VAASA UNIVERSITY OF APPLIED SCIENCES

TEXT-TO-SPEECH SOFTWARE COMPARISON

Ying Zheng

Technology and Communication
2010

VAASAN AMMATTIKORKEAKOULU
UNIVERSITY OF APPLIED SCIENCES
Degree Program of Information Technology

ABSTRACT

Author	Ying Zheng
Title	Text to Speech software Comparison
Year	2010
Language	English
Pages	42 + 4 Appendices
Name of Supervisor	Smail Menani

This work has been initiated on request of ABB aiming to improve the Text–to–Speech solution for e-learning programs in the company. In the thesis, the author reviews e-learning program and Text–to–Speech software use, prepares a requirements gathering survey, a requirement specification, products research, testing and conducts a Text–to–Speech evaluation survey resulting in suitable tool(s) for ABB. Testing of Text–to–Speech software tools was concentrated on the voice quality which meant naturalness of sounding and intelligibility of speech, and functional features.

Keywords TTS (Text–to–Speech), e-learning program,

ACKNOWLEDGMENTS

This thesis arose during four months of research. By that time, I have worked with great people whose contribution in assorted ways to the research and the thesis deserve special mention. It is a pleasure to convey my gratitude to them all in my humble acknowledgment.

In the first place, I would like to give great thanks to Auli Koivunen for offering the thesis title, supervision, and guidance from the very early stages of this project as well as giving me extraordinary experiences through out the work. She taught me how to express my ideas and the different ways to approach a research problem. Above all and most needed, she encouraged me unflinchingly and supported me in various ways. I am indebted to her more than she knows.

A special thanks to Dr. Smail Menani for his advice, and crucial contribution. His intuition exceptionally inspires and enriches my growth as a student. Dr. Smail Menani's recommendations have truly guided me to produce well thought out research.

Many thanks go in particular to Sylvie Moisy and Matias Pyy. I will always remember them with gratitude for their valuable advice in discussions, their great support for the whole project, and using their precious times to give critical comments.

I also benefited form the survey participants and ABB e-learning developers. Their recommendations, comments and rich experience in e-learning development helped me in my research.

Notes of thanks are also given to Nyholm Birgitta(BU LV Motors training manager), Maira Forsti (BU LV Motors training coordinator), and Qian Wu(BU LV Motors trainee), who helped with on the requirement documents.

Contents

1 Introduction

This introduction is written as a brief guide to the theme. In addition, it will present the research purpose, focus, and summary of the thesis.

1.1 Review

The project was initiated for ABB. ABB is a leader in power and automation technologies that enable utility and industry customers to improve performance while lowering environmental impact. The ABB Group of companies operates in about 100 countries and employs around 108,200 people. (http://www.abb.com)

ABB training, learning and development are provided to ABB employees, channel partners and clients in categories of People and leadership competencies, Business process and tools and Products, Technology and solutions.

E-learning programs offer web-based courses for employees and channel partners along with up-to-date technologies for existing and new products. As the main tool in e-learning course development, high quality Text–to–Speech software tool guarantees the high quality of the training and e-learning courses makes employees and channel partners the best result and fulfils expectations and requirements.

1.2 Objectives

The main objectives of this project were

- To describe the problems in creating speech for e-learning course
- To search the Text–to–Speech software tools on the market
- To identify Text–to–Speech software needs of e-learning program
- To implement the comparison testing

- To recommend the most appropriate Text–to–Speech software tools for ABB e-learning development.

The practical part of the project was a constructive researching and testing of the text–to–speech software tool features that would fulfill the requirements of ABB e-learning course development. The starting point of the research was the meeting with the project support team. The materials from meeting of the ABB e-learning developers' needs are used to form the requirements for the text–to–speech solution improvement. Requirements were used to later identify testing criteria of text–to–speech software tools.

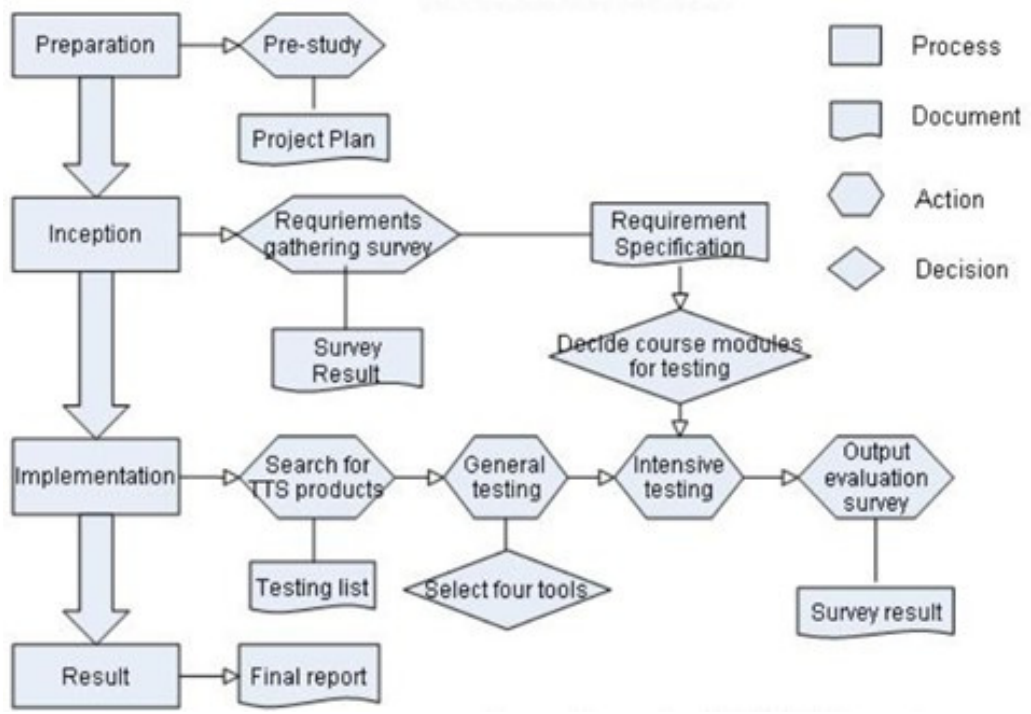


Figure 1.3.1 TTS comparison research flow chart.

1.3 Problem statement

ABB builds text–to–speech solutions to create the speech for ABB e-learning courses. Hiring the native speakers to record the speech is not a good option.

First of all, there are numerous speeches that needed to be recorded in e-learning courses. It is not economical to create spoken scripts that depend on human recording. Secondly, as technology develops, ABB e-learning courses are updated frequently. Thus, it is impossible to permanently retain the same native speaker to update the script of an e-learning course that needs to be frequently regenerated with updated course content. After that, e-learning courses are applied in different organizations, functions, and countries in ABB. Quite a number of courses need to be implemented in different languages, such as Chinese, Spanish, and French, etc. The multiple-language requirement makes it difficult to hire native speakers in all- kinds of e-learning course languages.

In contrast with human recording, the text–to–speech software tool is preferred for creating the speech for e-learning courses. In the past year, Loquendo TTS was used as a main tool in e-learning course development. However, it didn't fulfill the e-learning developers' needs, especially in user interface and voice quality.

On the other hand, it appears more text–to–speech software tools are continuously produced with new solutions and improvements for speech synthesis technology. The different functionalities of the software tools as well as increasing price competition make it important to compare the available text–to–speech tools.

Hence, the success of comparison and selection of text–to–speech product will promote the quality of ABB e-learning.

2 Text–to–Speech Synthesis

Speech synthesis is the transformation of text to speech. This transformation converts the text to synthetic speech that is as close to real speech as possible in compliance with the communication norms of special languages. [1] A computer system is used for the purpose of automatically generating speech output from data input which may include plain text, formatted text, or binary objects called a Speech Synthesizer and which can be implemented in software or hardware.

Speech Synthesis	The process of automatic generation of speech output from data input which may include plain text, formatted text, or binary objects.
Text-To-Speech	The process of automatic generation of speech output from text or annotated text input.

There are three generations of speech synthesis systems summarized by K.R. Aida – Zade, C. Ardil and A.M. Sharifove in the article *The main principles of Text–to–Speech Synthesis System* [1]: "During the first generation (1962-1977) formant synthesis of phonemes was the dominant technology. This technology made use of the rules based on phonetic decomposition of sentence to formant frequency contours. The intelligibility and naturalness were poor in such synthesis. In the second generation of speech synthesis methods (from 1977 to 1992) the diphones were represented with the LPC parameters. It was shown that good intelligibility of synthetic speech could be reliably obtained from text input by concatenating the appropriate diphone units. The intelligibility improved over formant synthesis, but the naturalness of the synthetic speech remained low. The third generation of speech synthesis technology is the period from 1992 to the present day. This generation is marked by the method of "unit selection synthesis" which was

introduced and perfected, by Sagisaka at ATR Labs. in Kyoto. The resulting synthetic speech of this period was close to human generated speech in terms of intelligibility and naturalness."

