Personalized Synthetic Voices for Speaking Impaired: Website and App

D. Erro[1,2], *I. Hernáez*[1], *A. Alonso*[1], *D. García-Lorenzo*[1], *E. Navas*[1], *J. Ye*[1], *H. Arzelus*[3], *I. Jauk*[4], *N.Q. Hy*[5], *C. Magariños*[6], *R. Pérez-Ramón*[1], *M. Sulír*[7], *X. Tian*[5], *X. Wang*[8]

[1]University of the Basque Country, Bilbao/Vitoria, Spain
[2]IKERBASQUE, Bilbao, Spain
[3]VicomTech-IK4, San Sebastián, Spain
[4]Technical University of Catalonia, Barcelona, Spain
[5]Nanyang Technological University, Singapore, Singapore
[6]University of Vigo, Vigo, Spain
[7]Technical University of Košice, Košice, Slovakia
[8]University of Science and Technology of China, Hefei, China

derro@aholab.ehu.es

Abstract

This paper describes the current state of the work that is being carried out in the framework of the ZureTTS project to give a personalized voice to people who cannot speak in their own. Despite the availability of tools and algorithms to synthesize speech and adapt it to new speakers, this process is affordable only for experts. To overcome this problem, we recently developed a web interface that assists users in doing so. At this point only healthy users can fully personalize the synthetic voice via adaptation, while impaired users can just manually tune a few dimensions of it. As a complement, we have launched an Android application that connects to the ZureTTS server and makes use of its functionalities in an intuitive way. Although many parts of the system need to be improved, its current version is publicly accessible and ready to be used.

Index Terms: speaker adaptation, multilingual statistical parametric speech synthesis, speaking aids

1. Introduction

The emergence of hidden Markov model (HMM) based speech synthesis [1, 2], powered by the release of the open-source HTS system [3], has enabled a wide variety of applications that were not feasible under the unit selection paradigm [4]. Among the advantages of the former, we would like to mention two because of their relevance in relation to this work: (i) its enormous flexibility for transformation or adaptation towards new speakers, emotions, styles, etc. [5, 6, 7]; (ii) its suitability for small devices (smartphones, tablet PCs, ...), given the relatively low footprint of the synthetic voices in terms of storage and the reasonable complexity of the synthesis engine.

One of the applications that have benefited the most from HMM-based speech synthesis is the design of speaking aids – in the form of text-to-speech (TTS) systems – for impaired people. Recent works [8, 9] have shown that technology is sufficiently mature to offer satisfactory performance in this field. Notably, the intelligibility scores given by a state-of-the-art HMM-based system are high in comparison to natural speech [1, 2], which provides a solid basis for a successful communication; furthermore, recent works have shown that quite simple transformations can make HMM-based synthesizers robust in very noisy environments [10]. On the other hand, the possibility of tuning the voice of the system according to the user's expectations or desires can be a good motivator toward technology acceptance by real-world users, as happens in other domains such as web search, e-commerce, and many others. In that sense, the flexibility of HMM-based synthesizers is a key property.

However, despite the existence of publicly accessible tools and databases, getting a personalized TTS is not straightforward at all for non-expert users. In response to this situation, the ZureTTS project [11] aimed at bringing speaker-adaptive speech synthesis technologies closer to non-expert users by means of a web interface that helps them obtain personalized synthetic voices with minimal effort and knowledge. This interface was conceived to be used, for instance, by people suffering from degenerative pathologies to create a "backup" of their voice before surgery or before the symptoms become audible; those who already cannot speak are provided with tools to manually and intuitively modify an available generic voice model. The core parts of the ZureTTS platform were developed in the framework of eNTERFACE'14, using existing technology and taking advantage of recent HTML5 functionalities. In parallel, preliminary generic voice models were created for seven languages: the four official Spanish languages (i.e. Castilian Spanish, Basque, Catalan and Galician), English, Chinese and Slovak. The server was also prepared to offer speech synthesis as a web service so that client applications can exploit the outcome of the adaptation process.

