Removal of Spectral Discontinuity in Concatenated Speech Waveform

Deepika Singh
Department of Computer Science and Engineering
Guru Nanak Dev Engineering College, Ludhiana, Punjab, India

Parminder Singh
Associate Professor, Department of Computer Science and 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 systems are known for producing high quality, natural-sounding speech as they generate speech by joining together waveforms of different speech units. This method of speech generation is quite practical. However the speech units that are being concatenated may have different spectra on either side of the concatenation points. Such mismatches are spectral in nature and give rise to spectral discontinuity in concatenated speech waveforms. The presence of such discontinuities can be very distracting to the listener and degrade the overall quality of output speech. This paper proposes a speech signal processing technique that deals with the problem of spectral discontinuity in the context of concatenated waveform synthesis. It involves the post-processing of the synthesized speech waveform in time domain. This technique is implemented on different single channel Punjabi wave audio files which were created by concatenating different Punjabi syllables. A listening test was conducted to evaluate the proposed technique, and it was observed that the spectral discontinuity is reduced to a large extent and the output speech sounds more natural with the reduction of audible noise.

General Terms
Technique for speech signal processing

Keywords
Speech waveform, Concatenative speech synthesis, Spectral discontinuity

1. 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 [1]. The mechanism for generating the human voice can be subdivided into three parts; the lungs, the vocal folds within the larynx, and the articulators (the parts of the vocal tract above the larynx consisting 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. These vibrations themselves can be represented by speech waveforms. Figure 1 shows a visual representation of vibrations typical of those in human speech - a speech waveform for a Punjabi word "ਦਰ".

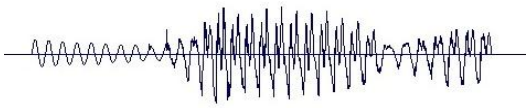

Figure 1: Example Speech Waveform for Punjabi word- "ਦਰ"

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 judged 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.

Three types of methods are mainly used for the purpose of synthesizing artificial speech- Articulatory Synthesis, Formant Synthesis and Concatenative Synthesis [2]. The articulatory and formant synthesis are the rule-based synthesis methods whereas the concatenative technique is a database-driven synthesis method. Articulatory synthesis uses a physical model of human speech production organs and articulators. Formant synthesis models the frequencies of speech signal based on source-filter model. In this method of speech synthesis, parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a speech waveform based on certain rules. Concatenative synthesis generates speech by concatenating recorded speech units and is described in more detail in Section 2.

The remainder of this paper is organized into 6 sections. Section 2 presents an overview of concatenative speech synthesis. In section 3, the problem of spectral discontinuity in the context of concatenative speech synthesis is discussed. Section 4 explains the stages of the technique proposed to remove audible spectral discontinuities in concatenated speech waveform. Section 5 evaluates the results of the proposed technique. Finally we end our paper with Conclusions and Future work in Section 6.

2. CONCATENATIVE SPEECH SYNTHESIS

Concatenative speech synthesis is currently the most practical method for the generation of realistic speech. 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: 1) Unit Selection Synthesis, 2) Diphone Synthesis and 3) Domain-Specific Synthesis [3]. The unit selection based systems may make use of very large databases of fluent speech units which may include phrases, words, syllables or phonemes. For the greatest fluency, finding the best available unit and choosing precise concatenation points is important. Diphone synthesis based systems generate speech by joining together diphones 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 diphones produce clear speech and desired prosody can be modelled for a particular context by using signal processing. The domain-specific synthesis is used in systems where the variety of output 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.

Despite the advantages, there are also a few problems associated with the concatenative synthesis method. Collecting all the required variations of speech units and labelling of speech samples and then constructing a large acoustic database are the major challenges. Memory requirements of this method are higher in comparison to other methods. Selection of the most suitable units for creating the desired utterance is time-consuming and labour-intensive. Another major problem in concatenative synthesis is that at some points occasional spectral discontinuity has to be tolerated. This usually happens when the units that are being concatenated have different prosodic and phonetic contexts. Such discontinuities may be audible as distortion and may also degrade the quality of speech generated to a large extent. Therefore, smoothing is necessary to lower this mismatch effect at the concatenation points.

