

Chapter 5

Synthesis Algorithms and Validation

An essential step in the study of pathological voices is re-synthesis; clear and immediate evidence of the success and accuracy of modeling efforts is provided by comparing the original and synthetic versions of the pathological voice. The effects of variations of each of the model parameters may be quickly evaluated perceptually by generating synthetic voice samples with an easily controlled synthesizer. Tests may be performed to validate analysis results, and experiments may be performed to determine the effects on the listener of variations and interactions of model parameters. In this section, the details of algorithms used to synthesize pathological vowels are described. Experiments confirming the success of synthesis are then explained.

5.1 Synthesis Algorithms

This section describes algorithms used by the synthesizers to regenerate a synthetic version time series of the original pathological vowels. Using the derived analysis model parameters describing the pathological voices (formants, glottal source waveform, aspiration level and spectral shape, tremor, HFPV, and low and high frequency power variation), a synthetic version was calculated for each original pathological voice sample. Most of the steps of the synthesis process have direct analogs in the analysis steps described in Chapter 2. The software synthesizer implements the most current algorithms.

5.1.1 Basic Waveform Generation

The modified LF model [31], with its ease of use and adaptability to a variety of waveforms, is currently chosen as the most useful source waveform model for synthesis of pathological voices. Using the estimated LF parameters as described in Section 2.2.2, a basic waveshape of the glottal flow derivative is calculated (Fig. 2.13 and Fig. 2.15) 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 LF waveshape is assumed to remain constant in the current implementations of the synthesizers. The effects of fundamental frequency changes due to tremor and HFPV are created by variation in the sample instants chosen for interpolation of the calculated basic LF waveshape, as described in Section 5.1.2.

5.1.2 Source Synthesis – Low Frequency Fundamental Frequency Variation

In order to simulate base (low frequency) variations in fundamental frequency, the source waveshape is effectively stretched or compressed in time such that the period of one fundamental frequency pulse in actual time is exactly the reciprocal of the desired instantaneous frequency. This changes the number of actual time samples interpolated on the LF pulse waveshape. To raise fundamental frequency, fewer samples are selected from the fitted LF pulse; to lower fundamental frequency, additional samples are selected. These interpolation points are chosen equally spaced along the LF waveshape, with their spacing inversely proportional to the desired frequency. The synthesizer provides several options for selection of the base frequency:

1. A constant value, such as the average of the low-pass filtered (tremor) frequency of the original voice (for example, the average value of the top curve in Fig. 2.21).
2. A sinusoidally varying frequency about the mean F0 value. The user selects the frequency of variation, and extent of variation (deviation).
3. A randomly varying frequency about the mean F0 value generated by low pass filtering of Gaussian noise. The user selects the extent of variation (deviation) and the filter cutoff, which effectively determines the mean frequency of variation.
4. The same tremor as the original voice. The base value of fundamental frequency is obtained from interpolation on the low pass filtered fundamental frequency track (tremor)

of the original voice (for example, the top curve in Fig. 2.21). The instant of interpolation on the tremor track is selected using the time of the first sample of the currently being constructed LF pulse in the simulated time series; fundamental frequency is not varied within a single source pulse.

To calculate the 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 in real time 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 fundamental frequency specified by the selected fundamental frequency generation method are smoothly produced, with no perceptually discernable jumps in frequency. By contrast, when fundamental frequency variation is implemented via simple truncation or addition of samples to the pulse, a quantization effect is generated, creating the impression of "steps" in fundamental frequency during linear changes in fundamental frequency.

5.1.3 Source Synthesis – High Frequency Fundamental Frequency Variation

High frequency fundamental frequency variations are simulated in the same manner as low frequency variations by effectively changing the instantaneous fundamental frequency with fundamental period modification. HFPV can be applied in the synthesizer independently of the low frequency fundamental frequency variations. As

each new fundamental frequency pulse is synthesized, the base fundamental period determined by any of the methods mentioned (Section 5.1.2) is perturbed by a random increment to lengthen or shorten it, thus modeling the measured HFPV (Sections 2.3.3-2.3.4). The random incremental change in fundamental period length is created by generating a random modification factor with Gaussian distribution, unity mean, and standard deviation determined by the desired level (usually the measured value) of HFPV. Setting synthesizer jitter to 100% implies the creation of a standard deviation in fundamental period length equal to the fundamental period. This modification factor is then applied to the base fundamental period to arrive at the final synthetic fundamental period

Setting the modification factor to get the desired level of jitter in the synthetic signal as measured by the fundamental frequency tracker and analysis software involves a complication. Unfortunately, setting the standard deviation of the modification factor exactly equal to the level implied by the desired HFPV does not produce this same level of HFPV in the resulting 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. 5.1, 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 (T_{pp}), which is measured by the fundamental frequency tracker, is determined by the sum of fractions of two random subintervals, as shown in Fig. 5.1 and Eq 1.

