UNIVERSITAT POLITÈCNICA DE VALÈNCIA
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

Deep Neural Networks for Automatic Speech-To-Speech Translation of Open Educational Resources

Author: Alejandro Pérez González de Martos

Directors: D. Alfons Juan Císcar
D. José Alberto Sanchis Navarro

April, 2022

Deep Neural Networks for Automatic Speech-To-Speech Translation of Open Educational Resources

Alejandro Pérez González de Martos

Thesis performed under the supervision of doctors Alfons Juan Císcar and Albert Sanchis Navarro and presented at the Universitat Politècnica de València in partial fulfilment of the requirements for the degree of *Doctor en Informàtica*.

València,
April 8, 2022

Work supported by the EU's H2020 research and innovation programme under grant agreement no. 761758 (X5gon), the Spanish government under grant RTI2018-094879-B-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", and the Erasmus+ Education program under grant agreement no. 20-226-093604-SCH.

Acknowledgements

En primer lugar, quiero expresar un especial agradecimiento a mis directores de tesis Alfons Juan y Albert Sanchis por todo su apoyo, confianza, esfuerzo y empatía durante la consecución de este trabajo. Me gustaría además extender este reconocimiento a Jorge Civera, y a ellos tres agradecerles enormemente la confianza depositada en mí al integrarme en su grupo de investigación MLLP-VRAIN y que han prorrogado categóricamente durante todos estos años. Estaré siempre en deuda con ellos. Aquí, por un lado, he tenido la oportunidad de aprender y experimentar algunas cosas sobre este apasionante campo de la inteligencia artificial y el aprendizaje automático. Por otro, he tenido la suerte de conocer a personas extraordinarias, por supuesto en el ámbito profesional, pero especialmente en lo personal. Hablo de Joan Albert, Gonçal, Adrià G., Nico, Miguel, Adrià M., Santi, Javi I., Javi J. y Pau. Si esta tesis sirve como excusa para celebrar una de nuestras largas noches de pizza y *TowerFall*, habrá merecido la pena.

También me gustaría agradecer profundamente y dedicar este trabajo y esfuerzo a mi familia, y en especial a mi madre Eva, a mi padre Manuel y a mi hermano Fernando. Ellos han sido y serán siempre una fuente de apoyo incondicional, confianza ciega e inmenso cariño en mi vida. Se lo debo todo.

Alejandro Pérez
Valencia
Febrero 2022

Abstract

In recent years, deep learning has fundamentally changed the landscapes of a number of areas in artificial intelligence, including computer vision, natural language processing, robotics, and game theory. In particular, the striking success of deep learning in a large variety of natural language processing (NLP) applications, including automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS), has resulted in major accuracy improvements, thus widening the applicability of these technologies in real-life settings.

At this point, it is clear that ASR and MT technologies can be utilized to produce cost-effective, high-quality multilingual subtitles of video contents of different kinds. This is particularly true in the case of transcription and translation of video lectures and other kinds of educational materials, in which the audio recording conditions are usually favorable for the ASR task, and there is a grammatically well-formed speech. However, although state-of-the-art neural approaches to TTS have shown to drastically improve the naturalness and quality of synthetic speech over conventional concatenative and parametric systems, it is still unclear whether this technology is already mature enough to improve accessibility and engagement in online learning, and particularly in the context of higher education. Furthermore, advanced topics in TTS such as cross-lingual voice cloning, incremental TTS or zero-shot speaker adaptation remain an open challenge in the field.

This thesis is about enhancing the performance and widening the applicability of modern neural TTS technologies in real-life settings, both in offline and streaming conditions, in the context of improving accessibility and engagement in online learning. Thus, particular emphasis is placed on speaker adaptation and cross-lingual voice cloning, as the input text corresponds to a translated utterance in this context.

Resumen

En los últimos años, el aprendizaje profundo ha cambiado significativamente el panorama en diversas áreas del campo de la inteligencia artificial, entre las que se incluyen la visión por computador, el procesamiento del lenguaje natural, robótica o teoría de juegos. En particular, el sorprendente éxito del aprendizaje profundo en múltiples aplicaciones del campo del procesamiento del lenguaje natural tales como el reconocimiento automático del habla (ASR), la traducción automática (MT) o la síntesis de voz (TTS), ha supuesto una mejora drástica en la precisión de estos sistemas, extendiendo así su implantación a un mayor rango de aplicaciones en la vida real.

En este momento, es evidente que las tecnologías de reconocimiento automático del habla y traducción automática pueden ser empleadas para producir, de forma efectiva, subtítulos multilingües de alta calidad de contenidos audiovisuales. Esto es particularmente cierto en el contexto de los vídeos educativos, donde las condiciones acústicas son normalmente favorables para los sistemas de ASR y el discurso está gramaticalmente bien formado. Sin embargo, en el caso de TTS, aunque los sistemas basados en redes neuronales han demostrado ser capaces de sintetizar voz de un realismo y calidad sin precedentes, todavía debe comprobarse si esta tecnología está lo suficientemente madura como para mejorar la accesibilidad y la participación en el aprendizaje en línea. Además, existen diversas tareas en el campo de la síntesis de voz que todavía suponen un reto, como la clonación de voz inter-lingüe, la síntesis incremental o la adaptación *zero-shot* a nuevos locutores.

Esta tesis aborda la mejora de las prestaciones de los sistemas actuales de síntesis de voz basados en redes neuronales, así como la extensión de su aplicación en diversos escenarios, en el contexto de mejorar la accesibilidad en el aprendizaje en línea. En este sentido, este trabajo presta especial atención a la adaptación a nuevos locutores y a la clonación de voz inter-lingüe, ya que los textos a sintetizar se corresponden, en este caso, a traducciones de intervenciones originalmente en otro idioma.

Resum

Durant aquests darrers anys, l'aprenentatge profund ha canviat significativament el panorama en diverses àrees del camp de la intel·ligència artificial, entre les quals s'inclouen la visió per computador, el processament del llenguatge natural, robòtica o la teoria de jocs. En particular, el sorprenent èxit de l'aprenentatge profund en múltiples aplicacions del camp del processament del llenguatge natural, com ara el reconeixement automàtic de la parla (ASR), la traducció automàtica (MT) o la síntesi de veu (TTS), ha suposat una millora dràstica en la precisió i qualitat d'aquests sistemes, estenent així la seva implantació a un ventall més ampli a la vida real.

En aquest moment, és evident que les tecnologies de reconeixement automàtic de la parla i traducció automàtica poden ser emprades per a produir, de forma efectiva, subtítols multilingües d'alta qualitat de continguts audiovisuals. Això és particularment cert en el context dels vídeos educatius, on les condicions acústiques són normalment favorables per als sistemes d'ASR i el discurs està gramaticalment ben format. No obstant això, al cas de TTS, encara que els sistemes basats en xarxes neuronals han demostrat ser capaços de sintetitzar veu d'un realisme i qualitat sense precedents, encara s'ha de comprovar si aquesta tecnologia és ja prou madura com per millorar l'accessibilitat i la participació en l'aprenentatge en línia. A més, hi ha diverses tasques al camp de la síntesi de veu que encara suposen un repte, com ara la clonació de veu inter-lingüe, la síntesi incremental o l'adaptació *zero-shot* a nous locutors.

Aquesta tesi aborda la millora de les prestacions dels sistemes actuals de síntesi de veu basats en xarxes neuronals, així com l'extensió de la seva aplicació en diversos escenaris, en el context de millorar l'accessibilitat en l'aprenentatge en línia. En aquest sentit, aquest treball presta especial atenció a l'adaptació a nous locutors i a la clonació de veu interlingüe, ja que els textos a sintetitzar es corresponen, en aquest cas, a traduccions d'intervencions originalment en un altre idioma.

Contents

Chapter 1

Introduction

1.1 Framework and motivation

This work is framed in the context of a series of EU and Spanish Government funded research and innovation projects related to the development and application of state-of-the-art speech and language technologies into different kind of online learning environments (OER repositories, MOOC platforms, etc). However, it is worth noting that the research work developed in this thesis is not part of any of these projects, which are just introduced in what follows to provide the chronological context preceding this work for a better understanding of the scientific and technological goals pursued in this thesis.

The first of these projects was the European Union's FP7 transLectures project (2011-2014). The transLectures project aimed to apply state-of-the-art automatic speech recognition (ASR) and machine translation (MT) technologies for transcribing and translating educational resources with best accuracy, and developing a platform for the seamless integration of such technologies into large video-lecture repositories. The goal was to break the language barrier for the consumption of online educational materials of all kinds (video-lectures, MOOCs, OER, etc) by providing accurate-enough automatic subtitles in different languages.

After transLectures ended in 2014, other research projects with similar goals followed: EU's CIP European Multiple MOOC Aggregator (EMMA, 2014-2016), MINECO's Multilingual Open Resources for Education (MORE, 2016-2018), EU's Horizon 2020 Cross Modal, Cross Cultural, Cross Lingual, Cross Domain and Cross Site Global OER Network (X5GON, 2017-2020), Multilingual subtitling of classrooms and plenary sessions (Multisub, 2019-2021)

supported by the Spanish Ministry of Science and Innovation, and finally EXPERT (2022) supported by the Erasmus+ Education programme. There has been therefore an ongoing line of research around improving these technologies for more than 10 years, not only in terms of using more powerful machine learning architectures and algorithms but also in terms of their applicability and adaptation to particular use cases and applications.

The results achieved in transLectures and EMMA were very successful, and it was clear after 2016 that ASR and MT technologies could be used to produce accurate enough multilingual subtitles of video lecture materials. In fact, since 2014 these technologies are being used in *UPV[Media]* (Universitat Politècnica de València's institutional video lecture repository) to generate automatic multilingual subtitles of the digital content generated by UPV lecturers. With these technologies at hand, one could think of additional steps in order to generate automatic *comprehensive* translations of these materials (i.e. as if they were initially conceived in the target language). To that end, the next natural step would be to integrate state-of-the-art text-to-speech (TTS) technologies into the existing speech translation (ASR + MT) pipeline, that is, to build automatic speech-to-speech (S2S) translation systems, which is the central aim of this work. The complete S2S translation pipeline (ASR + MT + TTS) will enable the *automatic dubbing* of educational resources by generating translated synthetic audio tracks of these materials. In this context, the TTS component should not only generate realistic speech in the target language in a predefined set of voices, but ideally keep the original lecturer voice characteristics even when he/she does not speak the target language (this is referred to as cross-lingual voice cloning).

There are a number of benefits introduced by enabling content dubbing with respect to subtitling. First, students do not need to split their attention between what is seen on the screen (presentation slides, blackboard annotations, etc.) and the subtitle text. Second, visually impaired people can also benefit from the translated content. Finally, in some cultures or countries there is a mainstream preference on dubbing over subtitling for the consumption of audiovisual materials [Zar00]. Other areas for which TTS tools have provided support include second language learning [God19], reading difficulties [Cam+20] and virtual humans [Chi+20].

In the recent years, deep learning based TTS has drastically improved the quality and naturalness of synthesized speech. In particular, the neural end-to-end approach to TTS has shown not only to be able to generate realistic and high quality synthetic speech [She+18], but has also significantly simplified the training pipeline and, thus, the expert knowledge (particularly regarding

signal processing) required to build TTS systems. However, neural TTS is a very recent technology that is only slowly being adopted in commercial systems as of 2018, when this thesis is conceived, where concatenative and parametric systems are still widely used due to their better robustness and efficiency. This is due on the one hand to the high computational requirements of neural TTS systems and, on the other hand, to some instability issues presented by encoder-decoder attention-based models. Both aspects will be addressed in detail along this thesis.

1.2 Scientific and technological goals

This section describes the scientific and technological goals pursued in this work.

1. Adapt neural end-to-end TTS architectures to multilingual and multi-speaker settings, particularly focusing on cross-lingual voice cloning, and assess the application of this technology for the automatic dubbing of Open Educational Resources (OER) in the context of higher education.

2. Improve robustness, efficiency and controllability of neural TTS models, focusing on its applicability in real-life production-ready environments.

3. Devise optimum ways of including a neural TTS component into a simultaneous speech-to-text translation pipeline, finding an optimum balance between response times and resulting speech naturalness.

4. Enable zero-shot TTS adaptation to unseen speakers in the context of speech-to-speech translation.

5. Integrate developed technologies into the UPV[Media] automatic transcription and translation production pipeline.

All in all, the main purpose of this thesis is to enhance the performance and widen the applicability of modern neural TTS systems in real-life settings, both in offline and streaming conditions.

1.3 Document structure

This document is structured in seven sequential chapters that cover the different topics and scientific and technological goals proposed in this thesis. First, Chapter 2 gives some preliminary concepts and background knowledge on the research fields covered by this thesis. The first goal concerning the adaptation of neural end-to-end TTS models to multilingual and multi-speaker settings is addressed in Chapter 3, where special emphasis is placed on cross-lingual voice cloning. The assessment of this technology and its application into the field of online learning in higher education is addressed in Chapter 4. Chapter 5 addresses the aspects of robustness, efficiency and cotrollability in neural TTS systems, which relates with the second goal. The third goal is addressed in Chapter 6, which discusses the adaptation of ASR, MT and TTS technologies to streaming conditions in the form of a simultaneous speech-to-speech (S2S) pipeline, focusing on the incremental TTS component. Last, Chapter 7 addresses the zero-shot adaptation of TTS models to unseen speakers in the context of speech-to-speech translation for online teaching, corresponding to the fourth goal. The integration of these technologies into UPV[Media] automatic translation pipeline, corresponding with the last goal considered in this thesis, is also addressed in Chapter 7. Finally, Chapter 8 gives a brief summary of the work described along the previous chapters, highlighting the scientific publications that endorse the scientific impact of the contributions of this thesis, as well as some concluding remarks and future work.

Although this work is made in collaboration with other members of the *Machine Learning and Language Processing* (MLLP) group from UPV, the theoretical framework, developments, experimentation and evaluations presented along this thesis have been carried out primarily (and in most cases exclusively) by the author of this work except when mentioned otherwise. When applicable, individual contributions are listed at the end of each chapter for the work done in collaboration with others.

The experienced reader can skip the preliminary concepts given in Chapter 2, with maybe the exception of Section 2.7 introducing the neural end-to-end approach to TTS. A sequential reading of the seven chapters of this document is encouraged if the reader wants to learn about the whole work. However, specific chapters can be read attending to the particular interests on the different aspects of neural TTS models addressed within this thesis.

Chapter 2

Preliminaries

In this chapter, some general concepts and terminology closely related to the machine learning models and algorithms used throughout this work is introduced. The intention is not to dig into the details, but instead present a brief introduction for each of the different concepts presented below.

The Machine Learning and Pattern Recognition fields are introduced in Section 2.1. Section 2.2 introduces the attention mechanism in the context of Sequence-to-Sequence problems. The most relevant aspects of the transformer architecture are briefly introduced in Section 2.3. Section 2.4 introduces generative adversarial networks. Then, Sections 2.5, 2.6 and 2.7 introduce, respectively, the ASR, MT and TTS natural language processing tasks. Finally, Section 2.8 introduces the Speech-to-Speech Translation task both under offline and online (streaming) conditions.

2.1 Machine Learning

Machine Learning is a branch of the wider *Artifical Intelligence* (AI) field. It focuses on the development of computer systems or applications that are able to learn from data or past experience to solve a particular task [Mur22].

Attending to the nature of the output variables, machine learning problems are mainly encompassed in two different types:

- **Classification:** Classification tasks aim to predict a label or set of labels representing a category $y \in \{1, ..., C\}$ (discrete variables or labels). In the research community, this has been historically known as *pattern recognition*, which refers to the task of finding regularities in data by using machine learning algorithms.

- **Regression:** Regression tasks deal with the estimation of numerical values $y \in \mathbb{R}$ (continuous variables). A simple example would be to predict the price of a house given its size and the distance to the city centre.

Machine learning algorithms can also be divided attending to the nature of the data used for training the underlying models into three main paradigms:

- **Supervised learning:** In supervised learning, models learn from a collection of N input-output pairs $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ known as *training set*. Formally, we intend to learn a probability distribution over the output variable conditioned on the input data, $p(y|x)$.
- **Unsupervised learning:** In unsupervised learning, models try to learn patterns from unlabeled data. In contrast to supervised learning, it intends to learn the prior distribution $p(x)$ of the data.
- **Reinforcement learning:** In reinforcement learning, models (also called *agents*) are trained to make a sequence of decisions (*actions*) in an uncertain, potentially complex environment. It differs from supervised learning in not needing a collection of input-output pairs to correct sub-optimal actions. Instead, the agent should learn an optimal policy that maximizes a reward function that accumulates from occasional rewards (or punishment) in response to the actions that it takes.

ASR and MT are classification tasks, since the output is a discrete variable (text sentence), while TTS is addressed as a regression task. All three are regularly addressed under the supervised learning paradigm, where large amounts of manually labeled data (i.e. transcribed speech for ASR and TTS, bilingual sentence pairs for MT) are used to train the machine learning models involved.

A machine learning model learns from data by (iteratively) adjusting its parameters θ to optimize an *objective function* that provides an objective measure of the performance of the model. By convention, objective functions are defined so that *lower is better*, and thus these are sometimes called *loss functions*. In regression tasks, a common training criterion is to minimize either the mean absolute error (MAE or $\ell 1$):

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{2.1}$$

or the mean squared error (MSE or $\ell 2$):

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.2}$$

between predicted $\hat{y}$ and ground truth y values, where $\hat{y}_i = p_\theta(x_i)$.

In contrast, in classification tasks it is common to minimize the negative log likelihood of the model's predictions (i.e. perform a maximum likelihood estimation), which is equivalent to minimizing the binary cross entropy of the model:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log \hat{y}_i \tag{2.3}$$

where y and $\hat{y}$ represent ground truth and predicted discrete probability distributions over the possible classes.

In the last ten years, deep learning models based on artificial neural networks (ANN) have been shown to outperform classical machine learning methods in a wide variety of tasks. In what follows, it is assumed the reader is already familiar with basic deep learning concepts and architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The reader is refered to [Mur22, Chapters 13, 14 and 15] for more details.

2.2 Sequence-to-Sequence with Attention Mechanism

In machine learning, sequential data problems where both the input and output are variable-length sequences are known as Sequence-to-Sequence (Seq2Seq) problems. For example, in neural end-to-end approaches to ASR, MT or TTS these are tackled as Seq2Seq problems in which the input corresponds to a sequence of acustic features (ASR) or words (MT, TTS) and the output is a sequence of words (ASR, MT) or acoustic features (TTS).