As the most logical subsequent step, we have recently developed the first version of a ZureTTS-compatible Android application specifically designed to facilitate communication to speaking impaired users, their only restriction being the need for a smartphone or a similar device. This first version, already available for download in Google Play, includes basic functionalities such as multilingual synthesis with text as input, playing of stored synthesized utterances associated to gestural inputs, basic manipulations, etc.

The next sections of this paper present both a technical and a functional overview of the ZureTTS platform (Section 2) and describe the first version of the new application that makes use of it (Section 3). Finally, we briefly describe the ongoing works and envisaged future extensions.

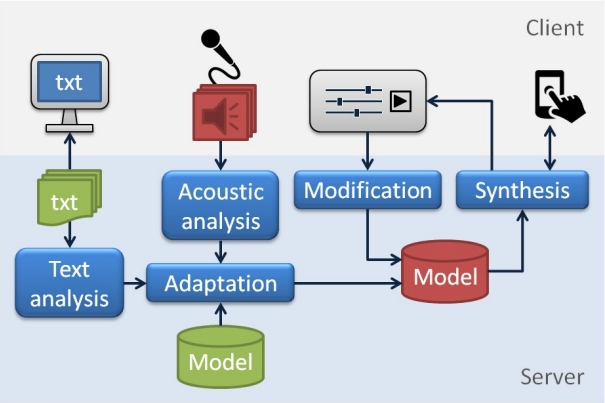

Figure 1: *Overview of the ZureTTS system.*

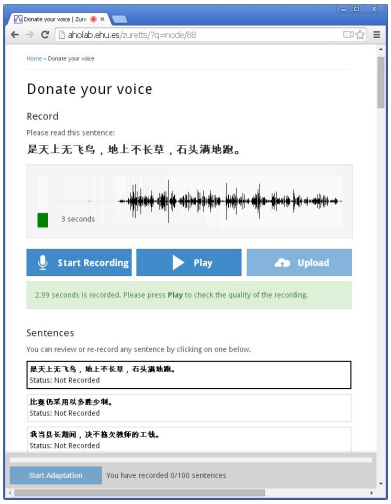

Figure 2: *Recording area of the ZureTTS website.*

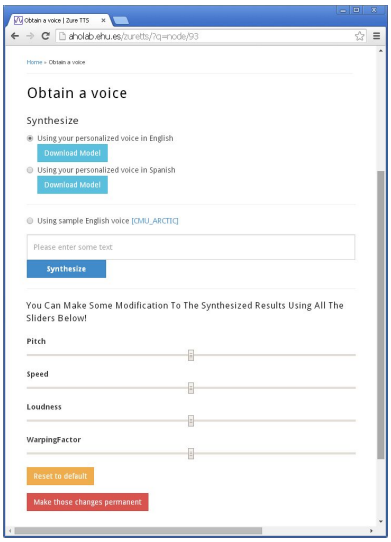

Figure 3: *User's internal area within ZureTTS website where generic or adapted voices can be manipulated.*

2. Overview of ZureTTS Platform

2.1. General architecture

The architecture of the ZureTTS platform is shown in Figure 1. A web interface communicates the client/user with a server that contains: the texts to be read and recorded by the user, a generic voice model in the target language, text and acoustic analysis tools (exactly those used to train the mentioned model), and scripts to adapt voice models to new data or transform them in several ways.

The flow of the interaction is the following. When a registered user accesses the web portal and selects a language for recording, the server sends a number of phonetically balanced sentences (about 100) to the client's screen. Each sentence has to be recorded and validated by the user, then transmitted to the server (see Figure 2). When the recording process has been completed, the server extracts acoustic information from the recordings and contextual information from their corresponding texts. Then, it adapts the baseline voice model to the user's data and lets the user try and synthesize some utterances from the adapted model. The user is also allowed to manually manipulate some of its properties: mean f_0 level, speed, loudness, and vocal tract length, among others (see Figure 3). When the user approves the manipulated synthetic voice, the changes are made permanent and the final model is stored in the user's internal zone.

The synthesis module of ZureTTS can be accessed by external applications, which can get synthetic speech in any of the offered languages and even manipulate basic synthesis parameters (f_0, speed, perceptual loudness and vocal tract length).

