Implementation of TDPSOLA to Add Emotions in Punjabi Speech

*Mamta Sharma, **Payal Sharma
*Student, RIMT, Mandi Gobindgarh
**Assistant Professor, GKU
mamta_sharma0101@rediffmail.com ,bhj.payal@gmail.com`

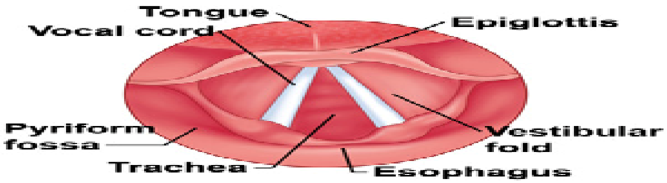
Abstract
This Paper concerns the process of embedding an emotion in the speech using Time Domain Pitch Synchronization Overlap and Add. The goal was to modify the pitch of the speech that can portray emotion with different levels of intensity. To achieve this, the system was based on theoretic frameworks developed by Psychologists to describe emotions. The basic goal is to perform synthesize the speech so that it sounds naturally. To increase the naturalness of the synthesized speech, the synthesized speech should deliver certain content in right emotion (e.g. good news are delivered in a happy voice), therefore making the speech and content more believable. Embedding emotion to the speech is a step in this direction. Embedding emotion seems to be straightforward initially but revealing the emotions adds more difficulties in studying and implementation.

Introduction

Speech: 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 (usually to about 10,000 different words) vocabularies. Each spoken word is created out of the phonetic combination of a limited set of vowel and consonant speech sound units.

Parameters:

Pitch:-Pitch is defined as "rate of Vibration of vocal folds



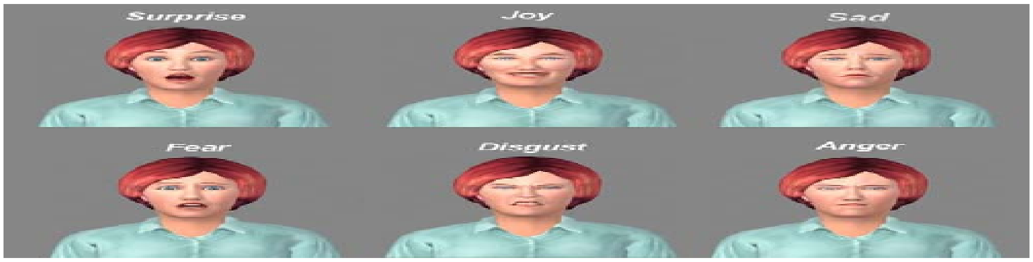
The Sound of the voice Changes as the rate of vibration varies.

How these Vibration and pitches are created: The Vibrations, and the speed at which they vibrate, are dependent on the length and thickness of the vocal cords, as well as the tightening and relaxation of the muscles surrounding them. The vocal fold is not only the only factor that affects the pitch. The pitch of someone's voice can also be affected by emotion, mood and inflection.

Duration: The duration or time characteristics can also be investigated at several levels from phoneme (segmental) durations to sentence level timing, speaking rate, and rhythm. The segmental duration is determined by a set of rules to determine correct timing. Usually, some inherent duration for phoneme is modified by rules between maximum and minimum durations. **Intensity:** The intensity pattern is perceived as a loudness of speech over the time. At syllable level, vowels are usually more intense than consonants and at a phrase level, syllables at the end of an utterance can become weaker in intensity.

Frequency:-Voiced sound has periodic structure i.e. their signal form repeats itself after time, and this is called pitch period and reciprocal of the same is called pitch frequency.

Emotion:-The number of possible emotion is very large. But there are five discrete emotional states which are commonly referred as primary or basic emotions and other are the altered or mixed form of emotion.



Primary emotion:-

Anger:-Anger in speech causes intensity with Dynamic Changes. The voice is breathy. The average pitch pattern is higher. The pitch range and the vibration are also wider than in normal Speech and the average speech rate is faster.

Happiness or Joy:-Happiness and Joy causes slightly increased intensity and articulation for content. Happiness also leads to increase in pitch. The peak values for pitch are higher than Basic emotion.

Fear or anxiety:-Fear or anxiety makes the intensity of Speech Lower with no dynamic changes. The speech rate is faster than in normal speech and contains pauses between words.

Secondary Emotion:-

Disgust or contempt:-Disgust or contempt and *in* speech also decreases the speech intensity and its dynamic range. The average pitch level and the speech rate are also lower compared to normal speech and the number of pauses is high.

Tiredness: - Tiredness causes a loss of elasticity of Articulatory muscles leading to lower voice and narrow.

