Chapter 2

Analysis of Pathological Voices

The study of pathological voices begins with the digitization and analysis of sample tokens. In this work, the analysis was performed on subjects vocalizing the sustained vowel /a/. Data used in this dissertation include a set of thirty-one pathological voices (randomly selected) representing a range of disorders, age, and gender. Additional voice data of normal voices simulating pathological conditions were also collected. Actual or simulated pathological voice data were then analyzed into the parameters described in Sections 2.1 – 2.5, which were found upon synthesis to provide in most cases a good basis for reproduction of the original sound. All analysis was performed offline using custom software specifically for the analysis of pathological voices and written in C and MATLAB (a commercial software package distributed by MathWorks TM). The steps of data analysis and synthesis are summarized in Fig. 2.1 and discussed in detail in the sections that follow. Briefly, the voice time series is first analyzed for formants (vocal tract resonances), which are then used to inverse filter the time series to obtain the source glottal flow derivative time series (driving waveform in the source-filter model). The source waveform is then fitted to a predefined waveshape: the LF model. Finally, the nonperiodic features of the time series (nonperiodic FM variations, nonperiodic AM variations, and aspiration noise) are quantified.

Processing begins with collection of voice data. The voices covered a range of disorders and age of both sexes, but excluded bicyclic cases. The pathological voice samples used in this study were collected at 20 kHz by Dr. B. Gerratt at UCLA using a one-inch B&K (Bruel & Kjaer) condenser microphone with a cathode follower preamplifier and antialiasing filter. This setup yields a flat frequency response from about 10 Hz to 20 kHz (the microphone/preamplifier proved not responsive to DC); it was used to record 1 second samples of patients vocalizing the sustained vowel /a/. These samples were then low-pass filtered offline with a FIR filter and decimated to 10 kHz before analysis. Additional normal and simulated pathological voice data were collected by the author for the study of vocal tract identification (Section 3.3) using a professional 0.8 cm condenser microphone element. These samples were digitized at 40 kHz using a sigma-delta A/D (which incorporates the antialiasing function).

The current five-component voice model provides the basis for analysis and synthesis of pathological voices. It includes the following components:

1. Formants. The basic source/filter model is used, which assumes an independent glottal source signal that is filtered with an all-pole resonator model of the vocal tract

2. Source waveform. The glottal source signal, which is obtained in analysis via inverse filtering, is fitted to the LF form [31].

3. Source nonperiodic FM (frequency modulation). Precision fundamental frequency tracking is used to establish both high and low frequency FM effects, which are then included in the synthetic version.

4. Source nonperiodic AM (amplitude modulation). Fundamental frequency pulse energy analysis establishes AM effects, which are then included in the synthetic version.

5. Source aspiration noise. Spectrally-shaped Gaussian white noise is measured in the original voice via cepstral filtering [24] and is added to the synthetic source time series.

In the following (Sections 2.1 – 2.6) the analysis of each of these components is discussed in detail.

2.1 Formant Analysis: LP Analysis and Inverse Filtering

The first step in the analysis of a pathological voice is the determination of formants via the established technique of LP (linear prediction analysis [28]) and application via inverse filtering to determine an estimate of the raw glottal flow derivative time series. This process attempts to perform system identification of both the source time series signal and the vocal tract model system in the source-filter model of speech production by merely using the final output time series of speech; the source/filter model is demonstrated in Fig. 1.2. Mathematically, the source signal and vocal tract system are just two transfer functions, and once they have been multiplied together to form the output signal, there is no way to segregate them again without additional knowledge. The process of identifying both glottal source time series and vocal tract system is ambitious, and it involves considerable intervention of apriori knowledge and manipulations beyond basic LP. The process has been fairly successful for normal voices, but encounters additional difficulties in the case of pathological voices. In some pathological voices, the method proved ineffective in establishing proper vowel quality, as shown by re-synthesis and perceptual testing, necessitating (sometimes unsuccessful) manual attempts to adjust formants.

The following briefly describes the basic process of LP. The results of LP/inverse filtering are then illustrated for idealized impulsive sources; ideal results are then contrasted with the results of application to more realistic LF (Section 2.2.2) sources. Finally, the apriori strategies and techniques attempting to improve the system identification are outlined. These techniques were collected or established by N. Antonanzas-Barroso and encapsulated by her in the computer program “invf” for source/vocal tract identification [1] at the UCLA Voicelab. This program was used to identify formants and raw glottal flow derivative in the first processing steps in this study (Fig. 2.1).

2.1.1. Basic LP/Inverse Filtering

Briefly, LP performs as follows; see [28] for details. Refer to Fig.2.2; in the figure and the following description, the notation of [28] is used. The basic process of LP attempts to form a P-th order linear FIR (finite impulse response) predictor that takes as its input the previous N samples of system output. That is, given the N previous samples s(n) of a system H, it attempts to predict the next sample. As shown in Fig.2.2, the predictor output s’(n) is subtracted from the system output s(n) to form an error signal e(n). The predictor (alpha’s) is derived by minimizing the summed error squared of e(n) over the selected analysis interval; this least squares minimization yields a set of P simultaneous linear equations in the P alpha’s which is easily solved. If the unknown system is assumed to consist of an all-pole IIR (infinite impulse response) system, as shown in Fig. 2.2, and the a’s of the unknown correspond to the alpha’s of the predictor, then the system formed by the predictor and subtractor form a new system A(z) that estimates the inverse of H(z). Ideally, then, it is possible to determine the predictor using LP, form the inverse system A(z) (which contains the formant estimates), and apply it to the original time series s(n) to inverse filter the voice and generate an estimate of the driving function u(n) which corresponds in our case to the glottal flow derivative (compare Fig. 2.2 to Fig.1.2).