To address Seq2Seq tasks, Encoder-Decoder models [SVL14] were originally proposed in the context of MT to map variable-length inputs and outputs using recurrent long short-term memory (LSTM) neural networks [HS97]. As illustrated in Figure 2.1, the encoder part of the model processes the input and maps it into a fixed length latent representation. Ideally, this encoded

representation holds all the input information needed to perform the task at hand (e.g. the translation). In the case of MT, the decoder model, starting from this encoded representation, generates the output sentence word by word. In this context, the decoder model can be seen as a language model conditioned on the encoded representation.

Formally, the encoder processes an input sequence $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_n]$ of N elements and returns state representations $\mathbf{z} = [\mathbf{z}_1, ..., \mathbf{z}_n]$, called *encoder hidden states* or *encoder outputs*. The decoder takes $\mathbf{z}_n$ and generates the output sequence $\mathbf{y} = [\mathbf{y}_1, ..., \mathbf{y}_T]$ left-to-right, as follows:

$$\mathbf{y}_t = \text{dec}(\mathbf{y}_1^{t-1}, \mathbf{z}_n) \tag{2.4}$$

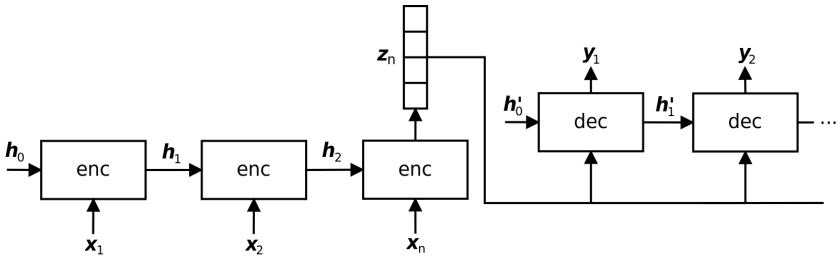


Figure 2.1: Encoder-Decoder architecture.

There is though a primary drawback to this architecture: it has very limited memory. A single vector of fixed dimensionality ($\mathbf{z}_n$) is used to hold all the relevant input information (which can be a sequence of dozens of words from a very large vocabulary). This issue is believed to be more of a problem when decoding particularly long sequences.

To overcome this issue, the *attention mechanism* was introduced for Seq2Seq models in 2015 [Bah+15]. The idea behind the attention mechanism is to replace the fixed context vector ($\mathbf{z}_n$) with a context vector $\mathbf{c}_t$ computed dynamically at each decoding step as:

$$\mathbf{c}_t = \sum_{i=1}^{N} \alpha_{t,i} \mathbf{z}_i \tag{2.5}$$

where α_t are the weights assigned to each encoder hidden state at step t, with the most relevant vectors being attributed the highest weights. These are computed using the softmax function to make them differentiable as follows:

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^{N} \exp(e_{t,k})} \tag{2.6}$$

ensuring that $0 \leq \alpha_{t,i} \leq 1$ for all i and that $\sum_{i=1}^{N} \alpha_{t,i} = 1$, where $e_{t,i}$ is an *attention score* (also known as *energy value*) computed according to some *scoring function*. The scoring function provides a measure of how relevant the encoder hidden state $\mathbf{z}_i$ is at decoding step t. In recurrent Seq2Seq models, it is common to compute this score using the previous decoder hidden state, as $e_{t,i} = a(\mathbf{h}_{t-1}, \mathbf{z}_i)$ for all i, where $a(\cdot)$ is an attention scoring function that can be implemented by a feed-forward network.

From a more general perspective, the attention mechanism can be seen as a soft dictionary look-up, in which we compare a *query* $\mathbf{q} \in \mathbb{R}^{d_q}$ to each *key* $\mathbf{k}_i \in \mathbb{R}^{d_k}$ in a collection of N *key-value pairs* $(\mathbf{k}_1, \mathbf{v}_1), ..., (\mathbf{k}_n, \mathbf{v}_n)$, and then retrieve the corresponding value $\mathbf{v}_i \in \mathbb{R}^{d_v}$. However, instead of retrieving a single value $\mathbf{v}_i$ we retrieve a convex combination of the values. Figure 2.2 shows a general overview of the attention mechanism.

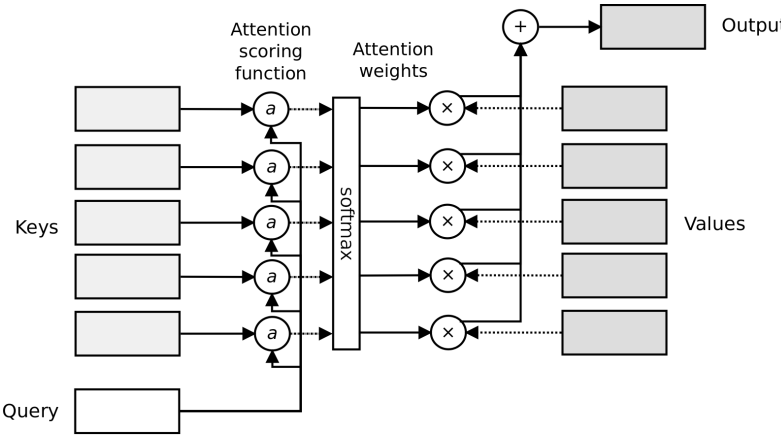


Figure 2.2: Attention mechanism overview adapted from [Zha+21].

The two more relevant options for $a(\cdot)$ are the *additive* [Bah+15] and the *scaled dot-product* [LPM15] scoring functions. The additive scoring function computes attention scores as follows:

$$a(\mathbf{q}, \mathbf{k}) = \mathbf{w}_v^\intercal \tanh(\mathbf{W}_q \mathbf{q} + \mathbf{W}_k \mathbf{k}) \in \mathbb{R} \tag{2.7}$$

where $\mathbf{w}_v \in \mathbb{R}^h$, $\mathbf{W}_q \in \mathbb{R}^{h \times q}$ and $\mathbf{W}_k \in \mathbb{R}^{h \times k}$ are learnable parameters of the model. Differently, the scaled dot-product scoring function requires that both $\mathbf{q}$ and $\mathbf{k}$ have the same dimension d, and it computes attention scores as:

$$a(\mathbf{q}, \mathbf{k}) = \frac{\mathbf{q}^\intercal \mathbf{k}}{\sqrt{d}} \in \mathbb{R} \tag{2.8}$$

2.3 Transformer

Transformers are a type of encoder-decoder model that have revolutionized the field of natural language processing (NLP) over the last few years, becoming the state-of-the-art in many NLP tasks such as language modeling, ASR, MT or TTS. Transformers are also being widely adopted in other fields thanks to their highly parallelizable architecture and their astonishing performance in many conditional sequence modeling tasks. In this work, a particular adaptation of transformers for the TTS task is introduced in Chapter 7.

The transformer architecture [Vas+17], depicted in Figure 2.3, follows an encoder-decoder structure but does not rely on recurrent nor convolutional layers in order to generate an output. Instead, it relies on the novel *self-attention* mechanism and *positional encodings* to allow the model to look into other positions of the sequence when processing a particular input. In addition, the transformer extends the regular attention mechanism with multiple attention heads for improved performance. This is known as *multi-head attention*. While these three concepts are briefly introduced in what follows, the reader is referred to [Mur22, Sec. 15.5] for further details.

Self-attention

In Section 2.2 it was shown how the decoder can attend to the input sequence in order to capture contextual embeddings of each input. In self-attention, the encoder attends to itself to compute *context aware* encodings of the same length as the input sequence, where each input is allowed to pay attention to all the positions of the input sequence.

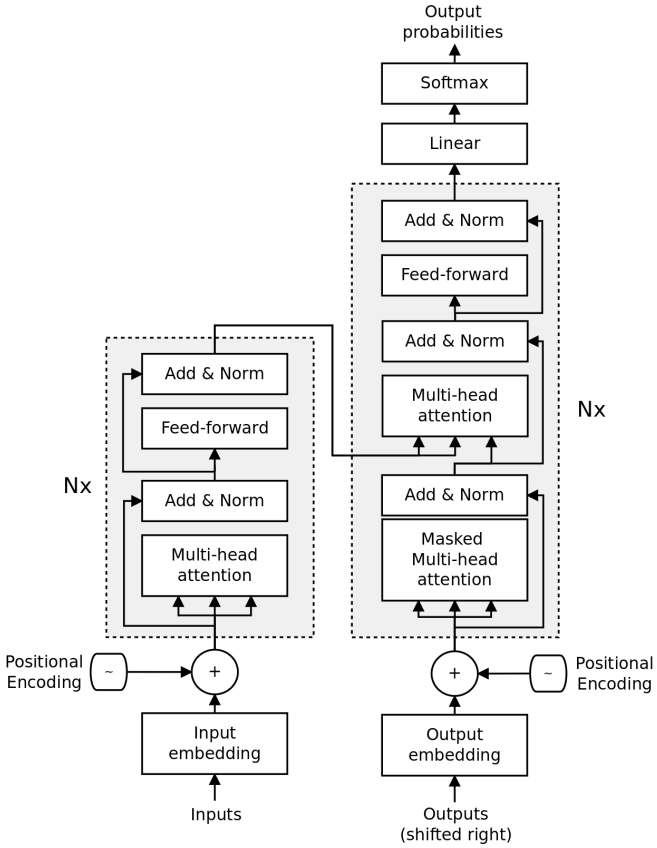


Figure 2.3: The original transformer architecture adapted from [Vas+17].

Formally, given an input sequence $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_n]$ where $\mathbf{x}_i \in \mathbb{R}^d$, self-attention computes an output sequence using $\mathbf{x}_i$ as the query, and $[\mathbf{x}_1, ..., \mathbf{x}_n]$ both as keys and values.

Multi-head attention

The basic idea behind the multi-head attention mechanism is to enable capturing different notions of similarity, such as capturing dependencies of various ranges (e.g. shorter-range vs. longer-range). To that end, multi-head attention allows the attention mechanism to jointly use different representation subspaces of queries, keys and values.

Given a query $\mathbf{q} \in \mathbb{R}^{d_q}$, a key $\mathbf{k} \in \mathbb{R}^{d_k}$ and a value $\mathbf{v} \in \mathbb{R}^{d_v}$, each attention head $\mathbf{h}_i$ is computed as:

$$\mathbf{h}_i = f(\mathbf{W}_i^{(q)}\mathbf{q}, \mathbf{W}_i^{(k)}\mathbf{k}, \mathbf{W}_i^{(v)}\mathbf{v}) \tag{2.9}$$

where $\mathbf{W}_i^{(*)}$ are learnable parameters and f is the attention pooling (e.g. additive or scaled dot-product attention). The multi-head attention output ($\mathbf{h}$) is a linear transformation via learnable parameters $\mathbf{W}_0$ of the concatenation of the n heads:

$$h = \mathbf{W}_0 \begin{bmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_n \end{bmatrix} \tag{2.10}$$

Positional encodings

The performance of vanilla self-attention can be suboptimal since attention is permutation invariant, and hence ignores the input sequence ordering. In the transformer architecture, position embeddings are used to give the order context to the non-recurrent multi-head attention architecture. In [Vas+17], it is proposed to add an absolute positional encoding based on the sine and cosine functions to the input sequences, which are computed as:

$$p_{i,2j} = \sin\left(\frac{i}{C^{2j/d}}\right), \quad p_{i,2j+1} = \cos\left(\frac{i}{C^{2j/d}}\right) \tag{2.11}$$

where C corresponds to a fixed maximum sequence length (e.g. 10000), i is the position, j is the embedding dimension index and d is the total number of dimensions. The advantage of this representation is two-fold. First, unlike learnable position embeddings, it can be computed for arbitrary length inputs up to C. Second, it allows the model to attend relative positions effortlessly, since for any fixed offset k, p_{i+k} can be represented as a linear function of p_i [Vas+17].

2.4 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are an approach to generative modeling using deep learning methods, designed by Ian Goodfellow et. al in 2014 [Goo+14].

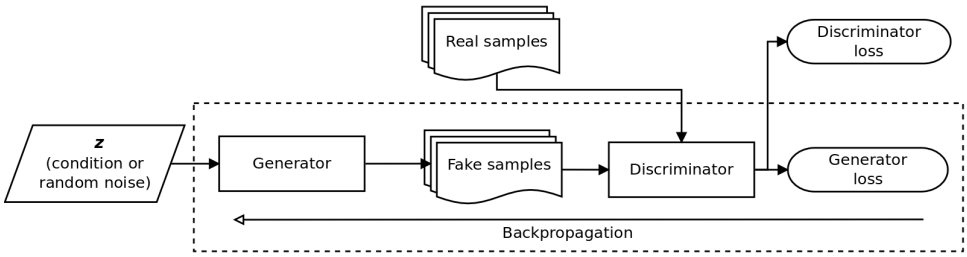


Figure 2.4: Generative Adversarial Network training scheme.

GANs rely on the idea that a generative model is good if it is difficult to distinguish real samples from fake (generated) samples. Thus, GANs consist of two competing networks: a generator G and a discriminator D. Both models play an adversarial game where the generator tries to fool the discriminator by generating samples that are close to the real distribution of the training data. These samples are accepted or rejected by the discriminator (a binary classifier) as real/fake and, in this process, the generator incrementally updates to improve itself to generate fake samples that are increasingly more realistic. At the same time, the discriminator is trained in a supervised way to learn how to differentiate between real and fake (generated) samples. The training scheme of a GAN is depicted in Figure 2.4.

In the original work, cross-entropy is used as the loss function for the discriminator. Thus, the generator is trained to minimize the following loss:

$$\mathcal{L}(G) = -\log D(G(\mathbf{z})) \tag{2.12}$$

where $\mathbf{z}$ is either a source of randomness (e.g. a normal distribution $z \sim \mathcal{N}(0, 1)$) for unconditional modeling or a conditioning input for conditional modeling. At the same time, the discriminator is trained to minimize the following loss:

$$\mathcal{L}(D) = -\log D(\mathbf{x}) - \log(1 - D(G(\mathbf{z}))) \tag{2.13}$$

where $\mathbf{x}$ is a real training sample. Thus, G and D are playing a *minimax* game with the following comprehensive objective function:

$$\max_{D} \min_{G} \{\mathbb{E}_{\mathbf{x}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}}[\log(1 - D(G(\mathbf{z})))]\} \tag{2.14}$$

In the recent years, GANs have played an important role in building efficient high-quality neural vocoders for the TTS task, which are extensively used along this work (Chapters 5, 6 and 7).

2.5 Automatic Speech Recognition

Automatic Speech Recognition (ASR) refers to the task of transcribing spoken words (speech) into written text (speech-to-text). Formally, ASR can be viewed as a classification problem in which an observed acoustic sequence $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T]$ is mapped into a word string $y = [y_1, y_2, ..., y_N]$, where y_i are words belonging to a vocabulary $\mathcal{Y}$. In the statistical approach to ASR, given a sequence of acoustic observations $\mathbf{x}$ our aim is to find the most probable word sequence $\hat{y}$ by maximizing the posterior probability $p(y|\mathbf{x})$:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}*} p(y|\mathbf{x}) = \arg\max_{y \in \mathcal{Y}*} \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})} = \arg\max_{y \in \mathcal{Y}*} p(\mathbf{x}|y)p(y) \tag{2.15}$$

where $p(y)$ is modelled by a *language model* (LM) and $p(\mathbf{x}|y)$ is modelled by an *acoustic model*. The statistical language model $p(y)$ estimates the probability for any given sequence of words y.

Traditionally, n-gram based models [KN95] have been widely used in language modeling, not only for ASR but also for many other tasks involving language processing. Nowadays, neural LMs based on recurrent or transformer architectures have become the state-of-the-art in language modeling due to their improved performance over n-gram LMs [Mik+11; Baq+20].

The acoustic model $p(\mathbf{x}|y)$ estimates the likelihood that the sequence of acoustic observations $\mathbf{x}$ has been generated by a sequence of words y. HMMs have been traditionally used to model $p(\mathbf{x}|y)$ [Lee88; Lee90], where state observation probabilities are frequently modelled using either Gaussian mixture models (GMM) or discriminative DNN models. Until deep learning techniques were applied, GMM-HMM had been the dominant framework for speech recognition.

In the recent years, HMM-free end-to-end ASR systems have been extensively investigated [GJ14; Cha+16; CPS16; Pra+19]. Differently from hybrid models, end-to-end models attempt to directly model $p(y|\mathbf{x})$. In particular, transformer-based end-to-end ASR architectures have achieved state-of-the-art performance [Mia+20; Wan+20; Gul+20] in many tasks, outperforming strong hybrid baseline systems [VA21].

2.6 Machine Translation

Machine Translation (MT) refers to the problem of automatically translating a source sentence x into a target sequence y:

$$x = x_1, x_2, ..., x_I \quad x_i \in \mathcal{X}$$

$$y = y_1, y_2, ..., y_J \quad y_j \in \mathcal{Y}$$

where x_i and y_j are words from the source and target sentences, and $\mathcal{X}$ and $\mathcal{Y}$ are the source and target vocabularies, respectively.

In the statistical approach for MT (SMT), this problem is tackled by finding the most probable target word sequence $\hat{y}$ given the source word sequence x:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}*} p(y|x) = \arg\max_{y \in \mathcal{Y}*} p(x|y)p(y) \tag{2.16}$$

where $p(y|x)$ (the probability of y being the translation of x) has been decomposed into a *translation model* $p(x|y)$ and a target *language model* $p(y)$. The translation model is trained on parallel text data (a collection of sentences and their corresponding translation), and is responsible for modeling the correlation between source and target sentences. On the other hand, the language model is trained on monolingual text data in the target language and models the well-formedness of the candidate translation y.

In the recent years, neural approaches for MT (NMT) based on encoder-decoder architectures have been proposed with very successful results, outperforming traditional SMT models and becoming the state-of-the-art in MT. NMT models [KB13; SVL14] directly calculate $p(y|x)$ as:

$$p(y|x) = \prod_{j=1}^{J} p(y_j|y_1^{j-1}, x) \tag{2.17}$$

and the most likely translation (Eq. 2.16) is usually obtained either by greedy or beam search decoding algorithms based on the left-to-right factorization of NMT models. Again, transformer-based NMT architectures have become the state-of-the-art also in machine translation [Vas+17], outperforming CNN and RNN architectures [Wu+16].

