Statistical text-to-speech synthesis of Spanish subtitles

S. Piqueras, M. A. del-Agua, A. Giménez, J. Civera, and A. Juan

MLLP, DSIC, Universitat Politècnica de València,
Camí de Vera s/n, 46022, València, Spain
{spiqueras,mdelagua,agimenez,jcivera,ajuan}@dsic.upv.es

Abstract. Online multimedia repositories are growing rapidly. However, language barriers are often difficult to overcome for many of the current and potential users. In this paper we describe a TTS Spanish system and we apply it to the synthesis of transcribed and translated video lectures. A statistical parametric speech synthesis system, in which the acoustic mapping is performed with either HMM-based or DNN-based acoustic models, has been developed. To the best of our knowledge, this is the first time that a DNN-based TTS system has been implemented for the synthesis of Spanish. A comparative objective evaluation between both models has been carried out. Our results show that DNN-based systems can reconstruct speech waveforms more accurately.

Keywords: video lectures, text-to-speech synthesis, accessibility

1 Introduction

The proliferation of online video lecture repositories over recent years is a phenomenon hard to ignore. In particular, in the field of education, universities around the world are making a huge effort in the recording and publication of video lectures. Some of the most successful online video lecture repositories are TED talks [21], VideoLectures.NET [28], Coursera [2], and Khan Academy [4], to name just a few.

These repositories are opened on a global scale, but their monolingual content creates a language barrier that is difficult to overcome, driving away many potential users. Although the most popular video lectures in these repositories are manually transcribed and translated by dedicated users in a collaborative effort, manual subtitling cannot keep pace with the increasing rhythm of video generation on the long term. This subtitling process becomes even more cumbersome when dealing with talks that include highly specialized vocabulary.

Recent advances in automatic speech recognition (ASR) [7] and machine translation (MT) [5,13,15] have pushed the scientific community to tackle more challenging subtitling tasks related to large video lecture repositories. Indeed, current state-of-the-art ASR and MT systems can provide accurate enough subtitles that can be manually revised with minimum effort, saving time and money.

In particular, the transLectures project [20] is aiming to develop high quality, cost-effective solutions for the transcription and translation of massive online repositories. This project has so far resulted in the release of the open-source transLectures -UPV toolkit [22]. Nevertheless, the availability of subtitles may not be enough to fully exploit video visualisation, since users are forced to split their attention between subtitles and lecture slides. In addition, visually impaired users cannot benefit from subtitles. In these cases, it would be much more convenient to be able to listen to the lecturer in the user's own language.

A text-to-speech (TTS) synthesizer is a system capable of generating an artificial speech track for a given text. State-of-the-art TTS systems usually employ one of two approaches: unit selection [11] or statistical parametric speech synthesis [34]. The TTS system presented here is based on the latter, as it is usually regarded as the most reliable synthesis approach when it comes to intelligibility [12], which is a key factor in our problem. However, the current reference Spanish TTS system publicly available only provides pre-trained HMM-based models [17].

In this work, two statistical TTS systems for Spanish are presented. The first of them is based on the conventional HMM acoustic modeling [30], while the second system implements state-of-the-art deep neural networks (DNN) for acoustic modeling [33]. These TTS systems were objectively evaluated on a real-life video-lecture repository. To the best of our knowledge, this evaluation has never been performed for the Spanish language. The best performing TTS system is intended to be applied to the generation of Spanish audio tracks on large video lecture repositories. The TTS-generated voice will be seamlessly integrated into the original video in order to allow users to concentrate on the video lecture content, keeping them from having to read subtitles.

The rest of this paper is organized as follows. Firstly, an overview of a TTS system is depicted in Section 2. Then, TTS systems, both HMM-based and DNN-based, are described in Section 3. Next, results on the objective evaluation with both TTS systems are reported in Section 4. Finally, concluding remarks and future research lines are wrapped up in Section 5.

2 System overview

In Figure 1 we provide an overview of the modules that make up our TTS system. We describe the modules involved in our system from the moment the subtitle file is received to the point the speech output is ready to be embedded.

In the first step, the subtitle file is divided into segments according to its timestamps. This division allows us to process large transcription files in parallel, and we will later concatenate the segments appropriately. Furthermore, silences between segments will not be passed to the synthesizer, so the generation is more efficient.

