Chapter 3

External Source Identification of the Vocal Tract

The technique proposed in Section 2.1 to identify the vocal tract and source characteristic relies on a combination of LP analysis, a set of empirical rules, and apriori knowledge of expected parameters for the vocal tract and source waveforms for normal voices. In the case of normal voices with the expected vocal tract formants and glottal source waveforms approximating the LF shape, the rules and apriori assumptions apply and the current technique works fairly well. However, in the case of pathological voices, difficulties have been experienced. In many cases, the automatic LP analysis and inverse filtering algorithms produce parameters which, upon re-synthesis, may result in inaccurate vowel quality. Even efforts to manually tweak these formant frequencies and bandwidths in the synthesizer to match the original voice vowel quality sometimes prove very difficult for expert lab workers.

This may not be surprising when the nature of the source/vocal tract model is considered. The source and vocal tract are represented as two blocks in the usual system diagram of Fig. 1.2. The glottal source is thought of as a time series and the vocal tract as a filter in the frequency domain. The source time series is filtered by the vocal tract, yielding the final voice. Mathematically, however, the source and vocal tract are both simply two indistinguishable signals or systems that are convolved in time. Thus, a perturbation in either the glottal source or the vocal tract could be compensated for by a perturbation of the opposite effect in the other and still produce the same resulting voice waveform. Thus, more than one seemingly “reasonable” source waveform could be combined with different vocal tracts to produce the same result. The same ambiguity problem manifests itself in the inverse (system identification) problem as difficulty achieving proper formant values. Fig.3.1 illustrates a specific example. The convolution of source and vocal tract to yield a voice is shown for three cases.; results from actual computation are shown. In the first case a normal voice /a/ is shown; here the typical LF source [12] is convolved with a typical vocal tract impulse response to yield a typical /a/. In the second case, the source has greatly reduced high frequency content, typical of a “breathy” /a/, but the vocal tract is normal; this gives rise to a vowel low in high frequency content. Even though the normal vocal tract formants are present, they are absent from the voice spectrum, as shown in Fig. 3.1, second row. In the third case, a normal LF [12] source is convolved with a vocal tract with almost no discernable formants. The second and third cases, however, have exactly the same resulting voice. The process of LP formant analysis and inverse filtering attempts to reconstruct the

source time series and vocal tract transfer function from the resulting voice (output) time series alone; in this case it is clearly impossible. Although this example is an exaggerated case, it nevertheless demonstrates the problem of source-vocal tract ambiguity.

One possible solution to the ambiguity is to apply an external stimulus of KNOWN form to the vocal tract and analyze the resulting response. The vocal tract could then, in theory, be identified unambiguously. Once the vocal tract has been properly identified, the voice can be inverse filtered to yield the actual glottal source waveform, which may or may not be the shape expected. In fact, external source identification of the vocal tract has already been applied in various approaches and levels of sophistication for about the last 60 years. Please refer to Section 1.1 for a summary of earlier works.

Given the success of previous investigators, the possibility of applying external source identification to aid in vocal tract identification for pathological voices seemed promising. A demonstration of external source excitation incorporating features from some of the previous efforts was successfully implemented and tested here to illustrate its application to vocal tract identification, which could then aid unambiguous source time series identification for the case of pathological voices.

3.1 Overview of External Stimulation

A simple experimental setup for demonstrating proof of principle of external source stimulation of the vocal tract is shown in Fig.3.2. The system consists of two IBM

PC compatible computers, stimulus amplifier, response amplifier, attenuater, acoustic ducting, microphone, and mounting jig. In operation, the following occurs:

1. External source signals (sinusoidal sweeps, pulses, etc) are computed and output by a D/A converter in PC1.

2. Stimulus signals are amplified and coupled to the output transducer.

3. Stimulus sound is ducted via acoustic conduit to the subject's vocal tract.

4. The resulting vocal tract response, with or without natural vocalization, is collected by a microphone.

5. Both the stimulus waveform (electrical) and the vocal tract response are sampled by A/D converters in PC2 and stored in memory.

