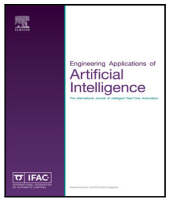
Engineering Applications of Artificial Intelligence 105 (2021) 104413

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Towards cross-lingual voice cloning in higher education

Alejandro Pérez [a,*], Gonçal Garcés Díaz-Munío [a], Adrià Giménez [a], Joan Albert Silvestre-Cerdà [a], Albert Sanchis [a], Jorge Civera [a], Manuel Jiménez [b], Carlos Turró [b], Alfons Juan [a]

[a] Machine Learning and Language Processing (MLLP) Group, Valencian Research Institute for Artificial Intelligence (VRAIN), Spain
[b] Media Services (ASIC), Universitat Politècnica de València, València, Spain

ARTICLE INFO

Keywords:
Text-to-speech
Multilinguality
Cross-lingual voice conversion
Educational resources
OER

ABSTRACT

The rapid progress of modern AI tools for automatic speech recognition and machine translation is leading to a progressive cost reduction to produce publishable subtitles for educational videos in multiple languages. Similarly, text-to-speech technology is experiencing large improvements in terms of quality, flexibility and capabilities. In particular, state-of-the-art systems are now capable of seamlessly dealing with multiple languages and speakers in an integrated manner, thus enabling lecturer's voice cloning in languages she/he might not even speak. This work is to report the experience gained on using such systems at the Universitat Politècnica de València (UPV), mainly as a guidance for other educational organizations willing to conduct similar studies. It builds on previous work on the UPV's main repository of educational videos, MediaUPV, to produce multilingual subtitles at scale and low cost. Here, a detailed account is given on how this work has been extended to also allow for massive machine dubbing of MediaUPV. This includes collecting 59 h of clean speech data from UPV's academic staff, and extending our production pipeline of subtitles with a state-of-the-art multilingual and multi-speaker text-to-speech system trained from the collected data. Our main result comes from an extensive, subjective evaluation of this system by lecturers contributing to data collection. In brief, it is shown that text-to-speech technology is not only mature enough for its application to MediaUPV, but also needed as soon as possible by students to improve its accessibility and bridge language barriers.

1. Introduction

Educational videos are now at the core of diverse online learning environments such as (content-agnostic) Massive Open Online Course (MOOC) platforms, dedicated tools and general platforms repurposed to support learning (Roll et al., 2018). Similarly, educational videos are becoming increasingly popular at universities, where they are often used to support blended learning (Fong et al., 2019; Morris et al., 2019).

A common challenge in both online and blended learning is how to produce multilingual video subtitles of publishable quality at scale and low cost. Clearly, a direct approach to this is to use modern AI tools for automatic speech recognition (ASR) and machine translation (MT): raw (automatic) subtitles in the *source* (spoken) language are produced first by an ASR system; then, they are machine-translated into a number of other, *target* languages of interest. It goes without saying that, depending on the quality of the ASR/MT systems involved, a greater or lesser effort has to be made to review (post-edit) the raw subtitles, especially those in the source language to avoid cumulative errors (i.e. early errors in the pipeline leading to further errors in later stages). This approach has been recently studied in the EU projects "transLectures: Transcription and Translation of Video Lectures" (Valor-Miró et al., 2015a), "EMMA: European Multiple MOOC Aggregator" (Valor-Miró et al., 2018) and "X5gon: Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site Global OER Network" (Iranzo et al., 2019; Jorge et al., 2020b). On the one hand, the results from transLectures and EMMA showed that the quality of the subtitles produced by modern, *task-adapted* ASR/MT systems makes raw subtitles worth post-editing. Indeed, it was found that, when compared to generating subtitles ex novo, post-editing saves approximately 25%–75% of reviewing time (Valor-Miró et al., 2018). On the other hand, the more recent results from X5gon, obtained by application of the very latest developments in ASR/MT to large Open Educational Resource (OER) repositories, have shown that we have now reached the point at which raw subtitles are often good enough for direct publication (Iranzo et al., 2019; Jorge et al., 2020b). Moreover, current research in ASR/MT is also showing that advanced systems no longer need prerecorded audio (*offline* setup), as they can now work with no significant degradation under the so-called *streaming* setup; that is, subject to the constraint that output must be delivered in nearly *real time,* only within a short delay (around one second) after the incoming audio stream (Jorge et al., 2020a). To us, all of this recent progress will soon result in a rapid increase of (raw) multilingual subtitles of publishable quality for

* Corresponding author.
E-mail address: alpegon2@vrain.upv.es (A. Pérez).

https://doi.org/10.1016/j.engappai.2021.104413
Received 11 September 2020; Received in revised form 12 July 2021; Accepted 29 July 2021
Available online xxxx

large repositories of educational videos, and also *live* lectures, either online or not, delivered under reasonable acoustic conditions.

Assuming a progressive reduction of the cost to produce publishable subtitles in diverse target languages, it is natural to also consider the use of text-to-speech (TTS) tools to efficiently dub the lecturer's speech in target languages she/he might not even speak. In fact, as with subtitles, synthesized speech has been used for many years to make content accessible to people with disabilities (W3C, 2018). Other areas for which TTS tools have provided support include second language learning (Godwin-Jones, 2019), reading difficulties (van Campen et al., 2020) and virtual humans (Chiou et al., 2020). Although these tools have been available and used for many years, it has not been until very recently that a plethora of contributions based on modern AI tools have dramatically improved and extended TTS capabilities. Indeed, the naturalness of the speech generated by state-of-the-art TTS systems is now known to rival that of human speech (Shen et al., 2018). Also, the most advanced TTS systems are capable of generating speech for multiple speakers and languages, even in the usual case in which speakers can only provide training data in just a few languages, thus enabling cross-lingual voice cloning in all target languages of interest (Zhang et al., 2019). To us, this particular feature is especially interesting to bridge language barriers at universities, as it opens the door to produce multilingual educational videos at scale with both publishable subtitles and cloned lecturer speech.

This work is to report the experience gained on (cross-lingual) voice cloning at the Universitat Politècnica de València (UPV) in recent years. It builds on past and ongoing work on using modern AI tools to produce multilingual subtitles and synthesized speech for the UPV's main repository of educational videos, *MediaUPV* (Valor-Miró et al., 2018; Pérez et al., 2019). Our first, pioneering tests using deep neural networks (DNNs) for Spanish TTS in MediaUPV were carried out by the end of transLectures (Piqueras et al., 2014). Albeit with some delay with respect to ASR and MT, at that time it was clear to us that TTS technology was on the brink of a breakthrough on both performance and capabilities. Thus, in order to properly assess what TTS progress can do for voice cloning at the UPV, two main actions were taken. On the one hand, a call for participation to the UPV's academic staff was made so as to collect *clean* lecturer speech data (i.e. with no background noise or artefacts), and later survey their opinions and suggestions on the potential application of TTS at the UPV. At this point it is worth noting that acquiring such a database of lecturer speech data was also seen as crucial in learning about patterns of language proficiency among UPV lecturers with good predisposition to use TTS. On the other hand, we began to monitor TTS progress, especially as regards to systems capable of dealing with multiple speakers and languages. After the acquisition of the database during the academic courses 2016–17 and 2017–18, a multilingual and multi-speaker TTS system was built from current state-of-the-art TTS technology adapted to the UPV case. This system was then used to voice-clone (part of) MediaUPV and survey what UPV lecturers participating in our study think about present TTS technology. For the survey, participants listened to human and synthetic voice, for their own and others' videos in MediaUPV, also including cross-lingual cloned voice. Although the survey originated many questions and thoughts from lecturers, the general view is that TTS technology is not only mature enough for its application at the UPV, but also needed as soon as possible, especially to bridge language barriers for foreign students.

