Prosody Modification of Recorded Speech in Time-Domain

Ishpreet Kaur
M.Tech, Computer Sci. Engg.
RIMT-IET, Mandi Gobindgarh,
Punjab, India

Manveen Kaur Oberoi
M.Tech, Computer Sci. Engg.
RIMT-IET, Mandi Gobindgarh,
Punjab, India

Simrat Kaur
Associate Professor
Computer Sci. dept, RIMT-IET,
Mandi Gobindgarh, India

ABSTRACT

Human language carries a lot of information. It can be thought of as comprising two channels – the words themselves and style in which they are spoken. The speaking style reflects the state of a person in a particular environment. And the style in which these words are spoken reflects the prosodic information in them. Prosody is a collection of factors that control the pitch, duration and intensity of a speech signal. These prosodic parameters can be controlled in domains like frequency domain and time domain. In this paper we will explore how prosodic information can be controlled and thus the speech can be modified in time domain.

Keywords

Prosody, Prosody modification, Time-domain.

1. INTRODUCTION

Speech is a common means of communication for people. Speech is the vocalized form of human communication. It is based upon the syntactic combination of lexical and names that are drawn from very large vocabularies. Each spoken word is created out of the phonetic combination of a limited set of vowel and consonant speech sound units. These vocabularies, the syntax which structures them and their set of speech sound units differ, creating the existence of many thousands of different types of mutually unintelligible human languages.

Speech processing is the study of speech signals and the processing methods of these signals. The signals are usually processed in a digital representation, so speech processing can be regarded as a special case of digital signal processing, applied to speech signal. For example, Speech recognition deals with the analysis of the linguistic content of a speech signal and its conversion into a computer-readable format and speech generation is the process which allows the transformation of a string of phonetic and prosodic symbols into a synthetic speech signal. A **text-to-speech (TTS)** system converts normal language text into speech.
Time domain is the analysis of mathematical functions, physical signals or time series of economic or environmental data, with respect to time. In the time domain, the signal or function's value is known for all real numbers, for the case of continuous time, or at various separate instants in the case of discrete time. In Time-domain representation of speech waveform, the horizontal axis represents time and the vertical axis represents amplitude. In Time-domain, prosody of the recorded speech can be controlled.

Prosody is one of the key components of Speech Synthesizers, which allows implementing complex weave of physical, phonetic effects that is being employed to express attitude, assumptions, and attention as a parallel channel in our daily speech communication.

Prosody is essentially a collection of factors that control the pitch, duration and intensity to convey non-lexical and pragmatic information in speech [1]. The objective of ***prosody modification*** is to change the pitch and duration of the sound units of speech so as to generate the output in the same form as if it is actually spoken or generated by people in conversation.

From listener point of view, prosody consists of systematic perception and recovery of speaker intentions based on [2]:

a) Pauses: To indicate phrases and separate the two words.
b) Pitch: Rate of vocal fold cycle as function of time.
c) Rate: Phoneme duration and time.
d) Loudness: Relative amplitude or volume.

The above parameters can be controlled in time- domain of any recorded speech.

2. CONTROLLING THE PROSODY FACTORS OF SPEECH IN TIME-DOMAIN

Various prosody parameters can be controlled in time domain for generated speech, and changing the various parameters like pitch, duration, amplitude and speech rate affect the waveform of the generated speech. The parameters are generally controlled to increase the *naturalness* of the speech. *Naturalness* describes how closely the output sounds like human speech [3, 4].

The prosody parameters which can be controlled in time-domain for speech are discussed below:

2.1 Pitch

Pitch is also perceived as Fundamental frequency. Frequency refers to the rate at which a sound wave vibrates. Pitch can be varied as per our requirements. For example, anger emotion in speech exhibits wide pitch range and high mean and sadness exhibits lower or narrow pitch range [3, 4]. For female voice, pitch range is higher than male voice.

2.2 Amplitude

Amplitude refers to the rate of loudness of sound wave. Amplitude can also be perceived as Intensity. Amplitude can be increased or decreased. The vertical axis in time-domain waveform represents the amplitude of sound/speech over time. For example, the tendency of the amplitude is increasing for anger, slight increase for happiness and decreasing tendency for sadness [4].