Two versions of LP are applied in practice: autocorrelation and covariance. The versions differ only in the assumption made about samples outside the analysis window (range of samples used to calculate the alpha’s). The autocorrelation method assumes samples outside the window are zero; it gives rise to a Toeplitz system [28] which is efficiently solved via Durbin’s algorithm. The covariance method does not assume zero values outside the analysis window, and gives rise to a more computationally complex system [28]. In practice, the autocorrelation method is applied to a range of several fundamental frequency periods of voice signal, and a first estimate of A(z) is obtained. The window of analysis is then restricted to what is believed to be the “closed phase” (assumed quiet period of the source), and a more accurate estimate of A(z) is attempted via the covariance method; the assumption here is that the interfering activity of the source dynamics is removed, which permits a better vocal tract estimate. In practice, however, the less impulse-like the source (the greater the portion of the fundamental frequency period that the source has nonzero activity), the less successful LP is in vocal tract identification. The technique is unable to segregate source from vocal tract system using only their convolved result, the output time series s(n). (See also the discussion in Section 3.1).

2.1.2 Idealized LP: Impulsive Sources

LP in both its variants accurately identifies the vocal tract if the source time series is a simple impulse train and the vocal tract transfer function is all-pole. That is, as expected, if the source has no dynamics (z or Laplace transform is unity), LP has no difficulty in assigning all detected dynamics to the only remaining part of the system: the vocal tract. That is, in the case of impulsive sources, there is no ambiguity between the source signal and the vocal tract system, and LP correctly identifies the vocal tract system. The performance of LP on impulsive sources is illustrated in Fig. 2.3 to Fig. 2.5.

A simple vocal tract with four resonators for the vowel /a/ is shown in Fig. 2.3. The impulse train source time series is shown at the top, the pole locations (roots of the resonators) in the middle plot, and the resulting voice output time series is shown at the bottom. For this ideal system, applying the inverse filter, which is simply the FIR filter formed from the reciprocal of H(z) vocal tract, successfully reproduces the impulse train input exactly.

Fig. 2.4 illustrates the application of autocorrelation LP to the impulse train system of Fig. 2.3. Using 10 fundamental frequency periods of the output of this system windowed with a Hamming window yields the pole position shown in the top plot; there is very good agreement with the true poles. Also, the prediction FIR filter, when applied to the output time series s(n), yields an accurate match s’(n) to the original signal, as shown in the middle plot. When the resulting inverse filter A(z) is applied to the output s(n), an accurate approximation to the input impulse train is obtained, as shown in the bottom plot. The application of the covariance method to the impulse train system also produces good results, as shown in Fig. 2.5. Accurate poles, prediction, and inverse filtering are obtained. Here a window length of one fundamental frequency period was used.

2.1.3 Non-impulsive Source Functions

When more realistic glottal source waveforms are employed, however, the results change drastically. The LP now has the impossible task of trying to separate the convolved source signal and vocal tract system. To illustrate, the system of Fig. 2.6 is constructed. This system is analogous to Fig. 2.3 except the impulse train input is now replaced with a realistic LF waveform (Section 2.2.2); the same vocal tract is used, and the output appears somewhat similar. The results of application of autocorrelation LP are shown in Fig. 2.7; the same sample window and position are used as in Fig. 2.4. As may be seen, however, the pole position now contains significant errors: the higher frequency poles become progressively less accurate, and a pair of real poles replaces one of the complex pairs. The inverse filtered signal now contains large ripples due to incomplete pole-zero cancellation. Only the prediction signal remains accurate.

Application of covariance LP alone results in little improvement. The results of covariance analysis are summarized in Fig. 2.8. In an effort to provide time separation of signal and system (see discussion in Section 3.1), the analysis window is shortened to 40 samples, which is about the length of time during which the source signal is zero (closed phase). The analysis window is then swept through all 100 possible positions and the locus, resulting from plotting the poles at each window position, is shown. As may be seen, the true pole positions are never achieved, and error is more pronounced at the higher frequencies; a pair of real roots often replaces a complex pair. The results of covariance LP using one of the more optimal window positions is shown in Fig.2.9 (compare to Fig.2.7). Again, poor pole accuracy results in formant ripple in the inverse filtered signal, which vaguely resembles the actual LF input; at least two of the uncanceled pole frequencies are clearly visible in the error signal. Prediction remains accurate.

2.1.4 Apriori Strategies

Several techniques have proven useful to improve the performance of LP, thus enabling more accurate separation of the glottal source signal. These methods allow the user to add known characteristics of vocal system to the simple least squares algorithm of LP.

