Voice Conversion Method for Myanmar Speech Synthesis

Tar yar Myo Tun
University of Computer Studies, Mandalay
taryarmyotun@gmail.com

Abstract

Speaker independent speech recognition is important for successful development of speech recognizer in most real world applications. This paper emphasizes speaker independent based on voice conversion usinga pitch shifting method,Pitch SynchronousOver Lap Add Algorithm (PSOLA).It depends on detecting the pitch of the signal (fundamental frequency) and changing it according to the target pitch period using time stretching with Pitch SynchronousOver Lap Add Algorithm (PSOLA), then resampling the signal in order to have the same play rate.The traditional voice activity detection algorithms which track only energy cannot successfully identify potential speech from input because the unwanted part of the speech also has some energy and appears to be speech. In the proposed system, VAD (Voice Activity Detection) that calculates energy of high frequency part separately as zero crossing rate to differentiate noise from speech is used.

***Keywords**: Myanmar Speech Synthesis, VAD, Energy, Zero Crossing Rate, TD-PSOLA*

1. Introduction

Speech recognition is the process of automatic extracting and determining linguistic information conveyed by a speech wave using computers. Speech recognition has tremendous growth over the last five decades due to the advances in signal processing, algorithms, new architectures and hardware. Speech processing and recognition are intensive areas of research due the wide variety of applications. Speech recognition is involved in our daily life activities like mobile applications, weather forecasting, agriculture, healthcare, video games etc.

East Asian languages arethe phonology of a tone language. It includes a number of pitch levels or contours that, along with segmental phonemes. Language Technology has a potential to play a major role in the process of learning a language. Myanmar language is like Chinese, Japanese, India and Thailand and so on. Myanmar is a kind of tonal languages. This means that all syllables in Myanmar have prosodic features that are an integral part of their pronunciation. Prosodic contrasts involve not only pitch, but also phonation. The need for speaker independent continuous speech to conversion system lies at the core of many rapidly growing application areas. A speaker independent system is intended for use by any speaker.

2. Related Works

In speech analysis, the voiced-unvoiced decision is usually performed in extracting the information from the speech signals. Two methods to separate the voiced- unvoiced parts of speech from a speech signal. They are zero crossing rate (ZCR) and energy. The results by dividing the speech sample into some segments and used the zero crossing rateand energycalculations to separate the voiced and unvoiced parts of speech. The results suggest that zero crossing rates are low for voiced part and high for unvoiced part where as the energy is high for voiced part and low for unvoiced part. Therefore, the methods are proved more effective in separation of voiced and unvoiced [2].

Voice morphing is the process of producingintermediate or hybrid voices between the utterances of two speakers. It can also be defined as the procedure of gradually transforming the voice of one speaker to that of another. Likeimage morphing, speech morphing aims to preserve the shared characteristics of the starting and final signals, while generating a smooth transition among them. Voice Conversion technique based on Pitch Synchronization Over-Lap Add (PSOLA) algorithm has been developed [1].

In the process of diphone-concatenation synthesis, space complexity and searching time are less than other techniques. The techniques improve the performance of text-to-speech in the Myanmar speech synthesis using the TD-PSOLA (Time Domain Pitch Synchronous Overlap-Add) method. It is based on the signalinto overlapping synchronized frames of the pitch period. The diphone-concatenation of the speech synthesis is to maintain the consistency and accuracy of the pitch marks of the speech signal and diphone database with integrated vowels and consonants of Myanmar language[5].

3. Myanmar language

The Myanmar language is the official language of Myanmar and is more than one thousand years old. The Burmese (Myanmar) language is a tonal and analytic language. The language utilized the Burmese script which derives from the Mon-scripts and ultimately from the Brahmin script. Standard Myanmar is based on the dialect spoken in the lower valleys of theIrrawaddy and Chindwin rivers. It is spoken in most of the country with slight regional variations. In Myanmarthere are 8 main races and 135 sub-races. Myanmar (Burmese) is the official language in Myanmar. A syllable isassigned a tone and each spoken syllable with a different tone will have a different lexical meaning.

