Building Text-To-Speech Voices in the Cloud

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

AT&T Labs – Research
180 Park Avenue, Florham Park, NJ - USA
{adc,tokken,yjkim,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 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. The system records and validates users' utterances, processes them to build a synthetic voice and provides a web service API to make the voice available to real-time applications through a scalable cloud-based processing platform. All the procedures are automated to avoid human intervention. We present experimental comparisons of voices built using the system.

Keywords: Text-to-Speech, Voice Building, Cloud Computing

1. Introduction

There is considerable interest in custom text-to-speech (TTS) voices. AT&T Labs–Research receives regular queries about the feasibility of building such voices, from a variety of sources.

Until recently it has been very difficult to make a good TTS voice. The recording component is somewhat complicated, and parts of the voice building procedures can require a substantial amount of speech engineers' labor and expertise.

Infrastructure for producing custom synthetic voices is not widely available. The open source Festival Speech Synthesis System (Black et al., 1999) has the capability, but the user is required to learn a considerable amount about the system before it is realistically possible to construct a synthetic voice. Currently it appears that users either prefer the supplied voices or are students or researchers building complete synthesis systems in new languages.

Another resource is the Speech Interactive Creation and Evaluation Toolkit (SPICE) from CMU which is a web-based system primarily for helping to develop speech technology in under-resourced languages, e.g., Afrikaans, Vietnamese, and Bulgarian. It assumes the user has some familiarity with speech processing.

Perhaps the closest system to what we describe here is ModelTalker (Bunnell et al., 2010). ModelTalker is designed primarily to help individuals who, for medical reasons, are likely to have severely reduced speaking ability within a period of time. In terms of speech recordings the scope is somewhat more limited than described here, and there is generally a need for more flexible support due to individual circumstances and often an initial lack of knowledge about the technology involved.

AT&T VoiceBuilder is targeted at technology-savvy users who have an awareness of speech technology, but are not necessarily speech specialists. It significantly lowers the barriers to building a high-quality custom voice by automating the process in a number of ways. Firstly, it reduces human intervention in the process of recording speech data by adoption of automatic speech recognition (ASR) techniques. Secondly, 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. Finally, once a voice is built, it is immediately available for use in the synthesizer via a web interface, and processing resources are allocated in the cloud.

2. TTS Voice Building

Building a high-quality unit selection TTS voice currently depends on two main factors. The first is a set of voice recordings, the second is the technology to turn the recordings into a synthetic voice.

The voice recording process requires the speaker to read a substantial amount of written text. The text used has properties that make it desirable for use in a synthesis system. Generally, several hours of material are recorded. Examples would be newspaper text or written dialogs. Voice quality will depend in part on the amount of material recorded. A high quality recording system and a quiet environment are recommended for best results. The process of recording is likely to span several hours, perhaps spread over a few days. Recording and speaking *consistency* are also important so that recordings are uniform in rate and quality. During recording an automatic check using ASR is done to verify the speaker's audio matches the text that was presented for reading.

Once a set of recordings is complete, the text and audio recordings are processed together. The audio is segmented into words and phonemes and vectors of features are assigned to each phoneme. Features include obvious characteristics such as pitch and duration, as well as less obvious markers, such as a feature identifying whether a particular consonant is before or after the vowel in a syllable. The set used in the build process has previously been evaluated as contributing to the speech quality of unit selection voices. A training phase is used to find appropriate weights, per phoneme, for the features.

Next the units and features are packaged together with the audio into a "voice module". At unit selection time this voice module is called upon to provide suitable units that are then evaluated for synthesis. The best available sequence of audio units is then concatenated together and output.

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 (Di Fabbrizio et al., 2009). All the speech processing and speech synthesis components (Figure 1, the three boxes in green on the right) are interfaced to the external world through the Speech Mashup Manager (SMM) and accessible as standard web services via a REST-style interface (Fielding, 2000) as well as a web browser graphical interface.

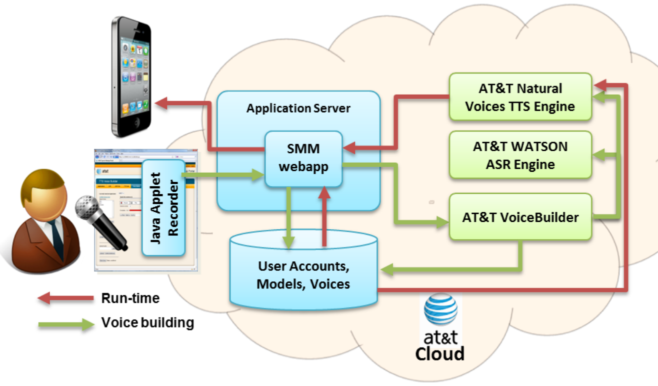


Figure 1: AT&T VoiceBuilder architecture

The AT&T VoiceBuilder page uses a Java applet hosted in the user's browser to record and upload speech. Real-time recording and network streaming is managed transparently by the applet.

