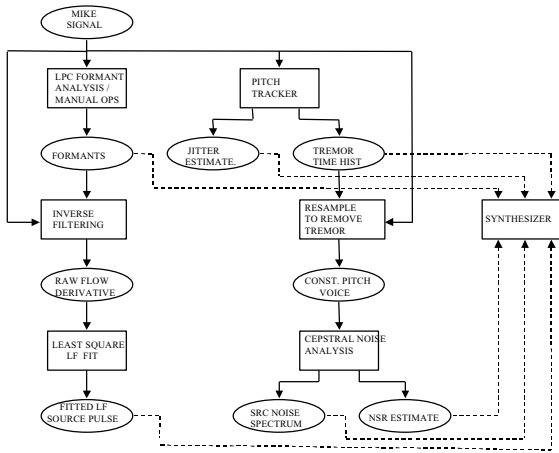


Figure 1. Overview of voice analysis, resulting in six control parameters as input to the synthesizer.

with user-selected number of points is fitted to the maxima/minima, and the "true" minima calculated from the parabolic vertex. Two fairly distinct types of pitch variation (FM modulation) are seen: low frequency (<10 Hz) changes associated with tremor, and high frequency (>10 Hz) cycle to cycle variation associated with HFPV. Failures of automatic pitch tracking appear as discontinuities in this plot, and are easily detected and rerun for correction.

The high and low frequency components of the pitch track are respectively segregated into HFPV and tremor time histories using high and low pass filters with a cutoff frequency of 10 Hz. Fig. 3 displays the result for one pitch track. To verify the success of pitch tracking, the statistical distribution of HFPV is displayed by histogramming the high pass filtered pitch periods. Successful tracking is characterized by a Gaussian distribution with one standard deviation of usually less than 1 or 2 percent. The measured standard deviation of pitch period, in units of percent, is also used to define the level of HFPV for later input to the synthesizer.

2.6 Ratio of aperiodic to periodic energy (NSR)

The relative amounts of aperiodic and periodic energy in the original voice are estimated using cepstral comb-lifering [1]: the cepstrum of the original voice is calculated, a comb lifter is calculated from the measured pitch and applied to the cepstrum to remove periodic energy, and the result transformed back to the frequency domain, yielding an estimate of the spectrum of aperiodic energy. The energy content of the aperiodic spectrum is subtracted from the total, yielding the periodic energy and thus the ratio of aperiodic to periodic energy (NSR). Note that the resultant aperiodic energy is largely due to aspiration noise, but may also include other effects such as HFPV, shimmer, etc.

2.6.1 Effect of tremor on cepstral peak width

The notch width of the comb lifter is a tradeoff. Wider notches remove "rahmonic peaks" that have been widened by both the variations in pitch frequency over the sample and the windowing effect of short duration time samples. However, wider notches

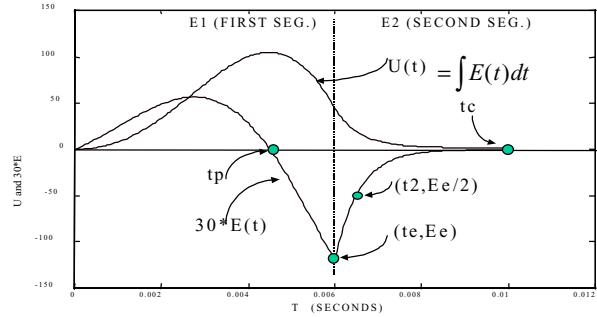


Figure 2. Modified LF model used to fit the calculated flow derivative [2].

also remove a greater portion of the background aperiodic energy underlying the notch, giving rise to lower than actual estimate of aperiodic energy. Narrow notches fail to completely remove "rahmonic" peaks widened by the pitch variation and windowing, resulting in periodic energy being included in the aperiodic estimate and thus giving rise to a higher than actual estimate of aperiodic energy.