Speech synthesis

Speech synthesis, and specifically emotional speech synthesis, is a challenging research topic. Two of the main challenges are that (1) there are numerous parameter values that can be selected during the generation of pitch, duration, and energy contours, and that (2) human evaluators are needed to evaluate synthesizers' performances.

EMOTIONAL SPEECH SYNTHESIS

Speech synthesis is usually done in a two step approach. First, the text gets analyzed by a natural language processing (NLP) module and converted into a phonemic representation aligned with a prosodic structure, which is then passed to a digital speech processing (DSP) component in order to generate a speech signal. We developed an emotional speech synthesis system on the basis of Mbrola [6]. In order to obtain as many speech samples as possible, we used two different phonemisation components, namely Text2Pho and open- Mary [7] for natural language processing. Emofilt acts as a transformer between the phonemisation (Text2Pho or open Mary) and the speech-generation component (Mbrola). The emotional simulation is achieved by a set of parameterized rules that describe manipulation of certain acoustic aspects of a speech signal. The rules were motivated by descriptions of emotional speech found in the literature. Before the rules are applied by Emofilt, the input phoneme chain gets syllablised by an algorithm based on sonority hierarchy. In addition, stressed syllables are identified as those that carry local pitch contour maxima [8]. For the experiments we synthesized the 10 sentences of the Berlin Emotional Database (cf. sec. 4), simulated 8 target emotions and emotion-related states (boredom, despair, fear, happiness, hot anger, joy, sadness, and yawning) plus neutral with Emofilt, using all seven German voices for Mbrola (4 female and 3 male), thus getting 1260 samples (10 _ 2 _ 9 _ 7). The following sections describe the modifications provided by EmoFilt.

TT TDPSOLA ALGORITHM:

This thesis concerns the process of embedding an emotion in the speech using Time Domain Pitch Synchronization Overlap and Add. The goal was to modify the pitch of the speech that can portray emotion with different levels of intensity. To achieve this, the system was based on theoretic frameworks developed by Psychologists to describe emotions. The basic goal is to perform synthesize the speech so that it sounds naturally. To increase the naturalness of the synthesized speech, the synthesized speech should deliver certain content in right emotion (e.g. good news are delivered in a happy voice), therefore making the speech and content more believable. Emotions can make the interaction with the computer more natural because the system reacts in ways that the user expects. Embedding emotion to the speech is a step in this direction. Embedding emotion seems to be straightforward initially but revealing the emotions adds more difficulties in studying and implementation. The biggest challenges in emotion speech resynthesis is the selection of modification parameters that will make humans perceive a targeted emotion. The best selection method is by using human raters. Our inspiration for this thesis is based on the paper written by [1].In the paper they described recognition for synthesis system to automatically select a set of possible parameter values that can be used to resynthesize emotional speech. The system deals with the synthesis using TD-PSOLA.The TD-PSOLA algorithm was proposed allowing pitch modification of a given speech signal without changing the time duration and vice versa. The TD-PSOLA consists mainly of the following three steps [1].

1. The analysis step, where the original speech signal is first divided into separate but often overlapping short term analysis signals (ST). Short term signals xm(n) are obtained from the digital speech waveform x(n) by multiplying the signal by a sequence of the pitch synchronous analysis window hm(n) as in Eq.

   $Xm\ (n) = hm(tm - n)\ x\ (n)$

   Where m is an index for the short-time signal
2. The windows, which are usually Hanning type, are centered on the successive instants tm, called pitch marks. These marks are set at a pitch-synchronous rate on the voiced parts of the signal and at a constant rate on the unvoiced parts.
3. The modification step, where each frame is modified according to the target. The synthesis steps are performed such that these segments are recombined by means of overlap adding.

   TD-PSOLA relies on the pitch synchronous decomposition of the signal into overlapping frames synchronized with pitch period. The main objective is thus to preserve the consistency of marks between neighboring frames with respect to the temporal structure of pitch periods. First, we improve pitch marking by eliminating mismatch errors which appear during rapid formant transitions. This is achieved by pruning pitch mark candidates whose distance with other candidates is clearly not consistent with the current pitch period. Pitch adjustment of a digitally-sampled audio file can be implemented simply using resampling. However, this completely alters the time scaling and cannot account for changes in the pitch and inflection of a voice over time, and thus cannot be considered. Instead, we shall use the more sophisticated

Pitch-Synchronous Overlap Add algorithm, which allows us to modify pitch without compromised information or modifying the time scaling.

