Proposing Virtual Phone Conversation System for Dumb and Hearing Impair Person

Pyae Phyo Thu, Htwe Nu Win
University of Computer Studies, Mandalay
pyaephyothu149@gmail.com, nanghtwenuwin@gmail.com

Abstract

The term virtual phone conversation system refers to the bidirectional Myanmar to Myanmar Speech-to-Text and Text-to-Speech translation system that runs in real-time on the mobile phones. With the used of Real-time Speech Recognizer and Real-time Speech Synthesizer, this bidirectional phone conversation translation capability helps the dumb and the hearing impaired person to speak on the mobile phone like the normal person. In this paper, we will propose the framework of the system, its work flow, its applicable methodologies and its challenges.

Keywords: *Virtual Phone Conversation System, Speech-to-Text Recognition, Speech Recognizer, Text-to-Speech Synthesizer, Speech Synthesizer*

1. Introduction

Mobile phone becomes the most widely used communication devices for the sake of anywhere and anytime usage. In recent years, the amount of mobile phone users in Myanmar becomes about 12 million among 56 million populations. This can be expected as it will increase in day by day. Now, almost everyone in a Myanmar household own one or more mobile devices and touch the rewards of ubiquitous terminology.

However, for a dumb and hearing impair person, they can't touch the advantages of mobile communication because of they can't hear and speak. Only counting the number of dumb-deaf person, there are over 20,000 dumb-deaf person in Yangon division. If they can speak on phone with some technical helps, the rate of mobile users will be arisen and this will reduce the boundary between the normal people and the dumb-deaf people. With the use of this proposed system, not only the dumb-deaf person but also the deaf-only will grasp what the opponent is talking on the phone conversation via the text message and they can talk back with text-to-speech translation.

In this paper, we will present in the following structure: the related work of the proposed system is summarized in Section 2. Later, Nature of Myanmar Language is described in Section 3. The description of the proposed system and its applicable methodologies are presented in Section 4. Challenges faces in the system are described in Section 5. In Section 6, we will propose the evaluation methods that will used in the system and expected results of the system and then conclude the paper in Section 7.

2. Related Work

There are many systems that work on the combination of DSP and NLP processing. But, this can only be one-sided: speech-to-text, text-to-speech and speech-to-speech issues. But, this proposed system work in bidirectional: speech-to-text; text-to-speech conversion. In this section, the research work on one-sided speech-to-text; text-to-speech and speech-to-speech conversion are presented.

In 2012, Amarendar Reddty Gundla, Karunakar Reddy S, N.L.Pratap, work on "Communication of Dumb & Blind People with T-T-S" [5]. It is discussed about the historical and the theoretical bases of system and their design explanation such as what text normalization and word pronunciation is and how they work. But detail explanation of methodologies used is omitted. In 2013, Myanmar Text-To-Speech Synthesis System that used Diphone-Concatenation Method: TD-PSOLA algorithm is presented by Ei Phyu Phyu Soe and Aye Thida [6]. It is based on the signal into overlapping synchronized frames of the pitch period. This paper illustrates TD-PSOLA algorithm is effectively improve the performance of TTS system by comparing the performance such as Naturalness, Intelligibility and the Speed of the system that didn't use the algorithm.

The work in [7] presents automatic speech recognition for continuous speech in Myanmar Language using DTW and HMM. DTW is used for the feature clustering and HMM is used for recognition. With the use of multiple speech feature extraction algorithms in the system, system addresses the issue of automatic word/sentence boundary detection in both quiet and noisy environments. Since DTW fixes the generalization nature of HMM, the combination of two methodologies improves the accuracy of recognition phase. In [14], it proposed the system called Verbmobil: a speaker-independent and bidirectional speech-to-speech translation system for spontaneous dialogs in mobile situations. It recognizes spoken input, analyses and translates it, and finally utters the translation. Verbmobil is a multilingual system that incorporates three speech recognizers and three speech synthesizers for German, English and Japanese. The distinguishing feature of Verbmobil is its multi-engine parsing architecture that uses of five concurrent translation engines: statistical translation, case-based translation, substring-based translation, dialog-act based translation, and semantic transfer. It met more than 80% of approximately correct translations and a 90% success rate for dialog tasks.

