Bipika Amatya

Emotional Speech from Machine

Master's Thesis of Mathematical Information Technology

22nd May, 2020

University of Jyvaskyla

Faculty of Information Technologies

Author: Bipika Amatya

Contact information: krispyqueen24@gmail.com

Supervisors: Vagan Terziyan, Jukka Saari

Title: Emotional Speech from Machine

Project: Master's thesis

Study line: Cognitive Computing and Collective Intelligence

Page count: 50

Abstract:

Emotional speech is the expressiveness in speech that is transmitted through changes in pitch, loudness, timbre, speech rate and pauses that convey emotion. Although the current TTS technology is capable of converting a given text into speech, they sound monotonous and lack emotion and naturalness. In order to improve artificial voices, application of emotion is highly evaluated. In this thesis, we will be creating a system that makes use of speech mark-up language to produce emotion in speech by analysing the tone of given text. For this purpose, we combine IBM tone analyser with TTS that accepts the speech mark-up language. In this research, we perform empirical study on two experimental implementation using two TTS and two speech mark-up language. The first combination involves IBM TTS and SSML and the second combination includes MARY TTS and EmotionML. The mark-ups are predefined in EmotionML for four major emotions namely anger, fear, joy and sadness and for SSML prosody value from previous study is used. Therefore, this study describes the two implementations and evaluate their output emotional speech synthesis which is then compares with human voice to define its perfection.

Keywords:

Emotional Text to speech, IBM Watson Tone Analyser, emotional speech, EmotionML, SSML, MARY, IBM Watson Text to Speech

GLOSSARY

SSML	Speech Synthesis Markup Language
TTS	Text to Speech
IBM	International Business Machines
API	Application programming interface
AI	Artificial Intelligence
HTML	Hypertext Markup Language
JSON	JavaScript Object Notation
SDK	Software development kit
TESS	Toronto emotional speech set
MARY	Modular Architecture for Research on speech sYnthesis
EmotionML	Emotion Markup Language

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Speech is a major way of verbal communication between humans which has also been popularly used in communication with computers and machines these days using speech synthesis (Lemmetty, 1999). Speech synthesis is a system that allows us to convert electronic texts into a synthetic speech (Aida–Zade et al., 2010). The speech synthesis consists of two main stages in its process namely text processing and generation of speech waveform. In the first stage of text processing, the input texts are translated into a phonetic and verbal representation, then in the second stage, these phonetic and prosodic information produces the acoustic waveforms. The source of input texts, for example, may be data from a word processor, paperback, email, web pages, a smartphone text message, or a newspaper scanned text. These strings of characters from source are then pre-processed and translated into phonetic representation, which is typically a string of phonemes, and then speech sound is finally produced (Chen & Jokinen, 2010). Figure 1 shows a simple TTS diagram. Since there is a growing usage of speech synthesizers in recent years, it has become more desirable for all to obtain its increased naturalness. The naturalness in speech is achieved by the use of expression and emotion in it (Schröder, 2001). Therefore, this study is intended to create emotion in speech.

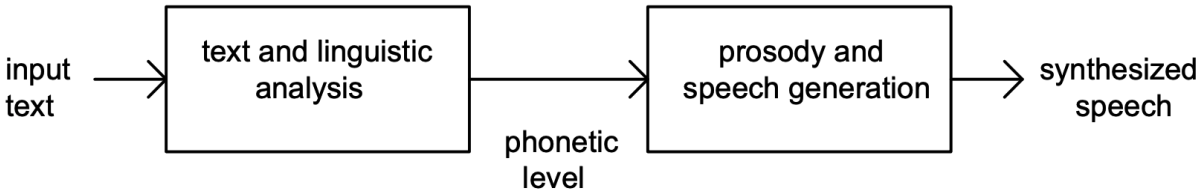


Figure 1. Simple text-to-speech synthesis procedure(Chen & Jokinen, 2010)

Expressivity in speech is obtained with the use of vocal prosody. Vocal prosody is a vital component in speech communication that makes use of speech features such as pitch, rate and loudness in order to coordinate with the expression of emotion. These three aspects of speech are also referred to as The Big Three of vocal prosody (Crumpton, 2015). We have used SSML to incorporate speech features to our text from IBM tone analyser. SSML element <prosody> is used for this purpose which is explained in chapter two. In this experimentation we utilize the rate, volume and pitch attribute of the prosody element of SSML to create respective emotions. The below Table 2. provides the characteristics of vocal prosody used for our emotions.

Emotion	**Pitch**	**Rate**	**Volume**
Anger	High	Fast	High
Fear	High	Fast	High
Joy	High	Moderate	High
Sadness	Low	Slow	Low

Table 1. Emotions and Associated Vocal Prosody Characteristics (Crumpton, 2015)

Emotion refers to a varying strength, tone, or pitch of voice that listeners use to draw conclusions regarding the speaker's emotional situation (Laukka, 2020). The human tends to convey such emotional behaviour in their voice. Despite the ability of TTS to accurately convert text into voice, these systems lack emotion and naturalness of human voice (Reddy et al., 2015). Literature study about emotional speech reveals that there have been various studies made for

creating emotional TTS. (Reddy et al., 2015) mentions in his article about such works done from 1989 to 2005 during his research work on creating emotional speech for Telugu language. The emotional speech system study for Spanish language was performed in paper (Aida–Zade et al., 2010). (Lee et al., 2017) made use of Tacotron (Wang et al., 2017) a TTS system and created an emotional speech synthesizer that was experimented on Korean dataset. (Asakura et al., 2019) proposed and demonstrated an emotional voice conversion method which converted neutral speech into emotional speech. But there are few empirical researches done on use of mark-up language in order to create emotional speech in terms of text tone. Constructing and exploring the usefulness of such methods may provide valuable information.

The purpose of this study is to create a natural voice from the speech synthesizer and compare whether or not the artificial voice produced sounds like the human voice with auto tagging of speech mark-ups. We will be examining if the artificial speech can carry accurate emotions of sadness, happiness, anger and fear and explore the quality, accuracy and cost of generated emotional speech from the constructed system.

This study contributes valuable theory of the successful implementation of technology that is required by artificial intelligence applications, such as machine support argents and human-type robots that use speech synthesis, in the human computer interface (HCI) framework (Hartmann et al., 2013). This program is useful in providing customer support, public announcements and human handicap assistance for mute people.

1.1 RESEARCH QUESTION AND OBJECTIVES

In this research topic, I will be investigating following research questions:

1. *How close the speech generated by a machine is to the human voice?*
2. *How accurate will the service be?*
3. *Will it be able to annotate pause, speed and pitch in the text where necessary?*
4. *What will be the cost of the service?*

The objective of the research is to analyse the possibilities to create a service that takes text as its input that is analysed for its tone and produces the output speech with that resulting tone emotion.

1.2 RESEARCH METHOD

For this research, I choose the design science research method. Design science method assists in creating a technology-based solution to real world problems. It helps create new knowledge or improve existing one by reviewing previous work and methods (Hevner et al., 2004). This thesis originated from real-life problems of a company Teleste Oyj. The company was seeking for an expressive human-like artificial voice for their product audio announcement system for rails.

