AT&T VoiceBuilder : A Cloud-based Text-To-Speech Voice Builder Tool

Yeon-Jun Kim, Thomas Okken, Alistair D. Conkie, Giuseppe Di Fabbrizio

AT&T Labs, Research - Inc.
180 Park Avenue, Florham Park, NJ - USA
{yjkim,tokken,adc,pino}@research.att.com

Abstract

The AT&T VoiceBuilder provides a new tool to researchers and practitioners who want to have their voices synthesized by a high–quality, commercial–grade text-to-speech (TTS) system without the need to install, configure, or manage speech processing software and equipment. It is implemented as a web service on the AT&T Speech Mashup Portal[1]. The system records, processes, and validates users' utterances, and provides a web service API to make the new voice immediately available to real-time applications. All the procedures are fully-automated to avoid human intervention.

Index Terms: Text-to-speech, Speech Mashup

1. Introduction

There has been for a long time great interest in custom TTS voices. Though AT&T Labs–Research receives regular queries about the feasibility of recording such voices, it has been very challenging to make a quality custom TTS voice until recently. It is not only because the recording component is somewhat complicated to get right, but also because many parts of the voice building procedure mandate a substantial amount of speech engineers' labor and expertise.

The AT&T VoiceBuilder significantly lowers the barrier to build custom voices by automating of the process in a number of ways. First, it reduces human intervention in the process of recording speech data by adoption of automatic speech recognition (ASR) techniques. Once sufficient data is collected and stored "in the cloud" for processing, the system largely automates the technical process of converting the raw speech to a form that can be used for a TTS Voice. This process uses many of the existing tools used to construct AT&T *Natural Voices*$^{\text{TM}}$voices. Once a voice is built, it is immediately available for use in the synthesizer via a web interface.

2. TTS Voice Building

The process of building a unit selection TTS voice consists largely of three modules: 1) spoken audio acquisition, 2) labeling audio segments with specified features and 3) compilation into a TTS voice index.

The first step is to decide the text to be read by the chosen speaker. This can vary widely and will depend on the language and intended uses of the TTS voice. Generally several hours of material are recorded to cover possible phone combinations and target domains. Examples would be newspaper text or written dialogs. A high quality recording system and a quiet environment are recommended for best results. During recording, it must be confirmed by someone that a speaker reads out the given text accurately. In this work, we introduce an automated way to monitor recording sessions using ASR, not shown in Kominek's system [1].

Once the recordings are made, speech processing can potentially take many weeks in order to align text and audio. The audio is segmented into individual speech sounds and vectors of features are assigned to each unit in the database. The specific set of features chosen and how they are used typically has a significant bearing on the quality of the resulting voice. The choice of units is not fixed and can be *phonemes* or *diphones* or some other unit. The feature information is compiled into a voice index while the audio is processed to allow efficient selection of chosen units.

For TTS synthesis, a unit selection voice and compatible unit selection synthesizer work together. The unit selection module is able to compare the specification requested by the input text with units in the database, via the index, and determine an optimal sequence of units. The selected sequence of units is concatenated with any optional signal processing then being applied before the output audio files are generated.

3. System Implementation

Figure 1 shows the overall system architecture. AT&T VoiceBuilder is part of a larger speech processing framework publicly available on the AT&T network cloud and accessible through the Speech Mashup Portal [2]. All the speech processing and speech synthesis components (Figure 1, boxes in green) are interfaced to the external world through the Speech Mashup Manager (SMM) and accessible as standard web services via a REST-style interface [3] as well as a web browser-based graphical interface.

