Fujitsu develops new speech synthesis technology

March 31 2014

Figure 1: Usage Scenarios for Speech Synthesis

Fujitsu Laboratories has announced development of speech synthesis technology that can create a variety of high-quality synthetic voices in a short period of time, and that can convey information in a tone that is appropriate for the usage scenario or usage environment.

Current speech synthesis technology, widely employed in society, is able to read-out a variety of texts, but in a monotone voice. For this reason there is a need for synthesized speech to be able to convey spoken text to listeners in accordance with the given circumstances, making it easy to understand.

Now Fujitsu Laboratories has developed speech synthesis technology in which the tone can be synthesized to fit the circumstances of the

situation in which it is being used. In addition, voices that have a high-level of clarity can be created in a short period of time – approximately 1/30th of the time required using previous technologies. As a result, speech can be conveyed according to the situation, such as using an alarming tone in emergencies, for example, or, when in a noisy environment, conveying information in a voice that resonates clearly. Moreover, voices that match a customer's expected preference, such as voices that are perceived to be endearing, or distinctive voices for particular characters, can be quickly created with high quality, making it easy to use synthesized speech in a wide variety of scenarios.

Background

Speech synthesis technology, which reads text out loud, is used in a wide range of situations, such as to broadcast constantly changing traffic conditions, for disaster prevention announcements by municipal authorities to convey information to local residents, and in art galleries and museums for audio guides. Speech synthesis is also used in automated interactive voice response systems that use automated voice response to guide phone enquiries, and in voice applications built into car navigation systems and other devices. In addition, in factories and other types of work sites, where it is an advantage for workers to keep their hands free so as not to disrupt their work, audio announcements are beginning to be used to convey messages (Figure 1).

Up until now, speech synthesis technology has only been able to read text out loud, but as the number of usage scenarios for speech synthesis technology increase, users have expressed a need for a variety of voices and tones that can be easily understood.

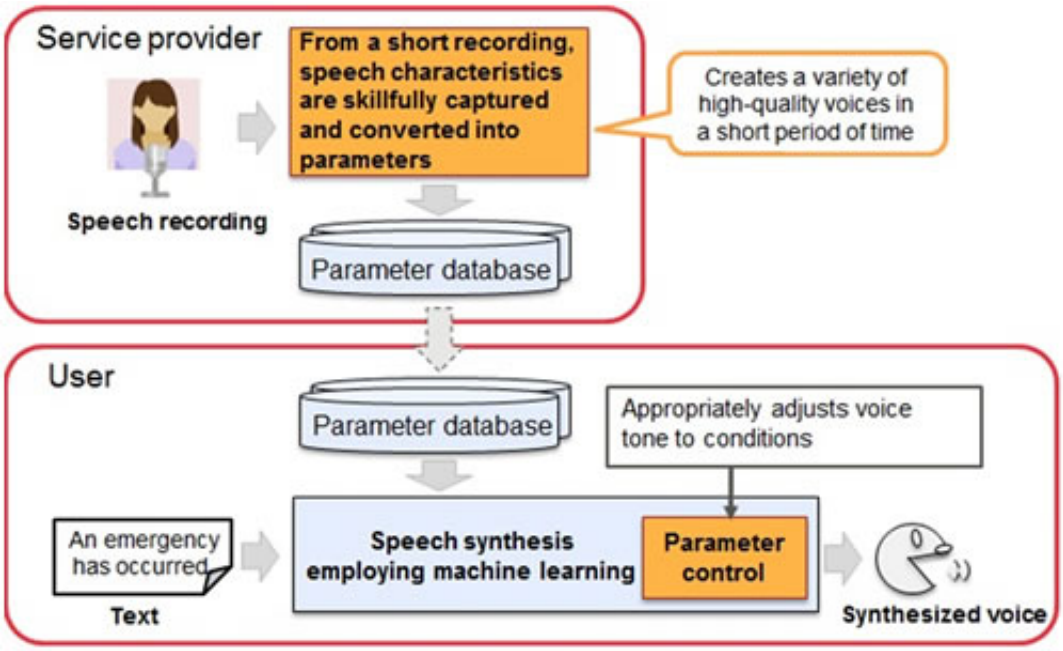

Figure 2: Overview of the newly developed speech synthesis

Issues

With conventional speech synthesis technology, it is possible to make simple adjustments to speed and pitch or other aspects of the voice, but it was difficult to synthesize the voice or tone to fit the usage scenario or usage environment. This results in the problem of not being able to sufficiently convey the desired information or fit with the image of the company or service being offered. In addition, in using high sound quality speech synthesis for voice services, there has been a need to be able to create new voices that match the service, but the problem has been that this could not be done quickly.