We are convinced that the experience gained on TTS at the UPV so far can be really valuable for other universities willing to conduct similar studies. Although this experience encompasses different, more or less technical aspects, the main technical contribution of this work is an effective production pipeline of subtitles and cloned voice, particularly in regard to multilingual and multi-speaker TTS technology. In Section 2, we begin with a description of the two main data sources considered in this work. First, the MediaUPV repository is introduced, from its origin and evolution to its current contents, especially from a linguistic perspective and paying attention to the way publishable multilingual subtitles have been produced cost-effectively. Second, our lecturer speech database for TTS is described in terms of acquisition protocol and basic statistics. Section 3 follows with a review of our production pipeline of subtitles and cloned voice, particularly in regard to TTS technology, its state-of-the-art and the way it was adapted to train a multilingual and multi-speaker TTS system from our lecturer speech database. Section 4 is devoted to the evaluation of this TTS system, the protocol and support platform we used to acquire the UPV lecturers' opinion on it, and the results obtained. Finally, the main conclusions drawn and future plans are given in Section 5.

2. MediaUPV and the DeX-TTS dataset

2.1. MediaUPV with multilingual subtitles

In a broad sense, the MediaUPV repository is a professional UPV service for the creation, storage, management and open dissemination of educational videos (Turró et al., 2009; MediaUPV, 2020). Launched in 2007, it was initially designed for UPV lecturers to produce high-quality short video recordings at dedicated UPV studios, with the aim of supporting blended learning through prerecorded "knowledge pills". These recordings, usually referred to as *poliMedias*, have also served as the main back-end video service for the UPV to provide MOOCs (UPVX, 2020), especially as an edX member since 2014 (UPValenciaX, 2020). In this respect, it is worth noting that UPV has become one of the most renowned MOOC providers in Spanish, with more than 85 MOOCs and 290 editions already completed, more than 2.3 million enrolments, and two of the 100 most popular online courses of all time (ClassCentral, 2020). Apart from poliMedias, MediaUPV has been expanded to include homemade videos produced by students and lecturers themselves, known as *poliTubes*, which are uploaded to it in much the same way as in YouTube. Finally, since joining the Opencast consortium in 2011, UPV has deployed lecture capture technology to 84 locations from which more than 600 h per year are being recorded and added to MediaUPV for their distribution to students only through a Sakai LMS (Turró et al., 2014; Opencast, 2020).

Although MediaUPV comprises diverse kinds of educational videos, this exploratory work focuses only on poliMedias due to their predominance and simplicity in terms of duration, speakers and audio quality. As indicated above, they are produced at dedicated UPV studios which, in brief, are just low-cost video production (4 × 4 metre) rooms equipped with a white backdrop, video camera, capture station, pocket microphone, lighting and AV equipment including a video mixer and an audio noise gate (Fig. 1, top). After choosing day and time of an appointment by an online booking system, the lecturer comes to a poliMedia studio with slides and delivers her/his presentation in front of the video camera, which is captured and synchronously embedded in real-time at the bottom-right corner of the computer's video output. Then, after metadata annotation, review and approval by the lecturer, the resulting poliMedia is uploaded to MediaUPV (see example in Fig. 1, bottom).

Supported by the UPV's *Docència en Xarxa (DeX)* stimulus plan for online teaching, the number of poliMedias uploaded to MediaUPV has been steadily increasing since 2007, up to 44096 videos and a total of 10601 recording hours in June 2020. As with face-to-face teaching sessions, the vast majority of poliMedias are produced in Spanish though, as shown in Table 1, they are also produced, to a much lesser extent, in Catalan (also known as Valencian in the Valencian Community) and English. In this regard, the UPV has recently approved an ambitious plan to promote multilingual teaching for the period 2020–2023 in which Catalan and English are specifically identified as top priorities for support (BOUPV20, 2020, pp 120–144). On the one hand, Catalan is an official yet minority language in the Valencian Community, and thus its protection is seen not only as an appreciation of cultural diversity, but also an obligation to reduce discrimination on the grounds of language

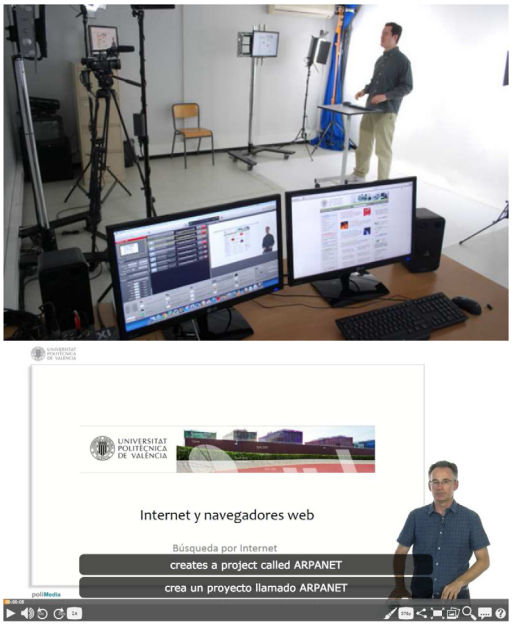


Fig. 1. A poliMedia studio (top) and example (bottom).

Table 1
Number of poliMedia videos and hours in Spanish, Catalan and English.

Language	Videos		Hours	
	No.	%	No.	%
Spanish	38 172	87	9 451	89
Catalan	1 333	3	232	2
English	4 591	10	918	7
Total	44 096	100	10 601	100

at the UPV. The case of English, on the other hand, is totally different. Increasing its use as a teaching language is clearly needed to strengthen the UPV's internationalization and competitiveness. It goes without saying that, for this plan to succeed, it will be good to have accurate and cost-effective means to fully convert basic (monolingual) poliMedias into trilingual learning objects.

Table 2 shows the number of lecturers producing poliMedias in each of the seven possible combinations of Spanish (es), Catalan (ca) and English (en): three of them monolingual (es, ca, en), three bilingual (es-ca, es-en, ca-en) and the trilingual case es-ca-en. It is worth noting that, for the figures in Table 2, only original recordings are considered, which in general are produced in a single language. Also, note that the percentages of monolingual, bilingual and trilingual lecturers are 91.9, 7.6 and 0.5, respectively. This means that a great majority of lecturers are producing poliMedias in a single language, Spanish in most cases, to support their face-to-face teaching sessions. Also worth noting is the fact that the number of lecturers producing poliMedias in English (872) is roughly 4 times that of poliMedias in Catalan (212), yet both languages account for a similar percentage of the total academic offer (BOUPV20, 2020, pp 120–144). This is because all Catalan-speaking learners are highly proficient in Spanish, and thus poliMedias in Spanish are also often used to support blended learning for Catalan-language groups. Needless to say, promoting multilingualism (in the UPV) means that all supported languages must be treated equally with regard to available resources.

