INTERSPEECH 2016: Show & Tell Contribution
September 8–12, 2016, San Francisco, USA

My-Own-Voice: A web service that allows you to create a Text-to-Speech voice from your own voice.

Fabrice Malfrere, Olivier Deroo, Emmanuelle Franques, Jonathan Hourez, Nicolas Mazars, Vincent Pagel, Geoffrey Wilfart

ACAPELA Group, 33 Boulevard Dolez, 7000 Mons, BELGIUM
mov-information@acapela-group.com

Abstract

My-Own-Voice is a service that provides a tool to end-users 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. The system records and validates users' utterances with Automatic Speech Recognition (ASR), to build an HMM or a Unit Selection synthetic voice. All the procedures are automated to avoid human intervention. We describe here the system for particular end-users about to lose the ability to speak with their own voice, who can now synthetically re-create it with the help of their speech therapist, enabling them to preserve this essential part of their identity.

Index Terms: Text-To-Speech; Speech Impairment

1. Introduction

There is considerable interest in custom Text-to-Speech (TTS) voices. CMU and AT&T Labs–Research were the first to provide a similar concept in [1][2] but these developments were meant for students and speech researchers for demonstration purposes. The *My-Own-Voice* concept is designed for people who are about to lose the ability to speak with their own voice (diagnosed with speech or language disorders resulting from ALS (Amyotrophic lateral sclerosis) or other conditions such as aphasia, dysarthria, throat cancer or apraxia). *My-Own-Voice* is already available in 13 languages (BE/NL Dutch , US/UK/AU English, CA/FR French, German, Italian, Norwegian, ES/US Spanish, Swedish).

2. Description

In 2010 Acapela Group participated in the Diyse ITEA-funded project [3] and developed the first web based voice creation tool. Following this R&D proof–of-concept which originally targeted general-purpose voice creation, Acapela Group officially launched a breakthrough service allowing users diagnosed with speech or language disorders to capture the essence of their voice before losing it. Users can keep speaking with their own voice rather than having to use a standard anonymous synthetic voice. Once a patient is diagnosed time is precious. Our goal was thus to develop the most convenient service, while ensuring that all the sounds we need to create a synthetic version of their voice could be recorded.

Most of today's commercial-grade TTS systems rely on Unit Selection technology. They are based on recordings by a highly-trained professional voice-talent in a sound-proof recording booth. Several thousands of sentences are part of the recipe that ensures the success of those industrial systems. But in this particular project, time is of the essence:

- HMM synthesizers start with 500 sentences
- Unit Selection voices can be built from 1500 sentences

3. System implementation

To allow individuals to capture their own voice, the first challenge was to create an entirely on–line process to enable recordings to be made by non-professional speakers, without any help from Acapela linguists, but with the support of the user's speech therapist. The application is web-based and the only technical requirement is a good quality headset microphone.

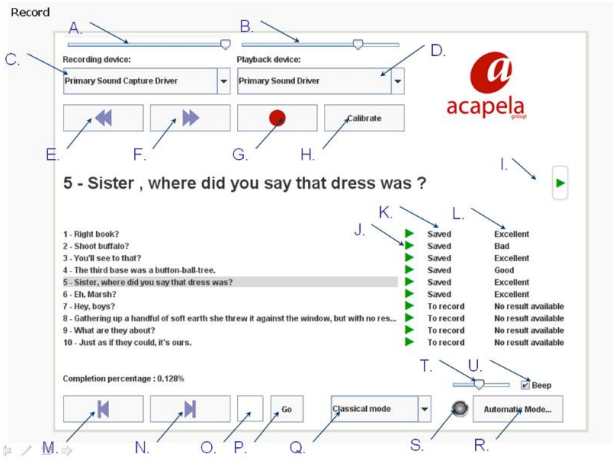

Figure 1: *My-Own-Voice* web interface

Figure 1 shows the interface where users:

- can control recording and playback levels (A. and B.), or select audio drivers (C. and D.), or do some microphone calibration (H.) because it's important to keep consistency between recordings;
- select which sentences to record (E., F., G.), and navigate through a list of sentences (M. N.)
- inspect ASR evaluation of the recordings (L.). They can listen to generic TTS sentence pronunciation (I.) or previously recorded sentences (K.),

During the recording, ASR detects if a pronunciation is different from the one expected by the system [4]. After the recording is analyzed, the words that may have been

pronounced unclearly or differently will be highlighted in red (Figure 2). Undesired pauses are indicated with a tilde. If a sentence is not recognized by the system, the user can skip it or listen to the generic TTS pronunciation example if unsure. Once ASR scores are satisfactory and a sufficiently large number of sentences has been recorded, the user can start building the voice. When the voice has been built successfully, it is available on-line.