3. Nature of Myanmar Language

Burmese belongs to the Lolo-Burmese sub-branch of the Tibeto-Burmese branch of the Sino-Tibetan language family. It is the official language of Myanmar, spoken by about 32 million as a first language and as a second language by over 10 million. It is used as national language and can be found widely in the media, government administration, and all levels of education around the country. Burmese language comprises a string of characters: consonants, medial, and vowels, written in sequence from left to right without the white spaces between words or between syllables [1]. The following sub-sections describe the Burmese Language Registers differ in spoken and written and its pronunciation.

3.1. Registers

The term registers means how the language represents in reading, writing and speaking. In Burmese language, the writing styles and the speaking styles are differed and can be categorized into a formal and a colloquial registers [1] described as follows:

- **Formal register (Literary High (H) form):** used for writing especially in official publications, literary works, and formal speech.
- **Colloquial register (Spoken Low (L) form):** is used for speaking especially in daily communications.

Table 1. Differences between Literary and Spoken Burmese Language

Literary (HIGH)	Spoken (LOW)
ကျေးဇူးတင်ပါတယ်	ကျေးဇူးပဲ
ဈေးမှ မုန့်အချို့ဝယ်ခဲ့ပါ	ဈေးက မုန့်ဝယ်ခဲ့ဦး
မေမေသည်ကျွန်တော်ကိုအလွန် ချစ်သည်	မေမေကကျွန်တော့်ကို အရမ်း ချစ်တယ်

Table 1 describes how the sentences vary in literary and spoken Burmese language. This proposed system is especially work on speech. The colloquial register (L form) will be used.

Table 2. Myanmar Character Sets and its Phonemes

<table>
<tr><th colspan="11">Basic Consonants</th></tr>
<tr><th>Group Name</th><th colspan="2">Unaspirated (with ‘h’ sound’)</th><th colspan="2">Aspirated (without ‘h’ sound’)</th><th colspan="4">Voiced (စကားသံနှင့်ပြောဆိုရသော ဒအသံ)</th><th colspan="2">Nasal (နှာသံ)</th></tr>
<tr><td>Gutturals (အာခေါင်သံ)</td><td>က</td><td>/k/</td><td>ခ</td><td>/k^h/</td><td>ဂ</td><td>/g/</td><td>ဃ</td><td>/g/</td><td>င</td><td>/ŋ/</td></tr>
<tr><td>Palastals (အာခေါင်မာအခုံး)</td><td>စ</td><td>/s/</td><td>ဆ</td><td>/s^h/</td><td>ဇ</td><td>/z/</td><td>ဈ</td><td>/z/</td><td>ည</td><td>/□/</td></tr>
<tr><td>Alveolars (သွားရင်းဖျားသည်း)</td><td>ဋ</td><td>/t/</td><td>ဌ</td><td>/t^h/</td><td>ဍ</td><td>/d/</td><td>ဎ</td><td>/d/</td><td>ဏ</td><td>/n/</td></tr>
<tr><td>Dentals (သွား)</td><td>တ</td><td>/t/</td><td>ထ</td><td>/t^h/</td><td>ဒ</td><td>/d/</td><td>ဓ</td><td>/d/</td><td>န</td><td>/n/</td></tr>
<tr><td>Labials (နှုတ်ခမ်းဖျားသံ)</td><td>ပ</td><td>/p/</td><td>ဖ</td><td>/p^h/</td><td>ဗ</td><td>/b/</td><td>ဘ</td><td>/b/</td><td>မ</td><td>/m/</td></tr>
<tr><td rowspan="2">Without Group</td><td>ယ</td><td>/j/</td><td>ရ</td><td>/j/</td><td>လ</td><td>/l/</td><td>ဝ</td><td>/w/</td><td>သ</td><td>/θ/</td></tr>
<tr><td colspan="2"></td><td>ဟ</td><td>/h/</td><td>ဠ</td><td>/l/</td><td>အ</td><td>/□/</td><td colspan="2"></td></tr>
<tr><th colspan="11">Vowels</th></tr>
<tr><td colspan="11">
<table><tr><td>-ိ</td><td>/i/</td><td>ေ-</td><td>/e/</td><td>--ယ်</td><td>/□/</td><td>--ာ</td><td>/à/</td><td>ေ--ာ်</td><td>/□/</td><td>ို</td><td>/o/</td><td>-ူ</td><td>/u/</td></tr></table>
</td></tr>
<tr><th colspan="11">Medial</th></tr>
<tr><td colspan="11">
<table><tr><td>-ျ</td><td>/j/</td><td>ြ</td><td>/j/</td><td>--ွ</td><td>/w/</td><td>--ှ</td><td>/-̥-/</td></tr></table>
</td></tr>
</table>