3. SPECTRAL DISCONTINUITY IN CONCATENATED SPEECH

When the join between two speech units is clearly audible, it refers to discontinuity. The mismatch in spectra of the speech units on either side of join causes this discontinuity. Audible spectral discontinuities in concatenated signals were researched in [4]. Signal components can change in a number of ways at the join; an abrupt termination of signal components, an abrupt onset of signal components and more subtle changes in signal components sustained across the join [5]. The synthesised speech can sound very natural if the discontinuities at the concatenation points are inaudible. But when these joins are audible, their presence can be very frustrating to the listener and it also reduces the overall perceived quality of synthesized speech.

In systems which use databases containing longer speech units and where the variety of output is limited, the problem of spectral discontinuity is less severe. This is because with longer speech units, there will be lesser concatenation points. However, in systems which create speech by combining large number of smaller speech units, the presence of spectral discontinuity at the concatenation boundaries is a major problem; since there is an increase in the number of joins, therefore, there is an increase in the number of discontinuities. There are a number of reasons for the presence of spectral discontinuities. Audible discontinuity may arise due to inconsistencies in fundamental frequencies, or different levels of loudness (energy of the segments), or due to the contextual differences and variations of speaking style of the speaker [6].

In order to avoid the problem of spectral discontinuity at concatenation boundaries, an appropriate signal processing technique must be applied. Ideally, a signal processing approach would include algorithms that will examine the synthetic speech waveform at concatenation points and then manipulate the waveform at these points to produce a more natural sounding continuity. In the next section we propose one such signal processing technique to reduce the effect of spectral discontinuities in the original acoustic signal.

4. PROPOSED TECHNIQUE

The block diagram in figure 2 gives an outline of the proposed technique.

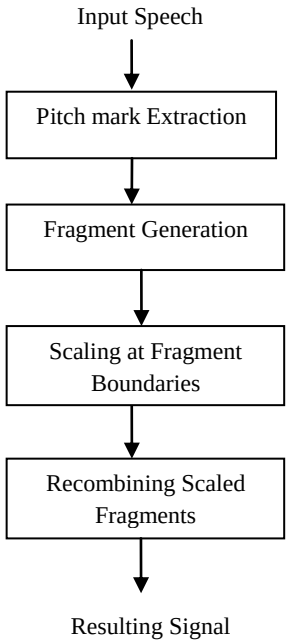

Figure 2: Block Diagram of the Proposed Technique

The proposed technique works on speech signals directly in time-domain and allows them to be processed and modified in real time. The input for this technique is a concatenated Punjabi speech waveform that is generated

by joining together different syllables of Punjabi. Our goal is to create as output a final speech waveform that is free from distortion even when its tempo is increased or decreased. The working of the proposed technique and the different stages involved in it will be discussed in the following sub-sections.

4.1 Extraction of Pitch-marks

The first stage of the proposed technique involves determining the pitch-mark locations within the input speech utterance. Pitch-marks are certain maximum points along a wave that correspond to glottal openings in the human voice box producing a rush of air, and usually mark one cycle of the particular vibration used to create the sound [8]. These appear as high peaks in speech waveforms and hence, can be located easily.

Pitch-marking is a crucial stage as its results can greatly affect the quality of the resulting signal. This is because pitch-marks are used to determine the centre and the width of each speech fragment that will be generated during the second stage. Studies show that synthetic pitch marks are the most important aspect that can cause quality degradation. So it becomes extremely important to use an efficient pitch-marking algorithm. We now describe our algorithm for the extraction of pitch-marks.

INPUT: Data chunk part of the original concatenated speech signal
OUTPUT: Pitch-marks present in the original signal
STEPS:

i. Store the data samples of the input wav audio signal in an array.
ii. Divide these original samples into small blocks of fixed length. We used a frame size of around 500 samples for determining pitch marks.
iii. Find the sample value with maximum amplitude in each of these fixed-length blocks in succession. These are the required pitch-marks as shown in figure 3(a).
iv. Preserve these pitch-marks for use during the next stage for the generation of fragments. These pitch marks are the analysis marks. Signal modification involves manipulations of analysis marks to generate synthesis marks. Synthesis marks are required for the generation of modified speech.

Note that speech is characterized as voiced and unvoiced. It is easier to find pitch marks in the voiced regions but unvoiced speech is mainly composed of random noise. However, pitch-marks must still be assigned in the unvoiced portions of the speech, although it is rather arbitrary where to place the pitch-marks.

