Speech Enhancement Using Synthesis and Adaptive Techniques

James Fosburgh and Scott Bradley

AME Department, University of Rochester

Objectives

The goal for this project was to create our own algorithm that would achieve speech signal enhancement using the following process:

- Detect phonemes of original speech signal
- Synthesize clean speech estimate from detected phonemes
- Use synthesized clean speech as target for adaptive noise estimation
- Subtract estimated noise from original noisy signal

Introduction

Speech signal enhancement is an important task in audio DSP, especially with the increase in popularity of voice controlled applications. Our proposed method for speech signal enhancement is based on the observation that humans are better at understanding speech they are more familiar with, such as the voices of family members or close friends, than speech they are unfamiliar with. We propose that computers are the same: if they have an estimate of what they think the voice is saying, they can use this estimate to enhance the quality of the original speech signal.

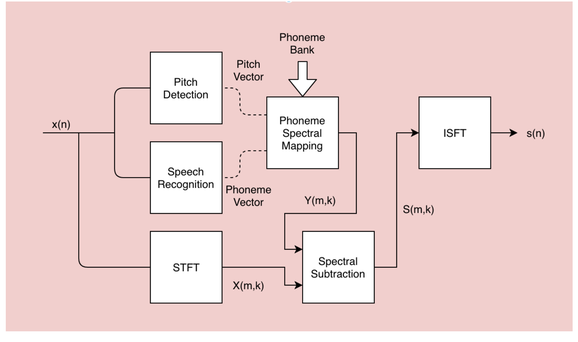

Figure 1: Block diagram of the proposed method.

Approach

The steps of our implemented approach are as follows:

- Run phoneme detection on noisy speech signal using PocketSphinx (CMU)
- Run Yin pitch detection algorithm on noisy speech signal
- For each frame of original signal, look up detected pitch and phoneme
- Using phase vocoder, pitch shift detected phoneme from bank to match detected pitch
- Concatenate pitch-shifted phonemes to synthesize estimated speech signal
- Use synthesized speech as target signal to estimate noise in original signal
- Use spectral subtraction to remove estimated noise from original signal

Speech Synthesis

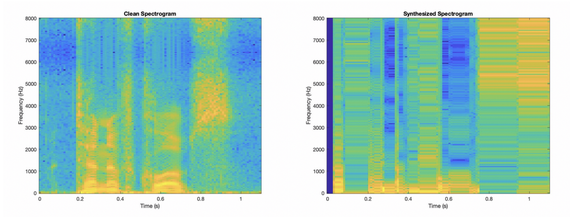

Figure 2: Spectrogram of clean speech vs. spectrogram of synthesized speech.

Concatenative synthesis is the process by which speech is synthesized by stringing together small portions of human speech. For our project, we have implemented unit selection synthesis, with phonemes being our smallest unit of sound. By combining the pitch vector returned by the Yin algorithm and the phoneme vector returned by PocketSphinx, we can first shift the pitch the phonemes from our recorded bank to match the input signal, then concatenate successive frames to produce a synthesized speech output.

Spectral Subtraction

Spectral subtraction can be described as the process of restoring a signal's magnitude spectrum through the subtraction of an estimated noise component. The noise is estimated in several different ways, and depends on the context of the implementation. In our implementation of spectral subtraction, we were able to use the other system components to create a more accurate noise estimation. By finding the difference between the two components, the original signal and the synthesized speech signal, we generate a continuous estimation of the noise component. This can be expressed as

$$S(m,k) = |X(m,k)| - N(m,k) \quad (1)$$

Where

$$N(m,k) = \beta N(m-1,k) + (1-\beta)P(m,k) \quad (2)$$

And

$$P(m,k) = max[0, |X(m,k)| - |Y(m,k)|] \quad (3)$$

Here, S(m,k) is the clean speech signal, X(m,k) is the noisy signal, Y(m,k) is the synthesized speech signal and N(m,k) is the estimated noise component. Beta is a smoothing factor that describes how quickly to change the noise component, and affects the performance of the system. We chose a beta value of 0.3 for our tests.
Looking at the results of the spectral subtraction section, we can see significant improvements in speech quality.

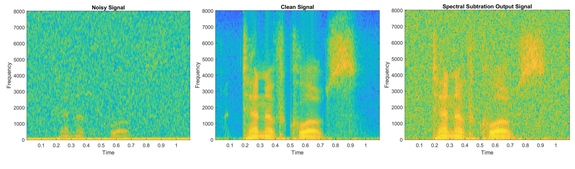

Figure 3: Spectral Subtraction Results.

Results

When using the original clean speech as the target signal during noise estimation, spectral subtraction shows significant improvement in quality of the speech signal.

As of now, our speech synthesis technique is not accurate enough to provide a target signal good enough to make our method of enhancement effective.

Conclusion

This implementation shows promise as the building blocks for a comprehensive speech enhancement system.

In order for this system to be fully implemented into a real-world application, there would need to be continued work on the training model of the phoneme bank.

The results of this system would be drastically improved by a more accurate synthesis signal, however this would leave the spectral subtraction component useless. Therefore, this system could be further improved by enhancing some of the components to the point where the other stages are no longer needed.

Future Work

In going forward, some of the work we plan to do is:

- Improve pitch shifting of speech
- Improve concatenative synthesis algorithm
- Improve quality of phoneme bank samples

UNIVERSITY of ROCHESTER