INTERSPEECH 2016 ADDENDUM

The following Show & Tell demonstration was not included in the Interspeech 2016 abstract book and online proceedings. It was presented on Sunday, 11 September 2016 during Show & Tell Session 6, paper code Sun-S&T-6-4.

MIVOQ-PTTS - A Revolutionary New Way of Thinking TTS

Piero Cosi[1], Giulio Paci[2], Giacomo Sommavilla[2], Fabio Tesser[1]

[1]ISTC-CNR, UOS Padova
[2]MIVOQ S.R.L., Padova, Italy
{piero.cosi, fabio.tesser}@pd.istc.cnr.it
{giulio.paci, giacomo.sommavilla}@mivoq.it

Abstract[1]

MIVOQ-PTTS is a new TTS project whose goal is to offer innovative services suitable for creating and using personalized synthetic voices. Users can autonomously create their own synthetic voices by accessing a web interface and recording some sentences; the voice creation procedure do not require any other human intervention. In this work we will introduce MIVOQ-PTTS main ideas and we will illustrate the current state of development of its first web demonstrator.

Index Terms: TTS, Personalization, Personalized Synthetic Voices

1. Introduction

A Text To Speech (TTS) system automatically generates speech from the corresponding text and its typical use comprises dynamic/automatic generation of speech messages (i.e.: train stations), talking machines (robots, avatars), call centers, audiobooks, GPS navigation systems, screen readers, videogames, etc.

Classical TTS voice building procedure is time consuming and expensive: by using typical voice creation methods, results are outstanding, but they require a large amount of recorded speech and intensive post-production work. Thus, most users are forced to choose a synthetic voice in a limited catalog and this is a barrier for those who need a TTS voice similar to their natural one.

MIVOQ Personalizable TTS (PTTS) allows fast and easy creation of custom synthetic voices, so the users can choose among a potentially infinite list of voices and they can even choose their own voice. This may help, for example, people with degenerative diseases[2] or people that are going to loose their voice due to surgeries such as a laryngectomy: if the patient can record some sentences before losing her/his voice, with PTTS he will be able to talk with her/his original voice using a speech-generating device.

2. MIVOQ-PTTS

The current prototype for MIVOQ PTTS (schematically drawn in Figure 1) consists of a web application where users can record sentences and automatically obtain a personalized TTS voice model that will possess the vocal characteristics of the author of the recordings.

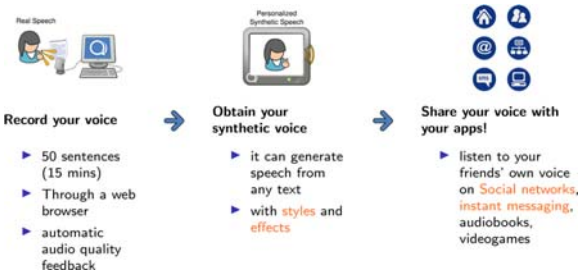

Figure 1: MIVOQ PTTS voice creation procedure.

Moreover, MIVOQ API are designed to create a "Voice Network Infrastructure" which allows the users to share their voice and to listen to their friends' synthetic voices.

Thus, MIVOQ-PTTS service can expand the classic idea of TTS (adding voice personalization and flexible TTS features, such as emotional styles, vocalizations and audio effects), allowing:

- voice preservation and voice banking;
- the creation of new applications in the fields of social networks, video games, animated movies, …;
- creative use of Text To Speech (e.g., dubbing of animated movies, interpret emoticons in a message through the generation of speech styles or vocalizations);
- the creation of email/sms reading services with the voice of the authors (and similar applications).

The technological value proposition of MIVOQ-PTTS consists on:

- the building of personalized voices in very short time (less than two hours with the current prototype, voice recording included);
- the opportunity for the users to record their voices in a quick and cheap way without the need for specialist equipment or a studio environment;
- the recording is carried on through a web application, no professional environment is needed (although using a professional setup will improve the final quality).

To record their voice, users will require a quiet room, free from background noise and with as little echo as possible. The better the quality of the recording conditions, the better the produced TTS voice will be. Users will be asked to read out prompts from specifically designed scripts, and the collected audio data will be used to build custom TTS voice. In order to

[1] Authors are in alphabetic order.
[2] MND (Motor Neuron Disease) or ALS (Amyotrophic Lateral Sclerosis)

build synthetic voices with less than 50/100 sentences MIVOQ-PTTS uses speaker adaptation techniques [3] integrated in a statistical parametric speech synthesis (SPSS) framework [4,5]. The recordings are carried out through a web interface and users will require one of the main browsers (Chrome, Firefox, IE or Opera). The web interface is working with most operating systems, included those running on smartphones and tablets. A notable exception is iOS, were no browser supports any method to record audio even when recording video is possible.

