Analysis and Digital Implementation of the Talk Box Effect

Yuan Chen

May 14th, 2012

Advisor: Professor Paul Cuff

Submitted in partial fulfillment of the requirements of the degree of

Bachelor of Science in Engineering

Department of Electrical Engineering

Princeton University

I hereby declare that this independent work report represents my own work in accordance with University regulations.

Yuan Chen

Analysis and Digital Implementation of the Talk Box Effect

Yuan Chen

Abstract

The talk box is an analog device which allows a human user to use his vocal tract to create an intelligible output from the input of a musical instrument, typically a keyboard or electric guitar. The intelligibility of a general speech signal results from the frequencies of its formant peaks, which correspond to the resonance characteristics of vocal tract. We model speech production as the output of the periodic glottal pulse excitation input into a linear, time-variant vocal tract system, and we use the complex cepstrum to separate source and filter. We examine the output from the talk box and identify formant peaks in the separated filter frequency response. Using cepstral liftering to calculate vocal tract impulse response, we create a digital implementation of the talk box which takes as input a speech signal and an instrument signal and produces intelligible output similar to that of the analog talk box.

Acknowledgements

I would like to thank my advisor, Professor Paul Cuff, for his guidance and support throughout the project. Time and again, he has provided me with innovative ideas to solve complex problems and thought-provoking questions to further my own perspective. I could not have accomplished this without his continued help, knowledge and insight.

I would also like to thank my friend Joey Edelmann for playing talk box guitar and providing vocals for the recordings used in the project. His knowledge of the talk box and expertise as a guitar player has been crucial to the process of analysis and design.

Finally, I would like to thank Wilson College for letting me use its recording facilities.

Table of Contents

1. Introduction

The talk box effects unit uses the resonant features of the human vocal tract to shape the output sound of an instrument. The unit uses a compression driver to produce the sound waveform of an instrument excitation signal and delivers the sound through a vinyl tube into the musician's mouth [1]. The musician shapes his/her mouth and vocal tract as if speaking (without physically creating sound) to give intelligibility to the outgoing instrument signal. The talk box differs from the vocoder effect in that the vocoder requires the musician to physically speak into a microphone and the voice signal controls a series of band-pass filters which shape the instrument signal [2]. The number of band-pass filters determines the resolution of the vocoder in frequency domain and in turn affects the quality of resemblance of the output signal to human speech. The problem of frequency resolution does not affect the talk box, since the musician uses the vocal tract directly to shape the output signal. The purpose of this project is to first analyze the characteristics of the output signal from the talk box unit to determine and confirm the features which give rise to its intelligibility. We then create a digital implementation of the talk box which extracts the vocal tract impulse response from an input of speech and uses this impulse response to filter the instrument input signal.

2. A Model of Human Speech Production

Fant summarizes the speech signal as "the response of the vocal tract to one or more excitation signals" [3]. For voiced speech, the excitation is the vibration of the vocal cords in a periodic manner caused by the buildup and release of air pressure in the glottis [4]. An adult speaker can change the frequency of vocal cord vibrations in a range of 50 to 500 Hz by varying the tension in the cords and the air pressure from the lungs [4]. Although the glottal pulse has a continuous shape waveform, we can qualitatively approximate it as an impulse train.

The glottal pulse impulse train excites the resonant frequencies of the vocal tract, which we model as a linear filter. Gold et al note that speech production "often involves varying degrees of nonlinear behavior, usually at the interface between excitation and filter" [5], though the effects are generally minor. The excitation-system model is most applicable to voiced speech, as only "the glottal vibrations of voiced sound are significantly modified by the acoustic resonances of the vocal tract" [4]. Unvoiced and mixed sounds involve the forced constriction of

the vocal tract at a specific point and are not as significantly affected by the vocal tract. The primary mechanism for articulation of speech is the variance of the vocal tract in time to change the impressions of vibratory patterns on airflow.

Thus we cannot model the vocal tract system in the long-time perspective as time-invariant. The shape of the vocal tract and the impulse response of the system changes as the speaker varies the sound he or she produces. We seek a linear, time-variant system model of the vocal tract so that we model speech production as a convolution of glottal excitation and vocal tract impulse response. To satisfy these constraints, we must consider a time-windowed representation of speech such that each window contains a single phonetic unit [4]. In the short-term, windowed perspective, the time-invariance condition on the vocal tract system is a more reasonable approximation.

3. Speech Intelligibility

The frequency domain analysis of the speech signal of a single phonetic unit shows a general envelope, and we view the vocal tract as an envelope function that shapes spectral features [4]. The peaks of the envelope are known as formants and indicate the resonant frequencies of the vocal tract [5]. For vowel sounds, which have the largest average amplitude and amplitude range [5] and thus have the best defined spectral envelopes, there are four to five formants of interest, with three of these formants necessary for recognition [4]. The spectral properties of unvoiced sounds do not show as well-defined of an envelope, and thus it is difficult to detect formants which give rise to their intelligibility.

By convention, we denote lowest frequency formant peak as F1 and subsequent frequencies as F2, F3, etc. In general, F1 contains most of the energy and pitch information of the speech signal, while higher formants account for the intelligibility of the speech [5]. There is a mismatch between energy and amplitude and contributions to intelligibility within the frequency band of highest ear sensitivity (300 Hz – 4 kHz). That is, the formant that contributes the most power and has the highest amplitude is not the same as the formant(s) that provides speech-like qualities to the signal. For an average speech signal, 84% of energy is found in frequencies under 1 kHz. On the other hand, F2 and F3, which contribute most to intelligibility of voiced sounds, are located in the frequency band 800 Hz – 3 kHz [5].

McLoughlin states that the region between 1 kHz and 4 kHz is the critical region for detecting speech intelligibility [5]. A signal that is low-pass filtered at frequency 1 kHz has approximately 25% of the original syllables as recognizable units, and removing all energy in the 1 kHz to 4 kHz band results in a signal that sounds approximately as full as the original signal but completely unintelligible. Conversely, high-pass filtering at 1 kHz leaves more than 75% of syllables recognizable and results in a speech signal that is qualitatively quieter and brighter (due to the lack of low frequency content) but one that still contains the original speech features [5]. Regardless of the excitation signal, the resemblance of a sound signal to a real unit of speech depends significantly on this frequency band. That is, in theory we can recreate intelligible speech by properly shaping this frequency band regardless of the frequency content of the original excitation signal.

4. Source-Filter Separation

Based on our model of voiced human speech production, speech consists of the glottal excitation pulse used as input to a linear, time variant vocal tract system and outputted as the final waveform. The talk box unit replaces the glottal excitation with the instrument signal. Thus the output signal of the talk box effect is the convolution of the instrument signal with the vocal tract impulse response. In order to perform this operation, we need to extract the system response from the voice signal, which itself is a convolution. We can use cepstral analysis and the liftering operation to perform this de-convolution [6].

Again from our model, we represent a general speech signal $s(n)$ as:

$$s(n) = e(n) * \theta(n), \tag{4.1}$$

where $e(n)$ is the excitation signal and $\theta(n)$ is the vocal tract impulse response. The goal of the de-convolution operation is to separate the two components of $s(n)$.

We represent the discrete-time Fourier Transform (DTFT) of the $s(n)$ as:

$$S(\omega) = E(\omega)\Theta(\omega), \tag{4.2}$$

where $E(\omega)$ and $\Theta(\omega)$ are the DTFT of the excitation and vocal tract impulse response, respectively.

The cepstrum is a representation of the speech signal with two important properties: first, the representative component signals are separated in the cepstrum and second, the representatives of the component signals are linearly combined in the cepstrum [6]. The operation transforms the multiplicative combination of the two signals in the frequency domain into a domain where the combination of the two signals is linear.