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

And

$$T_1 = T(1 + (PJ/100)R_1),$$

$$T_2 = T(1 + (PJ/100)R_2),$$

Where:

T_{pp} = measured negative peak to peak interval,

T_1, T_2 = first and second fundamental periods,

PJ = percent HFPV set in synthesizer,

R_1, R_2 = Gaussian random numbers with zero mean and $\sigma = 1.0$,

a = fractional position of negative peak within the fundamental frequency pulse = T_e/T ,

T = unmodified fundamental 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:

$$V = V1 + V2,$$

where the variances are:

$$V = (T \text{ PJf}/100)^2,$$

$$V1 = (a T \text{ PJ}/100)^2,$$

$$V2 = ((1-a) T \text{ PJ} /100)^2,$$

and PJf = resulting percent HFPV in Tpp. Solving for PJf as a function of PJ and peak position a yields the relationship in Eq. 2:

$$\text{PJf} = \text{PJ} (2a^2 - 2a + 1)^{0.5} \quad [2].$$

The validity of this relation was confirmed with a Monte Carlo MATLAB simulation of fundamental pulse peak-to-peak interval measurement. The expected measured fundamental frequency period of the synthetic voice was calculated using averages of 100,000 randomly generated pulses for each of a range of a values. For each pair of simulated pulses, the predicted fundamental frequency period (as measured between adjacent minima as shown in Fig. 5.1) was calculated. This measurement was repeated 100,000 times and then averaged; the whole process was repeated for values of 0.1, 0.2, ...1.0 corresponding to negative peak positions ranging from the beginning to the end of the fundamental pulse. Fig. 5.2 displays the result of the simulation. The circles show the result of the simulation, and the line is the standard deviation predicted by Equation 2. There is good agreement, which improves with more samples. Thus, a correction factor of $1/(2a^2 - 2a + 1)^{0.5}$ must be applied to the desired level of HFPV to obtain the value to use in the synthesizer simulation equations when simulating HFPV.

5.1.4 Source Synthesis – Low Frequency Power Variation

In a manner analogous to low frequency FM synthesis, provision is made for applying low frequency power modulation to the synthesized voice. The measured low frequency power variations (Section 2.4.3) of the original voice can be applied to the synthetic voice to generate the intensity variations perceived by the listener in the original voice. Signal power is proportional to the square of the signal voltage. In order to apply these variations, a gain correction time series is generated that is proportional to the square root of the low frequency power variation (upper dashed curve in Fig. 2.27). The gain correction is then applied to the synthesized signal to achieve a power variation approximating the original voice.

5.1.5 Source Synthesis – High Frequency Power Variation

Similar to the HFPV synthesis, high frequency power variations (shimmer) are available in the synthesizer. Shimmer is synthesized in a manner analogous to the way it is measured, as a perturbation of pulse power with a Gaussian distribution. To synthesize pulses with randomly varying power, a Gaussian random gain is generated and applied to the samples of each fundamental pulse (the same gain value is used over all the samples within a pulse). The applied gain has unity mean and standard deviation determined by the amount of desired shimmer.

As with HFPV, there are many methods of measuring shimmer [3]. Assuming shimmer is a small perturbation of fundamental period length with a Gaussian distribution, linearity allows conversion between several types of measures, including gain, power, and dB. The percentage power variation measured in the analysis of the original voice (Section 2.4) can be converted to shimmer in dB (used as input in the synthesizer) and a gain value for fundamental frequency pulses (used in the synthesis equations). The nonlinear relations between these quantities are linearized about the mean value of shimmer to yield simplified formulae. In general, probability distributions of a nonlinear function of a variable with Gaussian distribution are themselves not Gaussian. Small perturbations in the conversion equations used here, however, are Gaussian as a reasonable approximation, allowing the use of standard deviation as a measure of shimmer. Therefore, the quadratic relation between power and gain simplifies to the approximation:

$$PPS = 2 * GPS$$

Where

GPS = percent gain variation (linear)