2.3 Duration

Duration refers to the length of speech segments such phonemes and syllables. The horizontal axis represents the duration of the speech waveform. For example, tendency of duration is decreasing for joy, anger and fear; and increasing for grief. Longest durations are reported for sadness, average durations for neutral speech and fear, and short durations for joy and anger [3, 4].

2.4 Speech rate:

Speech rate refers to the speed or rhythm of a piece of music. For example, Anger exhibits fast speech rate, and sadness exhibits slow speech rate [4].

3. EXPERIMENTAL ANALYSIS

In our proposed work, a speech is recorded in Wave File Format (extension .wav). Wav file is read to obtain header information and amplitude values (sample points) in time-domain. Header information of a wave files gives information about important parameters like sample rate, bit rate etc.

The Figure1 shows the waveform for original recorded speech in Time-domain.

The *sample rate* is modified to vary pitch, and *amplitude* values are modified to change the loudness of the speech. Duration and speech rate are also controlled. The modification in amplitude values, duration and speech rate results in generating different emotions on the same neutral speech recording.

3.1 Pitch modification

The *pitch modification* is done to vary the pitch of the recorded wave sound file which results in producing different voices from same recording. In terms of wave file parameters, this fundamental frequency is known as sampling rate that is the rate at which the sound is to be played back in sample frames per second (i.e., Hertz). Sample rate is an important parameter of a wave file that can be used for the producing different voices from the same recording of speech. Male's recorded speech can be converted to female's or a child's speech by changing the value of this parameter and vice versa. This can be done as:

The sample rate is modified between 12 kHz and 15 kHz for female voice, sample rate lies between 8 kHz to 12 kHz for male voice and 20 kHz to 25 kHz for child.

3.2 Amplitude modification

Sample values (Amplitude values over time in time-domain) are modified as follows:

3.2.1 For anger

Sample values are modified as:
For the absolute values lying between 20000 and 25000, sample values are increased by approx 20-25% of their original values if they are positive, decreased by approx 20-25% of their original values if they are negative. For the absolute values lying between 0 and 20000, sample values are increased by approx 40-50% of their original values if they are positive, decreased by approx 40-50% of their original values if they are negative.

3.2.2 For happiness

Sample values are modified as:
For absolute values lying between 0 and 10000, sample values are increased by 70% of original values if they are positive and decreased by approx 70% of original values if they are negative. For absolute values lying between 10000 and 20000, the sample values are increased by approx 50% of their original values if they are positive, decreased by approx 50% of their original values if they are negative.

3.2.3 For sadness

Sample values are modified as:
The absolute sample values are decreased by approx 50-60% of their original values.

3.3 Speech rate and duration

For increasing the speech rate for anger, some samples are removed from the sound wav file and thus size of the sound file decreases. Thus, increasing in the speech rate results in decrease in duration also. For decreasing the speech rate for sadness, some samples are repeated and thus the size of the sound wave file increases. Thus, the decrease in speech rate causes the increase in duration also.