3.2. Phonology

The sequence of sounds builds the language’s phonology. For example, in English: the phonetics “koust” pronounce the word “coast”, the “'□d□ail” pronounce “agile” and in Burmese: ခလုတ် = ခ + လ + --ုတ် (Phonetics: k^həlo□□ = /k^hə/+ /l/ + /o□□/) [7]. Burmese phonology is generally constructed by means of combining the Consonants phoneme, Vowels phoneme and the Tones. Table 2 describes the Myanmar Character Sets and its Phonemes defined by the International Phonetic Association (IPA) [2, 6, 7, and 8].

3.3. Tones

Burmese is a tonal language, consisting of four tones (low, high, creaky, and checked) [1, 6]. Burmese Phonology involves not only this basis of the tone but also involves pitch, phonation, intensity (loudness), duration, and vowel quality [1]. Table 3 describes how the tones occur by using letter က [k] as a basis [2].

Table 3. Four Contrastive Tones in Burmese

Tone	Description	Example	
Low **နိမ့်သံ**	Low pitch	ကာ	/kà/
High **တက်သံ**	Slightly breathy, high pitch	ကား	/ká/
Creaky **သက်သံ**	Tense or Creaky, high pitch	ကန့်	/ka~/
Checked **တိုင်သံ**	Final glottal stop, high pitch	ကက်	/k□ □/

4. Proposed System

The proposed system in this paper incorporates with two subsystems: Speech-to-Text Recognizer and Text-to-Speech Synthesizer which are implement on the Dumb-Deaf persons' mobile phones. As shown in Figure 1, when someone dials to the phone that the system built, after the *"user"* hang on, the *"opponent"* will speak. At that time, Speech-to-Text Recognizer will work on to transform the incoming speech into recognized words which will display later on the user's phone display. Because of the user of the system can't talk back, he/she will text it back. At that time, the input of the system become text and that text will be translated into the corresponding speech by means of the Text-to-Speech Synthesizer of the system. So, although the user of the system can't hear and talk back on the phone conversation, the proposed system will listen to and speak on the phone on behalf of the user.

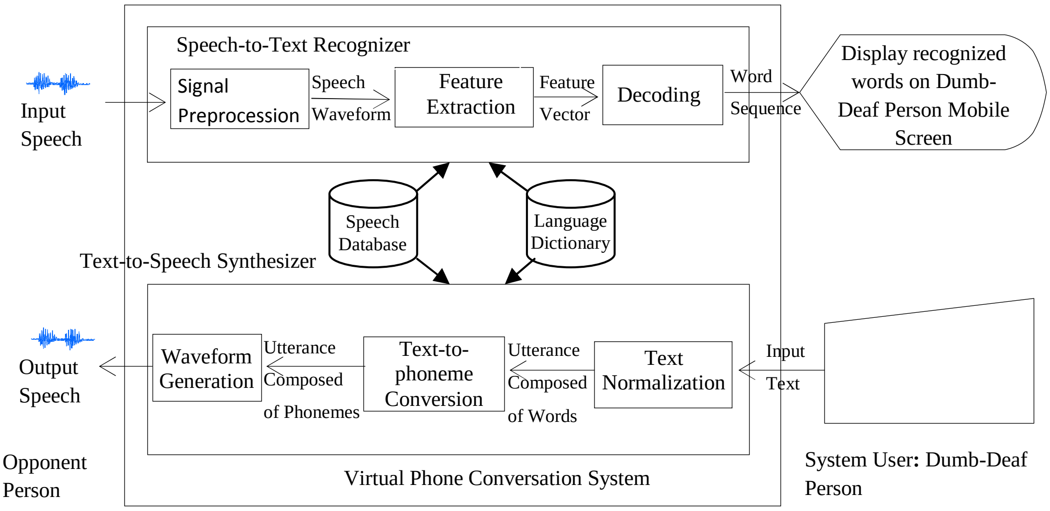
Speech-to-Text Recognizer
Input Speech
Signal Preprocession
Speech Waveform
Feature Extraction
Feature Vector
Decoding
Word Sequence
Display recognized words on Dumb-Deaf Person Mobile Screen
Speech Database
Language Dictionary
Text-to-Speech Synthesizer
Output Speech
Waveform Generation
Utterance Composed of Phonemes
Text-to-phoneme Conversion
Utterance Composed of Words
Text Normalization
Input Text
Opponent Person
Virtual Phone Conversation System
System User: Dumb-Deaf Person