4.2 Generation of Speech Fragments

The second stage of the proposed technique consists of generation of short-time speech fragments. Once the pitch-marks are extracted from the original speech signal, the signal is decomposed into small fragments using these pitch-marks. Each of these pitch-marks is taken as the centre of a different fragment in succession. Figure 3 shows the process of generation of speech fragments from pitch-marks in a speech signal. In this figure, N, N+1, N+2 and N+3 denote speech frames of original signal in succession and P1, P2, P3 and P4 are the respective pitch-marks in each of these fixed-length frames of samples. These pitch-marks give the sample values with maximum amplitudes and have been extracted using the algorithm described in Section 4.1.

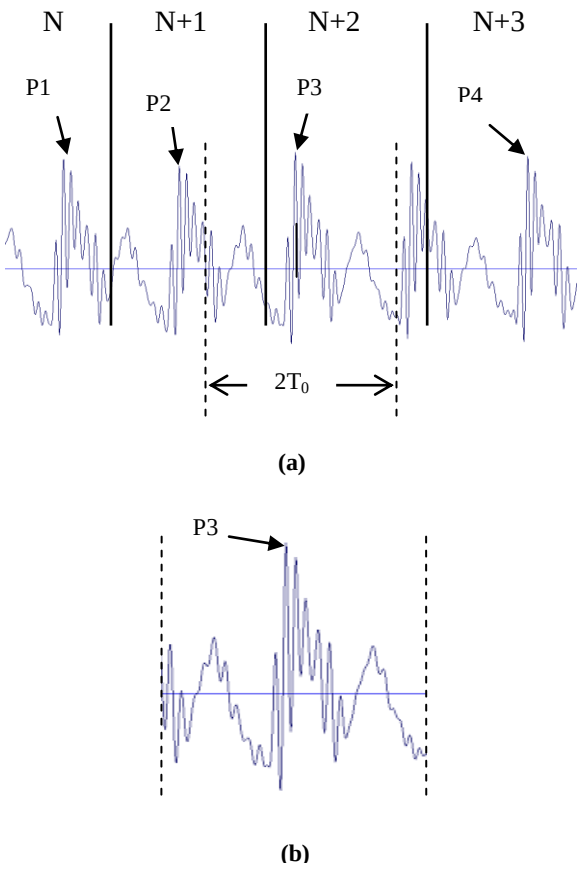

Figure 3: Generation of Speech Fragments- (a) Pitch-mark isolated in each of the fixed-length frames in succession; (b) a fragment lifted from around the pitch-mark P3 located in frame N+2

After a pitch-mark P3 has been located in frame N+2 in figure 3(a), a fragment is lifted from around this pitch-mark and is extended approximately one pitch-period in both directions. Note that generated fragment is of the size of twice the pitch period T_0 as shown in the Figure 3 (b).

4.3 Generation of the Final Speech Waveform

The third and final stage of our proposed method involves the creation of final speech where the speech fragments generated during the second stage are recombined. This involves adding together the fragments generated during the second stage with no overlap. It means fragments are added in such a way that the next fragment begins after the current fragment ends and there is no overlapping between any two fragments.

The most important aspect of this stage is to preserve the moments of principle vocal excitation. It is because this is what one hears the most and as these correspond to the pitch-marks, they are more or less unchanged from the original waveform segment. The resulting sound waveform will comprise of all the original fragments, spaced out according to the given pitch information. This stage of final waveform generation involves two important steps. These are- scaling of speech fragments and change of

tempo or play rate of the output speech. These steps are discussed in the following subsections.

4.3.1 Changing the Tempo of Output Speech Signal

The tempo of the resulting speech signal is also changed using the proposed technique. Tempo signifies the play rate or the utterance rate of the output speech. The tempo of speech, in turn, affects the duration of speech. A faster speech utterance is obtained by omitting certain constituent speech fragments while generating the output signal such that even when its tempo is increased, the loss of information content is minimal. For example, tempo is increased when every fifth speech frame is skipped from the original signal. Loss of speech information becomes quite obvious in this case. Similarly, a slower speech utterance is obtained by repeating certain constituent speech fragments. For example, when every third frame is repeated in the speech signal, play rate decreases and output speech produced is accompanied by an echo effect.