Myanmar language is said to have basically 33consonants, 12 vowels, other medial and consonant diphthongs. The basic consonants in Myanmar can be extended by medial. Syllables or words are formed by consonants combining with vowels. . The 33 consonants are represented by 26 phonemes since some consonantal letters represents the same phoneme. Myanmar has basically 12 vowels, 8 monophthongs and 4 diphthongs.Myanmar is a tonal language. This means that all syllables in Myanmar have prosodic features that are an integral part of their pronunciation.The general contrastive features of the four phonological tones offered by the analysis of their fundamental frequency can be described. Myanmar toneme is described with the variety of rate or duration.

However, some syllables can be formed by just consonants, without any vowel. Other characters in the Myanmar script include special characters. Therefore, it is too complex for Natural Language Processing (NLP) purpose.

4. Speech

Speech is the most natural form of human communication. Speech sounds are sensations of air pressure vibrations produced by air exhaled from the lungs and modulated and shaped by the vibrations of the glottal cords and the resonance of the vocal tract as the air is pushed out through the lips and nose.

4.1. Speech Features

Speech features and parameters can be listed as follows:

(a) Spectral envelope of speech is modeled by the frequency response of a linear prediction model of the vocal tract, or by the envelope of the DFT spectrum of speech or the output of a set of filter-banks.

(b) Speech formants, including formant frequencies and bandwidth, and their trajectories in time. Formants are the resonance frequencies of vocal tract cavity; where the spectral envelope’s peaks occur.

(c) The fundamental frequency of the opening and closing of glottal cords, i.e. the pitch.

(d) The temporal dynamics of speech parameters namely the time-variation of the spectral envelope, the formants, and the pitch.

(e) Intonation signals. Intonation signals are conveyed by the temporal dynamics of pitch across a segment of speech.

(f) Stress/emphasis patterns which are functions of pitch intonation, duration and energy.

4.2. Types of SpeechUtterances

Speech recognitions system can be separated by several different classes by describing what types of utterances they have ability to recognize. These classes are classified as the following:

Isolated Words: The speaker has to speak word-by-word into the system. Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on both sides of the sample window. It accepts single wordsor single utterance at a time. Isolated-word recognition systems, with short pauses between spoken words, are primarily used in small vocabulary command control applications.

Connected word systems (or more correctly 'connected utterances') aresimilar to isolatedwords, but allows separate utterances to be 'run-together' with a minimal pause between them.

Continuous Speech: Continuous speech means naturally spoken sentences, separated by minimum silence which is used for detecting boundaries. Recognizers with continuous speech capabilities are some of the most difficult to create because they utilize special methods to determine utterance boundaries.

4.3. Speaker Model

Speaker model for Speech recognition systems can either be speaker dependent or independent, and they can either accept isolated utterances or continuous speech.

4.3.1. Speaker Dependent

Speaker Dependent systems are trained with one speaker and recognition is done only for that speaker. Speaker dependent systems are designed around a specific speaker. This system is used by a exact speaker. So the system needs to recognize different speeches from one speaker. They generally are more accurate for the correct speaker, but much less accurate for other speakers. They assume the speaker will speak in a consistent voice and tempo. Speaker dependent system is designed for one speaker who has trained the system.

4.3.2. Speaker Independent

Speaker Independent systems are trained with one set of speakers and designed for a variety of speakers. A speaker independent system is intended for use by any speaker. Speaker independent recognition is designed for all users without prior training. A speaker independent speech interface would support many valuable applications like telephone directory assistance.In a speaker-independent system, it is necessary to normalize the absolute with respect to the average over the entire utterance, to reduce across-speaker differences.

The gender based differences in human speech are partially due to physiologicaldifferences such as vocal fold thickness or vocal tract length and partially due to differences in speaking style. The female speakers normally have higher formant frequencies as well as higher fundamental frequency (F0). But for male speakers, the F0 is lower, because of the qualities like aggressiveness, body size, self-assurance, and assertiveness. Speaker independent speech recognition works based on the premise that a person voice exhibits characteristics are unique to different speaker. The signal F0 contours of four lexical tones are shown in figure 1.

