Online Platform for obtaining Personalized Synthetic Voices

- Project Proposal for eNTERFACE’14 -

Daniel Erro & Inma Hernáez, UPV/EHU

Technological background

Speech synthesis technologies have evolved during the last decade from selection and concatenation paradigms [1] to statistical parametric ones [2][3]. The main advantage of hidden Markov model (HMM) based speech synthesis is its enormous flexibility for speaker/style adaptation [4], though it exhibits also some others: low footprint (<5MB in most cases!), smooth synthetic signals without annoying discontinuities, etc. The availability of an open source statistical parametric speech synthesis system, HTS [5], has played a key role in this recent technological evolution. Statistical parametric speech synthesis has enabled lots of new applications that were not possible in the previous technological states: voice reconstruction for people with several speech impairments [6], personalized speech-to-speech translation [7], noise-robust speech synthesis [8], etc.

[1] Hunt, A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database”, in Proc. ICASSP, vol. 1, pp. 373-376, 1996.

[2] H. Zen, K. Tokuda, A. W. Black, “Statistical parametric speech synthesis”, Speech Commun., vol. 51, no. 11, pp. 1039-1064, 2009.

[3] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, K. Oura, “Speech synthesis based on hidden markov models”, Proc. IEEE, vol. 101, no. 5, pp. 1234-1252, 2013.

[4] J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King, S. Renals, “Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis”, IEEE Trans. Audio, Speech, Lang. Process., vol. 17, no. 6, pp. 1208-1230, 2009.

[5] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda, “The HMM-based speech synthesis system version 2.0”, Proc. 6th ISCA Speech Synthesis Workshop, 2007. Online: http://hts.sp.nitech.ac.jp

[6] J. Yamagishi, C. Veaux, S. King, S. Renals, “Speech synthesis technologies for individuals with vocal disabilities: voice banking and reconstruction“, Acoustical Science & Technology, vol. 33, pp.1-5, 2012.

[7] J. Dines, H. Liang, L. Saheer, M. Gibson, W. Byrne, K. Oura, K. Tokuda, J. Yamagishi, S. King, M. Wester, T. Hirsimaki, R. Karhila, M. Kurimo, “Personalising Speech-to-Speech Translation: Unsupervised Cross-lingual Speaker Adaptation for HMM-based Speech Synthesis”, Computer Speech and Language, vol. 27, no. 2, pp. 420-437, 2013.

[8] D. Erro, T.C. Zorila, Y. Stylianou, E. Navas, I. Hernaez, “Statistical Synthesizer with Embedded Prosodic and Spectral Modifications to Generate Highly Intelligible Speech in Noise”, Proc. Interspeech, pp. 3557-3561, 2013.

Objective

The primary goal of this project is the design and development of a web interface that allows non-expert users to get their own personalized HTS-compatible synthetic voice. This interface could be used, for instance, by people suffering from degenerative pathologies before the symptoms are visible or before surgery to create a “backup” of their voice. To reach a broad audience, the system is intended to be multilingual as far as possible.

Available material

A preliminary version of this interface is already being developed at UPV/EHU. Therefore, the team is not expected to start from scratch but to improve and continue the work in progress, with special emphasis on including new functionalities.

The basic tools for developing the synthesis system will be based on HTS. A multilingual TTS (AhoTTS [9]) and related tools in Iberian languages (Spanish, Basque, Catalan, Galician) plus English will be provided by the team leaders.

The project does not involve recording new speech material. Speech databases with appropriate permissions (at least for research) in the aforementioned languages will be used. Participants are encouraged to bring databases together with HTS-compatible text analyzers in their own language to enrich the system.

[9] A. Alonso, I. Sainz, D. Erro, E. Navas, I. Hernaez, “Sistema de conversión texto a voz de código abierto para lenguas ibéricas”, Procesamiento del Lenguaje Natural, vol. 51, pp. 169-175, 2013.

Resources needed

The only requirements are (i) a web server to allocate the system along with the corresponding adaptation engine and (ii) standard microphones and headphones to test the system at the different stages of the development. The participants are expected to work on their own laptops.

Work plan

The work can be split into the following packages:

WP0 – Management

Global supervision.

WP1 – Offline system setup

The objective of this WP is providing the system with the elements needed to start operating. It involves two main tasks: (i) offline training of initial (possibly average) synthetic voices to be adapted to the users’ voice, and (ii) preparation of phonetically balanced adaptation corpora to be recorded by the users. Issues related to new languages will be addressed within this WP, too.

WP2 - Development and integration

As its name suggests, the goal of this WP is the development and integration of the basic scripts to control the recording process, transmit the recorded utterances to the server and automatically adapt the synthetic voice to them.

WP3 - Intelligent recording

The goal of this WP is the design of a graphical user interface and protocols that help detecting and rejecting unacceptable recordings (recordings exhibiting too much noise, reverb, saturation, etc.). Depending on the profiles of the participants, new functionalities will be included such as speech enhancement and verification of the recorded utterance using speech recognition and forced alignment.

WP4 – Post-editing of synthetic voices

This WP will aim at providing the users of the system with intuitive tools to “edit” the output personalized voice and modify some basic aspects of it: speaking rate, average pitch and range, vocal tract length, intensity of relevant frequency bands, etc.

WP5 – Evaluation

A formal evaluation is hard to conduct in such a short period. Therefore, this WP will aim at including specific modules to measure the users’ satisfaction and get feedback from them.

Team

Coordinators

Daniel Erro (UPV/EHU), Inma Hernáez (UPV/EHU)

Potential participants

Iñaki Sainz (UPV/EHU), Agustín Alonso (UPV/EHU), Carmen Magariños (UVigo), Igor Jauk (UPC), and people having good programming skills (C/C++, Java, scripting…) and working on one/many of the following topics:

- HMM-based speech synthesis
- Speech enhancement and/or de-reverberation
- Web development
- GUI design