ZureTTS: Online Platform for Obtaining Personalized Synthetic Voices

Daniel Erro*, Inma Hernáez*, Eva Navas*, Agustín Alonso, Haritz Arzelus, Igor Jauk, Nguyen Quy Hy,
Carmen Magariños, Rubén Pérez-Ramón, Martin Sulír, Xiaohai Tian, Xin Wang and Jianpei Ye

Abstract—**The primary goal of the ZureTTS project was the design and development of a web interface that allows non-expert users to get their own personalized synthetic voice with minimal effort. Based on the increasingly popular statistical parametric speech synthesis paradigm, the system was developed simultaneously for various languages: English, Spanish, Basque, Catalan, Galician, Slovak, Chinese, and German.**

Index Terms—**Statistical parametric speech synthesis, speaker adaptation.**

I. Introduction

SPEECH synthesis technologies have evolved during the last decade from selection and concatenation based paradigms [1] to statistical parametric ones [2], [3]. The main advantages of hidden Markov model (HMM) based speech synthesis are its enormous flexibility for speaker/style adaptation [4], the low footprint of the voice models (those of a high-quality voice can be stored in less than 5 MB in some cases!), the ability of generating smooth synthetic signals without annoying discontinuities, etc. Importantly, the availability of an open source statistical parametric speech synthesis system, HTS [5], has played a key role in this technological revolution. Statistical parametric speech synthesis has enabled many new applications that were not possible in the previous technological frameworks, such as the design of aids for people with severe speech impairments [6], personalized speech-to-speech translation [7], and noise-robust speech synthesis [8].

In parallel, the market of hand-held devices (smartphones, tablet PC's, etc.) has grown substantially, together with their capabilities. As a result, communication barriers are diminishing while new ways of entertainment are emerging. In these two contexts, low-footprint personalized text-to-speech (TTS) synthesizers arise as a very interesting outcome of years of research. Nevertheless, it is still hard for a non-expert user to get a personalized TTS. The ZureTTS project bridges the gap between users and speaker-adaptive speech synthesis technology by providing a web interface that helps them obtaining personalized synthetic voices in an intuitive and automatic manner. This interface can be used, for instance, by people suffering from degenerative pathologies to create a "backup" of their voice before surgery or before the symptoms are noticeable, or by anyone wishing a GPS to speak in his/her own voice or that of a relative.

The ZureTTS project was undertaken by an international team of researchers during eNTERFACE'14, covering up to 8 languages: English, Spanish, Basque, Catalan, Galician, Slovak, Chinese, and German. This article contains a detailed description of the project and the way it was developed. Section II presents a general overview of the technology and roughly describes the system in relation to it; Sections III, IV and V go into the details of the different tasks accomplished during the development; Section VI discusses some open issues and future perspectives, and it summarizes the main conclusions.

II. General framework

A. Technological framework

Fig. 1 shows the general diagram of a speaker adaptive HMM-based speech synthesis system such as HTS [5], which is the core of ZureTTS. Basically, HTS provides the tools to (i) learn a global statistical correspondence between labels extracted from text and features extracted from the acoustic realization of that text, (ii) adapt the trained statistical models to new incoming data from a particular speaker [4] and (iii) generate the most likely sequence of acoustic features given a specific set of labels [9] (these tools correspond to the blue blocks in Fig. 1).

The so-called labels are phonetic, linguistic or prosodic descriptors of the text, which are stored in a specific format that the HTS engine is able to understand. The information they contain is typically related to phones (code of the current one and those in the surroundings, time location, position in syllable, position in word, etc.), syllables (accent, position in word, position in phrase, etc.), words, phrases, pauses... The label extraction process (green block in Fig. 1) is obviously

*Project leaders. The remaining co-authors are equal contributors and have been listed in strict alphabetical order.

D. Erro, I. Hernáez, E. Navas, A. Alonso and J. Ye are with Aholab, University of the Basque Country, Alda. Urquijo s/n, 48013 Bilbao, Spain (contact e-mail: derro@aholab.ehu.es). D. Erro is also with IKERBASQUE, Basque Foundation for Science, 48013 Bilbao, Spain.

H. Arzelus is with VicomTech-IK4, Paseo Mikeletegi 57, Parques Tecnológicos de Euskadi - Gipuzkoa, 20009 San Sebastian, Spain.

I. Jauk is with the VEU group, Teoria Senyal i Comunicacions, Technical University of Catalonia, C/ Jordi Girona 1-3, 08034 Barcelona, Spain.

N.Q. Hy and X. Tian are with NTU-UBC Research Centre of Excellence in Active Living for the Elderly, Nanyang Technological University, 639798 Singapore.

C. Magariños is with Multimedia Technology Group, AtlantTIC, University of Vigo, Campus Lagoas-Marcosende s/n, 36310 Vigo, Spain.

R. Pérez-Ramón is with Laslab, University of the Basque Country, Facultad de Letras, Paseo de la Universidad 5, 01006 Vitoria, Spain.

M. Sulír is with Department of Electronics and Multimedia Communications, Technical University of Košice, Letná 9, 040 11 Košice, Slovakia.

X. Wang is with the National Engineering Laboratory of Speech and Language Information Processing, University of Science and Technology of China, Hefei, 230027, China.