2.7 Text-To-Speech

Text-to-speech, or speech synthesis, refers to the artificial production of human speech. The goal is to render symbolic linguistic representations (like text or phonetic transcriptions) into a speech waveform, resembling human speech. Thus, the input is a sequence of discrete values y corresponding to the text or phonetic sequence, and the output $\mathbf{x}$ is a sequence of continuous acoustic feature values. In neural based approaches, $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T]$ is usually the log-mel spectrogram, which is an intermediate lossy representation of the speech waveform. It can be formalized as the modeling of the posterior probability distribution $p(\mathbf{x}|y)$ of the spectrogram $\mathbf{x}$ given the linguistic features $y = [y_1, y_2, ..., y_N]$.

Until the last few years, the two main approaches dominating the TTS field were concatenative speech synthesis (CSS) and statistical parametric speech synthesis (SPSS). The CSS approach produces the target speech waveform by concatenating pieces of speech (called *units*) that are stored in a database [HB96]. Units are carefully selected by a unit selection algorithm. Although CSS can produce high quality and intelligible synthetic speech, it requires of a huge recording database in order to cover all possible combinations of speech units and speakers.

As an alternative to the CSS approach, SPSS generates speech using statistical models instead of relying on pre-recorded segments. SPSS systems are usually composed of three main components: a text analysis module, a parameter prediction module or acoustic model, and a vocoder module. Similarly to ASR systems, SPSS systems are trained on a collection of $<$*text,audio*$>$ pairs. Nevertheless, TTS recordings are usually performed in professional studio settings to assure the best audio quality and proper acoustic conditions (e.g. avoiding background noises, reverberations, etc).

In the early 2010s, some works first introduced DNNs into the SPSS paradigm by replacing the different HMM components with a corresponding DNN [ZSS13; Fan+14]. However, the main advances in speech synthesis naturalness and quality came in the late 2010s with the introduction of end-to-end autore-

gressive neural acoustic and vocoder models. WaveNet [Oor+16], Tacotron 1/2 [Wan+17; She+18] or Deep Voice 3 [Pin+18] are some of the pioneering end-to-end TTS models that rapidly gained popularity in the TTS research community. These models not only reached unprecedented levels of synthetic speech naturalness, but also simplified the training pipeline by directly taking character or phoneme sequences as inputs and mel spectrograms for the acoustic features.

These pioneer works towards end-to-end neural TTS followed a two-stage approach. In a first stage (text-to-spectrogram), an acoustic model generates an intermediate acoustic representation (log-mel spectrograms) from the input text. Then, in a second stage (spectrogram-to-wave), the final waveform is reconstructed conditioned on the predicted acoustic features by means of a separate vocoder model [She+18; Oor+16] or algorithm [Wan+17; GL83]. This two-stage approach has become the predominant paradigm in neural TTS. Figure 2.5 shows a generic two-stage neural TTS architecture comprised of an attention-based encoder-decoder acoustic model and a neural vocoder.

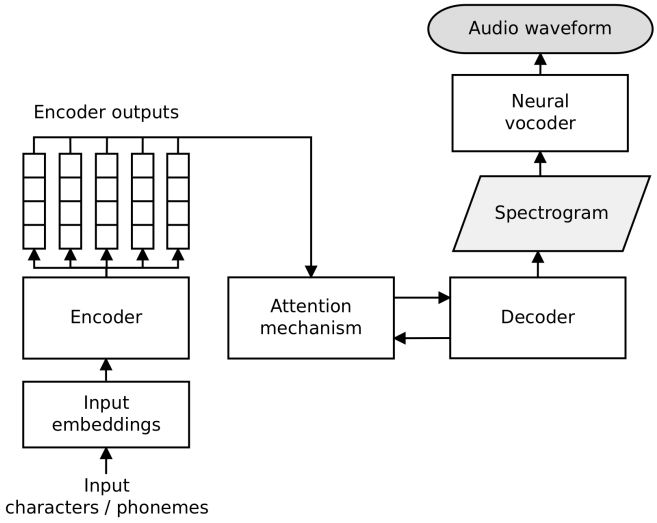


Figure 2.5: A generic two-stage neural TTS architecture comprising an attention-based encoder-decoder model and a neural vocoder.

2.7.1 Text-to-spectrogram

Pioneering works in end-to-end neural TTS systems were based on encoder-decoder architectures with attention [Wan+17; She+18; Pin+18]. In these, the encoder maps a sequence of input character or phoneme embeddings $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$ into a series of annotations or intermediate high-dimensional representations $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N]$. Each annotation $\mathbf{z}_i$ contains information about its corresponding input token $\mathbf{x}_i$ with respect to its neighboring tokens $\mathbf{x}_j$ for all $j \neq i$. Then, an autoregressive decoder iteratively outputs the spectrogram frames $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T]$, where an attention mechanism is used to condition $\mathbf{y}_t$ on the most relevant encoder hidden states $\mathbf{z}$ at step t.

The acoustic model is trained on a collection of $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{w}_i)\}_{i=1}^{N}$ $<$*text,audio*$>$ pairs to minimize either the $\ell 1$ or $\ell 2$ loss between the predicted $\hat{\mathbf{y}}$ and ground truth $\mathbf{y}$ spectrograms extracted from the audio waveforms $\mathbf{w}$. A *teacher forcing* algorithm is used for training the autoregressive model [Mur22, Sec. 15.2.4], where the ground truth spectrogram frame $\mathbf{y}_{t-1}$ is used to condition the decoder output at step t. This introduces a mismatch between training and inference conditions known as *exposure bias* [Liu+20; Ran+16] that will be tackled in more detail in Chapter 5.

2.7.2 Spectrogram-to-wave

In two-stage neural TTS pipelines, neural vocoder models are used to reconstruct the audio waveform given the predicted acoustic features. Differently from the text-to-spectrogram task, vocoder models are trained on a collection of untranscribed audios $\mathcal{D} = \{\mathbf{w}_i\}_{i=1}^{N}$ from which intermediate spectrogram representations can be directly obtained.

The autoregressive WaveNet [Oor+16] was the first neural audio generative model that was able to produce realistic speech with unprecedented levels of naturalness and audio quality. The WaveNet architecture (see Figure 2.6) is based on stacks of causal, dilated convolutional layers capable of achieving a wide receptive field. A dilated convolution skips input values with a certain step so that the filter is applied over an area larger than its length. The joint probability of a waveform $\mathbf{w}_i$ is here factorized as a product of conditional probabilities as follows:

$$p(\mathbf{w}_i) = \prod_{t=1}^{T} p(w_{i,t} | w_{i,1}, ..., w_{i,t-1}) \tag{2.18}$$

The conditional distribution $p(\mathbf{w}_i|\mathbf{y}_i)$, where $\mathbf{y}_i$ is the spectrogram representation of $\mathbf{w}_i$ can thus be modelled as:

$$p(\mathbf{w}_i|\mathbf{y}_i) = \prod_{t=1}^{T} p(w_{i,t}|w_{i,1}, ..., w_{i,t-1}, \mathbf{y}_i) \tag{2.19}$$

Nonetheless, WaveNet inference speed is remarkably slow due to its autoregressive nature. Digital speech signals tend to be sequences of extended length (usually 16K or more values per second), which brings in complications of computation and memory costs.

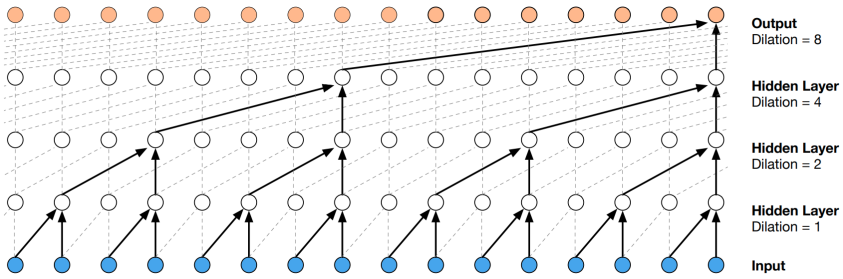


Figure 2.6: WaveNet architecture from [Oor+16].

Over the past 5 years, a lot of attention has been put on building faster and lighter neural vocoder architectures of comparable quality. Parallel WaveNet [Oor+18], FFTNet [Jin+18], WaveRNN [Kal+18], WaveGlow [PVC19] or LPCNet [VS19] are some of the first attempts to match WaveNet audio quality while bringing faster inference speeds. More recently, GAN-based vocoders made breakthrough improvements over autoregressive models in terms of inference speed and computation costs. MelGAN [Kum+19], Parallel WaveGAN [YSK20] or HiFi-GAN [KKB20] are some of the GAN-based models that have gained great popularity over the last few years.

2.7.3 Evaluation metrics

Evaluation of machine learning progress 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 [Hun90]. Analogously, the BLEU accuracy measure [Pap+02a] and the WER-inspired *Translation Edit Rate (TER)* metric [Sno+06] 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.

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 [Oor+16; She+18; Ren+19; Pin+18]. 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 [ITU94] 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 naturalness only, and on the basis of a *mean opinion score (MOS)* with a 95% confidence interval [She+18; Ren+19; Pin+18].

This way, subjective listening tests adopting either the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) methodology or mean opinion score (MOS) ratings have become the *de facto* standard to assess and compare the quality of TTS systems (particularly in terms of speech naturalness). However, when directly comparing two alternative methods, A/B testing is also widely used in the field.

Even though there are no generally-accepted metrics to evaluate TTS quality, a number of objective metrics from general audio processing are available to measure the quality of the synthetic speech. These become particularly useful

in the context of the isolated vocoding task, as it avoids the one-to-many mapping problem present in the more general TTS task.

Mel Cepstral Distortion (MCD)

Mel cepstral distortion (MCD) [Kub93] is a measure of how different two sequences of mel cepstra are:

$$MCD_K = \frac{1}{T} \sum_{t=1}^{T} \sqrt{\sum_{k=1}^{K} (m_{tk} - \hat{m}_{tk})^2} \tag{2.20}$$

where m_{tk} and $\hat{m}_{tk}$ are the k-th mel frequency cepstral coefficient (MFCC) of the t-th frame from the reference and predicted audio, respectively. The MCD is the sum of the squared differences over the first K MFCCs, skipping $k = 0$ (overall energy).

To measure the MCD between sequences of different length, dynamic time warping (DTW) [Mül07] can be used to compute the minimum MCD obtainable by efficiently considering all possible monotone alignments between the two sequences (MCD-DTW).

Perceptual Evaluation of Speech Quality (PESQ)

The Perceptual Evaluation of Speech Quality (PESQ) [Rix+01] is a computational model proposed to approximate human perception of audio quality which allows to predict the results of MOS evaluation. The PESQ, standardized as Recommendation ITU-T P.862 in 2001, is one of the most widely used measures for audio quality assessment. However, it was specifically developed for measuring distortions introduced by speech compression algorithms (codecs), thus compromising its performance in other domains (e.g. TTS).

F_0 root mean square error (F_0 RMSE)

The F_0 root mean square error (F_0 RMSE) is the RMSE between the original pitch contours and the reconstructed F_0 values.

2.8 Speech-To-Speech Translation

Speech-To-Speech Translation (S2ST or just S2S) has been one of the most challenging tasks in the field of natural language processing for many years [Lav+97; Wah13]. At this time, however, we are closer than ever to seeing accurate S2S systems thanks to the steady progress in deep learning for speech and language processing tasks, particularly ASR, MT and TTS. Indeed, although holistic, *end-to-end* systems are also being tried with good results, the more traditional *cascade* systems pipelining ASR, MT and TTS components, are now achieving impressive state-of-the-art results [Jia+19; Spe+19]. The cascaded approach allows for training strong independent ASR and MT systems, for which the abundancy of training data is clearly superior to that available for end-to-end systems.

Fueled by its immense applicability in real-life settings, research in simultaneous S2S is certainly gaining momentum. In this regard, there are at least three main research challenges of great importance to build effective simultaneous cascade systems. First of all, the ASR component has to work under a strict *streaming* regime; that is, subject to the constraint that output must be delivered in nearly real time, only within a short delay or *latency* after the incoming audio stream. Then, as with its predecessor, the MT component has to deliver its output almost immediately, in a *simultaneous* fashion, even if no "complete" input sentence is available for translation. Finally, the TTS component is faced with the same constraint: speech synthesis has to start from just a prefix of the "complete" input sentence. All in all, these constraints make conventional (offline) sentence-level ASR, MT and TTS systems impractical for simultaneous S2S [Sud+20]. Instead, they have to be replaced by well-fitted systems for *streaming ASR*, *simultaneous MT* and *incremental TTS*.

2.8.1 Streaming ASR

State-of-the-art ASR systems are based on the hybrid approach [YD14]. Bidirectional LSTM (BLSTM) networks have shown to deliver highly competitive results for acoustic modeling in a wide range of ASR tasks [GS05; CL15; CH16; ZSN16]. Likewise, transformer models have recently become the preferred choice for language modeling [Iri+19; Baq+20], though unidirectional LSTM recurrent neural networks are also widely used [Joz+16].

However, when we move from the offline (batch) to the streaming (online) setup, it is necessary to take into account a number of constraints imposed by the streaming scenario to efficiently manage DNNs. As expected in an offline

setup, the BLSTM architecture observes the complete acoustic sequence to estimate the likelihood for each frame in this sequence. However, this is not feasible in a streaming scenario under tight real-time constraints. In brief, we move from a sentence-based to a chunk-based training strategy, in which the input signal is processed by a sliding window over the audio stream [Jor+19; Jor+20]. The acoustic score of each audio frame is computed within the limited past and future contexts imposed by the sliding window. On the other hand, decoder hypotheses are also scored using a language model whose computational cost is kept under control by pruning and variance regularization techniques [Shi+14].

2.8.2 Simultaneous MT

In the S2S pipeline, the MT component is responsible of translating the source text, previously predicted by the ASR system, into the target (translated) text. Standard state-of-the-art transformer-based MT architectures [Vas+17; Bar+20] require the entire source sentence to be available before generating its corresponding translation. In the context of a streaming speech translation (ST) pipeline, the input is not a full semantic unit but a possibly infinite stream of words that are made available in a timely manner. Thus, two problems arise. First, there is the need of splitting the continuous stream of words produced by the ASR system into sentence-like units. MT systems are trained using sentence-aligned corpus, commonly no longer than 100-150 words, so the unbounded text stream needs to be split into (hopefully semantically self-contained) sentence-like chunks to be better translated. Secondly, the MT model needs to be adapted so that it can start translating from prefix-based input without the need of the complete sentence to become available.

The first problem can be solved by including a segmentation system that carries out the sentence-like splitting [Ira+20b]. The goal of this segmenter is to split the continuous stream of words generated by the upstream ASR system into non-overlapping chunks that maximize the accuracy of the downstream MT system. Once the segmenter emits an end-of-segment event, the MT encoder and decoder are reset to make a fresh start of the translation process.

To overcome the prefix-based translation, a number of simultaneous MT systems have been proposed with successful results [Ma+18; Ari+19; Elb+20; Zhe+20]. They are characterized by a translation policy that dictates, at each point, whether enough context is available, or if we must wait for additional input words to be available. These policies can be either fixed [Ma+18], if

they used a set of simple deterministic rules, or adaptative [Ari+19] if they also depend on the specific input words which are available.

2.8.3 Incremental TTS

Conventional TTS systems generate the synthetic speech corresponding to a text input comprising a full sentence or semantically self-contained unit. However, similarly to the case of streaming ASR and simultaneous MT, the TTS component of the simultaneous S2S pipeline should work until tight response-time constraints. Thus, the TTS should start the synthesis process from prefix-based input without the need of the complete translated sentence to become available. In this way, *incremental* TTS refers to the task of generating utterances in small linguistic units for the sake of real-time and low-latency applications [Ma+20].

Chapter 3

Cross-lingual Voice Cloning with Tacotron 2

3.1 Introduction

The first goal of this thesis is to adapt state-of-the-art end-to-end neural TTS architectures to multilingual and multi-speaker settings, paying particular attention to cross-lingual voice cloning. Cross-lingual voice cloning refers to producing synthetic speech for unseen speaker-language pairs (e.g. synthesizing English speech in the voice of a Spanish speaker), and remains a challenging task not as well-covered in TTS research. This aspect is very relevant for the application of TTS technologies into UPV[Media], where most of its contents are recorded in a single language, and then multilingual subtitles are automatically (or semi-automatically) produced by means of ASR and MT systems. In a final step, cross-lingual voice cloning can be used to enable automatic speech dubbing of UPV[Media] contents while preserving the original lecturer voice characteristics.

From all pioneering works in end-to-end neural TTS, Google's Tacotron 2 [She+18] is unquestionably the one that has gained most attention from the TTS research community. When paired with autoregressive neural vocoders such as WaveNet or WaveRNN [Oor+16; Kal+18], it has been shown to be able to rival human-recorded speech in terms of naturalness and audio quality. However, Tacotron 2 is not directly suitable for cross-lingual voice cloning, as its architecture is designed for building monolingual single-speaker systems.

This chapter addresses the adaptation of Tacotron 2 to support multiple speakers and languages, and to enable cross-lingual voice cloning for unseen speaker-

language pairs. This work has been conducted in parallel to Google's own extension to Tacotron 2 [Zha+19c] for this same purpose. Thus, although the proposed extension to the original architecture has much in common with [Zha+19c], it is considered an important contribution of this thesis.

3.2 Tacotron 2

Tacotron 2 follows the two-stage approach introduced in Section 2.7. It presents an autoregressive attention-based encoder-decoder text-to-spectrogram network that directly maps character or phoneme embedding sequences to mel-scale spectrograms. In a second stage, a WaveNet vocoder is used to generate the final waveform conditioned on the mel spectrogram features. Figure 3.1 illustrates the overall architecture of Tacotron 2.

The encoder maps the input character or phoneme embedding sequence $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$ into a series of annotations or intermediate representations $\mathbf{h} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N]$ known as *encoder hidden states*. Each hidden state $\mathbf{h}_i$ contains information about its corresponding input token $\mathbf{x}_i$ with respect to its neighboring tokens $\mathbf{x}_j$ for all $j \neq i$. It is comprised of a stack of 3 1-D convolutional layers with 5×1 filters (i.e. each filter spans 5 characters), followed by batch normalization [IS15] and ReLU activations. The output of the last convolutional layer, which can be seen as character n-grams, is passed into a single bidirectional LSTM (BLSTM or BiLSTM) layer [SP97; HS97] containing 256 units per direction to generate the hidden states $\mathbf{h}_{1:N}$. By feeding n-gram-like representations to the BLSTM, Tacotron 2 is able to better capture longer-term context from the input sequence.