We define the complex cepstrum of a signal as [6]:

$$\gamma_s(n) = \mathcal{F}^{-1}\{\log \mathcal{F}\{s(n)\}\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log S(\omega)\, e^{j\omega n} d\omega. \tag{4.3}$$

From equation (4.2), we substitute into equation (4.3) for the definition of $S(n)$.

$$\gamma_s(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\big(E(\omega)\Theta(\omega)\big)\, e^{j\omega n} d\omega$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\big(E(\omega)\big)\, e^{j\omega n} d\omega + \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\big(\Theta(\omega)\big)\, e^{j\omega n} d\omega$$

$$= \gamma_e(n) + \gamma_\theta(n). \tag{4.4}$$

From equation (4.4), we see that the complex cepstrum of the speech signal is just the linear combination of the cepstrum of the excitation signal and the cepstrum of the impulse response, since the operation of taking logarithms transforms multiplication into addition.

We can approximate the excitation signal $e(n)$ as an impulse train in time, and $E(\omega)$ is then an impulse train in frequency. Qualitatively, this means that $E(\omega)$ is the fast varying component of $S(\omega)$ and its cepstrum has high quefrency (the "time" axis of the cepstrum) components; that is, $\gamma_e(n)$ is non-zero for large values of n. Conversely, $\Theta(\omega)$ is a slow varying frequency envelope which shapes the impulse train and has only low quefrency components, meaning that it is non zero for small values of n. Although we cannot provide a formal proof of why the cepstrum of excitation and impulse response occupy different parts of the quefrency

domain, we intuitively show along the quefrency axis, at most one component of the speech signal cepstrum (excitation or vocal tract impulse response) has non-zero value [3].

Therefore, we can separate the two components of the speech signal by the liftering operation, filtering in the quefrency domain. By applying a low quefrency filter to the cepstrum of the speech signal, we preserve only the coefficients corresponding to the spectral envelope, from which we can compute the impulse response of the vocal tract. The complex cepstrum is an invertible operation and preserves phase, and the resulting impulse response following quefrency liftering is causal.

5. Recording Audio Signals

We record audio signals of a talk box output for analysis and to serve as comparison for the synthetically produced talk box output. In addition, we record the human voice and unmodified guitar audio to serve as input into the digital implementation of the talk box. The guitar we record is a Fender American Standard Telecaster amplified using the distortion channel of a Fender 2x12 SuperSonic tube amplifier. All equalization and gains are set to 50% of the maximum setting. The talk box uses the send output of the pre-amplifier as its guitar input. This input drives a speaker horn connected to a vinyl tube, which the musician places in his or her mouth to produce the talk box sound. Although the distortion channel of the amplifier yields an unmodified guitar signal that is qualitatively less clean in frequency content, we find that the talk box performs better with the distorted signal.

We record three different unmodified guitar signals which we use in synthesis of the talk box sound: the note C-262 playing at a rate of 60 beats per minute, the C major scale playing at 60 beats per minute and the C major scale playing at 120 beats per minute. We use the first of these signals to synthesize single words and sounds in which the throat impulse response remains invariant through the duration of the signal. We use the latter two signals to synthesize longer duration outputs in which the throat impulse response changes over time in order to test the implementation in a more realistic setting where we cannot preemptively segment the audio input into individual phonemes.

Of the different phonemes present in the English language, vowels have the largest average amplitude [5], have the best defined formant peaks, and most closely follow our LTI model of human speech production [6]. Accordingly, we initially focus on the analysis and

synthesis of talk box vowel sounds. There are twelve vowel sounds in the English language [6], and for each of these sounds, we record the talk box and human voice each producing both the isolated sound as well as the sound as a part of a word. We record the talk box signal with the guitar playing the C-262 note at a rate of 60 beats per minute, and we record the human voice speaking at a natural pitch, which we do not specify, at a rate of 60 beats per minute.

Similarly, we record one representative phoneme from each of the other phonetic categories in the English language. These other phonetic categories contain on average smaller amplitude and less energy than vowel sounds and are not necessarily continuous and time-invariant. For this reason, we do not seek to record isolated forms of these sounds and instead record them as part of a word. In each case, we record a sound in which the specific phoneme occurs before the vowel sound in the word, or in the case of the diphthong, replaces the vowel sound in the word. As with vowel sounds, we record a talk box version of the phoneme with the guitar playing the C-262 note at 60 beats per minute and a human voice version at a natural pitch speaking at a rate of 60 beats per minute.

In addition to the short time signals of single words and phonemes, we record longer duration talk box and voice signals corresponding to the longer duration unmodified guitar signals. We record the vocal scale ("Do-Re-Mi-Fa…") in talk box and human speech form at rates of 60 beats per minute and 120 beats per minute. We output all recorded audio files as stereo .mp3 files with identical channels, 44100 Hz sampling rate and 16 bit quantization depth.

6. Spectral and Cepstral Analysis of Vowels

For each of the recordings of isolated vowel sounds from both the human voice and talk box, we compute the Fourier transform and examine the frequency content of each signal for formant peaks. The talk box uses the human vocal tract as a filter for the instrument signal, and we expect the formant peaks for the talk box recording and human speech recording of the same vowel to be similar. The formant peaks need not be identical, since it is almost impossible for the same speaker to produce the same sound twice in exactly the same manner [4].

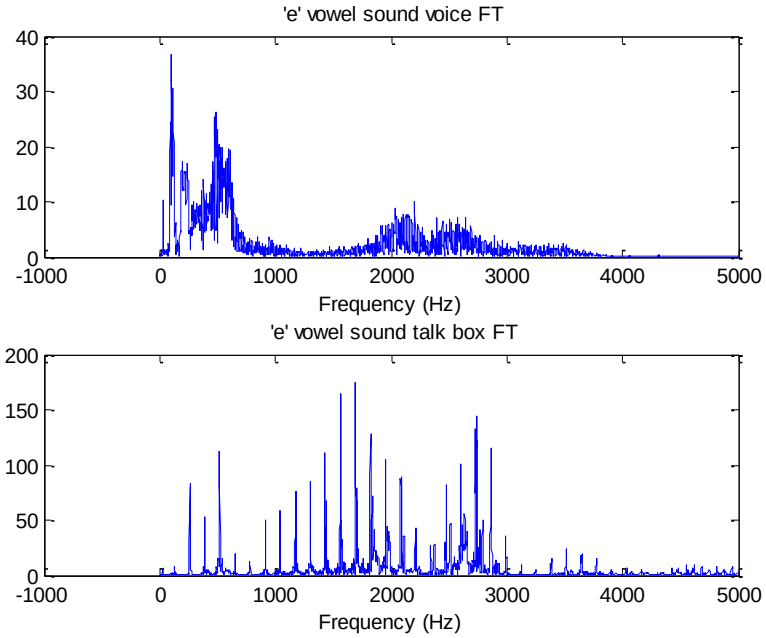

Figure 6-1: Fourier transform of the ‘e’ (“bait”) vowel sound (a – top) from human voice and (b – bottom) from talk box

Figure 6-1 shows the Fourier transform of the ‘e’ vowel sound from human voice (top-a) and the talk box (bottom-b). Although the signal can have frequency content up to 22050 Hz, we are interested in only the region from 0 to 5000 Hz as this is the region which affects the intelligibility of speech. We also see from the Fourier transform that the human voice has negligible frequency content above 4000 Hz. Furthermore, we do not consider the phase of the Fourier transform, since it does not contribute directly to speech intelligibility.

From figure 6-1a, we observe that the signal has highest amplitude frequency content at frequencies less than 1000 Hz. Specifically, there are local maxima at 96 Hz, 238 Hz, 490 Hz and 599 Hz. These peaks most likely correspond to the pitch of the human voice, especially the two peaks at 238 Hz and 490 Hz. In addition, we calculate the cumulative energy of the signal as a function of frequency. For this specific recording of human voice producing the ‘e’ vowel sound, we find that the frequency range from 0 to 1000 Hz contains 86% of the energy. This observation is consistent with the fact that formant peak F1, which occurs in the frequency range under 1000 Hz for vowels, accounts for most of the energy and pitch information.

The Fourier transform of the human voice recording also shows amplitudes which vary quickly frequency. Moreover, the spectrum indicates that there is content in all frequencies up to

4000 Hz without any specific pattern of peaks. We observe the overall envelope in frequency with peaks occurring roughly at 500 Hz, 2100 Hz and 2600 Hz. Because the Fourier transform varies so rapidly in frequency, we cannot use the local maxima directly to identify formant peaks. That is, we expect the envelope to be smooth in frequency with few local maxima. In figure 6-1a, we identify 2036 Hz and 2206 Hz as local maxima, but using the frequency envelope we instead estimate this peak to be around 2100 Hz. Since the spectrum varies rapidly in frequency, there are too many local maxima to consider each to be a contributing formant. Furthermore, we also cannot simply take the largest of these local maxima within a frequency region to be the formant peaks.