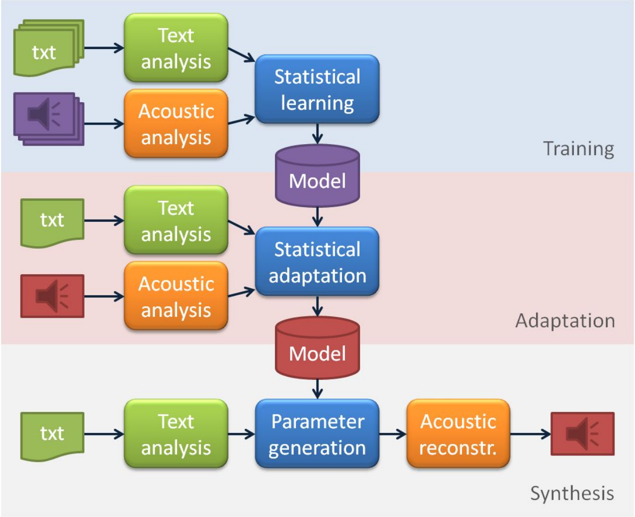


Fig. 1. General diagram of a speaker-adaptive statistical parametric speech synthesis system.

language-dependent. In principle, the front-end of any existing TTS system can be used to accomplish this task as long as the information it handles can be translated into the appropriate label format. Since this project deals with many languages, many language-specific text analyzers were used as will be explained in Section IV.

Vocoders are used to extract acoustic features from audio signals and also to reconstruct audio signals from acoustic features (orange blocks in Fig. 1). The statistical learning process implies the use of appropriate acoustic parameterizations exhibiting not only good speech analysis/reconstruction performance but also adequate mathematical properties. A typical state-of-the-art vocoder extracts acoustic features at three different levels: logarithm of the fundamental frequency (f_0), Mel-cepstral (MCEP) or linear prediction related representation of the spectral envelope, and degree of harmonicity of different spectral bands.

The statistical model that results from the training stage of a speaker-adaptive system (purple model in the top part of Fig. 1) is often an "average voice model" [4]. In other words, it is learned from data (recordings + texts) from a large number of speakers so that it covers the variability not only of the language but also of its many possible acoustic realizations. These kind of models are more easily adapted to a few recordings of a new unknown speaker than those obtained from one only speaker.

B. *Performance flow*

The overall structure and performance of the system is graphically described in Fig. 2 (for clarity, we have used the same colors as in Fig. 1). The users interact remotely with the system through a website, and the system itself is hosted on a server that carries out the core tasks and contains the necessary data.

1) The client side: Roughly speaking (see next section for more detailed information), the ZureTTS website allows registered users to record their voice in any of the available languages, visualize the recordings and listen to them for validation, transmit the recordings to the server, and tell the server to start the adaptation process. When the training process finishes, the website gives the user access to the artificial voice so that he/she can validate it or manually modify some intuitive aspects of it (average pitch, vocal tract length, speaking rate, loudness...) with an appropriate acoustic feedback. When the user approves the voice with the desired manual modifications, the website presents the final models of the synthetic voice for download.

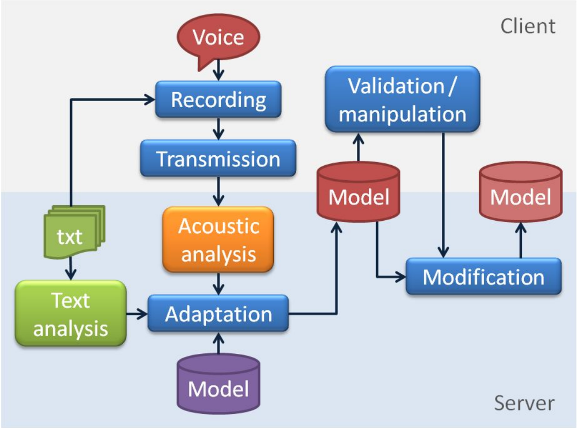
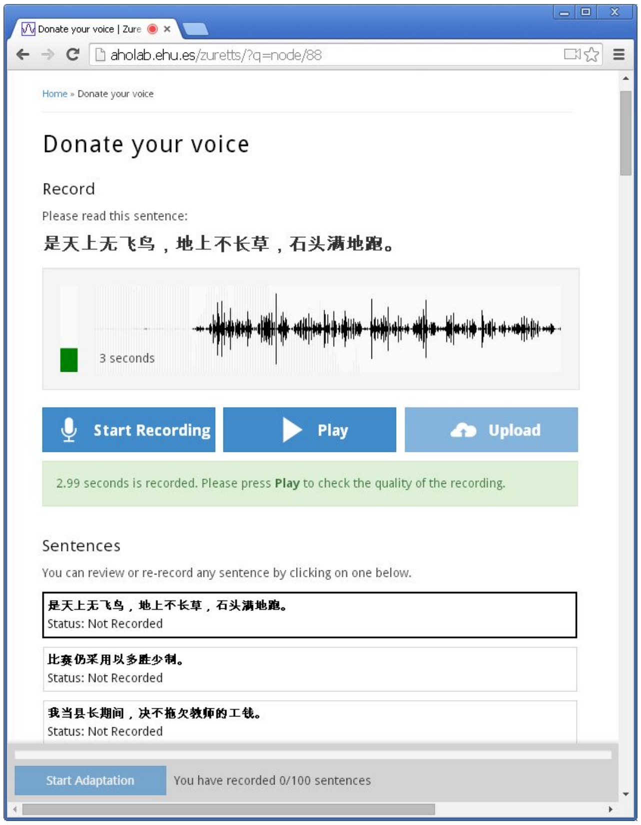


Fig. 2. General diagram of the system.

2) The server side: For each available language, the server contains an initial statistical model (either the model of a generic voice or, preferably, an average voice model [4]), the text analyzer and the vocoder that were used during the training of the initial voice, the scripts for adaptation and modification of models, and a phonetically balanced set of sentences to be recorded by the user. When a particular user selects a language, the server sends him/her a set of sentences (about 100) so that the recording process can be started. Then, when the user transmits the recordings to the server, the server carries out the adaptation (similarly as in the mid part of Fig. 1) and yields a "personalized" model. Since the adaptation process may take some time (depending on the size of the initial model and also on the concentration of user requests), the server alerts the user via e-mail when it has finished. If the user asks for any voice modification through the website after having previewed the result, the server embeds such modification into the model itself, making it permanent. Finally, it stores the "modified personalized" model in the user's internal zone so that it can be either downloaded or used within the ZureTTS portal. Last but not least, the server hosts a web service that allows an external application to connect and synthesize speech from a particular model.