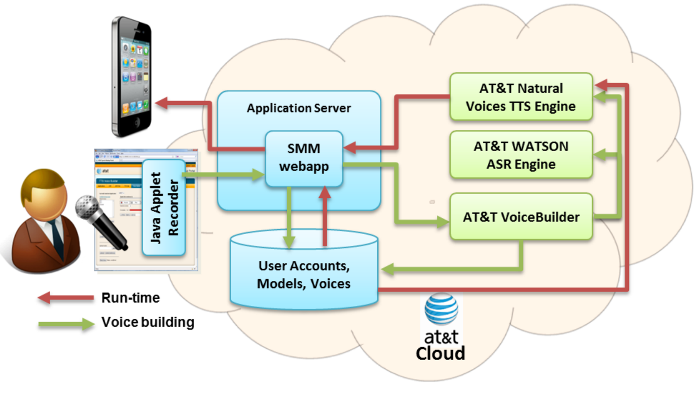
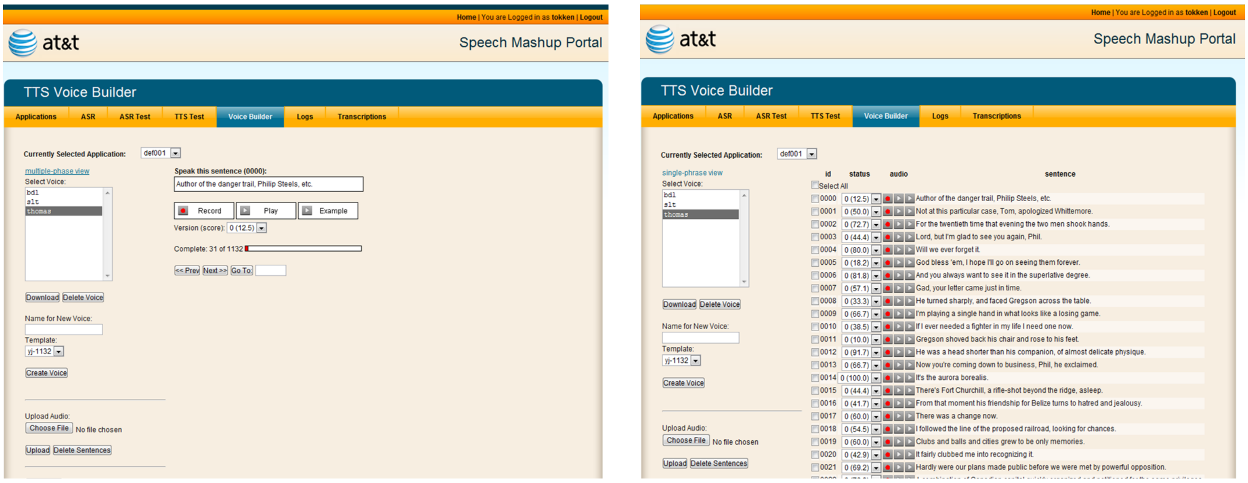


Figure 1: AT&T VoiceBuilder architecture

The AT&T VoiceBuilder web page uses a Java applet running in the user's browser to record and upload speech.

[1] https://service.research.att.com/smm



(a)

Currently Selected Application: def001

Select Voice: bdl, slt, thomas

Speak this sentence (0000): Author of the danger trail, Philip Steels, etc.

Version (score): 0 (12.5)

Complete: 31 of 1132

Name for New Voice:

Template: yj-1132

Upload Audio: No file chosen

(b)

Currently Selected Application: def001

Select Voice: bdl, slt, thomas

Template: yj-1132

Upload Audio: No file chosen

id	status	audio	sentence
□ Select All			
□ 0000	0 (12.5)	● ▷ ▷	Author of the danger trail, Philip Steels, etc.
□ 0001	0 (50.0)	● ▷ ▷	Not at this particular case, Tom, apologized Whittemore.
□ 0002	0 (72.7)	● ▷ ▷	For the twentieth time that evening the two men shook hands.
□ 0003	0 (44.4)	● ▷ ▷	Lord, but I'm glad to see you again, Phil.
□ 0004	0 (80.0)	● ▷ ▷	Will we ever forget it.
□ 0005	0 (18.2)	● ▷ ▷	God bless 'em, I hope I'll go on seeing them forever.
□ 0006	0 (81.8)	● ▷ ▷	And you always want to see it in the superlative degree.
□ 0007	0 (57.1)	● ▷ ▷	Gad, your letter came just in time.
□ 0008	0 (33.3)	● ▷ ▷	He turned sharply, and faced Gregson across the table.
□ 0009	0 (66.7)	● ▷ ▷	I'm playing a single hand in what looks like a losing game.
□ 0010	0 (38.5)	● ▷ ▷	If I ever needed a fighter in my life I need one now.
□ 0011	0 (10.0)	● ▷ ▷	Gregson shoved back his chair and rose to his feet.
□ 0012	0 (91.7)	● ▷ ▷	He was a head shorter than his companion, of almost delicate physique.
□ 0013	0 (66.7)	● ▷ ▷	Now you're coming down to business, Phil, he exclaimed.
□ 0014	0 (100.0)	● ▷ ▷	It's the aurora borealis.
□ 0015	0 (44.4)	● ▷ ▷	There's Fort Churchill, a rifle-shot beyond the ridge, asleep.
□ 0016	0 (41.7)	● ▷ ▷	From that moment his friendship for Belize turns to hatred and jealousy.
□ 0017	0 (60.0)	● ▷ ▷	There was a change now.
□ 0018	0 (54.5)	● ▷ ▷	I followed the line of the proposed railroad, looking for chances.
□ 0019	0 (60.0)	● ▷ ▷	Clubs and balls and cities grew to be only memories.
□ 0020	0 (42.9)	● ▷ ▷	It fairly clubbed me into recognizing it.
□ 0021	0 (69.2)	● ▷ ▷	Hardly were our plans made public before we were met by powerful opposition.

