Speech synthesis (also abbreviated as TTS, Text-to-Speech), unlike speech recognition, is not a technology that exploits the voice. It produces it. Synthetic voices are generally the final phase of the “voice assistant process” and are becoming increasingly popular, from youtubers and twitch streamers to what we support at Vivoka, enterprises complex projects.

Why are they getting that much interest would you say? Because they are a key component in the voice user experience (VUX), we’ll show you how.

What is speech synthesis? How does it work?

What uses does speech synthesis have?

How to choose and integrate speech synthesis?

Why is text-to-speech essential for voice?

What is speech synthesis? How does it work?

Speech synthesis (TTS) consists of the artificial production of human voices. The main use (and what induced its creation) is the ability to translate a text into spoken speech automatically.

How does it work?

Speech recognition systems use phonemes (the smallest units of sound) in the first place to cut out sentences. On the contrary, TTS is based on what are known as graphemes: the letters and groups of letters that transcribe a phoneme. This means that the basic resource is not the sound, but the text. This is usually done in two steps:

The first one will cut the text into sentences and words (our famous graphemes) and assign phonetic transcriptions, the pronunciation, to all these groups;

Once the different text/phonetic groups have been identified, the second step consists of converting these linguistic representations into sound. In other words, to read these indications to produce a voice that will read the information.

This is what voice synthesis can sound like: play the following soundcloud file to read out loud the following paragraph.

Sistem pengenalan ucapan menggunakan fonem (unit suara terkecil) di tempat pertama untuk memotong kalimat.

Sebaliknya, Text-to-Speech didasarkan pada apa yang dikenal sebagai grafem: huruf-huruf dan kelompok huruf yang mentranskripsikan fonem. Ini berarti bahwa sumber daya dasarnya bukanlah suara, tetapi teks. Ini biasanya dilakukan dalam dua langkah:

Yang pertama akan

1. Memotong teks menjadi kalimat dan kata-kata (grafem) dan menetapkan transkripsi fonetik, pengucapan, ke semua kelompok teks/fonetik.
2. Setelah kelompok teks/fonetik yang berbeda telah diidentifikasi,
3. Mengubah representasi linguistik ini menjadi suara. Dengan kata lain, membaca indikasi tersebut menghasilkan suara yang akan membacakan informasi tersebut.

Audio Player

00:00

00:00

Use Up/Down Arrow keys to increase or decrease volume.

[ Attention! Speech synthesis should not be confused with voice response systems, which are generally used in public transport for example. In the latter case, it is a database containing a large amount of voice information recorded by one or more operators. This information, which is very limited and contextual, will be read at key moments. For example a stop or a connection. This operation is therefore much simpler than a TTS which will try to really synthesize a voice for each text provided. This does not mean that TTS are not used in the transport sector! ]

Speech synthesis capabilities are limitless. For instance, we created a TTS voice with our Voice Development Kit to comment its demo video, take a look at it!

Video Player

00:00

04:35

What uses does speech synthesis have?

There is a multitude of applications where you can find speech synthesis. However, it is important to know that, originally, this technology was useful to help people with disabilities (particularly visually impaired) in their daily lives. For example, the very famous Stephen Hawking, because of his severe disability, used a speech synthesis system to communicate with the people around him (you can try it directly at this link).

Since then, many cases of use have been developed more or less close to the original virtue of TTS. For example, transport companies use this technology to transmit messages to passengers via voice, whether or not they are disabled. It is very easy today to find traces of TTS in our uses. You may also have noticed it in language translation engines. Speech synthesis technology in these allows to suggest the pronunciation of the translated information in order to complete the textual translation.

Another sector that integrates speech synthesis in embedded systems or cloud applications and keeps on revolutionizing uses is the broad field of IoT. Indeed, in a rapidly expanding universe, intelligent devices increasingly integrate TTS. On the one hand, it allows to improve the user experience. On the other hand, it improves accessibility and the intelligence of the interfaces. A strong example that continues to progress is the one of household appliances (AKA “appliances” in English), increasingly equipping consumer products and robots with voice.

How to choose and integrate speech synthesis?

In order to choose the right speech synthesis (text-to-speech), it is essential to take into account several criteria. These parameters are the following:

the language spoken;

the type of speaker;

the quality of the voice;

the supplier.

With this information, it is easier to select the right solution that meets your needs and constraints. Indeed, not all companies offering TTS have equivalent ranges. Thus, it is very important to source these partners well before you start. Next, the language and the type of voice are important criteria for the user experience proposed. Indeed, there must be consistency between the voice interface and what it should inspire.

On the integration side, speech synthesis technologies are also based on the notions of cloud, embedded or hybrid (also known as “on-premise”). You should keep in mind that embedded has technical limits in terms of sentence storage that a cloud will not have. However, while the cloud needs a connection, the embedded voice will work no matter what happens. Think about these parameters according to the nature of your projects. In transport for example, you should favour embedded speech synthesis to ensure a continuous service.

If you are looking for an embedded speech synthesis solution, we suggest you go to the Voice Development Kit page. This is our software development kit that gives you access to offline voice synthesis which is easy to configure and integrate.

Why is text-to-speech essential for voice?

Who today has never heard the voices of Siri, Alexa or the Google assistant? True ambassadors of “voice”, these assistants have all been directly equipped with voice synthesis in order to be able to respond to the user. This is not insignificant! It is precisely a question of strengthening the relationship between human and machine through a conversational… reciprocal link. The user talks to the assistant and the latter answers, as in a natural conversation between two or more humans. This component is more important than we imagine.

In fact, like any innovation, the adoption process is generally complex, especially when it brings a break in usage. The best way to gain acceptance for voice assistants was to offer new features that promote their use, but also to improve the user experience as much as possible by humanising the technology. These synthesised voices then made it possible to give an identity to the various assistants. They make it possible to differentiate them, but also to consider them as entities in their own right.

Beyond a simple functionality (the marketing course is about to begin), voices are now an integral part of the brand image! Some people even consider that the voice is a pillar of branding in the making. Indeed, it replaces images, which by nature are fixed (and over-represented in the media), with more engaging messages. An image is worth a thousand words, a voice is worth a thousand images?

The other interest for brands is to tell themselves that the pool of voice assistants is already large and that it tends to be larger. Then, isn’t it a good idea to go into voice with your own voice to reach such a large audience?