Ascent Correction in Text to Speech Using Segmentation Methodology

Miss.Nutan S.Raut, P.GDepartment of Computer science, Engg SGBAU, Amravati,India, nutan.raut@yahoo.co.in
Dr. Mrs.Sujata N.Kale, Assistant Professor in PG department, Applied Electronics SGBAU, India, sujatankale@rediffmail.com

Dr. V. M. Thakare, HOD Of Computer Science & Engg Dept,SGBAU, Amravati,India, vilthakare@yahoo.co.in

Abstract— Text-To-Speech System (TTS) is a computer-based system that automatically converts text into artificial human speech. To build a natural sounding speech synthesis system, it is essential that the text processing component produce an appropriate sequence of phonemic units corresponding to an arbitrary input text. Text processing and speech generation are two main components of a text to speech system. The process of transforming text into speech contains coarsely two phases: first the text goes through analysis and then the resulting information is used to generate the speech signal.proposed PSOLA(Pitch synchronous overlap add technique)can change flexibly the rhythm of speech naturalness without changing the details of the original speech segments to obtain a higher clarity and naturalness. MFCC is based on the human peripheral auditory system. The performance of the Mel-Frequency Cepstral Coefficients (MFCC) may be affected by the number of Filters and type of window.

Index Terms: Text to speech synthesis, target cost, concatenation cost, unit selection, MFCC,PSOLA.

1. INTRODUCTION

The Efficient Methodology for segmentation of speech signals in TTS(Text to speech) provides the optimal speech synthesis approach. A text to speech synthesis system converts a given text to corresponding speech output. Concatenative synthesis uses actual short segments of recorded speech that were cut from recordings and stored in an inventory (‘‘voice database’’), either as ‘‘waveforms’’ (uncoded), or encoded by a suitable speech coding method. Synthesis systems are commonly evaluated in terms of three characteristics: accuracy of rendering the input text, intelligibility of the resulting voice message, and perceived naturalness of the resulting speech.

2. BACKGROUND

Text-to-speech or speech synthesis does not fall into any one traditional academic discipline, and hence the background knowledge may vary greatly depending on a particular readers. The generation of the speech signal can also be divided into two sub-phases: the search of speech segments from a database, or the creation of these segments, and the implementation of the prosodic features. Prosody is a concept that contains the rhythm of speech, stress patterns and intonation. The idea behind text-to-speech is to “play back” messages that weren’t originally recorded. One step away from simple playback is to record a number of common words or phrases and recombine them, and this technique is frequently used in telephone dialogue services. Sometimes the result is acceptable, sometimes not, as often the artificially joined speech sounded stilted and jumpy. This allows a certain degree of flexibility, but falls short of open ended flexibility. Text-to-speech on the other hand, has the goal of being able to speak anything, regardless of whether the desired message was originally spoken or not.

3. PREVIOUS WORK DONE

Sudhakar Sangeetha [1] proposed the Spectral smoothing in which individual smoothness for syllable in Tamil text-to-speech, a time scale modification is carried out for each syllable. Smoothing at concatenation joints is performed using Mel-LPC. Concatenative unit selection speech synthesis systems have been found to be as intelligible as human speech. V. Ramu Reddy [2] proposed the two-stage FFNN model, first the tilt parameters are derived from the linguistic and production constraints using FFNN. Later, the derived tilt parameters are used for predicting the intonation contours. N.P.Narendra [3] proposed the Optimal weight tuning method based on genetic algorithm that can relate weights of subcosts with human perception. The genetic algorithm evaluates the fitness function in order to find the inputs which can produce minimum value of the fitness function Francesc Alıas [8] proposed weight tuning by using active interactive genetic algorithms and Obtaining subjective weight patterns at cluster level by Active interactive genetic algorithms (aiGAs). Adriana Stan [9] proposed the first-order all-pass frequency-warping function and The Bark and ERB scales using the first-order all-pass function.The Bark and equivalent rectangular bandwidth (ERB) scales are also well-known auditory scales. Aimilios Chalamandaris [10] proposed the NLP(Natural Language Processing unit) adapation module by using screen reading environments & take special care of natural language processing, speed & quality optimization. A number of methods can be used for automatic segmentation are Automatic segmentation, Fourier Transform, Short Term Energy of

Speech Signal, Minimum Phase Group Delay method for segmentation of speech into syllables, wavelet techniques, Word Chopper method.