The MediaUPV repository is a good example of how OER repositories are evolving in terms of size and complexity, especially at the linguistic level. This is why, (the poliMedia part of) it was chosen as a case study in the EU projects discussed in the introduction. By the second half of transLectures (2013–2014), poliMedia-adapted ASR/MT

Table 2
poliMedia lecturers for Spanish (es), Catalan (ca), English (en), bilingual combinations (es-ca, es-en, ca-en) and the trilingual case es-ca-en.

	Monolingual			Bilingual			Trilingual	Total
	es	ca	en	es-ca	es-en	ca-en	es-ca-en	
No.	2126	152	656	43	199	2	15	3193
%	66.6	4.8	20.5	1.3	6.2	0.1	0.5	100.0
Total (%)	91.9			7.6			0.5	100.0

Table 3
WER/BLEU scores provided by UPV and Google ASR/MT systems on poliMedias (es=Spanish, ca=Catalan, en=English, "es⇒ca"="Spanish to Catalan", etc.)

Systems	ASR WER (%)				MT BLEU (%)			
		es	ca	en		es⇒ca	es⇒en	ca⇒es
UPV	ASR	**9.1**	**12.6**	15.8	MT	**84.4**	33.5	**90.3**
Google	S2T standard	19.9	31.9	36.1	Translate	81.5	**36.8**	87.7
	S2T enhanced	n/a	n/a	**13.3**				
Δ%		118.7	153.2	−15.8		−3.4	9.8	−2.9

systems were already integrated into the MediaUPV production workflow to enrich all poliMedias with raw multilingual subtitles. At that time, however, it was felt that post-editing raw subtitles was still needed in many cases, and thus a user-friendly tool for reviewing was also integrated into the production workflow (Valor-Miró et al., 2015a; Silvestre-Cerdà et al., 2013; Pérez et al., 2015; Valor-Miró et al., 2015b). Being part of this workflow, subtitle post-editing was supported by the DeX stimulus plan, allowing each poliMedia to be reviewed not only by its author, but also by non-authors (e.g. users), with the author's approval prior to publication. Although this post-editing approach worked (and still works) well, poliMedias have been more and more published with no subtitle post-editing at all due to the increasing accuracy of new ASR/MT systems. Indeed, as indicated in the introduction, our latest results show that we have now reached the point at which raw subtitles are often good enough for direct publication. To be more precise, Table 3 provides some figures on the quality of the ASR and MT systems most used in the UPV at present. The reader is referred to Baquero-Arnal et al. (2020) and the references therein for details on UPV's ASR technology, resources and systems. As usual in ASR, transcription (source subtitles) quality is measured with an error metric known as *Word Error Rate (WER)* (Hunt, 1990). This metric counts the minimum (normalized) number of elementary word editing operations (insertions, deletions and substitutions of different words) required to transform an automatic transcription into its reference. Similarly, translation (target subtitles) quality in MT is usually assessed in terms of the *Bilingual Evaluation Understudy (BLEU)* accuracy criterion (Papineni et al., 2002). This criterion measures the degree of overlap between an automatic translation and a single correct reference translation, comparing isolated words and groups of up to four consecutive words. For comparison, Table 3 also provides analogous figures for general-purpose systems now commercially available from Google (*Google Cloud Speech-To-Text*, *standard* and *enhanced* if available; and *Google Translate*), and their relative value with respect to those built at the UPV ($\Delta\% = 100\,(G-U)/U$, where U and V are scores for UPV and Google systems, respectively).

For the analysis of results in Table 3, it should be pointed out first that there are no simple, error-free rules to decide, from WER and BLEU scores, whether raw subtitles are publishable or not. On the contrary, being a derivative work of an educational video owned by a lecturer, (raw) subtitles can be approved for publication only with the owner's consent and after review if desired. This is indeed the way in which publication consent has been sought for poliMedias since 2014 and, in doing so, it was soon realized that little or none subtitle post-editing was actually done as ASR/MT accuracy improved. To be precise, this was clearly observed for source subtitles produced by ASR systems with

Table 4
Participants contributing to clean speech data collection in Spanish (es), Catalan (ca), English (en), bilingual combinations (es-ca, es-en, ca-en) and the trilingual case.

	Monolingual			Bilingual			Trilingual	Total
	es	ca	en	es-ca	es-en	ca-en	es-ca-en	
Participants	36	1	4	16	22	3	16	98
Total	41			41			16	98

WER figures below 20%, as well as for target subtitles generated by MT systems with BLEU scores above 35% (Valor-Miró et al., 2018; Jorge et al., 2020b). It goes without saying that these thresholds must be taken with caution since, in principle, they are applicable only to poliMedias and the ASR/MT systems we are using. Coming back to Table 3 with these thresholds in mind, we see that UPV's raw subtitles are good enough for direct publication in all cases, except perhaps in the case of Spanish to English translation, whose BLEU score is slightly below 35%. In fact, the WER and BLEU scores for Spanish and Catalan, around 10% WER and above 84% BLEU, are far better than these thresholds. The comparatively higher WER for English might be due to the fact that most lecturers are not English native speakers, though we are convinced that WER results more similar to those of Spanish and Catalan would have been obtained from more in-domain data and thus better adapted models. In comparison, if we look at Google's results and their relative value, we see that Google's general-purpose systems are also fairly good, especially for MT and English ASR, though they are clearly behind the task-adapted ASR systems used at the UPV for Spanish and Catalan.

2.2. The DeX-TTS dataset

As indicated in the introduction, a call for participation was made to the UPV's academic staff under the DeX plan to collect *clean* lecturer speech data during the academic courses 2016–17 and 2017–18, which was answered by a total of 98 participants. Participants were all Spanish/Catalan-native speakers, 50 years old on average (with a standard deviation of 6) and equally distributed by gender. To this end, a number of sentences in Spanish, Catalan and English were first drawn from various sources (mainly newspapers, MOOCs and Wikipedia) and then reviewed for readability. Similarly to poliMedias, speech recordings were made under the same acoustic conditions at poliMedia studios, during two 90-minute sessions per participant. Participants were asked to record a minimum of 300 randomly drawn sentences in either one or two languages (with a minimum of 150 in each). In reality though, they were encouraged to record as many sentences as possible within the time available, not only in their mother tongue (typically Spanish or Catalan), but also in the other two languages under consideration, even if low-proficient (which is often the case in English); indeed, they were allowed to skip sentences when unsure about their correct pronunciation. As shown in Table 4, the net effect of this encouragement was more participants contributing in multiple languages rather than just one, which is different from what happens with poliMedias themselves (see Table 2), though good for our purposes.