3. MIVOQ-PTTS Web Infrastructure

The PTTS service consists of a web speech acquisition interface built on top of a RESTful API (PTTS API) and its infrastructure is illustrated in Figure 2. The web interface has been developed using HTML5 and getUserMedia API for recording, with a flash fallback where that API is not available. The backend is implemented in a cloud fashion, using loosely coupled modular components. Authentication is carried out using OAuth 2.0. Users can authenticate using login and password, or using Google or Facebook authentication. Mivoq act as an OAuth 2.0 provider itself in order to let applications use the service on users' behalf.

Once authenticated, users will be presented with text prompts: when the users record them, the web interface gets the parameters for the upload from the PTTS API and then sends the data directly to the audio repository (no proxy is involved), in order to limit the required bandwidth. The repository will proactively notify the backend that a new audio file is available immediately after the upload is complete. By using a context broker, subscribed components, such as those carrying out quality feedback measures, are notified about the upload and may notify other components about the of their analysis. Adding a new component is just as simple as adding proper subscriptions in the context broker. Whenever a new quality feedback is ready, it is notified to the speech acquisition interface using Websockets, so that it can be immediately shown to the users.

Using a very similar mechanism, it is possible to inform components carrying out voice models creation that enough good quality audio is available. Using PTTS API the components may retrieve voice models parameters for the upload, and directly upload the file to the voice model repository. On upload, the voice model repository proactively notifies the system about the event, which is then propagated to the speech interface and to the user.

Text-To-Speech service is provided by implementing Flexible and Adaptive Text To Speech API (FA-TTS). Users and applications will be able to use all those voices for which they have been granted access, which includes catalogue voices, their own voices and voices for which they have obtained access from other users. Through FA-TTS, styles and effects are exposed to the developers, so that it is possible to change several aspects of the voice (pitch, rate, tract scale) and apply several effects.

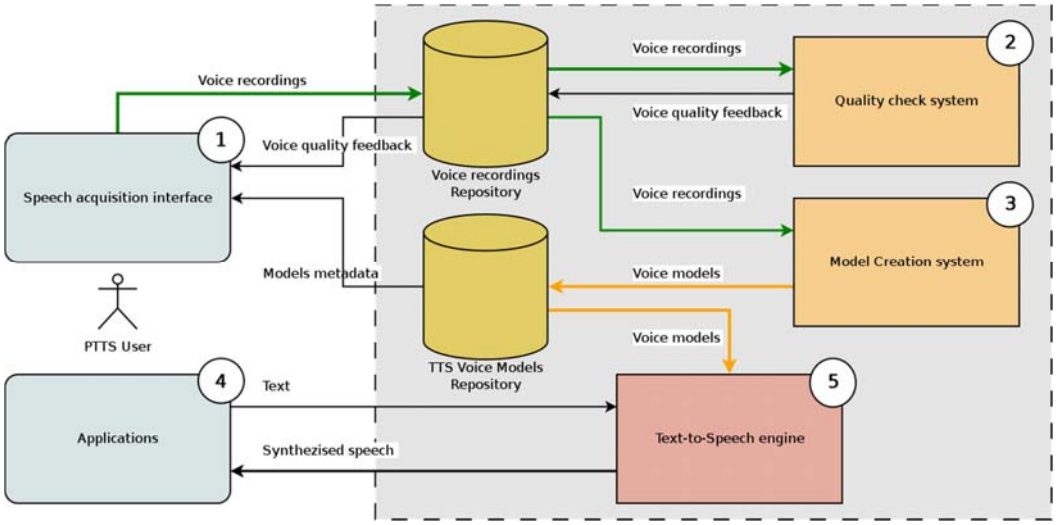

Figure 2: MIVOQ PTTS WEB infrastructure.

4. Conclusions

The MIVOQ-PTTS web demonstrator is up and running at https://www.mivoq.it and beta testers are welcome.

5. References

[1] MIVOQ WEB site: https://www.mivoq.it

[2] FA-TTS - website:http://lab.mediafi.org/discover-flexibleandadaptivetexttospeech-overview.html.

[3] Junichi Yamagishi, Takao Kobayashi, Yuji Nakano, Katsumi Ogata, and Juri Isogai. Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm. IEEE Transactions on Audio, Speech, and Language Processing, 17(1):66–83, January 2009.

[4] Heiga Zen, Keiichi Tokuda, and Alan W. Black. Statistical parametric speech synthesis. Speech Communication, 51(11):1039–1064, November 2009.

[5] The HMM-Based Speech Synthesis System (HTS). http://www.hts.sp.nitech.ac.jp/