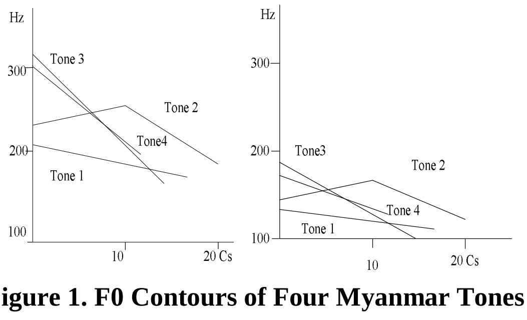


Figure 1. F0 Contours of Four Myanmar Tones with Uttered by Female Speaker (a) and Male Speaker (b)

5. Voice Activity Detection

VAD is widely used to identify the presence of speech in an input signal by marking the boundaries of speech and non-speech segments. VAD is a classification problem in which features of the audio signal are used to separate the input speech and non-speech. The end points accuracy is one of the major factors in recognition performance. VAD can be performed by two algorithms. The first algorithm uses signal features based on energy level and the second algorithm uses signal features based on the rate of zero crossings. The combination of both gives good result that is used in the proposed system. A basic VAD works on the principle of extracting measured features from the incoming audio signal, which is divided into frame size of 120ms duration. The extracted signal features based on energy level and zero crossing rate from the audio signal are then compared to a threshold and then VAD decision is computed. The accuracy and reliability of a VAD algorithm depends heavily on the decision thresholds.

STEP 1: If the feature of the input frame exceed the estimated threshold value, a VAD decision (VAD = 1) is computed which declares that speech is present.
STEP 2: Otherwise, a VAD decision (VAD = 0) is computed which declares the absence of speech in the input frame.

5.1. Choice of Frame Duration

Speech samples that are transmitted should be stored in a signal-buffer first. The length of the buffer may vary depending on the application.Let x(i) be the i-th sample of speech. If the length of the frame was N samples, then the j-th frame can be represented as,

$$f_j = \{x(i)\}_{i=(j-1).N+1}^{j.N} \qquad (1)$$

5.2. Energy of Frame

The energy of the speech signal provides a representation that reflects these amplitude variations.Voiced speech has most of its energy collected in the lower frequencies, most energy of the unvoiced speech is found in the higher frequencies.The most common way to calculate the full-band energy of a speech signal is

$$E_j = \frac{1}{N} . \sum_{i=(j-1).N+1}^{j.N} x^2(i) \qquad (2)$$

where, E_j – energy of the j-th frame and f_j is the j-th frame is under consideration.

5.3. Zero Crossing Rate

The rate at which zero crossings occur is a simple measure of the frequency content of a signal. Zero-crossing rate is a measure of number of times in a given time interval/frame that the amplitude of the speech signals passes through a value of zero.High frequencies imply high zero crossing rates, and low frequencies imply low zero-crossing rates. If the zero-crossing rate is high, the speech signal is unvoiced, while if the zero-crossing rate is low, the speech signal is voiced.

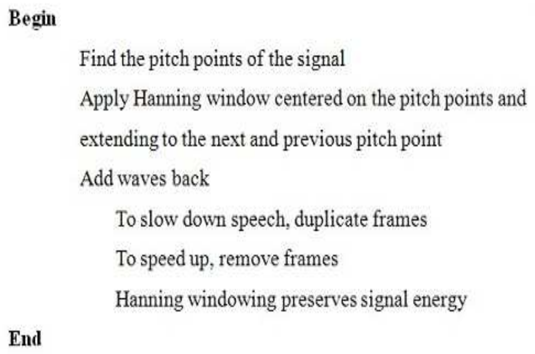
$$z_{s1}(m) = \frac{1}{L}\sum_{n=m-L+1}^{M}\left|\frac{sgn(s(n))-sgn(s(n-1))}{2}\right| \quad (3)$$

$where\ sgn(s(n)) = \begin{cases} +1, s(n) \geq 0 \\ -1, s(n) < 0 \end{cases}$and