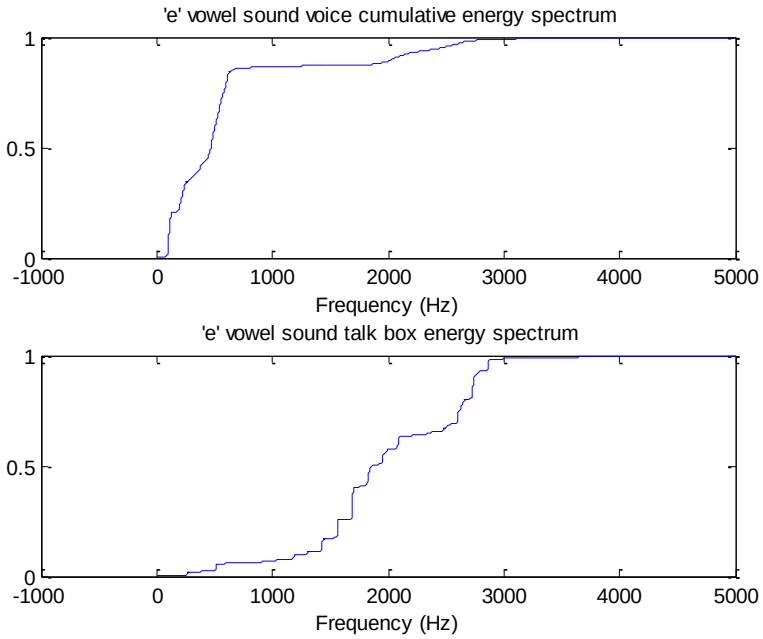

Figure 6-2: Cumulative energy distribution of the ‘e’ vowel (a – top) from human voice (b – bottom) from talk box

From figure 6-1b, we notice that unlike human speech the talk box signal has very specific large peaks corresponding to the harmonics of the guitar note C-262. Even though the fundamental pitch occurs in the region below 1000 Hz, there is also no concentration of energy in the lower frequencies for the talk box signal. As we see in figure 6-2b, the frequency range of 0 – 1000 Hz contains 6.7% of the energy for the signal. We also observe a basic envelope structure with three peaks centered at 519 Hz, 1692 Hz and 2734 Hz. Because the spectrum of the talk box signal is more periodic in frequency than the spectrum of human speech, the

magnitude of the Fourier transform does not vary as rapidly in frequency, and we take the peaks of the envelope to be the largest three local maximum peaks. Nevertheless, we note that we cannot directly extract a frequency envelope and formant peaks from the spectrum of the talk box signal. While we observe the general shape of a frequency envelope, we cannot conclude that the local maxima are the formant peaks, since the frequencies with highest amplitude in the talk box spectrum are always the harmonics of the instrument excitation input.

Formant peaks correspond to resonant frequencies of the throat impulse response, and we cannot calculate this response using only the Fourier transform of the speech signal. As figure 6-1a shows, the spectrum of human speech varies rapidly while the throat impulse response is an envelope which varies slowly in frequency by the definitions of our human speech model. In addition we cannot treat the largest peaks of the talk box spectrum to be the formant peaks since true frequency of the formant may occur in between the discrete peaks of the instrument excitation. To obtain a more reliable calculation of the throat impulse and frequency response, we use cepstral liftering to separate the two components of speech and talk box signals.

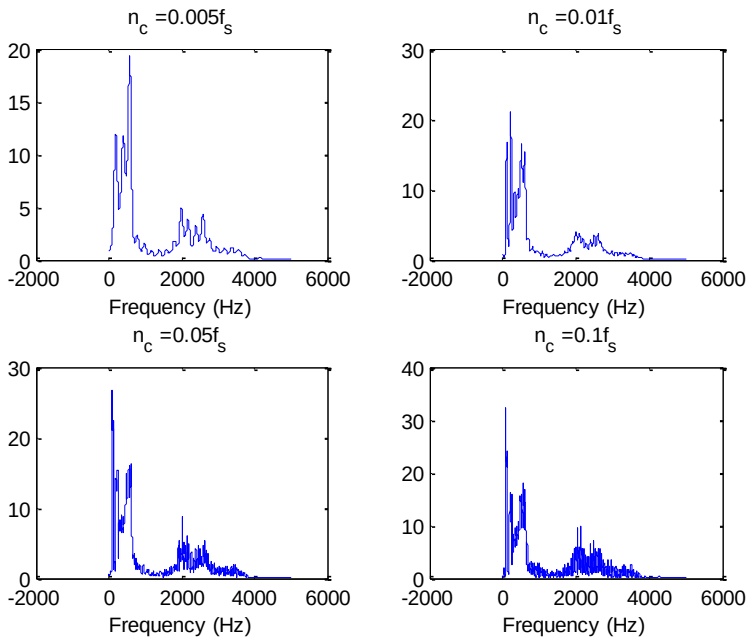

Figure 6-3: Fourier transform of the 'e' vowel sound from human voice after cepstral liftering (a – top left) with $n_c = .005f_s$ (b – top right) with $n_c = .01f_s$ (c – bottom left) with $n_c = .05f_s$ (d- bottom right) with $n_c = .1f_s$

For each audio clip of human speech and talk box output, we compute the complex cepstrum and apply an ideal low-time lifter, which zeros all cepstral coefficients above the cutoff quefrency n_c. We express n_c as a proportion of the sampling rate of the signal f_s. As we decrease n_c, the resulting spectrum varies more slowly in frequency, and we observe a more apparent frequency envelope. Figure 6-3d shows that with $n_c = .1f_s$ the resulting Fourier transform still varies rapidly in frequency and is similar to the Fourier transform of the unliftered speech signal. On the other hand, figure 6-3a shows that with $n_c = .005f_s$, the resulting Fourier transform does not have fast varying components in frequency. We take this liftered signal to be the frequency response (envelope function) of the throat and proceed with cepstral analysis using $n_c = .005f_s$.

The Fourier transform of the liftered 'e' vowel sound from human voice shows that the frequency region under 1000 Hz has the highest amplitude and still contains a significant portion of the signal's energy. Because F2 and F3 occur in the frequency range 800 Hz – 3000 Hz, with most of these specific formants occurring above 1000 Hz [5], we only consider the largest local maximum peak in the 0 – 1000 Hz region as the formant peak F1. We then consider the subsequent largest local maxima in the frequency region above 1000 Hz to be F2, F3, etc. The large amplitude of the other local maxima in the 0 – 1000 Hz band is an artifact of the property that this region contains most of the energy in natural human speech, and accordingly, we do not consider these peaks to be formants. From figure 6-4a, we identify 574 Hz, 1995 Hz and 2590 Hz as formants in the throat frequency response of the human voice recording, and from figure 6-4b, we identify 515 Hz, 1890 Hz and 2659 as formants in the throat frequency response of the talk box recording.

While the $n_c = .005f_s$ liftering condition removes the fast-varying component of the Fourier transform, it does not yield an equally smooth frequency envelope for all of the recorded clips. This lifter creates very smooth, well-defined frequency envelopes for certain vowels of the talk box output, and for others, the Fourier transform of the liftered signal still contain noticeable traces of the discrete frequency peaks of the un-liftered signal. For consistency, we choose to keep n_c the same for all signals. If we continue to decrease this cutoff quefrency, the envelopes that are already well defined under the current value of n_c becomes too slow varying in frequency, and we cannot accurately calculate the formant peaks.