Table 5 shows the number of sentences and duration in hours collected in our *DeX Text-To-Speech (Dex-TTS)* dataset of clean lecturer speech data. In total, it comprises 59 h of clean speech data from 47 K sentences uttered by 98 participants. Looking at it row by row, it can be seen that Spanish, Catalan and English account for around 61%, 15% and 24% of the data (both in terms of sentences and recorded speech), respectively. By columns, we can observe that most of the data comes from multilingual acquisitions, either bilingual (42%) or trilingual (23%), meaning that only some 35% of the data corresponds to monolingual participants.

The DeX-TTS dataset is undoubtedly a very valuable resource to test modern TTS technology at the UPV and also an example that can be easily replicated in other universities. On the one hand, TTS technology does not require vast amounts of manually transcribed speech data, as ASR does, but simply a relatively small corpus of clean speech. Indeed, our corpus is similar in size to those commonly used in TTS research (cf. Shen et al., 2018 and Ren et al., 2019). On the other hand, being produced at the UPV by its academic staff, the DeX-TTS dataset is an optimal resource to explore how a UPV lecturer's speech can be best cloned, not only in her/his mother tongue, but also in other languages she/he might not even speak. In this regard, our corpus can be considered a good example of linguistic diversity at a higher education institution, where the dominant official language (Spanish) coexists with a minority yet official language (Catalan) and English. As a result, the DeX-TTS dataset is rich in Spanish speech data but not so rich in Catalan and (non-native) English speech.

3. Cross-lingual voice cloning at the UPV

As described in the introduction, our work on (cross-lingual) voice cloning at the UPV relies on modern AI tools to produce cost-effective multilingual subtitles and synthesized speech for poliMedias. This is clearly illustrated by the production pipeline diagram shown in Fig. 2. The process begins with a new poliMedia uploaded to MediaUPV, including its speaker (lecturer) and (source) language IDs. The first pipeline step (ASR) consists in automatically transcribing the new poliMedia to produce raw source subtitles, which can be optionally reviewed (post-edited) if convenient. In the second step (MT), source subtitles (transcriptions) are machine-translated into a number of target languages (e.g. into Catalan and English if the source language was Spanish). As with transcriptions, target subtitles (translations) can also be post-edited if convenient. TTS comes as the third and final pipeline step; in it, the speaker is automatically voice-cloned (dubbed) for each target language from the corresponding target subtitles. Note that each of the three pipeline steps requires specific models that need to be trained in advance from appropriate training data. The reader is referred to Valor-Miró et al. (2018) for more details on the first two steps of the production pipeline. In what follows, our focus is on the TTS step, for which we assume (reviewed) translations to be available in each target language of interest.

Until a few years ago, conventional TTS systems consisted of diverse, handcrafted components requiring highly specialized expert knowledge of both acoustics and linguistics. Moreover, they were normally restricted to a single speaker and language, making them impractical for massive voice cloning even in just a single language. However, driven by the deep learning revolution and an increased interest among big technology companies, the field of TTS has recently seen large improvements in quality, flexibility and capabilities. In brief, conventional TTS approaches have been surpassed by *end-to-end neural network architectures* (Shen et al., 2018; Ren et al., 2019; Ping et al., 2018) and *neural vocoders* (Oord et al., 2016; Kalchbrenner et al., 2018b). In particular, Google's *Tacotron-2* has become the *de facto* standard architecture for end-to-end TTS (Shen et al., 2018). Compared to previous TTS technology, end-to-end TTS does not require highly specialized expert knowledge, achieves higher degrees of speech naturalness (Shen et al., 2018), and can be easily extended to deal with the general multilingual and multi-speaker setting (Zhang et al., 2019). As discussed in the introduction, this generality is a key feature of the new end-to-end neural architectures, as it opens the door to massive machine dubbing of educational videos, even in target languages of which the speaker has little or no command. With this idea in mind, an extension of Tacotron-2 (Shen et al., 2018) for multiple speakers and languages was developed after completing the DeX-TTS dataset, which is referred to below as *Tacotron2-UPV*. Its basic architecture is depicted in Fig. 3.

As shown in Fig. 3, Tacotron2-UPV is an auto-regressive sequence-to-sequence model with attention that predicts a sequence of mel spectrogram frames from an input phoneme sequence. The encoder consists of learned 512-dimensional phoneme embeddings that are

Table 5
Number of sentences and duration in hours of the clean speech data collected in Spanish (es), Catalan (ca), English (en), bilingual combinations and the trilingual case.

		Monolingual			Bilingual			Trilingual	Total	%
		es	ca	en	es-ca	es-en	ca-en	es-ca-en		
No. of	es	14.6	–	–	4.1	6.7	–	3.5	28.9	61
sentences	ca	–	0.3	–	2.7	–	0.5	3.8	7.3	15
(×1000)	en	–	–	1.0	–	5.5	0.6	4.0	11.1	24
Total			15.9			20.1		11.3	47.3	–
%		34			42			24	–	100
Duration	es	19.2	–	–	5.4	8.0	–	3.7	36.3	62
in hours	ca	–	0.4	–	3.4	–	0.6	4.1	8.5	14
	en	–	–	1.3	–	6.9	0.7	5.1	14.0	24
Total			20.9			25.0		12.9	58.8	–
%			36			42		22	–	100

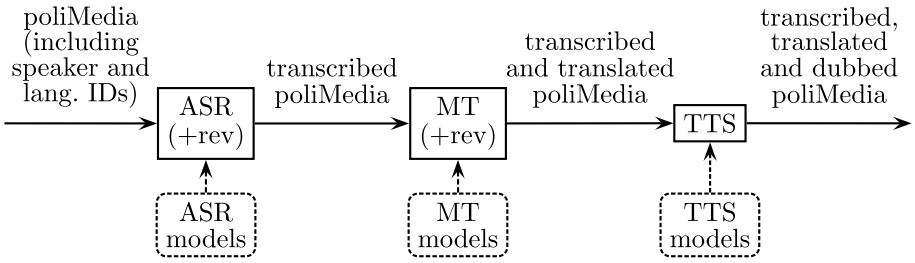


Fig. 2. Production pipeline of transcribed, translated and dubbed poliMedias.

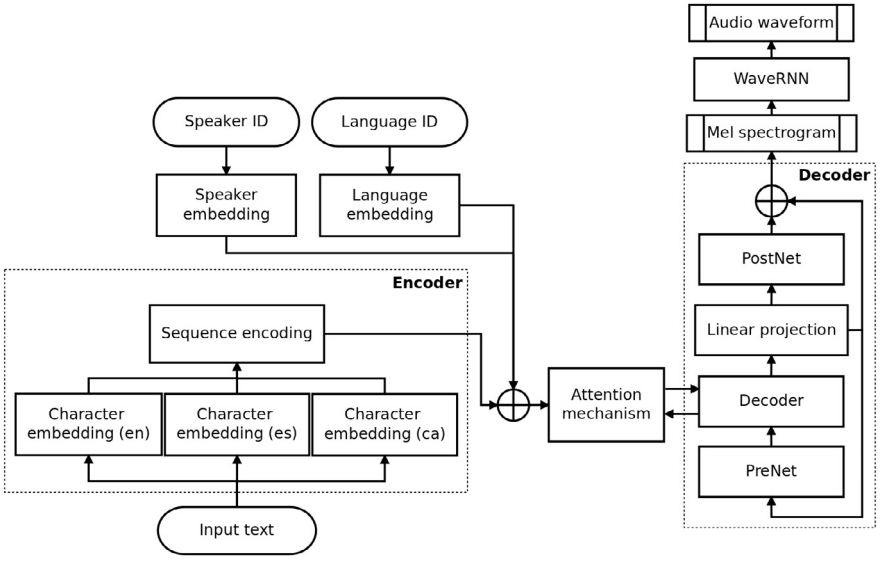


Fig. 3. Basic Tacotron2-UPV architecture.