1/L =the average value of the zero-crossing measure

6. Pitch Synchronous Over-Lap Add

One of the approaches in order to change the voice is to change the pitch of the voice; this is done by shifting the pitch of the voice using certain techniques like the Pitch Synchronous over Lap-Add (PSOLA) algorithm. There are several types of PSOLA such as Time Domain TD-PSOLA, Frequency Domain PSOLA (FD-PSOLA) and the Linear-Predictive PSOLA (LP-PSOLA). The purpose of TD-PSOLA (Time-Domain) is to modify the pitch or timing of a signal. The process of the TD-PSOLA algorithm is to find the pitch points of the signal and then apply the hamming window centred of the pitch points and extending to the next and previous pitch point. If the speeches want to slow down, the system defines the frame to double. If the speeches want to speed up, thesystem removes the frames in the signal.



```
Begin
    Find the pitch points of the signal
    Apply Hanning window centered on the pitch points and
    extending to the next and previous pitch point
    Add waves back
        To slow down speech, duplicate frames
        To speed up, remove frames
        Hanning windowing preserves signal energy
End
```

Figure 2. TD-PSOLA Algorithm

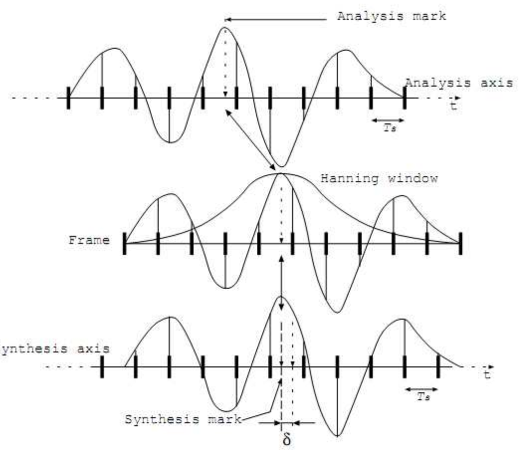


Figure 3. Matching of an Analysis Frame on Synthesis Time Axis

7. Proposed System Design

Our ear cannot response to very fast change of speech data content, we normally cut the speech data into sampling and frame segment before analysis. Human frequency range is between 20Hz and 20K Hz.Frames can be overlapped, normally the overlapping region ranges from 0 to 75% of the frame size.For speech recognition systems normally use for frequency range is 10 to16KHz, sample data bit is 12to16 bit and frame size or frame length is 10 to30ms. The spoken voice frequency lies between 300 to 3400 Hertz. The speech signal isrecorded with the sampling rate of 44100Hz. For reducing the noise signal, speech will be filtering. The output of a filter is a function not only of the input at the present time, but also of previous events. In this system, we use low pass filter with Gaussian window. Figure 4 shows the original signal and output signal of low pass filter with Gaussian window using sampling frequency 44100 Hz for 'ká', 'kã' , 'kâ'.

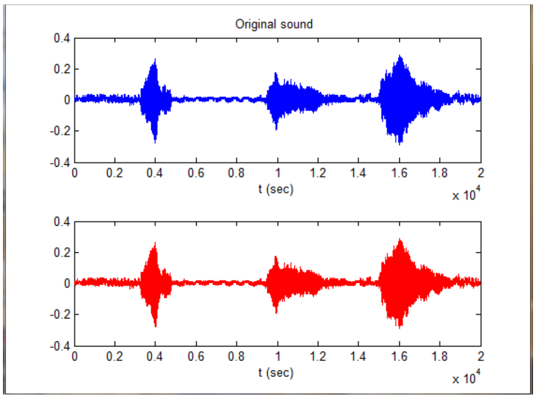


Figure 4. Input and Output Response of Low Pass Filter

Then a short piece of signal is cut out of the whole speech signal.This is done by multiplying the speech samples with a windowing function to cut out a short segment of the speech signal.The time for which the signal is considered for processing is called a window, and the data acquired in a window is called a frame. Features are extracted once every M ms, which is called frame rate, while the window duration is N ms. Typically N is bigger than M. Thus two consecutive frames have overlapping areas.After the preprocessing, VAD is used for end point detection. The process of separating the speech segments of an utterance from the background is called endpoint detection. Frame segment speech signal is calculated the short term energy and zero crossing rate for end point detection by VAD. If zero crossing rate is small and energy is high, we define speech signal is voiced. Otherwise Speech signal is unvoiced. If it is not sure for detection, we calculate subdivision of the frame segment again.