The next three sections describe the work we conducted to develop each part of the ZureTTS system, namely the website itself, the initial voice models and text analyzers for each language under consideration, and the synthetic voice manipulation procedures.



Donate your voice

Record

Please read this sentence:

是天上无飞鸟，地上不长草，石头满地跑。

3 seconds

2.99 seconds is recorded. Please press **Play** to check the quality of the recording.

Sentences

You can review or re-record any sentence by clicking on one below.

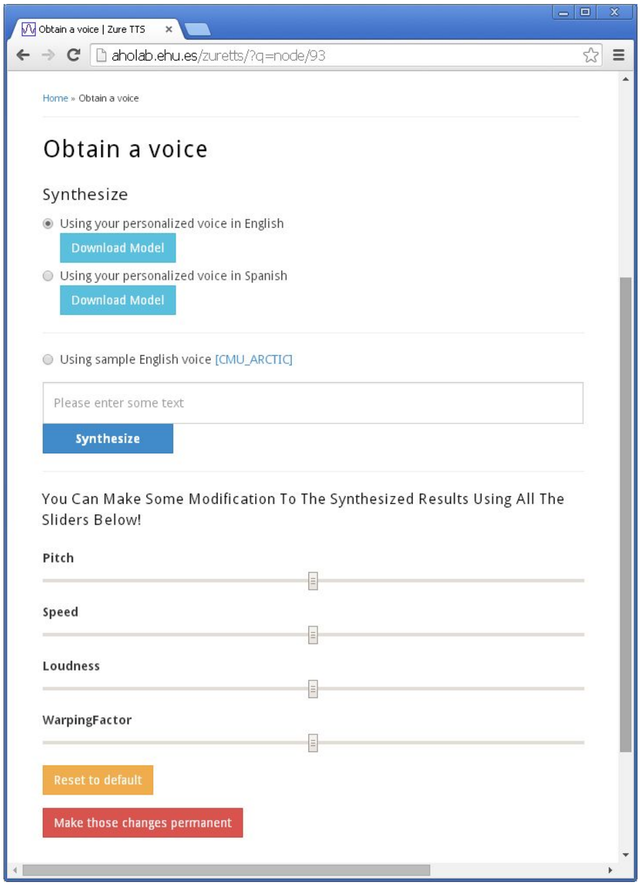
是天上无飞鸟，地上不长草，石头满地跑。
Status: Not Recorded

比赛仍采用以多胜少制。
Status: Not Recorded

我当县长期间，决不拖欠教师的工钱。
Status: Not Recorded

You have recorded 0/100 sentences

Fig. 3. The recording interface for Chinese language.



Obtain a voice

Synthesize

● Using your personalized voice in English

○ Using your personalized voice in Spanish

○ Using sample English voice [CMU_ARCTIC]

Please enter some text

You Can Make Some Modification To The Synthesized Results Using All The Sliders Below!

Fig. 4. Getting and trying personalized or generic artificial voices in different languages.

III. The ZureTTS Website

A. Functional description

The website includes both a public access area, providing information about the project and registration forms, and a private area available only to registered users. In this private area, two functional parts can be distinguished: the *"Donate your voice"* area (depicted in Fig. 3) and the *"Obtain a voice"* area (depicted in Fig. 4).

1) Donation area: When a registered user accesses the voice donation area (Fig. 3), he/she is firstly asked to choose the target language. Then, a speech recorder and a list of about 100 sentences are displayed on the screen. The user is expected to read aloud the sentences and record them in a silent environment. The recorder was designed to give the user useful visual feedback to help him/her control the level of the signals, thus avoiding saturation. In the current implementation, the user is forced to listen to each recorded sentence for validation before transmitting it to the server. By communicating with the central database, the system controls the number of uploaded/remaining recordings. When all of them have been received by the server, the user is given the chance to start the adaptation of the underlying voice models to the samples of his/her voice by pressing the *"Start adaptation"* button.

2) Voice obtaining area: This area is normally accessed when a user receives via e-mail the confirmation that the adaptation process has finished. As can be seen in Fig. 4, the user can then synthesize speech using his/her personalized synthetic voice in any of the languages for which the necessary utterances have been recorded and the adaptation has finished successfully (as an example, in Fig. 4 there are two personalized voices available: English and Spanish). Some generic synthetic voices are also available for registered users who do not want to have a personalized one (in Fig. 4 there is one generic voice in English). A number of modifications are also available in case the user wants to tune some intuitive aspects of the synthetic voice (more details are given in section V). These modifications are applied through sliders, and for moderate factors they do not alter the perceived identity of the voice significantly. Each time the *"Synthesize"* button is pressed, the system synthesizes the message written in the text area using the selected voice model with the modifications given by the current positions of the sliders. Once the modifications are made permanent by pressing the *"Make those changes permanent"* button, the user's voice models are overwritten and replaced by the modified ones.

B. Technical description

The frontend website was written in HTML, Javascript and CSS. It leverages the new Web Audio API of HTML5 to implement a plugin-free voice recorder directly on the webpage; currently this API has been adopted by several major web browsers such as Google Chrome and Mozilla Firefox and will be supported in the next versions of Microsoft

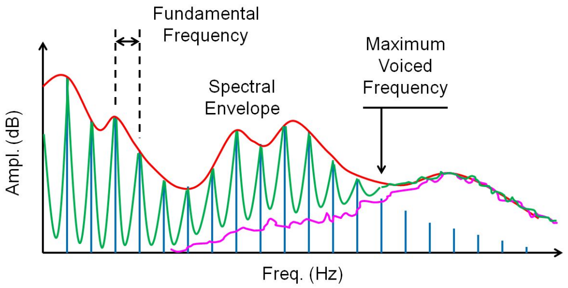


Fig. 5. Assuming that a given speech frame is the sum of a harmonic component (blue) and a noise component (magenta), which occupy the lower and the upper band of speech respectively, Ahocoder measures the log-f_0, the MCEP representation of the spectral envelope (red), and the maximum voiced frequency, defined as the boundary between harmonic and noisy bands.