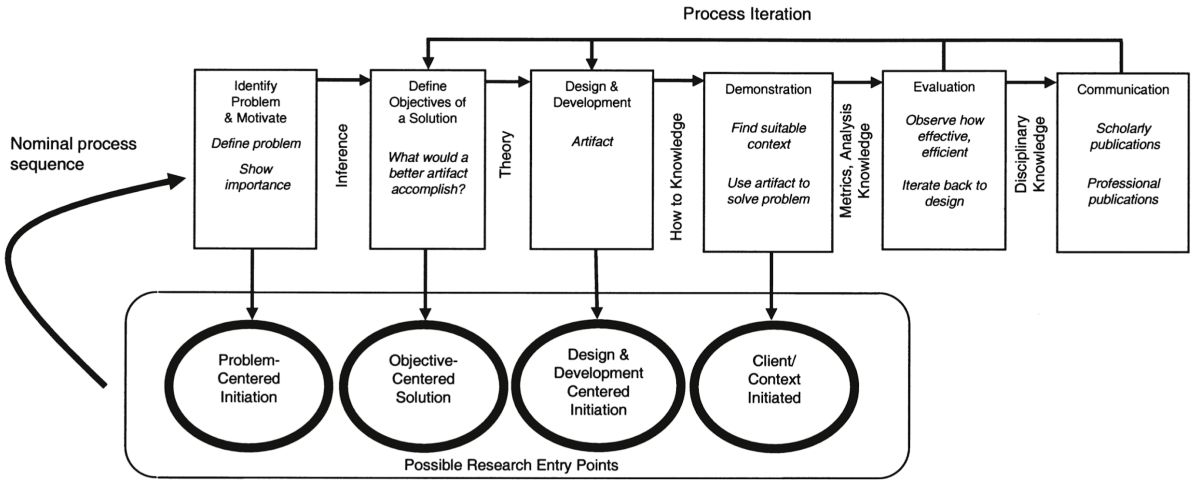


Figure 2. DSRM Process Model (Peffers et al., 2007)

Although the current text to speech is able to convert the text into speech, the quality of the output audio is not as pleasing as human voice. There is a need for increased naturalness in speech synthesis. The most obvious aspect that provides naturalness is the appropriate emotional expression while speaking. Human voice incorporates such expressions of joy, compassion, etc. in their speech. Therefore, my aim with my research work is to create a solution to this problem and develop a service that is capable of producing such emotional voices from the machine.

The solution development incorporates the use of IBM Watson tone analyser that analyses the emotion in the text. Specific tagging will be defined for each emotion then speech with emotional expressions will be generated. For the quality evaluation of the generated speech, recorded speech from humans for different emotions (Pichora-Fuller & Dupuis, 2020) will be compared. The similarities and differences between the audio from humans and TTS will be reported. The audio data will be analysed using a Praat software package. The result may have minor statistical variance as the different speakers have their own different pitch and frequency which will be ignored and only the major emotional flow will be considered while comparing.

The advantages of design science research methods are:

1. It develops and evaluates advised solutions for recognized problems.
2. It provides a nominal process for conducting DS research.
3. It suggests a model for the presentation of research outcomes.
4. It creates designing principles
5. Use of context specific methods or techniques.

Despite of being the most needed method for development, design science research method possesses few limitations which are enlisted below:

1. Research worth is dedicated to a certain demand of a group only which might not fit for all.
2. For example, if the research has been made for a company then the problem and scenario of that particular company is solved but the different company can have different needs. The research demand may be from different perspectives like manager, developer, consumers, stakeholders, etc.

1.3 THESIS STRUCTURE

The structure of this thesis is arranged in a way that this chapter gives a short introduction of the topic, explains the scope of the thesis, presents the research questions and also explains the structure of the thesis. Chapter2 presents the basic theory and concepts required to understand the thesis work. This chapter also gives an explanation of the choice of methods used for the experimental work. Chapter 3 is dedicated to the prior work done regarding the development of emotional speech generation. It explains the methods and experimental environment used for building the service. Chapter 4 presents the results obtained from the experiments. Chapter 5 consists of discussion on the results obtained and possibility for future enhancement. Finally, Chapter 6 present the conclusions drawn from the thesis work.

2 THEORY

The aim of this thesis is to design a system which combines the available modern tools. This chapter therefore explores the tool for providing aid in designing. To acquire knowledge on what features better fits we study and evaluate current devices. We identify that all of them provide some sorts of automated service that improves the development of system. We explain in more detail a few tools that we consider to be the most relevant. They are: IBM Watson TTS, Tone Analyzer, MARY TTS, SSML, EmotionML. These tools portray a significant importance in our experimentation as they fulfil some of the objectives that our program has to satisfy.

2.1 IBM WATSON TONE ANALYSER

The IBM Watson Tone Analyzer is a service API by IBM corporation that identifies emotional and language tones in written text (*IBM Watson Tone Analyzer*, 2020). This service is an artificial intelligence (AI) platform for business, professionals and academic study and implementation. It is one of the IBM Watson Applications that provides SDK's and cloud-based services (IBM Bluemix) for the developers and researchers. The service provides a limited free usage of 2,500 API calls per month and after that payment had to be made per API call as per their price listing. It accepts plain text, JSON, or HTML input up to 128 KB of text. IBM provides documentation to help researchers use the services. It provides business solutions such as customers' feedback review, improving customer interactions.

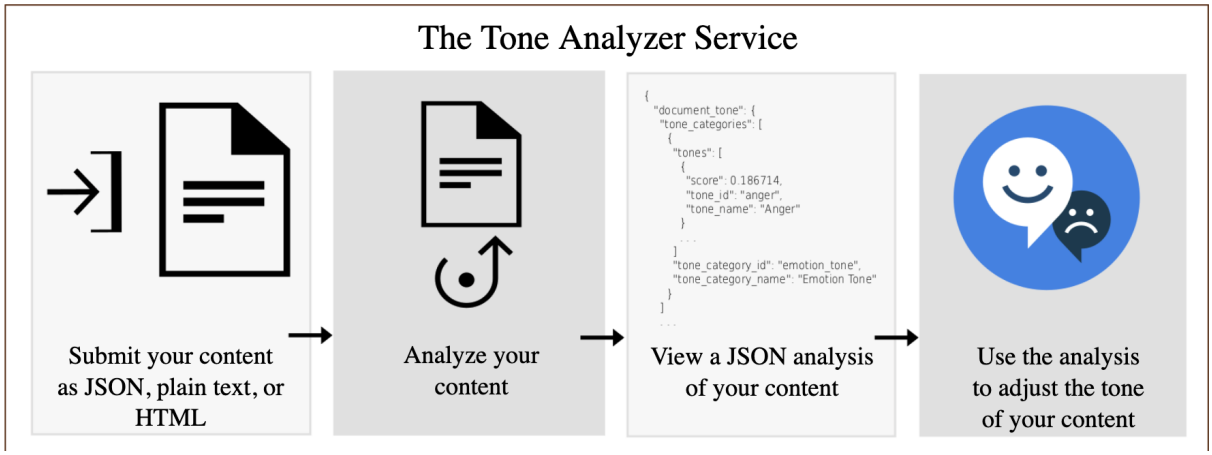


Figure 3. Basic flow of tone analyser

IBM Watson Tone Analyzer is used in this research to find the emotional tones. It offers seven general purpose tones namely anger, fear, joy, sadness, analytical, confident and tentative (*IBM General Purpose Tones*, 2020). The service is able to detect tones in both document level and sentence level. It can analyse various input texts such as email, messages or online reviews. The IBM WTA uses score to determine the acceptance of its return tones. The score range for each tone class lies between 0.5 to 1, whereas a score greater than 0.75 indicates a high possibility that this tone is present in the content. In this paper we only consider four major tones that are anger, fear, joy and sadness for our experimentation and the tone score above 0.6 is accepted.

Below table shows the IBM tones and its description.

Emotional Tone	**Description**
Anger	Anger is evoked due to injustice, conflict, humiliation, negligence, or betrayal. If anger is active, the individual attacks the target, verbally or physically. If anger is passive, the person silently sulks and feels tension and hostility.
Fear	Fear is a response to impending danger. It is a survival mechanism that is triggered as a reaction to some negative stimulus. Fear can be a mild caution or an extreme phobia.
Joy	Joy (or happiness) has shades of enjoyment, satisfaction, and pleasure. Joy brings a sense of well-being, inner peace, love, safety, and contentment.
Sadness	Sadness indicates a feeling of loss and disadvantage. When a person is quiet, less energetic, and withdrawn, it can be inferred that they feel sadness.