Segments are processed by the linguistic analysis module that has been developed for this work. The words are split into syllables, which are then converted to phonemes with a rule-based grapheme-to-phoneme algorithm. As Spanish

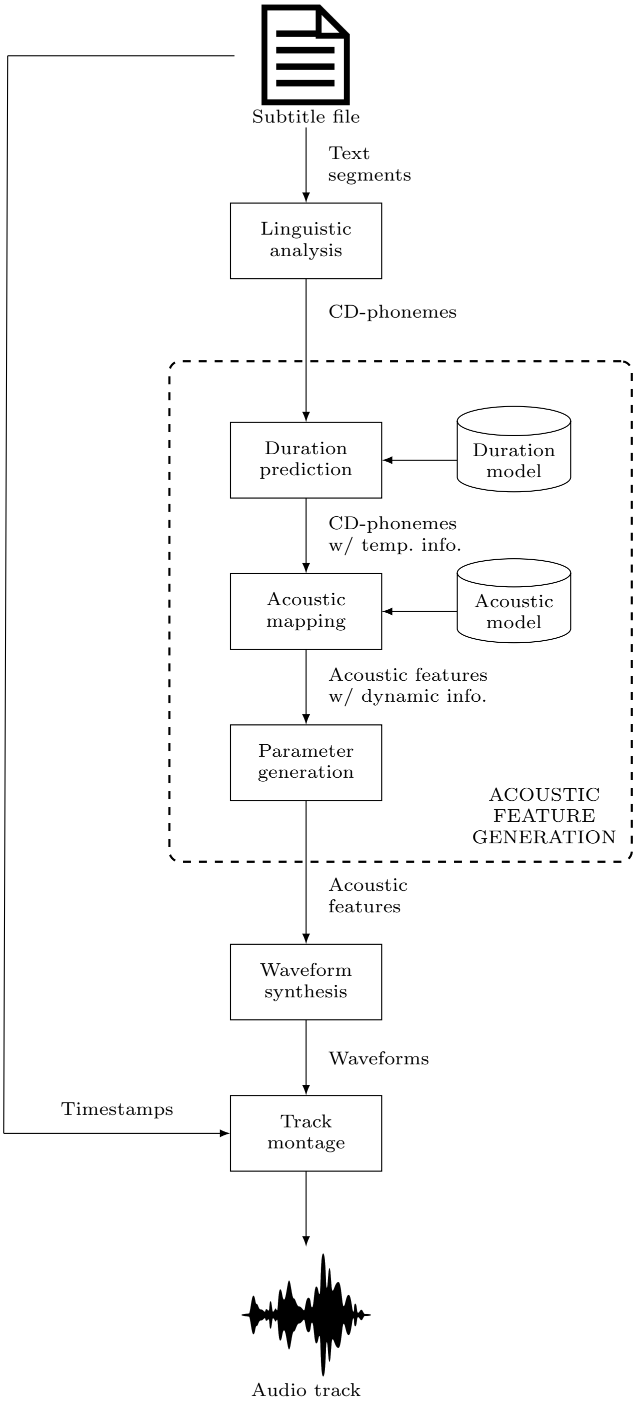

Fig. 1. System overview

orthography is highly phonemic, this conversion is carried out without much loss. Please note that while this approach can deal appropriately with Spanish words, it does not cover foreign words (e.g. proper nouns of people). The module also extracts contextual information of each phoneme, syllable and word, and then employs it to create context-dependent (CD) phonemes [31]. We have not included some of the higher-level features, which are used in other speech synthesizers, such as POS tagging, stress marks or ToBI endtones.

Next, an acoustic feature generation module converts the CD-phonemes into an acoustic parameter sequence. It is currently divided into three parts, which correspond to the duration generation module, the acoustic mapping module and the parameter generation module. These modules will be described in detail in Section 3.

The acoustic parameter sequence is then post-processed with a spectral enhancement algorithm and then sent to the vocoder to generate the audio segments. The vocoder's task is to reconstruct the speech waveforms from the acoustic parameter sequence. Our system uses a harmonics-plus-noise model based vocoder [8], and makes use of the free implementation provided by their authors [1]. In this vocoder, the spectral parameters are the Mel-frequency cepstral coefficients (*mfcc*), while the excitation parameters are the logarithm of the fundamental frequency ($\log F_0$), which determines the pitch, and the maximum voiced frequency (*mvf*).

The track montage module uses temporal annotations included in the subtitle file to create a new track by concatenating silence and synthesized audio segments. This track may later be embedded in the multimedia file as a side track, in order to allow the user to select their preferred language.

3 Acoustic modeling

In this section, the two main approaches to acoustic modeling investigated in this work, HMM and DNN, are described in detail. The reader familiar with ASR should note that in TTS, in contrast to ASR, the acoustic modeling process tackles the reverse problem of mapping acoustic features to CD-phonemes.