passed through a stack of three 1-D convolutional layers, followed by batch normalization and ReLU activations. The output of the last convolutional layer is processed by a single bidirectional LSTM layer to generate the encoder hidden states. To deal with multiple speakers and languages, independent speaker and language embeddings are introduced in a way similar to that in Zhang et al. (2019). These embeddings are concatenated to the encoder hidden states before being *consumed* by the attention mechanism. The original *location sensitive attention* is replaced by the *stepwise monotonic attention mechanism* proposed in He et al. (2019). This mechanism constrains input–output alignments to be monotonic with no skipping on inputs, and thus improving inference robustness and training convergence. The decoder is identical to the original Tacotron-2 autoregressive decoder. To this end, the spectrogram frame from time $t-1$ is first passed through a small PreNet comprising 2 fully connected layers of 256 hidden ReLU units each). Here, the PreNet serves as an information bottleneck to leverage the exposure bias problem introduced by the teacher forcing strategy used for training (Liu et al., 2020). Then, the PreNet output and the attention context vector are concatenated and passed through a stack of 2 uni-directional LSTM layers. The LSTM output and the attention context vector are concatenated again, and then linearly projected to the mel-scale dimension. The resulting predicted spectrogram is further improved by adding a residual computed from a convolutional PostNet comprising a stack of 5 1-D convolutional layers. Lastly, the final waveform is generated from the improved spectrogram by using a well-known public implementation of the neural WaveRNN vocoder (Kalchbrenner et al., 2018a; McCarthy, 2018).

At this point it is worth noting that the Tacotron2-UPV system was conceived and developed after completing the DeX-TTS dataset, in parallel yet independently of Google's own multilingual and multi-speaker extension to Tacotron-2 reported in Zhang et al. (2019). However, there are not many differences between the two systems. Indeed, the most

salient difference is in the way multiple languages are taken account of. In Tacotron2-UPV, separate grapheme embeddings per language are used to capture language-dependent particularities at shallow system layers (in the language embedding component), thus facilitating deeper layers (in the attention mechanism and decoder component) to focus more on language-independent patterns of the human voice. In contrast to this, a common set of grapheme/phoneme embeddings for all languages is used by Zhang et al. (2019). All in all, for the purposes of this work, using either Tacotron2-UPV or Google's extension to Tacotron-2 should not make a significant difference.

As indicated earlier on in this section, conventional TTS systems from the pre-deep learning era consisted of diverse signal-processing and linguistic components manually tuned by experts. In contrast to this, and as with *end-to-end* TTS models in general, Tacotron2-UPV can be trained with minimum human intervention (and expert knowledge) from an appropriate collection of <text, audio> pairs. In this regard and as noted above, the DeX-TTS dataset is a very valuable resource as it was acquired with this goal in mind. However, also as noted above, it is rich in Spanish but not so rich in Catalan and English, and thus a TTS system trained only from it will certainly be biased towards Spanish. This is not likely to be an issue for Catalan due to its high similarity to Spanish. However, it is certainly an issue for English, not only because of its comparatively lower degree of similarity, but also due to the limited level of fluency in the non-native English speech recorded. To compensate for this lack of (fluent) English speech data, we also included (part of) the *VCTK* corpus of multi-speaker native English speech for TTS (Yamagishi et al., 2019). More precisely, only speakers with American or British English accents were considered by just adding their speech data to the Tacotron2-UPV training set.

The actual training of the Tacotron2-UPV system was carried out after applying a few common preprocessing steps for TTS data. In particular, the DeX-TTS dataset was preprocessed by first trimming leading and trailing silences, and then applying certain basic audio filters to reduce noise and loudness variability among recordings. All Tacotron2-UPV components but the neural vocoder were jointly trained using an extended version of a publicly available implementation of the basic Tacotron-2 (Mama, 2018). Similarly, the neural vocoder was trained using an open-source implementation of WaveRNN (Kalchbrenner et al., 2018b) by McCarthy (2018). In this way, a complete, fully-trained Tacotron2-UPV system was built to enrich any poliMedia with machine-dubbed audio tracks in its target languages. In this regard, it is worth mentioning that, for the synthesized speech to be (more or less) in synchrony with the video image, machine dubbing is done at the sentence level and aligned in time with source sentences. It also must be noted that, although Tacotron2-UPV was developed thinking primarily about contributors to the DeX-TTS dataset, it can be applied to poliMedias by other authors as well by simply choosing appropriate target speakers.

4. Evaluation

Evaluation of machine learning progress by machine learners is generally driven by widely accepted, *objective* (well-defined) metrics that can be automatically computed by comparing system output and *ground truth* on a set of data samples not used for system training (*test set*). Being able to compute objective metrics in a fully automatic way is seen as a key factor to speed up progress, since not only can researchers thus compare their achievements easily and objectively, but also production of new, improved systems is accelerated by simply running a fully-automated training and testing loop. A good example of this is the WER metric, which has successfully driven the ASR field for decades (Hunt, 1990). Analogously, the BLEU accuracy measure (Papineni et al., 2002) and the WER-inspired *Translation Edit Rate (TER)* metric (Snover et al., 2006) have played a similar role in MT. Needless to say, most important of all for objective metrics is to be highly-correlated with human judgement.

Table 6
Naturalness MOS with 95% confidence intervals per language, including cross-lingual cloning (synthetic samples from lecturer-language pairs *unseen* in training).

Language	Naturalness MOS				
	Synthetic samples			Control	Evaluated
	Seen	Unseen	Total	samples	samples
Spanish	4.1 ± 0.1	3.9 ± 0.3	4.1 ± 0.1	4.5 ± 0.2	533
Catalan	4.2 ± 0.1	4.0 ± 0.1	4.1 ± 0.1	4.8 ± 0.1	551
English	3.6 ± 0.2	3.6 ± 0.1	3.6 ± 0.1	4.3 ± 0.2	594

In contrast to ASR and MT, no objective metrics have gained wide acceptance in TTS and, indeed, most recent work is assessed only by means of *subjective* evaluations (Shen et al., 2018; Ren et al., 2019; Ping et al., 2018). Generally speaking, (listening-type) subjective evaluations boil down to human participants listening to (real and synthetic) speech utterances and giving their feedback on the speech quality, either globally or in terms of individual factors. More precisely, the ITU-T Recommendation P.85 (ITU-T, 1994) is at the basis of most testing methods used for evaluating the subjective quality of synthetic speech. In it, the recommended testing method consists in asking subjects to express their opinion using one or more five-point opinion (Likert) scales. In addition to the overall quality scale, other scales can be considered for measuring listening effort, voice pleasantness, etc. However, by far the preferred way to test and compare current TTS systems is in terms of overall quality only, and on the basis of a *mean opinion score (MOS)* with a 95% confidence interval (Shen et al., 2018; Ren et al., 2019; Ping et al., 2018).