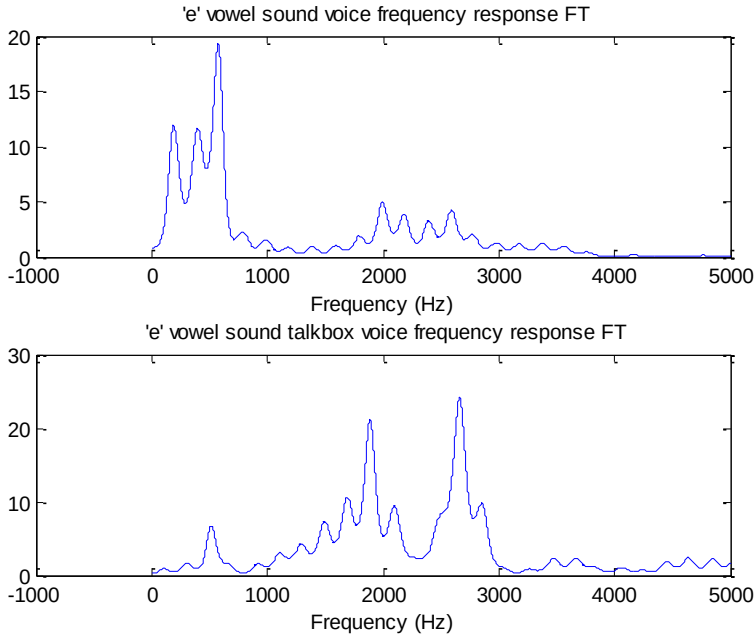

Figure 6-4: Fourier transform of the 'e' vowel sound liftered at $n_c = .005N$(a – top) from human voice (b – bottom) from talk box

We compute the frequency response by cepstral liftering of the throat for all vowels from both human voice and talk box sources and examine the spectrum for formant peaks. We report the results in table 6-1.

The average absolute error between the frequencies of formant peaks for the voice signal and talk box signal is 99.8 Hz for F1, 144.4 Hz for F2 and 155.0 Hz for F3. Human frequency discrimination, however, is more dependent on the relative error [5]. The average relative error is 19.6% for F1, 9.33% for F2 and 6.21% for F3. From the perspective of relative error, the talk box performs best at impressing the F2 and F3 formant peaks onto the input instrument signal. This observation is consistent with the claim that F2 and F3 account for intelligibility of speech, while F1 accounts for pitch and energy.

Table 6-1: Formant frequencies for vowel sounds from human voice and talk box sources

Vowel	Source	F1 (Hz)	F2 (Hz)	F3 (Hz)
e (hate)	Voice	574	1995	2590
	Talk Box	515	1890	2659
@ (at)	Voice	799	1700	2575
	Talk Box	799	1792	2597
i (eve)	Voice	293	2467	2863

	Talk Box	486	2093	2705
E (met)	Voice	670	1713	2660
	Talk Box	509	1685	2494
R (bird)	Voice	581	1392	1780
	Talk Box	495	1494	1889
I (it)	Voice	485	1958	2696
	Talk Box	509	1658	2496
o (boat)	Voice	479	1060	2262
	Talk Box	569	1256	2474
u (boot)	Voice	324	1114	2348
	Talk Box	400	1394	2375
a (father)	Voice	792	1200	2389
	Talk Box	597	1188	2581
c (all)	Voice	663	2749	3512
	Talk Box	893	2685	3668
U (foot)	Voice	438	1201	2293
	Talk Box	425	1370	2540
A (up)	Voice	645	1367	2531
	Talk Box	575	1378	2384

In addition, our calculation of the signals' cumulative energy distribution in the frequency domain shows that human speech signals and the throat impulse responses we calculate using cepstral liftering, have energy concentrated in the 0 – 1000 Hz frequency band. We do not observe this property in the talk box signals. This suggests that the talk box and throat impulse response does not amplify this low frequency band. The concentration of energy in this frequency range of human speech is a result of the excitation source and not the throat filter. One reason the calculated throat frequency response has larger magnitude in this region is because cepstral liftering does not exactly separate source and filter. The magnitude in the 0 – 1000 Hz region of the original human voice recording persists through the liftering process and distorts our calculation of the vocal tract frequency response.

We do not extensively analyze the frequency envelope and formant peaks of the talk box and human speech in other phonetic categories of the English language. By definition, diphthongs and glides have time variant impulse responses, and non-continuant phonemes occur naturally as part of a word which contains a vowel sound [6]. For these phonetic categories, we cannot use the long term Fourier transform since it does not express changes in the throat frequency response over time. Voiced continuant phonemes have a much smaller amplitude and

amplitude range compared to vowels, and unvoiced phonemes do not have well defined formant peaks [5]. In each case we can use the short-term Fourier transform and complex cepstrum to perform a similar form of analysis, but in order to obtain adequate resolution in time, we trade-off resolution in frequency. For voiced phonemes, whose production closely follows our source-filter model of human speech production, we expect the results of the analysis to be similar to the results from our analysis of vowels. While we later examine the performance of the talk box on the rest of the phonetic categories as part of testing our digital implementation, we nonetheless do not pursue this analysis in detail, due to limitations in time and frequency resolution.

7. Design and Implementation

Our analysis of vowel sounds shows that the talk box applies a filter with impulse response equal to the vocal tract impulse response to the input instrument source. The resulting output has identifiable formant peaks in the F2 and F3 frequency region corresponding to the formant peaks of human speech, which account for the output's intelligibility. While the talk box is an analog device which uses continuous-time input from both the instrument and the vinyl tube in the musician's mouth, our design uses only discrete-time signals. We create the discrete-time version of the instrument input through sampling, but we cannot however directly sample and reconstruct the shape and impulse response of the vocal tract. Instead, our design takes as input a sampled, discrete-time version of human speech and computes the vocal tract impulse response using cepstral liftering. Unlike the analog talk box, our digital implementation requires the musician to physically produce speech sounds. We implement our design as MATLAB functions which take as input a recording of the guitar audio and a recording of human speech. The final design interprets each of the inputs sequentially, so as to simulate one aspect real-time processing, and outputs the synthesized talk box signal.

Because we implement our design only as a simulation, we must perform additional preprocessing on the input signals. Just as the duration of the instrument input signal limits the duration of the output from the physical talk box, our simulation will output a signal that is equal in duration to the input instrument signal. The simulation ignores any vocal input that extends beyond this duration and outputs any portion of the input instrument that is longer than the voice input in its original form. Additionally, we must also ensure that the input audio clips correspond to each other in the time domain: we choose the number of samples of silence at the beginning of

each clip so that the true audio portion of each signal begins at the same time. The physical talk box does not encounter this problem because it uses the instrument signal and vocal tract impulse response directly, not recordings of instrument and human voice. We do not implement the preprocessing of inputs to ensure true audio begins simultaneously as part of the talk box function, since the physical device has no such capability, and instead perform the preprocessing on the audio clips before using them as input for the digital talk box.

7.1 Problem Definition and Design Objective

Given an instrument input signal $g(n)$ and a human speech input signal $s(n)$, we seek to estimate the mapping the vocal tract, $\mathcal{H}\{\cdot\}$, where $s(n) = \mathcal{H}\{e(n)\}$, $e(n)$ is a glottal pulse excitation, and $\mathcal{H}\{\cdot\}$ is causal, nonlinear and time variant. From our estimate $\widehat{\mathcal{H}}\{\cdot\}$, which is causal and linear but not necessarily time invariant, we then seek to synthesize a talk box output $y(n) = \widehat{\mathcal{H}}\{g(n)\} \approx \mathcal{H}\{g(n)\}$.

7.2 Vocal Tract Impulse Response Extraction

Given a segment of human speech $s(n)$, we first seek to separate the source and filter components. Recall that according to our LTI model of speech production:

$$s(n) = (e * \theta)(n), \tag{4.1}$$

$$S(\omega) = E(\omega)\Theta(\omega). \tag{4.2}$$

Initially, we make the assumption that the vocal tract impulse response is time invariant throughout the duration of the signal and we only need to calculate a single impulse response $\theta(n)$. Under this assumption, we also do not read the input sequentially and calculate $\theta(n)$ using all samples of $s(n)$. This static implementation does not simulate real-time processing since in the output signal, the n_0 sample depends on samples $n > n_0$ of the original input.

