THE WORLD OF SYNTHETIC VOICE CREATION AT A GLANCE

FEBRUARY 24, 2021

MADE BY: **SOVAN SINHA ROY**

TABLE OF CONTENTS

1. INTRODUCTION

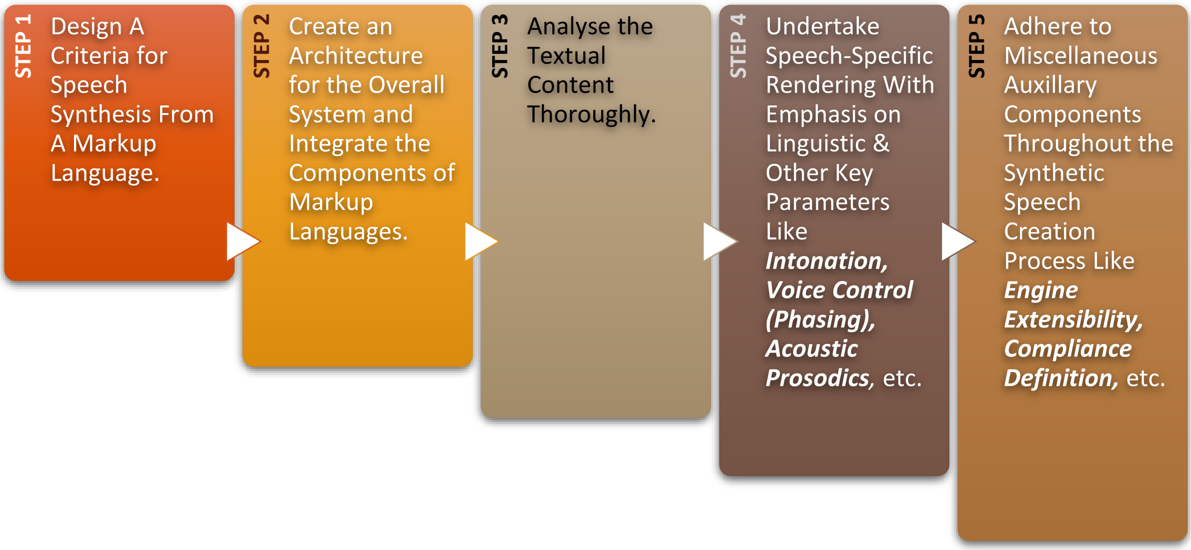
This project report titled **"The World of Synthetic Voice Creation at A Glance"** is based on the theme *Synthetic* or *Artificial Voice* that revolves around the principle of *"Speech Synthesis"* which is used to produce an artificial version of human speech. The output can be a real voice or a simulated voice that is capable of rendering human-sounding speech developed via deep learning mechanisms. We will try to elaborate more on this modern technology in this report by initially talking about its significance and core requirements, then showcasing the available technologies and existing products in this domain. Thereafter, we will highlight the future needs and technologies required to drive the future. Finally, there will be an insightful inference to sum it all up.

2. PROJECT FUNDAMENTALS

2.1 Brief Significance of the Synthetic Voice Creation

The prime significance of this technology at present is for providing voice assistance and facilitating prompt responses to voice commands (two-way interactions with the humans). Currently, the text-to-speech (TXT) synthesis is what is used widely for various purposes in this context but since it relies heavily on labelled data and human expertise, and also requires statistical models & specific rules, it is only limited to generic products at the moment with little to no tailoring to any application realm. However, it is a rapidly evolving domain of IT that is improving leaps & bounds steadily and promises to have a very bright future.

2.2 Core Requirements



STEP 1 Design A Criteria for Speech Synthesis From A Markup Language.

STEP 2 Create an Architecture for the Overall System and Integrate the Components of Markup Languages.

STEP 3 Analyse the Textual Content Thoroughly.

STEP 4 Undertake Speech-Specific Rendering With Emphasis on Linguistic & Other Key Parameters Like ***Intonation, Voice Control (Phasing), Acoustic Prosodics***, etc.

STEP 5 Adhere to Miscellaneous Auxillary Components Throughout the Synthetic Speech Creation Process Like ***Engine Extensibility, Compliance Definition***, etc.

The core requirements for this entire process can be subdivided into five broad categories: *Design Criteria, Architecture & Integration, Text Content, Speech-Specific Rendering,* and *Miscellaneous Auxiliary Components*.