4. PROPOSED METHODOLOGY

The Mel Frequency Cepstral Coefficients (MFCC) technique is used to extract features from the speech signal and compare the unknown speaker with the exits speaker in the database. The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency *t* measured in Hz, a subjective pitch is measured on a scale called the 'Mel Scale' .The mel frequency scale is a linear frequency spacing below 1000 Hz and logarithmic spacing above 1kHz.The MFCC technique in speech synthesis used for joining two speech segments S1 and S2. Represent S1 as a sequence of MFCC, Represent S2 as a sequence of MFCC. Join at the point where MFCCs of S1 and S2 have minimal Euclidean distance. The windowing block minimizes the discontinuities of the signal by tapering the beginning and end of each frame to zero. The FFT block converts each frame from the time domain to the frequency domain. In the Mel-frequency wrapping block, the signal is plotted against the Mel spectrum to mimic human hearing In the final step, the Cepstrum, the Mel-spectrum scale is converted back to standard frequency scale .

PSOLA refers to a family of signal processing techniques that are used to perform time-scale and pitch-scale modification of speech. These modifications are performed without performing any explicit source/filter separation. PSOLA works by dividing the speech waveform in small overlapping segments. To change the pitch of the signal the segments are moved further apart (to decrease the pitch) or closer together (to increase the pitch).To change the duration of the signal, the segments are then repeated multiple times(to increase the duration)or some are eliminated(to decrease the duration).The segmentation are then combined using the overlap add technique. PSOLA can be used to change the prosody of a speech signal.

All PSOLA techniques comprise of

- Isolate pitch periods in the original signal.
- Perform the required modification .
- Resynthesize the final waveform through an overlap-add operation.

5. ANALYSIS & DISCUSSION

In speech analysis, a sequence of pitch-marks is provided after filtering the speech signal. Voiced/unvoiced decision is based on the zero-crossing and the short time energy for each segment between two consecutive pitch marks. To select pitch marks among local extreme of the speech signal, a set of mark candidates given with all negative and positive peaks. The OLA synthesis is based on the superposition-addition of elementary signals in the new positions. These positions are determined by the height and the length of the synthesis signal. To increase the pitch, the individual pitch-synchronous frames are extracted and given to Hamming window. Then output frame moved close together and added up, whereas output frame moved further apart to decrease the pitch. Increasing the pitch will result in a shorter signal, so to keep constant duration duplicate frames need to be added. A fast resampling method is used to shift the frame precisely, where it will appear in the new signal using the pitch mark and the synthesis mark of a given frame. The analysis of Pitch synchronous mainly makes the voice synthesis unit synchronized. For PSOLA technology, the interception and superposition of short-time signal and the choice of time window length signal keep synchronous. Voiced segment has pith period, while the voiceless section of the signal is white noise, so it needs to distinguish the two types. When the pitch of voiced signals is marked, the pitch period of voiceless is generally a constant quantity to ensure consistency of the algorithm.

6. RESULT ANALYSIS

It is clear that TTS systems have come a long way towards delivering high-quality output to listeners. An optimal unit selection algorithm is used to reduce redundancy in the text corpus. The contribution of tilt parameters to PCPA features is significant in improving the quality of TTS system at perceptual level. The tonal context labeling propose can generate an F0 contour closer to that of natural speech than the other techniques. Examining different compositions of hybrid utterances, we found that almost all natural segments are replaced by statistical models in a HTTS based on the 5 MB baseline CTTS.