2.2. Implementation details

The front-end website is written in HTML, Javascript and CSS. It contains a plugin-free voice recorder based on the new Web Audio API of HTML5, which is supported by the latest versions of the most popular internet browsers. The back-end web service is written mainly in PHP, so as to be easily integrated within a Drupal framework. All tasks involving any significant computational load (i.e. text/acoustic analysis, adaptation, synthesis and modification) run on the server side, while only the recordings are made on the user's computer.

The interaction between the server and the client applications is based on the XML-RPC protocol: client applications send messages in XML format to the XML-RPC server of ZureTTS, which in turn sends XML responses containing encoded audio signals.

As for the core algorithms, the adaptation scripts are based on version 2.2 of HTS [3]. HTS models speech at phone level through 5-state context-dependent left-to-right hidden semi Markov models (HSMMs) [12]. Each state is characterized by a single multivariate Gaussian emission distribution defined by its mean vector and its diagonal covariance matrix. Discontinuous variables, such as the local fundamental frequency of the signal, are handled by means of multi-space distributions (MSD) [13]. Input vectors are assumed to include 1^{st} and 2^{nd}-order dynamics. The global variance of the acoustic parameters is modeled together with the parameters themselves [14]. When multiple voices are used to train a speaker-independent model (normally called an average voice model), speaker-adaptive training [15] is performed; to later adapt this model to new data, CMLLR and MAP algorithms are applied [16, 17]. It is worth mentioning, however, that these HMM adaptation algorithms are being pro-

gressively replaced by new standalone HSMM adaptation ones based on a mean squared error minimization criterion.

Several front-ends of existing TTS systems are used as language-specific text analyzers, as listed in Table 1; in the particular case of Chinese, a basic text analyzer was developed from scratch (see [11] for more details). In this early version of the system, the initial voices models were trained from available databases (shown in Table 1). Therefore, not all of them are average models. The Chinese and Galician databases were generously provided by iFlyTek and UVigo, respectively. A rapid version of Ahocoder [18] is used as high-quality acoustic analyzer for all languages. Ahocoder parameterizes speech frames into three different streams: $log - f_0$, Mel-cepstral representation of the spectral envelope, and maximum voiced frequency as a degree-of-harmonicity measure. The synthesis engine of ZureTTS includes a standard parameter generation module followed by the waveform reconstruction module of Ahocoder. Finally, the voice model manipulation tools apply a linear transformation defined from the user's input to the means and covariances of all states in the appropriate HSMM. For instance, f_0 increments are applied by summing the logarithm of the desired factor to the $log - f_0$ model means, whereas loudness and vocal tract length can be manipulated through an additive term and a multiplicative matrix in the cepstral domain, respectively [10, 19].

Table 1: *Tools and databases for each language.*

Lang.	Txt. analyzer	Database
English	Festival [20]	CMU Arctic [21]
Slovak	Festival [20]	[22]
Catalan	Festival [20]	Festcat [23]
Spanish	AhoTTS [24]	Albayzin [25]
Basque	AhoTTS [24]	AhoSyn [26]
Galician	Cotovia [27]	From UVigo
Chinese	New	From iFlyTek

3. A New Client Application

The speaking impaired community is heterogeneous in terms of age and access to technology. However, the recent development of the smartphone industry and of wireless connection networks has brought technology to the hands of a vast majority of the population. Thus, in line with the goals of ZureTTS, i.e. making speech synthesis technology accessible to nonexperts with minimal effort, we have developed an Android application that can access the ZureTTS server and get synthesis services from it in any of the available languages.