The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood, which can be simplified as two parameters, naturalness of sounding and intelligibility of speech. A Text–to–Speech system has to model both the generic, phonetic features that make speech intelligible, and the idiosyncratic, acoustic characteristics that make it human.

2.1 Overview of Speech Synthesis Processes

A Text–to–Speech system (or "engine") is composed of two main parts [2]: Texts–to–Phoneme (Natural Language Processing, NLP) and Phoneme–to–Speech (Digital Signal Processing, DSP).

TTS = NLP + DSP

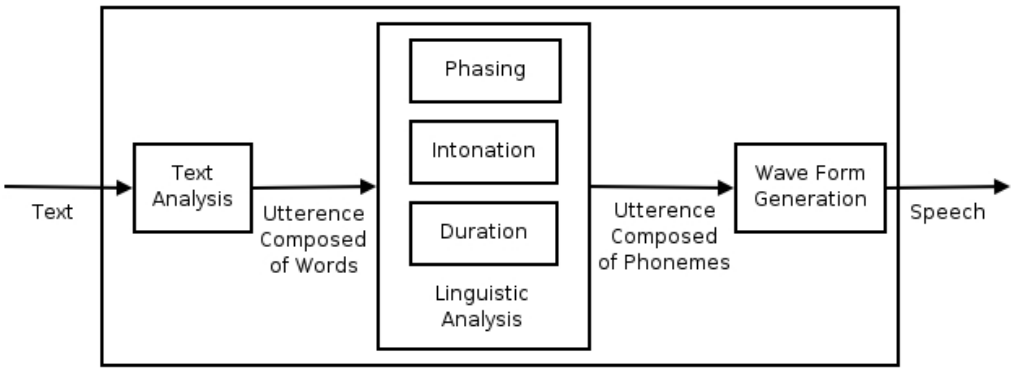


Figure 2.1 Overview of a typical TTS system

Texts–to–Phoneme: Also called a Grapheme–to–Phoneme conversion, the process of assigning phonetic transcription to words. The text must be converted into a linguistic representation that includes the phonemes to be produced, their duration,

the location of phrase boundaries, and the pitch / frequency contours for each phrase.

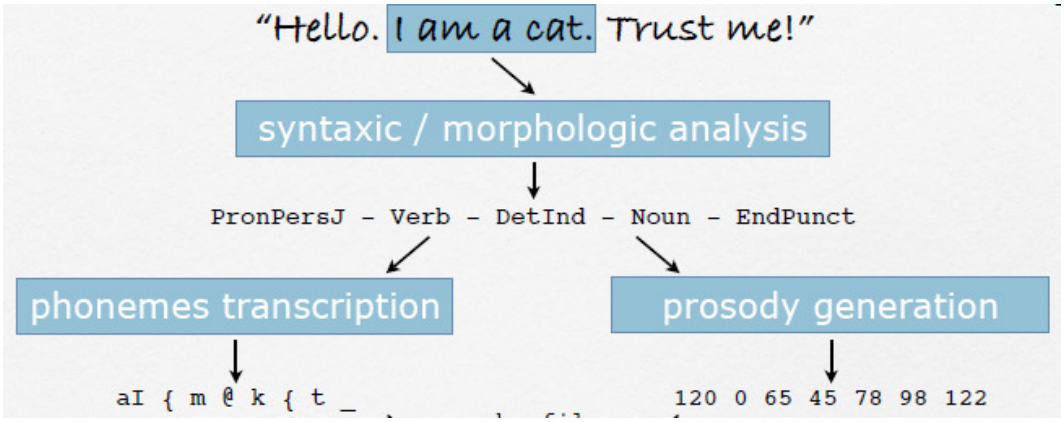


Figure 2.2 Texts – to – Phoneme.

Phoneme–to–Speech: The Phonetic transcription and prosody information obtained in the linguistic analysis stage are converted into an acoustic waveform.

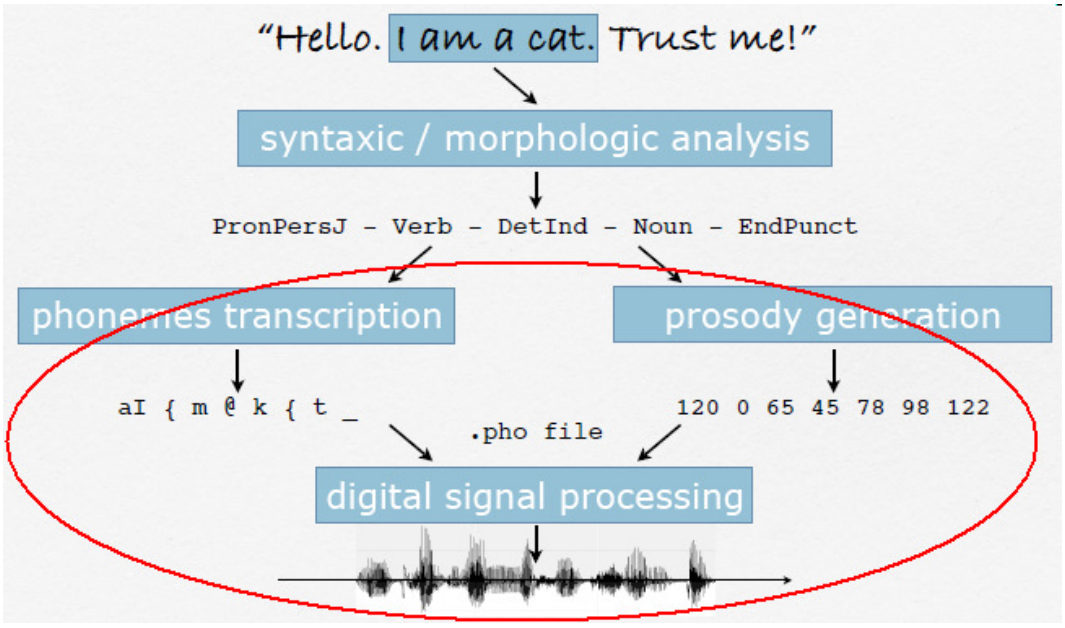


Figure 2.3 Phoneme – to – Speech.

While text is rich in phonetic information, it contains little or nothing about the vocal qualities that denote emotional states, moods, and variegations in emphasis or attitude. The elements of prosody (register, accentuation, intonation, and speed of delivery) are barely represented in the orthography (written representation) of a text. Yet without them, a synthesized voice sound monotonous and unnatural.

Concatenative synthesis and format synthesis are the two primary technologies to generate synthetic speech waveforms.

“Concatenative synthesis – Concatenative synthesis is based on the concatenation (or stringing together) of segments of recorded speech. Generally, concatenative synthesis produces the most natural-sounding synthesized speech. However, differences between natural variations in speech and the nature of the automated techniques for segmenting the waveforms sometimes result in audible glitches in the output. There are three main sub-types of concatenative synthesis.” [3]

“Formant synthesis – Formant synthesis does not use human speech samples at runtime. Instead, the synthesized speech output is created using additive synthesis and an acoustic model. Parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech. This method is sometimes called rules-based synthesis; however, many concatenative systems also have rules-based components. Many systems based on formant synthesis technology generate artificial, robotic-sounding speech that would never be mistaken for human speech. ” [3]

2.2 Speech synthesis within Windows Operating System

SAPI – Speech Application Programming Interface is designed for a software application to perform speech recognition and speech synthesis to work with the Microsoft Windows system.

Nowadays, SAPI4- and SAPI5-based speech systems are widely used in modern Windows systems. Text–to–Speech is the ability of the operating system to play back printed text as spoken words. [4] The driver installed with the operating system, which is called a Speech Synthesis engine, recognizes the text and uses synthesized voices which are pre-generated by a third-party manufacturer. Additional engines (for instance, certain jargon or vocabulary) are also available through third-party manufacturers. [4]

2.3 Markup Language for Text–to–Speech Synthesis

In order to make the most efficient use of computers in the processing of online text, it is necessary to have mechanisms for making the features that are deemed to be salient, but which might be difficult or impossible to automatically detect in a general way. [5]

The mark-up language provides a standard way to control aspects of speech, such as pronunciation, pitch, and rate. There are several mark-up languages in an XML-compliant format for the rendition of text as speech such as VXML (Voice Extensible Markup Language), STML (Spoken Text Markup language) and SSML (Speech Synthesis Mark-up Language).