1. Real roots are discarded. The vocal tract is assumed to consist of an all pole resonator model with no real poles.

2. Complex poles resulting from the covariance method occurring outside the unit circle are reflected back inside the unit circle by taking their reciprocal. It is known that the covariance method is not guaranteed to produce stable roots [25], while the vocal tract is a stable, passive system. The success of this strategy is shown in Fig. 2.10, which repeats the system of Fig.2.9. In the top plot the poles for a window position in the closed phase are shown; they all lie outside the unit circle. Reflecting them inside the unit circle places them into almost exact agreement with the true poles. The resulting inverse filtered signal (bottom) agrees well with the LF input, and prediction (after re-scaling) remains fairly good.

3. The length of the covariance analysis window can be varied in an effort to match a closed phase. Many pathological voices contain few if any quiet periods, however. The location of the closed phase (or closest approximation to a closed phase in the case of some pathological vowels) is chosen at the maxima of the residual.

4. The two poles per kilohertz rule is used to start analysis, but order of the system can then be optimized.

5. Provision for manual pole placement is made in the analysis software. Sometimes formant ripple in the inverse filtered signal can be reduced by manually tweaking pole positions. However, because of the high number of degrees of freedom (2 dimensions for each pole), this process is difficult to converge. Knowledge of the spectral tilts of typical voice signals can also give clues to improved pole placement.

6. A final resort is attempting to place formants by SABS (subjective analysis by synthesis). One of the synthesizers (Chapter 4) is used to compare the vowel quality of original and synthetic voices. However, experience shows this process is even more difficult than #3 above.

Formant analysis for pathological voices provides added problems over normal voices, invalidating assumptions made in normal LP. The source waveform may not be impulsive and lack quiet periods, making it impossible to apply covariance to separate signal and system. The source may be highly irregular from pulse to pulse, with the period and shape of each source pulse varying. For these and other reasons, application of LP to pathological voices is sometimes very difficult. External stimulation of the vocal tract (Chapter 3) may provide a means to faster and more accurate formant identification.

2.2. Fitting of Inverse Filtered Source

Having generated the raw inverse filtered source waveform, the next step in voice analysis is fitting the source to a mathematical model with adjustable parameters. The parameter values provide multiple benefits: They provide a means of comparing and rating voices and studying linguistic and voice quality effects. Furthermore, the parameter values are later used as input control values to both the hardware and software synthesizers. Several types of source models were investigated: filtered pulses, parabolic, and the LF model. The LF model had a greater ability to model the wider range of wave shapes found in pathological vowels. The hardware synthesizer (Section 4.1) was used to generate synthetic tokens of the vowel /a/ using each of these source models. Ultimately, a modified version of the LF model proved to be the most useful for analysis and synthesis of pathological vowels and was implemented in the software synthesizer (Section 4.2).

2.2.1 Simple Models

A variety of mathematical models have been used to fit the inverse filtered glottal flow derivative pulse [22,31] to match the wide range of waveforms found in pathological voices. These include simple impulses, impulses filtered by a variety of lead-lag filters, and polynomial fits. Simpler models have been shown to have shortcomings. For example, a two pole filtered impulse (which incorporates just two degrees of freedom) is shown by Deller [7] to have limitations in achieving the OQ (open quotient) vs SQ (speed quotient) relationship found in the source waveforms of even normal voices. Chasaide and Gobl [4] also show filtered impulses to have limitations: Their shape is time-reversed with respect to the normal pulse. That is, the large rate of change is on the rising edge rather than the falling edge.

One of the more useful simple models used in the early synthesis efforts with the hardware real-time synthesizer (Section 4.1) is the KGLOTT88 parabolic model [22]. The equations describing this model are shown in Fig. 2.11, and an example plot of a KGLOTT88 glottal flow pulse and its derivative are shown in Fig. 2.12. This model significantly improved perceptual fidelity over simpler filtered impulses.

2.2.2 LF Model Characteristics

Our earlier studies with a real-time synthesizer employing a simple parabolic source model showed considerable improvement in perceived fidelity when the simplified LF source model [31] was implemented [16]. The simplified LF model, including our modifications, is shown in Fig. 2.13. This model elaborates on the idea of using simple algebraic and trigonometric curves by employing an exponentially modified sinusoid for the first part (opening phase) of the glottal flow derivative pulse and a decaying exponential for the second part (closing phase). The result appears remarkably similar to actual flow derivative pulses of normal voices.

As a result of experience processing considerable amounts of voice data, we found it useful to use a set of physically observable parameters to define the pulse to be fitted rather than the actual LF parameters; the actual LF parameters can then be derived via the solution of simultaneous equations, in a manner similar to that used by Qi [31]. Our selection of observable parameters, which completely define the modified LF curve, results in a set of four values with an optional constraint.

tp = time of zero crossing

te = time of negative maxima

Ee = value of negative maxima

t2 = time of 50% on closing phase

m = linear coefficient of closing phase (OPTIONAL CONSTRAINT)

The first four values can be determined automatically from the time series. The optional fifth value *m* is calculated when the constraint is added that the glottal flow derivative pulse must return to zero. The original LF model contains the parameter *ta*, which is a measure of the return phase decay rate; it is the time interval from *te* to the intersection with the time axis of the tangent at *te*. The *ta* parameter in the original LF model is functionally replaced here with *t2*.