A similar tradeoff exists with sample duration. Short samples cause widened cepstral peaks. Longer samples alleviate the effect of windowing, but peaks may widen due to low frequency pitch variation (tremor) over the duration of the sample. The ideal solution is a long, pitch stable sample, allowing narrow cepstral notch filtering and producing a more accurate NSR estimate. Using the tremor time history, it is possible to resample the original voice and generate a pitch stable time series. Using the tremor time history, a vector of unevenly spaced resampling times is created by making the instantaneous sample interval inversely proportional to the instantaneous pitch frequency. Simple linear interpolation of the original time series on this modified time vector creates a version of the original voice with pitch variation removed. Other variations, such as formant modulation may, however, still be present in some cases.

Fig. 4 illustrates the effect of re-sampling on pitch variation for a tremulous voice female voice: the re-sampled voice shows greatly reduced pitch variation. Fig. 5 illustrates the effect of re-sampling on the power spectrum of the voice: the peaks associated with harmonic energy are greatly narrowed. The cepstral peaks are similarly narrowed, resulting in a more accurate estimate of the ratio of aperiodic to periodic energy. In the case illustrated, there is a 12 dB decrease in estimated NSR.

2.6.2 The aperiodic spectrum

Returning to the cepstral analysis, an estimate of the spectral shape of the source aperiodic component is calculated as in [1]. In order to generate the spectrum of the source alone, the vocal tract log magnitude frequency response is generated from the measured formants and subtracted from the aperiodic spectrum. The resulting source aperiodic spectrum is smoothed in the frequency domain with a 100 point triangular window moving average filter. The process is completed by fitting the remaining spectral shape with a 25-point piecewise-linear

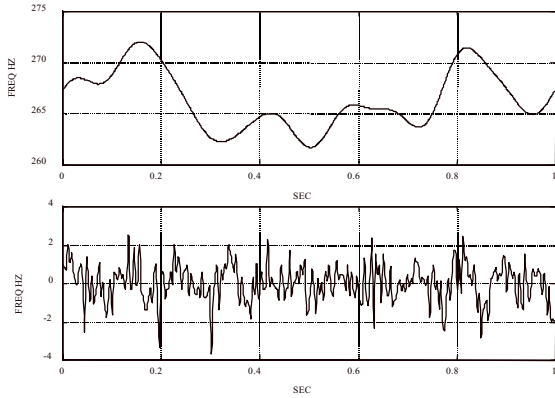


Figure 3. The measured pitch track is segregated into low frequency (< 10 Hz) tremor in the upper plot, and high frequency HFPV (> 10 Hz) in the lower plot.

model. In most cases the resulting spectrum is fairly flat. Source spectral calculation completes the analysis, and together with measured formants, pitch track, HFPV, NSR, and source flow derivative waveform are taken to model the original voice (Fig. 1).

3. SYNTHESIS OF PATHOLOGICAL VOICES

Using the derived six parameters describing a voice, a synthetic version is calculated using a time domain synthesizer written in MATLAB. Most of the synthesis is analogous to the analysis steps described above.

3.1 Basic waveform generation

Using the estimated LF parameters, a basic waveshape of the glottal flow derivative is calculated (Fig. 2), using a parametric time scale normalized to one pulse period. The amplitude is normalized to unity, and this waveshape is used throughout the simulated voice by concatenation. The effects of pitch changes and HFPV are created by variation in the sample instants chosen for interpolation of the basic waveshape.

3.2 Source synthesis - pitch variation due to tremor

In order to simulate low frequency variations in pitch, the source waveshape is effectively stretched or compressed in time such that the period of one pitch pulse is exactly the reciprocal of the instantaneous frequency at the beginning of each pulse.

To calculate specific samples for each pulse, the instantaneous frequency is used, along with the absolute finish time of the last sample of the previous pulse, to convert sample instants to phase arguments specifying abscissa values on the LF waveshape. The final LF samples are then generated via linear interpolation at these abscissa values. In this manner, changes in pitch specified by the pitch track from analysis are smoothly generated, with no perceptually discernable jumps in frequency. By contrast, when pitch variation is implemented via simple truncation or addition of samples to the pulse, a quantization effect is generated, creating the impression of "steps" in pitch during linear changes in pitch frequency.