SSML – Speech Synthesis Mark-up Language was developed at Edinburg University and was the first attempt in a TTS mark-up language. [5] SSML, known as a W3C [6] standard in 2004, is used to improve the quality of synthesized content. The essential role of the markup language is to provide authors of synthesizable content a standard way to control aspects of speech such as pronunciation, volume, pitch, and rate. across different synthesis-capable platforms. [7]

A Text-To-Speech system that supports the Speech Synthesis Mark-up Language will be responsible for rendering a document as spoken output and for using the information contained in the mark-up to render the document as intended by the author. [7]

```
<?xml version="1.0"?>
<!DOCTYPE speak PUBLIC "-//W3C//DTD SYNTHESIS 1.0//EN"
                "http://www.w3.org/TR/speech-synthesis/synthesis.dtd">
<speak version="1.0"
      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-synthesis/synthesis.xsd"
      xml:lang="en-US">

  <lexicon uri="http://www.example.com/lexicon.file"/>
  <lexicon uri="http://www.example.com/strange-words.file"
         type="media-type"/>
  ...
</speak>
```

Figure 2.3.1 Pronunciation Lexicon: “lexicon” elements

2.4 Text–to–Speech application

Currently, there are a number of applications; plug-ins and gadgets widely used as speech-synthesis technology tools. A great many Text–to–Speech systems in multiple languages are commonly used for desktop, server, telephone, and internet applications.

Modern speech synthesis technologies involve complicated and sophisticated methods and algorithms. [1] AT & T Bell Laboratories [8] (Lucent Technology) and the Centre for Speech Technology Research at, Edinburg University are perhaps two of the best known research organizations with long traditions in speech synthesis. In this day and age, it is still difficult to tell which approaches are more useful, though more and more speech synthesis systems appear on the market such as Neo Speech, Acapela Group, and Natural Soft.

Company	Location	Available languages
Natural Soft	Vancouver, BC Canada http://www.naturalreaders.com/	English, Canadian, Spanish, French, German, Italian, Swedish, Arabic
Loquendo	Italy http://www.loquendo.com	English, French, German, Italian, Portuguese, Russian, Spanish, Arabic, Danish, Dutch, Swedish, Finnish, Mandarin Chinese, Greek, Galician, Valencian, Polish
Acapela	December 2003: Acapela Group evolves from the strategic combination of three major European companies in vocal technologies: Babel Technologies (Belgium, 1997), Infovox (Sweden, 1983), and Elan Speech (France, 1980) http://www.acapela-group.com	English, French, German, Italian, Portuguese, Russian, Spanish, Arabic, Danish, Dutch, Finnish, Swedish, Norwegian, Czech, Greek, Polish, Turkish
NeoSpeech	California, U.S.A. http://www.neospeech.com/	English, Korean, Japanese, Chinese, Spanish, French* (under

		development)
AT&T	U.S.A. http://www.naturalvoices.att.com/	English, Spanish, Italian, German, French
IVOA	Poland, 2001 http://www.ivona.com	English, Romanian, Polish

3 Requirement Gathering

Requirements for a speech synthesis platform were formed by using two sources: discussions with the support team and a send requirement-gathering survey of ABB e-learning developers. Communication with the support team provided the theoretical requirements, which indicated what things should be done in common. On the other hand, the requirement gathering survey was launched in order to find out the everyday practical demands, and the testing criteria.

3.1 Requirements from support team

The support team was conducted by Head of Sales People Development and Training, and two e-learning developers from Process Automation Division (France), and Discrete Automation and Motion Division (Helsinki, Finland). Their diversity of experience and backgrounds contributed towards different needs of text–to–speech software tools can be chosen by the company.

The meeting with the support team discussed the minimum requirements of TTS software tools that would be chosen as the ABB standard and what questions would be designed in the requirements-gathering survey of e-learning developers. The support team pointed out that voice quality was the most significant criteria for choosing the text–to–speech software tools to create ABB e-learning course synthesized audio.

In respect that ABB e-learning developers are employed in different locations, the unionization of the synthesized audio features and software application updates can be achieved by a client-server architecture text–to–speech solution. No matter which TTS software application(s) will be chosen as the ABB standard tool(s), they will be kept update for voice quality, language availability, functional features, etc. Without the client-server architecture, it is difficult to implement the TTS engine for each end user in the ABB workplaces around the world.

In the questionnaire to e-learning developers, the voice quality would be specified in various aspects based on but ABB e-learning course content, for example the pronunciation of product terms. Furthermore, in order to use the software tool(s) as ABB global standard application, the variety of languages in the software user interface and voices in were also mandatory.

3.2 Requirements-gathering survey

3.2.1 Requirement-gathering implementation

Currently, there are about one hundred e-learning developers responsible for using Text–to–Speech software tools to create the synthetic out-puts embedded in numerous Web-based courses. The Web-based questionnaire was sent to every BU (Business Unit) to collect the specific user requirements. The survey started on 18th February and ended on 25th February 2010. (Appendix 3, The requirements-gathering survey link). The questions were designed to focus on the usability of current TTS tools, aspects of voice quality, functions of TTS software, and so on.

Question 1: “What TTS software have you used?”

Question 2: “In total hours, how long is the e-learning course you created last year with TTS tools?”

These two questions figured out how the e-learning developer experienced courses in creating and using the TTS software tools.

Question 3: “Besides English, do you need to create e-learning courses in other languages? If yes, please specify the other languages.”

Question 4: “How important are the following operational characteristics of TTS products to you?(Ease of installation, Integration with other software, Speed of program running, Online resources available, Accessibility of technical support)”

Question 5: “How important are the following functional features of TTS products to you?(Voice control flexibility, Switching between multiple voices, User lexicon, Availability of multiple languages, Use outside of company network, support for multiple document types, Text spelling check, Ability to create many audio files at once)”

The above questions were tailored to reveal the functions that e-learning developers perform when creating audio for e-learning courses using text–to–speech software tools. It should give a clue how important and how often these features affect the efficiency of e-learning developers’ work.

Question 6: “Overall, how do you rate the quality of the current TTS tool?”

Question 7: “What problems do you now have when using current TTS software, and what you would like to change?”

The questions were meant to discover the weakness of the TTS software currently used in company.

Question 8: “We are going to create a test module to compare TTS products. What are the features you would like to include in testing? If you have other alternatives, please specify them. (Pronunciation of technical terms, Pronunciation of products names, Pronunciation of abbreviations, Pronunciation of number sequences, Pronunciation of functions/formulas, Switching among different language voices)”

It specified the feature of voice quality of TTS software tools. Each aspect would be the criteria for evaluating voice quality, which was the most essential quality of the TTS software tool.

Question 9: “If you know of some TTS alternatives to test, please list in order of preference and give comments on them.”

This question was planned in advance of searching for TTS software tools.

3.2.2 Survey result

The survey got responses from 21 e-learning developers. The requirements gathering results were similar as the support team expected that the features of voice quality and user interface were the most required. The survey results were presented in two groups: voice features and non-voice features including all the functional features, operational features and supporting resources.

In terms of the voice features, read aloud in long text, technical terms, calculation, product names/unit and functions/formulas were the most demanded in the TTS software tool. Since most of the ABB e-learning courses involved a wide range of products and technologies, it makes great sense for improving the working efficiency and ABB e-learning course quality if the TTS software tool has high quality in these voice features.

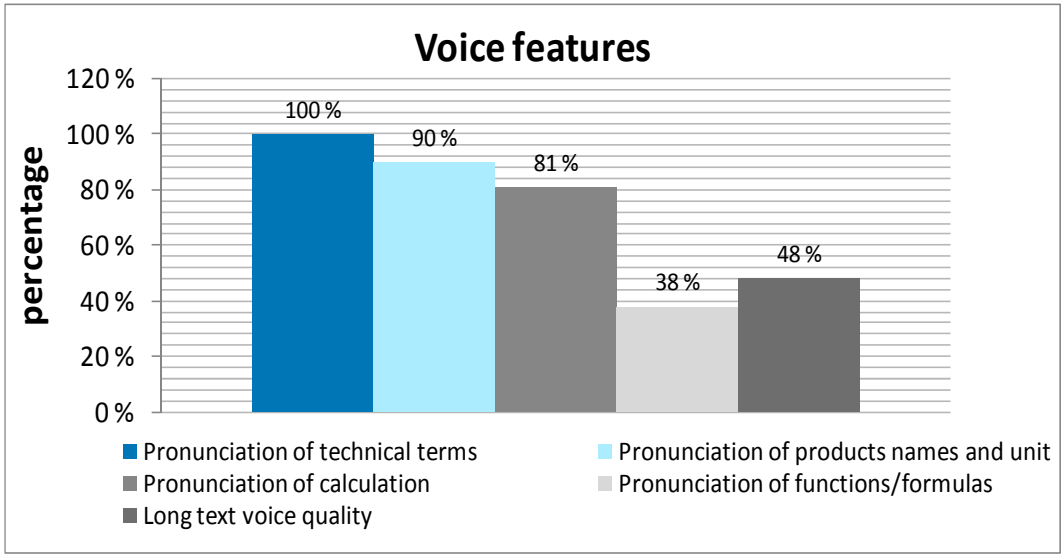


Figure 3.2.2.1 User demand on voice features

Besides the requirements in reading of text, the availability of multiple languages was one of the most significant requirements related to voice quality. The figure "Usage of multiple languages" indicated the usage of non-English languages in ABB e-learning course development. Leaving English aside, the usage of Spanish was 73%, and Russian, Chinese, French, German, and Italian were used in a wide range of e-leaning courses as well. Obviously, the demand for multiple languages was one of the most important criteria when evaluating text–to–speech software tools.