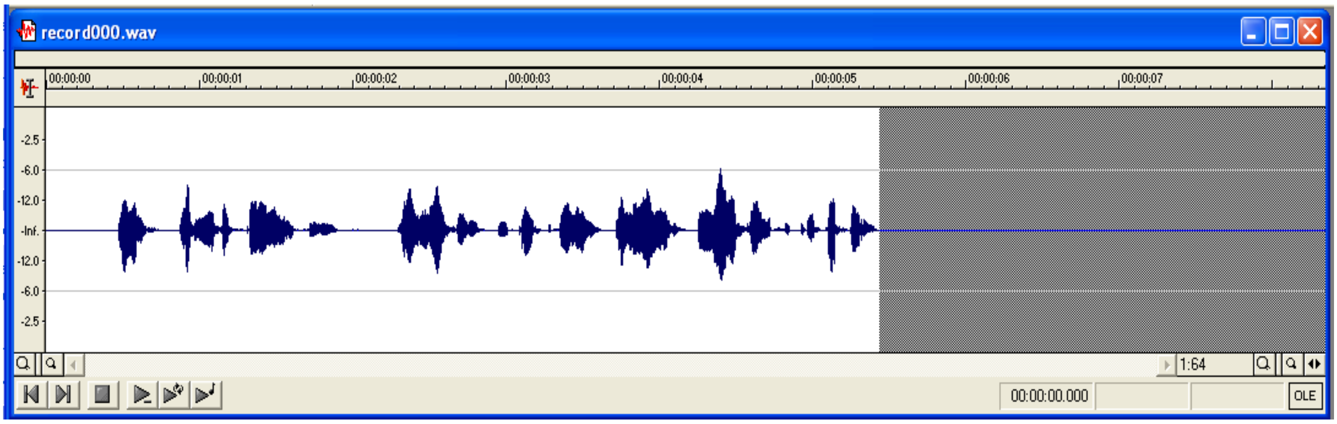

Fig 1: Speech Waveform of original sound wave in Time-domain

4. RESULTS

Here, waveforms resulted due to prosody modification are shown and discussed.

The Figure2 shows that prosody parameters pitch, amplitude, duration are controlled. The amplitude values are higher than those of original sound waveform. Pitch is increased here and duration is reduced from that of original sound wave from 1:64 seconds to 1:32 seconds which is shown in figure at the right of the bottom, which generally represents settings of parameters for Anger.

Figure3 shows the slight increase in amplitude and pitch in time-domain which generally represents settings of parameters for Happiness. Here, duration is almost same as original sound wave, but amplitude and pitch have slightly risen.

The Figure4 shows the increase in duration from that of original sound wave from 1:64 seconds to 1:128 seconds, and decrease in amplitude values from those of original sound wave which represents the settings for sadness emotion.