3.3 HFPV implementation

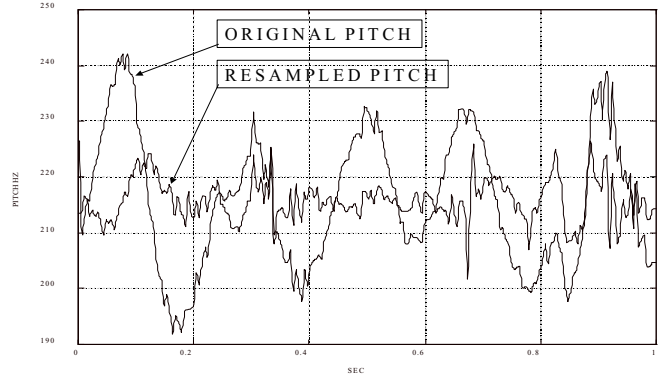


Figure 4. Pitch tracks of a tremulous voice. The original voice has a 5 Hz pitch tremor with a large 20 Hz deviation. The resampled signal, plotted on the same scale, has greatly reduced deviation.

HFPV effects are implemented by modifying the instantaneous frequency. Starting with the pitch period determined by the instantaneous frequency from the pitch track or set by the user, a random variation in pitch period is created by multiplying a random number with Gaussian distribution, unity mean, and variance determined by the desired level of HFPV. One standard deviation corresponds to 100%.

Unfortunately setting the random number variance equal to the desired level of HFPV does not produce this same level of HFPV in the synthesized source time series; when the HFPV analysis is applied to the synthetic signal produced, a smaller level of HFPV is always measured. The cause of this discrepancy is illustrated in Fig.6, which illustrates synthesis of two successive flow derivative waveforms. Note that although the length of each pulse is determined by a single random number, the peak to peak interval, which is measured by the pitch tracker, is determined by the sum of fractions of two random subintervals, as shown in Fig. 6 and Eq 1.

$$T_{pp} = (1 - a)T_1 + aT_2 \quad [1]$$

Where T_{pp} = measured negative peak to peak interval, T_1, T_2 = first and second pitch periods, respectively, generated by $T_i = T(1 + (PJ/100)R_i)$, PJ = percent HFPV set in synthesizer, R_i = Gaussian random number with zero mean and 1.0 sigma, a = fractional position of negative peak = T_e/T , T = unmodified pitch period, T_e = time of negative peak in pulse. The expected variance of T_{pp} is the sum of the variances of the two components: $V = V_1 + V_2$, where the variances are: $V = (T PJ/100)^2$, $V_1 = (a T PJ/100)^2$, $V_2 = ((1-a) T PJ/100)^2$, and PJ_f = resulting percent HFPV in T_{pp} . Solving for PJ_f as a function of PJ and peak position a yields the relationship in Eq. 2:

$$PJ_f = PJ \sqrt{2(a^2) - 2a + 1} \quad [2]$$

The validity of this relation was confirmed with a Monte Carlo MATLAB simulation of peak-to-peak interval measurement using averages of 100000 samples for a range of a . Thus, a correction factor of $1/\sqrt{2(a^2) - 2a + 1}$ must be applied in the synthesizer when simulating HFPV.

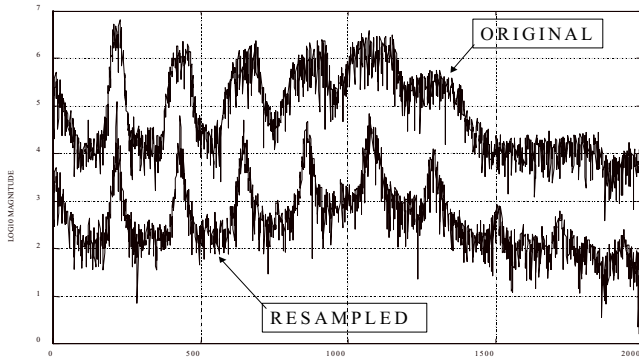


Figure 5. Power spectra of original tremulous voice (upper) and re-sampled voice (shifted down 2 units).

3.4 Aspiration noise implementation

The final step in source synthesis is the addition of noise to simulate aspiration at the glottis. The implicit assumption here is that aperiodic signal content other than HFPV is modeled by aspiration noise. Indications at this point are that this is approximately true for a subset of pathological voices. The statistical distribution, spectral shape, and energy level of this added noise are controlled to match the original voice.