This model seems to allow considerable flexibility to adjust to the wider variety of pulse forms found in pathological voices. For example, for breathy voices, the glottal flow derivatives we found had almost sinusoidal form; by extending *tp* (time of zero crossing) to larger values the model waveshape becomes more sinusoidal.

2.2.3 Modified LF Model Calculation

In operation, calculation of the LF model proceeds as follows:

1. A typical glottal flow derivative pulse is selected from the raw inverse filtered time series (Fig. 2.14). For this study, the waveform of the glottal flow derivative is assumed to be fixed for the entire one-second voice sample. Since time varying source pulses are not currently modeled, the selected pulse is chosen to be a representative one if there is variation present.

2. The major features described in Section 2.2.2 (shown in Fig. 2.15) are automatically determined from the data. Numerical methods are applied to obtain maxima, zero crossings, etc from the raw inverse filtered glottal flow derivative time series. The equations shown in Fig. 2.16 are then used to obtain the initial set of LF parameter values. The LF curve obtained is plotted and visually checked against the original raw flow derivative pulse, as shown in Fig. 2.17. This step was almost invariably successful.

3. The major features determined in Section 2.2.2 are then used as the starting point for least squares optimization using the MATLAB function ‘fmins’ which uses the simplex search method [26]. Starting at a reasonable first approximation resulted in much quicker convergence and reduction in the chance of spurious solutions (local minima). The summed error squared between the flow derivative pulse and the fitted LF curve is minimized by varying the four major feature parameters *tp, Ee, te,* and *t2* to their optimal values. An optional constraint was provided to force the fitted flow derivative pulse to zero; this results in an added equation for the parameter *m* (see Fig.2.13). The optimized major feature parameters are then converted to the LF parameters using the equation set shown in Fig.2.18. The first three equations, unfortunately, form a simultaneous nonlinear

set. However, by iteratively solving them in the proper sequence shown, each equation is solved for its variable of greatest effect and the process rapidly and reliably converges. The steps of data collection through LF fitting complete the periodic analysis of the voice signal, and they are summarized in the first column of Fig. 2.1.

2.3 Analysis of Fundamental Frequency Variations

Fundamental frequency variation is one of the most perceptually important acoustic measures of voices. Unlike other measures such as aspiration noise, small variations in fundamental frequency are not ignored. Efforts to provide ever increasing levels of synthesizer fidelity led to a successful approach to modeling fundamental frequency variation in the pathological /a/ vowel samples collected.

2.3.1 Fundamental Frequency Analysis Approach

In order to parameterize fundamental frequency variation for accurate synthesis, fundamental frequency variations are tracked in the time domain via interpolation to the precision permitted by the data and then broken down into low frequency (tremor) and HFPV (high frequency fundamental frequency variations). Previous investigations have used a variety of measures of HFPV, usually some type of average [3]. The current approach refines HFPV analysis and synthesis by modeling the variation in the fundamental frequency period as a Gaussian distribution: the frequency deviation excursions of HFPV are seen to be well-modeled by a Gaussian distribution in cases of successful tracking on short intervals of sustained /a/. In order to verify successful fundamental frequency tracking and aid later NSR (noise to signal ratio) calculations (Section 2.5), the original pathological voice is re-sampled to remove fundamental frequency variations. In summary, the following steps are performed:

1. Precision interpolating fundamental frequency tracking in the time domain.

2. Segregation of the fundamental frequency track into high and low frequency variations.

3. Gaussian modeling of the statistical properties of the high frequency fundamental frequency variations.

4. FM Demodulation of the original voice using measured fundamental frequency variations.

2.3.2 High Resolution Fundamental Frequency Tracking

Time domain fundamental frequency tracking is carried out over the entire original voice sample on a pulse by pulse basis, so that the short duration pulse period variations (HFPV) may be captured. Pulse period lengths are established by measuring the time interval between similar features in each pulse, such as maxima or minima points. For additional information, see Milenkovic [27] which describes application of correlation techniques to determine HFPV. In addition, Deem, Manning, Knack, and

Matesich [6] describe a process for determining period lengths in the time domain. Fundamental frequency tracking allows use of any of four signals for maxima/minima detection for cycle marking: original voice, glottal source, smoothed derivative of original voice, or smoothed differentiated glottal source. In difficult cases in which automatic loses tracking lock (fails to track accurately), the user may manually mark features on as many pulses as necessary to reestablish tracking lock. Interpolation between samples is used to determine fundamental frequency periods to less than one sample period (< 0.1 ms): a parabolic fit with a user-selected number of points is applied to the sample points surrounding the maxima/minima, and the "true" minima/minima is calculated from the fitted parabolic vertex. For some parts of the analysis, upsampling from 10 kHz to 40 kHz improves performance of the tracking algorithm. Upsampling is performed using the MATLAB (a commercial software package distributed by MathWorks TM) function “interp,” which uses symmetric filtering and minimizes the mean square error of the interpolated points [20]. Successful tracking is characterized by the absence of discontinuities in the resulting fundamental frequency versus time plot and Gaussian distribution in fundamental frequency period variations with standard deviation in the expected range of 0 to 1 percent. Fig. 2.19 illustrates automatic pulse feature selection for a typical signal; in this case, the selection of original voice minima yielded successful automatic tracking. Fig. 2.20 illustrates the resulting typical high-resolution fundamental frequency time series generated from the interpolating tracker.