We compute the complex cepstrum of $s(n)$ according to equation 4.3:

$$\gamma_s(n) = \mathcal{F}^{-1}\{\log \mathcal{F}\{s(n)\}\}. \tag{4.3}$$

In our implementation, we compute all Fourier transforms using the FFT algorithm and all inverse Fourier transforms using the IFFT algorithm. For a speech signal with N samples, the corresponding complex cepstrum also has N samples. Although we index the complex cepstrum beginning with index 1, we define the domain of $\gamma_s(n)$ to be $[-\frac{N}{2}, \frac{N}{2})$. The low quefrency coefficients then correspond to the indices near $\frac{N}{2}$. We design an ideal low quefrency lifter:

$$l(n) = \begin{cases} 1, & \left\lfloor \frac{N}{2} - n_c \right\rfloor \leq n \leq \left\lceil \frac{N}{2} + n_c \right\rceil. \\ 0, & otherwise \end{cases} \tag{7.1}$$

We use the floor and ceiling functions in equation 7.1 since both γ_s and l are discrete-time signals defined only for integer values of n.

Recall that we define the cutoff quefrency n_c as:

$$n_c = \beta f_s. \tag{7.2}$$

β defines the proportion of samples which are preserved in the liftering process and, we use $\beta = .005$ to calculate the frequency envelopes. The cutoff quefrency n_c is a proportion of the sampling frequency f_s so that regardless of the length of input clip, the lifter preserves components of the same quefrency.

From the complex cepstrum γ_s and the ideal low-quefrency lifter $l(n)$ we compute a liftered complex cepstrum:

$$\gamma_l(n) = \gamma_s(n) l(n). \tag{7.3}$$

We invert the process of calculating the complex cepstrum by taking the exponential of its Fourier transform [6]:

$$\Theta(\omega) = \exp(\mathcal{F}\{\gamma_l(n)\}). \tag{7.4}$$

Under the assumption that the vocal tract impulse response is invariant for the duration of input signal, we compute $\Theta(\omega)$ only once using equation 7.4.

If we relax such an assumption, we must calculate the frequency response of the vocal tract as a function of time $\Theta(\omega, n)$. We take short-time windowed versions of the voice input $s(n)$ and define a frame of the signal of length M ending at time m as [6]:

$$s(n;\ m) = s(n)w(m-n), \tag{7.5}$$

$$w(n) = \begin{cases} 1, & n = 0, 1, 2, \dots, M-1 \\ 0, & otherwise' \end{cases} \tag{7.6}$$

where $w(n)$ is a rectangular window of length M. The dynamic implementation which considers the input as frames of signal simulates real time processing since output sample n is a function of input up to sample $n + M'$. Here M is not the same value as M' because our implementation uses more than one frame at a time to calculate the impulse response. A small enough value of M yields frames of speech in which the vocal tract impulse response is approximately invariant through the duration of the frame. As we decrease M, we trade off resolution in frequency and quefrency for resolution in time. We experimentally find that $M = 3000$ for voice input sampled at 44100 Hz qualitatively yields the best performance.

For a given frame of speech $s(n; m)$, we calculate the vocal tract impulse response of the frame in the same way that we calculate the vocal tract impulse response of a time invariant speech signal:

$$\gamma_s(n; m) = \mathcal{F}^{-1}\{\log \mathcal{F}\{s(n; m)\}\}, \tag{7.7}$$

$$\gamma_l(n; m) = \gamma_s(n; m)l, \tag{7.8}$$

$$\Theta(\omega; m) = \mathcal{F}\{\theta(n; m)\} = \exp(\mathcal{F}\{\gamma_l(n; m)\}). \tag{7.9}$$

Equation 7.9 computes the impulse response of the vocal tract for a frame speech of length M ending at time m. By considering the speech signal as a sequence of frames, our implementation simulates this portion of processing in real time and updates the system impulse response to match the varying impulse response of the vocal tract over time.

7.3 Impulse Response Preprocessing

Recall our analysis of vowels which shows that the impulse response we calculate using cepstral liftering has high energy in the frequency band from 0 – 1000 Hz and that the same frequency band of the talk box spectrum does not contain a significant portion of the output signal. If we use the result of equation 7.4 (or equation 7.5 in the dynamic implementation), there is too much low frequency gain compared to the physical talk box. For the static design, we create a new piece wise frequency response:

$$\Theta(\omega) = \begin{cases} 1, & \omega < \dfrac{1000}{2\pi} \\ \Theta_{\text{old}}(\omega), & otherwise \end{cases}, \tag{7.10}$$

where $\Theta_{old}(\omega)$ is the original, unprocessed frequency response. Our analysis of vowels shows that the talk box does not consistently create the same F1 formant in its output as the F1 formant in the corresponding speech. In this preprocessing scheme, we completely ignore the effect of the system on the low frequency range.

In the dynamic implementation, we have the additional problem of normalizing the impulse response between the signal frames. The physical talk box does not directly affect the overall amplitude of the output signal. When the musician does not explicitly form his vocal tract, the talk box still produces the output signal without attenuating its amplitude. The analog of this situation in the digital implementation is the musician not producing speech. In this scenario, the corresponding frames of audio input signal have low amplitude and the resulting frequency responses have low gain across all frequencies. If we use the un-normalized impulse response, the output from the digital talk box will be attenuated and in some cases inaudible for these lower-amplitude frames. We do not desire for the output of the digital implemementation to have amplitude dependent on the amplitude of the voice input, as such behavior deviates from the behavior of the physical talk box.

We perform pre-processing for the dynamic implementation in the frequency domain to simultaneously decrease the low frequency gain and normalize the maximum gain of a given frame. Normalizing the maximum amplitude of each frame in time also scales down the resulting impulse and frequency response, but using such a scheme does not enable us to control the gain

of specific frequencies. For the dynamic implementation, we create the following normalized frame frequency response:

$$\Theta(\omega;m)=\begin{cases}\dfrac{0.3\Theta_{\text{old}}(\omega;m)}{\max\limits_{\Omega}\{|\Theta_{\text{old}}(\Omega;m)|:\Omega\le\frac{1000}{2\pi}\}}, & \omega\le\dfrac{1000}{2\pi}\\[2ex] \dfrac{\Theta_{\text{old}}(\omega;m)}{\max\limits_{\Omega}\{|\Theta_{\text{old}}(\Omega;m)|:\Omega>\frac{1000}{2\pi}\}}, & \omega>\dfrac{1000}{2\pi}\end{cases}. \tag{7.11}$$

Equation 7.11 assumes that the maximum magnitude of $|\Theta_{\text{old}}(\omega;m)| \neq 0$ for both the high frequency (above 1000 Hz) and low frequency bands. We account for this condition and set the gain of a specific frequency band to be constant if the maximum magnitude is zero in that frequency band. We experimentally determine that normalized gains of 0.3 for the low frequency band and 1.0 for high frequency yields the output which most closely resembles the output of the physical talk box.

7.4 Synthesis

We synthesize the output of the static design in a straight forward fashion. Equation 7.10 calculates the time invariant system impulse response $\theta(n)$, and we convolve the instrument input $g(n)$ with this impulse response to produce the output signal $y(n)$. We perform the convolution by multiplying the system frequency response and the Fourier transform of the instrument input since it is less computationally expensive to use the FFT algorithm and multiplication that to perform convolution directly.

For the dynamic design, we cannot use the same approach as in the static design to synthesize output. Although we consider $s(n)$ in sequential frames to calculate the time varying impulse response, we seek to treat $g(n)$ as a single signal. To perform convolution of the instrument and speech frame requires us to also break the instrument signal into individual frames. We find that this method produces output which has rough transitions between the frames and do not further pursue this implementation. Furthermore, we expect process of taking frames both in $s(n)$ and $g(n)$ to contribute to the poor quality of the output so that even we consider only $s(n)$ in sequential frames and treat $g(n)$ as a single signal, the output still has distinguishable transitions between frames.

To reduce the perception of the frame transitions, we first window the signal with 50% overlap between consecutive frames. Recall that $s(n;m)$ is a frame of signal $s(n)$ of length M ending at time m. The dynamic design segments the input audio into frames such that:

$$s_p(n) = s\left(n; \frac{Mp}{2} + M - 1\right), \tag{7.12}$$

where $p = 0, 1, 2, \ldots$ is the frame index. Under this windowing scheme, s_0 is on the interval $[0, M-1]$, s_1 is on the interval $\left[\frac{M}{2}, \frac{3M}{2} - 1\right]$, and in general, s_p is on the interval $\left[\frac{Mp}{2}, \frac{M(p+2)}{2} - 1\right]$.