PPS = percent shimmer in power

$$= 100 * \text{standard deviation in power} / \text{mean power}$$

The logarithmic relation between power and dB simplifies to the approximation:

$$PPS = 10 * \ln(10) * DBS = 23.0 * DBS$$

Where

DBS = shimmer in dB

= standard deviation of signal dB measure

5.1.6 Aspiration Noise Implementation

The final step in source synthesis is the addition of spectrally shaped Gaussian noise to simulate aspiration at the glottis. The current model assumes high frequency (>10 Hz) nonperiodic signal content other than HFPV and shimmer is modeled by aspiration noise. This assumption appears to be approximately true for a subset of pathological voices in which an excellent synthetic match to the original is obtained with aspiration noise. The Gaussian statistical distribution and the spectral shape of source aspiration noise are preset in the synthesizer to the measured values of the corresponding original voice. The energy level of aspiration noise relative to the periodic signal level can be fine-tuned by the user via the adjustments available in the synthesizer.

5.1.6.1 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.

5.1.6.2 Source Noise Energy Level

In order to complete the calculation for inclusion of aspiration noise, the relative gain of the aspiration noise signal relative to the glottal source signal must be found. The preset or user adjusted aspiration noise level in dB is used to find the correct gain value. It is calculated using the relative energies of the glottal source and aspiration noise time series before they are summed to obtain the final synthetic source time series. The nominal value of aspiration noise to apply in order to achieve the best match to the original voice is determined via the cepstral filtering method described in Section 2.5.2.

5.1.7 Vocal Tract Model

The final step in voice synthesis is applying the vocal tract filter to the glottal flow derivative time series, which at this point includes the adjusted LF waveform and the selected levels of nonperiodic features, such as AM, FM, and aspiration noise. Currently, the synthesizer uses fixed formants for the entire time series. The formants determined in the analysis (Section 2.1) are converted to all-pole resonator filters, and applied to the source time series to generate the final synthetic time series. The synthesizer automatically normalizes the amplitude of the maximum excursion of the final time series signal to the full range of the D/A used for sound generation, thus minimizing quantization effects while preventing clipping.

5.2 Synthesis Validation

With skillful adjustment of synthesizer parameters (including aspiration noise, HFPV, and shimmer) it is possible to achieve synthetic samples that are very close to the original; in some cases, synthetic voices are indistinguishable from the original. Since one of the initial motivations for this project was creation of synthetic vowels as perceptually close to the original as possible, considerable effort was made to objectively and perceptually compare the resulting synthetic vowels with the originals after which they were modeled. In this section, the success of several aspects of analysis/synthesis is evaluated with tests addressing the nonperiodic model parameters. In order to objectively evaluate the accuracy and consistency of the overall analysis/synthesis process, the processing loop is closed by re-analyzing the synthetic voices with the same software used to analyze the original pathological voices. The levels of nonperiodic components in the synthetic versions are then checked to guarantee values consistent with original values.

5.2.1 Aspiration Noise (AN) Verification

In the absence of AM and FM modulations, the cepstral NSR measurement of the synthetic voice should reflect the value of shaped source noise set in the synthesizer when the voice was created, since any nonperiodic energy should be entirely due to this aspiration noise. For each of the 31 voices, synthetic versions were created with the levels of AM and FM modulation set to zero, and the level of aspiration noise set to that measured in the original voice. Using the same noise analysis procedure used on the original voice, the synthetic NSR was measured. The result is shown in Fig. 5.3, in which

the measured synthetic NSR is plotted against the measured original NSR for all 31 cases. The original voices span a measured NSR range of about -25 dB to -5 dB. Over this range, the agreement between natural and synthetic NSR is within about 1 dB, which is well within perceptible limits, as approximately determined by varying this parameter on the synthesizer and comparing the resulting vowels. Thus, the process of measurement and synthesis of aspiration noise appears consistent.

5.2.2 HFPV Verification

In a manner similar to the NSR verification, HFPV in the synthetic voice was checked against the value set in the synthesizer (which was the measured value in the original voice). The measured values of HFPV in the synthetic voices achieved agreement with that of the original voice to within 0.1%, which is well within perceptible limits. Thus, the process of measurement and synthesis of HFPV appears consistent.