Internet Explorer. The synthesis page also leverages several capabilities of HTML5 such as audio tag, range input for better compatibility with new browsers and mobile devices. Several popular Javascript and CSS libraries are also used in the frontend such as Jquery, Knockoutjs and Bootstrap.

The backend webservice was mainly written in PHP, so as to be integrated with the mature user management system of Drupal. The webservice is responsible for generating JSON data (recording status, etc.) for the web interface, accepting inputs from the web interface and coordinating different other processing scripts for data collection, adaptation and synthesis.

With an aim towards extending the usability of the results of ZureTTS, we also implemented a multilingual-TTS web service based on the XML-RPC protocol by exploiting the potential of the Drupal framework. Thus, any XML-RPC client (an external application, for instance) can synthesize speech in the offered languages and even manipulate its parameters (as in Fig. 4) just by sending messages in the appropriate XML format to the XML-RPC sever of ZureTTS, which in turn sends XML responses containing encoded audio signals.

IV. Language-Specific Voices Models and Tools

This task of ZureTTS involved the development of initial (average when possible) voice models and text processing tools for all the languages under consideration. A common framework was assumed in terms of HMM topology and acoustic analyzer (vocoder). This approach enabled a more homogeneous implementation of the different functionalities of the system and facilitated some operations, such as those involved in the voice post-editing process (which will be explained in Section V). Excluding these acoustic modeling aspects, voices and tools are language-dependent, so the subsections below describe how the problem was tackled for each language.

The model topology we used is the "standard" one in HTS v2.2, namely 5-state context-dependent left-to-right hidden semi Markov models (HSMMs = HMMs with explicit state duration distributions) [10] where each state is characterized by a multivariate Gaussian emission distribution given by a mean vector and a diagonal covariance matrix. When dealing with parameters that can take discontinuous values, such as the local fundamental frequency of the signal, multi-space distributions (MSD) [11] were used. The vectors used to train the model were appended 1^{st} and 2^{nd}-order dynamics, and the global variance of the acoustic parameters was modeled together with the parameters themselves [9].

The vocoder we used in ZureTTS is called Ahocoder [12]. As depicted in Fig. 5, it handles three streams of acoustic information: log-f_0, MCEP coefficients, and the so-called maximum voiced frequency, which stands for the local degree of harmonicity of the signal. The relationship between MCEP coefficients $\{c_i\}_{i=0...p}$ and spectral envelope $S(\omega)$ can be formulated as follows[1]:

$$\log S(\omega) = \sum_{i=0}^{p} c_i \cos(i \cdot \text{mel}(\omega)) \qquad (1)$$

where $\text{mel}(\omega)$ is the Mel-scaled version of ω. All speech signals were digitized at 16 kHz sampling rate and were analyzed/reconstructed at 5 ms frame shift. The order of the MCEP parameterization was always equal to 39, which resulted in 42 static parameters per frame (plus their dynamics).

A. English

An average voice model for English was trained from the 7 voices (2 female + 5 male, 1132 utterances each) in the CMU ARCTIC database [13], similarly as in the HTS demo scripts [14] (except for the vocoder).

The text analysis tools were taken from the well known Festival speech synthesis system [15], developed by the University of Edinburgh.

B. Spanish

In order to train an average voice model for Spanish, we used the "phonetic" subset of the Albayzin speech database [16]. It contains a total of 6800 utterances and 204 different speakers, each of which recorded either 160, 50 or 25 utterances. The phonetic segmentation of the database was not available, so it was carried out automatically via forced alignment of HMMs using HTK [17].

The text analysis tool we employed was taken from AhoTTS, the open-source TTS system developed by Aholab [18]. The structure of the HTS labels was similar to the one described in [19].

C. Basque

The average voice model was trained from the databases described in [20]. They contain a total of 9 voices (5 female, 4 male), all of which include 1 hour of speech except for two (female and male) which include 6 hours each. It is noteworthy that due to time constraints, only half of the voices were used to train the initial model.

Similarly as in Spanish, the front-end of AhoTTS [18] was used for text analysis and label extraction according to the specifications in [19].

[1] Some authors include a multiplicative factor 2 for $i > 0$. In this case such factor is omitted to make the MCEP coefficients compatible with some of the transformations described in Section V.

D. Catalan

Thanks to the Festcat project [21], an open database of 10 different voices (5 female, 5 male) was available for Catalan language [22]. The amount of speech material was 10 hours for two of the voices therein (female and male) and 1 hour for the remaining ones. The speakers were all professional. As well as for Basque, due to time constraints, only half of the voices were used during training.

For text analysis we used a Catalan front-end compatible with Festival [15] that had been released in the framework of the Festcat project [21] too. It included word normalization, a lexicon, a letter-to-sound (L2S) converter and a part-of-speech (POS) tagger.

E. Galician

Due to the lack of freely available speech synthesis databases for Galician languages, we used a single voice provided by the University of Vigo to train the initial model of the system. It contained 1316 utterances (1 hour 13 minutes of speech) recorded by a professional male speaker.

For text analysis, we integrated the front-end of Cotovia [23], the TTS system developed by GTM, University of Vigo.