Basically, the application communicates with the synthesis engine of ZureTTS by means of the XML-RPC protocol. Thus, the application acts like a client and all the processing is carried out on the server. The user can configure the call to the synthesis module so as to use either a generic voice or his/her own adapted voice. It is also allowed to use special settings regarding the mean f_0, the speed, etc. Apart from this, the application has the following functionalities:

- *Conversation mode*. It contains a set of pre-defined utterances which are supposed to be the most frequently used during a conversation: "yes", "no", "I don't know", "ok", "wait a moment, please", etc. It also keeps a record of the utterances synthesized in response the user's typed input. The record is sorted by frequency of use so that usual utterances do not have to be synthesized again.
- *Gestural input*. To avoid typing text, the user can assign a simple picture to each stored utterance. Thus, the device plays the utterance when the user draws the corresponding picture on the screen (see Figure 4, bottom-left plot).
- *Synthesis settings*. The synthesis settings can be modified at any time without altering the models in the ZureTTS server (see Figure 4, bottom-right plot).
- When there is no internet conection, the app can use Android's local TTS system, thus keeping the main functionalities active.

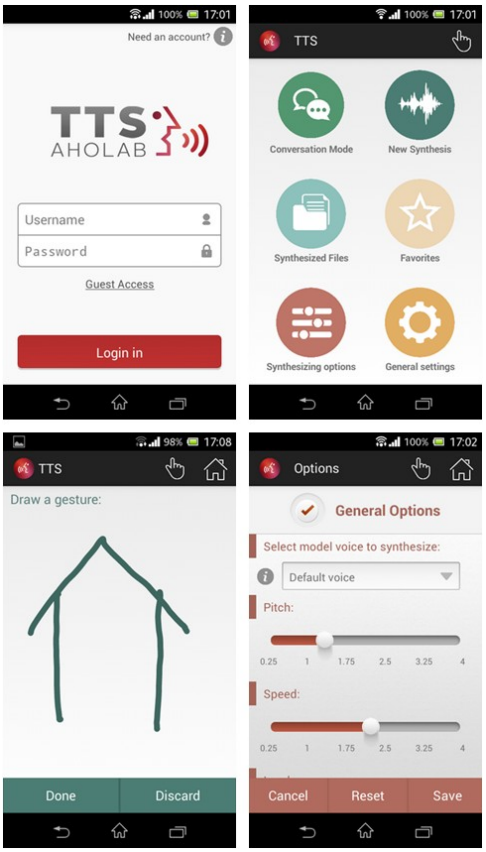

Figure 4: *Screenshots of the app. Top-left: main entrance for a registered user. Top-right: main menu. Bottom-left: gestural input. Bottom-right: some of the synthesis settings.*

The first version of this *app* is compatible with Android v4.0 (Ice Cream Sandwich) or higher and has been released in Google Play for free under the name "TTS Aholab". It was recently honored with the Start BiskayApp award 2014 (at regional level).

4. Conclusions

This paper has presented ZureTTS, a platform to obtain personalized synthetic voices in several languages with minimal effort and no expert supervision. It includes a multilingual web interface, a web service server that contains the tools and data to get everything working, and, since recently, an associated An-

droid application that provides the ZureTTS users with online personalized synthesis services.

Among the future works to be carried out in the short/mid term, we can mention the following. First, we will completely replace the adaptation scripts of HTS v2.2 by novel HSMM adaptation algorithms that are currently in the final stage of development. Given the social focus of the work, we are also conducting research to adapt synthetic voices to incomplete or corrupted data, i.e. speech recordings from disordered voices with audible symptoms but no total impairment. Second, we are open to incorporating new languages to the system, as long as suitable TTS front-ends and speech databases can be found. The Chinese text analyzer requires further work as well since it was developed from scratch in a very short period. It is also necessary to improve the underlying voice models of the system (average or not), especially for some particular languages, which will surely lead to immediate perceptual quality improvements. In addition, we intend to use the platform to investigate cross-lingual adaptation algorithms, which would allow users to record speech in one language and get personalized synthetic voices in multiple languages. Robustness against noisy adaptation data is also an issue that deserves attention. Finally, we want to improve the catalog of existing voices and the way it is shown. Regarding the Android application, we plan to include standard pictograms as input buttons to facilitate its use in daily communicative situations. We will also conduct a formal evaluation in a clinical environment.

5. Acknowledgements

This work has been partially funded by the Spanish Ministry of Economy and Competitiveness (SpeechTech4All, TEC2012-38939-C03; DIACEX, FFI 2012-31597; FPI program), with FEDER support, and the Basque Government (ZURETTS+, S-PE11UN081). The ZureTTS website was developed during eNTERFACE'14 ISCA Training School, sponsored by ISCA, EURASIP, RTTH, the local and regional Council of Bilbao and Metro-Bilbao.