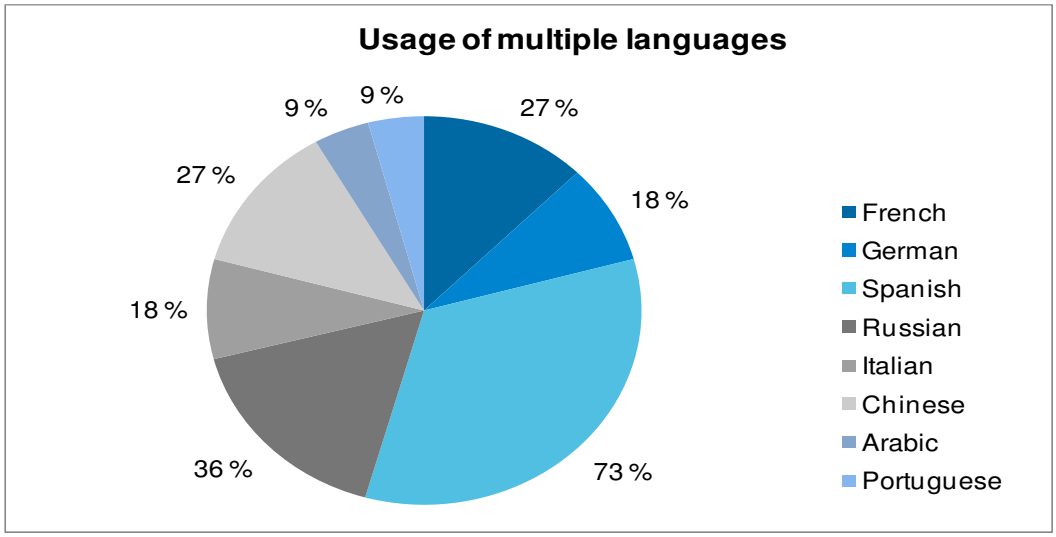


Figure 3.2.2.2 Usage of multiple languages

On the other hand, the requirements for non-voice features were summarized based on the order of their importance (Figures in appendix 2). Generally, the features not viewed as "Not important" would be classified as first priority requirements for voice quality features. Nevertheless, the requirements of "Accessibility of technical support" and "Online resources available" received responses of "Not important" by 10% of the respondents, so it would make great sense for users to develop the e-learning courses with successful support. It would directly affect the e-learning course quality. Hence, these two requirements must

be met in the TTS software tool, which meant the requirements were in the first priority.

With answers rating lower than 50%, features were treated as third level requirements in the comparison of TTS software tools. However, in this day and age, almost every software application performs in a high speed of computing environment, so the requirement of “Speed of program running” can be met by most TTS software tools. Nevertheless, it should be considered as second priority because 86% of the respondents rated it as “Important”. In addition, there were some specific software tools used in the e-learning course development, such as Articulate, which was not the common one integrated in the popular TTS software tools. The requirement of “Integration of other software” was better kept as a second-priority demand which was not related to the main criteria.

These requirements would be arranged following the priority of the requirement specification.

3.3 Requirement specification

This section arranged the requirements for selecting the text–to–speech software tool. According to the survey results, we created a table of features. Each feature has a unique identifier which was used during the whole project in each document, and it would provide traceability through all documents. Each feature was prioritized from 1 (highest) to 3 (lowest). Priority 1 means the feature is obligatory, 2 means it should exist, and 3 stood for it would be nice to have.

3.3.1 General

As the results from requirements gathering, the voice quality, user interface and some functional features should be treated as essential requirements for software tools to be considered as potential candidate ones. Other requirements were mandatory, but if not fulfilled, they must be compensated with equally useful features.

3.3.2 Use Cases

In this project, the use case methodology was applied on a general level in order to clarify the usability of text–to–speech software tools in e-learning development.

The Text–to–Speech synthetic audio development system should have client-server architecture. This is intended for two roles: the e-learning course developer (Figure 3.3.2.1) and the administrator (Figure 3.3.2.2). The main difference in use for the e-learning course developer and the higher-level administrator is being able to perform the Text–to–Speech engine maintenance, configuration, and defining the ABB e-learning course user lexicon.