The pitch correction method involves the following basic steps:

- Detection of original pitch
- Parsing of desired pitch frequencies
- Correction of pitch

Signal Model Analysis/Synthesis

Signal modeling techniques model sounds as a sum of elementary sinusoidal components called partials. These techniques start by extracting partials, in the form of time dependent magnitude and instantaneous phase data, from a signal via time-frequency analysis, usually a Fourier transformation transform based analysis, such as the short-time Fourier transform (STFT). The decomposition into individual partials is expressed in the additive synthesis

Results

The experiment is performed on Punjabi voices both for male and female. We achieved very clear synthesized emotional speech through TD-PSOLA for many cases. The TD-PSOLA method is not effective for stochastic speech signal as the frequency domain peak-picking cannot estimate a modulation rate in the aspiration noise source. The threshold value is still not perfect as we do more testing on different voices and utterances using TD-PSOLA. The pitch marking of the wave is very limited to the testing voices.

The results show that the proposed system is promising for selecting parameters for embedding emotional in speech. Considering the significantly different performances for different emotions, and the differences observed between human and machine perception of emotions, however, at this stage we prefer to view the proposed automated evaluation more as a Preprocessing step than a replacement to human evaluations.

Conclusion and Future Work

Our future research will be directed towards the design of more robust systems, more sophisticated parameter modifications, and experimenting with different parameter selection techniques and additional emotions and embedding emotion in a stochastic speech.

References

[1] Murtaza Bulut, Sungbok Lee and Shrikanth Narayanan, "Recognition For Synthesis: Automatic Parameter Selection For Resynthesis Of Emotional Speech From Neutral Speech", ICASP, pp. 4629-4632, Sep. 2008.

[2] Stevens, K. : Models of speech production. In: Crocker, M., J., Ed., Encyclopedia of
Acoustics, pp. 1565-1578, John Wiley & Sons. Inc., 1997.

[3] Fant, G.: Acoustical Analysis of Speech. In: Crocker, M., J., Ed., Encyclopedia of Acoustics, pp. 1589-1598, John Wiley & Sons, Inc., 1997.

[4] Deller, J., R., Proakis, J., G., Hansen, J., H., L.: Discrete-Time Processing of Speech Signals, Macmillan Publishing Company, 1993.

[5] Juang, B., H., Ed.: The Past, Present, and Future of Speech Processing. In: IEEE Signal
Processing Magazine, vol. 15, no. 3., pp. 24-48, May 1998.

[6] Spanias, A., S.: Speech Coding: A Tutorial Review. In: Proceedings of the IEEE, vol. 82, no.
10, pp. 1541-1582, October 1994.

[7] Makhoul, J.: Linear Prediction: A Tutorial Review. In: Proceedings of the IEEE, vol. 63, no.
4, pp. 561-580, April 1975.

[8] Kay, S., M., Marple, S., L.: Spectrum Analysis - A Modern Perspective. In: Proceedings of the IEEE, vol. 69, no. 11, pp. 1380-1419, November 1981.

[9] Markel, J., D., Gray, A., H.: A Linear Prediction Vocoder Simulation Based upon the Autocorrelation Method. In: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-22, no. 2, pp. 124-134, April 1974.

[10] Gray, A., H., Markel, J., D.: A Spectral-Flatness Measure for Studying the Autocorrelation Method of Linear
Prediction of Speech Analysis. In: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-2, no. 3, pp. 207
-

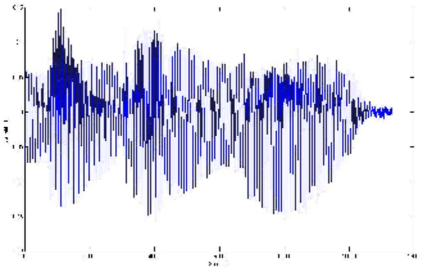
1. MATLAB Implementation

MATLAB files:

1. detect_vuv.m : it perform the detection of voiced and unvoiced samples in speech wave.
2. energy.m : it calculates the energy of the input frames of speech wave.
3. find_pmarks.m : it calculates and marks the pitches at the peaks in the short time energy function.
4. plot_pmarks.m : it plots the pitch marked by the „find_pmarks.m‟.
5. tdpsola.m : it perform the time domain pitch synchronous overlap add function for embedding emotion in speech.
6. test.m : integrated file to perform the whole process with an ease.
7. window.m : it calculates the windowed coefficients of the speech frames passed through window.
8. zerocr_frm.m : it calculates the average zero crossing rate of the input speech frames.

MATLAB figures:

1. 1.Speech waveforms used for test



2. Pitch marks on the original speech with the threshold value as

a. Tscale:1.5

b. Pscale: 0.7