*Table 2. Descriptions of emotional tones (*IBM General Purpose Tones*, 2020)*

2.2 IBM WATSON TEXT TO SPEECH

The IBM Watson Text to Speech is another service API by IBM corporation that converts written text into speech. This service platform is also available for business, professionals and academic study and implementation. It is one of the IBM Watson Applications that provides SDK's and cloud-based services (IBM Bluemix) for the developers and researchers. The service provides a limited free usage of 10,000 Characters per Month and after that users are charged per thousand characters. It supports the XML-based Speech Synthesis Mark-up Language (SSML), allow to customize phonetics, can produces audio in Ogg or WebM with the Opus or Vorbis codec, WAV, FLAC, MP3 (MPEG), l16 (PCM), mulaw, or basic format. The service supports the languages of Arabic, Brazilian Portuguese, Mandarin Chinese, Dutch, English (dialects of the United Kingdom and the United States), French, German, Italian, Japanese, Korean, Spanish (dialects of Castile, Latin American, and North American) (*IBM Watson Text to Speech*, 2020). The flow of service is shown below Figure 4.

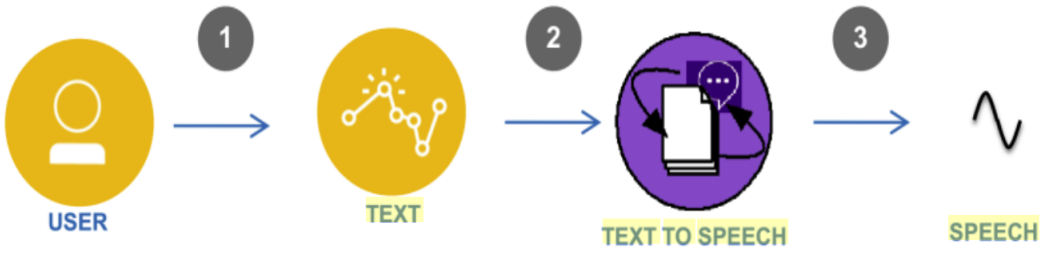


Figure 4. IBM Watson Text to speech (Santiago et al., 2017)

The API is capable of converting written text in SSML format into audible speech generating outputs in various speaking styles, pronunciation, pitch and speech rate. The SSML formatted text as per tone is fed to TTS and it produces the emotional audio file.

2.3 SPEECH SYNTHESIS MARKUP LANGUAGE (SSML)

Speech Synthesis Mark-up Language (SSML) is an XML-based mark-up language which provides a standard for the control of speech aspects such as pronunciation, volume, pitch, rate, etc. Voice Browser working group developed SSML standards which is published in W3C Recommendation 7th September 2010 version 1.1 (Recommendation, 2010). Below illustrated are the SSML elements that we use in this research:

2.3.1 "speak" Root Element

speak is a root element. The speak element can only contain text to be rendered and the following elements: audio, break, emphasis, lang, lexicon, lookup, mark, meta, metadata, p, phoneme, prosody, say-as, sub, s, token, voice, w.

Example:

```
<?xml version="1.0"?>
<speak version="1.1"
      xmlns="http://www.w3.org/2001/10/synthesis"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
                  http://www.w3.org/TR/speech-synthesis11/synthesis.xsd"
      xml:lang="en-US">
  ... the body ...
</speak>
```

2.3.2 “p” and “s” Elements

A p element represents a paragraph. An s element represents a sentence.

Example:

```
<?xml version="1.0"?>
<speak version="1.1" xmlns="http://www.w3.org/2001/10/synthesis"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
                  http://www.w3.org/TR/speech-synthesis11/synthesis.xsd"
      xml:lang="en-US">
  <p>
    <s>This is the first sentence of the paragraph.</s>
    <s>Here's another sentence.</s>
  </p>
</speak>
```

2.3.3 “voice” Element

It requests change in speaking voice. The voice feature attributes are:

- **gender**: *optional* attribute indicating the preferred gender of the voice to speak the contained text. Enumerated values are: "**male**", "**female**", "**neutral**", or the empty string "".
- **age**: *optional* attribute indicating the preferred age of speaker voice in years. Acceptable values are non-negative integers or the empty string "".
- **variant**: *optional* attribute indicating a preferred variant of the other voice characteristics. (e.g. the second male child voice). Valid values are positive integers or the empty string "".
- **name**: *optional* attribute indicating a processor-specific voice name. The value *may* be a space-separated list of names ordered from top preference down or the empty string "" excluding white space.

- **languages**: *optional* attribute indicating the list of languages the voice is desired to speak. The value *must* be either the empty string "" or a space-separated list of languages, with *optional* accent indication per language.

Example:

```
<?xml version="1.0"?>
<speak version="1.1" xmlns="http://www.w3.org/2001/10/synthesis"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
                  http://www.w3.org/TR/speech-synthesis11/synthesis.xsd"
      xml:lang="en-US">
  <voice gender="female" languages="en-US" required="languages
gender variant">Mary had a little lamb,</voice>
  <!-- now request a different female child's voice -->
  <voice gender="female" variant="2">
  Its fleece was white as snow.
  </voice>
  <!-- processor-specific voice selection -->
  <voice name="Mike" required="name">I want to be like
Mike.</voice>
</speak>
```

2.3.4 “emphasis” Element

The emphasis element requests that the contained text be spoken with emphasis. The attributes are:

- **level**: the *optional* **level** attribute indicates the strength of emphasis to be applied. Defined values are **"strong"**, **"moderate"**, **"none"** and **"reduced"**. The default **level** is **"moderate"**.

Example:

```
<?xml version="1.0"?>
<speak version="1.1" xmlns="http://www.w3.org/2001/10/synthesis"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
http://www.w3.org/TR/speech-synthesis11/synthesis.xsd"
      xml:lang="en-US">
  That is a <emphasis> big </emphasis> car!
  That is a <emphasis level="strong"> huge </emphasis>
  bank account!
</speak>
```

2.3.5 “break” Element

The break element is an empty element that controls the pausing or other prosodic boundaries between tokens. The attributes on this element are:

- **strength**: this is an *optional* attribute having one of the following values: **"none"**, **"x-weak"**, **"weak"**, **"medium"** (default value), **"strong"**, or **"x-strong"**.
- **time**: this is an *optional* attribute indicating the duration of a pause to be inserted in the output in seconds or milliseconds. e.g. "250ms", "3s".

Example:

```
<?xml version="1.0"?>
<speak version="1.1" xmlns="http://www.w3.org/2001/10/synthesis"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
                  http://www.w3.org/TR/speech-synthesis11/synthesis.xsd"
      xml:lang="en-US">
```

```
 Take a deep breath <break/>
 then continue.
 Press 1 or wait for the tone. <break time="3s"/>
 I didn't hear you! <break strength="weak"/> Please repeat.
</speak>
```

2.3.6 “prosody” Element

The prosody element permits control of the pitch, speaking rate and volume of the speech output. The attributes, all *optional*, are:

- **pitch**: the baseline pitch for the contained text. Legal values are: a number followed by "Hz", a relative change or **"x-low"**, **"low"**, **"medium"**, **"high"**, **"x-high"**, or **"default"**. Labels **"x-low"** through **"x-high"** represent a sequence of monotonically non-decreasing pitch levels.
- **contour**: sets the actual pitch contour for the contained text. The pitch contour is defined as a set of white space-separated targets at specified time positions in the speech output. In each pair of the form (time position, target), the first value is a percentage of the period of the contained text (a number followed by "%") and the second value is the value of the **pitch** attribute (a number followed by "Hz", a relative change, or a label value). Time position values outside 0% to 100% are ignored. If a pitch value is not defined for 0% or 100% then the nearest pitch target is copied. All relative values for the pitch are relative to the pitch value just before the contained text.
- **range**: the pitch range (variability) for the contained text. Legal values are: a number followed by "Hz", a relative change or **"x-low"**, **"low"**, **"medium"**, **"high"**, **"x-high"**, or **"default"**. Labels **"x-low"** through

"x-high" represent a sequence of monotonically non-decreasing pitch ranges.