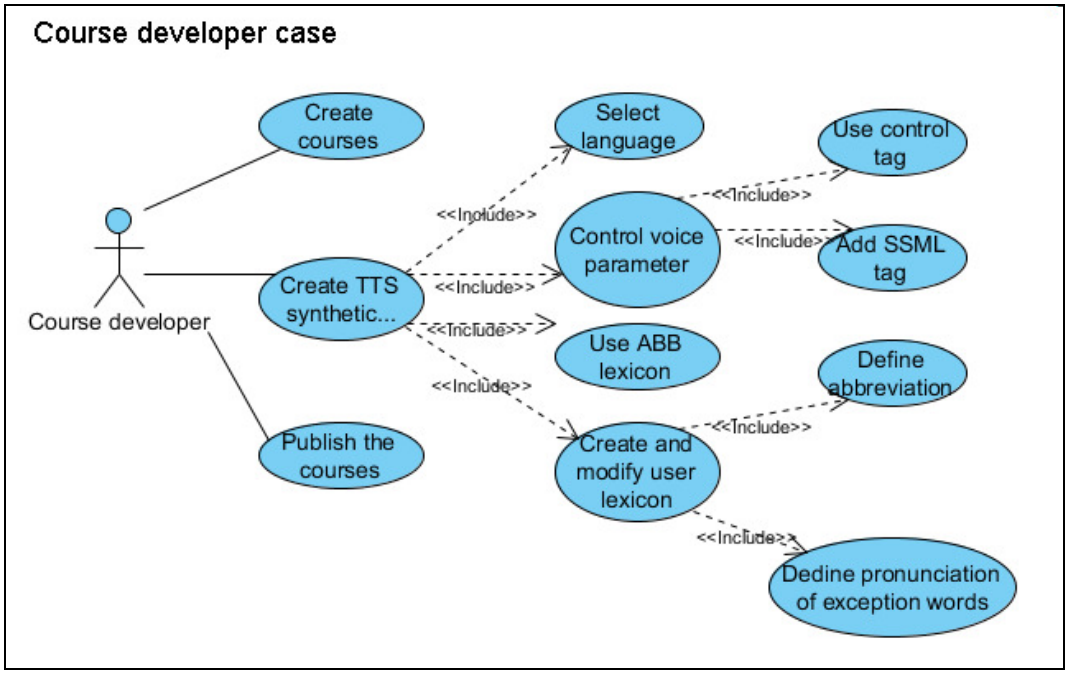


Figure 3.3.2.1 E-learning course developer use case

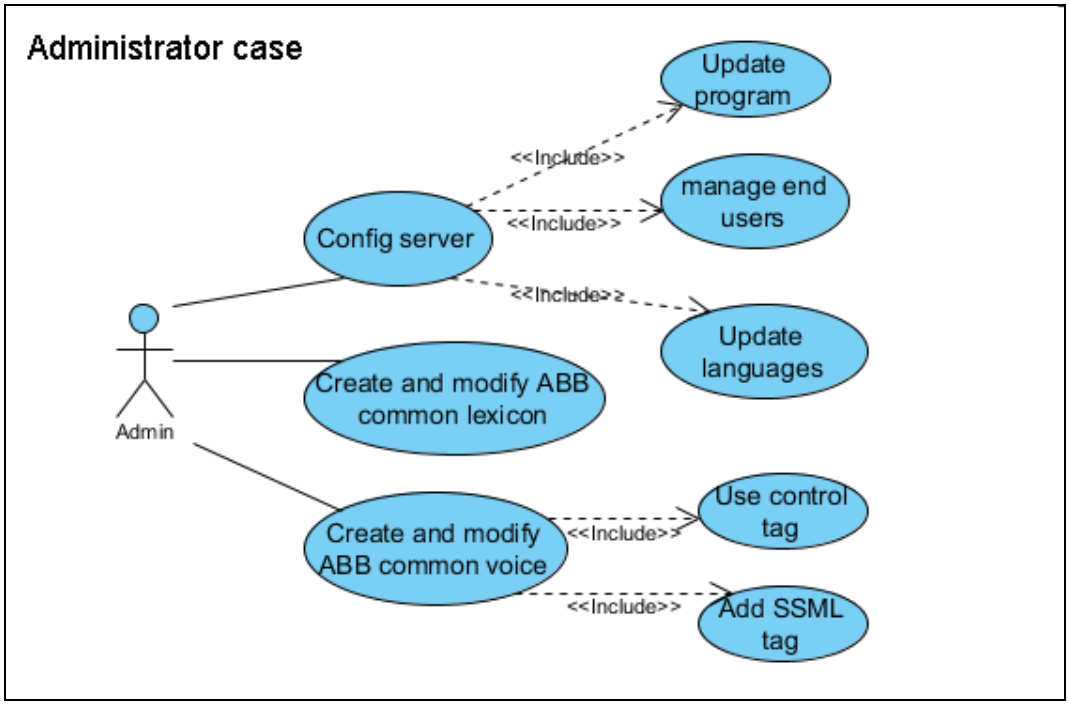


Figure 3.3.2.2 E-learning administrator use case

3.3.3 Requirements arrangement

The requirements were arranged in five groups: user interface, voice quality, functional feature, operational feature and support resource. These features would be evaluated by generating the sample Text–to–Speech system outputs.

I. Voice quality

REQ.	DESCRIPTION	PRIO
1.1	Overall voice quality (Long text)	1
1.2	Pronunciation of technical terms	1
1.3	Pronunciation of product names and unit	1
1.4	Pronunciation of calculation	1
1.5	Pronunciation of formulas	1
1.6	Non-English languages available (Chinese, German, Italian, Spanish, French, Russian, Portuguese, Arabic)	1

II. User interface

REQ.	DESCRIPTION	PRIO
2.1	Ease of use	1
2.2	Flexibility of voice control (set pitch, timbre, pause in the speech)	1
2.3	Ease of Mark-up Language setting	2

III. Functional requirements

REQ.	DESCRIPTION	PRIO
3.1	User-definable lexicon	1
3.2	Language switching within the text	2
3.3	Integration with other software used in e-learning course development	2
3.4	Support for multiple document types	2
3.5	Switching among the multiple voices	3
3.6	Ability to create many audio file at once	3

3.7	Text spelling check	3

IV. Operational requirements

REQ.	DESCRIPTION	PRIO
4.1	Speed of program running	2
4.2	Usage outside corporate network	3
4.3	Usage on demand	3
4.4	Ease of installation	3
4.5	Server – Client architecture	1

V. Supporting resource

REQ.	DESCRIPTION	PRIO
5.1	Accessibility of technical support	1
5.2	Online resources available	1

4 Testing

This section was dedicated to select software tools and test them in English. The testing was planned in two parts: (1) general testing and evaluation, and (2) intensive testing. This chapter went through the candidate Text–to–Speech software tools' selection and elimination, the description of the test environment, and then proceeds with each of the tools test results. At the end of this section, the test results are evaluated.

4.1 Candidate Selection and Elimination

The list of candidates for a Text–to–Speech software tool has been made from web search results and ABB e-learning developers' recommendations. The initial list of candidate software tools had 11 entries. During the general evaluation, four candidate software tools were selected for intensive testing.

Overall, the following eleven Text–to–Speech tools were evaluated.

1. Acapela Virtual Speaker
2. Neo Speech
3. Verbose
4. TextAloud
5. Loquendo TTS
6. Natural Reader
7. IVONA Reader
8. Alive Text to Speech
9. Nuauce Dragon Naturally Speaking 10.0
10. AT&T Natural Voices Desktop
11. ReadPlease

Four candidate software tools were selected for intensive testing:

- Loquendo TTS

- NeoSpeech Voice Text
- Acapela Virtual Speaker
- IVONA Reader

Although evaluating the financial impact was not as straight forward as it might appear in this research, during the search for Text–to–Speech software tools, there were a number of open-source tools competing with the commercial tools on the market. Unfortunately, most of the open source tools were capable of satisfying only a part of requirements. They would implement some module of Text–to–Speech software, for example text–to–speech conversion, multiple languages switching, multiple text formats, but rarely more than that. Sufficiently powerful open source Text–to–Speech software tools weren't found in the research.

Due to the limited descriptive information and no trial version available on some of the commercial products' home pages, the author had to contact the sales personnel to ask for trial versions. Owing to licensing issues and costs, AT&T didn't offer a trial version.

Bases on the two essential requirements, voice quality and ease of user interface, the remaining four commercial products were selected for intensive testing. The general evaluation of Text–to–Speech software tools were listed in the summary table (Appendix 1, General evaluation).

4.2 Intensive test Environment

The test plan was to go through the voice features that were crucial for company requirements. In order to fairly compare the voice quality, each TTS software tool was set to the same voice parameters for output of the audio file. The specific e-learning course modules were selected for testing as text samples.

System environment and output audio parameter setting:

Operating system	Voice language	Audio format	Mark up language
Win XP Pro 32-bit	US English, male	16 kHz	SSML 1.0

4.3 Test sample selection

The TTS software tools were tested by generating the output of the specific e-learning course modules. As mentioned in the requirements gathering, the voice features were the most significant to evaluate in the candidate software tools. The text modules are based on the real ABB e-learning courses content which includes complex sentences, product names, formulas, technical words, etc.

With company requirements for voice features, the testing text samples were arranged in six groups, as following.

Long text reading	*Motors with converters for VSD, slide 8*
Pronunciation of calculation	*Energy appraisal - The marketing kit, slide 26, Machines example*
Pronunciation of technical terms	*ACS850-04 product specification, slide 23, One slot for communication*

	options
Pronunciation of formulas	*AC drive basics - Process control and various control methods, Torque, slide 20*
Pronunciation of product name and unit	*ACS850-04 product specification, slide 14, Operating conditions*
Language switching among the text	*G964e Advanced ATEX*

4.4 Analysis of testing results

For the purpose of producing a successful comparison of candidate TTS software tools, the author evaluated the candidate tools in two main areas: the voice quality and non – voice features (which contained user interface, functional features, and operational features). The voice quality comparison was achieved by a TTS output evaluation survey, and the non-voice features were appraised in the process of generating the testing samples into synthetic outputs.

4.4.1 Voice quality (TTS output) evaluation

4.4.1.1 Evaluation method

The voice quality evaluation might be difficult because of subjective speaking behavior. With regards to this, the author created a multimedia survey avoiding a personal subjective analysis of voice quality. The survey was sent to ABB e-learning developers, including the native speakers (Appendix 4, TTS output evaluation survey link).

- The question was designed to present the each feature of voice quality with the company requirements.

- A set of four synthetic output audios were embedded randomly in each question.
- The names of the TTS software tool producing each output were invisible.
- Four outputs in each question were compared by the listener
- The listeners selected the best output in each question

4.4.1.2 Survey results

The survey started on 21st May and ended on 28th May 2010, and 35 responses were collected from different countries (Figure 4.4.1.1). The average rate of each Text–to–Speech tool was selected in the five evaluation questions. The survey results showed NeoSpeech was selected mostly, three times more than any other candidate software tools. Acapela (Virtual Speaker) was behind IVONA. (Figure 4.4.1.2)

