Prosody Modification for Neutral to Angry Speech Conversion

Guneesh Vats (2021122007)
Prashant Gupta (2020102030)
Rohan Madineni (2020102066)

1. Introduction

Speech is the natural mode of communication between human beings which contains multiple dimensions of information encoded within it. It is also an important outcome of the emotional state of the speaker. The emotional state of a speaker is accompanied by physiological changes affecting respiration, phonation, and articulation which manifests as and is mostly perceived through the prosodic or supra-segmental nature of the speech signal such as pitch, energy, and duration of the signal. Our objective is to synthesize the quality of a target emotion(anger) in that of neutral speech through modification of prosodic parameters.

Such modification has various applications. It can be used to add expressiveness to computer-generated speech which can make the speech more interesting to listen to. This can be used in computer-voiced storytelling for example. It can also be used in railway announcements and other circumstances where expressive speech is needed.

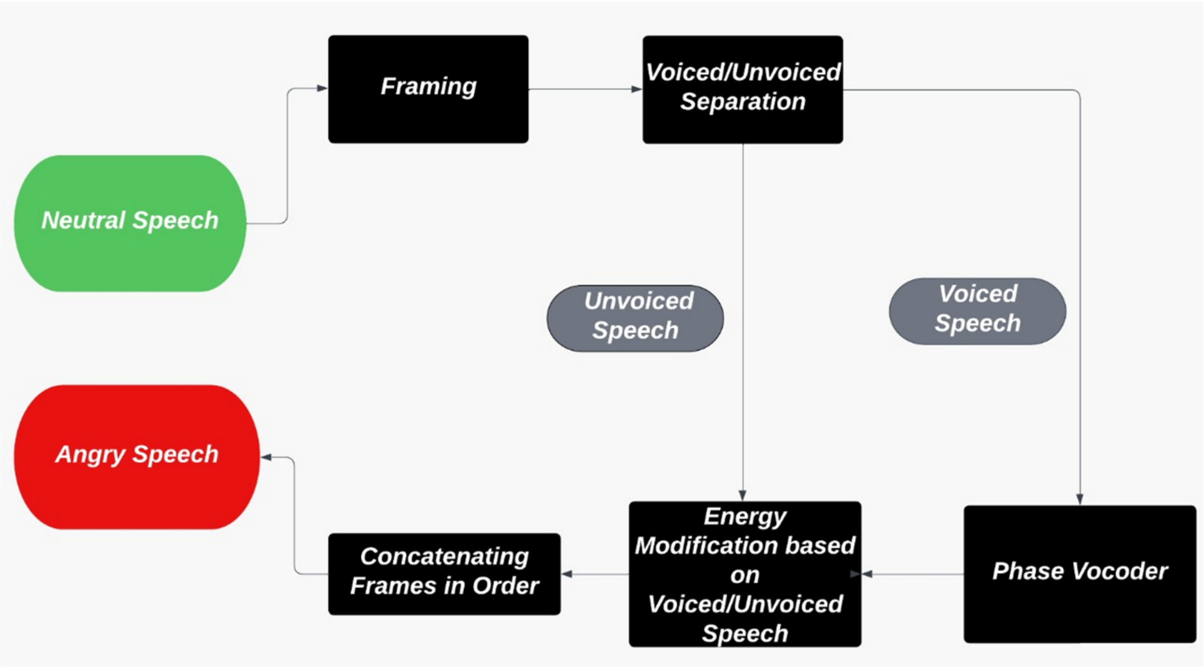
2. Prosody Analysis of Angry and Neutral Speech

In the literature, prosodic features are recognized as major correlates of vocal emotions. For active emotions like Anger, it has been found that pitch and energy values are high as compared to passive emotions like sadness. Conversely, duration is found to be shorter than that of neutral speech and emotions like happiness, sadness, compassion, and sarcasm. In comparison to neutral speech, Anger is also produced with higher intensity and pitch.

For our project, using existing literature as well as experimentally analyzing the effects of prosodic parameter modification in our system, we arrived at the following modification factors for neutral to angry speech conversion.

Prosodic Parameter	**Ratio of anger to neutral speech for given prosodic parameter**
Energy	1.7
Frequency	1.1
Duration	0.7

3. Method


Framing
Voiced/Unvoiced Separation
Neutral Speech
Unvoiced Speech
Voiced Speech
Angry Speech
Energy Modification based on Voiced/Unvoiced Speech
Concatenating Frames in Order
Phase Vocoder

We make the signal undergo windowing to obtain a series of frames. Then, the voiced and unvoiced parts of the speech are separated from each frame. This is done as only the voiced parts possess pitch and undergo most of the increase in energy or intensity during loud/angry speech. This can be implemented by setting an energy threshold that is 0.1 times the mean energy of the signal. The voiced frames are passed through the phase vocoder to modify the duration and pitch parameters through the time-scaling and pitch-shifting functions of the phase vocoder (explained in detail in 3.1 and 3.2 respectively). This is done to increase the pitch and shorten the duration which are characteristics of angry speech. Note that the phase vocoder performs the pitch shifting and time scaling on smaller frames within each voiced frame passed to it. It does this by first windowing each frame using a Hanning window. A rectangular window contains large amounts of energy in its side lobes in the frequency domain. This causes large windowing effects on the speech signal. The use of Hanning window, which contains good side lobe compression(energy is focused on the

main lobe), is done to reduce the altering of the signal from the windowing effect.
All the signal frames then undergo energy modification, with voiced frames undergoing the maximum increase in energy. The frames are then concatenated in their original order to obtain an angry-sounding modified version of the original neutral speech.