The encoder hidden states are consumed by an attention network which, at each decoding step t, computes a fixed-length *context vector* $\mathbf{c}_t$ conditioned on the full encoded sequence $\mathbf{h}_{1:N}$ and the current attention network state. Similarly to attention-based MT models, the attention network learns to focus on the most relevant subset of $\mathbf{h}_{1:N}$ to predict $\mathbf{y}_t$. To that end, Tacotron 2 extends the additive attention mechanism [Bah+15] by using cumulative attention weights from previous decoder timesteps to encourage the model to move forward consistently through the input sequence. This attention mechanism is known as *location-sensitive attention* [Cho+15], and it will be described in more detail hereinafter.

An autoregressive LSTM decoder, helped by the attention network, predicts the output mel spectrogram $\hat{\mathbf{y}}^{pre} = [\hat{\mathbf{y}}_1^{pre}, \hat{\mathbf{y}}_2^{pre}, ..., \hat{\mathbf{y}}_T^{pre}]$ one frame at a time (the term *pre* is used here to indicate $\hat{\mathbf{y}}$ predictions before the Post-Net). Dur-

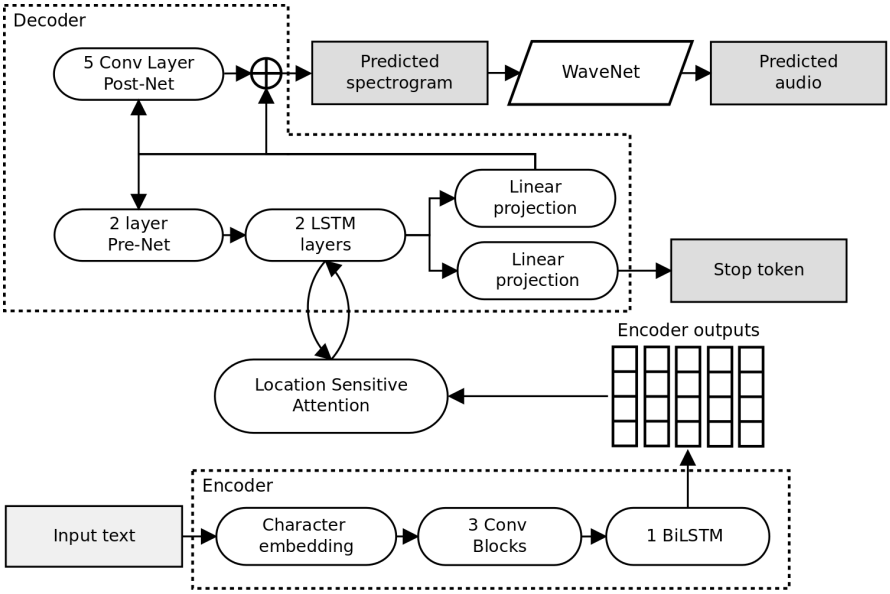


Figure 3.1: Tacotron 2 architecture.

ing training, the previous step ground truth frame $\mathbf{y}_{t-1}$ is used to compute $\hat{\mathbf{y}}_t^{pre}$ (this is known as teacher forcing training). However, $\mathbf{y}_{t-1}$ is first passed through a small Pre-Net comprising two feed forward layers with high dropout rates. This Pre-Net acts as an information bottleneck in the teacher forcing setting. As consecutive spectrogram frames are highly correlated, this bottleneck is essential for learning attention, as it encourages the decoder to make use of the information coming from the input sequence (through $\mathbf{c}_t$). Otherwise, the powerful LSTM decoder would learn to compute $\hat{\mathbf{y}}_t^{pre}$ based solely on $\mathbf{y}_{t-1}$ and disregard $\mathbf{c}_t$. The Pre-Net output and the attention context $\mathbf{c}_t$ are concatenated and passed through a stack of 2 LSTM layers. Then, the LSTM output and $\mathbf{c}_t$ are concatenated and linearly projected to the mel spectrogram dimension. At the same time, the LSTM output and $\mathbf{c}_t$ are also concatenated and linearly projected down to a scalar followed by a sigmoid activation to predict the probability $\hat{\pi}_t$ that the output sequence has completed (*stop token*). Finally, the predicted spectrogram frames $\hat{\mathbf{y}}_{1:T}^{pre}$ are refined by means of a 5-layer convolutional Post-Net that predicts a residual to add to the predicted frames to enhance the spectrogram reconstruction quality. The model is trained to minimize the summed $\ell 2$ loss from before and after the Post-Net between predicted $\hat{\mathbf{y}}$ and ground truth $\mathbf{y}$ spectrograms.

Formally, let $\mathbf{x}$ and $\mathbf{y}$ be the input phoneme embedding and the output mel spectrogram frame sequences, respectively:

$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N], \mathbf{x}_i \in \mathbb{R}^{d_x}$$
$$\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T], \mathbf{y}_i \in \mathbb{R}^{d_y}$$

where d_x and d_y are the embedding and the spectrogram dimensions, respectively. The encoder, described above, computes the annotations $\mathbf{h}$ by passing $\mathbf{x}$ through a series of convolutional and recurrent layers:

$$h = \text{enc}(\mathbf{x}) = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N], \mathbf{h}_i \in \mathbb{R}^{d_h} \tag{3.1}$$

The output mel spectrogram sequence $\hat{\mathbf{y}}^{pre}$ (before the Post-Net) is predicted sequentially in an autoregressive manner, as follows:

$$\hat{\mathbf{y}}_t^{pre} = \text{dec}(\mathbf{y}_{t-1}, \mathbf{s}_{t-1}, \mathbf{c}_t) \qquad (training) \tag{3.2}$$

$$\hat{\mathbf{y}}_t^{pre} = \text{dec}(\hat{\mathbf{y}}_{t-1}^{pre}, \mathbf{s}_{t-1}, \mathbf{c}_t) \qquad (inference) \tag{3.3}$$

where $\mathbf{s}_{t-1}$ is the RNN decoder hidden state after step $t-1$, and $\mathbf{c}_t$ is the attention context vector at step t. The latter is computed as the weighted sum of the encoder hidden states as follows:

$$\mathbf{c}_t = \sum_{j=1}^{N} \alpha_{t,j} \mathbf{h}_j \tag{3.4}$$

where α are the *alignments* or *attention weights*. Figure 3.2 illustrates computed α values for a given training sample.

The softmax function is used to compute the attention weights:

$$\alpha_{t,j} = \frac{\exp(e_{t,j})}{\sum_{k=1}^{N} \exp(e_{t,k})} \tag{3.5}$$

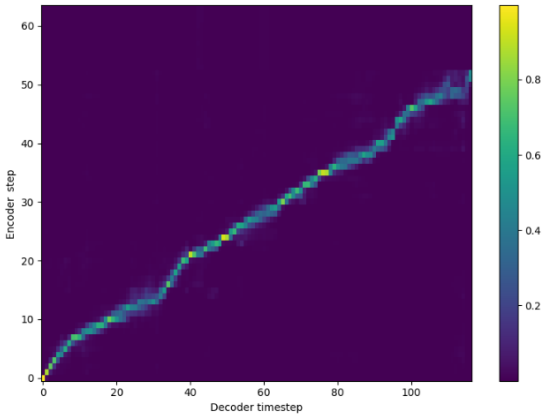


Figure 3.2: Location-sensitive attention weights (α) for a random training sample.

where $e_{t,j}$ is a score usually referred to as *energy*, and it can be interpreted as how relevant is $\mathbf{h}_j$ to predict $\hat{\mathbf{y}}_t$. The location-sensitive attention mechanism used in Tacotron 2 computes $e_{t,j}$ as:

$$e_{t,j} = a(\mathbf{s}_{t-1}, \tilde{\alpha}_{t-1}, \mathbf{h}_j) \tag{3.6}$$

which is most similar to the hybrid attention mechanism proposed in [Cho+15], but cumulative alignments ($\tilde{\alpha}_{t-1}$) are used instead of previous alignments (α_{t-1}) to compute the attention scores:

$$\tilde{\alpha}_{t-1} = \sum_{i=1}^{t-1} \alpha_i \tag{3.7}$$

$$a(\mathbf{s}_{t-1}, \tilde{\alpha}_{t-1}, \mathbf{h}_j) = \mathbf{w}_a^\mathsf{T} \tanh(\mathbf{W}_a \mathbf{s}_{t-1} + \mathbf{V}_a \mathbf{h}_j + \mathbf{U}_a \mathbf{f}_{t,j}) \tag{3.8}$$

where $\mathbf{w}_a$, $\mathbf{W}_a$, $\mathbf{V}_a$ and $\mathbf{U}_a$ are learnable parameters and $\mathbf{f}_{t,j} \in \mathbb{R}^{d_f}$ are location features computed by convolving cumulative alignments $\tilde{\alpha}_{t-1}$ with a matrix $\mathbf{F} \in \mathbb{R}^{d_f \times r}$:

$$\mathbf{f}_t = \mathbf{F} * \tilde{\alpha}_{t-1} \tag{3.9}$$

The process of computing location features $\mathbf{f}_t$ by convolving $\tilde{\alpha}_{t-1}$ is illustrated in Figure 3.3. Tacotron 2 uses 32 1-D convolution filters of length 31 to compute $\mathbf{f}_t$.

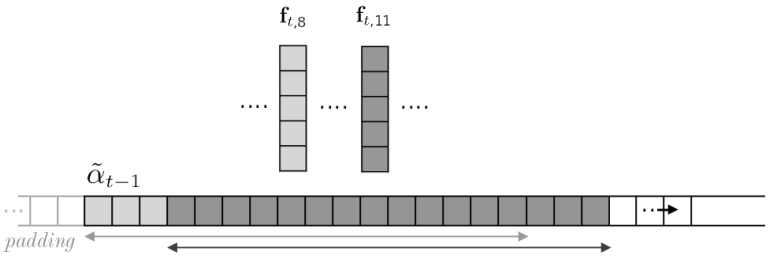


Figure 3.3: Computation of location features $\mathbf{f}_t$ by 5 15x1 convolution filters.

The model is trained to minimize the $\ell 2$ loss between predicted and ground truth spectrogram frames both before and after the Post-Net ($\mathcal{L}_{mel}$). Additionally, a binary cross-entropy loss is used for stop token prediction ($\mathcal{L}_{stop}$). Finally, a L2-regularization loss is included to prevent overfitting:

$$\mathcal{L}_{mel} = \frac{1}{2N}\sum_{n=1}^{N}(\mathbf{y}_n - \hat{\mathbf{y}}_n^{pre})^2 + \frac{1}{2N}\sum_{n=1}^{N}(\mathbf{y}_n - \hat{\mathbf{y}}_n^{post})^2 \tag{3.10}$$

$$\mathcal{L}_{stop} = -\frac{1}{N}\sum_{n=1}^{N}\pi_n \log \hat{\pi}_n \tag{3.11}$$

$$\mathcal{L} = \mathcal{L}_{mel} + \mathcal{L}_{stop} + \lambda \sum_{j=1}^{W} w_j^2 \tag{3.12}$$

where π_n are binary stop values indicating if the output sequence is completed (i.e. 1 for padded values) or not, w_j are the trainable weights of the network and λ is the L2 regularization weight.

3.3 Extending Tacotron 2 with cross-lingual voice cloning capabilities

In multilingual and multi-speaker scenarios such as UPV[Media], in which the considered set of lecturers or speakers $\mathcal{S}$ speak one or more languages $\ell \subseteq \mathcal{L}$, one can collect a training dataset $\mathcal{D} = \{(\ell_i, s_i, \mathbf{x}_i, \mathbf{w}_i)\}_{i=1}^{N};\ \ell_i \in \mathcal{L}, s_i \in \mathcal{S}$ where $(\mathbf{x}, \mathbf{w})$ are *(text,audio)* speech recordings pairs. However, Tacotron 2 is not able to handle multiple speakers and languages. In such context, these two aspects are essential not only to make the most out of the training data available, but also to enable cross-lingual voice cloning capabilities.

To that end, a multilingual multi-speaker extension to Tacotron 2 is proposed. To help the model transfer voices across languages, it is important to provide the network with speaker and language information at points in which most of the modeling power (i.e. network weights) is shared among them. Thus, the main challenge when introducing language and speaker information involves disentangling language attributes from speaker identities.

First, inspired in Deep Voice 2 [Ari+17], low dimensional trainable speaker embeddings are introduced to condition the spectrogram generation on the speaker identity. To our knowledge, Deep Voice 2 is the first work introducing trainable speaker embeddings to generate different voices from a single neural TTS model. However, it presents a more traditional pipeline comprising separate duration, frequency and vocal models in which the embeddings are attached independently to each of these components. For Tacotron 2, we opted to broadcast-concatenate speaker embeddings to the encoder hidden states so that the encoder layers can learn speaker-agnostic representations of the phoneme sequences, which we find helps generalization.

To account for the multiple languages, language-specific phoneme embeddings are used. This means for example that a phoneme æ uses a different embedding depending on the input language: $æ_{en}$, $æ_{es}$, etc. We find this choice effective provided that the different languages are well covered by a number of speakers.

As mentioned in the introduction to this chapter, these modifications are in line with the cross-lingual voice cloning architecture proposed in [Zha+19c], which is published shortly afterwards the prosecution of this work. The main differences with respect to the proposed architecture are the following. First, equivalent grapheme/phoneme embeddings are shared across languages, where small language embeddings are also incorporated to the input. Also, an adversarially trained speaker classifier is introduced to further help the model

disentangle speakers from languages when only one training speaker is available for some of the languages.

3.4 Overcoming the exposure bias and attention failures

Despite the fact that Tacotron 2 undoubtedly produces synthetic speech with unprecedented levels of naturalness, some of its architectural designs lead to different instability conditions that, in some cases, prevent it from being used in production environments where there is small tolerance to errors. This is particularly notable when the considered dataset presents some challenging conditions (e.g. variability of acoustic conditions, multiple speakers, scarce training data, etc).

Most of these errors at inference time are produced when the attention mechanism is not able to correctly follow a monotonic-like path along the input sequence, causing undesired outputs containing skipped phonemes, repetitions or mumbling speech [ZLD18; He+19]. In some extreme cases, the attention collapses and the generated output contains nothing but unintelligible gibberish. Figure 3.4 illustrates some of the aforementioned cases.

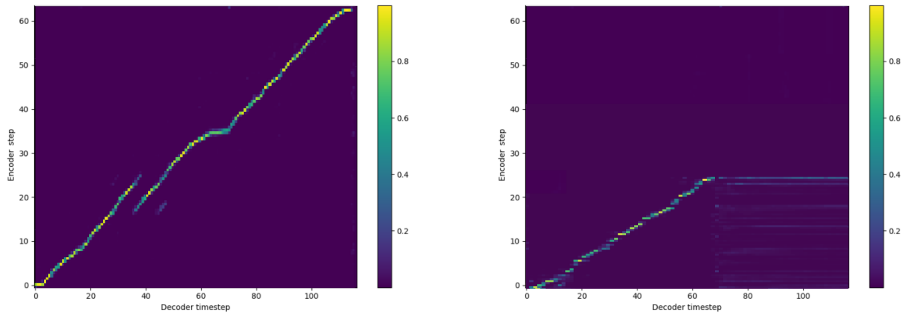


Figure 3.4: Attention failures causing phoneme repetitions (left) or unintelligible output (right).

Such robustness issues are strongly related to the following Tacotron 2 architectural designs. On the one hand, the teacher forcing training procedure introduces a discrepancy between training and inference conditions. During training, at timestep t the decoder is fed with previous step ground truth spectrogram frame $\mathbf{y}_{t-1}$, while the predicted frame $\hat{\mathbf{y}}_{t-1}^{pre}$ is used at inference time. This results in an unpredictable performance for out-of-domain test data [Liu+20]. Scheduled sampling [Ben+15] brings an alternative training

scheme to overcome this problem. However, the use of scheduled sampling entails some negative effects caused by the misalignment between recorded and predicted speech [Liu+20], introducing artifacts and lowering generated speech quality in the resulting models. After the execution of this work, some methods have been proposed to address this issue, which include using a teacher-student training framework [Liu+20] or a forward-backward decoding regularization method [Zhe+19].

On the other hand, although the location-sensitive attention (LSA) mechanism proposed in Tacotron 2 encourages the attention to move forward along the input sequence, it does not explicitly introduce any constraints to exploit the monotonic nature of the TTS task. Humans read out text sequentially, focusing only on a local subset of characters at a given time. This is referred to as the *locality* property of the TTS task. LSA alignment weights at step t (α_t) could potentially focus equally (i.e. assign the same weight) all encoder hidden states, which may lead to an attention collapse. To address these issues and prevent the attention mechanism to repeat or skip phonemes or collapse, some alternative attention mechanisms have been proposed.

Forward attention [ZLD18] encourages monotonicity and completeness implicitly by reweighing the alignments by previous ones using forward variables. Particularly, alignment weights at step t are computed as:

$$\hat{\alpha}_{t,j} = \frac{\exp(e_{t,j})}{\sum_{k=1}^{N} \exp(e_{t,k})} \tag{3.13}$$

$$\bar{\alpha}_{t,j} = (\alpha_{t-1,j} + \alpha_{t-1,j-1}) \cdot \hat{\alpha}_{t,j} \tag{3.14}$$

$$\alpha_{t,j} = \frac{\bar{\alpha}_{t,j}}{\sum_{k=1}^{N} \bar{\alpha}_{t,k}} \tag{3.15}$$

where $\hat{\alpha}_t$ are the regular softmax normalized energy scores before forward reweighting, and $\bar{\alpha}_t$ are the reweighted forward attention weights. The final attention weights for step t (α_t) are given by normalizing $\bar{\alpha}_t$ values to sum 1. This mechanism has been proven to improve convergence speed and inference robustness, drastically reducing the number of attention failures compared to LSA.