2.2.1 Design Criteria

The markup language for speech synthesis can be developed within the following broad design criteria. They are ordered from highest to lowest in priority. In the event that two goals conflict, the higher priority goal takes precedence. Specific technical requirements are addressed in the following sections.

1. The markup language for speech synthesis will enable consistent control of voice output by speech synthesizers for use in voice browsing and also other contexts. Consistent rendering of speech synthesis markup is possible across multiple platforms and multiple speech synthesis implementations.
2. The markup language for speech synthesis can be an XML Application and shall be interoperable with relevant W3C specifications (competent international consortium).
3. The markup language for speech synthesis should be appropriate for speech output from a wide range of computer applications with varying speech content.
4. The markup language for speech synthesis should be internationalized to enable speech output of a large number of languages.
5. It should be easy to automatically generate, author by hand, and process documents using the markup language for speech synthesis.
6. All features of the markup language for speech synthesis should be implementable with existing, generally available technology. Anticipated capabilities should be considered to ensure future extensibility (but are not required to be covered in the specification).
7. The number of optional features in the markup language for speech synthesis should be kept to an absolute minimum, ideally zero. For optional features, a reasonable rendering behavior should be available when not implemented fully by a speech synthesizer.
8. Documents written in the markup language for speech synthesis should be human-legible and reasonably clear and the specification should avoid unnecessary terseness.
9. The element set should avoid unnecessary differences with HTML, XHTML, ACSS, and other relevant specifications.
10. The markup language for speech synthesis specification should be prepared quickly, where appropriate derivation from existing and applied specifications is done (code reengineering).

2.2.2 Architecture and Integration

Speech Generation from Markup Languages

It must be practical to generate speech synthesis output from a wide range of existing document representations. Most importantly speech output from HTML, HTML plus ACSS/CSS, XHTML, XML plus XSL, and DOM must be possible.

Speech Generation from Other Applications

It must be practical for a wide range of applications to automatically generate speech synthesis output. Key examples include ***voice browsers, email readers, web browsers, and accessibility applications***.

Integration with Other Voice Markup

The speech synthesis markup must be interoperable with other relevant specifications like one of the most common ones available that are developed by the *W3C Voice Browser Working Group*. It must be possible to embed speech synthesis markup into the dialog markup for prompt generation and other spoken output. It must be possible to utilize pronunciations defined in a standard pronunciation format. Moreover, it must also be possible to utilize speech synthesis markup for universal access.

Mono & Multi-Modal Outputs

The speech synthesis markup must be appropriate in the context of an audio-output-only (mono-modal) user interaction. On top of that, the speech synthesis markup must also be appropriate in the context of a multi-modal system output, most importantly, in combination with visual output. Where appropriate, synchronization of speech and other output should be supported with SMIL or another relevant and related standard.

2.2.3 Text Content

Document Structure

The speech synthesis markup must support the ability to indicate document structure in a way that is instructive to a speech synthesizer for rendering the document. The specification must provide a well-defined set of document structure elements. At a minimum, it must be possible to mark paragraph and sentence structures. Dialog types and other structural elements with distinctive spoken styles should also be considered in the specification process.

Mono & Multi-Lingual Documents

The speech synthesis markup must support the ability to incorporate and render the text of a single language in a single document and to mark the language content appropriately. Furthermore, the speech synthesis markup should also support the ability to incorporate and render the text of more than one language in a single document where those languages are supported by the speech synthesizer. The levels in document structure in which language change is permitted shall be determined during the specification process when the definition of the speech synthesis document structure emerges.

Phonemic Pronunciations

The speech synthesis markup must provide the ability to specify pronunciation entities as sequences of phonemes (smallest unit of sound). Phonetic pronunciation models may also be considered for this.

Reference to Externally Defined Pronunciations

The speech synthesis markup should support the ability to reference externally defined pronunciation or lexicon documents. In particular, if the Voice Browser Working Group defines a lexicon format it must be possible to reference it from the speech synthesis markup. [*In the absence of a Working Group proposal, there are no obvious candidates for standard externally referenceable lexicon formats*].

Out-of-Vocabulary Handling

A nice to have feature is perhaps the speech synthesis markup support which is a mechanism to request particular handling of out-of-vocabulary text or other unpronounceable text. (This may lead to an API design issue, but this is beyond the scope of this artifact).

Acoustic-Phonetic Sequences

The speech synthesis markup may provide a mechanism to exactly specify the desired acoustic-phonetic rendering of a given text segment. This may be accomplished with a sequence of high-level phonetic and phonemic symbols, accompanied by a piece of detailed acoustic information for rendering the phonetic and phonemic symbols such as *duration, pitch movement, intensity*, etc.