F. Slovak

The initial average voice model was trained from a total of 17903 utterances (more than 36 hours of speech). This material contained 18 different voices taken from several databases:

- A big Slovak speech database composed by 4526 phonetically balanced sentences, all of them spoken by two different speakers, female and male (about 6 hours of speech per speaker), and recorded in a professional studio. This database had been specifically designed for synthesis purposes [24].
- A smaller database containing 330 phonetically balanced sentences recorded by both a female and a male speaker (40-50 minutes of speech each).
- A speech recognition database with 14 different voices (7 female and 7 male) and a variable number of utterances per speaker (between 469 and 810, 80-140 minutes).

Similarly as for English and Catalan, the Slovak text analyzer used in this project was based on Festival [15]. It included a big pronunciation dictionary containing about 150k problematic Slovak words (a word is considered problematic when there is a mismatch between the written form and its corresponding canonical pronunciation), a rule-based L2S conversion system and a set of token-to-word rules for basic numerals (from zero to several billion).

G. Chinese

Two databases generously provided by iFLYTEK Co. Ltd. were utilized to train a Chinese average voice model. Both databases had been recorded in neutral reading style with 16 kHz sampling rate and 16 bits per sample. The first one contained 1000 utterances (110 minutes) from a male speaker, the average utterance duration being 6.6 seconds. The second one contained 1000 utterances (186 minutes) from a female speaker, with average duration 11.2 seconds. The texts of the two databases were different.

For a correct performance of the TTS system, the Mandarin Chinese text analyzer must parse the input text into a composite structure where not only the phoneme and tone for every syllable but also the prosodic structure of the whole sentence is specified. Unfortunately, such text analyzer is not currently available; thus we built a Mandarin Chinese text analyzer for this project. The text analysis procedure includes: (i) word segmentation, (ii) POS tagging, (iii) grammatical parsing, (iv) L2S conversion, and (v) prosodic prediction. The last two steps are parallel but they both depend on the results of the first three steps.

To implement steps (i)-(iii), an open-source parser called ctbparser [25] was utilized. This parser is based on conditional random fields (CRF). The provided CRF models for word segmentation, POS tagging, and grammatical parsing were trained on The Chinese TreeBank [26] and have been reported to achieve good results on the three tasks [25].

For the L2S conversion, every Chinese character (we assume it as one syllable) must be converted into the composition of phonemes and tone, or pinyin. This conversion can be implemented through a search in a pronunciation dictionary. However, Chinese is well known for polyphones: the pronunciation of one syllable may differ in different contexts. Therefore, we built a hierarchical dictionary and adopted a simple search strategy: if the current grammatical unit, such as a word or phrase, can be found in the dictionary, the corresponding pinyin sequence is used; otherwise, the pinyin of all the syllables of the grammatical unit are retrieved and concatenated into a pinyin sequence.

Getting the pinyin sequence is not enough for high-quality speech synthesis in Mandarin Chinese. Another necessary component is the prosodic hierarchy [27]: some adjacent characters should be pronounced as a single prosodic word and several prosodic words form one single prosodic phrase. Such hierarchy resembles the grammatical structure of a sentence; thus it is possible to derive the prosodic structure based on the results of the grammatical parsing mentioned above. In this project, we adopted grammatical features similar to those mentioned in [28] and used decision trees [29] to build the prosodic prediction model. The model was trained from 1900 sentences in the aforementioned corpus. Tests performed using the remaining 100 sentences showed that the performance of the prosodic prediction model achieved similar results as those reported in [28].

H. German

The German initial voice has been created by using only one voice called "Petra" [30]. The Petra corpus is an extract of the German BITS database [31]. It contains a total of 399 sentences and it was originally recorded, transcribed and segmented for the use in the BOSS speech synthesis system [32]. Thus, a set of scripts had to be implemented in order to translate the original files of the corpus (in BOSS-XML format) into an HTS-compatible label format. Basically we extracted the same type of information as in [19], with very few differences. The same scripts are used in combination with the BOSS text analyzer at synthesis time. This part of the work was not yet fully operational at the time of writing this paper.

V. User-Driven Post-Edition of Synthetic Voices

The purpose of this part of the project was to provide the algorithms to manually manipulate intuitive aspects of synthetic speech. In accordance with the interface described in Section III, the modifications must be applicable: a) on the acoustic parameters of a given signal, and b) on the models that generate such parameters. When performing modifications on the acoustic parameters, the user can listen to the modified version of the signal and tune the parameters according to his/her desires until the results are satisfactory. At that moment, the system will "print" the new values on the HMM, thus making them permanent. In this regard, linear transforms are optimal: given a set of acoustic vectors $\{\mathbf{x}_t\}$ generated by an HMM, we can either transform directly the vectors, $\hat{\mathbf{x}}_t = \mathbf{A}\mathbf{x}_t + \mathbf{b}$, or transform the mean vectors and covariance matrices of all the HMM states:

$$\hat{\boldsymbol{\mu}} = \check{\mathbf{A}}\boldsymbol{\mu} + \check{\mathbf{b}} \ \ , \quad \hat{\boldsymbol{\Sigma}} = \check{\mathbf{A}}\boldsymbol{\Sigma}\check{\mathbf{A}}^\top \tag{2}$$

where

$$\check{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A} \end{bmatrix} \ , \quad \check{\mathbf{b}} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \tag{3}$$

Note that the block structure of $\check{\mathbf{A}}$ and $\check{\mathbf{b}}$ allows transforming both the static and the dynamic parts of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. For consistency, the global variance models are also transformed through eq. (2) for null bias vector $\check{\mathbf{b}}$ and for $\check{\mathbf{A}}$ equal to the element-wise product of $\mathbf{A}$ by itself.

Focusing on spectral modifications, the choice of MCEP coefficients as acoustic parameterization is particularly useful in this context because it has been shown that, in the MCEP domain, frequency warping[2] and amplitude scaling[3] operations can be formulated as a product by a matrix and an additive bias term, respectively [33], [34].