We define $\theta_p(n)$ as the impulse response we calculate from frame s_p. As a result of the overlap between consecutive frames, for an input voice signal of length N and n on the interval $\left[\frac{M}{2}, \left\lceil\frac{2N}{M}\right\rceil\frac{M}{2} - 1\right]$, there are two impulse responses corresponding to time n. We define the impulse response h_n at time n to be the linear interpolation of the two calculated frame impulse responses if n is on this interval:

$$h_n(v) = \begin{cases} \frac{M - 2\alpha(n)}{M}\,\theta_{p(n)-1}(v) + \frac{2\alpha(n)}{M}\,\theta_{p(n)}(v), & \frac{M}{2} \le n \le \left\lceil\frac{2N}{M}\right\rceil\frac{M}{2} - 1 \\ \theta_{p(n)}(v), & otherwise \end{cases}. \tag{7.13}$$

$p(n)$ is a function which maps a sample to its corresponding frame index, and $\alpha(n)$ is a function which maps a sample to its relative index in the overlap region:

$$p(n) = \left\lceil\frac{2n}{M}\right\rceil. \tag{7.14}$$

$$\alpha(n) = n - \frac{Mp(n)}{2}. \tag{7.15}$$

Because the system we seek to estimate $\mathcal{H}$ is causal, $\theta_{p(n)}(v) = 0,\ v < 0, \forall n$, and it follows that $h_n(v) = 0,\ v < 0, \forall n$. By definition, h_n is the time variant impulse response of our system $\widehat{\mathcal{H}}$, and therefore $\widehat{\mathcal{H}}$ is also causal. For each frame s_p of length M, our design calculates

the impulse response θ_p to have length M as well, and it follows that h_n also has length M. Since the system is causal, the output y at sample n_0 does not depend on the input $g(n)$ for $n > n_0$, and since the impulse response has fixed length M, the output y at sample n_0 also does not depend on the input $g(n)$ for $n < n_0 - M + 1$. From these properties, we derive an expression for the output $y(n)$ in terms of the input $g(n)$:

$$\begin{aligned} y(n) = {} & g(n-M+1)h_{n-M+1}(M-1) + g(n-M+2)h_{n-M+2}(M-2) \\ & + g(n-M+3)h_{n-M+3}(M-3) + \ \ldots + g(n)h_n(0). \end{aligned} \tag{7.16}$$

Equation 7.16 has the closed form:

$$y(n) = \sum_{k=n-M+1}^{n} g(k)h_k(n-k). \tag{7.17}$$

Our dynamic design uses the equation 7.17, a modification of discrete-time convolution with different limits of summation and time variant impulse response to synthesize the output $y(n)$ directly in the time domain.

7.5 Design Summary

We create a static and dynamic implementation of the talk box. The static implementation assumes that the vocal tract has an invariant impulse response for the duration of the voice signal $s(n)$ while the dynamic implementation relaxes this assumption. Our static implementation computes the complex cepstrum of the voice input $s(n)$ (4.3) and applies a low quefrency lifter $l(n)$ (7.3) to estimate the vocal tract frequency response $\Theta(\omega)$ (7.4). We preprocess $\Theta(\omega)$ to attenuate low frequency gain (7.10) to more closely mimic the behavior of the physical talk box. This overall design implements an LTI system with impulse response $\theta(n)$. We synthesize the output $y(n)$ by convolving the input instrument signal $g(n)$ and $\theta(n)$, and we perform this operation as multiplication in the frequency domain, which is less computationally expensive than direct convolution.

The dynamic implementation segments $s(n)$ into frames of length M (7.5) with 50% overlap between consecutive frames s_p and s_{p+1} (7.12). For each individual frame s_p, we calculate the complex cepstrum (7.7) and apply a low quefrency lifter $l(n)$ (7.8) to estimate the frequency response $\Theta_p(\omega)$. We normalize each $\Theta_p(\omega)$ in the frequency domain (7.11) to attenuate low frequency gain and ensure that the average magnitude of $s(n)$ within a frame does not directly affect the average magnitude of output. Due to the 50% overlap between frames, we calculate two impulse responses corresponding to each sample except at the beginning and end of $s(n)$. We linearly interpolate between these two impulse responses to determine $h_n(v)$, the system's time variant impulse response (7.13). The dynamic implementation synthesizes output $y(n)$ directly in the time domain using a modified discrete-time convolution summation which accounts for the system's causal, finite and time variant impulse response. We summarize the dynamic design in figure 7-1.

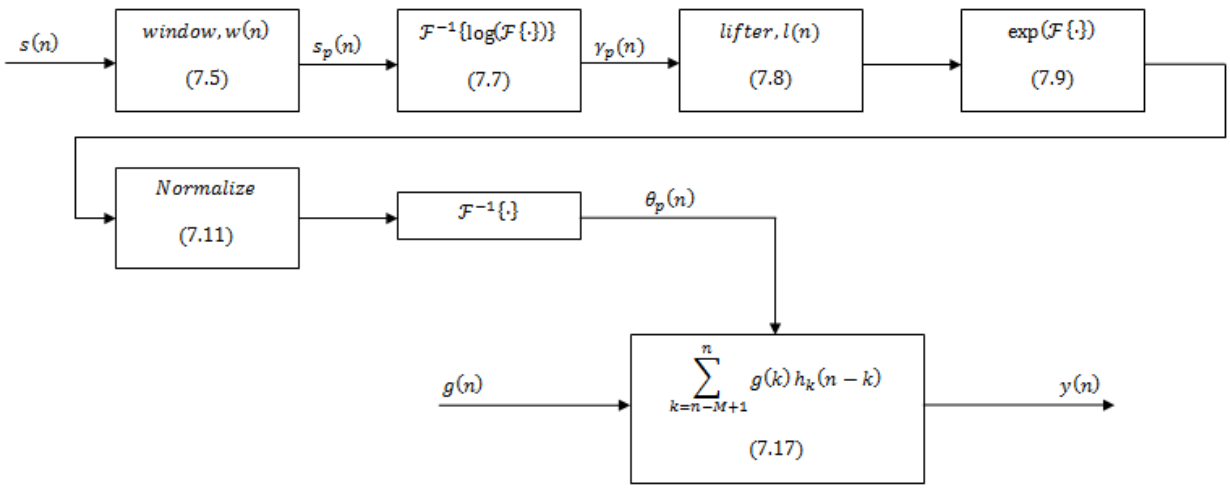

Figure 7-1: Block diagram of the digital talk box (dynamic implementation)

8. Performance Results

8.1 Performance on Isolated Vowels

We test our designs using clips of isolated vowel sounds for $s(n)$ and the C-262 guitar note for $g(n)$. Each $s(n)$ is 1 second in duration, sampled at 44100 Hz, and preprocessed outside of our design so that the guitar and voice begin at the same time. $g(n)$ is also 1 second in duration and sampled at 44100 Hz. For each synthesized clip $y(n)$, we lifter the signal with $n_c = .005f_s$ and

compute its Fourier transform, which we show for the ‘e’ vowel in figure 8-1. Since the physical talk box does not impress the F1 formants as consistently as F2 and F3, we only consider the effect of the digital implementations on the frequencies of the F2 and F3 peaks, and we report the results in table 8-1.