Special Text Constructs

The speech synthesis markup must provide the ability to mark a set of common text constructs that require special handling by speech synthesizers. A mechanism should also be provided to indicate locale or other information (when required) that enables a speech synthesizer to incorporate dates and other locale-sensitive constructs.

Spelling-Literal Output

The speech synthesis markup must provide the ability to mark regions of text for "spelled" or literal output, as appropriate to the text language.

Non-Speech Output

The speech synthesis markup must provide the ability to incorporate non-speech audio output. This may include references to audio files (e.g., wave and MIDI files) that are linked inline. This may also include the generation of a set of defined audio samples such as touch-tone or other commonly used prompt sounds.

2.2.4 Speech-Specific Rendering

Speaking Voice Control

The speech synthesis markup must provide the ability to indicate a speaking voice for a document or for regions of text within a document. A set of common speaking voice (or speech font) characteristics must be defined and may include gender, age, name, and instance selection (where multiple voices have common characteristics, e.g., two male voices).

Emphasis

The speech synthesis markup must provide the ability to mark words and other regions of text for spoken emphasis (also referred to as *prominence* or *stress*).

Intonation Control

The speech synthesis markup may provide the ability to mark words and other regions of text with intonational characteristics including **boundary tones** (rise or fall at sentence/phrase end) and **sentential intonation** (movements across phrases/sentences).

Acoustic Prosodics

The speech synthesis markup must provide the ability to mark regions of text with acoustic characteristics such as pitch, pitch range, speaking rate, and volume.

Synchronized Facial Animation

The speech synthesis markup may provide the ability to mark the text with features that enhance synchronized facial animation. Features may include positions of physical facial features (e.g., lip rounding, jaw position, eyebrow movements), timing data, and expressions (e.g., smile).

Spatial Audio

The speech synthesis markup may provide a mechanism for generating spatial audio (also known as 3D audio). For instance, this could be a request that the voice output be in the upper-right quadrant forward of the listener. It may also allow the voice location to shift over time.

2.2.5 Miscellaneous Auxiliary Components

Compliance Definition

The specification must address the issue of compliance by defining the sets of features that must be implemented for a system to be considered compliant with the specification. Where appropriate, compliance criteria may be defined with variants for different contexts or environments.

Event Generation

The speech synthesis markup may provide methods to mark points in text output or segments of text output to generate callbacks, an event notification, or other information that can be used to track the progress of text output, that determine timing and location of barge-in for an appropriate resume, that can be used to trigger other activities, or that can be used to synchronize speech output with other output modalities. The mechanisms by which event notifications are issues are actually outside the scope of the speech synthesis markup specification.

Pause/Resume Behavior

The speech synthesis markup specification may define the behavior of implementations with respect to pausing and resuming audio output. Beyond the typical instant stop/start model (a tape player paradigm) some consideration could be given to specifying word boundaries or other locations where pausing is reasonable for a listener. Similarly, the markup may enable a mechanism to indicate appropriate locations to resume output that may be different from the pause location.

Comments

The speech synthesis markup must support a mechanism for inline comments (Presumably the parent markup language, e.g., XML, will provide such a mechanism).

Engine Extensibility

The speech synthesis markup may need to define a mechanism by which specific speech synthesizer implementations can provide enhancements or non-standard extensions without affecting the core specification behavior.

3. CURRENT TECHNOLOGIES IN USE

There are many software tools & applications that are currently available to serve the purpose of TTS creation and/or conversion but before we elucidate on that, let us recapitulate how it all works. So, to qualify for inclusion in the TTS category, a product must:

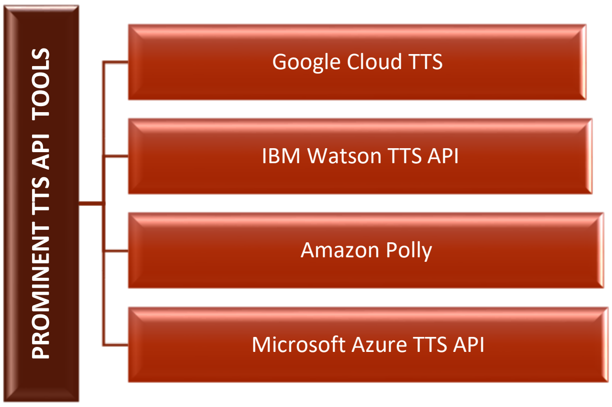
- Convert written text to natural-sounding speech.
- Integrate with applications and websites via a connector such as an API.
- Control aspects of the synthesized voice, such as volume, pitch, and emotion.