Figure 1. Framework of the Virtual Phone Conversation System

4.1. Speech-to-Text Recognizer and its Applicable Methodologies

Speech-to-Text recognition system that will be used in this proposed system is similar to the standard speech recognition system. Like the standard speech recognition system, this speech-to-text recognizer work as follow:

1. **Signal Preprocessing**: the input Myanmar speech signal is discretized with a sampling frequency of e.g. 16 kHz. And then, signals that have frequencies below 100 Hz which tend to contain noise are filtered with a high-pass filter. And then speech is analyzed using the frame size and shift in the range of 10-30 ms to extract speaker information.
2. **Feature Extraction**: In this module, acoustic observations are performed to extract over time frames of uniform length. Within these frames, the speech signal is assumed to be stationary. For the acoustic samples in the window, multi-dimensional feature vector is calculated. On the time window, a fast Fourier transformation is performed, moving into the spectral domain.
3. **Decoding**: In this module, decoding calculates which sequence of Myanmar words is most likely to match to the acoustic signal represented by the feature vectors. At that time, available information sources are: speech database which contains the statistical representations of each of the distinct sounds that make up a Myanmar word, a dictionary,

typically a list of words and the phoneme sequences they consist of, and a language with represent the word or word sequence likelihoods. Then, it performs a search to find the most possible sequence of words and result the probable Myanmar word sequences.

Emerging methodologies that can be used in this speech-to-text recognizer of the system are: Hidden Markov Model (HMM), Dynamic Time Warping (DTW), and Artificial Neural Networks (ANN) [3, 10]. Among them, HMM is mostly used and its give more accurate results [9, 11, 12, 13]. The summarizations of these methodologies describe in Table 4 [3].

Table 4. Speech Recognition Methodologies

Techniques	Description
HMM	-Statistical Markov model, -Present in simplest dynamic Bayesian network, -Useful for isolated word recognition, continuous speech recognition and training HMM
DTW	-Algorithm for measuring similarity between two temporal sequences, -Well-known technique to find an optimal alignment between two given (time-dependent) sequences
ANN	-Defined as a model of reasoning based on the human brain, -Highly non-linear modeling, -Widely used in solving various classifications and forecasting problems, -Require less formal statistical training, -Have ability to detect all possible interactions between predictor variables, and the availability of multiple training algorithms

4.2. Text-to-Speech Synthesizer and its Applicable Methodologies

Text-to-Speech Synthesizer converts Myanmar speech into the corresponding text. As shown in the Figure 1, text-to-speech synthesizer of the proposed system contains three modules:

1. **Text Normalization**: incoming raw Myanmar texts type by the dumb-deaf user is isolate into the equivalent of written-out words. Then, searches for numbers, times, dates, and other symbolic representations. These are analyzed and converted to Myanmar words. For example: ဖုန်းဆက်နော် - ဖုန်း + ဆက် + နော်
2. **Text-to-Phoneme Conversion**: this module recovers the syntactic constituency and semantic features of Myanmar words, phrases, clauses and sentences, which are important for both pronunciation and prosodic choices in the successive processes. Then, the phonetic sequence of Myanmar words such as ဖုန်း /p^{h}ó□□/, ဆက် /s^{h}a□/, နော် /n□̀/ [2] are produced.
3. **Waveform Generation**: In this module, the phonetic sequences are converted into the probable Myanmar speech which will transmit to the opponent’s devices.

The two primary technologies generating synthetic speech waveforms are concatenative synthesis and formant synthesis [6] and there also Articulatory synthesis and HMM-based synthesis. The summarizations of these technologies are describes in Table 5 [4].

Table 5. Speech Synthesizer Methodologies

Techniques	Description
Concatenative synthesis	-Based on the concatenation of segments of recorded speech, -Produces the most natural-sounding synthesized speech
Formant synthesis	-Speech output is created using additive synthesis and an acoustic model, -Formant-synthesized speech can be reliably intelligible, even at very high speeds, -Smaller programs than concatenative systems
Articulatory synthesis	-Computational techniques for synthesizing speech based on models of the human vocal tract and the articulation processes occurring there, -Have not been incorporated into commercial speech synthesis systems
HMM-based synthesis	-Synthesis method based on hidden Markov models, -Frequency spectrum, fundamental frequency and duration (prosody) of speech are modeled simultaneously by HMMs,

	-Speech waveforms are generated from HMMs themselves based on the maximum likelihood criterion

5. System Challenges

Although this system will help the people with a hearing impairment to be able to talk on mobile phones, there are many challenges arise for the researcher in the case of real-time processing. These are:

1. **Time**: to be fast enough in converting written language into speech and spoken conversation into text
2. **Accuracy**: to be able to give the accurate speech-to-text or text-to-speech results
3. **Resources Capabilities**: due to the capabilities of the mobile phones such as CPU speed, Memory and Power Resource, resulting performance will be limited.