To assess the Tacotron2-UPV system described in Section 3, a call for participation was made to the 98 lecturers contributing to the DeX-TTS dataset (Section 2.2), which was answered by nearly half of them (47). The evaluation procedure was designed around a *test set* of 8820 speech samples synthesized by Tacotron2-UPV. They correspond to 98 lecturers, times 3 languages per lecturer, times 30 sentences for each lecturer-language pair, with sentences randomly picked from poliMedia subtitles not used for Tacotron2-UPV training. Note that many test samples were produced by cross-lingual voice cloning since nearly half (42%) of all lecturer-language pairs were not covered by training data in the DeX-TTS dataset (see Table 4). With this test set at hand, participants were asked to register at a web platform for them to proceed with the evaluation from a user home page (Fig. 4).

As shown in Fig. 4, the evaluation procedure consisted of four parts: *1. Naturalness*, *2. Speaker similarity*, *3. Real or synthetic* and *4. Survey*. It was suggested to start with parts one and two, then optionally move to part three, and finally answer the survey in part four. With the help of a brief progress indicator in each part, participants were allowed to stop and resume the procedure as they wished. In what follows, procedural details and evaluation results are provided for each part separately.

4.1. Naturalness

Naturalness refers to overall speech quality, that is, the main criterion by which current TTS systems are tested and compared. Using a five-point (star) opinion scale, participants were asked to rate the naturalness of a minimum of 50 samples randomly drawn from the test set (Fig. 5). For validation purposes, truly natural (human) speech recordings were also included as control samples among synthetic ones, at random with a ratio of one human recording per six evaluated samples.

Table 6 shows, for each language, the naturalness MOS with 95% confidence intervals for both synthetic and control samples, as well as the number of evaluated samples. The *seen* and *unseen* columns refer to synthetic samples from lecturer-language pairs used and not used, respectively, for Tacotron2-UPV training.

From the results in Table 6, it can be observed that the naturalness MOS on the synthetic speech produced by Tacotron2-UPV is in general

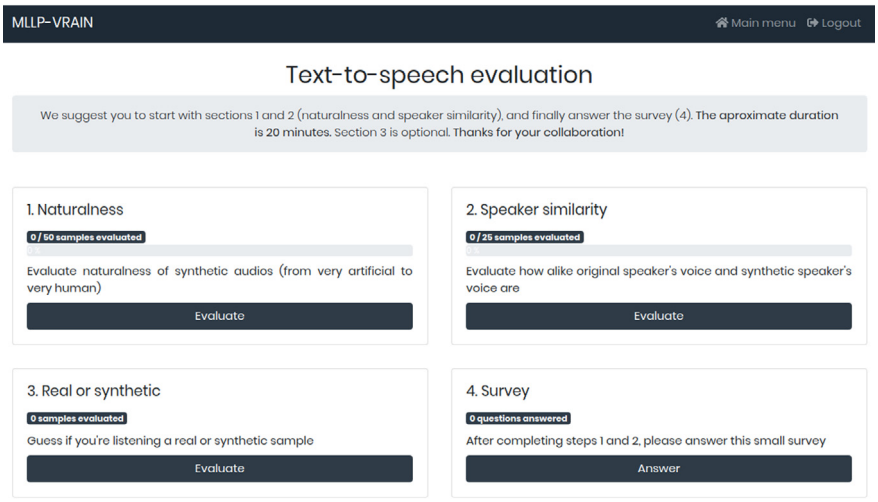


Fig. 4. Home page of the evaluation platform.

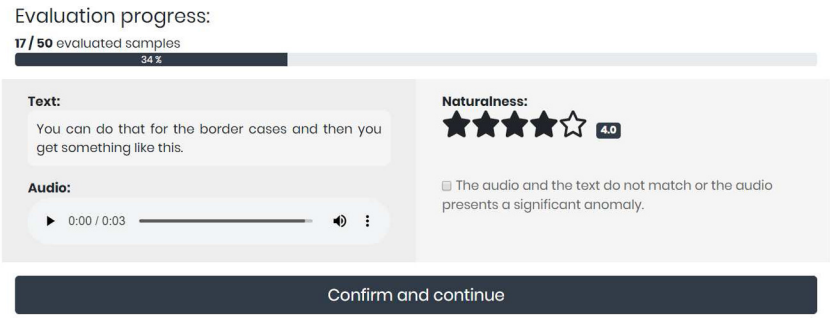


Fig. 5. Naturalness evaluation interface.

fairly good though, as expected, not as good as human speech. In particular, the naturalness of synthetic Spanish and Catalan was judged to be at the very same high rate of 4.1, slightly but significantly below that of human Spanish (4.5) and Catalan (4.8). Similarly, the naturalness of synthetic English was rated at 3.6, again slightly but significantly below that of human speech (4.3). These comparatively lower rates for (synthetic and human) English are certainly due to the non-nativeness nature of the English recordings in the DeX-TTS dataset, from which we get, not surprisingly, a (realistic) non-native bias for English in Tacotron2-UPV. In spite of using the VCTK corpus, this clearly shows that our TTS system effectively learns to mimic the actual non-native speech of UPV lecturers. In any case, summarizing, a main conclusion from Table 6 is that Tacotron2-UPV produces highly natural synthetic speech, not far from human speech. Moreover, by comparing the seen and unseen rates for each language, we see that, in general, synthetic speech naturalness does not depend significantly on which specific lecturer-language pairs were covered in the training data. In other words, Tacotron2-UPV has effectively learned to transfer (clone) lecturer voices from source languages (e.g. mother tongue) to target languages they might not even speak.

At this point, it is worth noting that the above results in terms of naturalness for Tacotron2-UPV do not differ a lot from those reported for Google's own multilingual and multi-speaker extension to Tacotron-2 (Zhang et al., 2019). In brief, Google's results for seen speaker-language pairs are slightly closer to those for control samples while, for unseen pairs, our results are closer. Although Google's naturalness figures are not directly comparable to ours due to the very different nature of the task and data resources considered, the reader is referred to Zhang et al. (2019) for more details.

Table 7
Speaker similarity MOS with 95% confidence intervals per language, for test samples produced from seen and unseen lecturer-language pairs of training data.

Language	Speaker similarity MOS		Evaluated samples
	Seen	Unseen	
Spanish	4.2 ± 0.1	4.0 ± 0.5	324
Catalan	4.1 ± 0.2	4.0 ± 0.2	284
English	3.7 ± 0.2	3.4 ± 0.2	299

4.2. Speaker similarity

Although naturalness is without question the main criterion to judge synthetic speech goodness, it falls short in measuring how similar original (human) and cloned (synthetic) voices actually are. This is particularly relevant for cross-lingual voice cloning since, as pointed out above, it seems that Tacotron2-UPV is capable of cloning voice for unseen lecturer-language pairs almost as well as for seen ones. Needless to say, as this is a feature only available to the most advanced TTS systems, it deserves empirical confirmation. To this end, the second part of the evaluation procedure consisted in rating, on a five-star opinion scale, the speaker similarity between test and training samples. Broadly speaking, speaker similarity is an ill-defined similarity measure depending on diverse perceptual speaker features such as rate, tone, texture or intonation. Each pair of test and training samples was picked at random from the same speaker, but not necessarily from the same language. Participants were asked to do this for a minimum of 25 test samples. Table 7 shows the speaker similarity MOS with 95% confidence intervals for the seen and unseen lecturer-language pairs separately, and the number of evaluated samples.