After each recording is uploaded, it transfers the recorded utterance to AT&T *WATSON*^{SM} ASR (Goffin et al., 2005) to match the spoken utterance with the expected text.

The ASR includes a noise and garbage model to detect poor recording conditions as well as a language model designed to report spoken utterances that deviate from the expected sequence of phonemes. If a recording receives a low *confidence score* from ASR, the user can record the sentence again till the quality is satisfactory. During a 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 allows users to create audio recordings themselves, if they prefer, and upload them in batch mode. In this case, users may achieve better audio quality from their submissions, but don't benefit from the interactive checking.

Once a sufficiently large number of sentences have been recorded with sufficiently good 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 AT&T VoiceBuilder process at any time, and cancel or restart it if necessary. When the AT&T VoiceBuilder 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-based web service.

4. Current Status

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 prompt 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 build procedure usually takes around 2–3 hours on the AT&T cloud computing environment.

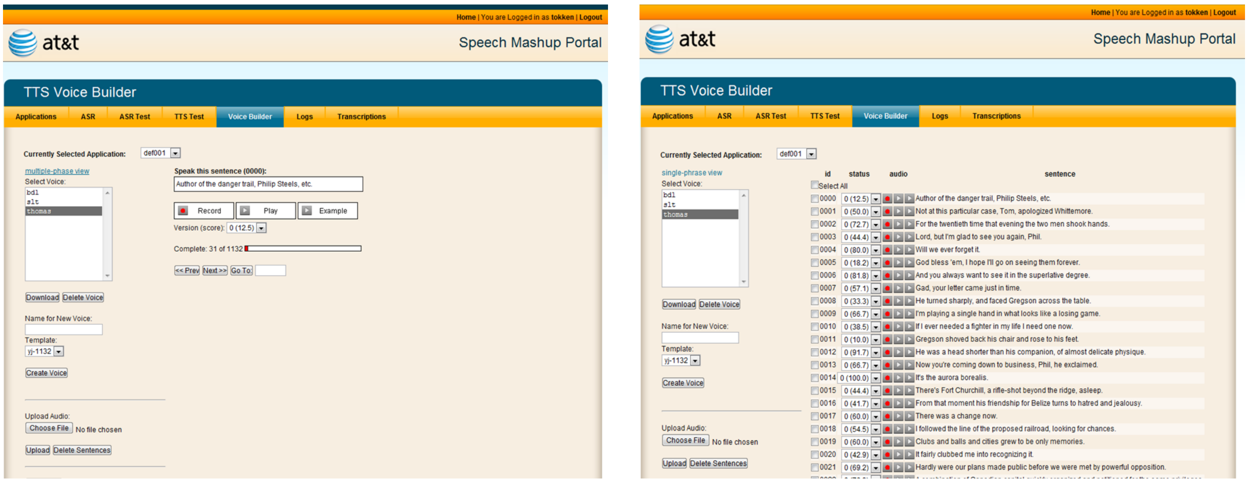
5. Evaluation

We conducted three experiments to compare the AT&T VoiceBuilder system with the Festival system and Festival system voices. Listeners were asked to rate examples of synthetic speech. Ten sentences from the Harvard (IEEE Subcommittee on Subjective Measurements, 1969) phonetically-balanced sentence set were selected as input text for the synthesizer (see Table 1). The same set of sentences was used for all three experiments. Sentences were presented in pairs, with the order of the sentences within pairs randomized. Additionally the order in which sentence pairs were presented was randomized.

Prompt	Text
harvard_151	The empty flask stood on the tin tray.
harvard_152	A speedy man can beat this track mark.
harvard_153	He broke a new shoelace that day.
harvard_154	The coffee stand is too high for the couch.
harvard_155	The urge to write short stories is rare.
harvard_156	The pencils have all been used.
harvard_157	The pirates seized the crew of the lost ship.
harvard_158	We tried to replace the coin but failed.
harvard_159	She sewed the torn coat quite neatly.
harvard_160	The sofa cushion is red and of light weight.