6. Evaluation Methods and Expected Results

The performance of the proposed system is evaluated in terms of accuracy, speed, naturalness and intelligibility. Accuracy may be measured in terms of performance accuracy which is usually rated with word error rate (WER), whereas speed is measured with the real time factor. Naturalness describes how closely the output sounds like human speech, while intelligibility is the ease with which the output is understood. Accuracy and speed are important qualities of speech-to-text recognition of the proposed system and naturalness and intelligibility are the qualities of text-to-speech synthesizer of the proposed system.

The system is expected to operate on real time processing: high speed, high accuracy and high performance. According to the [9, 11, 12, 13], HMM is the current widely used methodologies in both speech-to-text and text-to-speech conversion and it give the high performance than any other methodologies. Neural network is also widely used methodologies that give high performance. So, it can be assumed that if these two methodologies are combined and applied some new contribution; and then if it use in this system, it will reach to the system's expected result of nearly real-time processing.

7. Conclusion

Major contribution of this proposed system is to operate on nearly real-time. This system not only fulfills the needs of hearing impair person to speak efficiently on the mobile devices especially in the emergency cases to save their own lives but also arise the mobile users' rate of the nation that will especially affect the mobile phone vendors and the nation budget. This paper will lead to the trends towards the speech-to-speech translation on mobile environment.

References

[1] http://en.wikipedia.org/wiki/Burmese_language

[2] http://en.wikipedia.org/wiki/Burmese_alphabet

[3] http://en.wikipedia.org/wiki/Speech_recognition

[4] http://en.wikipedia.org/wiki/Speech_synthesis

[5] Amarendar Reddty Gundla, Karunakar Reddy S, N.L.Pratap, "*Communication of Dumb & Blind People with T-T-S*", IJRCCT, ISSN 2278-5841, Vol 1, Issue 6, November 2012

[6] Ei Phyu Phyu Soe, Aye Thida, "*Myanmar Text-To-Speech Synthesis System Using Diphone-Concatenation Method*", International Journal of Computational Linguistics and Natural Language Processing, Vol 2 Issue 5 May 2013, ISSN 2279 – 0756

[7] Ingyin Khaing, "*Myanmar Continuous Speech Recognition System Based on DTW and HMM*", IJIET, Vol. 2 Issue 1 February 2013, ISSN: 2319 – 1058

[8] International Phonetic Association, "*Phonetic description and the IPA chart*", Handbook of the International Phonetic Association: a guide to the use of the international phonetic alphabet, Cambridge University Press, 1999

[9] Nobuaki Minematsu, Yukiko Fujisawa, Seiichi Nakagawa, "*Performance Comparison among HMM, DTW, and Human Abilities in Terms of Identifying Stress Patterns of Word Utterances*", Graduate School of Engineering, University of Tokyo Department of Information and Computer Sciences, Toyohashi University of Technology

[10] Pahini A. Trivedi, "*Introduction to Various Algorithms of Speech Recognition: Hidden Markov Model, Dynamic Time Warping and Artificial Neural Networks*", 2014 IJEDR | Volume 2, Issue 4 | ISSN: 2321-9939

[11] Pieter Vermeulaen, Etienne Barnard, Mark Fanty, Younghon Yan, Ronald Cole, "*A comparison of HMM and Neural Network Approaches to Real World Telephone Speech Applications*", Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, USA

[12] Vimala.C, Dr.V.Radha, "*A Review on Speech Recognition Challenges and Approaches*", World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741, Vol. 2, No. 1, 1-7, 2012

[13] Vincent Vanhoucke, Geoffrey Hinton, "*Deep Neural Networks for Acoustic Modeling in Speech Recognition*", IEEE Signal Processing Magazine, November 2012

[14] Wolfgang Wahlster, DFKI GmbH, Saarbrücken, Germany, “*Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of Final Verbmobil System*”, Documentation of Verbmobil project (1993-2000) in Germany