2.3.3 Analysis into HFPV and Tremor

In order to analyze the voice sample for FM characteristics, the fundamental frequency time series is examined for both its frequency content and statistical features. Two fairly distinct types of fundamental frequency 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. The selection of the 10 Hz cutoff is arbitrary, but seems to give satisfactory results when synthesized tokens are studied in perceptual tests. The tremor variations are perceptually associated with an unsteady voice, while HFPV gives rise to the perception of roughness. The high and low frequency components of the fundamental frequency track are respectively segregated into HFPV and tremor time histories using high and low pass filters with a cutoff frequency of 10 Hz respectively. High and low frequency components of the fundamental frequency track are separated in the frequency domain using the windowing technique on the Fourier transform, thus taking advantage of using the entire fundamental frequency time series offline (non-real-time). Fig. 2.21 displays the result for the fundamental frequency track of Fig. 2.20.

2.3.4 Gaussian Modeling of HFPV

Once tremor has been removed from the fundamental frequency track, the remaining higher frequency variations appear to be well modeled by a Gaussian distribution. To verify the success of fundamental frequency tracking, the statistical distribution of HFPV is displayed by histogramming the frequency deviation of the high pass filtered fundamental frequency time series. Fig. 2.22 displays the histogram for the same voice signal as Fig.2.20. Successful tracking is characterized by a Gaussian distribution with one standard deviation of usually less than 1 or 2 percent. Bimodal distributions and large values of deviation are almost always associated with loss of lock in the fundamental frequency tracking algorithm. The measured standard deviation of fundamental frequency period, in units of percent, is also used to define the level of HFPV for later input to the synthesizer.

2.3.5 FM Demodulation

In order to investigate the effects of frequency modulation on measurement of aspiration noise (Section 2.5), the original voice time series is resampled to create versions of the original voice with (a) low frequency (tremor) removed and (b) all frequency variations removed. This is a refinement of the approach used in [14]. Using the fundamental frequency time series, a vector of unevenly spaced re-sampling times is created by making the instantaneous sample interval inversely proportional to the instantaneous fundamental frequency frequency. Simple linear interpolation of the original time series on this modified time vector creates a version of the original voice where all fundamental frequency periods are forced to be the same length; that is, fundamental frequency variations are removed. To remove only low frequency fundamental frequency variation, the low-pass filtered fundamental frequency time series is used to generate the modified time vector; to remove all fundamental frequency variation, the unfiltered fundamental frequency time series is used. In order to verify the success of this FM demodulation, fundamental frequency tracking is re-performed on the re-sampled voices. Fig. 2.23 is an example of the effect of removing the tremor from the original voice. Note that the low frequency fundamental frequency variations in the re-sampled voice are greatly reduced. Fig. 2.24 shows the results of repeating fundamental frequency tracking on a successful re-sampling of the original voice to remove all fundamental frequency variation. The upper plot shows that fundamental frequency period variation has been reduced to less than 0.2 Hz, and the frequency is essentially constant at 266.4 Hz, as shown by the points clustered about this frequency. The lower plot shows that HFPV is less than 0.05%. Success of the process is also verified by listening to the re-sampled time series. Tremor removal creates a much steadier sounding token. Removal of all FM creates a token sounding almost synthetic in its stability. (Interestingly, however, removal of all fundamental frequency variation in some cases still leaves voice quality variations synchronous with the original tremor variations; these may be due to formant modulation or other effects.)

2.4 Analysis of Fundamental Frequency Pulse Power Variation

In a manner analogous to fundamental frequency variation, power variations in the fundamental frequency pulses of the vowel /a/ are analyzed. This approach uses a similar treatment of both the AM and FM variations.

2.4.1 Power Analysis Approach

Power analysis seeks to quantify the variations in amplitude, power, and energy within the pathological vowel sample. A rationale for measurement is constructed: the boundaries of fundamental frequency pulses are located with the aid of the results from fundamental frequency tracking (Section 2.3.2). Having defined the pulses, it is easy to calculate various measures of amplitude, energy, and power. Power is selected for subsequent analysis, since it is probably highly correlated with perception. As with fundamental frequency, power variations are segregated into high and low frequency phenomena. The high frequency variations (shimmer) are found to be Gaussian. Removal of AM variations from the original voice provides verification of processing and permits testing the effects of AM on NSR (Section 2.5.5).

2.4.2 Power Tracking

Analysis of the original pathological voice continues after fundamental frequency tracking with analysis of the amplitude, energy, and power of fundamental frequency pulses. The results of fundamental frequency tracking are used as a starting point for estimating the maximum amplitude, sum of the absolute value of samples, energy, and power of each fundamental frequency pulse.