Table 8
Confusion matrices on the *real or synthetic* test for each language and overall.

Language	Actual condition	Guessed condition		Total samples
		Real	Synthetic	
Spanish	Real	79%	21%	48
	Synthetic	48%	52%	73
Catalan	Real	72%	28%	29
	Synthetic	41%	59%	61
English	Real	66%	34%	32
	Synthetic	32%	68%	69
Overall	Real	73%	27%	109
	Synthetic	40%	60%	203

From the results in Table 7, we can confirm that cross-lingual voice cloning by Tacotron2-UPV works almost as well as conventional voice cloning from seen lecturer-language pairs. Although minor (not significant) yet consistent MOS differences show a slight preference for cloned voice in the seen case, to us this is rather a confirmation that current TTS technology can be safely used for cross-lingual machine dubbing.

As with naturalness, when comparing Tacotron2-UPV with Google's own extension to Tacotron-2 (Zhang et al., 2019), results are mixed. On the one hand, in the seen case, Google reports slightly higher speaker similarity MOS figures. On the other hand, in the unseen case, our speaker similarity results are better. However, as above, these differences are most likely due to the very different nature of the task and data resources considered.

4.3. Real or synthetic

As an extra check to validate MOS results on naturalness and speaker similarity, participants were also invited to optionally run a sort of Turing test to try to guess whether a given speech sample is real (human) or synthetic. This was done in the third part of the evaluation procedure, from speech samples picked at random with a ratio of two synthetic samples per each real one. Table 8 shows the resulting confusion matrix for each language and overall.

Although the number of evaluated samples is modest, we see that the participants misclassified 48%, 41%, 32% and 40% of the synthetic samples in, respectively, Spanish, Catalan, English and overall. Note that the results for Spanish are particularly good since the participants were roughly as accurate as simply deciding at random (for synthetic samples). The results for Catalan and English are also good, though not as good as those for Spanish, particularly in English. This is most likely due to a comparatively lower number of training samples in Catalan and English, and also to the heterogeneity of the English training data. All in all, these results again confirm that the quality of the speech synthesized by Tacotron2-UPV is really close to human speech.

4.4. Questionnaire and comments

The fourth and final part of the evaluation procedure consisted of just two Yes or No control questions on the acceptance of TTS technology, each accompanied by a box for free-text comments and suggestions. Table 9 shows these two control questions and the Yes or No votes received.

As shown in Table 9, all participants think that machine dubbing is useful to improve accessibility and engagement in online educational materials. Also, almost all of them would accept their educational materials to be automatically dubbed in different languages using Tacotron2-UPV.

Apart from the Yes or No feedback, each question originated many comments by participants. On the one hand, we received sixteen comments to the first question: four of them pointed out that there is still room for improvement in pronunciation, nine others were just very positive feedback on the speech synthesis quality and, finally, three comments suggested extending our work to *full* machine translation of poliMedias including slides. On the other hand, thirteen comments were made to the second question: seven of them were to encourage us to deploy TTS technology into production without delay, while the six other comments just requested that lecturers be allowed to review and approve their machine-dubbed materials prior publication. Summarizing, the general view of our study is that TTS technology is not only mature enough for its application at the UPV, but also needed as soon as possible.

Table 9
Final questions and answers on the acceptance of TTS technology.

Questions:	Yes	No
Do you think that the shown automatic dubbing technology can be useful to improve accessibility and engagement in online educational materials?	47	0
Would you accept your educational materials to be automatically dubbed in different languages using this technology?	46	1

5. Conclusions and future work

This work has reported the experience gained on the use of TTS technology at the UPV in recent years, mainly as a guidance for other educational organizations also interested in testing this technology. We have first described the main UPV repository of educational videos, MediaUPV, particularly its largest part of high-quality short video recordings known as poliMedias, which are extensively used by UPV lecturers as "knowledge pills" to support blended learning. Due to its relevance to this work, the poliMedia (sub-)repository has been described in detail: origin, recording procedure and current contents, especially from a linguistic perspective and paying attention to the way publishable multilingual subtitles have been produced cost-effectively. Then we have focused on the main data resource needed to build an in-house, repository-adapted (cross-lingual) TTS system: our lecturer speech database for TTS, its acquisition protocol and basic statistics. This has been followed by a review of our production pipeline of subtitles and cloned voice, particularly in regards to TTS technology, its state-of-the-art and the way it was adapted to train a multilingual and multi-speaker TTS system from our lecturer speech database. Finally, an extensive, subjective evaluation of this TTS system has been reported, including the protocol and support platform we used to acquire the UPV lecturers' opinion on it, and the results obtained. Summarizing, these results show that TTS technology is mature enough for massive machine dubbing of educational videos, even in the cross-lingual case. To us, the door has been opened to producing multilingual educational videos at scale and low cost with both publishable subtitles and cloned lecturer speech of high quality.

Although TTS technology is mature enough for its deployment and operation in higher education, challenges still exist that should be addressed in the (near) future. At the UPV, our most immediate goal is to fully deploy this technology into MediaUPV, not only for poliMedias, but also for poliTubes and lecture recordings (with good audio quality), and thus support the rapidly increasing demand of all these educational video formats. In a later stage, this should be followed by adapting our production pipeline to the streaming setup, so as to extend its applicability to live lecturing, either online or not, delivered under reasonable acoustic conditions. We think that these goals will be achieved sooner rather than later. More importantly, we are convinced that similar developments can be easily made at other educational organizations, either using in-house systems as we do, or third-party systems, if available, under reasonable cost and usage conditions. More generally, multilingual TTS technologies have a large potential for a variety of non-educational application scenarios in further fields such as the media industry, academic conferences and professional translation and interpreting.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We wish first to thank all UPV lecturers who made this study possible. We are also very grateful for the funding support received by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 761758 (X5gon), the Spanish government under grant RTI2018-094879-B-I00 (Multisub, MCIU/AEI/FEDER), and the Universitat Politècnica de València's, Spain PAID-01-17 R&D support programme. Funding for open access charge: CRUE-Universitat Politècnica de València.

References

Roll, I., Russell, D.M., Gašević, D., 2018. Learning at scale. Int. J. Artif. Intell. Educ. 28 (4), 471–477.

Fong, M., Dodson, S., Harandi, N.M., Seo, K., Yoon, D., Roll, I., Fels, S., 2019. Instructors desire student activity, literacy, and video quality analytics to improve video-based blended courses. In: Proc. of the Sixth ACM Conference on Learning @ Scale (L@S).

Morris, N.P., Swinnerton, B., Coop, T., 2019. Lecture recordings to support learning: A contested space between students and teachers. Comput. Educ. 140.

