Interspeech 2018
2-6 September 2018, Hyderabad

The IBM Virtual Voice Creator

Alexander Sorin[1], *Slava Shechtman*[1], *Zvi Kons*[1], *Ron Hoory*[1], *Shay Ben-David*[1], *Joe Pavitt*[2], *Shai Rozenberg*[1], *Carmel Rabinovitz*[1], *Tal Drory*[1]

[1]IBM Research – Haifa, Israel
[2]IBM Research – Hursley, UK
sorin@il.ibm.com, slava@il.ibm.com

Abstract

The IBM Virtual Voice Creator (IVVC) is an end-to-end cloud-based solution for TTS voice customization and voiceover generation in games and animated movies. The solution is based on the IBM expressive TTS technology with built-in online voice transformation capabilities. It is endowed with an interactive web GUI studio.

IVVC lets the users create unique voice personas according to their needs and imagination, and control the vocal performance of the virtual speakers. IVVC provides a powerful set of controls over the voice characteristics, including the vocal tract, glottal pulse, breathiness, pitch, rate, and special voice effects. IVVC also allows the user to control emotional style and emphasis in the synthesized speech.

The virtual voice design and performance controls are interactive, intuitive, fast and do not require any special skills.

Index Terms: speech synthesis, TTS, voice transformation, glottal vocoder, expressive prosody, voiceover

1. Introduction

Voiceover for game and animated movie characters remains the major dependency on human actors and physical recordings in the whole production cycle. The legacy voiceover process is lengthy, expensive and inflexible [1].

Modern Text-to-Speech Synthesis (TTS) technology is an attractive alternative to the legacy voiceover, significantly reducing or completely removing the dependency on human voice actors and recording studios. A fundamental barrier that limits the use of TTS for the voiceover production is that a TTS system can speak in a limited number of voices prepared in advance by the TTS provider.

The IBM Virtual Voice Creator (IVVC) is an end-to-end cloud-based voiceover solution endowed with an interactive web GUI studio: "*https://ivva-tts.sl.haifa.il.ibm.com*". The solution is based on and expands the IBM expressive TTS technology integrating a glottal vocoder with built-in online voice transformation capabilities [2]. IVVC lets the users create unique voice personas according to their needs and imagination and control the vocal performance of the artificial voice actors. Unlimited number of different human voices and exaggerated cartoonish ones can be derived from a single standard TTS voice. The voice design and performance control processes are easy, fast and do not require any special skills.

The IVVC offers the following features differentiating it from other TTS-based voiceover solutions such as [3] and [4]. The IVVC provides a powerful set of controls over the voice characteristics, unlike a basic set limited to the rate, pitch and sometimes the vocal tract length available in other solutions. IVVC also makes it possible to control the emotional style and emphasis in the generated speech.

IVVC show and tell presentation will include a live demonstration of its GUI studio with an active participation of the audience. We encourage the reader to explore the capabilities offered by IVVC web GUI studio "*https://ivva-tts.sl.haifa.il.ibm.com*" and check some examples available in IVVC_examples.html file attached to this publication.

2. IVVC architecture

The IVVC solution is built up of three layers as shown in Figure 1. The layers communicate with each other via HTTP API.

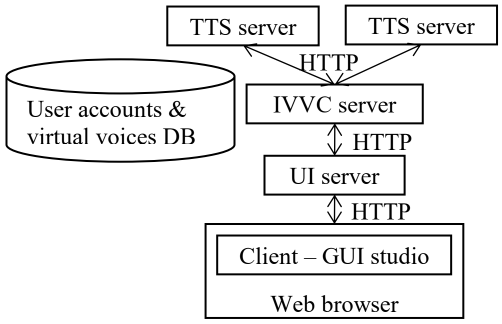


Figure 1: *High-level diagram of IVVC architecture*

The user interface (UI) layer employs a UI server that manages UI clients. Each client is an instance of the IVVC interactive web GUI studio running on the user's web browser.

The IVVC server is a middleware layer that manages user accounts populated with user defined virtual voice personas (or simply *virtual voices*), and dispatches speech synthesis requests to the TTS servers. A virtual voice is a light-weight object containing a reference to an original TTS voice and a set of global voice transformation configuration parameters. It is possible to create a workgroup account and share virtual voices within the workgroup. A synthesis request sent to the IVVC server from the UI server contains the virtual voice identifier and the input text. The same synthesis API call can be used to synthesize speech online from the user's application.

The TTS servers perform the synthesis in combination with online voice transformations specified via an SSML-style markup attached to the input text. Hence, the TTS servers can generate speech in newly created virtual voices without re-initializing. The IVVC server modifies the synthesis request format converting the transformation associated with a virtual voice to the SSML-style markup.

3. IVVC capabilities overview

In addition to listing IVVC functions we provide a very brief description of the underlying TTS and voice transformation technology elements and refer the reader to our previous works where this technology is presented.

Most of the functions described below are available after the user logs into his/her account. Signing in is not required to design a virtual voice, store it temporary and listen to the synthesized speech audio.

3.1. Virtual voice design

The Voice Design GUI allows the user to select one of the original TTS voices as a basis and set up a global voice transformation by moving the control sliders shown in Figure 2. The user enters a test sentence, moves the sliders and plays back the synthesized audio to assess the voice transformation effect. It is also possible to download the synthesized audio.