The first step of amplitude analysis is segregation of the fundamental frequency pulses of the original time waveform. The set of features within each cycle described in fundamental frequency tracking, such as pulse maxima, is assumed to exist between minima of each fundamental frequency pulse. The adjacent minima of the signal's amplitude envelope to either side of this center are used to define the boundaries of the fundamental frequency pulse. The set of features within each cycle described in Section 2.3.2, such as pulse maxima, are assumed to exist between minima of each fundamental frequency pulse. The adjacent minima of the signal's amplitude envelope to either side of the tracking feature are used to define the boundaries of the fundamental frequency pulse. The intent is to separate pulses by placing pulse boundaries so that the maxima of voice power in each pulse occurs between tracking features of the pulse and the minima of power occurs near the boundaries. This provides a natural separation of normal voice pulses and a reasonable approximation for pathological voices. Minor variations in the selection of pulse boundaries have little effect on the remainder of the power analysis.

Analysis proceeds via generation of the envelope of absolute value and power of the original voice. Starting at the features used for fundamental frequency tracking, the corresponding power envelope minima before and after the feature are located. For each fundamental frequency track feature there exists a power minima, so the resulting power minima instants (time values at the minima) are interspersed between the fundamental frequency tracking feature instants (time values at the features). The envelope minima thus determined form a natural boundary for fundamental frequency pulses. Fig. 2.25 displays the absolute value of a short segment of a voice time series showing the fundamental frequency track maxima and the pulse boundaries (envelope minima) selected by the algorithm. In practice, power tracking proved to be a far easier task than fundamental frequency tracking; no manual interventions or alternative approaches were ever required

Having defined pulse boundaries, four measures of pulse strength are calculated: maximum amplitude, signal average, indicated by sum of absolute value, energy, and power. The maximum amplitude is the absolute value of the greatest extent (plus or minus) of the original voice samples within the pulse (between the pulse demarcations.) The sum of the absolute value is the addition of the absolute value of all the samples within the pulse. The energy is the sum of the squares of all samples in a pulse. The power is the energy divided by the number of samples within the pulse. Generally, all four measures track each other. Power was selected as the most useful, as it is probably most closely correlated with perceived signal strength. Fig.2.26 displays power for the same signal shown in Fig. 2.25.

2.4.3 Power Time Series Analysis

In a manner analogous to the fundamental frequency time series analysis, the original voice amplitude (AM) features are analyzed. Pulse power variations are now analyzed into a low frequency power time series track and a high frequency shimmer measure. The power measure described in Section 2.4.2 is selected as the basis of this analysis over the other three measures, since it should be most closely related to the perceived signal level. A cutoff frequency is selected (usually 10 Hz), and a low pass FIR filter is constructed and applied to the power time series. The resulting low passed signal defines the low frequency power time series. The difference between the original power time series and the low frequency power time series defines the shimmer time series; the standard deviation of the shimmer time series estimates the amount of shimmer present. Fig. 2.27 illustrates the high and low frequency power variations. In the upper plot of Fig. 2.27, the low pass filtered power time series is shown as a dotted line; the original power time series, deviating above and below the low pass filtered signal, is plotted on top with a solid line. The original time series power closely follows the low pass filtered version, differing only by small high frequency variations. In the lower part of Fig. 2.27, the difference between the high and low pass filtered signals is plotted; it is essentially equal to the high pass filtered signal. Specific hallmark peaks in Fig. 2.27 illustrate this; the negative-going shimmer peaks in the top of Fig. 2.27 appearing at about 0.12 and 0.32 sec also appear directly below in the lower plot. A positive peak at about 0.34 sec also appears in both plots.

2.4.4 Gaussian Modeling of Shimmer

In the same manner as HFPV, the high frequency power variations are histogrammed, as shown in Fig.2.28. Again, a Gaussian distribution seems to model the high frequency power variations for our quasi-steady vowels. The standard deviation of the distribution, measured in percentage power variation, provides a measure of shimmer. Power variations were usually larger than HFPV, typically a few percent. Unlike HFPV, bimodal or other unusual distributions signifying problems with power tracking were seldom seen, as power tracking was a much more robust process. Again, measured shimmer values were passed on to the synthesizer software for use in perceptual studies.

2.4.5 AM Demodulation

In a manner analogous to FM demodulation, analysis continues with AM demodulation. The AM variations may be minimized in the original voice by rescaling the samples in each fundamental frequency pulse in the time series such that all fundamental frequency pulses exhibit the same power level. Using the data collected in AM analysis, a scaling factor is calculated for each pulse to normalize its power level to the mean power level measured over the entire original voice one second sample. The resulting demodulated time series exhibits constant power over the 1-second sample. The modified time series is resubmitted to the AM analysis software to verify successful demodulation. Fig. 2.29 displays the result of a successful demodulation; note that the measured power level of the demodulated voice is constant. The other measures (sum, maxima, or energy) are not necessarily held constant, but are approximately constant. Informal listening by the authors of the AM demodulated waveform verifies that apparent changes in perceived volume level have been removed. The analysis steps of AM and FM demodulation complete the first stage of the nonperiodic analysis, and are represented in the top center area of Fig. 2.1.