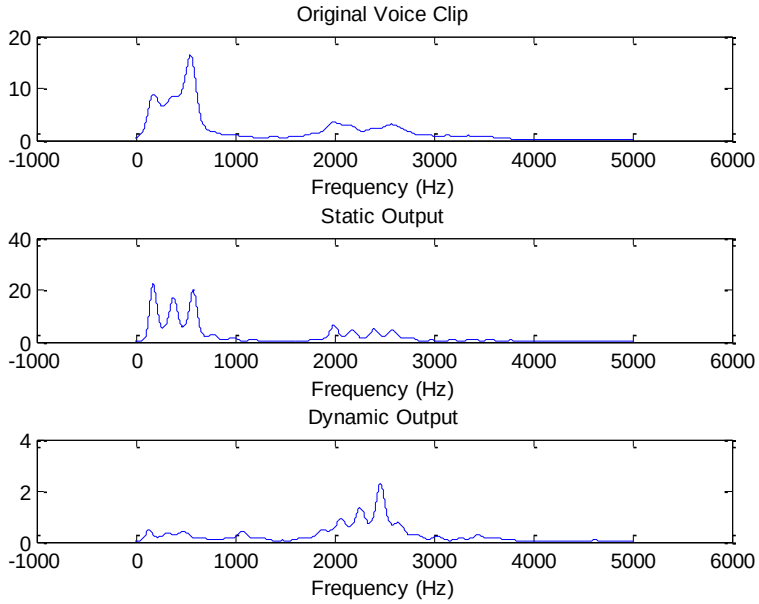

Figure 8-1: Fourier transform of liftered signals of the ‘e’ vowel sound (a – top) from human speech (b – bottom) from static synthesis (c – bottom) from dynamic synthesis

The static implementation output has an average relative error in frequency (with reference to the formant peaks of $s(n)$ in table 6-1) of 3.0% for F2 and 3.5% for F3. The dynamic implementation output has an average relative error in frequency of 3.7% for F2 and 3.2% for F3. Recall that the physical talk box has an average relative error in frequency of 9.33% for F2 and 6.21% for F3. Our design consistently creates the F2 and F3 peaks in the output. The lower relative errors for our designs do not mean that the digital implementations are better at impressing formants onto a instrument input $g(n)$ because we use the same $s(n)$ to create the output and to calculate the relative error. When we calculate the relative error for the physical talk box, the reference signal $s(n)$ is an attempted replication of the vocal tract during the production of the talk box output. It is nearly impossible for a speaker to produce the same sound

twice in exactly the same way [4], and we expect this to contribute to the error of the physical talk box.

Table 8-1: Frequencies of F2 and F3 formant peaks of vowels for static and dynamic implementation of the talk box

Vowel	Implementation	F2 (Hz)	F3 (Hz)
e (hate)	Static	1992	2401
	Dynamic	2061	2463
@ (at)	Static	1605	2586
	Dynamic	1689	2485
i (eve)	Static	2396	2578
	Dynamic	2402	2584
E (met)	Static	1786	2588
	Dynamic	1704	2490
R (bird)	Static	1389	1793
	Dynamic	1348	1763
I (it)	Static	1995	2593
	Dynamic	1952	2722
o (obey)	Static	1189	2381
	Dynamic	1189	2379
u (boot)	Static	1182	2382
	Dynamic	1065	2345
a (father)	Static	1195	2390
	Dynamic	1094	2482
c (all)	Static	2786	3570
	Dynamic	2754	3560
U (foot)	Static	1195	2377
	Dynamic	1103	2283
A (up)	Static	1372	2571
	Dynamic	1375	2564

There are qualitative imperfections in the synthetic outputs from both implementations compared to the output of the physical talk box. The output from the static design contains an echo throughout the duration of the signal, and the gain of the static design is also too small in the high frequency range relative to the gain in the low frequency range. The output from the dynamic design contains noticeable transitions between the frames even with normalization. The transitions do not appear to affect the intelligibility of the talk box signal.

8.2 Performance on Words Containing Non-Vowels

We also test the performance of our designs for $s(n)$ as clips of words with non-vowels. Each $s(n)$ is 1 second in duration, sampled at 44100 Hz, and preprocessed outside of our design so that the instrument and voice begin at the same time. The vocal tract impulse response is not invariant throughout the duration of $s(n)$, and we do not expect the static implementation to perform much more poorly than the physical talk box. We do not calculate relative error for performance on these words, since not all phonemes have well-defined formants, and the formant peaks change in time. While a spectrogram is appropriate for displaying frequency over time, it requires a tradeoff in frequency resolution for time resolution. We cannot calculate the spectrogram with enough resolution in frequency and time to precisely determine the formant peaks.

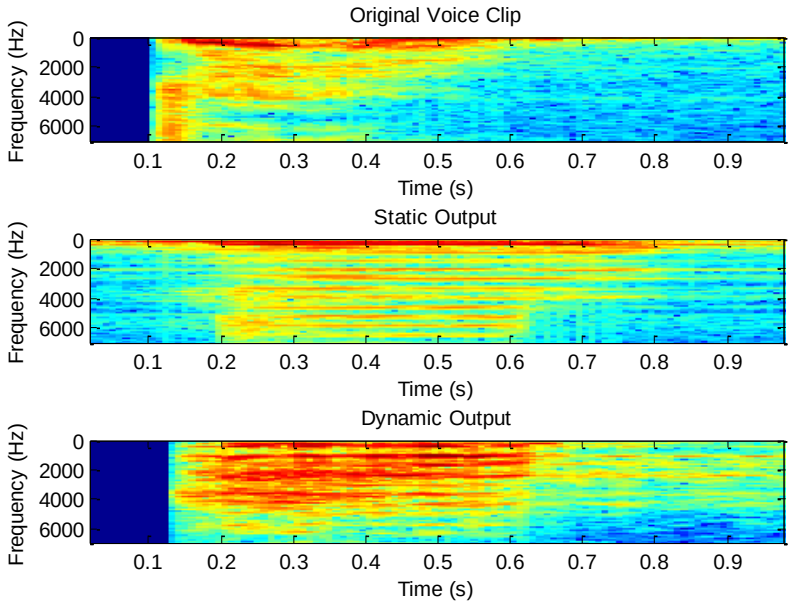

Figure 8-2: Spectrogram of the word "Jam" (a – top) from original voice clip (b – middle) from static synthesis (c – bottom) from dynamic synthesis

When the signal $s(n)$ has time-varying frequency content, the static implementation is unsuccessful in creating an output with similar frequency content. In figure 8-2a, we clearly see the frequency content difference between the affricate /J/ at 0.1 seconds and the vowel /@/ at 0.15 seconds. Figure 8-2b shows that the frequency content of the static output does not seem to vary with the input voice signal, and any change in the spectrum over time most likely comes

from the changing spectrum of the instrument input $g(n)$. Figure 8-2c shows that the frequency content of the dynamic output seems to vary, but it is not clear that the output's spectral changes over time correspond to the spectral changes of $s(n)$. For example, we do not observe the affricate /J/ frequency pattern in the spectrogram of the dynamic output.

Qualitatively, the static output has a more perceivable echo and sounds unintelligible for the voice input "jam". In the case of the word containing the unvoiced fricative (/f/ in "four"), the voiced fricative (/v/ in "vote"), and the diphthong (/O/ in "out"), the static output contains intelligible portions of the primary vowel sound of the word. In cases of the other phonetic categories, the output is unintelligible, similar to the output for the word "jam". For input signals from all the non-vowel phonetic categories, the static output also contains part of the pitch component of $s(n)$. From its performance on words and non-vowels, we conclude that the static design is inappropriate for inputs $s(n)$ with varying frequency content and is a poor implementation of the talk box.

The dynamic design successfully and consistently produces the affricate (/J/ in "jam"), semi-vowel glide (/j/ in "you"), voiced fricative (/v/ in "vote"), voiced plosive (/b/ in "bait") and diphthong (/O/ in "out") in the . It does not successfully and consistently produce the unvoiced fricative (/f/ in "four"), unvoiced plosive (/p/ in "pay") and nasal (/m/ in "me"). The physical talk box is only unsuccessful in producing the unvoiced fricative and unvoiced plosive, and in both cases the physical talk box output contains the voiced analog of unvoiced phoneme. Since the unvoiced phonemes do not involve glottal pulse excitation, its production deviates from our general model of speech production, and we do not expect to reliably calculate the vocal tract frequency response from audio of these phonemes. The nasal phoneme may cause problems in calculating the impulse response because its formants are lower in energy and result from resonance of the nasal cavity [6].

8.3 Performance of Dynamic Implementation for Longer Duration Inputs

Finally, we test the dynamic design with longer duration input signals $s(n)$ and $g(n)$, to simulate more realistic input conditions. $s(n)$ is the vocal scale ("Do Re Mi Fa So La Ti Do") played at 60 beats per minute and 120 beats per minute. $g(n)$ is the C major scale beginning on the note C-262 played at 60 beats per minute and 120 beats per minute.