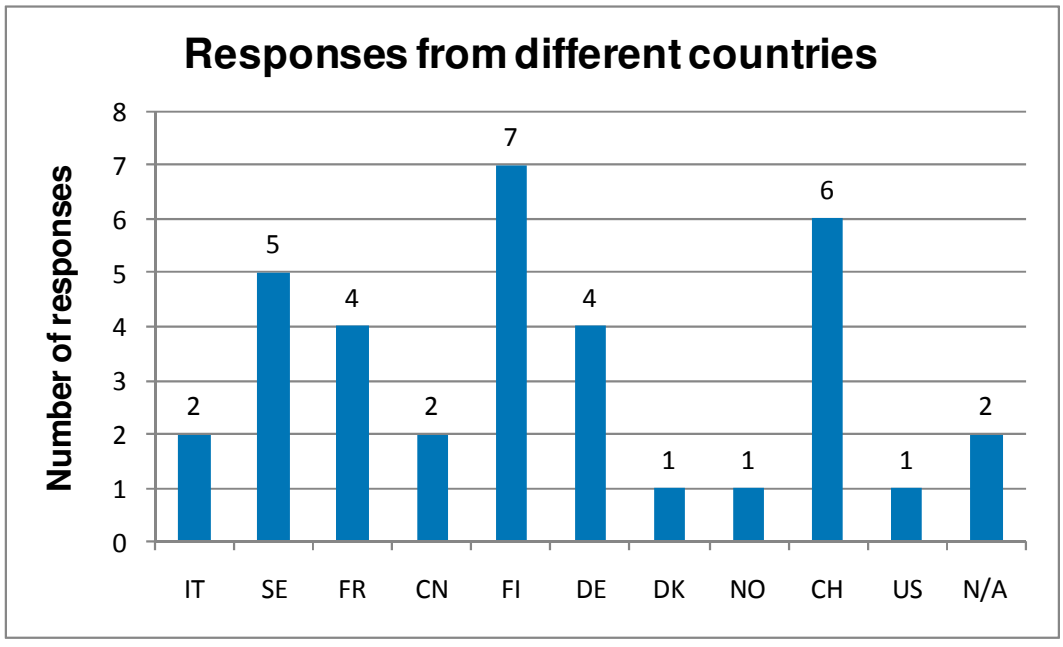


Figure 4.4.1.1 Responses from different countries

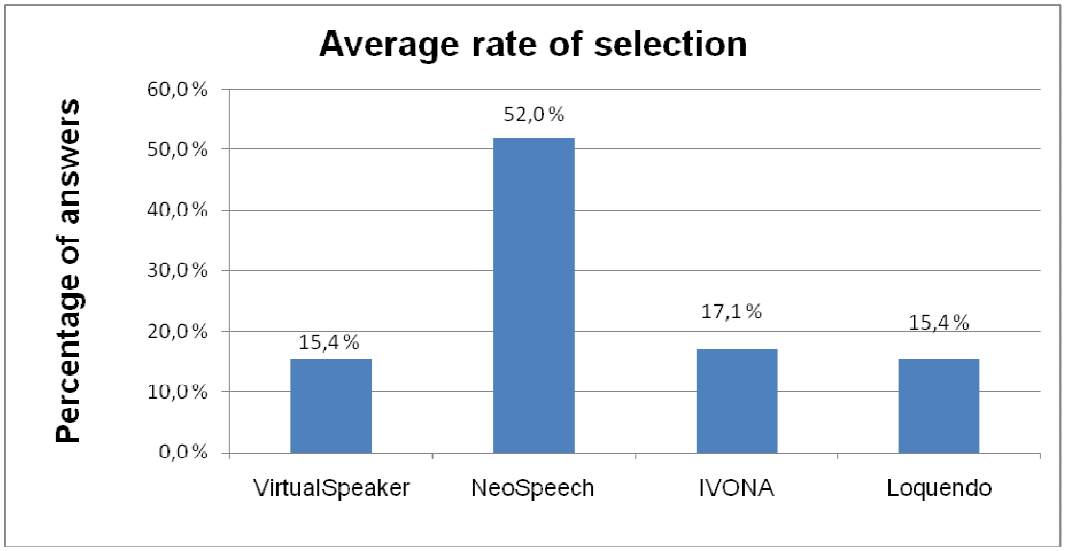


Figure 4.4.1.2 Average rate of tools selection

Viewing the figures (Figure 4.4.1.3) with each voice features, Neo Speech was seen as the most acceptable according to the highest performance in each features. IVONA outputs were a bit more popular than Loquendo and Acapela Virtual Speaker in “Long sentence text”, “Calculation sentences”, and “Technical terms”. Acapela performed lowest in terms of “Product names/Unit” and “Technical terms”, but beat Loquendo in “Long sentence text” and “Formulas”. Loquendo only beat Acapela and IVONA in “Product names/Unit”, and it was merely acceptable in other features. This survey results were only directed towards voice quality. These should be considered together with non-voice quality features in the final results.

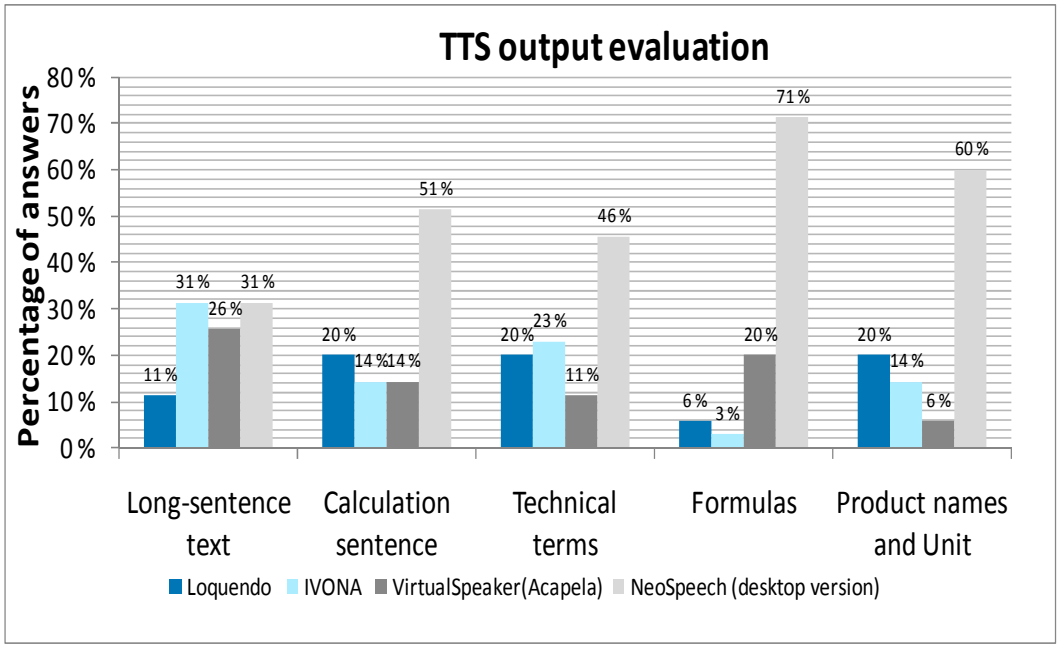


Figure 4.4.1.3 Evaluation of TTS voice quality

4.4.2 Non - voice features evaluation

The conclusions of the testing of TTS software features other than voice quality were presented in form of a table with assigned scores (Table 1.). The non-voice quality features were listed in priority throughout the table.

Table1. The score table of test result

Prio		Acapela	Neo Speech	IVONA	Loquendo
1	User Interface	4	4	4	3
1	User lexicon	3	4	2	3
1	Voice control ability	5	4	3	4
1	Accessibility of technical support	4	4	4	4
1	Online resource available	3	4	4	5

1	Available for non-English languages	4	3	2	5
2	Speech synthesis mark-up language	4	4	3	4
2	Language switching within the text	5	4	2	4
2	Support for multiple document type	4	4	4	4
2	Speed and stability of program running	4	4	4	3
2	Integration with other e-learning software	0	3	2	0
3	Dynamic switching between multiple voices	5	4	2	4
3	Usage on demand	4	4	0	3
3	Use outside of company network	4	4	3	3
3	Ability to create many audio files at once	3	3	3	4
3	Ease of installation	3	4	4	3
	Overall	3,7	3,8	2,9	3,5

From the view of overall, Neo Speech scored highest among the candidate tools, having great and stable performance along with the feature requirements. Acapela beat Loquendo overall and in six categories. IVONA came in last with an overall score under 3. In features prioritized on first class, Loquendo beat both Neo Speech and Acapela by a tiny margin. IVONA still scored the lowest, particularly with low quality in “User lexicon” and “Available for non-English languages”. The following analysis concentrated on the performance differences in each tool among the features.

