Echo: A Voice Recognition and Playback System

Blake Oberfeld, Davy Huang, Allison Ramsey, Arjun Patel, Kate Ryan

Abstract

The manipulation of signals is vital to many disciplines and necessary to understand complex phenomena. In the music industry speech manipulation is used to autotune so that artists can sing a wider range of pitches. In telecommunications, speech manipulation is used to disguise speech, so that the identity of a speaker is protected. Signal processing is also vital to ensure clarity in hearing aids so that speech is not clipped or muffled. Delving into signal processing, the purpose of this *Echo* project was to design and implement a digital channel vocoder by exploring speech coding, determining voicing and pitch, and "echoing" resynthesized speech with varied parameters to generate male, female, monotone, whispered, and pitchless speech.

Echo relies on the source-filter model of speech, such that speech is considered as the output of a time-varying filter excited by a source signal, where both source and filter are controlled almost independently by the speaker. It gives real-time analysis of the source and filter aspects of speech. *Echo* analyzes the voicing source by further breaking it down into a fundamental frequency, pitch vector and envelope matrix. These three components are interpreted from the subject's speech by *Echo*. *Echo* then modifies the pitch vector according to user input and synthesizes speech. By making this change, the characteristics of the speech signal are changed and interpreted as different sounds. *Echo* is used to understand the role of source and/or filter in monotone, whispered, male, and female utterances.

The voice echo and modification device uses the fundamental frequency of a subject's speech and other components to manipulate recorded speech. The device is able to change male speech into female speech and vice versa. There are slider features that allow more specific manipulation of the recorded speech. The manipulated audio is automatically outputted to speakers so that the effects can be heard. The voice echo and modification device manipulates speech signals, producing sounds the user desires in a simple and intuitive interface.

Methods

As a digital device, *Echo* was executed in MATLAB and rendered user-friendly through GUI display. *Echo* characterizes and processes speech signals through data acquisition, channel vocoder analyzer, channel vocoder synthesizer, and playback functionalities (Figure 1).

Data Acquisition

First, a speech signal is recorded and processed by the device. *Echo* collects speech signals with a quality equivalent to compact disks (CDs) at a sampling rate of 44.1kHz and a 16-bit depth. However, to minimize the lag accompanied by processing such a large volume of recorded data, the speech signal is resampled at 8kHz and an antialiasing filter (4kHz cutoff) is applied so that no frequencies are erroneously identified.

Channel Vocoder Analyzer

Applying the source-filter model of speech, *Echo*'s channel vocoder analyzer characterizes the pitch source and the time-varying filter of the utterance. The vocoder analyzer

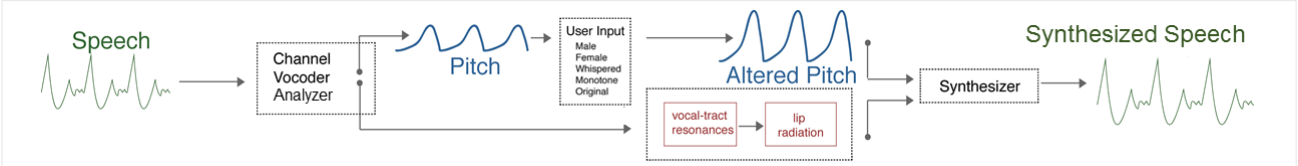

Figure 1: Block diagram overview of Echo. Echo receives a speech signal as input, analyzes the signal by using a channel vocoder oriented around the source-filter model of speech. Echo then receives user input and modifies the pitch vector of the speech in its channel vocoder synthesizer. The output is synthesized speech and graphical representations of the modifications Echo made.