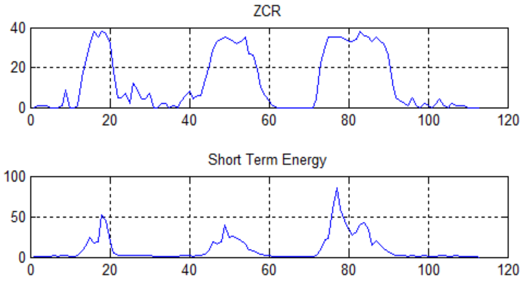


Figure 5. Energy and Zero Crossing Rate

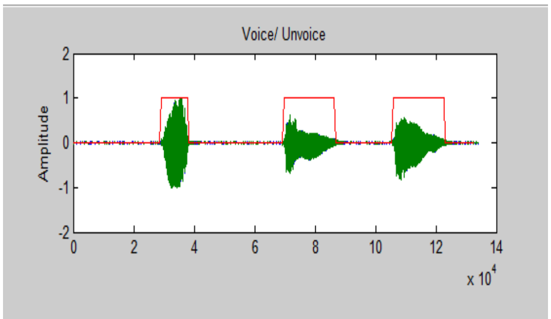


Figure 6.Voice/Unvoice detection

TD-PSOLA requires an exact marking of pitch points in a time domain signal. Pitch marking any part within a pitch period is okay as long as the algorithm marks the same point for every frame.

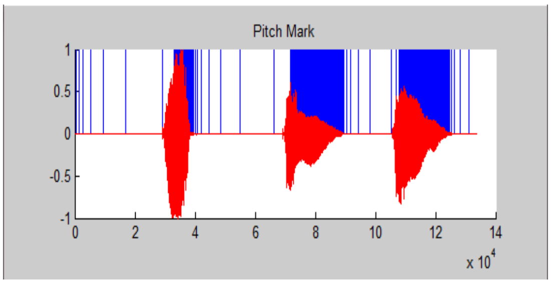


Figure 7. Marking of Pitch Points

The pitch was detected in order to determine the pitchfrequency of the source signal and the target, thenthespeech is converted from female speech to male speechorvisa versa.