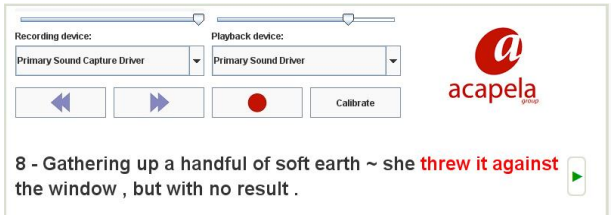

Figure 2: Speech Recognition feedback

4. Early adopters and first results

My-Own-Voice was officially launched on March 2015 at CSUN San Diego and first statistics show a strong and persistent interest from the patients and market:

- More than 110 accounts opened
- More than 35 voices created
- Multiple languages: Dutch, English, French, German, Italian, Norwegian, Spanish, Swedish

Eric, Garmt (both diagnosed with ALS) and Peter (suffering from cancer) have decided to create their own synthetic voice and have participated as pioneers in the first *My-Own-Voice* experiences. Videos of these users' testimonials are available on the *My-Own-Voice* web site.

4.1. First Field experiences with Garmt

Garmt : '*The voice from Acapela is an important part of me. No, that is not a grammatical error. My voice is a part of who I am, part of my identity. ALS took that away quite fast. Within months, the synthetic voice sounded more like myself than I did. The first few samples that I played for my friends convinced me that the time investment was worth it, as they honestly thought I was speaking myself instead of artificially. Now, I use it almost exclusively to communicate. I am being taken apart by this disease but this part of me is safe.*'

4.2. First Field experiences with Eric

The voice creation depends on the patient's physical condition and can meet more obstacles like with Eric's experience: *"Since the day I was diagnosed with ALS 2 years ago, my brother Josh and I learned my voice would be rapidly declining and very soon I would be unable to speak. This would frighten anyone, but I have always been a man of many words so it was incredibly crushing to me. My brother took action and we focused on finding the best synthetic voice out there that could replicate my voice so I could 'hear' myself again. Unfortunately, by that point, I was unable to complete the clear speaking hours necessary to create my own voice. I was depressed again. Then Acapela Group suggested if my brother had a similar voice to mine that he could do the speaking and then we could deepen the voice to sound like mine! The voice is extremely realistic."*

In such a case HMM technology [5] is beneficial as it can separate vocal tract parameters from the prosodic model. Training the latter can prove difficult when somebody does not have the stamina to record enough data. Vocal tract parameters on the other hand can be tuned to resemble the original speaker with less recorded data.

4.3. Multilingual databases with Peter

Peter speaks German, Spanish and English. When a surgery to treat his cancer threatened to take away his voice, he took steps to ensure he would not lose his ability to effectively communicate with family, friends and co-workers. Working with the *My-Own-Voice* service, he captured his voice in all three languages. A foreign accent can be captured by the models.

4.4. Analysis from use cases

From an economic point of view it is clear that such a system has to be as automatic and self-sufficient as possible. But from an ethical point of view, Acapela must provide situation-dependent support as every case is unique. For example:

- Acapela monitors the recordings and more specifically the beginning of new sessions to make sure that the basic quality requirements are met. If not we contact the speaker.
- Acapela provides guidance for technological options (Unit Selection; HMM; HMM with prosody transplant).

5. Conclusions

A functional cloud-based voice-building system has been presented, which enables users diagnosed with speech or language disorders to capture the essence of their voice before losing it. They can create their own custom TTS voice by using Acapela's ASR and TTS technologies. Today, the *My-Own-Voice* website is ready to record, fully guiding the user and the speech therapist through the full process, based on sentences easy to deal with for those without any specific knowledge of speech technology.

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] Alistair Conkie, Yeon-Jun Kim, Thomas Okken, Giuseppe Di Fabbrizio, "Building Text-To-Speech Voices in the Cloud", The eighth international conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, May 21-27, 2012.

[3] https://itea3.org/project/diy-smart-experiences.html

[4] F. Mafrere, O. Deroo, T. Dutoit, C. Ris, 2003, "Phonetic alignment: speech-synthesis-based versus Viterbi-based", Speech Communication, 2003, vol. 40, n°4, pp. 503-517.

[5] Heiga Zen, Keiichi Tokuda, and Alan W. Black. "Statistical parametric speech synthesis." Speech Communication, 51(11):1039 – 1064, 2009. ISSN 0167-6393