Table of features prioritized 1

Prio		Acapela	Neo Speech	IVONA	Loquendo
1	User Interface	4	4	4	3
1	User lexicon	3	4	2	3
1	Voice control ability	5	4	3	4
1	Accessibility of technical support	4	4	4	4

1	Online resource available	3	4	4	5
1	Available for non-English languages	4	3	2	5
	Overall	3,8	3,8	3,2	4

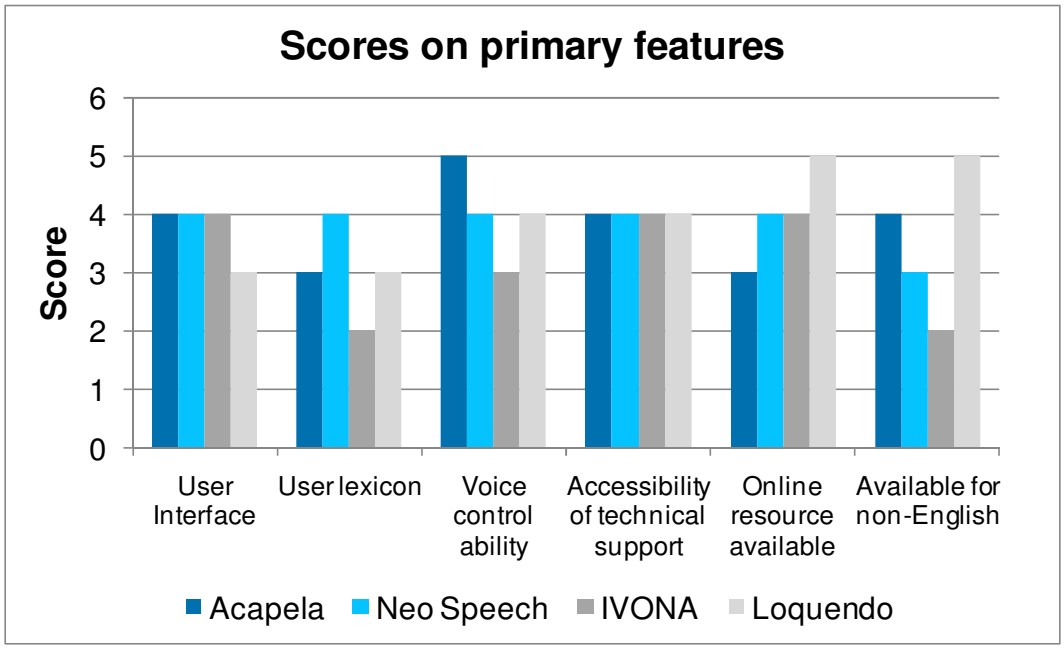


Figure 4.4.2.1 Evaluation of non-voice features with first priority

- User Interface and Voice control ability

In general, "User interface" in each of the four candidate software tools was friendly. According to the e-learning developers' feedback, Loquendo was deemed not easy to use which meant the user interface may not be friendly enough.

Acapela got the highest score in "Voice control ability" because unlike the common voice control functions in many TTS software tools. Acapela allows the user to customize the control tag instead of typing SSML to enhance the text read

aloud (pauses, sounds, speed…). It gives users options to adjust the vocal effects much more flexibly than the other candidate TTS software tools.

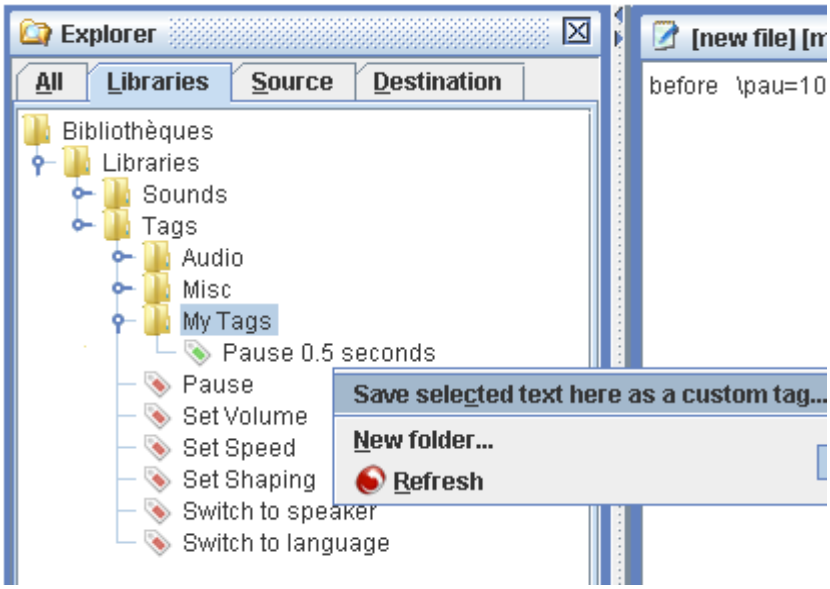


Figure 4.4.2.2 Custom control tags.

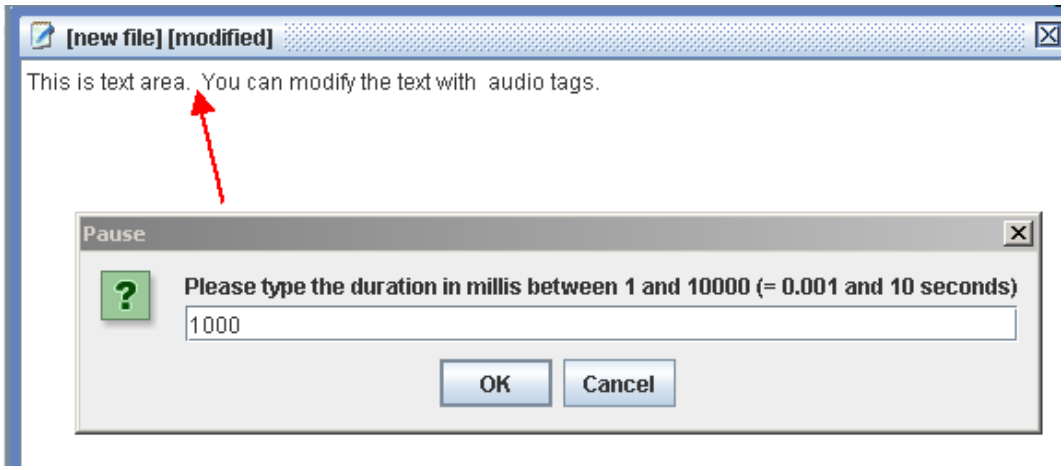


Figure 4.4.2.3 Usage of control tag

- User lexicon

A user lexicon was implemented in each of the four candidate software tools. In Acapela, the lexicon was presented in 'txt' format, which might be difficult for typical users to edit, so it got a 4. Loquendo user lexicon was not stable according to the e-learning developers' feedback. IVONA got 2 because it failed the requirement of the ABB Text–to–Speech system that it must be constructed with client-server architecture in order to implement common functions such as the ABB standard lexicon.

- Accessibility of technical support and Online resource availability

Since the Loquendo server version had been used in ABB for more than a year, in addition to the common support documents, Loquendo had already provided on online forum service regarding TTS solutions for ABB e-learning developers. Therefore, it was evaluated as the best of the candidate software tools. The results of other three candidate software tools should be treated more tolerant.

- Available for non-English languages

Loquendo got 5 because of the variety of languages. It covered English, French, German, Italian, Portuguese, Russian, Spanish, Arabic, Danish, Dutch, Swedish, Finnish, Mandarin Chinese, Greek, Galician, Valencian, and Polish. Compared to Loquendo, Acapela offered multiple languages as well, except for Chinese, which would be one of the main languages in e-learning courses. NeoSpeech was at a relative disadvantage in this aspect since it only offered in English, Spanish, Chinese, Korean, Japanese, and French (under development). IVONA only provided English in languages need so that it should be eliminated in this aspect.

5 Results

5.1 Reporting to commissioner

The results of the project have been reviewed as the project progressed. The research results were presented as a presentation to ABB. The presentation contained a walk-through of the factors that brought research on Text–to–Speech software tools for the ABB e-learning program. The comparison results were summarized in the table below.

Tool	Strength	Weakness
Acapela	• Cover most kinds of languages • Good voice quality • Flexibility of voice control (eg. Custom voice control tag) • Desktop version & Server version	• Chinese language is not available
NeoSpeech	• Very high quality natural voice • Integration with Adobe Captive 4.0 • Desktop version & Server version • Ease to use	• Only English, Chinese, Spanish, Korean, and Japanese available (French is under development)
Loquendo	• Cover most kinds of languages	• User interface is not

	• Great support resources • Desktop version & server version	friendly • Server was not stable
IVONA	• Good voice quality • Integration with Skype, MS word • Quick-response technical support	• Only English, Polish, Romanian available • Preferred to Web-based use • Weak user lexicon • Server version not unavailable

Table 5.1.1 Summary of tools comparison

5.2 Summary and Recommendation

- **NeoSpeech** had great testing performance in all areas, especially in voice quality. Besides English, NeoSpeech provides high quality Chinese language voice as well. The friendly user interface made for efficient and good quality work.

- **Acapela** was outstanding in meeting the multiple languages requirements. It covered all the non-English languages except Chinese. The flexible voice control ability and voice quality were more competent than the other candidate software tools.

- **Loquendo** was generally good in voice quality and variety of languages. It has been used in ABB for one year, and the good connection between ABB and Loquendo may be helpful for price negotiation.

- **IVONA** should be eliminated because of the low-level of comprehensive features. It was more suitable for personal text–to–speech purposes and for Web plug applications.