This system provides for flexible artificial stimulation and recording of vocal tract responses in addition to the usual simple recording of pathological voice sounds on a single channel with a microphone.

A system model of the experimental setup, which proved useful, is illustrated in Fig. 3.3. In this model, the vocal tract is assumed to be a series element with the stimulus; a more detailed model might include parallel signal paths (representative of reflections of the acoustic waves). When the stimulus is present, S0 is closed. As shown, the stimulus time series is sampled by A/D channel one and passed through the dynamics of the amplifier, transducer, and environmental acoustics, which are modeled in H1(s) and H2(s). The sound stimulus may then stimulate or not stimulate the vocal tract depending

on whether or not the subject's mouth is open; this is modeled by S1 open or closed. Whether the subject is vocalizing or not is modeled by S2 closed or open. The glottal source signal is summed to the input of the vocal tract transfer function. The resulting acoustic signal is collected by a microphone and passed to A/D channel 2 via signal conditioning; these dynamics are represented by H3(s).

Using this setup, many experiments in vocal tract identification are possible. One approach to vocal tract identification proceeds as follows:

1. Time series of stimulus and response is acquired with S0 = CLOSED, S1 = CLOSED, S2=OPEN (stimulus on, mouth shut, glottis open).

2. Time series of stimulus and response is acquired with S0 = CLOSED, S1 = OPEN, S2 = OPEN (stimulus on, mouth open, glottis open). The vocal tract is held in the configuration for the desired vowel, eg. /a/, but the speaker is not vocalizing.

3. Using the stimulus/response of step 1, the transfer function H(s) = R(s)/X(s) mouth shut = H1(s)H2(s)H3(s) can be established.

4. Using the stimulus/response of step 2, the vocal tract model V(s) can then be calculated as V(s) = (R(s)/X(s) mouth open)/H(s). That is, the transfer function of the vocal tract can be recovered by dividing the transfer function of mouth open by the transfer function of mouth shut.

3.2 Validation with a Simple Tube Model

The first step in validation of the experimental setup and validation of the approach is to analyze the formants of a simple quarter wave tube model (shown in Fig.3.2 below the vocal tract depiction), which occur at:

F = C/4L, 3C/4L, 5C/4L ,

…. Where

C = Speed of sound (m/s)

L = Length of tube (m)

For this experiment, two time series were collected:

1. ha(t) is the time series response of the tube model to a 300 – 10,000 Hz chirp sampled at 40kHz.

2. hb(t) is the time series response of the same tube model, physical setup, and chirp, except a rag was solidly crammed into the tube.

Although ha and hb are not time/phase correlated, it is still possible to estimate the magnitude of V(s), ie, formant peaks of the tube.

From the model of Fig. 3.3,

$$|V(j\omega)| = |HA(j\omega)|/|HB(j\omega)|$$

The spectra of ha and hb are combined to form an estimate of the magnitude of V(s); the result is shown in Fig. 3.4. The following may be observed:

1. Formant peaks are prominently revealed, with magnitudes ranging from 60 dB at low frequencies to 10 dB at high frequencies. The peaks occur at frequencies of 660 Hz, 1980 Hz, … expected for an 11 cm tube (using C = 300m/s). Thus, the setup provides sufficient energy coupling to accurately reveal formants.

2. The more prominent contributions of H1, H2, and H3 to HA and HB are visible as the 20 dB wide peak from 0 – 5kHz, the fine 5dB ripples spaced at 94 Hz visible at low frequency (due to resonance inside the acoustic conduit), and 5dB peaks spaced at 300Hz at the higher frequencies. These "background" effects not due to the vocal tract are effectively removed when V is calculated. Thus, the effects of the transducers, conduit, room, etc are effectively minimized by the subtraction of the two spectra, resulting in the relatively smooth response in $V(j\omega)$.