5.2.3 Effect of AN on HFPV

Another relevant question is the degree of interaction between aspiration noise and HFPV. The addition of aspiration noise to the source time series would be expected to affect the measurement of HFPV due to perturbation of the position of time domain features (eg. peaks) detected by the fundamental frequency tracker. The relevant question is how significant is the effect for the levels of aspiration noise and HFPV measured in the set of original pathological voices. To assess the increment in measured HFPV due to

the inclusion of aspiration noise in the synthetic voices, a set of 31 voices was synthesized with the original levels of HFPV (Sections 2.3.3 and 5.1.3) plus the level of aspiration noise set to the NSR level measured in the original voice before any demodulation (this represents the worst case of additive noise). The FM analysis was then carried out on these synthetic voices with both aspiration noise and HFPV. The result is shown in Fig. 5.4, which plots measured HFPV in the synthetic voices with aspiration noise versus the level of HFPV in the synthetic voices without aspiration noise (Sections 2.5 and 5.1.6). As can be seen, there is an increment in HFPV of about 0.2%, which was near the limit of perception.

5.2.4 Effect of HFPV on AN

Similarly, the effect of HFPV on measured aspiration noise is addressed. The increment in measured NSR due to the addition of HFPV at the level measured in the original voice was evaluated. Starting with synthetic voices with aspiration noise only (Section 5.1.6), HFPV was added and the resulting NSR measured. The result is displayed in Fig. 5.5, which plots the cepstral NSR of synthetic voices with HFPV versus those without. The result appears to be about a 4 dB increment in NSR, which seems consistent with the result of Fig. 2.32.

5.2.5 SABS for Aspiration Noise

Pilot perceptual experiments were conducted comparing original voice samples with synthetic vowels. The effect of FM demodulation on the accuracy of NSR measurement was demonstrated. Listeners (who were demonstrated the effects of NSR parameter variation) attempted to match synthetic samples to the original ones by varying the synthetic aspiration noise level. The synthetic HFPV was turned off for this test. The results are displayed in Figs. 5.6, 5.7, and 5.8 which plot the mean level of aspiration noise listeners chose to match the perceptual effect of the original samples versus the original measured cepstral NSR. Fig. 5.6 displays the result for the original voice. Fig. 5.7 displays the result for the cepstral NSR measurement on the voices with tremor removed. Fig. 5.8 displays the result for the voices with both AM and all FM removed. There is a good indication of correlation with the original voice (Pearson = 0.51). However, the correlation increases when tremor is removed (Pearson = .71), and then increases again when all AM and FM is removed (Pearson = 0.87). In addition, the best-fit line moves from as much as 10 dB off (from perfect correlation) in the case of the original voice, to within 2 dB in the case with all AM and FM removed. Thus, the major disagreement between cepstral measured NSR and listener-set aspiration level is accounted for by FM modulation

5.2.6 SABS for HFPV

In a same manner as with aspiration noise, SABS pilot tests were conducted to vary HFPV. With the level of aspiration noise (which proved to be more perceptually distinguishable than HFPV for the 31 voices) first set for best match to the original,

listeners adjusted the level of HFPV to improve the match to the original. In most cases, it proved more difficult to set HFPV when compared to aspiration noise. The results are displayed in Fig. 5.9, which plots the mean of HFPV set on the synthesizer to match the original sample versus measured HFPV in the original voice. The level of correlation (Pearson coefficient = 0.403) is lower than that of aspiration noise.

5.3 Summary

This Chapter described the efforts for re-synthesis of pathological vowels. The algorithms for implementing synthesis of model parameters derived in analysis defined in Chapter 2 (LF source parameters, formants, aspiration noise, etc.) have been described. Validity of the overall analysis/synthesis process was tested by closing the loop with re-analysis of synthesizer outputs and with listener comparisons of original and synthetic vowels. Key findings include the fact that AM and FM demodulation improves the agreement between measured levels of aspiration noise and levels set by listeners in SABS (subjective analysis by synthesis) tests. The effect of AM demodulation was much less than FM demodulation. Tests showed less correlation between measured and listener-set HFPV levels in SABS tests than was observed for aspiration noise.