7. CONCLUSION

Genetic algorithm is used to adjust the weights of subcost such that the ranking obtained from the total cost of all instances and the ranking obtained from perceptual preference tests are nearly same. Fitness function is designed such that weights of the cost function can select the units which are perceptual preferred by listeners. Spectral discontinuities are lowered at unit boundaries based on the Mel-LPC method. Thus unrestricted TTS have been developed. TTS system produces the synthesized speech with naturalness and good quality using MFCC and PSOLA.

8. FUTURE WORK

The Efficient Methodology for segmentation of speech signals in TTS(Text to speech), provides the optimal speech synthesis approach. Prosody plays an important role in an improving the quality of text-to-speech synthesis (TTS) system both in terms of naturalness and intelligibility. Prosody refers to duration, intonation and intensity patterns of speech for the sequence of syllables, words and phrases. All the methods have a common goal to proposed TTS system can be generalized to all the languages & also to improve the quality of speech.

There is a scope of further improvement in terms try to improve this system to be a text independent speaker identification system. The segmentation performance, and the better segmentation rates can be obtained by increasing the number of training speech sentences for future work, and also include the text with video.

REFERENCES

1. Sudhakar Sangeetha,Sekar Jothilakshmi," Syllable based text to speech synthesis system using auto associative neural network prosody prediction ",Springer Science+BusinessMedia,Vol.17,No.2,PP.91-98,June 2014.
2. V. Ramu Reddy and K. Sreenivasa Rao," Two-stage intonation modeling using feedforward neural networks for syllable based text-to-speech synthesis", Science Direct(SciVerse ScienceDirect) on Computer Speech and Language, VOL. 27, NO. 5, PP.1105-1126. Aug-2013.
3. N.P.Narendra,K.Sreenivasa Rao," Optimal weight tuning method for unit selection cost functions in syllable based text-to-speech synthesis". Science Direct on Applied Soft Computing,VOL.13, NO. 2, PP.773-781, Feb-2013.
4. N. P. NARENDRA and K. SREENIVASA RAO," Syllable Specific Unit Selection Cost Functions for Text-to-Speech Synthesis", ACM Trans. Speech Lang. Process , Vol. 9, No. 3, : November 2012.
5. Vataya Chunwijitra ,Takashi Nose,Takao Kobayashi ," A tone-modeling technique using a quantized F0context to improve tone correctness in a verage-voice-based speech synthesis", Science Direct (SciVerse ScienceDirect) On Speech communication", Vol.54, No.2, PP.245-255 , Feb-2012.
6. Archana Balyan ,S. S. Agrawal , Amita Dev," Automatic phonetic segmentation of Hindi speech using hidden Markov model", Springer Science, vol.27, No.4,P.P.543-549. Nov-12.
7. Stas Tiomkin,david Malah,Life Fellow,"A Hybrid Text to speech system that combines concatenative & statistical synthesis unit", IEEE transaction on audio, speech and language Programming,VOL.19, NO. 5, PP.1278-1287, July 2011.
8. Francesc Alıas, Lluıs, Formiga , Xavier Llora,"Efficient and reliable perceptual weight tuning for unit-selection text-to-speech synthesis based on active interactive genetic algorithms: A proof-of-concept", ScienceDirect On Speech Communication,VOL.53, NO. 5, PP.786-800, May-2011.
9. Adriana Stan, Junichi Yamagishi, Simon King , Matthew Aylett," The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate", ScienceDirect On speech Communication, Vol.53,No.3,PP.442-450,March2011.
10. Aimilios Chalamandaris, Sotiris Karabetsos ,"A Unit Selection Text-to-Speech Synthesis System Optimized for Use with Screen Readers", IEEE Transaction on Consumer Electronics,Vol. 56, No. 3, PP.1890-1897,August 2010.