3.4.1 Source noise spectral shaping

White noise with Gaussian distribution and unity variance is first generated. A 100-tap FIR filter is synthesized to match the spectral shape of the original source (25 point piecewise linear approximation determined from analysis); the noise is passed through the filter to match the original noise source shape.

3.4.2 Source noise energy level

Besides the spectral shape, the source noise energy level is adjusted such that the ratio of synthetic source Gaussian noise energy to energy of the HFPV periodic source is equal to the measured NSR of the resampled source. This presumes, as is tested below, that HFPV has a small effect on measured NSR. It also presumes that aspiration noise is the major component of non-periodic signal. All non-periodic effects in the original voice are being modeled with HFPV equal to the original HFPV and synthetic aspiration noise to match the measured NSR. The calculated source Gaussian noise level is used to calculate a noise signal gain, which is applied to the noise time series, which is then added to the LF time series to generate the source function.

3.5 Vocal tract model

Formants determined in the analysis are converted to an all-pole resonator filter, and applied to the source time series to generate the final synthetic time series. The synthesizer automatically normalizes the signal to the full range of the D/A.

4. RESULTS

Synthetic voices are analyzed and compared with the original voices in the following ways: 1. The HFPV level of the synthetic voice in the absence of a Gaussian noise source is verified equal to the measured level. 2. The synthetic NSR, in

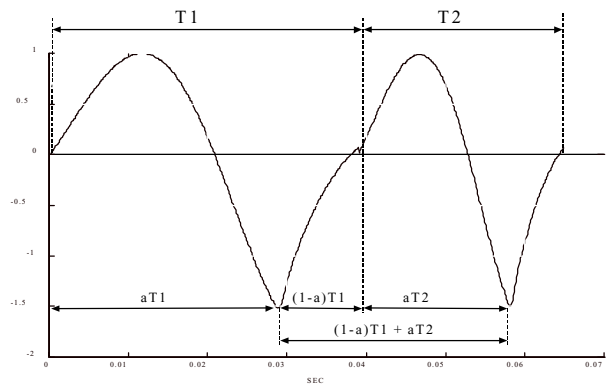


Figure 6. Two synthetic flow derivative waveforms.

the absence of HFPV, is measured and verified equal to the level of Gaussian source noise. 3. Listening experiments are conducted to compare listener adjusted synthetic source HFPV and Gaussian noise levels with measured original values. Comparisons reveal that HFPV and aspiration noise seem to have little effect on each other when present at the levels measured in pathological voices. Initial analysis by synthesis experiments reveal correlation between original measured HFPV and Gaussian noise levels with listener-set values. The combined functionality of all synthesizer parameters gives rise to synthetic signals which are of high quality and in some cases indistinguishable from the original.

5. SUMMARY

Pathological voices are analyzed and resynthesized with aperiodic qualities modeled in terms of FM modulation and a Gaussian noise source, which seem to account for a major portion of aperiodic components in the signal. Detailed tracking data allows segregation of pitch variations into low frequency (tremor) and high frequency pitch variation (HFPV) time series. Tremor data is used to resample the original voice into a quasi-constant pitch signal, which results in a more accurate source noise estimate using the noise analysis algorithm described by de Krom [1]. Gaussian distributions are used for both source HFPV and aspiration noise models. Combined analysis parameters are used to drive a formant synthesizer, resulting in improved perceived fidelity. Work was supported in part by NIH/NIDCD grant DC01797.

6. REFERENCES

1. Krom, Guus de, 1993. "A Cepstrum-Based Technique for Determining a Harmonics -to-Noise Ratio in Speech Signals," JSHR 93, Vol 36, 254-266.
2. Qi, Y., and Bi, N. 1994. "A simplified approximation of the four-parameter LF model of voice source," JASA 96, 1182-1185.
3. The inverse filter program developed by Norma Antonanzas can be investigated online at the following website: www.surgery.medsch.ucla.edu/glottalaffairs/software_of_the_boga.htm