About the New Technology

Now Fujitsu Laboratories has developed speech synthesis technology in

which the tone can be synthesized to fit the situation, and with which a variety of high-quality voices can be quickly created. The conventional synthesis methods used thus far involved connecting large volumes of prerecorded speech waveforms. To make the synthesis process more flexible, Fujitsu Laboratories used a method to synthesize speech in which multiple characteristics, such as voice quality, intonation, and pauses, are skillfully captured and converted into parameters.

The new technology has the following features (Figure 2).

1. Synthesis of voice tones appropriate to the circumstances

Synthesizing speech with tones precisely appropriate for the situation was achieved by reflecting to parameters the difference between normal and distinctive voice tones, such as voices warning of danger or speaking slowly and clearly. These are not simply uniform adjustments for speed, pitch, and vibrancy of the voice, so information is able to be conveyed using very realistic expression. As a result, the voice can change to an alarming tone in an emergency, or to a tone that is clear and easy to understand, according to the ambient level of noise.

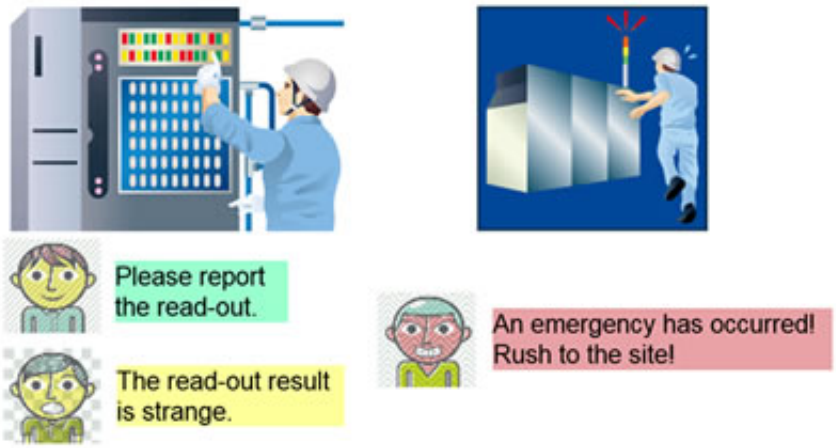

Figure 3: Usage examples of the newly developed speech synthesis

2. Creation of a variety of high-quality voices in a short period of time

Using technology that converts parameters and an algorithm that employs machine learning, voice characteristics are able to be efficiently extracted, so that only a small amount of speech needs to be recorded. This enables a variety of high-quality voices to be created in approximately 1/30th of the time previously required (in an internal comparison). As a result, new speech synthesis voices can be custom-made and delivered in a short period of time.

Results

Using this new technology in, for example, a system that conveys audio messages on the status of a system that is in operation to workers in a factory, routine messages could be conveyed in an ordinary tone of voice, error messages could be conveyed in a concerned tone, and emergency message could be conveyed in a very alarming tone (Figure 3). In addition, depending on the level of ambient noise, the voice can be changed to one that is clear and easy to understand, so that, even in very noisy environments, the announcement from the speakers can be clearly heard. As a result, it can be used by municipal authorities for announcements in disaster prevention situations, for which demand has been rising in recent years, and for other applications. For that application, routine local notices could be broadcast in a calm voice, whereas, in the event of a disaster, in accordance with the severity of the circumstances, information could be broadcast in an alarming tone.

In addition, for all types of voice services, voices that perfectly match the customer's preference, such as voices that are perceived to be endearing, or distinctive voices for particular characters, can be used. Moreover, taking advantage of this speech synthesis technology that enables it to reflect the characteristics of individual voices with just a small amount of recorded speech, it could also be used in medical and welfare applications. For example, if someone was about to lose his or her voice because of an illness, by recording that person's voice in advance, that person would be able to converse anytime using his or her own synthesized voice.

Provided by Fujitsu