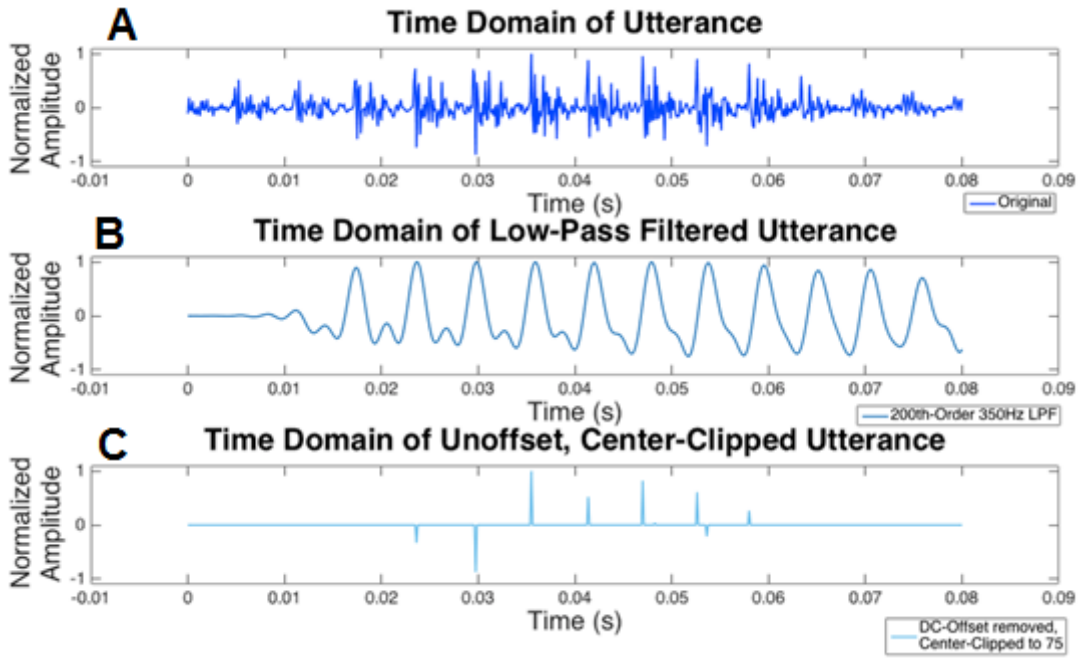

Figure 2: Time domain graphs of original utterance frame (A), frame filtered at 200th-order 350Hz LPF (B), and frame with DC offset removed and center-clipped with 75% threshold (C).

processes signals longer than two seconds into small overlapping frames around 30ms in length. This frame length is actually determined by the decimation rate (10) and the number (18) of 65th order frequency bands of 200 Hz bandwidth. These parameters were chosen based on research and field standards separately suggested by Talkin[1] and Greenberg[2].

To distinguish the voicing source, the pitch detector portion of the analyzer each frame of data is passed through a 200th order 350 Hz low pass filter to remove high frequency noise (Figure 2A,B). Because human adults speak within the range of 85-255Hz, the filter will not distort the relevant pitch values[3,4]. Next, any DC offset is removed from the signal by subtracting the mean, and the signal is center clipped with a 75% threshold in order to suppress peaks from the vocal tract[1,2] (Figure 2C). Once processed, the signal is autocorrelated with itself in order to identify the fundamental frequency in the signal implied by its harmonic frequencies (Figure 3). The zeroth lag of the correlation corresponds to 0Hz, while the next highest peak after this corresponds to the fundamental peak of the frame. The accuracy of this pitch detection was checked in Table 1. *Echo* automatically determines this fundamental peak and then finds the associated fundamental frequency—the effective pitch of the signal in that frame.

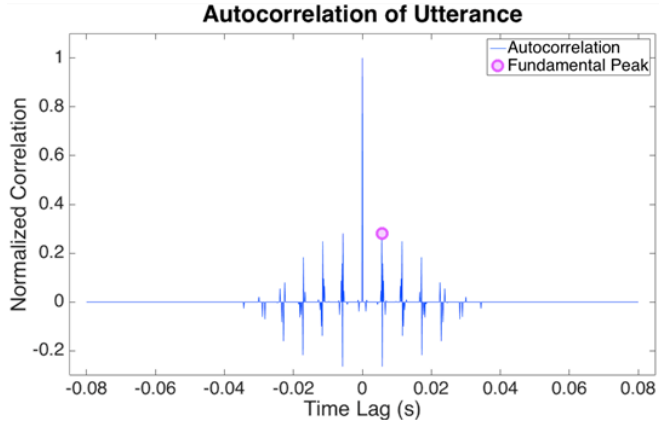

Figure 3: Autocorrelated utterance frame with demarcated fundamental peak.