3.1 HMM-based

The conventional approach to acoustic modeling in speech synthesis is to perform the acoustic mapping through context-dependent Hidden Markov Models with explicit duration, also known as Hidden Semi-Markov Models (HSMMs). In the generation step, first the state durations for each phoneme are predicted by a Gaussian distribution model. Then, an HMM model is selected. Finally, the means and variances of the output acoustic parameter vector are generated by the HMM model. In order to avoid the discontinuities that would arise from a maximum likelihood approach, the acoustic parameter sequence is smoothed with the introduction of dynamic features and the use of the maximum likelihood parameter generation (MLPG) algorithm [24].

As CD-phonemes often have high dimensionality, training a CD-HMM for each possible combination of text analysis features is unrealistic and would result into poorly estimated HMMs. By way of solution, context clustering techniques at a state-level are used. Clustering is performed by means of binary decision trees. In the training phase, the Minimum Description Length (MDL) criterion is used to construct these decision trees [19]. The size of the trees can be controlled through the penalty term α (where α is typically set to 1). As the spectral and excitation parameters have different context dependency, separate trees are built for each one. This approach allows our model to handle unseen contexts.

An extra problem emerges from the modelization of the non-continuous parameters log F_0 and *mvf.* These parameters are defined in the regions known as "voiced", and undefined in the regions known as "unvoiced". Log F_0 has been modeled with a multi-space probability distribution [25], while the *mvf* parameter was added as an extra stream and modeled with a continuous distribution, as suggested in [8]. The *mvf* values were interpolated in the unvoiced frames.

3.2 DNN-based

DNNs have been successfully applied to acoustic modeling in ASR tasks [10]. DNNs map frame features, including textual and temporal features, to acoustic features in a feed-forward approach. The textual information is composed of binary features, such as *is-current-syllable-accented*, and numerical features, such as *number-of-phonemes-in-current-word.* Four temporal features are defined, corresponding to the position of the current frame (forward and backward) in the current phoneme, the duration of the phoneme and the duration of the whole segment. Similar to the HSMMs method, the duration of the phonemes is predicted by an external Gaussian model. However, in contrast to HMM models, all the parameters for every possible CD-phoneme will be generated by the same network. This joint modeling procedure results in a more robust estimation, which produces better generalization [33].

In order to deal with the voiced/unvoiced (V/UV) discontinuity problem, a continuous explicit modeling approach has been used for both log F_0 and *mvf.* An extra bit of the output is used to classify the frame as voiced or unvoiced. To produce smoother parameter trajectories, the output includes dynamic information (first and second derivatives) of the parameter sequence. The network output is assumed to be the mean vector of a Gaussian posterior distribution, and is combined with a precomputed variance vector to generate the acoustic feature vector through the MLPG algorithm used in HMM synthesis. A single variance for each output is estimated from all the training samples.

4 Experimental results

In this section, the corpus employed in our experiments is described. Then, the evaluation measures are presented along with the experimental setup. Finally, comparative results between HMM-based and DNN-based TTS systems are reported and discussed.

4.1 Corpus description

The data used for our experiments has been extracted from the poli[Media] repository, which contains over $2,000$ hours of video lectures. poli[Media] is a recent, innovative service for the creation and distribution of multimedia educational content at the UPV [16,27] mostly in Spanish, but also in Catalan and English. It is primarily designed to allow UPV lecturers to record their courses in short videos lasting up to 10 minutes, accompanied by time-aligned slides.

The production process of a poli[Media] repository has been carefully designed to achieve both a high rate of production and a fine quality, comparable to a TV production but at a lower cost. However, this repository was not specifically recorded with synthesizer training in mind, and so audio conditions are far from perfect. Furthermor, the recordings contain speaker hesitations, unfinished words and various noises (i.e. coughs).

The complete poli[Media] repository has been automatically transcribed using the open-source ASR system called transLectures-UPV toolkit [22]. In order to train this ASR system, a set of 100 hours of video lectures were manually transcribed. From this set, a subset of 40 videos with 2320 utterances by a single male native Spanish speaker was selected. After removing the silences from the videos, 6 hours of speech remain for our experiments.

From this subset, 49 utterances were used for testing purposes. The remaining 2271 utterances were used to train the HMM-based system. In the case of the DNN-based system, 2171 utterances were devoted to pretraining and fine-tuning stages, while 100 utterances were reserved as a validation set in order to avoid overfitting. Phoneme alignments were automatically performed by the best acoustic model deployed in the **transLectures** project at month 24 [26].