3. When the chirp ends at 10000 hz, V turns to noise, as expected, since there is no stimulus energy available to reveal formants. This is excactly analogous to the failure of a pathological glottal source to reveal vocal tract formants.

4. As frequency increases, the tube's formants peaks shift and change shape, possibly due to the non-ideal realization of the tube.

3.3 Validation with a Vocal Tract

Having verified the setup for a tube model, a normal male (the author) vocal tract was tested next. In this case the open mouth vocalizing /a/ replaces the tube. In order to

achieve the same benefit of background cancellation, the mouth is either open or closed and silent (analogous to the rag stuffed into the tube). A rag was not used in the mouth. In order to compare ES (external source) analysis with the usual LP (linear prediction analysis) and FFT (fast Fourier transform), the sequencing scheme depicted in Fig. 3.8 was used. The external source executed a 0.2s pulse (to allow for possible impulse testing) followed by a 0.3s linear sine chirp from 300 to 4kHz.; the sequence repeats at about 2Hz. Simultaneously, the subject performs the following actions (each for about 1/3 of the test duration of 4 sec):

1. Mouth open vocalizing a normal /a/.

2. Mouth and glottis held in the same position, but not vocalizing.

3. Mouth shut and voice silent.

The subject's state is labeled in Fig.3.8. Both the vocalization and response of the vocal tract to external stimulation is clearly visible. The effects of the formants F2, F3, and F4 are clearly visible in the mouth open chirps in the microphone time series, and are almost absent from the mouth shut chirps (residual peaks may be due to nasal aperture effects or acoustic penetration of the closed lips).

The same spectral analysis and subtraction performed on the tube model is repeated on the vocal tract and displayed in Fig. 3.5. The top curves display the vocal tract and background (mouth open and shut) chirp responses, and the bottom curve illustrates the resulting spectral difference attributed to the vocal tract. Again, excellent signal to noise ratio is observed in the spectrum, with many fine details clearly visible. In

addition to the expected approximate formant frequencies for /a/, additional peaks are observed.

Because the subtraction process does not simultaneously measure mouth open/shut responses, the question arises as to how much movement of the articulators (which determine $V(j\omega)$) occurs between measurements. This problem is addressed in Fig.3.6, which repeats the calculation using two different mouth open chirps. As can be seen, there are slight shifts on the order of 100Hz in F2 and F4, while F1 and F3 remained within 10Hz; other peaks (some of which may not be vocal tract formants) also shifted. In this case, the subject was fairly successful in holding the articulators fixed. In the experiments that follow, this comparison is maintained to estimate time variations. If variations prove excessive, established advanced signal processing techniques that may be applied to the segments of simultaneous voice and ES to identify the vocal tract.

Another question that immediately arises is: How does ES analysis compare with standard FFT and LP (Section 2.1)? To address this question, FFT and LP are applied to the mouth open, voiced, external source quiet segment (0.3s – 0.8s) in Fig.3.8. The result is shown in Fig.3.7. Here the resulting ES calculations of Fig.3.6 are shown in the top curves, and the results of FFT and LP are shown in the bottom two curves. Several observations are possible:

1. All the formant peaks revealed by FFT/LP are present in essentially the same positions as the ES data. In this case, however, there are frequency shifts of greater magnitude than observed in Fig. 3.6; F2 appears about 200 Hz lower in the LP.

2. The ES data contains many substantial peaks above F4 that are almost invisible in the FFT/LP data. These are apparently resonances that are not excited by the glottal source.

Figures 3.9 and 3.11 continue this comparison with similar results. These shifts may be due to an even greater involuntary shift of articulators that may occur from the vocalizing state to the silent state.

In summary, it appears that the ES source data provides very precise vocal tract resonance information. Peak positions are even more sharply defined than in the FFT/LP data. Resonances not revealed with FFT/LP are shown with ES analysis. Problems with articulator drift may be overcome with more advanced testing/processing approaches.