[1] Talkin, D. (1995). Chapter 14: A Robust Algorithm for Pitch Tracking (RAPT), Speech Coding and Synthesis, Elsevier Science B.V.(p495-516).
[2] Greenberg, J. (2007). Speech Coding and Laboratory 2, Biomedical Signal and Image Processing MIT Open Courseware. Cambrigde: Harvard-MIT Division of Health Sciences and Technology HST.582J.
[3] Titze, I.R. (1994). Principles of Voice Production, Prentice Hall (p. 188).
[4] Baken, R. J. (1987). Clinical Measurement of Speech and Voice. London: Taylor and Francis Ltd. (p. 177).

Iterating over N frames of the utterance, this ultimately generates an N-dimensional pitch vector that acts as the speech source.

To distinguish the filter portion of speech, the analyzer characterizes band envelope values for each frame such that the response of the time-varying filter is represented in an NxB matrix of band envelope values, where N is the number of frames and B is the number of frequency bands.

Channel Vocoder Synthesizer and Playback

Before synthesizing speech, the pitch source can be changed by user input in the GUI so that male, female, monotone, whispered, or pitchless speech is generated. As motivated by Table 1 and because male and female voices are voiced between 85-180Hz and 165-255Hz respectively[3,4], scaling the pitch vector by three-quarters appropriately shifts a female voice into the frequency range for males, while a scale of five-thirds shifts male voice into the high frequency range suitable for women. Monotone voice passes a constant-valued 100Hz pitch vector to the filter. Moreover, whispered voice is generated with white noise, an aperiodic signal with random frequencies of equal intensities, while pitchless speech is synthesized from a pitch vector of zeros. Speech signal is then produced in the channel vocoder synthesizer by running the pitch vector through the time-varying filter as characterized by the analyzer. The synthesized speech signal can be heard and seen in spectrograms and time-domain plots on the GUI.

Table 1. *Echo* generated voice pitch for stressed vowels from different subjects, accurately found within the accepted ranges for adult female and male voice (165-255Hz and 85-180Hz respectively[3,4]).

Subject	Fundamental Frequency (Hz)	Subject	Fundamental Frequency (Hz)
Female 1	216.216	Male 1	112.676
Female 2	210.526	Male 2	115.942
Female 3	205.128	Male MIT Recording[2]	114.285

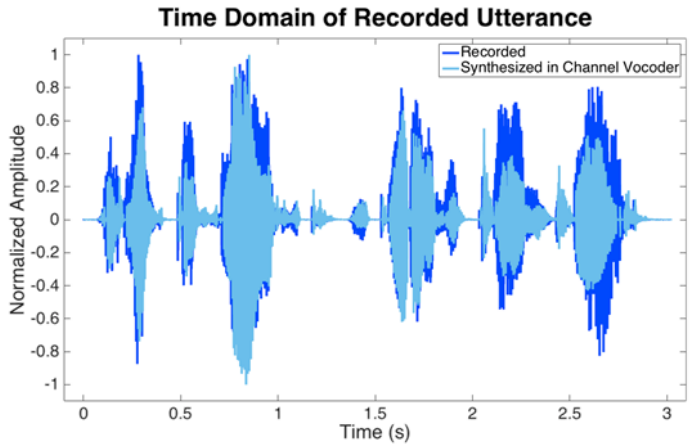

Figure 4: Original and synthesized utterances for "The empty flask stood on the tin tray."

Results

The *Echo* device was successfully implemented and is able to characterize the pitch source and time-varying filter in speech, as well as to generate male, female, monotone, whispered, and pitchless speech with both accuracy and precision.

Resynthesizing Unaltered Speech

Assessing against the gold standard of resynthesizing exactly the original recording, it is clear that the device performs very well and that these results are reasonable. Examining the processing of an utterance from the MIT sound library[2], the synthesized speech and the original speech closely match for the majority of the signal, but especially at the end, around 2.25s and later, there is a slight discrepancy between the signal magnitudes (Figure 4). The high performance, with some notable discrepancies, is confirmed in highly similar spectrograms for these utterances, though the synthesized original does seem have been slightly smoothed in intensity (Figure 5A,B). Sources for these discrepancies will be examined in the *Discussion.*