The prominent technologies in current use are namely ***Amazon Polly, IBM Watson (TTS API), Microsoft Azure TTS,*** and ***Google Cloud TTS***. Now, let us briefly discuss these API tools.

a) **Google Cloud TTS**: This software application or API from Google is probably the most renowned & sophisticated facility at the moment that facilitates developers to synthesize natural-sounding speech with 220+ voices, available in several languages and variants. It harnesses the amalgamation of **DeepMind's** groundbreaking research in **WaveNet** and Google's powerful neural networks to render high-fidelity audio.
b) **IBM Watson TTS API**: IBM Watson provides another very popular TTS API that harnesses its latest neural speech synthesizing techniques to convert written text to natural-sounding speech. It allows the creation of custom voices that can be used to create a uniquely branded voice adapted to a target speaker of one's own choosing.
c) **Amazon Polly**: Amazon Polly is another eminent service at our disposal today that is powered by the *AWS* (Amazon Web Services) department of Amazon. It is a rapidly evolving service that apart from the standard TTS services also offers a Neural Text-To-Speech (NTTS) service that can play the traditional important roles of a **Newscaster** and/or an **Interviewer** via its advanced machine learning algorithms.

d) **Microsoft Azure TTS**: This widely used service from Microsoft is highly efficient in rendering high-quality human-like synthesized speech in various languages and variants which can either consist of standard or neural voices. It leverages Speech SDK (Software Developer Kit) and/or REST API for speech synthesis. It also provides an asynchronous synthesis of long audio as per the requirements.

Moreover, it is noteworthy that there are many smart speakers and digital assistants like **Amazon Alexa, Apple's Siri, Google Assistant,** and **Microsoft Cortona** available today that are making voice technology mainstream, making more & more people familiarize with its use for undertaking various tasks & activities. Synthetic Voice underpins the functioning of these aforesaid smart devices as they operate under the overarching speech synthesis mechanism of their respective brands or creators.



4. FUTURE TECHNOLOGIES

We have already discussed about the TSS in detail earlier. An immensely beneficial future development in this regard can be ***'Within or In-Text Document Natural Language Generation'.*** The speech synthesis markup may someday provide a mechanism to allow on-the-fly generation/modification of output text. For example, based on dialog context or based on what a user said to a dialog system, the speech output may choose the appropriate verb/noun to echo the user's spoken words. Another example could be the use of style sheets to apply style rules to control how things like dates are transformed before being spoken.

Artificial voice creation or synthesis will become prevalent in the future for the discharge of various professional duties from *the ingestion of documents* (for details verification & validation) to *note-making for doctors and students* to help them jot down everything that they need, etc.

As **Amazon** has very rightly put forward recently, ***'Ambient or Ubiquitous Computing'*** is indeed the need of the future especially in the world of voice synthesis & recognition technology where we bridge the gap between the undertakings of the human & digital components in a business operation. Voice is becoming an interface of its own, moving beyond the smart devices to our homes and soon, it will promulgate to many other quotidian contexts and perhaps even more beyond our imagination like glamorous roles of *singing,* *acting*, *coaching/mentoring,* etc. So simply put, in the times ahead, with voice technology, businesses across the globe can make their teams more productive, support better customer experiences, and even strengthen the security of their operations.

SUMMARY

To wrap up in a nutshell, synthetic voice creation & recognition is the future that will assist mankind in various spheres or sectors of operations from the ingestion of documents to note-making that is clear, comprehensive, and absolutely flawless for the needs, wants, and desires of the end-users. We have seen the core components required for the TTS technology which is apparently the most prevalent synthetic voice technology on planet earth right now. We also discussed about the significance of it concisely in this artifact. Today, we have several TTS providers, the prominent ones include Google Cloud TTS, Amazon Polly, IBM Watson TTS API, and Azure TTS API, they have all been discussed briefly in this report. Furthermore, it is noteworthy that the most widely used digital assistants today like Google Assistant, Alexa, and Siri all operate under the overarching principles of this technology in discourse.

Hence, the future looks really rosy and with a few advancements like Ubiquitous Computing, In-text Natural Language Generation, etc we could be looking at something that will perhaps become a household name for many quotidian tasks & activities and even a few glitzy ones like singing & acting, etc. So, all in all, sky is the limit here and there are many possibilities possible!