However, forward attention does still suffer from occasional attention errors such as skips or repetitions. A similar yet better performing mechanism is

the stepwise monotonic attention (SMA) proposed in [He+19], which builds upon monotonic attention and restricts the hard-aligned position at each step to move at most one step, ensuring all hidden states will be covered during the decoding process. At each timestep t, the probability of moving one step forward or stay unmoved is sampled from a Bernoulli distribution with probability $p_{t,j} = \sigma(e_{t,j})$ (where σ is the sigmoid activation function). The alignment weights are therefore calculated as:

$$\alpha_{t,j} = \alpha_{t-1,j-1}(1 - p_{t,j-1}) + \alpha_{t-1,j}p_{t,j} \tag{3.16}$$

The stepwise monotonic attention mechanism provides a highly stable inference process, greatly improving Tacotron 2 and other attention-based models robustness at inference time. Consequently, the original location-sensitive attention is replaced by the stepwise monotonic attention as the proposed attention mechanism for the extended Tacotron 2 model.

3.5 Improving stop token prediction

Another frequent instability issue at inference time experienced with Tacotron 2 is that in some cases the model is not able to correctly predict the stop token. This results, in exceptional cases, in an infinite loop that produces unintelligible mumbling after predicting the given utterance.

To improve the robustness of the stop token prediction we replace the binary cross-entropy loss with an $\ell 2$ loss, where instead of predicting binary 1 and 0 values for padded and non-padded sequences, the stop token is encoded as a real number starting at 0.0 and reaching 1.0 at the end of the sequence [Lat+19]. Then, two fixed threshold values are adjusted empirically to control generation stop. We use a soft-threshold (e.g. 0.9) so that, after a safe number of additional steps (e.g. 20), the inference autoregressive loop is stopped. We also define a hard-threshold (e.g. 0.98) so that, when reached, the generation is forced to stop immediately.

3.6 Proposed model and general training procedure

The modifications to the original Tacotron 2 architecture described in Sections 3.3 and 3.4 were introduced into a well-known open-source Tacotron 2 TensorFlow implementation [Mam18]. In short, we introduce trainable speaker embeddings and language-specific phoneme embeddings to account for the different speakers and languages. The stepwise monotonic attention is used in favor of location-sensitive attention, which brings improved convergence speed and inference robustness. We also use the improved stop token prediction detailed in Section 3.5. Finally, the more efficient WaveRNN [Kal+18] neural vocoder is used to generate the final waveform from the predicted spectrograms. It brings an inference speedup of near 10× over WaveNet while producing high fidelity audio of comparable quality [Kal+18]. The resulting multilingual and multi-speaker Tacotron 2 architecture is depicted in Figure 3.5.

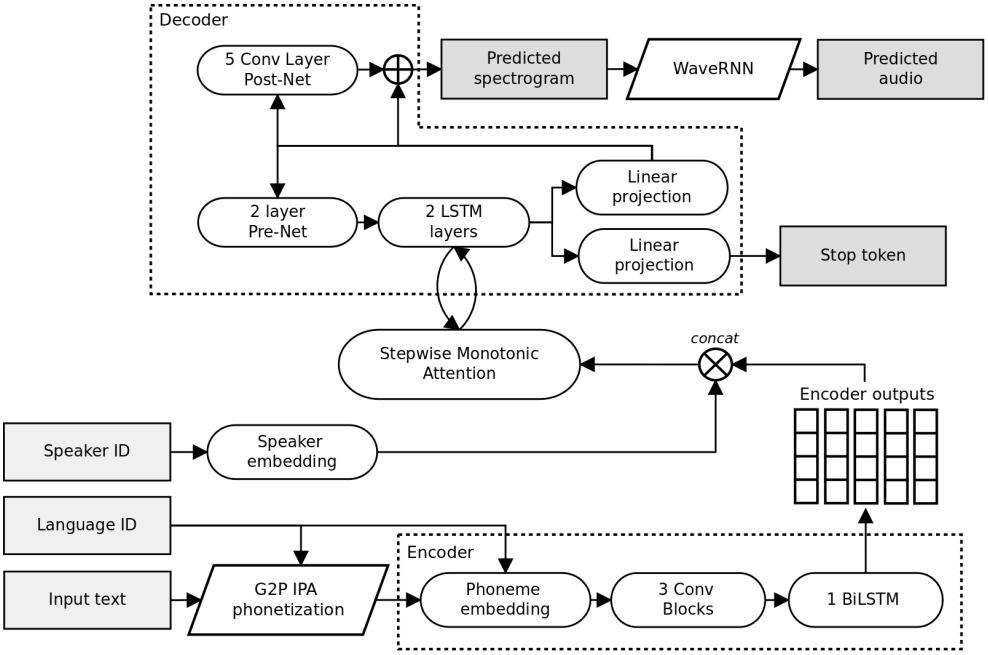


Figure 3.5: Proposed cross-lingual voice cloning extension to Tacotron 2.

The general training procedure is as follows. First, the text-to-spectrogram network is trained on the collection of recorded samples. Then, the model is run in teacher forcing mode to predict the *ground-truth aligned* (GTA) spec-

trograms $\hat{\mathbf{y}}$ (i.e. the ground truth frame $\mathbf{y}_{t-1}$ is used to predict $\hat{\mathbf{y}}_t$) from the training data. This is to ensure that each predicted frame exactly aligns with its corresponding waveform samples. Finally, the WaveRNN vocoder is trained on the GTA spectrograms to predict the final waveform samples. As predicted spectrograms $\hat{\mathbf{y}}$ tend to be oversmoothed and less detailed than $\mathbf{y}$, training the vocoder on GTA predictions reduces the gap between training and inference conditions of the vocoder model resulting in improved audio quality.

3.7 Conclusions

In this chapter we have addressed the adaptation of end-to-end neural TTS architectures to multilingual and multi-speaker settings, focusing on the cross-lingual voice cloning task, which corresponds to the first goal raised in this thesis. This included an extensive review of the state-of-the-art in neural TTS and a comprehensive experimentation with the Tacotron 2 architecture.

By including trainable speaker embeddings and language-specific phoneme embeddings, Tacotron 2 can be trained on multilingual and multi-speaker datasets, also allowing for voice transfering across languages. During the experimentation with Tacotron 2, a number of instability conditions at inference time mainly related to the teacher forcing training paradigm were identified. In particular, we have addressed the attention and stop token prediction failures as a result of the poor generalization of the model to inference conditions (exposure bias).

The evaluation of the proposed extension to Tacotron 2 is addressed next in Chapter 4, where 47 UPV lecturers assess different aspects of the resulting cross-lingual voice cloning model trained on a multilingual and multi-speaker dataset.

Chapter 4

Cross-lingual Voice Cloning for UPV[Media]

4.1 Introduction

As mentioned earlier, state-of-the-art ASR and MT technologies are used since 2014 to produce cost-effective multilingual subtitles of publishable quality at the UPV[Media] platform. With the objective of enriching UPV[Media] video lectures and learning pills with automatically voice-cloned (dubbed or voiceover) versions in Spanish, Catalan and English (the main teaching languages in UPV), a TTS component with cross-lingual voice cloning capabilities can be added to the existing ASR + MT pipeline. In this process, target subtitles (translations) can be post-edited if convenient. In this work, the focus is put on the TTS step, for which it is assumed (reviewed) translations to be available in each target language of interest.

The objective of this chapter is to assess the performance of the cross-lingual voice cloning model proposed in Chapter 3, and also whether this technology can contribute improving accessibility and engagement in online learning, particularly in the context of higher education. To that end, a call for participation is made to the UPV's academic staff under the *Docència en Xarxa* (DeX) plan to collect *clean* lecturer speech data during the academic courses 2016–17 and 2017–18. The collected data is used to train the extended Tacotron 2 model proposed in Section 3.6, which is subjectively evaluated by 47 UPV lecturers.

The rest of this chapter is organized as follows. First, the UPV[Media] platform (UPV's main repository of educational videos) is presented in Section 4.2. Then, the above mentioned DeX call for participation and the collected TTS

dataset are described in detail in Section 4.3. The training process of the proposed cross-lingual voice cloning system on the DeX-TTS dataset is detailed in Section 4.4. Section 4.5 covers the subjective assessment of this technology by 47 UPV lecturers. Finally, some concluding remarks are given in Section 4.6.

4.2 The UPV[Media] platform

In a broad sense, UPV[Media] is a professional UPV service for the creation, storage, management and open dissemination of educational videos [Tur+09; Med20]. 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 [UPV20b], especially as an edX member since 2014 [UPV20a]. 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 enrollments, and two of the 100 most popular online courses of all time [Cla20] as of June 2020. Apart from poliMedias, UPV [Media] 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 hours per year are being recorded and added to UPV[Media] for their distribution to students only through a Sakai LMS [Tur+14; Ope20].

Although UPV[Media] 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 (4x4 meter) 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 (Figure 4.1). 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 UPV[Media] (see example in Figure 4.2).

Figure 4.1: [UPV] Media studio for the recording of poliMedias.

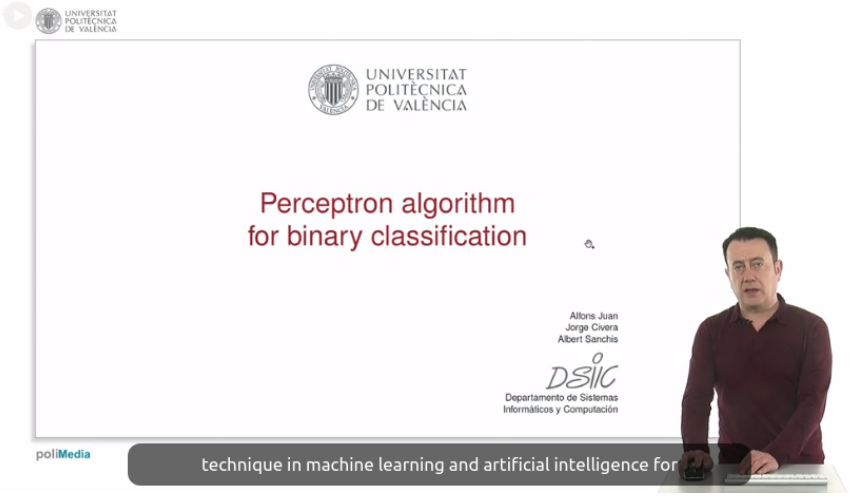


Figure 4.2: A poliMedia with automatic subtitles.

Supported by the UPV's *Docència en Xarxa* stimulus plan for online teaching, the number of poliMedias uploaded to UPV[Media] 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 4.1, 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 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 [BOU20, 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 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 is good to have accurate and cost-effective means to fully convert basic (monolingual) poliMedias into trilingual learning objects.

Table 4.1: Number of poliMedia videos and hours in Spanish, Catalan and English.

	Videos		Hours	
Language	No.	%	No.	%
Spanish	38172	87	9451	89
Catalan	1333	3	232	2
English	4591	10	918	7
Total	44096	100	10601	100

Table 4.2 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 4.2, 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 [BOU20,

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.

Table 4.2: 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	
	es	ca	en	es-ca	es-en	ca-en	es-ca-en	Total
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

The UPV[Media] 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 systems were already integrated into the UPV[Media] 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 [Val+15b; Sil+13a; Pér+15b; Val+15a]. 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, we have now reached the point at which raw subtitles are often good enough for direct publication.

4.3 The Docència en Xarxa multilingual TTS dataset

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 4.3, 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 4.2), though good for our purposes.

Table 4.3: 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	
	es	ca	en	es-ca	es-en	ca-en	es-ca-en	Total
Participants	36	1	4	16	22	3	16	98
Total	41			41			16	98

Table 4.4 shows the number of sentences and duration in hours collected in our *DeX–TTS* dataset of clean lecturer speech data. In total, it comprises 59 hours of clean speech data from 47K 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.

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, this corpus is similar in size to those commonly used in TTS research (cf. [She+18] and [Ren+19]). 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

Table 4.4: 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	
		es	ca	en	es-ca	es-en	ca-en	es-ca-en	Total (%)
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 (34)			20.1 (42)			11.3 (24)	47.3 (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 (36)			25.0 (42)			12.9 (22)	58.8 (100)

this regard, the DeX-TTS 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.

4.4 Model training

This section details the training process of the cross-lingual voice cloning TTS system proposed in Chapter 3 on the DeX-TTS dataset. The DeX-TTS dataset contains speech recordings accompanied by its corresponding transcriptions covering English, Spanish and Catalan from 98 UPV lecturers.

In order to optimize DeX-TTS recordings for speech synthesis, some basic digital signal processing (DSP) operations were applied to the audio recordings. First, dynamic range compression was applied in order to uniform loudness of louder and quieter fragments by mapping the natural dynamic range of the audio signals to a smaller range. Then, high-pass filters were used to reduce low frequency noises. Leading and trailing silence was removed from all recordings. Audio samples were downsampled to 22kHz, and 100-bin log magnitude mel-scale spectrograms were extracted with Hann windowing, 50ms window length, 12.5ms hop size and 1024 point Fourier transform. The spectrograms were normalized to lay within the $[-4.0, 4.0]$ range.

Regarding text preprocessing, texts were applied basic cleaning operations (removing special characters) and lowercasing for all three languages. Then, phoneme sequences were extracted from the normalized texts using the well-known open source speech synthesizer tool eSpeak NG [DD], which includes a predefined set of grapheme-to-phoneme (G2P) conversion rules for many languages (including English, Spanish and Catalan). The use of phoneme instead of grapheme sequences as inputs to the acoustic model has probably a negligible effect in the case of Spanish or Catalan, where there is almost a one-to-one mapping between phonemes and graphemes. However, in the case of English, the spelling of a word does not directly correspond to its pronunciation, and the use of G2P tools help relieve the TTS model from learning complex phonetic rules (particularly regarding vowels) from the training corpora.

When not mentioned otherwise, default Tacotron 2 hyperparameters were used for training the DeX-TTS cross-lingual voice cloning models. These are summarized in Table 4.5. Trainable 64-dim speaker embeddings were broadcast-concatenated to the text encoder outputs to condition spectrogram prediction on the speaker identity. The stepwise monotonic attention was used in favor of the location-sensitive attention mechanism. Similarly to [Wan+17], a reduction factor $r = 2$ was used (that is, predicting 2 frames per decoding step) to reduce GPU memory footprint and aid attention convergence. The Adam optimizer [KB15] with learning rate decay, starting at 0.001 for 100K steps and decaying to 5×10^{-5} over 100K additional steps was used. The models were trained on a single-GPU machine for a total of 250K steps with a batch size of 32. The $\ell 1$ loss was used instead of $\ell 2$ loss as $\ell 1$ presents improved robustness to outliers and tends to create less blurry spectrograms.

Finally, an autoregressive WaveRNN neural vocoder model was trained on the DeX-TTS recordings. A well-known open source implementation of the WaveRNN neural vocoder [McC18] was used for this purpose, which presents a slightly modified architecture with respect to the original work. A raw 10-bit model was trained on the ground truth aligned (GTA) log-magnitude mel-scale spectrograms (i.e. the predicted spectrograms $\hat{y}$ from a trained acoustic model under teacher forcing settings) with a batch size of 64 and a learning rate of 0.0001 over 800K steps.

Table 4.5: Cross-lingual voice cloning Tacotron 2 hyperparameters for DeX-TTS.

Common	Training mode	Teacher forcing
	Batch size	32
	Reduction factor	2
	Initial learning rate	0.001
	Final learning rate	5×10^{-5}
	Spectrogram loss	$\ell 1$
	Stop token loss	$\ell 2$
	Optimizer	Adam(0.9, 0.999, 1×10^{-6})
	LSTM zone-out prob	0.1
Audio processing	Sampling rate (Hz)	22050
	Pre-emphasis	0.98
	Windowing	Hanning
	Window size	1024
	Hop size	256
	Mel channels	100
	Mel frequency upper bound	8000
	Mel frequency lower bound	0
Inputs	Phoneme embedding dim	512
	Speaker embedding dim	64
Encoder	Conv kernel	5×1
	Conv dim	$[512, 512, 512]$
	Conv layers dropout prob	0.5
	Bi-LSTM dim	256×2
Attention	Type	Stepwise monotonic attention
	Attention dim	128
Decoder	Pre-Net dims	$[256, 128]$
	Pre-Net dropout prob	$[0.5, 0.5]$
	LSTM dim	$[1024, 1024]$
	Post-Net conv kernel	5×1
	Post-Net conv dim	$[512, 512, 512, 512, 100]$

4.5 Evaluation

To assess the DeX-TTS cross-lingual voice cloning system described in previous sections, a call for participation was made to the 98 lecturers contributing to the DeX-TTS dataset (Section 4.3), which was answered by nearly half of them (47). The evaluation procedure was designed around a *test set* of 8820 speech samples synthesized by the trained TTS system described in Section 4.4. 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 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 4.3). 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 (Figure 4.3).

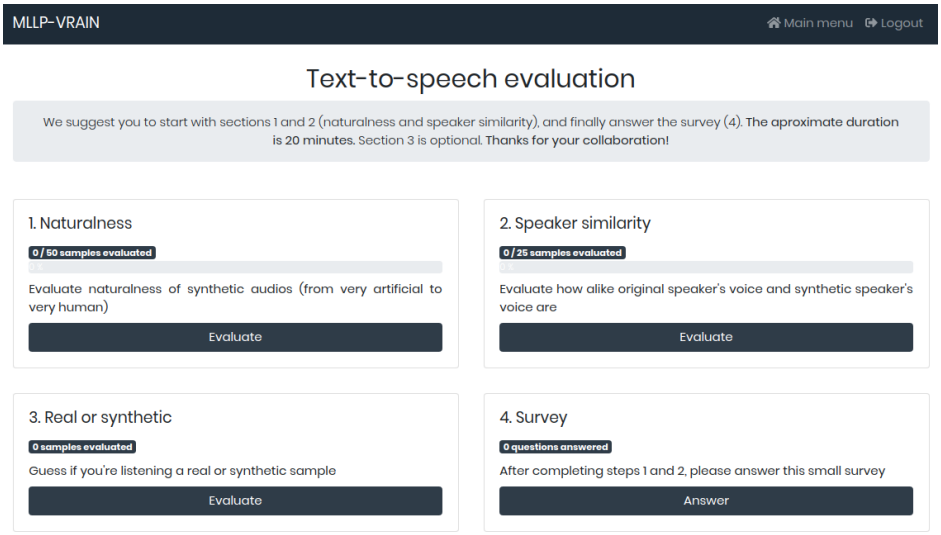


Figure 4.3: Home page of the evaluation platform.