4.3.2 Scaling of Speech Fragments

Spectral discontinuities can be perceptible at various points in the final speech waveform. One possible reason may be the abrupt beginning or ending of the waveforms of speech segments. This abrupt onset and offset occurs since units are taken from different prosodic and phonetic contexts. It is important to deal with these audible discontinuities, since their presence degrades the quality of speech. In the proposed technique, we used waveform scaling along the time axis to remove the audible distortion. This waveform scaling is performed at the beginning and at the end for every composite speech fragment that were generated during the second stage. Last few sample values of the previous speech fragment are scaled down so that there is a gradual fading of the signal. Similarly, first few sample values are scaled up at the beginning of the current speech fragment, so as to avoid an abrupt start of the signal. Around 20 samples were scaled on both sides of the speech fragment. This method affects duration and spectrum, although no degradation of the naturalness is caused when the scaling ratio is nearly 1. Note that further signal modification leads to the degradation of output signal. Using this method, the audible noise is suppressed to a great extent and the quality of the signal is improved.

5. RESULTS

Our aim was to remove the spectral discontinuity from the original concatenated Punjabi speech waveform. The proposed technique was analysed to check if the desired results were produced. For this purpose, we used a few 16-bit mono channel concatenated Punjabi wav audio files.

5.1 Mean Opinion Score (MOS)

The working of this technique is similar to the pitch synchronous analysis and synthesis framework of Time Domain Pitch-Synchronous Overlap and Add (TD-PSOLA) technique. So the results of the two techniques were compared using a listening test. Although the resulting speech sounds were similar in both these techniques, but audible distortion was present at some points in the speech generated by TD-PSOLA. This is because no signal modification was done to reduce this audible distortion. However, in our proposed technique, waveform scaling along the time axis was applied for all the speech fragments in order to deal with spectral discontinuities. After scaling the sample values at the beginning and end of every speech fragment, it was observed that the distortion was reduced to a large extent and the audible quality also improved.

A listener test was also conducted to evaluate the results of the proposed technique. 6 listeners were asked to rate the quality of the speech synthesized using the proposed technique for a number of sentences. Each listener was required to give each sentence a rating on a scale from 1 to 5, where 1 represents the lowest perceived audio quality whereas 5 represents the highest perceived audio quality, as shown in the Table 1. Mean Opinion Score (MOS) is the arithmetic mean of all individual scores and gives the numerical indication of the perceived audio quality.

Table 1. Parameters for Mean Opinion Score (MOS)

MOS	Quality	Distortion
5	Excellent	Imperceptible
4	Good	Slightly imperceptible
3	Fair	Slightly Annoying
2	Poor	Annoying
1	Bad	Very Annoying

The rating given by each listener suggests that the proposed technique is successful in achieving our goal of generating an output speech with minimal audible distortion.

5.2 Analysis of Speech Tempo

It was noted that the speech waveform synthesized using the proposed technique was visually similar to the waveform of original speech (Figure 4). This is because of the preservation of pitch-marks.

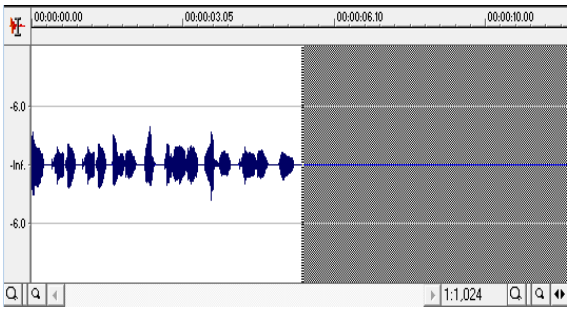

Figure 4: Original Speech Waveform

As mentioned before, the utterance rate of the original speech was also modified with the help of proposed method which, in turn, affected its overall duration. A slower speech was produced with the repetition of certain speech fragments increasing the overall duration of the final speech. Figure 5 shows the waveform of speech with slower tempo. A great elongation of the synthesized signal causes the echo effect. Similarly, omitting certain fragments in the final speech signal resulted in the production of a faster speech, reducing the overall duration. However, omission of fragments causes the loss of speech information content. Figure 6 shows the waveform of speech with faster utterance rate. Note the change in duration with the change in the tempo of speech in the waveforms.