- **rate**: a change in the speaking rate for the contained text. Legal values are: a non-negative percentage or **"x-slow"**, **"slow"**, **"medium"**, **"fast"**, **"x-fast"**, or **"default"**. Labels **"x-slow"** through **"x-fast"** represent a sequence of monotonically non-decreasing speaking rates. When the value is a non-negative percentage it acts as a multiplier of the default rate. For example, a value of 100% means no change in speaking rate, a value of 200% means a speaking rate twice the default rate, and a value of 50% means a speaking rate of half the default rate.
- **duration**: a value in seconds or milliseconds for the desired time to take to read the contained text. e.g. "250ms", "3s".
- **volume**: the volume for the contained text. Legal values are: a number preceded by "+" or "-" and immediately followed by "dB"; or **"silent"**, **"x-soft"**, **"soft"**, **"medium"**, **"loud"**, **"x-loud"**, or **"default"**. The default is +0.0dB. Specifying a value of **"silent"** amounts to specifying minus infinity decibels (dB).

Example:

```
<speak version="1.1"
      xmlns="http://www.w3.org/2001/10/synthesis"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
                  http://www.w3.org/TR/speech-synthesis11/synthesis.xsd"
      xml:lang="en-US">
  <s>I am speaking this at the default volume for this voice.</s>
  <s><prosody volume="+6dB">
```

```
    I am speaking this at approximately twice the original signal
amplitude.
  </prosody></s>
  <s><prosody volume="-6dB">
    I am speaking this at approximately half the original signal
amplitude.
  </prosody></s>
  The price of XYZ is <prosody rate="90%">$45</prosody>
  <prosody contour="(0%,+20Hz) (10%,+30%) (40%,+10Hz)">
</speak>
```

2.4 EMOTION MARKUP LANGUAGE (EMOTIONML)

EmotionML is a mark-up language to provide support for generating emotional voices. It allows us to enclose emotion to the text. The main elements of Emotion mark-up as of W3C Recommendation 22 May 2014 (Recommendation, 2014a) are:

2.4.1 “emotionml” element

This is a root element. It requires its namespace and version attributes. Other optional attributes are category-set, dimension-set, appraisal-set and action-tendency-set.

Example:

```
<emo:emotionml version="1.0"
xmlns:emo="http://www.w3.org/2009/10/emotionml">
 ...
</emo:emotionml>
```

2.4.2 “emotion” element

It represents a single emotion annotation. It must contain one of these four children <category> or <dimension> or <appraisal> or <action-tendency>.

Example:

2.4.3 “category” element

Emotion is described using this element. It must contain a name attribute in it.

Example:

```
<emotion category-
set="http://www.example.com/custom/category/interpersonal-
stances.xml#voc">
   <category name="distant"/>
</emotion>
```

2.5 MARY TTS

MARY (Modular Architecture for Research on Speech sYnthesis) is an open source German text-to-speech system (Schröder & Trouvain, 2003). It allows processing step-by-step and provides access to partial processing results. It provides a web interface that facilitates users with an interactive environment to explore the impact of a particular piece of information on the performance of a processing phase in question. It is suitable for both technical and non-technical users. It is capable of parsing speech synthesis mark-up such as SABLE (Sproat et al., 1998) and EmotionML.

The architecture of the MARY TtS system is similar to a typical TTS architecture as described by Dutoit (Dutoit, 1997) as cited by (Schröder & Trouvain, 2003).

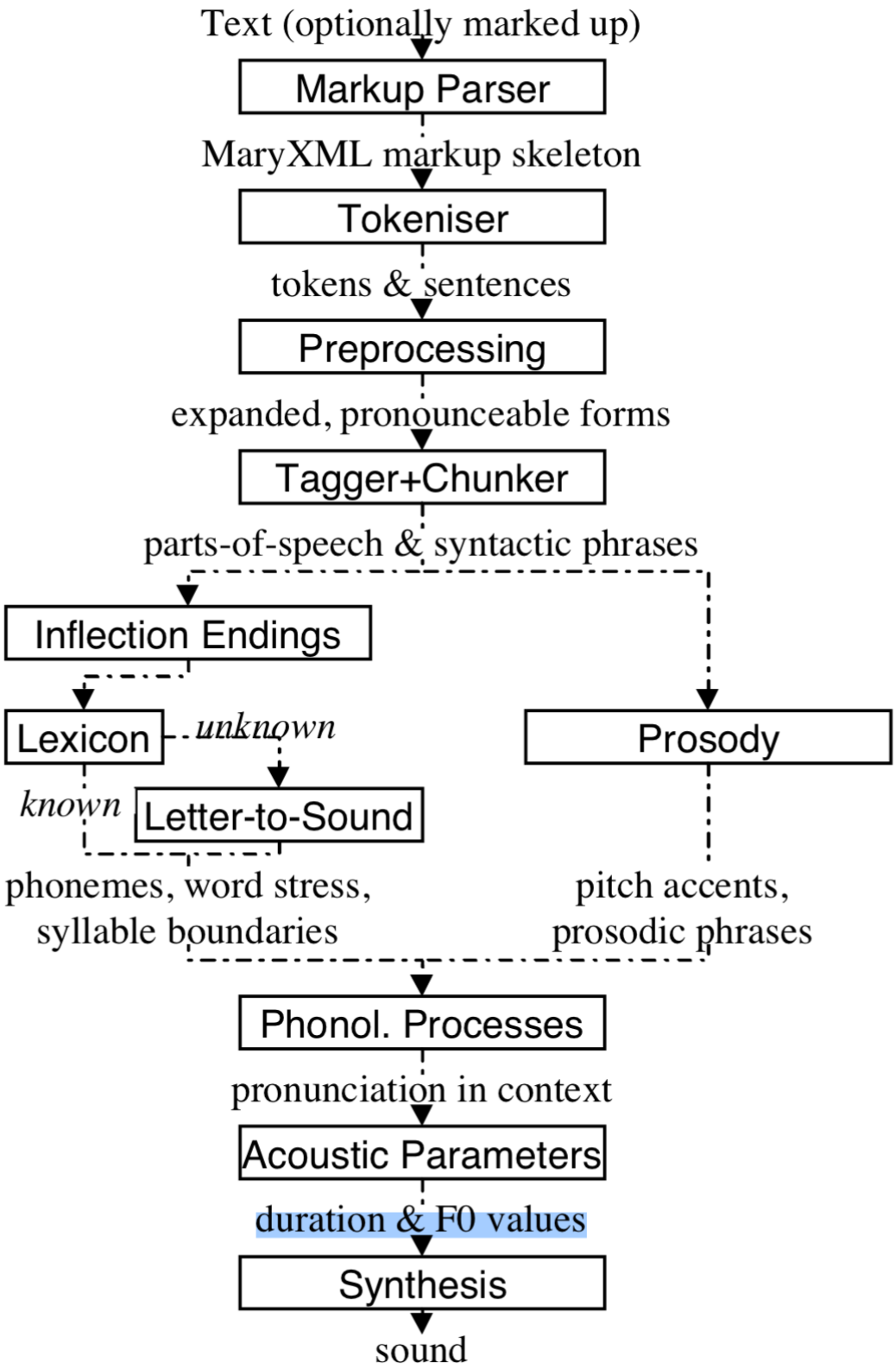


Figure 5. The architecture of the MARY TtS system (Schröder & Trouvain, 2003)

MARY TTS provides support for Emotional ML for speech synthesis.

2.6 ACAPELA GROUP TEXT TO SPEECH

This Text to speech is commercially provided by acapela group. They carry an expertise in creating digitized voices. It provides personalized voice tuning options with the use of their acapela tags (*Acapela Group*, 2020.). These tags are placed in the beginning of the text that is to be tuned. The acapela tags are listed below (*Acapela Tags*, 2020.):

2.6.1 Phonetic Tags

Same word can have different pronunciation depending on the places. Therefore, this tag allows users to change the pronunciation of the word.

Tag: \prn= {numeric value}\

2.6.2 Speed Tag

This tag allows us to change the speed of the voice spoken.

Tag: \rspd={numeric value}\

2.6.3 Voice shaping tag

This allows us to produce voices for different age groups. This tag basically helps alter the pitch of the voice.

Tag: \vct={numeric value}\

2.6.4 Spelling tag

This tag allows the given word to spell out.

Tag: \rms={numeric value}\

2.6.5 Pause tag

This tag lets us define the pause between the text.

Tag: \pau={numeric value}\

2.6.6 Audio tag

This allows the insertion of any audio or music in the text.