As shown in Figure 4.3, 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 proce-

dure as they wished. In what follows, procedural details and evaluation results are provided for each part separately.

4.5.1 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 (Figure 4.4). 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.

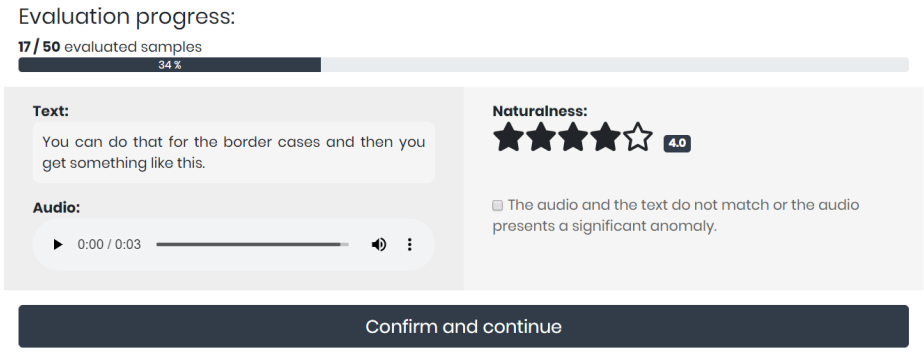


Figure 4.4: Naturalness evaluation interface.

Table 4.6 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 model training.

From the results in Table 4.6, it can be observed that the naturalness MOS on the synthetic speech produced by the TTS model is in general 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

Table 4.6: Naturalness MOS with 95% confidence intervals per language, including cross-lingual cloning (synthetic samples from lecturer-language pairs *unseen* in training).

	Naturalness MOS				
	- *(Synthetic samples)* -			Control	Evaluated
Language	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

any case, summarizing, a main conclusion from Table 4.6 is that the proposed cross-lingual voice cloning system 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, the system has effectively learned to transfer (clone) lecturer voices from source languages (e.g. mother tongue) to target languages they might not even speak.

4.5.2 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 the system 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 4.7 shows the speaker similarity MOS with 95% confidence intervals for the seen and unseen lecturer-language pairs separately, and the number of evaluated samples.

From the results in Table 4.7, we can confirm that cross-lingual voice cloning works almost as well as conventional voice cloning from seen lecturer-language

Table 4.7: Speaker similarity MOS with 95% confidence intervals per language, for test samples produced from seen and unseen lecturer-language pairs of training data.

	Speaker similarity MOS		Evaluated
Language	Seen	Unseen	samples
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

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.

4.5.3 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 4.8 shows the resulting confusion matrix for each language and overall.

Table 4.8: Confusion matrices on the *real or synthetic* test for each language and overall.

	Actual	Guessed condition		Total
Language	condition	Real	Synthetic	samples
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

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 this system is really close to human speech.

4.5.4 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 4.9 shows these two control questions and the Yes or No votes received.

Table 4.9: 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

As shown in Table 4.9, 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 this technology.

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 this 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 this study is that TTS technology is not only mature enough for its application at the UPV, but also needed as soon as possible.

4.6 Conclusions

In this chapter, the synthetic speech naturalness and speaker similarity of the cross-lingual voice cloning system proposed in Chapter 3 has been assessed by 47 UPV lecturers with very positive outcomes, where synthetic samples have been confused in many cases with human-recorded speech (see Table 4.8). All 47 UPV lecturers participating in the subjective evaluation of this system conclude that the shown automatic voice cloning technology can be useful to improve accessibility and engagement in online learning. We would like to thank the effort made by UPV lecturers and staff participating in the *Docència en Xarxa* calls and in the assessment of this technology. The work presented in Chapters 3 and 4 has led to one of the main scientific publications derived from this thesis ([Pér+21]).

Nevertheless, autoregressive models inherently suffer from slow inference speeds, particularly in the audio generation task, as speech waveforms contain tens of thousands of samples per second of audio (usually 16000 or more). This can become an important obstacle when it comes to deploying these models in production-ready settings. Also, it is difficult to have control over different aspects of the speech such as the voice speed or the intonation, which can be useful for certain applications. Both efficiency and controllability of neural TTS models are addressed next in Chapter 5.

The following work was done in collaboration with others:

- The collection of the DeX-TTS dataset introduced in Section 4.3 was organized and led by Santiago Piqueras and Alejandro Pérez.
- The following international journal article, which was derived from this work, was prepared in collaboration with other members of the MLLP and the UPV Media Services (ASIC):
  - Pérez-González-de-Martos, A., Díaz-Munío, G. G., Giménez, A., Silvestre-Cerdà, J. A., Sanchis, A., Civera, J., Jiménez, M., Turró, C. & Juan, A. (2021). *Towards cross-lingual voice cloning in higher education.* Engineering Applications of Artificial Intelligence, 105, 104413.

Chapter 5

Robust, Efficient and Controllable Neural Text-To-Speech

5.1 Introduction

Pioneering works on end-to-end neural TTS were based on encoder-decoder architectures that, helped by an attention mechanism, learn to map a character or phoneme input sequence into an intermediate acoustic representation (e.g. mel-scale spectrograms). Thus, the speech generation is generally split in two stages: a text-to-spectrogram stage which generates the acoustic features from the input text (acoustic model) and a spectrogram-to-wave stage, in which a vocoder (usually a separate neural-based model) reconstructs the final waveform conditioned on the predicted acoustic features.

However, as discussed in Section 3.4, attention-based acoustic models can present different stability issues at inference time [Liu+20; He+19; Zhe+19; Bat+20]. These can be attributed to the poor generalization of the attention mechanisms to inference conditions as a result of the exposure bias (train-test discrepancy) caused by the teacher forcing training strategy [Liu+20]. To overcome such issues, alternative attention mechanisms that exploit the monotonic nature of the TTS task have been proposed with different success [ZLD18; He+19; Bat+20].

Furthermore, although autoregressive TTS acoustic models such as Tacotron 2 can produce highly realistic and natural speech, the decoding process is hardly parallelizable and results in slow inference speeds. This also applies to au-

toregressive neural vocoders such as WaveNet or WaveRNN, with inference speeds around 100× and 10× RTF on GPU respectively [Gov+19]. In addition, although some works leverage unsupervised learning to control ill-defined aspects such as the speaking style [Wan+18c; Hsu+19], it is not straightforward to directly control general attributes of the speech such as the speaking rate or prosody (pitch contour, pauses) in the autoregressive generation.

In this chapter, the focus is put on improving robustness, efficiency and controllability of modern state-of-the-art neural TTS technologies to its large-scale use in production-ready environments.

5.2 Non-autoregressive TTS with explicit duration modeling

Replacing the attention mechanism

The attention mechanism is a core component of autoregressive end-to-end neural TTS models, as it allows to learn a mapping (soft alignments) between input and output sequences [She+18; Pin+18; Li+19]. However, it leads to an unpredictable performance at inference time [Liu+20; He+19; Zhe+19; Bat+20]. Recent works replace the attention mechanism used in end-to-end neural TTS systems by explicitly predicting phoneme durations, more similar to the alignment model in traditional parametric systems [Ren+19; Yu+20; Eli+20; Łań20]. In this approach, ground truth alignments (phoneme durations) need to be provided by an external model (e.g. an autoregressive attention-based TTS model or a forced-aligner tool). Then, encoder hidden states are expanded accordingly to match the length of the target mel-spectrogram, thus avoiding the issues usually affecting attention-based models (repetitions, skips or collapse).

Formally, let $\mathbf{h} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N]$ be the encoded input sequence, where N is the length of the input sequence. Let $\mathbf{d} = [d_1, d_2, ..., d_N], d_i \in \mathbb{Z}$ be the target phoneme durations extracted from an external model or aligner tool, where $\sum_{i=1}^{N} d_i = T$ and T is the length of the output mel-spectrogram sequence. The encoded sequence $\mathbf{h}$ is expanded according to $\mathbf{d}$ to match the mel-spectrogram length, where $\mathbf{h}_i$ is repeated d_i times. For example, let $\mathbf{h} = [\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3, \mathbf{h}_4]$ and $\mathbf{d} = [2, 2, 3, 1]$, then the expanded sequence $\mathbf{h}_{exp}$ becomes $\mathbf{h}_{exp} = [\mathbf{h}_1, \mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_2, \mathbf{h}_3, \mathbf{h}_3, \mathbf{h}_3, \mathbf{h}_4]$. During training, target durations are used to expand the encoded sequence $\mathbf{h}$, and a phoneme duration predictor module is trained along with the feature prediction network to minimize $\ell 1$ or

$\ell 2$ loss between the predicted and target durations. Figure 5.1 illustrates the hidden state expansion procedure.

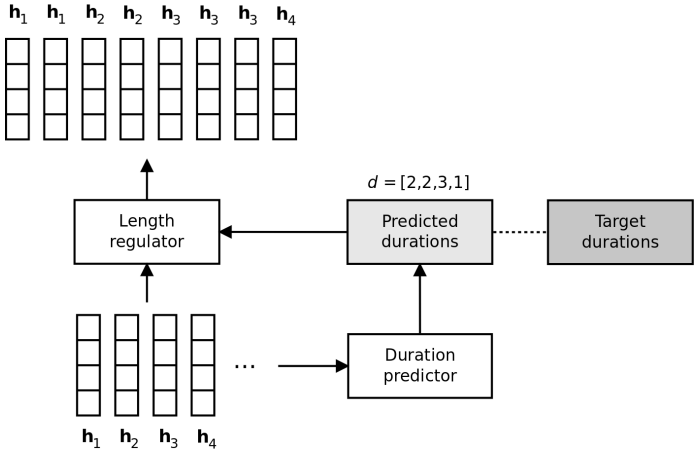


Figure 5.1: Explicit phoneme duration modeling with duration predictor and state expansion (dashed lines correspond to training-only connections).

Explicit duration modeling not only avoids the stability issues arising in attention-based models, but also allows for fine-grained control over speaking rate (voice speed). A simple hyperparameter α can be used to control the voice speed at inference time, proportionally adjusting predicted phoneme durations before the state expansion operation (e.g. $\mathbf{d} = \alpha\mathbf{d}$).

Microsoft's FastSpeech [Ren+19] paved the way for explicit duration modeling in modern neural TTS architectures. In their work, a stack of Feed-Forward Transformer (FFT) blocks (consisting of self-attention and a 1D convolution) are used both for the encoder and decoder modules, allowing for parallel mel-spectrogram generation and thus speeding up the inference process. The clear advantages of explicit duration modeling spread quickly to alternative neural architectures, such as DurIAN [Yu+20], Parallel Tacotron [Eli+20] or FastPitch [Łań20] among others.

Pitch and energy prediction

Text-to-speech is a large-scale inverse process: a highly compressed input source (text) is decompressed into audio (or acoustic features). This is similar to other machine learning generative tasks such as image generation (e.g. given the word *cat*, generate an image of a cat). In such tasks, the one-to-many mapping problem arises (e.g. there are potentially infinite possible ways of saying *hello*). Under these conditions, training machine learning models using regular training objectives (such as minimizing $\ell 1$ or $\ell 2$ loss between predicted and target samples) usually lead to unacceptable or suboptimal solutions (e.g. blurry images in the case of images, flat prosody in the case of speech). This is particularly noticeable in non-autoregressive TTS models, as the teacher forcing training procedure helps alleviating this problem in autoregressive TTS models.

To reduce the information gap (i.e. the input just contains partial information to predict the target) and alleviate the one-to-many mapping problem in non-autoregressive TTS, additional variability information of the speech utterance can be provided. In particular, frame-wise pitch (F_0) and energy values can be extracted from the original utterances and be provided as additional conditioning inputs [Ari+17; Ren+21; Łań20]. Both pitch and energy can be utilized in a similar manner as phoneme durations are predicted and used during training and inference processes. However, differently from the state expansion carried out to account for phoneme durations, pitch and energy information can be effectively included in the encoded sequence $\mathbf{h}_{1:N}$ by means of pitch and energy embeddings.

Again, the explicit modeling of pitch and energy does not only alleviate the one-to-many mapping problem in non-autoregressive TTS, but also allow for fine-grained control of these attributes. For example, to control expressiveness, one could manually increase or decrease predicted pitch and energy variance at inference time.

There are alternative ways to model pitch prediction. In FastSpeech 2 [Ren+21], the continuous wavelet transform (CWT) is used to decompose the continuous pitch series into a pitch spectrogram [Sun+13], and this is taken as the training target for the pitch predictor. Pitch and energy values for each frame are discretized to 256 possible values and encoded into a sequence of one-hot vectors. On the contrary, in FastPitch [Łań20] frame-wise F_0 values are standardized to mean of 0 and standard deviation of 1 and averaged over every input symbol using the extracted durations $\mathbf{d}$. Here, the normalized symbol-averaged pitch values are directly taken as targets.

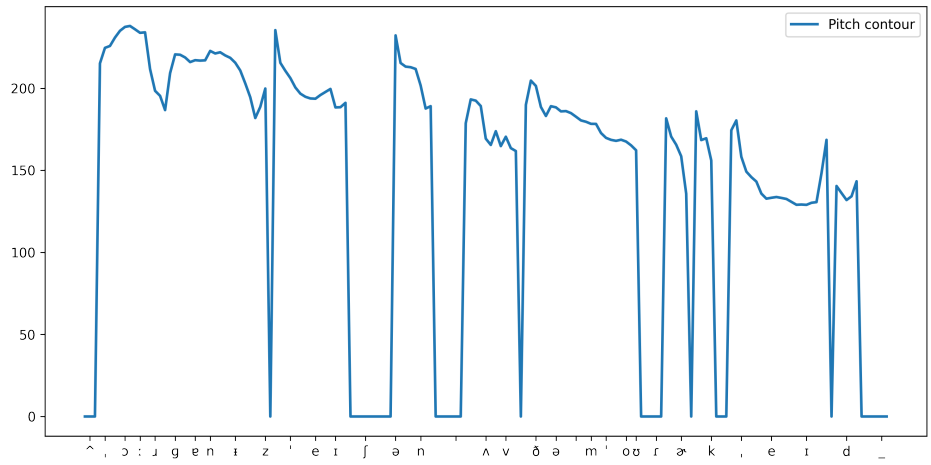


Figure 5.2: Pitch contour (F_0) of a random LJSpeech utterance.

5.3 GAN-based neural vocoders

Autoregressive generative vocoder models such as WaveNet or WaveRNN have shown much superior performance when compared to traditional parametric vocoders [Oor+16; Kal+18]. However, speech waveforms contain tens of thousands of samples per second of audio (usually 16000 or more), thus making non-parallelizable autoregressive models suffer from slow inference speeds.

In the recent years, a lot of effort has been put on producing lighter and more efficient neural vocoder models of comparable quality. Some of the first relevant works along these lines are the LPCNet [VS19], Parallel WaveNet [Oor+18] or WaveGlow [PVC19] vocoders.

More recently, a whole new family of GAN-based vocoder models have been proposed with great success. GANs are generative models that are composed of two separate neural networks: a generator (G) and a discriminator (D) [Goo+14]. The generator G learns a distribution of realistic waveforms by trying to deceive the discriminator (or discriminators) D to recognize the generator samples as real. Parallel WaveGAN [YSK20], MelGAN [Kum+19], Multi-band MelGAN [Yan+20a] or HiFi-GAN [KKB20] are some of the most popular and best performing GAN-based vocoder models.

In particular, the recently published HiFi-GAN vocoder has gained popularity as it has been shown to produce both efficient and high-fidelity speech syn-

thesis, outperforming other vocoders such as WaveGlow or MelGAN in terms of perceived audio quality (MOS) [KKB20]. In its larger version (V1), HiFi-GAN synthesizes human-quality speech audio at speed of 3.7MHz on a single NVIDIA V100 GPU (e.g. 0.006 RTF at 22kHz sampling rate) [KKB20].

HiFi-GAN generator is a fully convolutional network that predicts the final audio waveform conditioned on mel-spectrogram inputs. The mel-spectrogram frames are upsampled by means of a stack of 1-D transposed convolutions to match the temporal resolution of raw waveforms. Each transposed convolution is followed by a multi-receptive field fusion (MRF) module, which observes patterns of various lengths in parallel and returns the sum of outputs from multiple residual blocks with different kernel sizes and dilation rates.

For the adversarial training framework, HiFi-GAN uses two different discriminators: multi-scale (MSD) and multi-period (MPD) discriminators. This is similar to the work in [Bińkowski+20]. The MSD is taken from MelGAN [Kum+19], and comprises a mixture of sub-discriminators operating on different raw audio input scales. The MPD is again a mixture of sub-discriminators each handling a portion of periodic signals of input audio. In addition to the adversarial loss, HiFi-GAN uses a mel-spectrogram reconstruction loss as the $\ell 1$ distance between mel-spectrograms of synthetic and ground truth waveforms to improve the stability and efficiency of the adversarial training process.

5.4 The Blizzard Challenge 2021

The following section is devoted to summarize our participation in the Blizzard Challenge 2021, an international TTS challenge organized annually since 2005 with the purpose of better understanding and comparing different TTS technologies applied to the same provided training dataset. The proposed system [PSJ21] follows the two-stage predominant paradigm in neural TTS, and it is composed of a novel non-autoregressive acoustic model with explicit duration modeling and a HiFi-GAN neural vocoder.

5.4.1 Introduction

The Blizzard Challenge, organized annually since 2005, has the purpose of better understanding and comparing different TTS technologies applied to the same provided training dataset. Since its inception, renowned IT companies and institutions involved in TTS research have been participating in the different editions. In the 2021 edition, the task SH1 consisted of building a Spanish

system from about 5 hours of studio-quality recordings from a native female speaker. The organization allowed to further include up to a total of 100 hours of speech recordings from other sources for training the TTS models.

The proposed system is composed of a non-autoregressive neural acoustic model with explicit duration modeling and a GAN-based neural vocoder. The acoustic model was initially developed from ForwardTacotron[1], to which we introduced several modifications based on different recently published works, which will be detailed in Section 5.4.4. The acoustic model takes the phoneme sequence as inputs and generates an intermediate speech representation (mel-spectrogram), which can be seen as a lossy compressed version of the audio signal. Then, the vocoder model is responsible of reconstructing the final speech waveform conditioned on the mel-spectrograms. For the vocoder model, a public implementation of HiFi-GAN [SJF20] was used, which is capable of producing high speech audio quality significantly faster than real-time both on GPU and CPU.