The vocal tract modification is implemented by a piece-wise linear frequency warping [2] where three input frequency nodes are fixed. Three *Vocal Tract* control sliders change the output nodes to allow independent modifications in low, middle and high frequency bands. Moving all the sliders to the left/right is perceived as vocal tract shrinking/stretching.

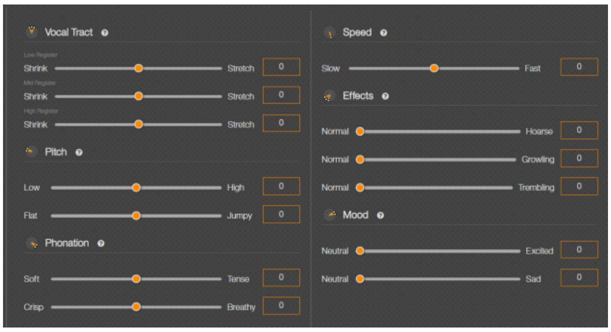


Figure 2: *Workarea of Voice Design page*

The F0 average level and dynamic range are controlled by the two respective sliders combined in the *Pitch* group. The *Speed* slider changes the speech rate.

The upper slider of the *Phonation* group makes the voice sounding softer or tenser by modifying the glottal pulse shapes as proposed in [2]. The lower slider controls the level of the synthesized aspiration noise [2] thus changing the perceived breathiness level.

Three sliders combined in the *Effects* group control certain voice qualities implemented as modifications of the synthesized glottal source signal at the pitch cycle level. Hoarseness and growling are implemented as randomized jitter and shimmer. Trembling is implemented as a low-frequency modulation of the F0 contour.

The *Mood* sliders make the voice inherently excited or sad. These controls are mutually exclusive. The expressive synthesis is implemented by switching between target prosody models [5] trained on a single-speaker emotional data, with a global non-linear mapping to the respective virtual voice. The user-controlled strength of the emotion is also incorporated in the mapping.

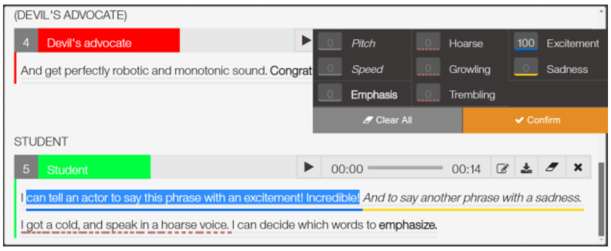
The user can save the current virtual voice (defined by the original voice selected and the control slider positions), in order to use it for speech synthesis as described below.

3.2. Offline synthesis using Script-to-Audio GUI

The Script-to-Audio GUI allows the user to upload a script containing the dialogues of the game/movie characters, select speaker turns (*sections*), assign corresponding virtual voices to the sections, synthesize and then download the resultant speech audio for integration into the game/movie sound track.

It is possible to tune the vocal performance within a section by selecting a part of it and altering the default synthesis along several axes as shown in Figure 3. These axes include pitch level, rate, vocal effects, emotion and word emphasis. The latter is implemented in the synthesis as described in [6].

Immediate audio feedback is available at the section level. It is also possible to play back all the sections sequentially and download all the section-wise audio files at once.


(DEVIL'S ADVOCATE)

4 Devil's advocate

And get perfectly robotic and monotonic sound. Congrat

0	*Pitch*	0	Hoarse	100	Excitement
0	*Speed*	0	Growling	0	Sadness
0	Emphasis	0	Trembling		

STUDENT

5 Student 00:00 00:14

I can tell an actor to say this phrase with an excitement! Incredible! *And to say another phrase with a sadness.*

I got a cold, and speak in a hoarse voice. I can decide which words to **emphasize.**

Figure 3: *Vocal performance control on Script-to-Audio page*

3.3. On-line speech synthesis using IVVC HTTP API

Applications that require online synthesis with virtual voices, such as chatbots and talking toys, can call the IVVC synthesis HTTP API directly.

4. Acknowledgements

The authors are thankful to Netanel Eder and Netta Shani for their contribution to the IVVC user experience design.

5. References

[1] S. Horowitz, and S. R. Looney, *The Essential Guide to Game Audio: The Theory and Practice of Sound for Games*. Taylor & Francis, 2014.

[2] A. Sorin, S. Shechtman, and A. Rendel, "Semi-parametric concatenative TTS with instant voice modification capabilities," in *INTERSPEECH 2017 – 97th Annual Conference of the International Speech Communication Association, August 20-24, Stockholm, Sweden, Proceedings*, 2017, pp. 1373–1377.

[3] Amazon Lumberyard TTS Cloud Gem (April 5, 2018). Retrieved from https://docs.aws.amazon.com/lumberyard/latest/userguide/cloud-canvas-cloud-gem-text-to-speech-cgp.html

[4] Crosstales RT-Voice (April 5, 2018). Retrieved from https://www.crosstales.com/media/data/assets/rtvoice/webgl/

[5] L. R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Using Deep Bidirectional Recurrent Neural Networks for Prosodic-Target Prediction in a Unit-Selection Text-to-Speech System", in *INTERSPEECH 2015 - 96th Annual Conference of the International Speech Communication Association, September 6-10, Dresden, Germany, Proceedings*, 2015, pp. 1606–1610

[6] S. Shechtman and M. Mordechay, "Emphatic speech prosody prediction with deep LSTM networks," in *ICASSP 2018 – IEEE International Conference on Acoustics, Speech and Signal Processing, April 15-20, Calgary, Alberta, Canada, Proceedings*, 2018