4.2 Evaluation measures

The comparative evaluation of our TTS systems was performed in terms of well-known objective error measures. These measures are mean Mel-cepstral distortion [14] (MMCD), voiced/unvoiced error rate and root mean squared error (RMSE) in log F_0. In the latter case, the RMSE was only computed for the frames where the system had correctly guessed whether the frame was voiced or unvoiced. Phoneme durations were set to match those from the natural speech, rather than being generated by the Gaussian model described in Section 3.

It should be noticed that while these objective values are frequently used in the TTS research field to compare the performance of the acoustic models, they do not perfectly correlate with the naturalness of the synthesized speech [23].

4.3 Experimental setup

For training purposes, audio was extracted from the video and downsampled from 44100Hz to 16000Hz. Every 5 milliseconds, 40 Mel-cepstral coefficients, log F_0 and maximum voiced frequency values were extracted using AhoCoder tools [1]. The *mvf* parameter was interpolated in the unvoiced regions for both models, while the log F_0 was interpolated for the DNN explicit voicing. The acoustic parameter vectors were then augmented with the information of the first and second derivatives. The textual analysis information was the same for both models.

The HMM system was composed of 5-state, no-skip models with diagonal covariance matrices. A total of 1017 different questions were used for the construction of the decision trees. For comparison purposes, we trained 3 HMM-based systems modifying the parameter α which controls the number of nodes of the decision trees (with $\alpha = 0.5$, 1.0 and 2.0). The training was performed using the most recent stable version (2.2) of the the HMM-based Speech Synthesis System (HTS) [3].

In the case of the DNN-based system, the number of neurons in the input layer was 169, while the number of neurons in the output layer was 127, corresponding to 39 *mfcc* plus energy, log F_0, *mvf*, first and second derivatives and the V/UV bit. Inputs to the DNN were normalized to have zero mean and one variance, while outputs were normalized between 0.01 and 0.99. Different neural network sizes were tested by changing the number of hidden layers (1, 2, 3 or 4) and the number of neurons per layer (128, 256, 512 or 1024). The sigmoid activation function was used in the hidden and output layers. Neural networks with more than one hidden layer were pretrained using a discriminative approach [18], and then fine-tuned with a stochastic minibatch backpropagation algorithm [6]. The error criterion in both steps was the mean squared error (MSE). The training was performed with a CUDA-based GPU implementation, part of a development version of the **trans**Lectures toolkit.

4.4 Results and Discussion

Table 1 shows the objective evaluation measures computed for each DNN configuration, together with the results of the best HMM model. For every DNN configuration, the optimal number of neurons per layer has been selected so that the evaluation measure is optimized. We can see that DNN-based systems systematically achieve better results in every measure than HMM-based systems. The optimal number of layers is unclear, since the evaluation measures exhibit different behaviour. The V/UV error rate performs better when using simpler architectures, while the spectral parameters benefit more from a complex architecture.

Table 1. Comparison between HMM-based and DNN-based acoustic models.

System	# layers	RMSE log F_0	MMCD	V/UV Error rate
HMM	-	0.190	6.987	13.35
DNN	1	0.183	6.792	12.08
	2	0.183	6.702	12.27
	3	0.184	6.678	12.36
	4	0.184	6.679	12.42

5 Conclusions and future work

We have presented a novel text-to-speech system for the synthesis of Spanish subtitles in video lectures. We have reviewed the statistical speech synthesis framework and discussed why it is appropriate for our task. We have described the whole system and presented two different approaches to performing acoustic mapping: HMM-based and DNN-based. We have performed a series of experiments to compare the performance of both approaches. Objective measures show that the best DNN systems consistently outperform the HMM systems.

Currently, our next steps include the training of a female Spanish voice and the integration of the system in the UPV video lecture platform poli[Media]. Once integrated, subjective evaluation of the intelligibility and naturalness of the voices will be carried out. Future work also includes the exploration of other network topologies [9], incorporating variance modeling into the DNNs [32], cross-lingual speaker adaptation [29] and a more in-depth linguistic analysis.

Acknowledgments The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 287755 (trans**Lectures**) and ICT Policy Support Programme (ICT PSP/2007-2013) as part of the Competitiveness and Innovation Framework Programme (CIP) under grant agreement no 621030 (EMMA), and the Spanish MINECO Active2Trans (TIN2012-31723) research project.