3.4 Improvement for Pathological Source

The next question that naturally arises is how does ES perform on pathological voices? In order to obtain some experience applying ES to a simulated pathological condition, ES analysis was applied to normal male and female voices with the “breathy” condition in which the glottis does not obtain complete closure. Under these conditions higher frequency formants may be less stimulated by the more sinusoidal source waveform of the breathy condition, and FFT/LP (Section 2.1) may not reveal these peaks for determination of synthesizer settings (Chapter 4). This in turn leads to the problems of achieving proper vowel quality in simulation, discussed in the introduction to this chapter.

In testing the breathy condition, another task was added to the subject's set of actions shown in Fig.3.8. The subjects simulated the breathy condition at the start of the test, saying "haaaa" softly, attempting to achieve a sinusoidal driving function with consequent minimal high frequency formant excitation. The subjects transitioned to the normal /a/, effectively pronouncing "haaa-AAAH". This was followed by the usual mouth open, voice silent, mouth shut sequence. The result contains ES quiet segments with a breathy segment (samples 20000-35000) and a normal /a/ (samples 50000-65000), and then the voice silent/ES active segments. All these segments are recorded in one time series, as shown in Fig.3.10 (which is again the author's voice.)

The results of ES breathy testing are shown in Figs. 3.9 (normal) and 3.9b (breathy). In the normal voice of Fig.3.9, there is again fairly good agreement between ES and FFT/LP data. The FFT shows fundamental frequency harmonics of the modal /a/ well into the higher frequencies above 2500Hz; LP reveals clear peaks for F3 and F4, which are within about 100 Hz of the ES peaks. In Fig. 3.9b, the breathy FFT/LP data are shown; in this case harmonic peaks disappear at a much lower frequency. Here the LP peaks are almost absent, and the presence of F3 and F4 are revealed only by aspiration noise peaks in the FFT. The ES data, of course, remains the same, clearly revealing F3 and F4. In this case, ES is far superior to LP.

The same test is repeated with a normal female voice in Figs 3.11 (normal) and 3.11b (breathy). Again, fundamental frequency harmonics disappear at a lower frequency

in the breathy case. In this case, the LP does somewhat better in the breathy case, but the peaks of F3 and F4 are still reduced.

3.5 Summary

A major limitation of formant analysis and inverse filtering for pathological voices is the reliance upon the (possibly spectrally deficient) source to reveal formant information. With use of only the glottal source input to the vocal tract, segregation of the source waveform from the vocal tract using LP and inverse filtering may be ambiguous in the case of pathological voices, resulting in difficulty in achieving good vowel quality in synthesis. Without source energy representation across the entire spectrum of interest (which is supplied by the sharp return phase of a normal voice), resonances in pathological voices may not be detected with LP and other techniques. By externally stimulating the vocal tract with a spectrally rich source (chirp, impulse, white noise, etc), all resonances are clearly detected. Information obtained from ES analysis may then be combined with other approaches to yield more accurate formants and inverse filtered waveforms. The application of ES testing was successfully demonstrated via a progressive series of experiments. Formants of a known physical form (tube) were detected at the expected frequencies. Application of ES testing to the vocal tract revealed high resolution detection of the expected formants. In addition, comparison of ES test results with traditional LP and FFT analysis of the same voices reveals ES testing produces higher resolution, more detail, and additional resonances not detected at all by

LP and FFT. Experiments with simulated pathological (breathy) voices demonstrate a particular case in which ES testing improves formant estimation. Problems of articulator movement during ES testing may be solved by any of several technical approaches. Thus, the value of ES testing for pathological voice analysis is illustrated.

Two limitations or differences from typical FFT/LP vocal tract identification were observed. The typical 12 dB per decade decrease in signal observed in the FFT of a typical voice is not seen. This is expected as the chirp does not fall off in intensity with increasing frequency as the typical glottal source does, and because this ES measurement technique cancels out any variation in source intensity with frequency via spectral subtraction (Section 3.1). In addition, there are extra resonances in the ES data; the source of these peaks has yet to be identified.