3.1 Time Scaling in Phase Vocoder

If we wish to time scale the signal by modifying the sampling rate, the pitch gets affected as a consequence. To time scale without a shift in pitch, the phase vocoder changes the spacing or alignment of the frames. This means the sampling rate is left unchanged and the pitch is hence preserved. As we wish to decrease the duration of the signal, we compress the signal by decreasing the time interval between successive frames or increasing their overlap. This however causes discontinuities in the signal as parts of the signal belonging to different instants in time get aligned between different frames. This is audible to the human ear and is perceived as a glitch or an unpleasant buzzing noise. The phase vocoder gets rid of these glitches by adjusting the phases of each frame such that the discontinuities are not present in the modified frame alignment version of the signal.

3.1.1 Phase Modification Method

Applying FFT on the signal divides the signal into n frequency bins. A signal with a frequency between two bins is smeared. We can get the true frequency using the phase difference between two successive frames in the bin which is represented as, $(\Delta\phi_a[k])_i$ where k is the bin index and i is the frame index. The method is as follows:

If hop_a is the number of samples between the start of two successive frames in the unmodified signal. The time interval between two successive frames, Δt_a is the hop size hop_a divided by the sampling rate frequency, f_s. The frequency deviation of the signal from the bin frequency is given by,

$$(\Delta\omega[k])_i = \frac{(\phi_a[k])_i - (\phi_a[k])_{i-1}}{\Delta t_a} - \omega_{bin}[k]$$

The phase information given by FFT is wrapped i.e it is from -pi to pi, so we calculate the wrapped frequency deviation as,

$$\left(\Delta\omega_{wrapped}[k]\right)_i = \mathrm{mod}[((\Delta\omega[k])_i + \pi), 2\pi] - \pi$$

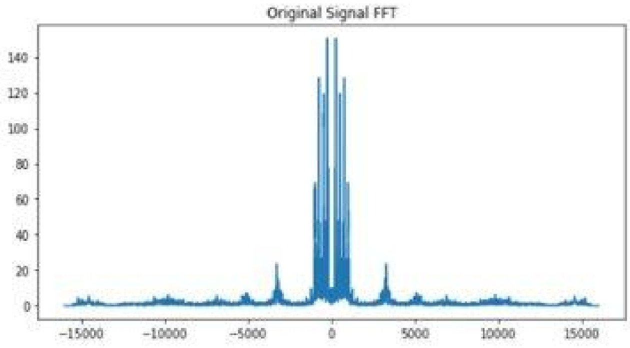
then the true frequency is given by,

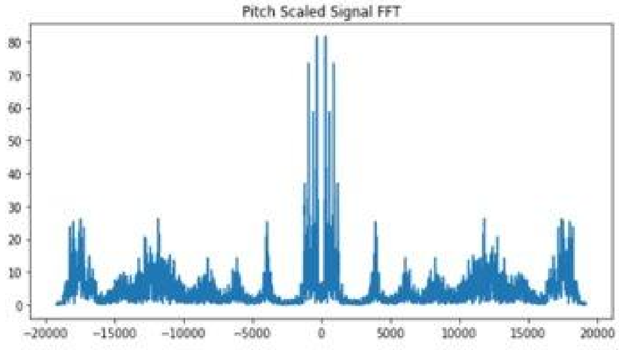
$$(\omega_{true}[k])_i = \omega_{bin}[k] + \left(\Delta\omega_{wrapped}[k]\right)_i$$

$(\phi_s[k])_i = (\phi_s[k])_{i-1} + \Delta t_s \times (\omega_{true}[k])_i$ where Δt_s is the required time interval between two frames in their new alignment. The synthesized phase for the preceding frame, $(\phi_{i-1})_s$ is already known as our algorithm adjusts the phase of each frame in succession.

3.2 Pitch Shifting using Phase Vocoder

Resampling the original signal at a higher rate will reduce the duration and may sound unnatural. To avoid this we first, time scale the signal prior to resampling by a factor equal to the pitch shift factor. The time-scaled signal is resampled at the required rate (= original rate*pitch shift factor) to achieve the pitch shift. This, in turn, returns the signal to its original duration giving us a pitch-shifted signal without change in duration.