Tag: \aud=”{filepath}”\

2.6.7 Voice Switch tag

We can switch among the available speakers with these tags.

Tag: \vce=speaker={speaker name}\

2.6.8 Expressive voices

It allows the emotional voices. Currently, It supports happy, sad, bad guy, from afar and up close voices for speaker Will for the US English language (*Expressive Synthetic Speech*, n.d.).

Tag: \vce=speaker={speakername-expression}\

3 EXPERIMENTAL DESIGN AND IMPLEMENTATION

In this section, we present the details about our experimentation. We set up to three implementational experiments in this study among which one was incomplete. All the experiments utilize the same datasets and system design diagram.

3.1 DATASETS

For all three experimental work, we used announcement text from the passenger information system of the user client database. These texts are the sentences for announcing information on railway platforms to their travelling passengers. For comparing the emotions accuracy of generated audio with TESS (Pichora-Fuller & Dupuis, 2020), the same phrases as spoken in TESS audio were generated from the experimentation.

3.2 SYSTEM DESIGN DIAGRAM

All the experiments follow the same system design and the change made for each implementation were with the use of TTS and speech mark-up languages. The input for all the experiment can be any form of electronics text and the output generated is emotional speech.

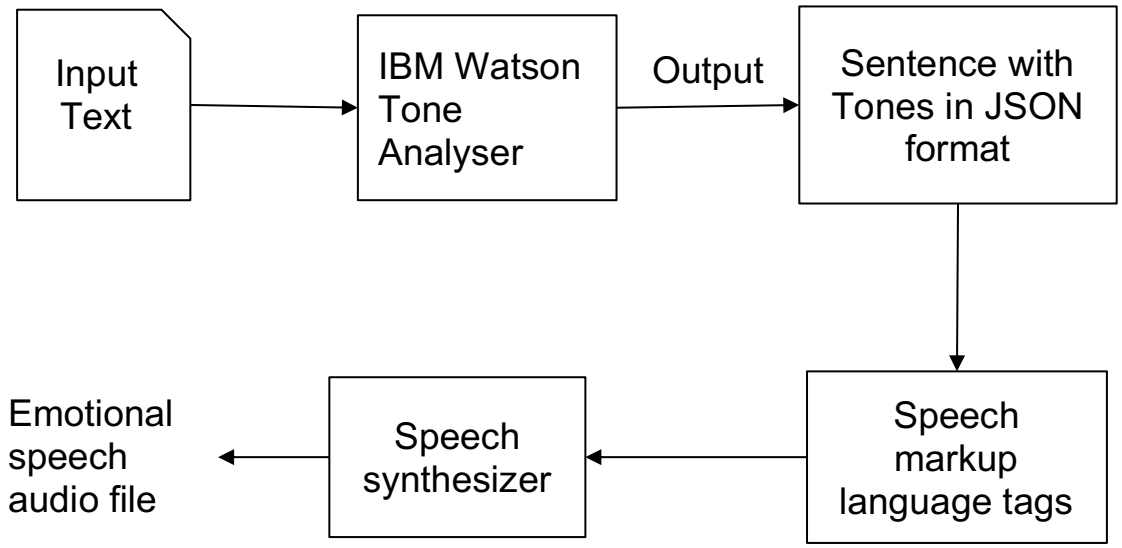


Figure 6. System design diagram

3.3. FIRST IMPLEMENTATION WITH ACAPELA

First, we enlist the equipment of implementation, then the system design of the experiment is explained.

3.3.1 Equipment

- IBM Wastson tone analyser
- Acapela group TTS
- Acapela group Tags

3.3.2 Design

This experiment follows the Figure 6 system design diagram. The implementation was started with the use of Acapela group TTS that was chosen by a case company as the TTS provided a great variety of languages and support. But this TTS was not implemented. The reason for the failure to implement the Acapela group TTS was that it was built in the C/C++ language.

In order to make it work together with our experimentation, the C wrapper was required. Due to the lack of time, we could not proceed with this TTS for our research. Therefore, we opt out for alternative TTS.

3.4 SECOND IMPLEMENTATION WITH IBM

We first enlist all equipment that are used for this implementation, then the system design of the experiment is explained.

3.4.1 Equipment

- IBM Wastson tone analyser
- IBM Wastson text to speech
- SSML

3.4.2 Design

The first block of system consists of IBM Waston tone analyser as shown in Figure 6. The text input is first fed to a tone analyser from which we receive the output in JSON format. This JSON data contains results of text with its tone and score value of the tone. The JSON data provides the tone in both document level and sentence level with their respective tone score. For this research we prioritize sentence level tone and in absence of tone value in sentence we utilize document level tone results. The acceptable score for synthesis was above 0.6. Then, Speech Synthesis mark-up language (SSML) was used to annotate the emotional effect on the text. Specific value to the attributes of the prosody element of SSML was created for four emotions of anger, fear, joy and sadness. The attributes of prosody, such as pitch variables and speaking speed have been previously analysed (Amir et al., 2001; Mcgilloway et al., n.d.) and implemented in emotional speech synthesis. Therefore, the values of pitch, rate and volume attribute of prosody was decided on the basis of the study made by

(Crumpton, 2015) on the use of prosody for emotions expression. Therefore, prosody attribute values for this study was derived from the previous study and are enlisted in Table 1 (Crumpton, 2015). Table 3 illustrates the prosody value used in this experiment. These values for each emotion were used to enclose the sentence as per their tone.

After these sentences were enclosed inside a predefined SSML, the annotated text was then submitted to IBM TTS and the audio file was generated. IBM TTS provides the support for SSML. This audio file is further evaluated for its quality of emotion that it possessed. For this experimentation lite version of cloud IBM Watson TTS service was used. The api key from IBM was used to make API calls to synthesize the input text.

	Pitch	**Rate**	**Volume**
Anger	x-high	fast	x-loud
Fear	high	fast	x-loud
Joy	high	medium	x-loud
Sad	x-low	x-slow	x-soft

Table 3. SSML values used to create emotion

3.5 THIRD IMPLEMENTATION WITH MARY

First all equipment that was used for this implementation are enlisted, then the system design of the experiment is explained.

3.5.1 Equipment

- IBM Wastson tone analyser
- MARY TTS
- EmotionML

3.5.2 Design

This implementation follows the system diagram of Figure 6. The first block of system consists of IBM Waston tone analyser which analyses input text, then the text with tone results in it was obtained in JSON format data. In this experiment we use MARY TTS which has support for Emotion mark-up language. Emotion mark-up language incorporates all four major emotions: anger, fear, joy and sadness inside the emotion element with the category name values of “anger”, “fear”, “happiness” and “sadness”. The category set value used for our experiment is “http://www.w3.org/TR/emotion-voc/xml#big6” (Recommendation, 2014b). Therefore, we combine MARY and EmotionML.

In this experiment, we first consider the sentence level tone and when this detail is not available, we choose the tone details from document level. The toned texts along with tone score is received in JSON format from IBM Watson tone analyser. The acceptable tone score here is also above 0.6. The text from JSON data is then enclosed inside the emotion element of EmotionML with category name as per resulting tone. These annotated texts are then synthesized using MARY. The audio files were generated in wav file format which is further evaluated for its quality of emotion that it possessed. For this experiment MARY was installed in a local machine.

4 RESULT

This section illustrates the result achieved from our study. The focus of our study was to investigate the quality of audios for its naturalness, cost and accuracy of the system. This purpose of our study was achieved by descriptive analysis of output audios from our study and its comparison with pre-recorded human voices of the similar feature. This chapter presents the results of the audio analysis for the four stated emotion categories.

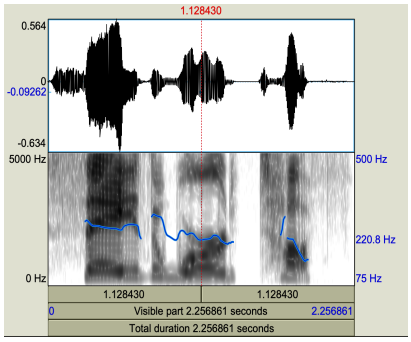
The descriptive audio images including waveform, spectrogram and pitch were reported. The presentation of the findings is arranged by the four research questions. The audio images of emotional audios from Toronto emotional speech set (TESS) Collection (Pichora-Fuller & Dupuis, 2020) and emotional audio from our two experimentation were used to answer research questions.