Figure 2: The AT&T VOICEBUILDER web interface

Real–time recording and network streaming activities are managed in the background by the applet, which only shows the recording activities in the browser. After each recording is uploaded, the applet transfers the recorded utterance to the AT&T *WATSON*™ ASR [4] to match the spoken utterance with the expected text.

Users are encouraged to use a high quality microphone and to record their sessions in quiet indoor locations. The ASR includes a noise and garbage model to detect poor recording conditions as well as a language model designated to detect spoken utterances different from the expected sequence of phonemes. If a recording receives a low *confidence score* from the ASR, the user can record the sentence again till the quality is satisfactory; the server keeps the most recent recordings (up to 10 utterances/sentence) and allows the user to choose which version they prefer. During the recording session, reference utterance recordings can be played back to reduce discrepancies between what the system expected and what a speaker actually utters.

The system also provides users an alternative way to submit their voice from their audio equipment instead of recording their voice using our web-interface. When a user has quality audio equipment, the system allows a user to create audio recordings on their own and upload them in bulk bundled in a zip file. In this case, users could expect better synthesis outcome from their submissions, however, they should make sure that each audio file matches the corresponding sentence. The recordings uploaded in this manner, and their scores, may then be reviewed, and, if necessary, updated, in the Voice Builder page.

Once a sufficiently large number of sentences has been recorded with sufficiently large scores, the user can launch the voice building procedure by clicking a button on the AT&T VOICEBUILDER page. The procedure can continue running in the background even after the user logs off from the portal. The user can monitor the status of the current Voice Builder process at any time, and cancel and restart it if necessary. When the Voice Builder has completed successfully, it deposits the generated voice files in the user's section of the SMM file system, along with the user's audio recordings, ASR grammars, and log files. The voice can then be tested using the portal's 'TTS Test' page, and used via the portal's TTS REST service.

4. Demonstration

Users can access the AT&T VOICEBUILDER system and create or manage their TTS voices by registering an AT&T Speech Mashup account with a standard web browser. Among many applications on the SMM, the AT&T VOICEBUILDER menu can be found on the 'TTS test' application. AS shown in Figure 2 (a), users can name their TTS voices and record the given prompts one at a time or upload speakers' audio in bulk. For each utterance, there are two play buttons: one to play back the current user's recording, and the second to listen to the reference recording from a professional speaker.

Figure 2 (b) shows the list of the whole recording session so that users can confirm their audio before they start the voice building procedure. The current system allows users to submit audio for the given text only. Finally, there is a button to create a custom TTS voice on the left corner of the web-page. An e-mail notification will be delivered at the completion of the procedure. The whole procedure usually takes around a couple of hours on the AT&T cloud computing environment.

5. Conclusions

This demonstration illustrates an implemented cloud-based voice building system which enables researchers and practitioners to create their own custom TTS voice by utilizing AT&T's world-leading ASR and TTS technologies. Future releases will include support for other languages.

6. References

[1] J. Kominek, T. Schultz, and A. W. Black, "Voice building from insufficient data - classroom experiences with web-based language development tools," in *6th ISCA Workshop on Speech Synthesis (SSW-6)*, Bonn, Germany, August 22-24 2007.

[2] G. Di Fabbrizio, T. Okken, and J. G. Wilpon, "A speech mashup framework for multimodal mobile services," in *11th International Conference on Multimodal Interfaces (ICMI 2009)*, 2009, pp. 71–78.

[3] R. T. Fielding, "REST: architectural styles and the design of network-based software architectures," Doctoral dissertation, University of California, Irvine, 2000.

[4] V. Goffin, C. Allauzen, E. Bocchieri, D. Hakkani-Tür, A. Ljolje, S. Parthasarathy, M. Rahim, G. Riccardi, and M. Saraclar, "The AT&T WATSON Speech Recognizer," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2005.