6. References

[1] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[3] "Hidden Markov model based speech synthesis system (HTS)." [Online]. Available: http://hts.sp.nitech.ac.jp

[4] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, 1996, pp. 373–376.

[5] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and 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.

[6] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech Communication*, vol. 52, no. 5, pp. 394–404, 2010.

[7] N. Obin, P. Lanchantin, A. Lacheret, and X. Rodet, "Discrete/continuous modelling of speaking style in HMM-based speech synthesis: Design and evaluation," in *Proc. Interspeech*, 2011, pp. 2785–2788.

[8] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1–5, 2012.

[9] S. Creer, S. Cunningham, P. Green, and J. Yamagishi, "Building personalised synthetic voices for individuals with severe speech impairment," *Computer Speech & Language*, vol. 27, no. 6, pp. 1178–1193, 2013.

[10] D. Erro, T.-C. Zorila, and Y. Stylianou, "Enhancing the intelligibility of statistically generated synthetic speech by means of noise-independent modifications," *IEEE/ACM Trans. Audio, Speech, & Lang. Process.*, vol. 22, no. 12, pp. 2101–2111, 2014.

[11] D. Erro, I. Hernáez, E. Navas, A. Alonso, H. Arzelus, I. Jauk, N. Q. Hy, C. Magariños, R. Pérez-Ramón, M. Sulír, X. Tian, X. Wang, and J. Ye, "ZureTTS: Online platform for obtaining personalized synthetic voices," in *Proc. eNTERFACE'14*, 2014.

[12] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 825–834, 2007.

[13] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. Syst.*, vol. E85-D, no. 3, pp. 455–464, 2002.

[14] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[15] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: a maximum likelihood approach to speaker normalization," in *Proc. ICASSP*, vol. 2, 1997, pp. 1043–1046.

[16] G. M.J.F., "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75–98, 1998.

[17] J. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Trans. Speech & Audio Process.*, vol. 2, no. 2, pp. 291–298, 1994.

[18] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal Sel. Topics in Signal Process.*, vol. 8, no. 2, pp. 184–194, 2014.

[19] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech & Audio Processing*, vol. 13, pp. 930–944, 2005.

[20] A. Black, P. Taylor, and R. Caley, "The Festival speech synthesis system: System documentation," 1997. [Online]. Available: http://www.cstr.ed.ac.uk/projects/festival

[21] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223–224.

[22] M. Sulír and J. Juhár, "Design of an optimal male and female slovak speech database for HMM-based speech synthesis," in *Proc. 7th Int. Workshop on Multimedia and Signal Process.*, 2013, pp. 5–8.

[23] A. Bonafonte, J. Adell, I. Esquerra, S. Gallego, A. Moreno, and J. Pérez, "Corpus and voices for Catalan speech synthesis," in *Proc. LREC*, 2008, pp. 3325–3329.

[24] I. Sainz, D. Erro, E. Navas, I. Hernáez, J. Sánchez, I. Saratxaga, I. Odriozola, and I. Luengo, "Aholab speech synthesizers for albayzin2010," in *Proc. FALA'2010*, 2010, pp. 343–348.

[25] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, and J. B. M. no, "Albayzin speech database: design of the phonetic corpus," in *Proc. 3rd European Conf. on Speech Commun. and Tech.*, 1993, pp. 175–178.

[26] I. Sainz, D. Erro, E. Navas, I. Hernáez, J. Sánchez, I. Saratxaga, and I. Odriozola, "Versatile speech databases for high quality synthesis for Basque," in *Proc. 8th Int. Conf. on Language Resources and Eval.*, 2012, pp. 3308–3312.

[27] E. Rodríguez-Banga, C. García-Mateo, F. J. Méndez-Pazó, M. González-González, and C. Magariños-Iglesias, "Cotovia: an open source TTS for Galician and Spanish," in *Proc. IberSpeech*, 2012.