The rest of this section is organized as follows. First, the data processing and the different tools used for this purpose are detailed in Section 5.4.2. Then, the proposed forced-aligner autoencoder model used to extract phoneme durations is presented in Section 5.4.3. The acoustic TTS model architecture is described in detail in Section 5.4.4. Section 5.4.5 introduces the GAN-based vocoder model used. Finally, the results of the subjective evaluation test are given in detail and discussed in Section 5.4.6.

5.4.2 Data processing

In the 2021 edition, the organization provided participants with about 5 hours of studio-quality recordings (after triming leading and trailing silence) from a native Spanish female speaker and their corresponding transcriptions. The audio samples were provided in 48kHz, PCM 16-bit format. Table 5.1 describes in detail the dataset released for the Blizzard Challenge 2021. As mentioned above, the total duration of the recordings was computed after triming leading and trailing silence from all samples.

Table 5.1: Blizzard Challenge 2021 dataset.

Set	Samples	Number of words	Duration
SH1	4920	50.0 K	5.2 h
SS1	10	154	96 s

[1] `https://github.com/as-ideas/ForwardTacotron`

The text preprocessing procedure was carried out as follows: first, the text transcriptions were normalized (lowercasing, removed special characters, etc.), and then phoneme sequences were extracted from the normalized texts using the well-known open-source speech synthesizer tool eSpeak NG[2].

Regarding the audio processing, all the audio recordings were resampled to 22kHz and leading and trailing silence was removed. Then, 100 bin log magnitude Mel-scale spectrograms with Hann windowing, 50ms window length, 12.5ms hop size and 1024 point Fourier transform were extracted from the audio samples. The spectrograms were finally min-max normalized to lay within the $[0.0, 4.0]$ range.

Last, phoneme durations were extracted by training a separated forced-aligner autoencoder model on the same dataset. This model is described next in Section 5.4.3.

5.4.3 Forced-aligner autoencoder model

Similarly to other non-attentive TTS models with explicit duration modeling [Ren+19; Yu+20; Łań20; She+20; Ren+21], the proposed TTS acoustic model requires of pre-existing phoneme durations (in frames) which are learnt during training. To extract phoneme durations, a monotonic phoneme to frame alignment must be extracted from a separated model. Usually, this is achieved by training a separated attention-based TTS model on the same data (Tacotron, Transformer TTS [Li+19], etc.) or by using an external forced alignment tool with pre-trained models like the Montreal Forced Aligner [McA+17]. The former has been found to provide slightly more convenient alignments as the aligner model is trained on the same TTS regression task as the acoustic model, as opposed to a pure classification task.

Nevertheless, under suboptimal conditions (inaccurate transcriptions, noisy recordings, small datasets, etc.) the training convergence of attention-based TTS models is not guaranteed, particularly when the attention mechanism does not constrain the alignments to be monotonic [He+19; Bat+20]. Also, the training of the attention-based model becomes computationally more expensive than training the final TTS acoustic model.

For these reasons, and inspired by Axel Springer Ideas Engineering's DeepForcedAligner[3], a forced-aligner autoencoder model is proposed. It makes use of an auxiliar connectionist temporal classification (CTC) loss [Gra+06] to find

[2] `http://espeak.sourceforge.net`
[3] `https://github.com/as-ideas/DeepForcedAligner`

the alignments between the spectrogram frames and the phoneme sequences. This model is trained on the same speech data as the TTS model. Also, the autoencoder framework is able to refine the CTC (or pure speech recognition) alignments and make them more suitable for the TTS task. Last but not least, it provides enhanced robustness and significantly faster convergence than attention-based TTS models.

Figure 5.3 depicts the proposed forced-aligner model. It is composed of two interconnected modules following an autoencoder framework, which are trained end-to-end with the help of an auxiliar CTC loss. The speech-to-text (STT) encoder module is a simple speech recognition model. The input spectrogram frames are passed through a stack of 5 1-D convolutional layers, followed by batch normalization and ReLU activations. The output of the last convolution layer is then processed by a single-layer bidirectional LSTM, followed by a final linear projection layer with softmax activations. An auxiliar CTC loss computed over the ground-truth phoneme sequences is used on the softmax outputs to help the convergence of the STT module. Then, a simple TTS module plays the decoder role of the autoencoder framework. This module takes the STT softmax outputs (phoneme class probabilities) as inputs, and aims to reconstruct the original spectrogram frames. It is composed of 2 bidirectional LSTM layers followed by a linear projection to the spectrogram dimension. The mean absolute error (MAE) between the ground-truth and the generated spectrograms is backpropagated through the entire autoencoder model helping refining the STT alignments for the TTS task.

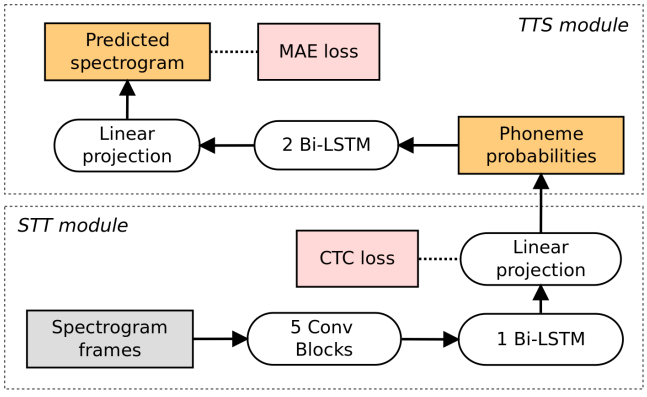


Figure 5.3: Forced-aligner autoencoder architecture overview.

To extract phoneme durations, the decoder (TTS) module is discarded, and the speech-to-text encoder is solely used to calculate phoneme posteriors. Then, Dijkstra's algorithm can be used to find the most likely monotonic path through the sequence of phoneme probabilities.

5.4.4 Acoustic model

The proposed acoustic model was initially based on ForwardTacotron[4], an open-source non-autoregressive variant of Tacotron [Wan+17] inspired on parallel TTS models like FastSpeech [Ren+19] or DurIAN [Yu+20]. The original ForwardTacotron architecture was composed of two Pre-Net bottleneck layers, a CBHG encoder [Wan+17], a variance duration predictor similar to [Ren+19], a 2-layer BiLSTM decoder and a convolutional residual Post-Net as in [She+18]. However, following more recent works, several modifications are introduced to this original architecture. The final architecture is depicted in Figure 5.4, and the introduced modifications are described in detail in what follows.

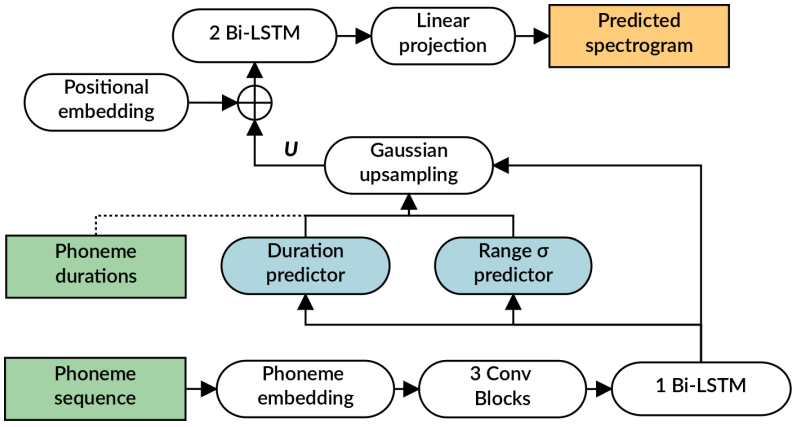


Figure 5.4: Proposed TTS acoustic model architecture for the Blizzard Challenge 2021. Dashed lines are training-only connections.

First, the encoder module is replaced with the simplified architecture proposed in Tacotron 2 [She+18] shown at the bottom of Figure 5.4. The encoder module consists of learned 512-dimensional phoneme embeddings that are 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

[4] https://github.com/as-ideas/ForwardTacotron

processed by a single bidirectional LSTM layer to generate the encoder hidden states. These states are later expanded attending to phoneme durations and consumed by a decoder to generate the spectrogram frames.

Second, the vanilla upsampling through repetition (also known as length regulator module) is replaced in favor of the Gaussian upsampling approach recently proposed in [She+20], which has been shown to improve speech naturalness. In the Gaussian upsampling, encoder hidden states $\mathbf{h} = [\mathbf{h}_1, ..., \mathbf{h}_N]$ are expanded into the upsampled sequence $\mathbf{u} = [\mathbf{u}_1, ..., \mathbf{u}_T]$ as follows:

$$c_i = \frac{d_i}{2} + \sum_{j=1}^{i-1} d_j \tag{5.1}$$

$$w_{t,i} = \frac{\mathcal{N}(t; c_i, \sigma_i^2)}{\sum_{j=1}^{N} \mathcal{N}(t; c_j, \sigma_j^2)} \tag{5.2}$$

$$\mathbf{u}_t = \sum_{j=1}^{N} w_{t,i} \mathbf{h}_i \tag{5.3}$$

where $\mathbf{d} \in \mathbb{Z}^N$ are phoneme-level integer duration values (in frames), and the range $\sigma \in \mathbb{R}^N$ is a learnable parameter of the model.

In addition, after the upsampling, Transformer-style sinusoidal positional embeddings of a frame position with respect to the current phoneme are added [Eli+20]. Also, the recurrent layer of the variance predictor module is discarded as we empirically found improved robustness on phoneme duration predictions, in line with [Ren+19].

Finally, the convolutional residual Post-Net is removed as we found it not to bring any noticeable improvements on the spectrogram reconstruction task when using a bidirectional LSTM decoder.

It is also worth mentioning we did not include explicit modeling of pitch and energy in this case, as we found no improvements in terms of naturalness nor audio quality. We believe the high quality recordings of the training data, produced in a controlled professional-setting environment, made it unnecessary to further leverage the one-to-many mapping problem to improve resulting audio quality, particularly when the vocoder is fine-tuned on the GTA spectrograms.

Table 5.2: Subjective evaluation parts.

Part (Section)	Aspect	Metric
Part 1 and 2	Speaker similarity	MOS (1-5)
Part 3 and 4	Naturalness	MOS (1-5)
Part 5 (Sharvard)	Intelligibility	WER %
Part 6 (SUS)	Intelligibility	WER %

The text-to-spectrogram models are trained using a combination of the $\ell 1$ loss and the *structural similarity index measure* (SSIM) [Wan+04] between the predicted and the target spectrograms, and *Hubber loss* for logarithmic duration prediction [VD20].

5.4.5 Vocoder model

The official public HiFi-GAN implementation[5] is used for reconstructing the audio waveform conditioned on the generated spectrograms. We choose the HiFi-GAN V1 ($h_u = 512$), which is the larger HiFi-GAN model proposed in the original paper and brings the better audio quality compared to the reduced V2 and V3 models [SJF20]. It was trained only on the audio recordings provided by the organization. The training procedure comprises two steps. Initially, the vocoder model is trained on the extracted ground-truth spectrograms for 500K steps. Then, after the acoustic model is trained, ground-truth aligned (GTA) spectrograms are generated for the training dataset and the HiFi-GAN model is fine-tuned on the acoustic model outputs for an additional 100K steps, which helps reducing the artifacts induced by the mismatch between the vocoder training and inference conditions and brings slightly better audio quality for the TTS task.

5.4.6 Subjective results

A total of 12 participating teams submitted their generated test samples to be evaluated in the subjective listening test. Table 5.2 describes the different aspects considered for the subjective evaluation test. Our system was assigned the letter J, while R corresponds to the original audio recordings.

The listeners participating in the subjective test could be divided into three different groups:

[5] `https://github.com/jik876/hifi-gan`

- SP: Paid participants (native speakers of Spanish)
- SE: Volunteer speech experts (self-identified as such)
- SR: Rest of volunteers

Naturalness

In the speech naturalness test (parts 3 and 4), listeners listened to one sample and chose a score which represented how natural or unnatural the sentence sounded on a scale of 1 (completely unnatural) to 5 (completely natural).

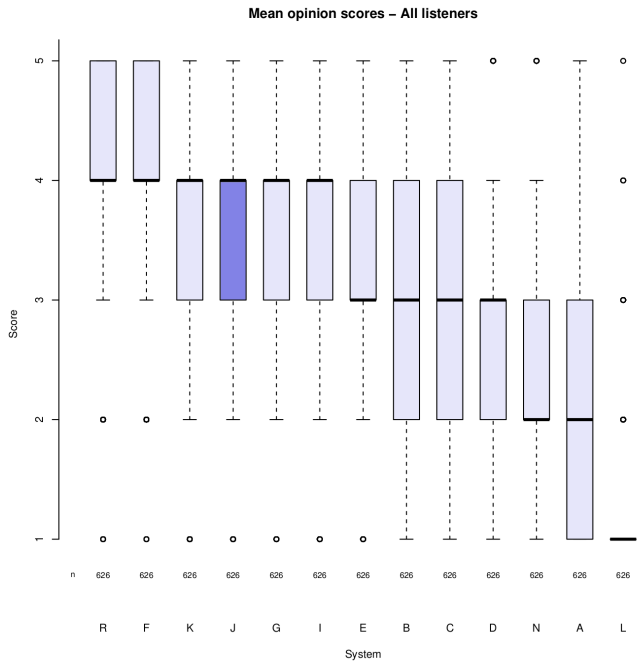


Figure 5.5: Blizzard 2021 naturalness MOS for all participants (all listeners).

The boxplot evaluation results of all systems on speech naturalness from all listeners is showed in Figure 5.5. In this case, Microsoft's system (F) performed clearly better than other systems. It is though worth noting their system was trained using an additional non-public dataset comprising 80 hours of high quality Spanish recordings [Liu+21]. Our system was scored with a naturalness MOS of 3.61 in terms of speech naturalness, while the real recordings were

scored with 4.21. As can be seen in Figure 5.5, our system performance is comparable to that of systems K (Samsung Research China), G (CPQD-Unicamp) and I (IOA-ThinkIT), only outperformed by system F (Microsoft). Among these, systems F, G and I included external data for training the acoustic and vocoder models. This is a very positive result, especially considering both the acoustic and vocoder models were trained with a limited amount of data (∼5 hours) compared to what is common in two-stage neural TTS pipelines (20 hours or more) [IJ17; VYM17].

Speaker similarity

In the speaker similarity test (parts 1 and 2), listeners could play 2 reference samples of the original speaker and one synthetic sample. They chose a response that represented how similar the synthetic voice sounded to the voice in the reference samples on a scale from 1 (sounds like a totally different person) to 5 (sounds like exactly the same person).

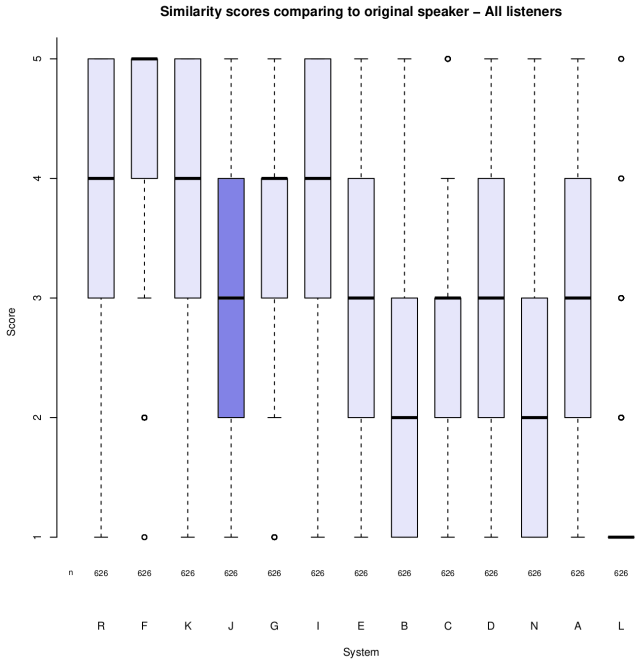


Figure 5.6: Speaker similarity scores for all participants (all listeners).

The boxplot evaluation results of all systems on speaker similarity from all listeners is showed in Figure 5.6. System F, K and I performed better than other systems. Despite the fact that our system (J) was only trained with the data provided by the Blizzard Challenge 2021 organizers, it was scored with a speaker similarity MOS of 3.29 compared to 4.07 from the original recordings (R).

Intelligibility test

Finally, an intelligibility test was carried out in parts 5 and 6. The goal of this test was just to determine whether or not the synthetic speech was understandable. Listeners heard one utterance in each part and typed in what they heard. Listeners were allowed to listen to each sentence only once. The sentences were specially designed to test the intelligibility of the synthetic speech: the sentences and the reference natural recordings for part 5 came from the Sharvard corpus, and the SUS test samples for part 6 were kindly provided by TALP-UPC and Aholab-EHU research laboratories.

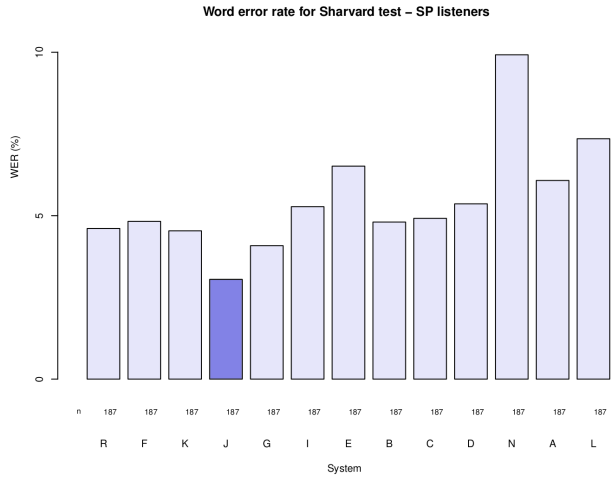


Figure 5.7: Intelligibility Word Error Rates (WER) for the Sharvard intelligibility test.

