Volume 4, Issue 5, May 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Prosody Modification of its Output Speech Signal

HARPREET KAUR **PARMINDER SINGH**
M.Tech .Student Associate Professor
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
GURU NANAK DEV ENGINEERING COLLEGE, LUDHIANA (PUNJAB) INDIA

ABSTRACT-Speech synthesis systems which involve concatenation of recorded speech units are currently very popular. These methods are known for producing high quality, natural-resound speech as they generate speech by joining together speech signal of different speech units. This technique of speech generation is quite practical. But the speech units that are being concatenated in Punjabi language may have different discontinuity in signal. The presence of such discontinuities can be very distract to the listener and degrade the overall quality of output speech. This paper deals with the problem of discontinuity of speech signal in order to increase its naturalness. This technique is implemented on different Punjabi wave audio files which were created by concatenating different Punjabi syllables.

Keywords-Speech signal, Pitch, Duration, Intensity and discrete-time Fourier transform.

I. INTRODUCTION

Speech is the most primary form of communication used by human beings to express their thoughts, feelings and ideas. Speech production involves a series of complex movements that alter and mould the basic tone created by human voice into specific sounds. The mechanism for generating the human voice can be subdivided into three part that is; the lungs, the vocal folds within the larynx, and the articulators (the parts of the vocal tract above the larynx exist of tongue, palate, cheek, lips, nose and teeth). Speech sounds are created when air pumped from the lung causes vibratory activity in the human vocal tract. Speech is the most natural form of human communication. It is one of the most facts-laid signals. Speech sounds have a rich and many layered temporal-spectral variation that convey intention, expression, tone, accent, speaker identity, gender, age, style of speaking, condition of health of the speaker and emotion. Speech is an immensely information-rich signal exploiting frequency-modulated, amplitude-modulated and time-modulated conveys (e.g. resonance movements, harmonics and noise, pitch intonation, power, duration) to convey information about words, speaker identity, accent, expression, way of speech, emotion and the condition of health of the speaker.A computer system with the ability to convert written text into speech is known as Text-To-Speech (TTS) synthesis system. The quality of a speech synthesizer is decide by naturalness, which refers to the similarity of generated speech to the real human voice and intelligibility, which refers to the ability of generated speech to be understood. The main goal of researchers and linguists is to create ideal speech synthesis systems which are both natural and intelligible.

II. PARAMETERS OF SPEECH

As basic parameters in speech processing we regard pitch, extent, intensity, voice quality, signal to noise ratio and voice activity detection. This section gives a brief description of various types of Speech parameters in speech signal.

Pitch – Pitch is our perceptual interpretation of frequency. The lowness or highness of sound is its Pitch. The pitch of sound depends on how quick the object vibrates. Something that vibrates steadily makes a low-pitched sound. Something that vibrates very quick makes a high-pitched sound. The pitch pattern over a sentence (intonation) in natural speech is a combination of many factors. The pitch contour hangs on the meaning of the sentence. In normal speech the pitch slightly decreases toward the end of the sentence and when the sentence is in a question form, the pitch pattern will raise to the end of sentence. In the end of sentence there may also be a continuation rise which indicates that there is more speech to come. A lift or drop in fundamental frequency can also indicate a stressed syllable.Finally, the pitch contour is also pompous by gender, physical and emotional state, and attitude of the speaker.

Duration – The duration or time characteristics can also be investigated at several levels from phoneme. Span to sentence elevation timing, speaking rate, and rhythm. The segmental duration is determined by a set of rules to determinate timing correct. Usually some inherent duration for phoneme is modified by rules between maximum and minimum durations or time. For example, consonants in non-word-initial position are shortened, emphasized words are significantly stretch, or a stressed vowel or son rant preceded by a voiceless plosive is lengthened (Klatt 1987, Allen et al. 1987). In general, the phoneme duration differs due to neighbouring phonemes. At sentence level, the speech rate, rhythm, and right placing of pauses for correct phrase boundaries are important. For example, a absent phrase boundary just makes speech sound rushed which is not as bad as an extra boundary which can be confusing (Donovan 1996).

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.

The fundamental frequency is highly related with intensity pattern in speech. The intensity of a voiced sound goes up in proportion to fundamental frequency (Klatt 1987).