The Praat software package (Boersma, 2002) was used to visualize audio information. Praat is a free software for speech analysis. The sound in this software system is displayed in two split windows. The upper half shows a sound waveform while the lower half allows us to visualize spectrogram, the pitch contour, intensity and formant contours (Boersma, 2002). In this study, we analyse spectrogram and pitch information. These data on audio images are utilized to study the quality of generated speech in each emotion category stated in this study. Figure 7 shows the waveform, spectrogram and pitch of each emotion category audio obtained for the same line of text.

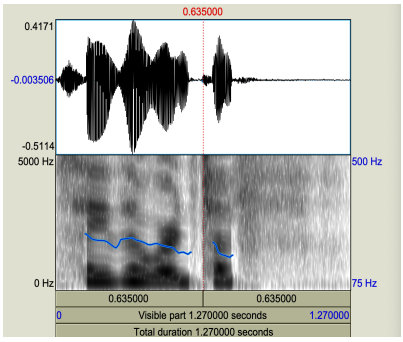
Audio from TESS

Audio from MARY TTS

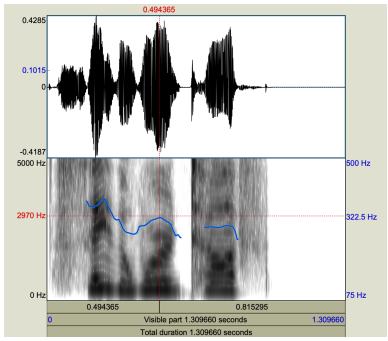
Audio from IBM TTS



(a)

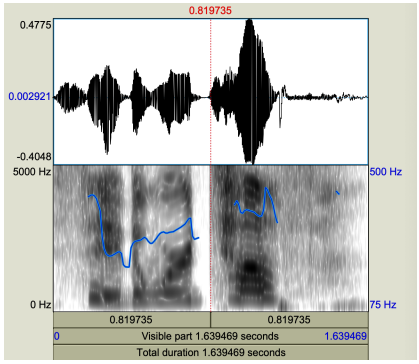


(b)

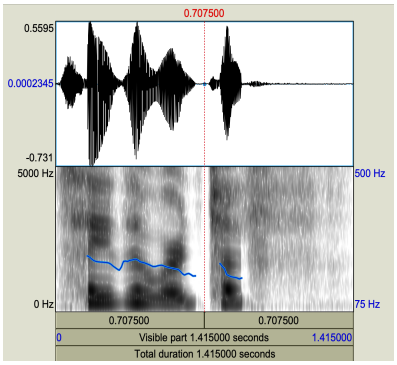


(c)

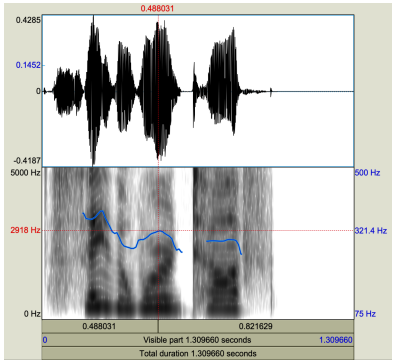
Anger



(d)

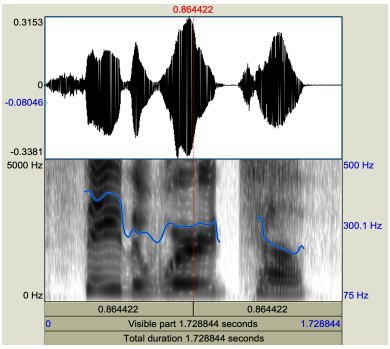


(e)

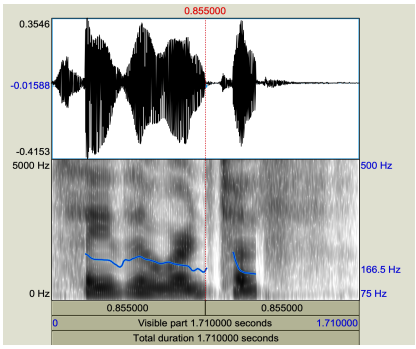


(f)

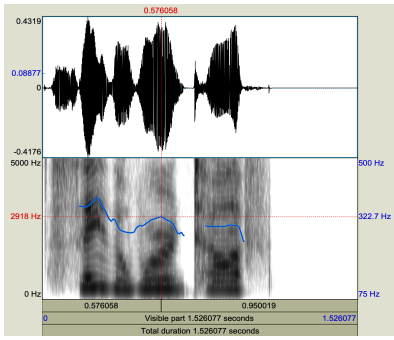
Fear



(g)

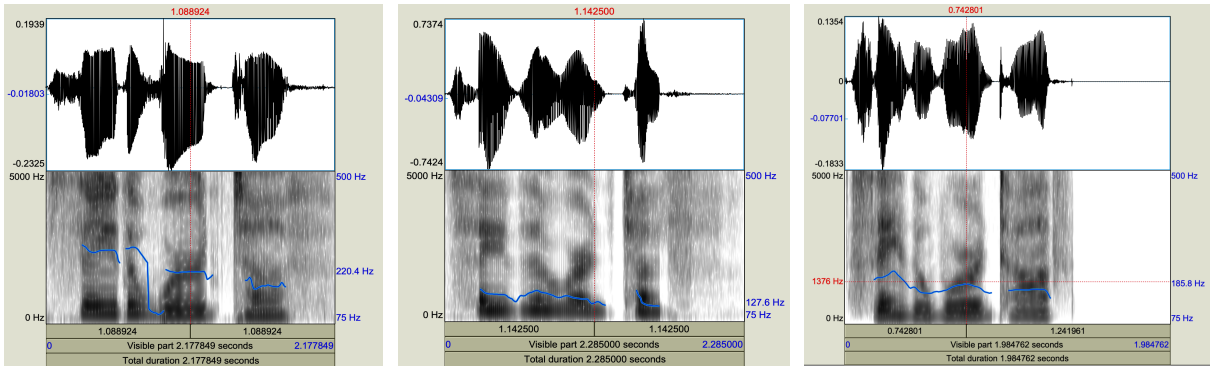


(h)



(i)

Happy



(j) (k) (l)