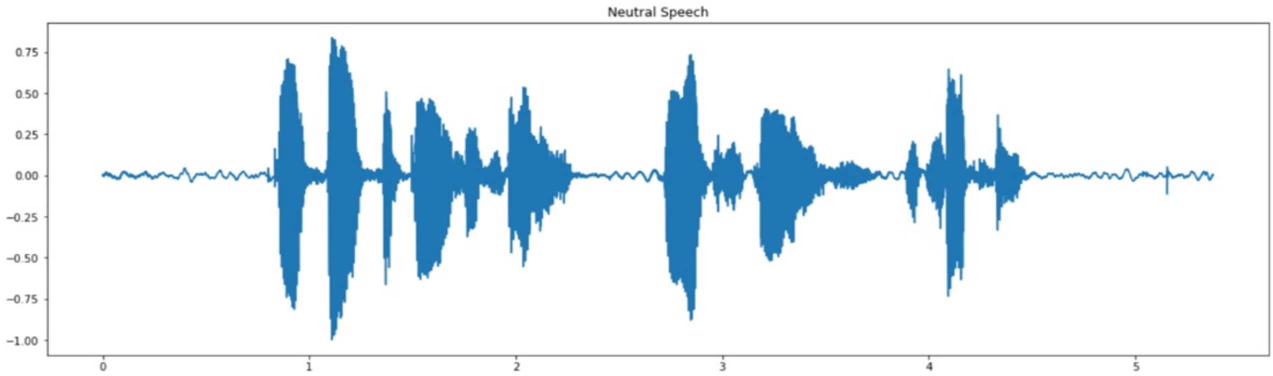




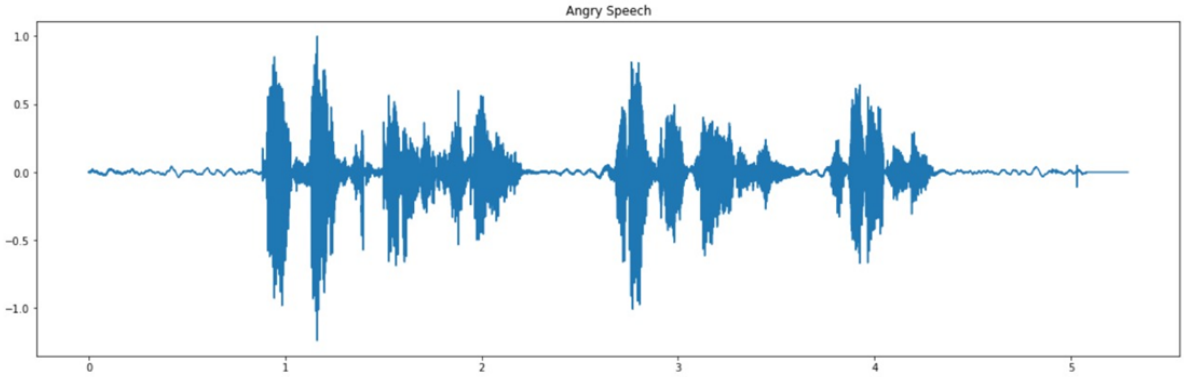
4. Results and Evaluation:

We performed the neutral to angry speech conversion on a total of 8 speakers including 4 male and 4 female speakers. We observed an increase in pitch and energy with a decrease in duration. The plots for one such conversion are given below.

Neutral Speech Signal:



Modified Angry-Sounding Signal:



4.1 Objective evaluation

We conducted the objective evaluation of the success of conversion for 4 male and 4 female speakers by tabulating measurable parameters, namely duration, mean pitch, and mean energy.

The observations are as follows:

Male Speakers:

	Neutral			Angry		
	Mean pitch	Mean Energy	Duration	Mean Pitch	Mean Energy	Duration
Spk 1	60.9	0.8	5.37	71.4	1.4	4.62
Spk 2	55.5	0.67	4.7	61.77	1.3	4.07
Spk 3	104	1.06	5.72	110	1.5	4.94
Spk 4	110	1.01	5.17	121	1.16	4.34

Female Speakers:

	Neutral			Angry		
	Mean pitch	Mean Energy	Duration	Mean Pitch	Mean Energy	Duration
Spk 1	263	0.9	1.55	289	1.3	1.32
Spk 2	229	0.47	2.71	256	1.06	2.29
Spk 3	144	1.4	2.36	159	2.46	1.97
Spk 4	118	1.2	2.89	129	1.9	2.46

4.2 Subjective evaluation

On a scale of 1 to 5 we, with the help of our TA, rated the success of the neutral to anger conversion and acquired the following results:

Rating scale:

Rating	Speech Quality
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

Ratings:

	Male	**Female**
1	4	2
2	3	3
3	3	2
4	2	3

The above results show that the conversion success was satisfactory.

5. Summary and Conclusions

We built a system that exploits prosodic parameters to modify neutral speech to produce an angry-sounding version of it. Pitch, duration, and energy parameters were used in our conversion. We reached the conclusion that the majority of the rise in speech energy occurs in the voiced parts of speech. This observation along with pitch being a property of only voiced signals naturally led us to subject the voiced frames to the majority of the modification. The conversion was fairly successful.