Synthesizing Gender-Based Speech

Examining gender speech both validates the functionality of the device and offers insight into differences between masculine and feminine speech signals. Comparing the spectrograms of the synthesized speech to the original recording, it is clear that, as described in *Methods*, the artificial male speech has a pitch that has been scaled by a factor of three-quarters, and this is seen in the resonant frequencies of the vowel phonemes that have become about 75% of their original values (Figure 5C). Similarly, because synthetic female speech has a pitch that has

been scaled by a factor of two, the resonant frequencies of the vowel phonemes have become about 200% of their original values (Figure 5D). Quantitatively, in the phenome at 0.448s, the lowest resonant frequency in the original utterance is 187.5Hz; while in the synthesized speech, it is 140.6Hz and 312.5Hz for male and female respectively. This corresponds to 0.749 and 1.667 scale factors respectively, evidencing the successful, albeit imperfect, implementation of *Echo*.

Synthesizing Other Speech Forms

In this project, monotone speech connotes a constant pitch value of 100Hz. Qualitatively, this utterance becomes robotic sounding—as if pitch or intonation cannot be expressed. Comparing the spectrogram of the synthesized monotone speech to the original recording, the effect of changing the voicing source this way is clear: the resonant frequencies of the vowel phonemes have been replaced with continuous resonances of 100Hz, as seen in the banding (Figure 5E).

Meanwhile, by passing white noise through the filter, whispered speech appears to be unvoiced in the spectrogram—in that there is no periodic structure recorded in the spectrogram (Figure 5F).

As to whether passing a zero pitch vector through the time-varying filter for speech would also effectively generate a whispered utterance, pitchless voice sounds and appears identical to whispered voice (Figure 5F). This is reinforced by the curriculum of Greenberg[2]. However, we would like to note that the audio recordings of both whispered and pitchless speech do not seem as unvoiced as the spectrograms imply. In this way, looking closer at both whispered and pitchless spectrograms, we believe that some smoothing occurs, increasing the intensity of background noise and decreasing the specificity of the power distribution in the signal.