III. IMPROVING NATURALNESS OF TTS GENERATED SPEECH

Text To Speech synthesis is an application to convert the written text in a language into speech. The text to speech conversion process enables user to enter text and output it with sound. The text inputted may be written by the operator or it may be scanned paper that is converted to speech. Speech can be produced by several different methods. The first task of all TTS systems is to convert input data to proper form for a synthesizer. All of these have some benefits and deficiencies. The methods are usually rate into three groups:

- Articulatory synthesis, which attempts to model the human speech production system directly.
- Formant synthesis, which models the pole frequencies of speech signal or transfer function of vocal tract based on source-filter-model.
- Concatenative synthesis, which uses different length pre-recorded samples derived from natural speech.

The formant and concatenative methods are the most commonly used in present synthesis systems. The formant synthesis was dominant for long time, but today the concatenative method is becoming popular. The articulatory method is still complicated for high quality implementations, but may arise as a potential method in the future.

(a) CONCATENATIVE SYNTHSIS

Concatenative speech synthesis is currently the most practical method for the generation of realistic speech Connecting pre-recorded natural utterances is probably the easiest way to produce intelligible and natural sounding synthetic speech. However, concatenative synthesizers are usually limited to one speaker and one voice and usually require more memory capacity than other methods. The concatenative synthesis is simpler than rule-based synthesis because there is no need to determine speech production rules. Concatenation-based systems produce very natural sounding continuous speech, since in this method, databases of pre-recorded speech sounds are referred and waveforms of appropriate speech units are joined together to form any sentence. For simplicity, words or other speech units are stored as sampled waveforms in the acoustic databases. A large number of utterances can be created by referring to the databases and selecting suitable words or phrases according to the given context. Concatenative synthesis has three sub-categories. Unit Selection Synthesis, Dip hone Synthesis and Domain-Specific Synthesis.

(b) UNIT SELECTION SYNTHESIS

The unit selection based systems may make use of very large databases of fluent speech units which may include phrases, words, syllables and phonemes. For the greatest fluency, finding the best available unit and choosing precise concatenation points is important.

(c) DIPHONE SYNTHSIS

Diaphone synthesis based systems generate speech by joining together diaphones which are context-sensitive unit that extend from the middle of the stable region of one phoneme to the middle of the stable region of the following one. The speech synthesis systems using dip hones produce clear speech and desired prosody can be modelled for a particular context by using signal processing.

(d) DOMAIN SPECFIC SYNTHSIS

The domain-specific synthesis is used in systems where the variety of result is limited to a particular domain such as weather reports, airport announcements and digital clocks. Systems based on this technique are limited by the number of words and phrases in the database and are known to produce speech sounds of very high quality.

IV. SURVEY OF THE TECHIQUES

I. **Dinesh Kumar et al. 2005** specified two phases for the speech: speech recognition & speech synthesis. The main aim of this research was to develop a system which speaks the handwritten Punjabi word now. The research for Punjabi word recognition is limited to 2460 Punjabi characters only (i.e. only for words available in database). He suggested a system or technique used for recognition was Support Vector Machine & for speech synthesis technique used is CTTS (ConcatenativeText-to-Speech). For recognition, the approached was database independent. But for speech synthesis, the approach is database dependent.

   Macon et al. 1997 proposed that sinusoidal models were used successfully in singing voice synthesis. He suggested that the synthesis of singing differs from speech synthesis in many forms. In singing, the intelligibility of the phonemic message was often secondary to the intonation and musical qualities. Vowels were usually sustained longer in singing than in normal speech and naturally independent controlling of pitch and loudness was also required.

II. **Gaved 1993** discussed approach called pronunciation by analogy where a novel word was recognized as parts of the known words and the part pronunciations were built up to produce the pronunciation of a new word. In many speech markup languages information of correct pronunciation could be given separately.

III. **Donovan 1996** described that the pitch pattern or fundamental frequency over a sentence in natural speech was a combination of many factors. The pitch contour was depended on the meaning of the sentence. In normal speech the pitch slightly decreases toward the end of the sentence and when the sentence was in a question form, the pitch pattern would raise to the end of sentence. In the end of sentence there might also be a continuation rise which indicates that there was more speech to come. A raise or sink in fundamental frequency indicated a stressed syllable. The pitch contour was also affected by gender, physical, attitude and emotional state of the speaker.