Table 1: Ten test sentences selected from the Harvard set of phonetically-balanced sentences

Subjects were web-users who visited the test page and submitted responses to the questions.



(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

Name for New Voice:

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

The tests were web-based A/B comparisons on a 5 point scale. The five choices were: Strongly prefer A (A++), Prefer A (A+), No preference, Prefer B (B+), or Strongly prefer B (B++). Apart from the audio judgments, subjects were asked two additional questions. First, whether they listened via headphones or loudspeaker. Second, whether they considered themselves a native speaker of English or not.

Evaluation was done using Comparative Mean Opinion Score (CMOS). CMOS measures the average score for a stimulus or group of stimuli, where we assign "Strongly prefer" a value of 2 and "No preference" a value of 0. A strong preference for one of a pair corresponds to a strong negative preference for the other.

5.1. First Test

The first experiment involved comparing the Festival "nitech_us_slt_arctic_hts" female voice with a voice built from the ARCTIC "SLT" speech database (Kominek and Black, 2004) using the AT&T VoiceBuilder framework. Both voices are based on the same recorded speech database. The Festival voice was chosen because it is the default on the test system which was a Linux distribution running RedHat Fedora 12 (Constantine).

Test one had 77 participants, 59 (77%) were native speakers of English, 34 (44%) listened via headphones, 43 (56%) via loudspeaker(s). The overall result was a CMOS score of 1.32 (on a -2 to 2 scale) for the AT&T VoiceBuilder system. (The corresponding score for the Festival system was -1.32).

This large difference can probably be attributed to the signal processing used by the Festival voice.

Breaking down the results according to scores in Figure 3, there are many responses showing a preference for the SLT (non-Festival) voice (1 and 2), and very few responses indicating a preference for the Festival voice (-2 to 0).

Looking in more detail at the per-sentence results we see in Figure 4, there is a degree of variability among individual sentences but the results indicate a consistent preference for the SLT (non-Festival) voice.

There is only a very slight order bias effect with a CMOS preference of 0.036 for the second member of a pair, presumably because the stimuli were notably different.

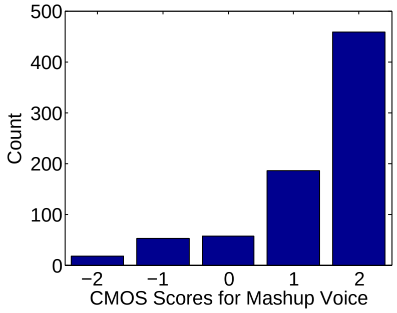


Figure 3: Experiment 1 score distribution for the mashup voice

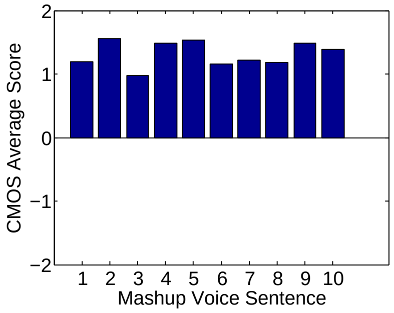


Figure 4: Experiment 1 per-sentence CMOS variation for the mashup voice

5.2. Second Test

The second test compared two voices built with the AT&T VoiceBuilder. One voice used 300 sentences from the ARCTIC database male voice "BDL" and the second female voice "AS" was built using the AT&T Voice-

Builder system by a representative volunteer (i.e., not a speech synthesis professional) who used the same 300 sentences. We expect differences here to reflect primarily the differences in voice quality and recording experience.

For this test there were 81 participants, 59 (73%) were native speakers of English, 32 (40%) listened via headphones, 49 (60%) via loudspeaker(s). The overall result was a CMOS score of 0.47 (on a -2 to 2 scale) for the AS voice, -0.47 for the BDL voice.