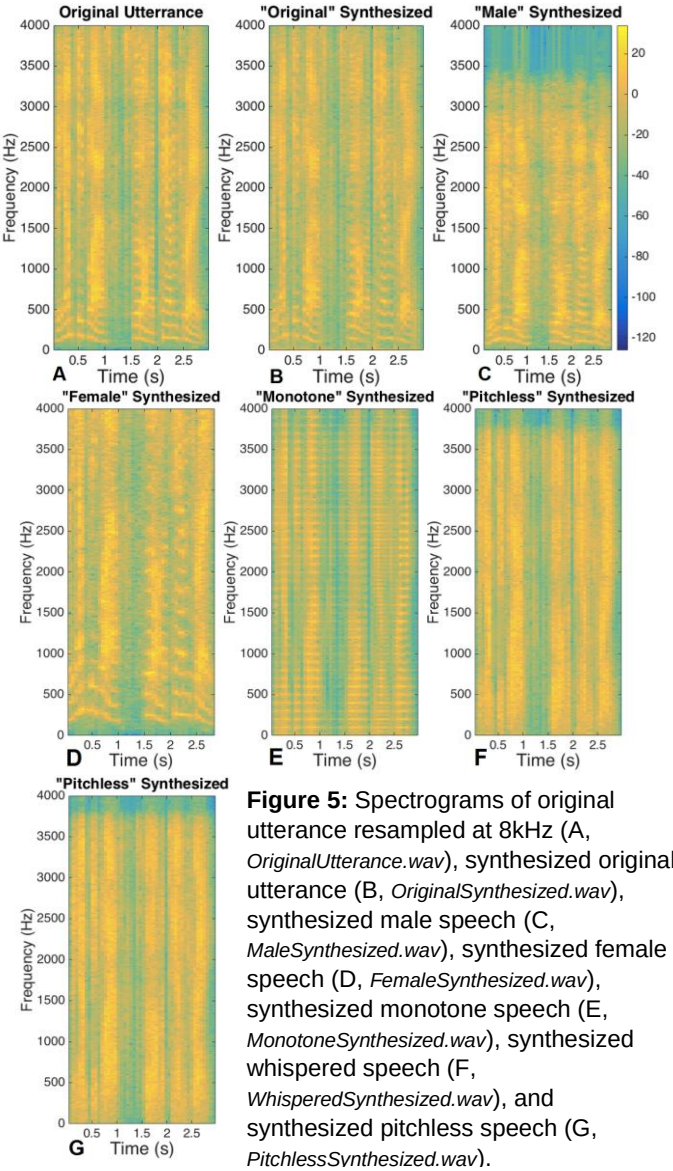

Figure 5: Spectrograms of original utterance resampled at 8kHz (A, *OriginalUtterance.wav*), synthesized original utterance (B, *OriginalSynthesized.wav*), synthesized male speech (C, *MaleSynthesized.wav*), synthesized female speech (D, *FemaleSynthesized.wav*), synthesized monotone speech (E, *MonotoneSynthesized.wav*), synthesized whispered speech (F, *WhisperedSynthesized.wav*), and synthesized pitchless speech (G, *PitchlessSynthesized.wav*).

Discussion

Device Performance

Ultimately, *Echo* performs well when compared to the gold standard measurements of accuracy and precision for voice modification devices. The parameters for success are accuracy—that synthesized speech closely approximates the original utterance—and precision—that the results are reproducible across iterations of the program. In this way, accuracy was checked throughout the course of the design and implementation of *Echo* by analyzing the degree of similarity between time-domain plots and spectrograms of the original speech and the synthesized "original" speech, as well as between the sounds of the two speech signals. As examined for above and as reiterated time and time again by real-time recording data, the synthesized "original" speech very closely approximated the original utterance.

Moreover, these results were easy repeated when, in real-time recordings, the subjects repeated the phrase “the empty flask stood on the tin tray” multiple times in order to ensure success of *Echo* was time-invariant. There was high conservation of results across all tests. *Echo* works ideally with input signals from 80Hz to 320 Hz, and this is reasonable because that is the range of adult human voice[3,4].

The discrepancies in the synthesized signal are due to sources of error in *Echo.* One source of error includes the assumptions of the source-filter model of speech—presuming the independence of source and filter. This assumption should be relatively sound though because the source-filter model is readily accepted and frequently used in the field. In fact, discrepancies observed in the figures and in resynthesized speech that “did not sound quite right” are best accounted for by the resampling of the data, the non-ideality of a 200th order LPF, and the 30ms windowing approach taken to create frames. Even so, the discrepancies between the original and manufactured speech are quite minimal, and *Echo* is able to operate with precise, accurate results that are close to, albeit not exactly, the same as the original speech.

Conclusions on Speech Forms

Although analyzing of the synthesized gender speech signals validated the functionality of the device, it also offers insight into differences between masculine and feminine speech signals. Perhaps, most significantly, we noticed that, although the device successfully changed the pitch of the utterance, this did not always render a convincing gender-swap. This is explained by the fact that gender voice may be largely frequency dependent—higher frequencies for females and lower for males—but that there also exists some difference in the time-varying filter between genders. This may include differences in lip radiation, vocal tract resonances, cadence, speech patterns, or even anatomical differences relevant to speech production. In order to produce even more convincing gender swaps, we hope that future development of the project will not only assess the effect of changing pitch source, but also filter in the source-filter model of speech.

In comparison, the convincing synthesis of unvoiced speech from both the “whispered” and “pitchless” forms suggests that unvoiced speech relies exclusively on the filter aspect of speech. In pitchless speech, this is because passing zeroes still generates comprehendible speech—and more so, this speech sounds very much like a whisper. The equivalence of these two signals can be explained by the idea that “whispered” signal acts to pass white noise through the filter in order to generate unvoiced speech, while passing the zero vector in the "pitchless" signal through the filter has the same effect of forcing white noise for the source and generating unvoiced speech[2].

Future Work

These conclusions and *Echo’s* functionality motivate further research and improved design not only of the channel vocoder, but in all applications of the source-filter speech model of speech, including hearing devices, voice changers and automated gender recognition algorithms. As a stepping stone, *Echo* could easily be modified and used for voice recognition against both gender and source (i.e. human or “intelligent assistant” like Siri). This should also include balancing resampling at higher rate while minimizing lag, incorporating increasingly ideal filters, and improving decimation rate and band processing.

Having accomplished the design and implementation of the *Echo* project—having explored speech coding, determined voicing and pitch, and “echoed” resynthesized speech with varied parameters to generate male, female, monotone, whispered, and pitchless speech—we are excited to continue our study of signals and systems, especially in the context of biological applications.