2.5 Analysis of Source Aspiration Noise

Having accounted for two known sources of nonperiodicity in the original voice, analysis continues with estimation of the ratio of nonperiodic energy in the original voice to periodic energy (NSR). Another source expected to account for the major part of the remaining nonperiodic energy in most pathological voices is aspiration noise generated by turbulent airflow through the glottis. (Possible additional remaining sources not yet measured are variation in the glottal waveform and variation in formants.) As with the flow derivative, this aspiration noise is spectrally shaped by the vocal tract. Both the ratio of nonperiodic to periodic energy (NSR) and the spectral shape the aspiration noise are approximated.

2.5.1 Noise Measurement Approach

Measurement of the aspiration noise component of a voice is performed using a modification of the cepstral notch filtering technique described in [24]. The original de Krom algorithm generated SNR (signal-to-noise ratio) in voices as a function of frequency; cepstral filtering and spectral peak envelope tracking were used to arrive at noise spectra and SNR functions The assumption in this work is that the nonperiodic energy measured and reflected in the NSR is due mainly to aspiration noise. A limitation of this approach is the tradeoff between the inaccuracies due to the windowing effects of a short time sample and the instabilities (which include AM and FM) of the voice in a long time sample. Both of these effects tend to increase NSR to levels significantly above those set in SABS tests in which the listeners match synthetic voice samples to the original by varying the parameters of the synthesizer. The current approach removes AM and FM variations (Sections 2.4.5 and 2.3.5) before the cepstral noise measurement, thus allowing long samples to reduce windowing effects but maintaining stability of the voice over the longer sampling interval. This results in more accurate aspiration noise level estimates which are much closer to those set in the SABS tests. The effects of removal of FM and AM of both high and low frequency were tested individually and in combinations. Changes in measured NSR of as much as 12 dB were observed. The effect of removal of FM components was observed to be much greater than that of AM components. Effectively, the process of removal of AM and FM components from the natural voice helps to bridge the gap between natural and synthetic voices in a stepwise approach. Removal of each component renders a result that sounds progressively more like the synthetic version, and it aids in the identification of any unaccounted remaining differences between the original and synthetic voices.

As part of the cepstral noise analysis [24], the power spectrum of the glottal source aspiration noise component is estimated. After removal of the periodic signal from the voice sample, a spectrum of the non-periodic components is produced, which is then inverse-filtered to remove the vocal tract effects, yielding the source nonperiodic spectrum alone. The shape of this spectrum is usually flat (without significant systematic variation of power with frequency) to within a few dB; a spectrum that is not flat would indicate aspiration noise with more complex frequency content. The spectrum is captured in a 25 point piecewise linear approximation.

2.5.2 Cepstral NSR Calculation

The relative amounts of nonperiodic and periodic energy in the original voice are estimated using the basic cepstral comb-liftering approach outlined in [24]. The following steps refer to Fig. 2.30, which summarize successive steps in the NSR analysis for one typical pathological voice (male, fundamental frequency variation from 118 Hz to 124 Hz, NSR at –24dB). First, the spectrum of the original voice is generated, as shown in Fig. 2.30a; the periodic components appear as the usual harmonic peaks. The spectrum is transformed to the cepstrum, as shown in Fig. 2.30b; now the periodic components from the spectrum are further condensed into the peaks of the cepstrum (eg., the large peak at 0.008 sec in Fig. 2.30c). Fig. 2.30c provides an expanded view of the area with highest periodic energy, at low cepstral time range. Using autocorrelation, the fundamental frequency is estimated and used to construct a periodic notch filter (comb-lifter) to apply to the cepstrum to remove the periodic components. Application of the comb-lifter to the cepstrum removes the periodic energy in the peaks, as shown in Fig. 2.30d, which shows its application to the cepstrum of Fig. 2.30c. (Note: the width of the cepstral notch seemed to have minimal effect on NSR measurement, so long as it is wide enough to encompass the peak, since the spectral density of the noise component is much lower.) After removing the periodic components, transformation back to the spectrum yields the spectrum of the nonperiodic components alone. Fig. 2.30e re-displays the original spectrum and the resulting noise (nonperiodic) spectrum, which appears immediately below and partly overlapped with the original spectrum. Also shown in Fig. 2.30e is the magnitude plot of the vocal tract determined independently from the LPC analysis. (Successful analysis of stable voices is reflected in matching shapes in the spectra, as all three should contain the vocal tract shaping). At this point, Parseval's rule (which simply states that summed signal energy is the same in both the frequency and time domains) is applied to calculate NSR. The energy in the noise spectrum is simply summed and subtracted from the original spectrum energy to yield periodic energy, and hence the ratio of nonperiodic to periodic energy in the original time series.

2.5.3 Source Noise Spectrum

Finally, an estimate of the spectral shape of the source noise alone is calculated. In order to generate the spectrum of the source alone, the vocal tract log magnitude frequency response is generated from the all-pole filter used to model the vocal tract (generated in inverse filtering) and subtracted from the nonperiodic spectrum. The resulting source nonperiodic 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 model. In most cases, the resulting spectrum is fairly flat. The final source spectrum and its fit are shown in Fig. 2.30f.