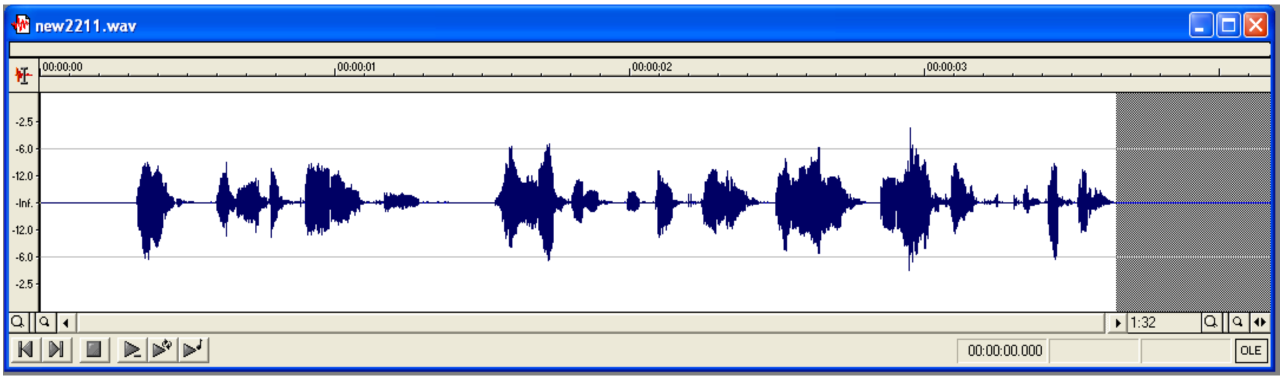

Fig 2: Speech waveform for Emotion Anger in Time-domain

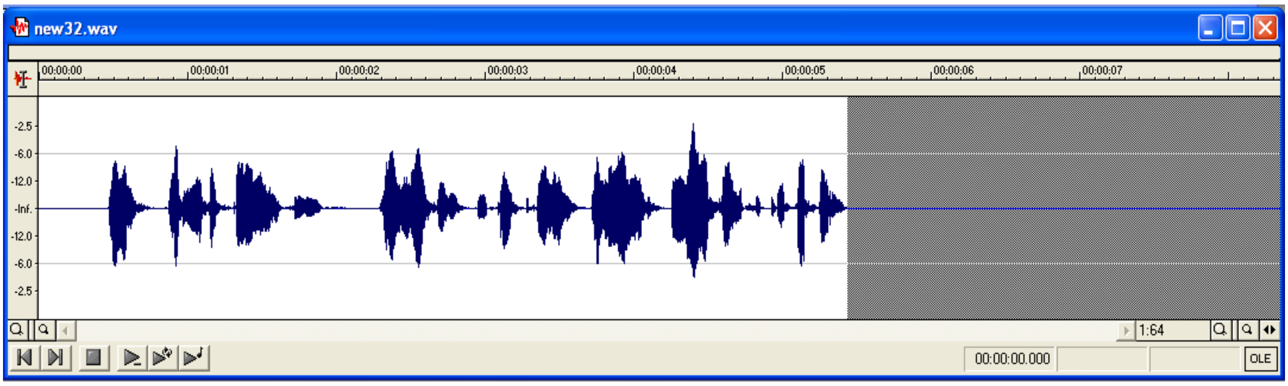

Fig 3: Speech Waveform for Emotion Happiness in Time-domain

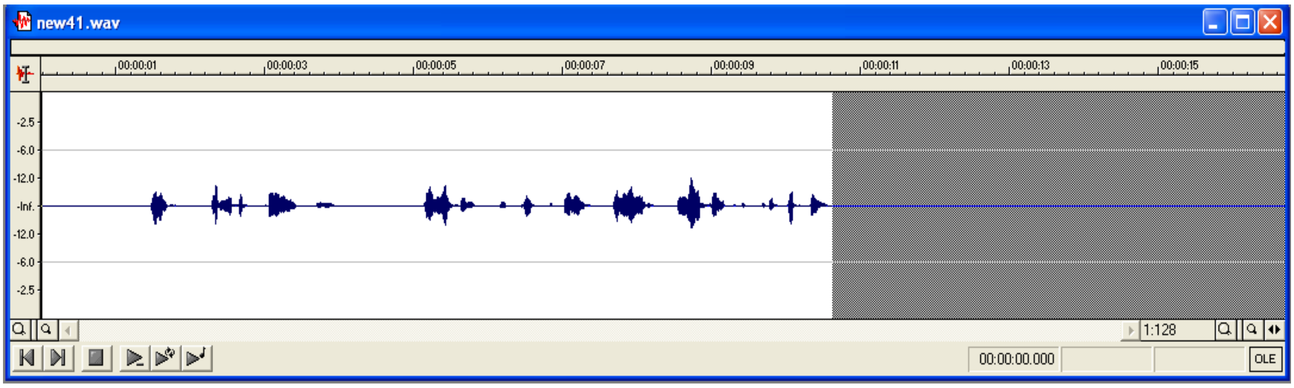

Fig4: Speech waveform for Emotion Sadness in Time-domain

5. CONCLUSION AND DISCUSSION

In this paper various prosody parameters like pitch, duration and amplitude have been described. We have shown that how the speech can be modified by controlling various prosody parameters in time domain. It has been shown that prosodic information is essential for a reliable detection of the underlying state of the speaker. We analyzed by our experiment that there were sharp changes in the amplitude and other prosody parameters for anger and sadness but there was slight increase in amplitude for happiness. Thus the recognition rates for anger and sadness are higher than happiness Future work will concentrate on exploring more behaviors or states of the speech and also improving the results for happiness.

6. REFERENCES

[1] Hoult, C. May 2004. Emotion in speech synthesis.

[2] Chandak, M.B., Dharaskar, R.V. and Thakre, V.M. 2010. Text to Speech Synthesis with Prosody feature: Implementation of Emotion in Speech Output using Forward Parsing. In the Proceedings of International Journal of Computer Science and Security.

[3] Thakur, S.K., Satao, K.J. 2011. Study of various kinds of Speech Synthesizer Technologies and Expressions For Expressive Text To Speech Conversion System. In the Proceedings of International Journal of Advanced Engineering Sciences and Technologies.

[4] Ishpreet Kaur, Manveen Kaur Oberoi, Simrat Kaur. 1-2 Nov 2012. A Survey on Approaches to Emotional Speech Synthesis and Parameters to control Emotions. In the Proceedings of Speech, Image, Biomedical and Information Processing 2012 (SIBIP 2012), International Conference at Chitkara University, Nov2012.