IV. **Abadjieva et al. 1993** stated that the speaker's feelings and emotional state affect speech in many forms and the proper implementation of these features in synthesized speech might increase the quality considerably. With text-to-speech systems this was rather difficult because written text usually contains no information of these features. This kind of information might be provided to a synthesizer with some specific control characters or character strings. The users of speech synthesizers might also need to express their feelings in "real-time".

V. **Deepika 2012** proposed a speech signal processing technique that deals with the problem of spectral discontinuity in the context of concatenated waveform synthesis. It involved the post-processing of the synthesized speech waveform in time domain. This technique was implemented on different single channel Punjabi wave audio files which were created by concatenating different Punjabi syllables. A listening test was conducted to assess the proposed technique, and it was seen that the spectral discontinuity is reduced to a large extent and the output speech sounds more natural with the reduction of audible noise.

VI. **Harald Höge et al. 1993** stated that basic parameters in speech processing was pitch, duration, intensity, voice quality, signal to noise ratio, voice activity detection and strength of Lombard effect. Taking in account all factors the performance of many published algorithms to extract those parameters from the speech signal automatically is not known. A framework based on fierce evaluation was proposed to push algorithmic research and to make progress comparable.

VII. **Pavol Partila et al.** described that recognition of speech made in a particular emotional state and examined the impact of person’s emotional state on the fundamental speech signal frequency. Vocal chords create audio signals which carry information coded with human language. This process was known as human speech. Based on a speech signal several speakers’ attributes such as sex, age, speech disorders (stuttering or cluttering) and emotional state can be determined. As for emotions, only about 10% of speaker’s emotional state or state of mind is expressed by means of speech.

VIII. **O.F.Krivnova, et al 2003** .stated strategy and ways of F0 contour generation in TTS system for Russian language are described. The system was developed in Lomonosov Moscow State University and based on two methods: concatenation of allophones waveforms and prosodic rules to control pitch, duration and intensity. These potentate form a part of speech control module which carries out the interface function, bridging the gap between the outputs of

IX. text linguistic processing and the input of speech signal generation module. As a result each segment (allophone) in a phrase being synthesized is attributed by at least two F0 point as its starting and ending values. Three and even more F0 values can be assigned to the phone if it is necessary. Signal generation is execute according to the phrase control file, which describes the phrase as a order of allophones code names with assigned duration, energy and fundamental frequency significance technology.

X. **Jianhua Tao et al.2003** proposed significant number of activities in the area of Chinese Natural Language Processing including the language resource construction and assessment. He summarized the major ways and key technologies in Natural Language Processing, which encompasses both text processing and speech processing by appendix. The Chinese Language resources, including speech data, linguistic data, evaluation data and language toolkits which are elaborately constructed for Chinese Natural Language Processing related fields and some language resource consortiums are also introduced in it. Aimed to promote the development of corpus-based technologies, many resource consortiums complete themselves to create and collect many kinds of resources. The goal of these organizations is to set up a universal and well accepted Chinese resources database so that to push forward the Chinese Natural Language Processing.

XI. **Alexandre Trilla 2009** depicted the usage of Natural Language processing techniques in the production of voice from an input text. Text-To-Speech synthesis, and the inverse process, which is the production of a written text transcription from an input voice utterance that is Automatic Speech Recognition.

XII. **Cowie et al. 1996** suggested that sadness in speech decreases the speech intensity and its dynamic changes. The average pitch was at the same level as in neutral speech, but there are almost no aggressive changes. The articulation clarity and the speech rate were also decreased. High ratio of halt to phonation time also occurs. Grief was an extreme form of sadness where the average pitch was lowered and the pitch range was very narrow. Speech rate was very slow and pauses for almost a half of the total speaking time.

XIII. **Waters et al. 1993** suggested that fluent speech was also emphasized and punctuated by facial expressions. With visual information added to synthesized speech it was also possible to increase the intelligibility significantly, especially when the auditory speech was degraded by noise, bandwidth filtering, or hearing impairment.

XIV. **Beskow 1996** suggested that speech communication relies not only on consultation, but also on visual information. Facial movements, such as eye blinking, head nodding, smiling, grinning and eyebrow rising gave important additional information of the speaker's emotional state. The emotional state might be even resulted from facial expression without use of any sound.