References

1. Ahocoder. http://aholab.ehu.es/ahocoder
2. Coursera. http://www.coursera.org
3. HMM-Based Speech Synthesis System (HTS). http://hts.sp.nitech.ac.jp
4. Khan Academy. http://www.khanacademy.org
5. Axelrod, A., He, X., Gao, J.: Domain adaptation via pseudo in-domain data selection. In: Proc. of EMNLP. pp. 355–362 (2011)
6. Bottou, L.: Stochastic gradient learning in neural networks. In: Proceedings of Neuro-Nîmes 91. EC2, Nimes, France (1991)
7. Dahl, G.E., Yu, D., Deng, L., Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing 20(1), 30–42 (2012)
8. Erro, D., Sainz, I., Navas, E., Hernaez, I.: Harmonics plus noise model based vocoder for statistical parametric speech synthesis. IEEE Journal of Selected Topics in Signal Processing 8(2), 184–194 (2014)
9. Fan, Y., Qian, Y., Xie, F., Soong, F.: TTS synthesis with bidirectional LSTM based recurrent neural networks. In: Proc. of Interspeech. p. Submitted (2014)
10. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine 29(6), 82–97 (2012)

11. Hunt, A.J., Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. of ICASSP. vol. 1, pp. 373–376 (1996)
12. King, S.: Measuring a decade of progress in text-to-speech. Loquens 1(1), e006 (2014)
13. Koehn, P.: Statistical Machine Translation. Cambridge University Press (2010)
14. Kominek, J., Schultz, T., Black, A.W.: Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In: Proc. of SLTU. pp. 63–68 (2008)
15. Lopez, A.: Statistical machine translation. ACM Computing Surveys 40(3), 8:1–8:49 (2008)
16. poliMedia: The polimedia video-lecture repository (2007), http://media.upv.es
17. Sainz, I., Erro, D., Navas, E., Hernáez, I., Sánchez, J., Saratxaga, I.: Aholab speech synthesizer for albayzin 2012 speech synthesis evaluation. In: Proc. of IberSPEECH. pp. 645–652 (2012)
18. Seide, F., Li, G., Chen, X., Yu, D.: Feature engineering in context-dependent dnn for conversational speech transcription. In: Proc. of ASRU. pp. 24–29 (2011)
19. Shinoda, K., Watanabe, T.: MDL-based context-dependent subword modeling for speech recognition. Journal of The Acoustical Society of Japan 21(2), 79–86 (2000)
20. Silvestre-Cerdà, J.A., et al.: Translectures. In: Proc. of IberSPEECH. pp. 345–351 (2012)
21. TED Ideas worth spreading, http://www.ted.com
22. The transLectures-UPV Team.: The transLectures-UPV toolkit (TLK). http://translectures.eu/tlk
23. Toda, T., Black, A.W., Tokuda, K.: Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis. In: Proc. of ISCA Speech Synthesis Workshop (2004)
24. Tokuda, K., Kobayashi, T., Imai, S.: Speech parameter generation from hmm using dynamic features. In: Proc. of ICASSP. vol. 1, pp. 660–663 (1995)
25. Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T.: Multi-space probability distribution HMM. IEICE Transactions on Information and Systems 85(3), 455–464 (2002)
26. transLectures: D3.1.2: Second report on massive adaptation, http://www.translectures.eu/wp-content/uploads/2014/01/transLectures-D3.1.2-15Nov2013.pdf
27. Turró, C., Ferrando, M., Busquets, J., Cañero, A.: Polimedia: a system for successful video e-learning. In: Proc. of EUNIS (2009)
28. Videolectures.NET: Exchange ideas and share knowledge, http://www.videolectures.net
29. Wu, Y.J., King, S., Tokuda, K.: Cross-lingual speaker adaptation for HMM-based speech synthesis. In: Proc. of ISCSLP. pp. 1–4 (2008)
30. Yamagishi, J.: An introduction to HMM-based speech synthesis. Tech. rep., Centre for Speech Technology Research (2006), https://wiki.inf.ed.ac.uk/twiki/pub/CSTR/TrajectoryModelling/HTS-Introduction.pdf
31. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.: Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In: Proc. of Eurospeech. pp. 2347–2350 (1999)
32. Zen, H., Senior, A.: Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In: Proc. of ICASSP. pp. 3872–3876 (2014)
33. Zen, H., Senior, A., Schuster, M.: Statistical parametric speech synthesis using deep neural networks. In: Proc. of ICASSP. pp. 7962–7966 (2013)
34. Zen, H., Tokuda, K., Black, A.W.: Statistical parametric speech synthesis. Speech Communication 51(11), 1039–1064 (2009)