Sad

Figure 7. Audio from TESS, MARY TTS and IBM TTS

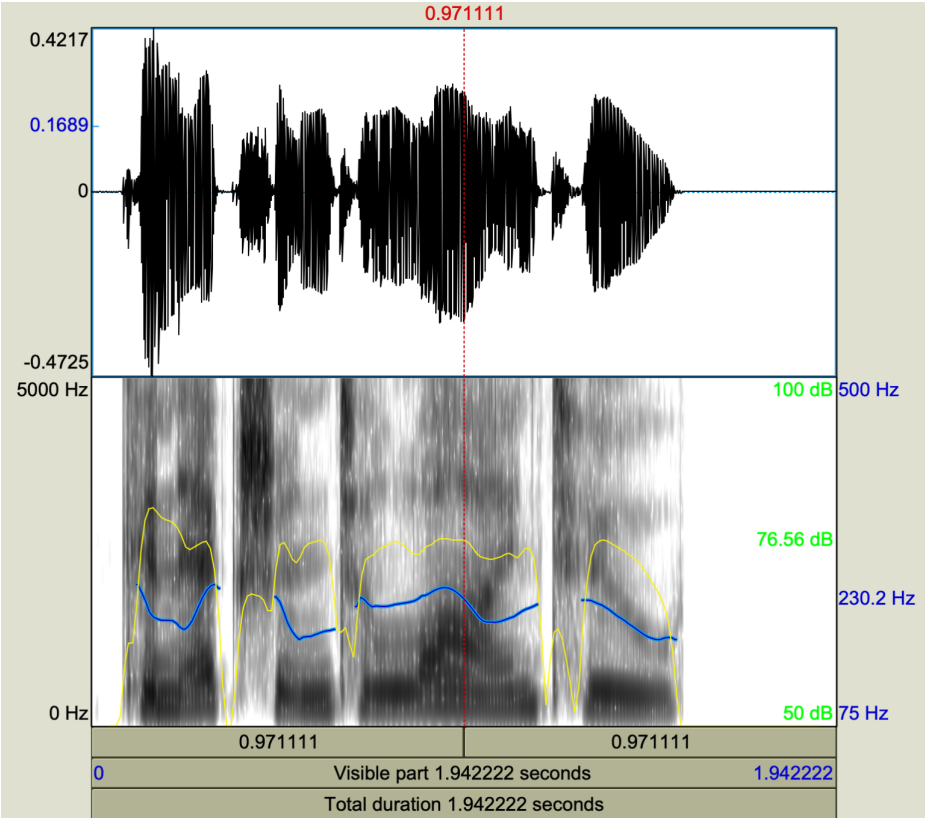


Figure 8. IBM Watson TTS and SSML

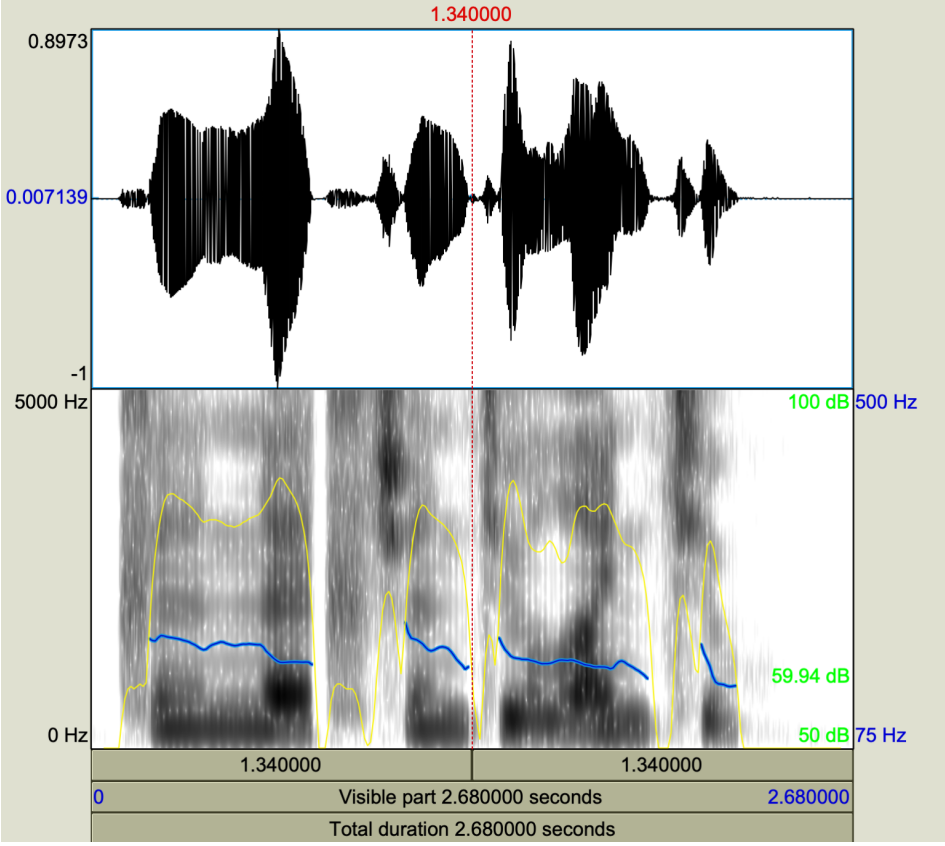


Figure 9. MARY TTS and EmotionML

Descriptive and inferential analysis were used to investigate the four research questions of this study. The four research questions are:

1. How close the speech generated by a machine is to the human voice?
2. How accurate will the service be?
3. Will it be able to annotate pause, speed and pitch in the text where necessary?
4. What will be the cost of the service?

To investigate the first research question, the above Figure 7 was used to compare the recorded human audio and the audio from this study. Second and third, research questions were explored by examining Figure 8 and 9. Lastly

for the fourth research question an execution time for various length of character was observed, few of which are listed in Table 4.

The first research questions examine the audio images in Figure 7. The left-hand side audios belong to TESS (Pichora-Fuller & Dupuis, 2020), the audio images in the middle column of Figure 7 are generated from MARY TTS using EmotionalML tags and the right-hand side audio images are obtained from IBM Watson TTS with the use of SSML for emotion tagging in our study. The upper window of each Figure 7(a), 7(b), 7(c), 7(d), 7(e), 7(f), 7(g), 7(h), 7(i), 7(j), 7(k), 7(l) provides the waveform of audio and the lower window of provides the spectrogram of the audio seen in the gray colour and the pitch contours in the blue lines. It has the audio total time information at the bottom and provides frequency information in Hertz (Hz). All the audio in images are for the same piece of text spoken in a female voice. From the Figure 7 above, we can see that there exist differences in the audios.

For the second and third research question, Figure 8 and 9 were explored. Figure 8 is an audio image of the first experimental implementation which uses IBM Waston TTS and SSML while Figure 9 is the audio from MARY TTS using EmotionML. From the figures we can see that the audio from MARY seems weak in comparison to the audio from IBM TTS. MARY possesses low pitch and low volume audio while IBM TTS is higher in pitch and volume. This was also noticed while listening to the audios. Also, the spectrogram is denser in IBM TTS and has better voice quality.

For the fourth research question, text with multiple character length were used for speech synthesizing in both experiments and execution time was observed and listed in Table 4 below.

Character length	IBM Execution time (sec)	MARY Execution time (sec)
117	8.138	2.047
23	4.512	1.763
204	11.503	2.353
152	9.085	3.275
20	4.400	1.432

Table 4. Execution time of TTS

In this chapter, an introduction was given regarding the analysis that was to be discussed. This was followed by an image analysis of the sample audio clips. The results revealed that there persisted a big difference in audios in terms emphasis for each word, total time for speech and frequency value.

5 DISCUSSION

The earlier chapter reported the presentation and analysis of data. This chapter consists of a summary of the study, discussion on the findings, implications for practice, and conclusions. The later sections expand concepts that are studied to provide a further understanding of the use of speech mark-up languages and their practices. The purpose of understanding of volume, pitch and speed behaviours and their impact on creating acoustic emotions is presented in the suggestions for further research. Finally, a synthesizing statement is offered to capture the substance and scope of what has been attempted in this research.

This chapter begins with a summary of the purpose and structure of the study and is followed by the major findings related to the emotional speech. The conclusions from the findings of this study are discussed in relation to the definition, function, and characteristics of a good theory. Finally, implications for practice and recommendations for further research are presented and discussed.

The purpose of this study was to analyse the quality and accuracy of emotional audio generated from a system set up that takes plain text as its first input, analyses for its tones, encloses text with emotion annotation as per its tone and finally generates speech output that conveys emotion.

The study included speech images that were acquired to visualize and inspect the audio characteristics. The spectrogram and pitch information of audio images data as in Figure 7. were compared with each other in order to answer our research question. Along with that audio image analysis, audio listening was conducted to deduce the expression impression of audio based on real-life knowledge to summarize what the speech was expressing.

This study involves four research questions that are enlisted below:

1. How close the speech generated by a machine is to the human voice?
2. How accurate will the service be?
3. Will it be able to annotate pause, speed and pitch in the text where necessary?
4. What will be the cost of the service?

These RQ are tested for its hypothesis and are discussed as below:

RQ 1. "How close the emotional speech generated by a machine will be to the human voice?"

- Hypothesis

  The produced speech from the machine will be able to provide prosody.

- Outcome produced

  The two different quality of audios were generated from our two-experiment set up. The audio produced from both the setup possessed prosody.

- Discussion

  Although the implementation was able to provide noticeable prosodic effect in the audio, those prosody did not fully give the clear idea of the emotion that it was trying to convey. The objective of this study was to obtain emotional voices with the use of prosody which was not achieved.

RQ 2. "How accurate will the service be?"