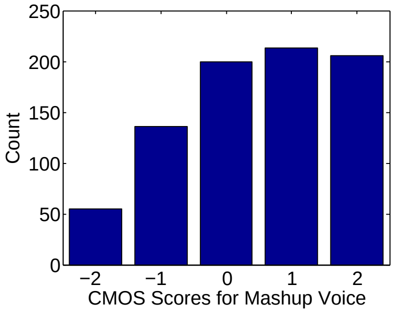


Figure 5: Experiment 2 score distribution for the AS voice

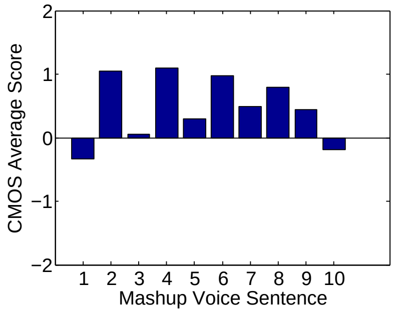


Figure 6: Experiment 2 per-sentence CMOS variation for the AS voice

In this case Figure 5 shows a more evenly balanced set of scores. The notable point is perhaps that there are few cases where the BDL voice is strongly preferred. In Figure 6 for this experiment we observe cases where particular sentences are preferred on one system or the other, with the overall trend favoring the AS voice which is preferred on average for eight of the ten sentences.

This time the order effect is 0.31. That is, listeners have a preference for the second sample in a pair on average.

5.3. Third Test

The third experiment compared the same "AS" voice (300 sentences) with the highest quality ARCTIC database "SLT" voice (1200 sentences) we could build, using all available data. It is intended to indicate an approximate upper limit of improvement possible within the system.

For this third test there were 61 participants, 50 (82%) were native speakers of English, 26 (43%) listened via headphones, 35 (57%) via loudspeaker(s). The overall result was a CMOS score of 0.68 (on a -2 to 2 scale) for the SLT voice, -0.68 for the AS voice. Hence the 1200 sentence SLT voice is clearly preferred over the AS voice.

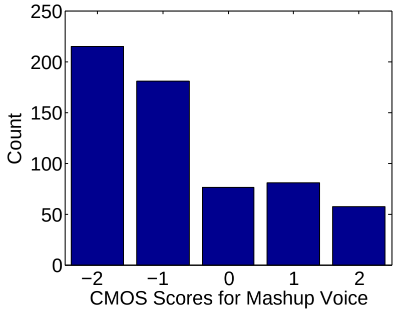


Figure 7: Experiment 3 score distribution for the AS voice

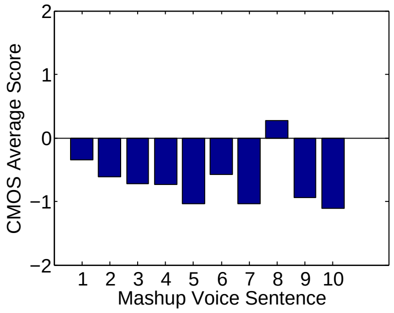


Figure 8: Experiment 3 per-sentence CMOS variation for the AS voice

Here Figure 7 indicates that the bulk of the responses favor the large SLT voice, and in more detail Figure 8 indicates that sentence by sentence the SLT voice is almost always preferred, with AS being preferred for only one sentence.

The order effect this time is 0.13 on the CMOS scale, again for the second stimulus.

6. Conclusions

This paper introduces the AT&T VoiceBuilder cloud-based voice building system, which enables researchers and practitioners to create their own custom TTS voice by utilizing AT&T's ASR and TTS technologies.

A representative voice from the system performs well in comparison with reference publicly available voices. Results indicate that additional recordings are likely to yield higher quality.

Future work will focus on maximizing quality and on reducing the amount of recording needed, using adaptation techniques.

7. Acknowledgements

We thank the developers of the ARCTIC databases and the Festival Speech Synthesis System for making available their data and software.

8. References

A. W. Black, P. Taylor, and R. Caley. (1999). The Festival Speech Synthesis System. Technical report.

H. T. Bunnell, J. Lilley, C. Pennington, B. Moyers, and J. Polikoff. (2010). The ModelTalker System. In *Proceedings of The Blizzard Challenge Workshop*, Nara, Japan.

Giuseppe Di Fabbrizio, Thomas Okken, and Jay G. Wilpon. (2009). A speech mashup framework for multimodal mobile services. In *ICMI*, pages 71–78.

Roy Thomas Fielding. (2000). *REST: Architectural Styles and the Design of Network-based Software Architectures*. Doctoral dissertation, University of California, Irvine.

Vincent Goffin, Cyril Allauzen, Enrico Bocchieri, Dilek Hakkani-Tür, Andrej Ljolje, S. Parthasarathy, Mazin Rahim, Giuseppe Riccardi, and Murat Saraclar. (2005). The AT&T WATSON Speech Recognizer. In *ICASSP*, Philadelphia, USA.

John Kominek and Alan W. Black. (2004). The CMU Arctic speech databases. In *SSW5*, pages 223–224, Pittsburgh, USA.