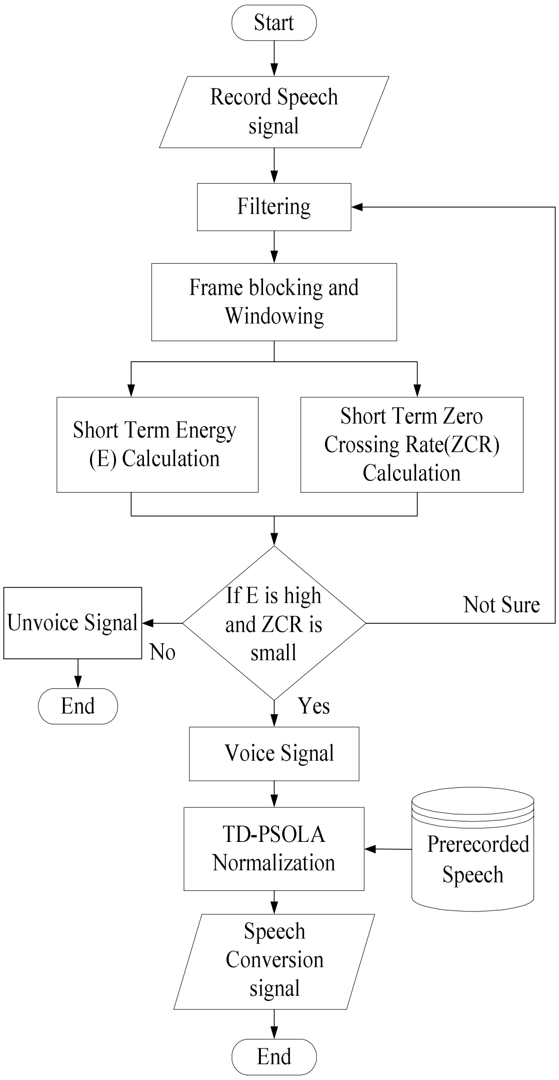


Figure 8. Proposed System Design

The obtained results when converting thefemale speech to male one has shown that the pitch frequency will decrease and the signal must be expanded,while when converting the male signal to female one thepitch frequency will increase and so the signal must becompressed.Pitch shifting by time stretching and resampling involves simply performing a time stretch as described earlier with the PSOLA and then resampling in order toretum sound length to its original value. Expanding thesound by time stretch then resampling creates a higherpitch, while compressing and resampling creates a deeperpitch.As pitch increases, pitch period shrinks. As pitch decreases, pitch period expands. And alsoIncrease number of PSOLA iterations (overlaps) to increase duration. Decrease number of PSOLA iterations (overlaps) to decrease duration.Time stretching using the PSOLA algorithm and resampling in order to change the pitch were applied to voice conversion for speech synthesis.

8. Conclusion

The speaker independent processing of spontaneously spokenhuman-to-human dialogues is a special challenge to presentautomatic speech recognition systems. In this paper, voice conversion method based on TD-PSOLA for Myanmar speech synthesis is demonstrated. Voice conversion for this model is only used for adult human with normal speech rate. This paper discusses Voice Conversion technique based on Pitch Synchronization Over-Lap Add (PSOLA) algorithm has been developed in MATLAB. VAD determines which parts of a voice signal are actual data and which are unvoice based on short term energy and zero crossing rate. A TD-PSOLA procedure, automaticvoice conversion, is used for speech normalization of male-female acoustic differences.This system needs to understand which parts of signal will be modified and how to modify them to normalizefor speaker independent. In the future, there is still needs some improvement of the simulation model in order to provide a more resemble compared to the real world and another algorithm for acoustic – phonetic recognition. This paper reports the progress in an on-going research towards achieving large vocabulary, speaker independent, speech to text recognition, speech to speech recognition system.

Acknowledgement

I am very gradeful to Dr. Khin Thidar Lynn and Dr Mie Mie Thaw for fruitful discussion during the preparation of this paper and also special thank to Rector, Professors and colleges from University of Computer Studies, Mandalay. I would like to thank all my teachers, fiends and my husband who taught and help me during the period of studies and thesis in Computer University, Mandalay.

References

[1] T.Jayasankar, DrR.Thangarajan and Dr.J.Arputha Vijaya, *“Automatic Continuous Speech Segmentation to Improve Tamil Text-to-Speech Synthesis”,* International Journal of Computer Applications (0975 – 8887) ,Volume 25– No.1, July 2011.

[2] Bachu R.G., Kopparthi S., Adapa B and Barkana B.D, “*Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal*”, Electrical Engineering Department, University of Bridgeport

[3] AshwiniSongar and Mrs B. Harita,”*MATLAB based Voice Conversion Model using PSOLA Algorithm”*, International Journal of Digital Application and Contemporary research, Volume 1, Issue 8, March 2013

[4] Mohammad Abushariah, , Raja Ainon, Roziati Zainuddin, Moustafa Elshafei and Othman Khalifa, “Arabic Speaker Independent Continuous Automatic Speech Recognition Based on a Phonetically Rich and Balanced Speech Corpus”, The International Arab Journal of Information Technology, Vol. 9, No. 1, January 2012

[5] EiPhyuPhyuSoe, Aye Thida, “*Diphone-Concatenation Speech Synthesis for Myanmar Language”*,International Journal of Science, Engineering and Technology Research (IJSETR), Volume 2, Issue 5, May 2013

[6] Moe Pwint and FarookSattar, “*Speech/Nonspeech Detection Using Minimal Walsh Basis Functions*” ,EURASIP Journal on Audio, Speech, and Music Processing, Mark Clements, Volume 2007

[7] Kl´araVicsi and Gy¨orgy Szasz´ak,*“Automatic Segmentation of Continuous Speech on Word Level based on Supra-segmental features”,Laboratory of Speech Acoustics. Budapest, Hungary*