- Hypothesis

  The system will be able produce audio with enough emotion accuracy.

- Outcome produced

  Two audios from two experiment of which images are shown in Figure 7.
- Discussion

  The output audios do not possess the accuracy of the define emotion, hence falsifying the hypothesis. Especially for joy and fear emotion, audios were close to anger as we can see in figure 7.

RQ 3. "Will it be able to annotate pause, speed and pitch in the text where necessary?"

- Hypothesis

  The system will maintain a pause after a full stop. Speed and pitch for a sentence.
- Outcome produced

  Both the audio posses adequate pause, speech and pitch in their speech for neutral emotion.
- Discussion

  However, MARY showed significant amount of fluctuation in speed and pitch for the text with multiple tone. The case was less severe with IBM TTS compared to MARY.

RQ 4. "What will be the cost of the service?"

- Hypothesis

  The processing time should be less than 5 seconds.
- Outcome produced

  The execution time to synthesize text with speech mark-up tags was calculated for different character length sentences. This execution time is illustrated in Table 4.

- Discussion

  This service is for the purpose of announcement in the railway system. So, the processing time matters for the announcement service everywhere. It cannot take too long to synthesize speech. For example, if the announcement before the next stop is required then it should be fast enough to deliver the information before reaching that stop.

The research questions one, two and three were answered from the audio image data. These questions were answered using the results from a spectrogram of the output audios. The findings resulting from these audio images reveal a significant amount of differences in amplitude over time. The audio from EmotionML looks very weak compared to that from SSML. The audio of SSML tends to possess a similar pattern as the human voice but with lower frequency for the emotion of anger and happiness. The research question four were answered based on execution observation. The result infers that the use of a larger number of characters for input text required the longer time for speech synthesis.

The findings of this study contain a design and implementation of a system that is useful for the study of HCI. The study indicates TTS compatibility with speech mark-up language. This study offers insight into the quality of speech generated by two TTS and the support provided by each TTS for the speech mark-up language. It will also give the researchers a better idea on working with IBM Watson tone analyser, text to speech, MARY TTS and mark-up language. In particular, this study suggests SSML supports better user control.

6 CONCLUSION

In this thesis, a combination system of TTS, speech mark-up languages and speech tone analyser were designed for voice tuning and the output audios were validated for intended emotion. The case of a TTS application is presented, and its objectives of availability, scalability, reliability, and needed resources are defined. The study was performed using two mark-up languages for setting up emotion namely SSML and EmotionML, and two TTS namely IBM Watson TTS and MARY TTS for speech synthesizing. The study investigation revealed that the emotion category provided in EmotionML does not produce accurate emotion expression. The use of SSML produced better emotional voice. The study result showed that deficient quality of audio was generated with the use of MARY TTS. On the other hand, IBM TTS has a better quality of voice. The process time of the system was also examined and found out that the MARY performed better that IBM TTS in this case. We were able to construct our designed system in this study. However, there are immense amount of work that could be done to further enhance the system.

7 FUTURE WORK

The goal of this study was to construct a combinational system of text tone analyser, speech mark-up languages and speech synthesizer and investigate the quality of speech generated from this set up. The system was constructed and tested for data. The output data was carried out to test four research questions relating to this goal. The information was studied, and findings were resulted from the examination of the data. The findings hold some limitations. The first limitation is that the research covers only four emotional characteristics. The second limitation is that the cost of execution tends to increase with increasing character length. The third limitation is that the current attributes of prosody generate poor quality of speech with emotions. Further research in this field should aim at these limitations and update the capabilities of solutions to improve the rate and pitch combination of audio along with handling an increasing and unexpected number of input text.

This study has thus provided some interesting considerations and answers to the central research question. The thesis gives a collection of different approaches of speech synthesis with the mark-up languages.

REFERENCES

Acapela group. (n.d.). https://www.acapela-group.com/about-us/

Acapela Tags. (n.d.). https://www.acapela-group.com/voices/voice-tuning/

Aida–Zade, K. R., Ardil, C., & Sharifova, A. M. (2010). *The Main Principles Of Text-To-Speech Synthesis System*. https://doi.org/10.5281/zenodo.1070638

Amir, N., Kerret, O., & Karlinski, D. (2001). *Classifying emotions in speech: A comparison of methods*. 4.

Asakura, T., Akama, S., Shimokawara, E., Yamaguchi, T., & Yamamoto, S. (2019). Emotional Speech Generator by using Generative Adversarial Networks. *Proceedings of the Tenth International Symposium on Information and Communication Technology*, 9–14.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Chen, F., & Jokinen, K. (2010). *Speech Technology*. Springer. https://doi.org/10.1007/978-0-387-73819-2

Crumpton, J. J. (2015). *Use of vocal prosody to express emotions in robotic speech*. Mississippi State University.

Dutoit, T. (1997). *An introduction to text-to-speech synthesis* (Vol. 3). Springer Science & Business Media.

Expressive Synthetic Speech. (n.d.). Retrieved 22 May 2020, from http://emosamples.syntheticspeech.de/

Hartmann, K., Siegert, I., Philippou-Hübner, D., & Wendemuth, A. (2013). Emotion Detection in HCI: From Speech Features to Emotion Space. *IFAC Proceedings Volumes; 12th IFAC Symposium on Analysis, Design, and Evaluation of Human-Machine Systems*, *46*(15), 288–295. https://doi.org/10.3182/20130811-5-US-2037.00049

Hevner, A., R, A., March, S., T, S., Park, Park, J., Ram, & Sudha. (2004). Design Science in Information Systems Research. *Management Information Systems Quarterly*, *28*, 75.

IBM General purpose tones. (2020). https://cloud.ibm.com/docs/tone-analyzer?topic=tone-analyzer-utgpe#tones-tone

IBM Watson Text to Speech. (2020). https://cloud.ibm.com/docs/services/text-to-speech?topic=text-to-speech-about

IBM Watson Tone Analyzer. (2020). https://cloud.ibm.com/docs/services/tone-analyzer?topic=tone-analyzer-about#about

Laukka, P. (2020). *Vocal Communication of Emotion* (V. Zeigler-Hill & T. K. Shackelford, Eds.; pp. 5725–5730). Springer International Publishing. https://doi.org/10.1007/978-3-319-24612-3_562

Lee, Y., Rabiee, A., & Lee, S.-Y. (2017). Emotional End-to-End Neural Speech Synthesizer. *ArXiv:1711.05447 [Cs, Eess]*. http://arxiv.org/abs/1711.05447

Lemmetty, S. (1999). Review of speech synthesis technology. *Helsinki University of Technology*, *320*, 79–90.

Mcgilloway, S., Cowie, R., Douglas-cowie, E., Gielen, S., Westerdijk, M., & Stroeve, S. (n.d.). *ISCA Archive APPROACHING AUTOMATIC RECOGNITION OF EMOTION FROM VOICE: A ROUGH BENCHMARK*.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pichora-Fuller, M., & Dupuis, K. (2020). *Toronto emotional speech set (TESS)*. https://doi.org/10.5683/SP2/E8H2MF

Recommendation, W. (2010). *SSML*. https://www.w3.org/TR/speech-synthesis/

Recommendation, W. (2014a). *EmotionML*. https://www.w3.org/TR/emotionml/#s2.1.1

Recommendation, W. (2014b). *Mechanism for referring to vocabularies*. https://www.w3.org/TR/emotionml/#s3.2

Reddy, M. G., Harikrishna, D. M., Rao, K. S., & Manjunath, K. E. (2015). *Telugu emotional story speech synthesis using SABLE markup language*. 331–335. https://doi.org/10.1109/SPACES.2015.7058278

Schröder, M. (2001). Emotional speech synthesis: A review. *Seventh European Conference on Speech Communication and Technology*.

Schröder, M., & Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, *6*(4), 365–377.

Sproat, R., Hunt, A., Ostendorf, M., Taylor, P., Black, A. W., Lenzo, K., & Edgington, M. (1998). *SABLE: A standard for TTS markup*.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., & Bengio, S. (2017). Tacotron: Towards end-to-end speech synthesis. *ArXiv Preprint ArXiv:1703.10135*.