Valor-Miró, J.D., Silvestre-Cerdà, J.A., Civera, J., Turró, C., Juan, A., 2015a. Efficient generation of high-quality multilingual subtitles for video lecture repositories. In: Proc. of the 10th European Conf. on Technology Enhanced Learning (EC-TEL). pp. 485–490.

Valor-Miró, J.D., Baquero-Arnal, P., Civera, J., Turró, C., Juan, A., 2018. Multilingual videos for MOOCs and OER. J. Educ. Technol. Soc. 21 (2), 1–12.

Iranzo, J., Pérez, A., Juan, A., et al., 2019. X5gon Deliverable 3.4: Early Support for Cross-Lingual OER. Tech. rep., Universitat Politècnica de València, https://www.x5gon.org/science/deliverables.

Jorge, J., Pérez, A., Juan, A., et al., 2020b. X5gon Deliverable 3.5: Final Support for Cross-Lingual OER. Tech. rep., Universitat Politècnica de València, www.x5gon.org/science/deliverables.

Jorge, J., Giménez, A., et al., 2020a. LSTM-based one-pass decoder for low-latency streaming. In: Proc. of 2020 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). pp. 7814–7818.

W3C, 2018. Web content accessibility guidelines (WCAG) 2.1. www.w3.org/TR/WCAG21.

Godwin-Jones, R., 2019. In a world of SMART technology, why learn another language? J. Educ. Technol. Soc. 22 (2), 4–13.

van Campen, C.A.K., Segers, E., Verhoeven, L., 2020. Effects of audio support on multimedia learning processes and outcomes in students with dyslexia. Comput. Educ. 150.

Chiou, E.K., Schroeder, N.L., Craig, S.D., 2020. How we trust, perceive, and learn from virtual humans: The influence of voice quality. Comput. Educ. 146.

Shen, J., Pang, R., et al., 2018. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In: Proc. of 2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). pp. 4779–4783.

Zhang, Y., Weiss, R.J., et al., 2019. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. In: Proc. of Interspeech 2019. pp. 2080–2084.

Pérez, A., Iranzo, J., Juan, A., 2019. X5gon Deliverable 5.2: Second Report on Piloting. Tech. rep., Universitat Politècnica de València, www.x5gon.org/science/deliverables.

Piqueras, S., Del-Agua, M.A., Giménez, A., Civera, J., Juan, A., 2014. Statistical text-to-speech synthesis of spanish subtitles. In: Proc. of the 2nd Int. Conf. on Advances in Speech and Language Technologies for Iberian Languages (IberSpeech), Vol. 8854. pp. 40–48.

Turró, C., Ferrando-Bataller, M., Busquets, J., Cañero, A., 2009. Polimedia: a system for successful video e-learning. In: Proc. of the EUNIS Annual Congress.

MediaUPV, 2020. The MediaUPV repository. https://media.upv.es, Retrieved on June 2020.

UPVX, 2020. UPVX: The MOOC initiative at the UPV. https://www.upvx.es, Retrieved on June 2020.

UPValenciaX, 2020. UPValenciaX: UPV as an edX member. https://www.edx.org/school/upvalenciax, Retrieved on June 2020.

ClassCentral, 2020. The 100 most popular online courses of all time (2020). https://www.classcentral.com/report/most-popular-online-courses, Retrieved on June 2020.

Turró, C., Despujol, I., Cañero, A., Busquets, J., 2014. Deployment and analysis of lecture recording in engineering education. In: Proc. of 2014 IEEE Frontiers in Education Conference (FIE). pp. 1–5.

Opencast, 2020. Opencast. https://opencast.org, Retrieved on June 2020.

BOUPV20, 2020. Official bulletin of the UPV. http://hdl.handle.net/10251/145577, Retrieved on June 2020 (in Catalan and Spanish).

Silvestre-Cerdà, J.A., Pérez, A., Jiménez, M., Turró, C., Juan, A., Civera, J., 2013. A system architecture to support cost-effective transcription and translation of large video lecture repositories. In: Proc. of 2013 IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC). pp. 3994–3999.

Pérez, A., Silvestre-Cerdà, J.A., Valor-Miró, J.D., Civera, J., Juan, A., 2015. MLLP transcription and translation platform. In: Proc. of the 10th European Conf. on Technology Enhanced Learning (EC-TEL).

Valor-Miró, J.D., Silvestre-Cerdà, J.A., Civera, J., Turró, C., Juan, A., 2015b. Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories. Speech Commun. 74, 65–75.

Baquero-Arnal, P., Jorge, J., Giménez, A., Silvestre-Cerdà, J.A., Iranzo-Sánchez, J., Sanchis, A., Civera, J., Juan, A., 2020. Improved hybrid streaming ASR with transformer language models. In: Proc. of 21st Annual Conf. of the Intl. Speech Communication Association (InterSpeech 2020). Shanghai (China), pp. 2127–2131. URL http://dx.doi.org/10.21437/Interspeech.2020-2770.

Hunt, M.J., 1990. Figures of merit for assessing connected-word recognisers. Speech Commun. 9 (4), 329–336.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. BLEU: a method for automatic evaluation of machine translation. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 311–318.

Ren, Y., Ruan, Y., et al., 2019. FastSpeech: Fast, robust and controllabletext to speech. In: Proc. of the 33rd Conf. on Neural Information Processing Systems (NeurIPS).

Ping, W., Peng, K., et al., 2018. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In: Proc. of the Sixth Int. Conf. on Learning Representations (ICLR).

Oord, A.v., Dieleman, S., et al., 2016. WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.pdf.

Kalchbrenner, N., Elsen, E., et al., 2018b. Efficient neural audio synthesis. In: Proc. of the 35th International Conference on Machine Learning (ICML 2018), Vol. PMLR 80. pp. 2410–2419.

He, M., Deng, Y., He, L., 2019. Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural TTS. pp. 1293–1297. http://dx.doi.org/10.21437/Interspeech.2019-1972.

Liu, R., Sisman, B., Li, J., Bao, F., Gao, G., Li, H., 2020. Teacher-student training for robust tacotron-based TTS. http://dx.doi.org/10.1109/ICASSP40776.2020.9054681.

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., van den Oord, A., Dieleman, S., Kavukcuoglu, K., 2018a. Efficient neural audio synthesis. In: Dy, J., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 80, PMLR, pp. 2410–2419, URL http://proceedings.mlr.press/v80/kalchbrenner18a.html.

McCarthy, O., 2018. WaveRNN. github.com/fatchord/WaveRNN.

Yamagishi, J., Veaux, C., MacDonald, K., 2019. CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (Version 0.92). Centre for Speech Technology Research (CSTR), University of Edinburgh.

Mama, R., 2018. Tacotron-2. github.com/Rayhane-mamah/Tacotron-2.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J., 2006. A study of translation edit rate with targeted human annotation. In: Proc. of the Association for Machine Translation in the Americas (AMTA). pp. 223–231.

ITU-T, ITU-T, 1994. ITU-T Recommendation P.85: A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices. Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T), https://www.itu.int/rec/T-REC-P.85-199406-I/en.