XV. **Akshay S. Utane et al 2013** implemented speech emotion recognition systems using several classifiers. The classifiers used to distinguish attribute like emotions such as neutral ,surprise ,anger ,happy, sad, fearful, disgust ,etc. emotional speech samples are used as database for emotion recognition from speech and extracted features from speech samples are prosodic and spectral features such as pitch, energy, formants, speech rate ,(MFCC) Mel frequency cepstrum coefficient and linear prediction cepstrum coefficient (LPCC).the performance of classifiers represented by extracted features. Advantages and performance of speech emotion recognition system using different types of classifiers are also discussed

.

IV. DIGITAL SIGNAL PROCESSING TECHIQUES

Digital signal processing is a mathematical manipulation of an information signal to modify and improve signal. They are of many types some are discussed here.

Discrete-time Fourier transform (DTFT) is one of the specific forms of analysis. It transforms one function into another, which is called the frequency domain depiction, or simply the "DTFT", of the original function. The DTFT requires an input function that is discrete. Such inputs are often created by digitally sampling a continuous function, like a person's voice. The DTFT frequency domain representation is always a periodic function. Since one period of the function contains all of the unique information, it is sometimes convenient to say that the DTFT is a transform to a "finite" frequency-domain, rather than to the entire actual line.

Discrete Fourier transform (**DFT**) converts a finite list of equally spaced samples of a function into the list of coefficients of a finite combination of compound sinusoids, ordered by their frequencies, that has those same sample merits. It can be said to convert the sampled function from its native domain to the domain. The input samples are complex numbers (in practice, usually real numbers, and the output coefficients are complex as well. The frequencies of the output sinusoids are integer multiples of a basic frequency, whose corresponding period is the length of the sampling interval. The mixture of sinusoids obtained through the DFT is therefore periodic with that same period. The DFT vary from the (DTFT) in that its input and output sequences are both finite; it is therefore said to be the Fourier analysis of finite-domain discrete-time functions.

Bilinear transform is used in digital signal processing and discrete-time control theory to transform continuous-time system representations to discrete-time and vice versa .It is a special case of a conformal mapping often used to convert a transfer function $H_a(s)$of a (LTI) filter in the continuous-time domain to a Transfer function $H_d(z)$of a linear, Shift-invariant filter in the discrete time domain.

V. CONCULSIONS

In this paper work, we have presented a survey of all techniques that had been used for the modification of a TTS speech signal to enhance its naturalness. The papers used in this shows the importance of the parameter selected in speech processing. Approached algorithms have been presented. Still the framework to evaluate all the mentioned parameters, have to be set up.

REFERENCES

[1] Luthra, S. & Singh, P., “Punjabi Speech Generation System based on Phonemes”, International Journal of Computer Applications, Vol. 4, No. 13, pp. 40-44, July 2012.

[2] Kumar, D. & Rana, N., “Speech Synthesis System for Online Handwritten Punjabi Word: An Implementation of SVM & Concatenative TTS”, International Journal of Computer Applications, Vol. 26, No. 2, pp. 13-17, July 2011.

[3] Rabiner, L. & Schafer, R., “Introduction to Digital Speech Processing”, Vol. 1, No. 1-2, pp.1-194, 2007.

[4] Bjorkan, I., “Speech Generation and Modification in Concatenative Speech Synthesis”, PhD. Thesis.

[5] Department of Electronics and Concatenative Speech Synthesis”, M.Sc. Thesis, University of Crete, Greece, pp. 1-18, 2010.

[6] Singh, D. & Singh, P., “Removal of Spectral Discontinuity in Concatenated Speech waveform”, International Journal of Computer Applications, Vol. 53, No. 16, pp. 13-17, September 2012.

[7] Mousa, A., “Voice Conversion Using Pitch-Shifting Algorithm by Time Stretching with PSOLA and Re-sampling”, Journal of Electrical Engineering, Vol. 61, No. 1, pp. 57-61, 2010.

[8] Kaur, I. & Kaur, M., “Prosody Modification of Recorded Speech in Time-Domain”, International Journal of Computer Applications, pp. 25-28, 2012.

[9] “Pitch Shifting”, Available: http://www.katjaas.nl/pitchshift/pitchshift.html, [Accessed: Oct 28, 2013].