As depicted in Fig. 4, users are shown a textbox where they can type any sentence to test the trained synthetic voice. They are also presented with a set of sliders to manipulate the four aspects of voice that we describe next, and a button to make the modifications permanent.

1) Speech rate: The effect of the modification can be previewed by reconstructing the waveform from the generated parameters at a different frame rate, i.e. the frame shift is multiplied by the lengthening factor d. Modifications are imposed to the duration model by multiplying its means and covariances by d and d^2, respectively. In addition, the dynamic parts of the acoustic models are multiplied by $1/d$ (deltas) or $1/d^2$ (delta-deltas) for consistency.

2) Average pitch: This modification was implemented through a linear transform of the $\log f_0$ stream where $\mathbf{A} = [1]$ and $\mathbf{b} = [\log \kappa]$, κ being the selected pitch factor.

3) Vocal tract length: Among the existing types of parametric frequency warping curves [33], we chose the one based on the bilinear function as in [34]. In the MCEP domain, the corresponding transformation matrix $\mathbf{A}$ takes only one parameter α as input (see [34] and [35] for details):

$$\mathbf{A} = \begin{bmatrix} 1 & \alpha & \alpha^2 & \dots \\ 0 & 1-\alpha^2 & 2\alpha - 2\alpha^3 & \dots \\ 0 & -\alpha + \alpha^3 & 1 - 4\alpha^2 + 3\alpha^4 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{4}$$

Positive values of α produce a vocal tract length reduction, i.e. move the spectral events to higher frequencies[4], and vice versa.

4) Loudness: Even without increasing the power of a signal, its perceptual loudness (and also its intelligibility) can be increased by reallocating energy in some specific bands. In [8] a filter enhancing the band 1–4 kHz was used for this exact purpose. In this work, loudness modification was implemented through a bias vector $\mathbf{b}$ that contains the MCEP representation of the mentioned filter, multiplied by a weighting factor ℓ tuned by the user.

Apart from these four simple modifications, algorithms were developed to allow more expert users to modify the spectrum of the synthetic voice via arbitrary piecewise linear frequency warping and amplitude scaling curves (for simplicity, their corresponding graphical controls have not been included in Fig. 4). These curves will be referred to as $W(\omega)$ and $A(\omega)$ respectively. Basically, the user "draws" them by choosing the position of a number of reference points in a 2D surface. To translate these curves onto the MCEP domain, they are first resampled at a sufficient resolution ($K = 257$ frequency bins between $\omega = 0$ and $\omega = \pi$). The warping matrix $\mathbf{A}$ is given by $\mathbf{C} \cdot \mathbf{M} \cdot \mathbf{S}$, where $\mathbf{M}$ is a $K \times K$ sparse matrix that contains the correspondence between source and target frequency bins (similarly as in [36]), $\mathbf{S}$ is the matrix that transforms a MCEP vector into K log-amplitude spectral bins, and $\mathbf{C}$ is the matrix that converts bins into MCEP coefficients. In accordance with eq. (1), the elements of $\mathbf{S}$ are given by

$$s_{k,i} = \cos\left(i \cdot \mathrm{mel}\left(\pi k/K\right)\right) \ , \quad 0 \le k \le K \ , \quad 0 \le i \le p \tag{5}$$

and $\mathbf{C}$ can be expressed as

$$\mathbf{C} = \left(\mathbf{S}^\top \mathbf{S} + \lambda \mathbf{R}\right)^{-1} \mathbf{S}^\top \tag{6}$$

where $\lambda = 2 \cdot 10^{-4}$ and $\mathbf{R}$ is a diagonal perturbation matrix whose i^{th} element is $r_{i,i} = 8\pi^2 i^2$ (see [36] for more details). The amplitude scaling vector $\mathbf{b}$ that corresponds to $A(\omega)$ can be obtained[5] by multiplying $\mathbf{C}$ by a K-dimensional vector containing the log-amplitude spectral bins of $A(\omega)$.

VI. Discussion and Conclusion

In this paper we have described the ZureTTS system, a system than includes a set of tools and algorithms covering

[2] Frequency warping modifies the frequency axis of speech spectrum according to a specific mapping function.

[3] Amplitude scaling can be understood as a filtering process that modifies the log-amplitude spectrum of speech.

[4] A similar formulation is typically used to implement the $\mathrm{mel}(\cdot)$ function of eq. (1).

[5] The MCEP representation of the aforementioned loudness filter was calculated exactly this way (though for efficiency it was calculated offline only once, then stored at code level).

the main goals of our project, i.e., the easy creation by non-expert users of a personalized synthetic voice. There are, however, some aspects of the performance that deserve an explicit discussion.

As we have mentioned in Section II-B, users record their voice while reading a set of phonetically-balanced sentences displayed on the screen. Since the texts of these sentences are known, the context labels needed by HTS to perform adaptation are easy to obtain. Nevertheless, it is necessary to align such labels with the input recordings. As labels are defined at phone level, phone segmentation is needed whenever a user transmits a set of recordings to the server and presses the *"Start adaptation"* button. This was not an issue for those languages where Festival-based text analyzers were being used (namely English, Catalan and Slovak), since Festival already provides the necessary tools to do this (it is even capable of detecting and considering the speaker's short pauses that are not consistent with the punctuation marks of the text). For the remaining languages, a solution had to be investigated. The problem of using speech recognition technology was that it required either the availability of appropriate pre-trained speaker-independent recognition models for each language, which would hinder the incorporation of new languages, or the ability to train them exclusively from the user's recorded material, which could be too scarce for an accurate modeling. Hence, we decided to use the initial voice models (synthesis HSMMs), which are obviously available for all languages and are supposed to be of high accuracy, in forced alignment mode. To overcome the possible spectral mismatch between the initial models and the input voice, we followed the strategy described in [35], which consists of the following steps:

1) Get phone durations via forced alignment between the input acoustic vectors and the HSMMs (the mathematical formulation can be found in [35]).
2) Calculate the vocal tract length factor α (see section V-3) that makes the acoustic vectors maximally closer to a similar sentence generated from the HSMM with the current durations.
3) Transform the input vectors through the matrix given by eq. (4) and go back to the first step until convergence is reached.