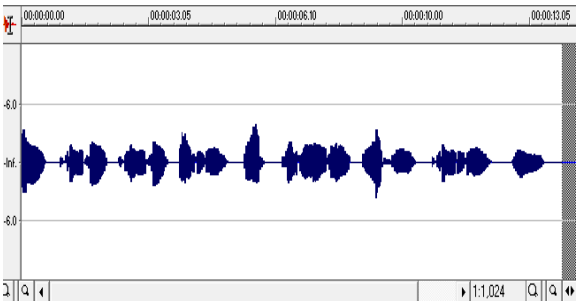

Figure 5: Speech Waveform with Slower Play Rate

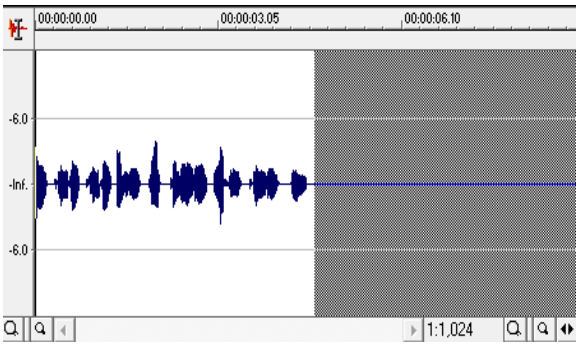

Figure 6: Speech Waveform with Faster Play Rate

6. CONCLUSIONS AND FUTURE WORK

Concatenation-based systems may not always produce consistent perceptual continuity due to spectral discontinuities at the concatenation points. In this paper, we have proposed a method for the removal of spectral discontinuity. Although the proposed technique works in a manner similar to TD-PSOLA but the results produced by these two methods are different. This was confirmed by conducting a listening test. It was concluded that TD-PSOLA technique is not suitable for dealing with spectral discontinuities, whereas the proposed technique performs waveform scaling at both ends of each speech fragment while creating the final speech. Using the proposed technique, the play rate of resulting speech was also increased or decreased, thus, affecting its overall duration.

There are also a few potential areas for future improvement. One of the key points is that the technique proposed here works only for mono channel wav files. This technique can be modified in future so that it can also be used for stereo channel wav files. Another important point is that this technique works with a limited database. In future, this technique can be further enhanced for use with larger database in concatenation-based speech systems.

7. REFERENCES

[1] Honda. M. (2003), "Human Speech Production Mechanisms", NTT Technical Review, Vol. 1, No. 3, pp. 24-29.

[2] Tabet. Y. And Boughazi. M. (2011), "Speech Synthesis Techniques: A Survey", 7th International Workshop on Systems, Signal Processing and their Applications, pp. 67-70.

[3] Thakur. S. K. and Satao. K. J. (2011), "Study of Various kinds of Speech Synthesizer Technologies and Expression for Expressive Text To Speech Conversion System", International Journal of Advanced Engineering Sciences and Technologies, Vol. 8, No. 2, pp. 301-305.

[4] Chappell. D. And Hansen. J. (2002), "A Comparison of Spectral Smoothing Methods for Segment Concatenation Based Speech Synthesis", Speech Communication, Vol. 36, pp. 343-374.

[5] Kirkpatrick. B. (2010), "Spectral Discontinuity in Concatenative Speech Synthesis - Perception. Join Costs and Feature Transformations", PhD. Thesis, Dublin City University, pp. 1-63.

[6] Klabbers. E. and Veldhuis. R. (2001), "Reducing Audible Spectral Discontinuities", IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 1, pp. 39-51.

[7] White. S. (2003), "Visualizing Speech Synthesis", Bachelor's Thesis, pp. 4-9.

[8] Lemmetty. S. (1999), "Review of Speech Synthesis Technology", Master's Thesis, Department of Electrical and Communication Engineering. Helsinki University of Technology, pp. 28-46.

[9] Bjorkan. I. (2010), "Speech Generation and Modification in Concatenative Speech Synthesis", PhD. Thesis. Department of Electronics and Concatenative Speech Synthesis", M.Sc. Thesis, University of Crete, Greece, pp. 1-18.

[10] Visagie. A. (2004), "Speech Generation in a Spoken Dialogue System", Master's Thesis, University of Stellenbosch, South Africa, pp. 35-91.

[11] Wouters. J. And Macon. M. (2001), "Control of Spectral Dynamics in Concatenative Speech Synthesis", IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 1, pp. 30-38.

[12] Klabbers. E. (1997), "High Quality Output Speech Generation through Advanced Phrase Concatenation", Proceedings of the Cost Workshop on Speech Technology in the Public Telephone Network: Where are we today?, Rhodes, Greece, Vol. 1, No. 88, pp. 85-88.

[13] Rabiner. L. And Schafer. R. (2007), "Introduction to Digital Speech Processing", Vol. 1, No. 1-2, pp.1-194.

[14] Mousa. A. (2010), "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.

[15] Plumpe, M. And Meredith, S. (1998), "Which is More Important in a Concatenative Text to Speech System- Pitch, Duration or Spectral Discontinuity?", Proceedings of the third ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan, Australia.