Consequently, NeoSpeech and Acapela can be chosen to fulfill the courses' language needs. NeoSpeech might be considered as the main tool to create the courses in English and Chinese. Acapela is good choice to be used as an additional tool to create courses in other non-English languages except Chinese. In fact that some e-learning developers were used to creating synthetic audio with Loquendo, it is better to keep it as an additional TTS tool for a period of time while the developers are learning to use the new TTS tool(s).

Although this research evaluated most of the TTS products on the market, the synthesis technology is growing fast, and Text–to–Speech software tools are being upgrade day by day, so the following questions can be studied and tracked in the future.

1. The integration of other e-learning course development tools such as Articulate.

2. The solution for Text–to–Speech software functioning in a VPN network environment. Or how to configure the server in order to make the end user function in VPN network environment?

3. Use of the Synthesis Speech Markup languages.

4. The voice quality and language availability of the main TTS products.

5.3 Outcome

The research results were accepted. ABB is going to negotiate with the candidate software company to decide on the final selection of Text–to–Speech software tool(s). The author may continue to participate in the process of final selection, new software tool(s) implementation and possibly gather additional information.

References

[1] K.R Aida – Zade, C. Ardil and A.M. Sharifova, *The main principles of Text–to–Speech Synthesis System*, International Journal of Signal Processing 6,1 2010

[2] Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, *Progress in Speech Synthesis*. Springer: 1997. ISBN 0-387-94701-9

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

[4] "*How to configure and use Text-to-Speech in Windows XP and in Windows Vista*", Support.microsoft.com. 2007-05-07. Retrieved 2010-02-17

[5] Richard Sproat, Paul Taylor, Michael Tanenblatt, Amy Isard, *A markup language for Text–to–Speech Synthesis*, Bell Laboratories, Lucent Technologies, Centre for Speech Technology Research, University of Edinburg

[6] W3C, World Wide Web Consortium, http://www.w3.org/

[7] Speech Synthesis Markup Language Specification, version 1.0
http://www.w3.org/TR/speech-synthesis

[8] http://www.bell-labs.com/projects/tts

Appendices

Appendix 1 Text–to–Speech tools general evaluation summary

Appendix 2 Requirements gathering survey results (regarding non – voice features)

Appendix 3 Link to the requirements gathering survey:

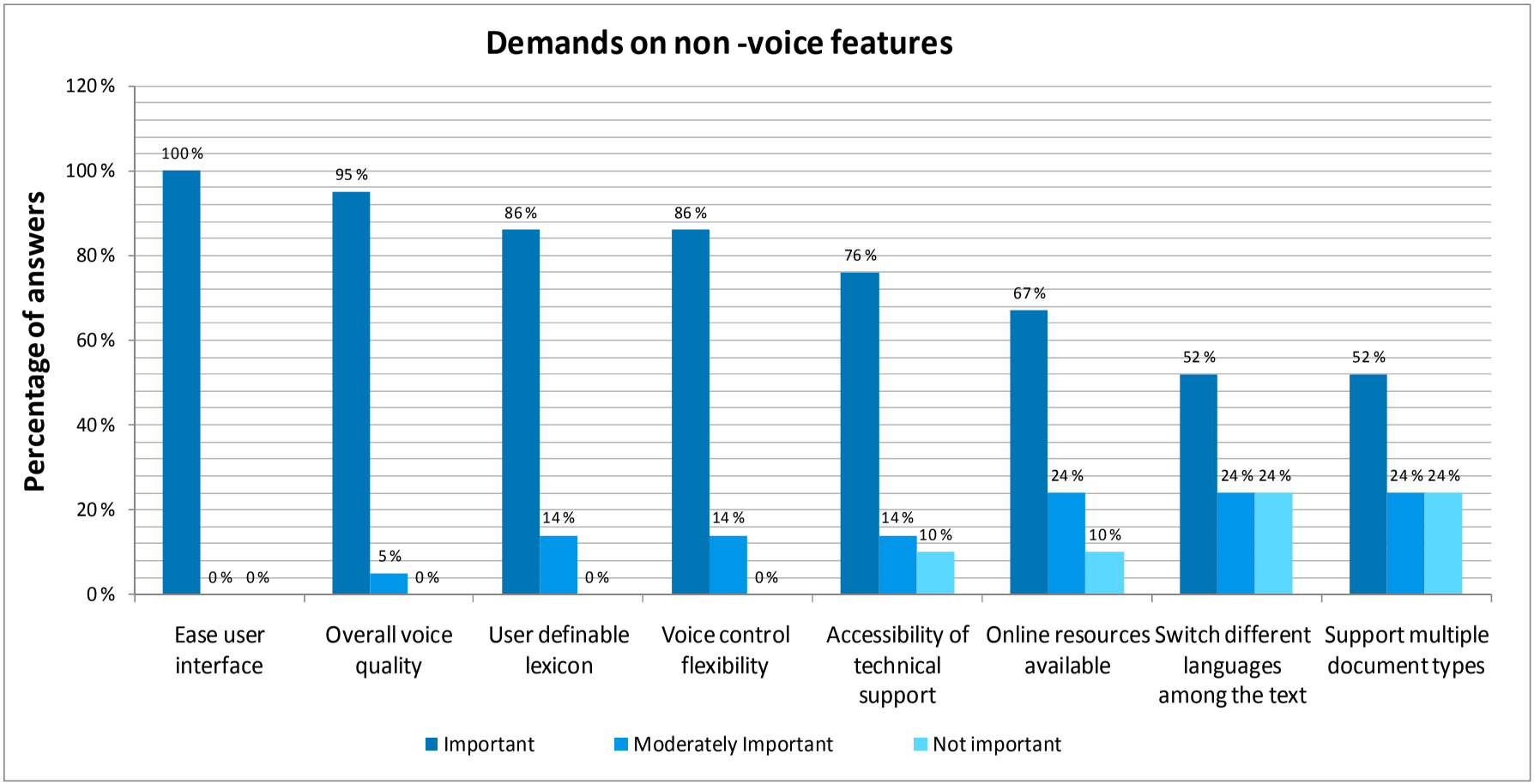
http://www300.abb.com/GLOBAL/GAD/GAD01366.NSF/viewUNID/0F3FB7253CBF7AFAC12576C800432408?OpenDocument

Appendix 4 Link to the TTS output evaluation survey:
http://www.surveygizmo.com/s3/299507/TTSoutputSurvey

Software	Potential	Comments	Communicating	License Price
Acapela VitualSpeaker	Y	• Multiple languages available • Easy language switching within the text • Several audio output formats (8 kHz, 11 kHz, 16 kHz, 22 kHz, 44 kHz, PCM, mp3, vox, A-law) • Customize speech control tag • User lexicon	Evaluation version (full voice)	N/A
NeoSpeech VoiceText TTS	Y	• High quality natural voice • Multiple languages available (English, Korean, Japanese, Chinese, Spanish) • Expressive control • Flexible audio output formats(8 kHz, 11 kHz, 16 kHz, PCM, Mu-law, A-law) • SAPI 5.1 supported • Desktop version & server version available • Ease of user interface • Integration with Adobe Captive 4.0 • User lexicon	Trail Version (English voice)	$ 1250 for one voice (including the desktop program)
Verbose	N	• Lack of functions • User lexicon unavailable		$ 19.99

		• Web use preferred		
TextAloud	N	• Basic functions • Natural voice and language work in collaboration with AT&T • User lexicon unavailable		$ 29.95,additional fee with AT&T voice
Loquendo TTS	Y	• Multiple languages • User lexicon • Server version available • Not ease of user interface • Ease of support resource •		N/A
NaturalReader (Free version)	N	• Multiple languages • Natural voice unavailable • Basic functions	N/A	Free for Basic Edition Additional charge for advance version
IVONA Reader	Y	• Languages (English, Polish, Romanian) • Ease of user interface • User lexicon • Good voice quality • Integration with Skype, MS word		Purchase voice

		• Quick-response Technical support		
Aalive Text to Speech	N	• Microsoft TTS engine • Simple user interface • Lack of voice output formats		N/A
Dragon Naturally Speaking 10.0	N	• Speech recognition	Responded four months later	N/A
AT&T NaturalVoice Desktop	N	• No trial version available	N/A	N/A
ReadPlease	N	• Web-based application	Trial version	N/A


Demands on non -voice features
Percentage of answers
120 %
100 %
80 %
60 %
40 %
20 %
0 %
100 %
0 %
0 %
95 %
5 %
0 %
86 %
14 %
0 %
86 %
14 %
0 %
76 %
14 %
10 %
67 %
24 %
10 %
52 %
24 %
24 %
52 %
24 %
24 %
Ease user interface
Overall voice quality
User definable lexicon
Voice control flexibility
Accessibility of technical support
Online resources available
Switch different languages among the text
Support multiple document types
Important
Moderately Important
Not important

APPENDIX 3