This procedure resulted in highly satisfactory phone segmentations while getting some interesting information about the vocal tract length of the input voice[6]. The main limitation of this method is that it does not consider the inclusion of short pauses. This issue should be addressed for a more accurate overall performance of the system for the involved languages.

We also studied the possible use of the output likelihoods of HSMM forced alignment for utterance verification. Note that in the current implementation of the system the user him/herself is asked to validate the recordings before transmission, which makes the system prone to undesired problems. Unfortunately, we were not able to implement an accurate detector of mispronounced sentences within the duration of the ZureTTS project. Future works will possibly address this issue.

Another limitation of the system is the lack of control over the quality of the input (presumably home-made) recordings. Given the profiles of the project participants, this aspect was not explicitly tackled, but informal tests indicate that passing the recorded signals through a Wiener filter before acoustic analysis avoids gross noise-related artifacts and does not substantially harm the performance of the system in terms of adaptation. In any case, it is logical to suppose that users will take care of the recording conditions as long as they want to get a high-quality personalized voice.

As detailed in Section IV, language-specific initial voice models were trained from databases that were available to the project participants with adequate permissions. In some cases the databases were not optimal for this task; in some others the amount of training material was not sufficiently large; also, time constraints imposed the need of using only a subset of some relatively large databases. Consequently, there are some significant performance differences between languages. Most of the models are currently being retrained or properly trained, which is likely to result into significant performance improvements in the short term.

The acquisition of a powerful dedicated server is undoubtedly one of the necessary changes to be carried out in the near future. Currently, the system runs in a mid-performance server and requests from different users are processed sequentially on a first-in first-out basis, which sometimes results in unnecessary delays.

Finally, beyond its immediate practical goals, ZureTTS provides a framework to investigate very interesting challenges such as dynamically incorporating the acquired knowledge (basically from the recordings) into the average voice models, or designing new adaptation strategies that reduce the number of sentences that the user has to record for a successful adaptation. When properly advertised, the ZureTTS web portal is expected to facilitate the communication and interaction between researchers and users. In accordance with the market trends it is still necessary, however, to provide the users not only with web services such as those described in Section III, but also with user-friendly "apps" that are compatible with the personalized voices yielded by ZureTTS. With regard to this, an Android version of AhoTTS [18] will be launched soon.

ACKNOWLEDGMENT

The authors would like to thank the support of the Basque Government (ZURETTS+, S-PE11UN081), the Spanish Ministry of Economy and Competitiveness (SpeechTech4All, TEC2012-38939-C03-03, and FPI program), the European Regional Development Fund, the Ministry of Education of the Slovak Republic and Slovak Academy of Sciences (VEGA, 1/0386/12), the Media Development Authority of Singapore (Grant MDA/IDM/2012/8/8-2 VOL 01), the sponsors of eNTERFACE'14 (ISCA, UPV/EHU, EURASIP, RTTH, IKERBASQUE, the Province and City Council of Bilbao, and Metro Bilbao), iFLYTEK Co. Ltd. for providing the Mandarin speech synthesis corpus, and the developers and maintainers of HTS for making its code publicly accessible.

[6] This privileged information is not yet exploited in the current implementation of the system.

References

[1] 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.

[2] H. Zen, K. Tokuda, and A. 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, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[4] 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.

[5] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. 6th ISCA Speech Synthesis Workshop*, 2007, pp. 294–299.

[6] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: voice banking and reconstruction," *Acoustical Science & Tech.*, vol. 33, no. 1, 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. Hirsimki, R. Karhila, and M. Kurimo, "Personalising speech-to-speech translation: unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis," *Computer Speech & Lang.*, vol. 27, no. 2, pp. 420 – 437, 2013.

[8] D. Erro, T. Zorilă, Y. Stylianou, E. Navas, and I. Hernáez, "Statistical synthesizer with embedded prosodic and spectral modifications to generate highly intelligible speech in noise," in *Proc. Interspeech*, 2013, pp. 3557–3561.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] [Online]. Available: http://hts.sp.nitech.ac.jp

[15] [Online]. Available: http://www.cstr.ed.ac.uk/projects/festival

[16] 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.

[17] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge University Engineering Department, 2006.

[18] A. Alonso, I. Sainz, D. Erro, E. Navas, and I. Hernáez, "Sistema de conversión texto a voz de código abierto para lenguas ibéricas," *Procesamiento del Lenguaje Natural*, vol. 51, pp. 169–175, 2013.

[19] D. Erro, I. Sainz, I. Luengo, I. Odriozola, J. Sánchez, I. Saratxaga, E. Navas, and I. Hernáez, "HMM-based speech synthesis in Basque language using HTS," in *Proc. FALA*, 2010, pp. 67–70.

[20] 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.

[21] [Online]. Available: http://festcat.talp.cat

[22] 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.

[23] 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.

[24] 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.

[25] [Online]. Available: http://sourceforge.net/projects/ctbparser/

[26] N. Xue, F. Xia, F.-D. Chiou, and M. Palmer, "The penn chinese treebank: Phrase structure annotation of a large corpus," *Nat. Lang. Eng.*, vol. 11, no. 2, pp. 207–238, 2005.

[27] A. Li, "Chinese prosody and prosodic labeling of spontaneous speech," in *Speech Prosody*, 2002.