From figure 8-3, we observe that both the analog and digital talk box outputs for $s(n)$ equal to the 60 beat per minute vocal scale and $g(n)$ equal to the 60 beat per minute C major scale have spectral changes which correspond to spectral changes in the input voice signal. The spectrogram for the analog talk box output (8-3b) appears smoother in time than the spectrogram for the digital talk box output (8-3c). Qualitatively, we can hear and distinguish every transition between consecutive frames in the digital implementation's output. As with shorter duration signals the transitions do not seem to affect the output's intelligibility. We can perceive all individual phonetic units in the digital talk box output with inputs at 60 beats per minute. This includes the nasal /m/ in "mi" which the implementation could not produce with the isolated, short duration voice input. We observe similar results for the 120 beats per minute analog of the inputs and output.

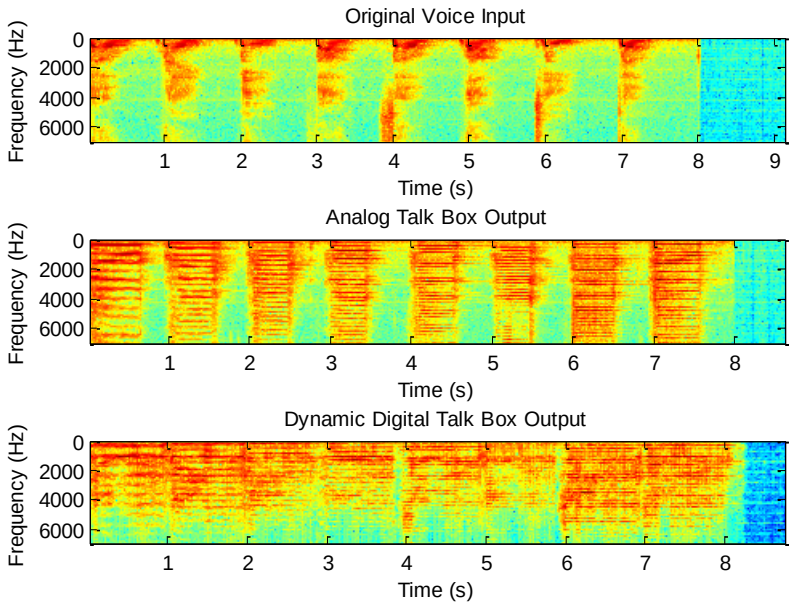

Figure 8-3: Spectrogram of vocal scale at 60 beats per minute from (a – top) from original voice input (b – middle) from analog talk box output (c – bottom) from digital talk box output

From the standpoint of producing output with intelligible phonetic units, the digital design can reproduce the performance of the analog talk box. In the MATLAB environment, the digital talk box implementation requires time greater than the duration of the inputs to compute the output. Part of the expense in computation could result from inherent overhead in the

environment. We can decrease the amount of required computation for the digital talk box by reducing the sampling rate of the input signals. At the current sampling rate of 44100 Hz, the digital design cannot produce output quickly enough for true real-time applications.

9. Conclusions and Further Considerations

The analysis of output from the analog talk box shows that the device impresses formant peak frequencies of the vocal tract impulse response onto the input instrument signal to give rise to intelligibility. Specifically, the device creates the F2 and F3 formants with low relative error in frequency with reference to the F2 and F3 formants of the vocal tract response, while it impresses the F1 formant with higher relative error. Recall that F2 and F3 are more important for signal intelligibility and that F1 carries the signals pitch content. We therefore expect the observed behavior of the analog talk box, as the output is intelligible but the device does not use the vocal tract to alter the pitch of the instrument input. In addition, the analog talk box intelligibly reproduces all voiced phonemes which occur in words containing multiple phonetic units. The device successfully captures the time variant vocal tract impulse response and produces intelligible output from vocal tract input of arbitrary length containing many different phonemes.

We seek to create a digital implementation of the talk box which takes as input a speech signal $s(n)$ and an instrument signal $g(n)$ and F2 and F3 formants from $s(n)$ onto the output with the same pitch content as $g(n)$. We use cepstral liftering to de-convolve the excitation source $e(n)$ from the vocal tract impulse response $\theta(n)$ which, according to our model of human speech production, comprise $s(n)$. Under the assumption that the vocal tract response is time invariant for the duration of $s(n)$, we first create a static implementation which uses the entirety of the speech signal to calculate the impulse response with maximum frequency resolution. Relaxing the assumption of vocal tract time invariance, we also create a dynamic implementation which considers $s(n)$ in frames with 50% overlap between frames. The dynamic design considers the vocal tract impulse response to be time variant and calculates the impulse response of the system $h_n(v)$ to be the linearly interpolation in time of the normalized vocal tract impulse responses in the corresponding frame overlap region. The dynamic design trades of resolution in frequency for resolution in time to calculate a time-variant impulse response.

Both the static design and dynamic design successfully impress the F2 and F3 formant peaks from speech input of isolated vowels onto the output. The static design output contains some of the pitch components from $s(n)$ and has a perceivable echo throughout the duration of the signal. The dynamic design output has noticeable transitions between frames even with normalization. The outputs from both implementations are intelligible, and this suggests that the tradeoff in frequency resolution of the dynamic implementation does not perceivably affect the intelligibility of its output. The static design fails to produce intelligible output for speech input of words containing other phonetic units. The dynamic design performs very similarly to the analog talk box in terms of performance on non-vowel phonemes. In our performance tests of voice inputs of single words, the dynamic design successfully produces all phonetic categories that the analog talk box can produce with the exception of the nasal phonetic units. From our performance tests using longer duration voice inputs, we observe that the dynamic design does successfully produce nasal phoneme, which differs from the implementation's behavior in the short duration voice input tests.

Our dynamic digital design has the ability to produce all phonetic categories the analog talk box can produce. The performance of the digital design seems to depend on the alignment of the inputs $s(n)$ and $g(n)$. Additionally, because this design has limited resolution in time, taking a frame every 1500 samples in a signal sampled at 44100 Hz, the time at which the phoneme occurs within a frame also affects the system's ability to produce that phoneme in the output. Whereas improper alignment of the inputs in the tests using 1 second clips of single words may cause the digital implementation's inability to intelligibly produce the nasal phoneme, proper alignment of the inputs allows the digital implementation to successfully produce this exact phoneme in the test using longer duration voice clips. The digital talk box seems less robust than the analog talk box to disturbances in the timing of the inputs.

The dynamic design also simulates processing in real time since it considers $s(n)$ in successive frames. The estimation of the vocal tract impulse response requires two FFT calculations and two inverse FFT calculations for each frame. In MATLAB however, the overall processing time is longer than the duration of the input $s(n)$. Since the total computation time depends on the number of samples per frame and the total number of frames, we can directly reduce the processing time by reducing the sampling rate. A current limitation of the design is the long calculation time, and in order to truly perform the processing in real time, we must

further optimize the algorithm. Another consideration for future work is to remove the noticeable transitions between the frames. Finally, efforts toward physical implementation should consider the alignment of instrument and speech inputs and continue to improve the system's time resolution in order to more reliably produce phonemes from non-stationary speech. Overall our design demonstrates that it is possible to digitally implement the talk box effect using frames of speech input to estimate a time variant vocal tract impulse response.

References

1. Fortner, Stephen. (2011, May). Talkbox 101. [Online]. Available: http://search.proquest.com/docview/862562336?accountid=13314

2. Anderton, Craig. (2008, May). Vocoder Basics. [Online]. Available: http://search.proquest.com/docview/199528533?accountid=13314

3. Gold, Ben et al. Speech and Audio Signal Processing, 2nd ed. Hoboken, NJ: Wiley, 2011.

4. Lingard, R. Electronic synthesis of speech, 1st ed. Cambridge, United Kingdom: University Press, 1985.

5. McLoughlin, Ian. Applied Speech and Audio Processing, 1st ed. Cambridge, United Kingdom: University Press, 2009.

6. Deller, John R et al. Discrete-Time Processing of Speech Signals, 1st ed. New York: Macmillan, 1993.