2.5.4 Effect of FM on NSR Measurement

The cepstral filtering noise analysis [24] implicitly assumes the signal being analyzed is stable in frequency. If there is variation in the fundamental frequency over the sample analyzed, the harmonic peaks of the spectrum and cepstrum will be broadened significantly compared to the width of the notch filter used to remove them. For example, if a 1ms cepstral notch width is used with a voice with fundamental frequency of 100 Hz, it only requires a 5% change in fundamental frequency to completely displace the rahmonic peak out of its intended notch, thus completely corrupting the NSR measurement. Smaller fundamental frequency changes would still significantly affect the NSR calculation.

The setting of the notch width of the comb lifter is a tradeoff. Wider notches remove “rahmonic peaks” that have been widened by both the variations in fundamental frequency over the sample and the windowing effect of short duration time samples. However, wider notches also remove a greater portion of the background nonperiodic energy underlying the notch, giving rise to lower than actual estimate of nonperiodic energy. Narrow notches fail to completely remove rahmonic peaks widened by the fundamental frequency variation and windowing, resulting in periodic energy being included in the nonperiodic estimate and thus giving rise to a higher than actual estimate of nonperiodic energy.

An analogous tradeoff exists with sample duration. Short samples cause widened cepstral peaks. Longer samples alleviate the effect of windowing, but peaks may widen due to fundamental frequency variation (tremor and HFPV) over the duration of the sample. The ideal solution is a long, fundamental frequency stable sample, allowing narrow cepstral notch filtering and producing a more accurate NSR estimate. Thus, using the FM demodulated original voice (Section 2.3.5), it should be possible to improve the accuracy of the NSR estimate and hence the agreement between the SABS listener-set and the measured aspiration noise levels. This turns out to be the case. Fig. 2.31 shows the relative spectral shapes of original (top curve) and demodulated signals, and it illustrates the effect of resampling. The three curves are offset vertically by an arbitrary amount for clarity. The peaks associated with harmonic energy are narrowed by tremor removal (middle curve); removal of all FM narrows the peaks further (bottom curve). The resulting reduction in spillover from the cepstral notch filter reduces the NSR measurement from the original –8.5 dB, to –13.9 dB with tremor removed, and further to –18.3 dB with all FM removed. The improved NSR estimate more closely agrees with the listener-set aspiration noise level of –22dB for the synthetic version in the SABS experiments (Section 5.2).

2.5.5 Effect of AM on NSR Measurement

In order to test the effect of removal of AM variations on the NSR estimate, the AM demodulated original voice was also re-analyzed. Removal of AM variations does not seem to affect harmonic peak widths and NSR estimates nearly as much as removal of FM variations. In most cases, the effect is under 1 dB. (Section 2.5.6). In general, amplitude modulation with a random signal at the low levels characterized by volume and shimmer variation in a pathological voice would not be expected to vary harmonic peak width nearly as much as frequency modulation.

2.5.6 NSR Measurement Results

In order to test the affects of various combinations of tremor, HFPV, and AM demodulation, six versions of the original voice are created:

1. Unmodified original voice.
2. Original voice with AM demodulation
3. Original voice with FM tremor demodulation
4. Original voice with FM tremor and AM demodulation
5. Original voice with complete FM demodulation
6. Original voice with complete FM and AM demodulation

The NSR value, (here assumed to be entirely due to aspiration noise) is calculated for each version of 31 pathological voices. The result is plotted in Fig. 2.32. In the figure, the cepstral NSR values for each of the six versions are plotted versus case number, with the case numbers sorted in order of ascending NSR of the unmodified original voice. The

NSR values of each version are connected with line segments for ease of comparison. Several effects clearly emerge:

1. The process with the largest and most consistent effect is tremor removal, with cepstral NSR reductions in the 5 to 10 dB range. (Note the decrease from the top curve pair to the middle curve pair in Fig. 2.32).

2. Removal of HFPV provides an additional decrease in cepstral NSR of 1 dB to 5 dB, with an average of about 2dB. (Note the decrease from the middle curve pair to the bottom curve pair in Fig. 2.32).

2. The incremental decrease in cepstral NSR due to AM demodulation is very small (usually less than 1 dB). (Note the small difference between each of the two members within each of the three main curve pairs).

2.6 Summary

The steps of the analysis phase of voice processing have been detailed. Using the source-filter model of voice production as a basis, voices have been parameterized into formants, LF source waveform, fundamental frequency time history, amplitude time history, FM nonperiodic effects, AM nonperiodic effects, and aspiration noise. The limitations and practical aspects of LP analysis for inverse filtering and the determination of formants have been described. Source waveform fitting using the LF model has been detailed. The nonperiodic features of pathological voices have been expressed in terms of AM and FM variations and aspiration noise. Gaussian distributions have been shown to model the small, high frequency effects in AM and FM variations. Aspiration noise has been found to be well modeled with spectrally shaped Gaussian noise.

The results of analysis form a set of parameters that may be used both for comparison of voices and for the generation of synthetic versions for perceptual testing. The following Chapters treat the subject of synthesis of pathological voices.