[28] Y.-Q. Shao, Z.-F. Sui, J.-Q. Han, and Y.-F. Wu, "A study on chinese prosodic hierarchy prediction based on dependency grammar analysis," *Journal of Chinese Information Process.*, vol. 2, p. 020, 2008.

[29] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[30] D. Moers and P. Wagner, "The TTS voice "Petra"," Bielefeld University, Tech. Rep., 2010.

[31] [Online]. Available: http://www.bas.uni-muenchen.de/Forschung/BITS

[32] [Online]. Available: http://sourceforge.net/projects/boss-synth

[33] 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.

[34] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Trans. Audio, Speech, & Lang. Process.*, vol. 21, no. 3, pp. 556–566, 2013.

[35] D. Erro, A. Alonso, L. Serrano, E. Navas, and I. Hernáez, "New method for rapid vocal tract length adaptation in HMM-based speech synthesis," in *8th ISCA Speech Synthesis Workshop*, 2013, pp. 125–128.

[36] T.-C. Zorilă, D. Erro, and I. Hernaez, "Improving the quality of standard GMM-based voice conversion systems by considering physically motivated linear transformations," *Commun. in Computer & Inf. Science*, vol. 328, pp. 30–39, 2012.

Daniel Erro$^{(6)}$ received his Telecommunication Eng. degree from Public University of Navarre, Pamplona, Spain, in 2003 and his Ph.D. degree from Technical University of Catalonia, Barcelona, Spain, in 2008. Currently he holds an Ikerbasque Research Fellowship at Aholab, University of the Basque country, Bilbao, Spain. His research interests include speech analysis, modeling, modification, conversion, reconstruction and synthesis. He was the general chair of eNTERFACE'14.

Inma Hernáez received the Telecommunication Eng. degree from the Technical University of Catalonia, Barcelona, Spain, and the Ph.D. degree from the University of the Basque Country, Bilbao, Spain, in 1987 and 1995, respectively. She is a Full Professor in the Faculty of Engineering, University of the Basque Country. She is founding member of the Aholab Signal Processing Research Group. Her research interests are signal processing and all aspects related to speech processing. She is highly involved in the development of speech resources and technologies for the Basque language. She was the general co-chair of eNTERFACE'14.

Eva Navas received the Telecommunication Eng. degree and the Ph.D. degree from the University of the Basque Country, Bilbao, Spain. Since 1999, she has been a researcher at AhoLab and an associate professor at the Faculty of Industrial and Telecommunication Engineering in Bilbao. Her research is focused on expressive speech characterization, recognition, and generation.

Agustín Alonso$^{(5)}$ received his Telecommunication Eng. degree in 2010 and his M.Sc. degree in 2013, both from the University of the Basque Country, Bilbao, Spain. Currently he is a PhD student at Aholab Signal Processing Laboratory, University of the Basque Country, focusing on speech synthesis, transformation and conversion.

Haritz Arzelus(1) received his Computer Eng. degree from the University of the Basque Country, San Sebastian, in 2009. Since October 2009, he has been working in Vicomtech-IK4 as a researcher in speech and language processing technologies on local, national and European projects such BERBATEK, SAVAS and SUMAT. He participated actively in the creation of Ubertitles S.L. (2013), a company oriented to provide automatic subtitling, developing and integrating the technology on which it is based.

Igor Jauk(4) received his Master degree in Phonetics and Computational Linguistics at the University of Bonn, Germany, in 2010. After a research period at the Bielefeld University, Germany, in artificial intelligence and dialogue systems and at the Pompeu Fabra University, Spain, in audiovisual prosody, he now holds an FPU grant for a Ph.D. degree at the Technical University of Catalonia, Barcelona, Spain. His research interests are expressive speech synthesis and information retrieval.

Nguyen Quy Hy(10) received his bachelor in Computer Science from Nanyang Technological University, Singapore, in 2003. He is currently a Master student in the same university. His research interests include all aspects of software engineering and speaker-adaptive expressive speech synthesis.

Carmen Magariños(3) received her Telecommunication Eng. degree in 2011, and her M.Sc. degree in 2014, both from the University of Vigo, Spain. Currently she works as a PhD student at the Multimedia Technology Group of the University of Vigo. Her research interests are focused on speech technology, mainly on HMM-based speech synthesis, hybrid models and speaker adaptation.

Rubén Pérez-Ramón(2) finished his studies on Spanish Language and Literature at Universidad Autónoma de Madrid, Spain, in 2010. He completed a Masters degree on Speech Technology and another one on Judicial Phonetics, and has collaborated with CSIC in several punditries. He is currently preparing his Ph.D. thesis at the University of the Basque Country, Vitoria, Spain.

Martin Sulír(9) received his M.Sc. (Eng.) degree in the field of Telecommunications in 2012 at the Department of Electronics and Multimedia Communications of the Faculty of Electrical Engineering and Informatics at the Technical University of Košice. He is currently PhD student at the same department. His research interests include text-to-speech synthesis systems.

Xiaohai Tian(11) received his Computer Application and Technology degree from Northwestern Polytechnical University, Xian, China, in 2011. Currently he pursues his Ph.D degree at School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include speech signal processing, voice conversion and speech synthesis.

Xin Wang(8) received his B.Eng. degree from University of Electronic Science and Technology of China, Chengdu, China, in 2012. Currently he pursues a master's degree in University of Science and Technology of China, Hefei, China. His research interests include speech synthesis in Text-to-Speech and Concept-to-Speech.

Jianpei Ye(7) received the Telecommunication Eng. degree and the M.Sc. degree in Space Science and Technology from the University of the Basque Country, Bilbao, Spain, in 2013 and 2014, respectively. He is currently a junior researcher in Aholab, University of the Basque Country. His research interests include statistical approaches to speech processing such as voice conversion, speech analysis and speech synthesis.