Figure 5.7 shows the Word Error Rates (WER) of all teams for the Sharvard intelligibility test, where our system achieved the lowest transcription error (3.0%). This result emphasizes the good performance of the proposed system regarding word pronunciation even though it was trained with a limited amount

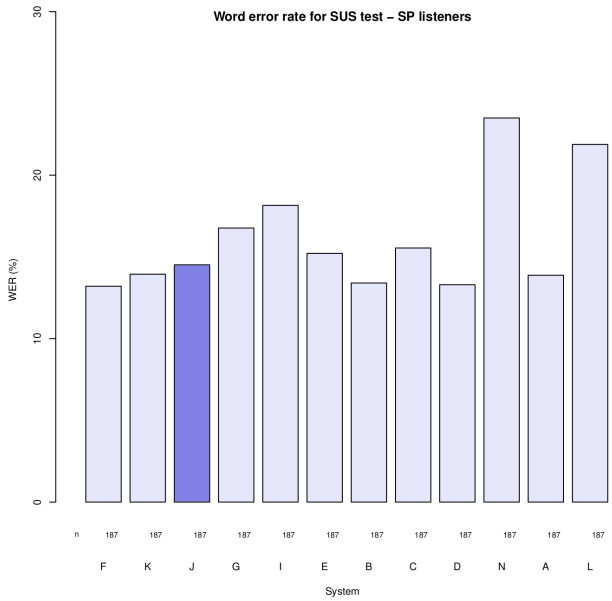


Figure 5.8: Intelligibility Word Error Rates (WER) for the SUS intelligibility test.

of speech data. Figure 5.8 shows the WER for the SUS intelligibility test, where the performance is similar to the best performing systems.

5.5 Conclusions

This chapter has addressed robustness, efficiency and controllability in modern neural TTS models. The limitations of attentive autoregressive acoustic models introduced in Sections 3.4 and 5.2, particularly regarding these three aspects, can be overcome by explicitly modeling phoneme durations as discussed in Section 5.2. Also, explicit modeling of pitch (F_0) and energy have shown not only to alleviate the one-to-many mapping problem in non-autoregressive TTS models, but also to allow for controllability of those aspects at inference time.

Section 5.3 introduced the more efficient family of GAN-based vocoders, significantly outperforming autoregressive vocoders in terms of inference speed while generating synthetic speech of comparable (when not better) quality.

Finally, the performance of these models has been assessed by participating in the 2021 edition of the Blizzard Challenge, a renowned international challenge

aiming to better understand and compare different TTS technologies. The proposed novel non-autoregressive TTS model with explicit duration modeling achieved excellent results in the subjective evaluation tests, only outperformed by one participating team in terms of speech naturalness.

Chapter 6

Simultaneous Speech-To-Speech Translation

6.1 Introduction

The preceding chapters addressed, on the one hand, the cross-lingual voice cloning task using modern neural TTS architectures and, on the other hand, enhancing robustness, efficiency and controllability of those models. In this chapter, the more ambitious task of building an automatic *simultaneous* (also referred to as *real-time*) Speech-To-Speech translation pipeline is explored, which aligns with the third goal pursued in this thesis.

The proposed cascaded simultaneous S2S system is composed of a streaming ASR, a simultaneous MT and an incremental TTS components. In order to assess the performance of the proposed system, experiments are carried out on the Europarl-ST v1.1 corpus [Ira+20a], a multilingual spoken language translation dataset built from the European Parliament debates archive covering 9 different official European languages. However, for brevity and to ease the subjective evaluation process only English and Spanish languages are considered (both directions: English to Spanish and Spanish to English translations). Nonetheless, the results achieved for this specific language pair can be easily extended to other European languages. In this context, general purpose streaming ASR and simultaneous MT systems covering English and Spanish are developed to satisfy the leading part of the S2S pipeline (i.e. simultaneous speech translation). Yet, the focus is set on the incremental TTS component. The primary goal concerns the development and assessment of an efficient incremental TTS model with cross-lingual voice cloning capabilities as the last component of the cascaded simultaneous S2S pipeline. The TTS system is

trained jointly on the Europarl-ST English and Spanish subsets with the objective of cloning English speakers voices into Spanish and viceversa, so that the translated speech retains the original speaker voice characteristics.

To summarize, the rest of this chapter is organized as follows. First, the Europarl-ST corpus and the English and Spanish subsets used for training the TTS component are described in Section 6.2. Then, the streaming ASR, simultaneous MT and incremental TTS components of the S2S pipeline are described and individually evaluated in Sections 6.3, 6.4 and 6.5, respectively. The latency introduced by the different components and the overall latency of the S2S system is evaluated in Section 6.6. Finally, some conclusions are given in Section 6.7.

6.2 The Europarl-ST dataset

Europarl-ST [Ira+20a] is a multilingual spoken language translation dataset built from the European Parliament debates archive in the period between 2008 and 2012. In its more recent version (v1.1), it contains paired <*audio, transcription, translation*> samples covering 9 different official European languages. The transcription in the original source language of the Members of the European Parliament (MEP) or invited speakers interventions and their corresponding translations are available for different source-target language pairs.

English and Spanish test subsets will be used to assess the performance of the streaming ASR and simultaneous MT components of the S2S system in the Europarl-ST task. To train TTS models, English and Spanish *train*, *dev* and *test* subsets Europarl-ST are jointly considered. The *train-noisy* subsets of this dataset are discarded. The resulting number of hours, words and speakers per language are given in Table 6.1.

Table 6.1: English and Spanish Europarl-ST subsets considered to train TTS models.

Source language	Hours	Words	Speakers
English	72.5	723K	282
Spanish	27.5	276K	85
Total	100.0	1M	367

However, EP recordings present some challenging conditions for the purposes of building TTS systems out of this data. These contain noisy, spontaneous

speech and present other undesired characteristics such as reverberations, leaked room noise and deficient recording settings. In addition, the available transcriptions are not 100% verbatim and often miss hesitations, word repetitions and other disfluencies that are present in the speech track.

6.3 Streaming ASR

In order to adapt the acoustic model to the streaming setup, the one-pass decoder approach presented in [Jor+19; Jor+20] is followed. General-purpose English and Spanish streaming ASR systems are trained on a collection of thousands of speech transcribed hours from public and private datasets according to the recipe described in [Baq+20] using the transLectures-UPV toolkit (TLK) [del+14] and TensorFlow.

The performance of the ASR systems is evaluated in terms of Word Error Rate (WER) on the Europarl-ST English and Spanish test sets. Table 6.2 provides WER figures in percentage for online and streaming ASR systems, showing only a small degradation as a result of the limited speech context available to the latter. The competitive performance shown by the ASR component enables applying MT to the ASR output with minimum error propagation.

Table 6.2: ASR results in terms of WER [%] on the English and Spanish Europarl-ST test sets.

	WER (%)	
System	English	Spanish
Offline	12.7	9.8
Streaming	13.4	10.7

6.4 Simultaneous Machine Translation

First, the continuous stream of words generated by the upstream ASR system is split into non-overlapping chunks that maximize the accuracy of the downstream MT system following the segmentation system proposed in [Ira+20b], which uses a sliding window over the ASR output to carry out the segmentation.

Then, due to its simplicy and effectiveness, the fixed wait-k policy [Ma+18] is followed. This policy waits for k words to be processed in the source text be-

fore starting the translation process. Then, every word processed in the source text generates a word in the target text. Additionally, the effective multi-k approach [Elb+20] is used to train systems that seamlessly may switch between different k at inference time, by sampling a random value for k for each training batch. This approach has been shown to obtain competitive results compared with the latest adaptative policies. An additional advantage of choosing a fixed policy is that the latency between words is fixed and consistent. In contrast, it has been observed that adaptative policies sometimes have spurious, longer than usual delays when translating between some input words. This fact greatly hinders the naturalness of the downstream TTS system, and becomes also an important reason for our preference for a fixed policy over an adaptative one.

As with the ASR component, general-purpose English → Spanish and Spanish → English simultaneous MT systems were trained on millions of sentence pairs from OPUS parallel public datasets [Tie12] under abovementioned settings. The translation quality of the developed simultaneous MT systems is compared to that of the corresponding offline systems on the Europarl-ST corpus in terms of BLEU [Pap+02b]. Table 6.3 reports comparative BLEU [Pap+02b] scores on the Europarl-ST corpus between the conventional offline system and a range of simultaneous wait-k MT systems as a function of k.

Table 6.3: BLEU scores of the offline and simultaneous wait-k MT systems on the Europarl-ST test sets.

System	k	English-Spanish	Spanish-English
Offline	-	47.4	41.3
Wait-k	1	32.2	31.0
	2	37.8	33.9
	4	42.6	35.1
	8	44.1	36.0
	32	44.3	36.1
	100	44.3	36.2

As shown in Table 6.3, there is a relevant gap in performance between the offline and simultaneous MT systems, which decreases as the value of k increases. As can be seen, values of k higher than 8 bring negligible quality improvements. The BLEU difference between both systems when the simultaneous wait-k system works in offline mode ($k = 100$) is due to the use of an unidirectional encoder instead of a bidirectional one.

In streaming settings, translation quality and response-time should be balanced. For lower response-time setups, such as k=2 and k=4, the incurred gap ranges from 5 to 9 BLEU points. All in all, the BLEU scores achieved by simultaneous MT systems ($\simeq$ 34-43) indicate that the quality of these systems is accurate enough for a TTS component downstream.

6.5 Incremental Multilingual Text-To-Speech

The streaming context also requires TTS models to work loosely below real-time in terms of computation time. Thus, efficiency of both acoustic and vocoder models becomes a crucial aspect for neural-based architectures. For the acoustic model, a non-autoregressive model most similar to that described in Section 5.4.4 is devised. The Multi-band MelGAN [Yan+20a] is used as the vocoder model of the TTS component, which brings significant inference speed improvements over similar full-band vocoders while keeping comparable synthesis audio quality. In this work, the incremental prefix-to-prefix framework proposed in [Ma+20] is followed with minor modifications.

6.5.1 Adapted prefix-to-prefix framework

When considering the tight response time constraints of a simultaneous S2S pipeline, it is unpractical to wait until the translated sentence is completed to start the synthesis process. This would incur in significant delays for the speech synthesis, particularly for long utterances. Inspired on the prefix-to-prefix framework adopted for simultaneous MT [Ma+18], [Ma+20] proposes an adaptation for the incremental TTS task. Under this framework adapted to our particular case, the spectrogram and waveform are incrementally generated as:

$$y_t = \Phi(x_{\leq g(t)}, y_{<t}) \tag{6.1}$$

$$w_t = \Psi(y_{\leq h(t)}, w_{<t}) \tag{6.2}$$

where x, y and w represent the input text, the speech spectrogram and the audio waveform, respectively; and $g(t)$ and $h(t)$ are monotonic functions that define the number of words in Eq. 6.1 or frames in Eq. 6.2, being conditioned on when generating the outputs for the t^{th} word.

For the spectrogram generation $g(t)$, lookahead-k policy proposed in [Ma+20] is followed:

$$g(t) = \min(t + k, |x|). \tag{6.3}$$

A maximum history context size δ_y is introduced to limit the computational complexity of the incremental text-to-spectrogram inference process. Thus, Eq. 6.1 becomes:

$$y_t = \Phi(x_{t-\delta_y}^{g(t)}, y_{t-\delta_y}^{t-1}). \tag{6.4}$$

In the waveform generation step, just a small trailing context from the previous chunk is included as an additional conditioning context for the next step. In particular, we set $h(t) = t$ and define a maximum history context size δ_w corresponding to the number of trailing frames from word $t-1$ that is used as additional history context. The use of future context for the vocoding step is discarded.

6.5.2 Model architecture

The TTS component follows the predominant two-stage paradigm of modern neural TTS architectures. The acoustic model proposed for this task is a non-autoregressive model most similar to that described in Section 5.4.4. Speaker and language embeddings are introduced to account for the multiple speakers and languages considered in the dataset. Input texts are converted into IPA (International Phonetic Alphabet) phoneme sequences, where equivalent phonemes are shared across languages. The language embedding helps disambiguate phoneme pronunciations between different languages.

In addition, an adversarial speaker classifier is used to discourage encoder hidden states from containing speaker information [Zha+19c]. The speaker classifier is trained to classify each encoder hidden state into its corresponding speaker ID for a given utterance. Then, a gradient reversal layer [Gan+16] is used to push the encoder to generate speaker-agnostic representations. This layer does nothing during the forward propagation but inverts the sign of the gradients flowing from the speaker classifier during the backpropagation. The overall model architecture is depicted in Figure 6.1.

The acoustic model is trained to minimize a combination of the $\ell 1$ distance and the *structural similarity index measure* (SSIM) between the predicted and the target spectrograms. Additionally, the *Hubber* loss is used for logarithmic

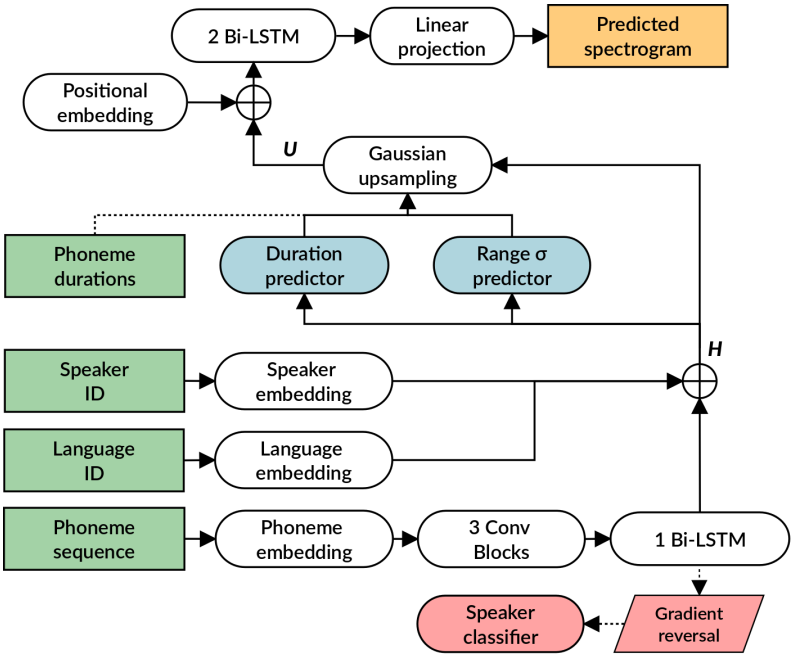


Figure 6.1: Proposed multilingual multi-speaker TTS model architecture. Dashed lines are training-only connections.

duration prediction [VD20], and cross-entropy loss is used for training the adversarial speaker classifier.

A Multi-band MelGAN vocoder is trained for reconstructing the audio waveform conditioned on the generated spectrograms, which is able to generate high quality speech with a real-time factor of 0.03 on CPU [Yan+20a].

6.5.3 Experiments

A baseline reference full-sentence model and lookahead models with values $k \in \{0, 1, 2\}$ are trained on the English and Spanish Europarl-ST subsets described in Table 6.1. For training the lookahead-k models, at each training step the first n words of each sample are selected at random (where $n \in \mathbb{N}$ can be a different value for different samples). Then, the lookahead context is limited to a maximum of k subsequent words (when available). Phoneme durations are used to determine the spectrogram frames corresponding to dif-

ferent words. This training procedure of the lookahead-k models is compliant with the inference conditions for different values of k.

All audio recordings are resampled to 22kHz. Then, 100-bin log magnitude mel-scale spectrograms are extracted from audio samples with Hann windowing, 50ms window length, 12.5ms hop size and 1024 point Fourier transform. Phoneme durations are predicted from a separate autoregressive attention-based Tacotron 2 model trained on the same data. All acoustic models are trained with a batch size of 48 and an initial learning rate of 5×10^{-4}, which is decayed by half every 40K steps for a total of 160K steps. The Adam optimizer [KB15] is used with default parameters and gradient clipping at 1.

To reduce the harmful effects of noisy samples (i.e. containing significant background noise or inaccurate transcriptions), a dynamic loss masking approach is devised. The purpose of this dynamic loss masking approach is to discard gradients coming from samples for which the mel-spectrogram reconstruction loss is significantly higher than a certain threshold. However, this threshold should be adapted dynamically through the training process as the model learns to better reconstruct target spectrograms. In this case, this threshold is computed dynamically as the median plus λ times the median absolute deviation of the $\ell 1$ loss from the last 5 batches, where λ is manually adjusted empirically. The dynamic loss masking is enabled after the first 5K training steps.

At inference time, the maximum history context size parameter δ_y introduced in Section 6.5.1 is set to 6 words in all experiments. Ma et al. [Ma+20] also define a chunk-level inference procedure so that the minimum number of words to synthesize at each step is such that the current chunk contains at least l phonemes (where l is a manually defined hyperparameter). This parameter l is set to 6 both for Spanish and English. The effective values of δ_y and l are ∞ for the full-sentence scenario.

Regarding the vocoder, a single Multi-band MelGAN vocoder is trained on the same data. To that end, the public Multi-band MelGAN implementation[1] is used.

[1] `https://github.com/kan-bayashi/ParallelWaveGAN`

6.5.4 Evaluation

To evaluate the naturalness and quality of the synthetic speech, 20 particularly long utterances are used as a test set (10 for each language), where each sample corresponds to a different speaker. We evaluated the speech naturalness of both regular and cross-lingual samples through a 5-scale Mean Opinion Score (MOS) subjective listening test. We also assessed the cross-lingual voice cloning capabilities of the proposed model by analyzing the speaker similarity between the synthetic and ground-truth utterances from the same speaker. Ten native Spanish speakers with proficiency of the English language participated in the subjective listening test. The reader is encouraged to listen to the cross-lingual synthetic samples[2].

Speech naturalness

The degradation in terms of speech naturalness caused by limiting both past and future context for each configuration is evaluated. Tables 6.4 and 6.5 show the 5-scale naturalness MOS with 95% confidence intervals (CI) for the different model configurations, both for English and Spanish synthesis, for regular and cross-lingual samples respectively. Cross-lingual samples refers to those for which the speaker-language pair is unseen during training.

A small degradation can be appreciated on the speech naturalness for the cross-lingual synthesis (e.g. 4.0 compared to 4.2 for the full-sentence case). Nevertheless, this gap is small and endorses the satisfactory performance of the proposed cross-lingual approach. It can be also noted how speech naturalness is degraded for incremental TTS configurations. This might be attributed to two principal factors. First, the impact of limiting the context both to past and future words. Second, the unnatural duration of pauses introduced between consecutive chunks as a result of the proposed incremental prefix-to-prefix approach.

[2]All generated cross-lingual synthetic samples will be temporarily available at `https://mllp.upv.es/interspeech21-demo/`

Table 6.4: Speech naturalness MOS of regular samples with 95% CI.

Regular	English	Spanish
k=0	3.03 ± 0.22	2.67 ± 0.23
k=1	3.17 ± 0.22	3.44 ± 0.21
k=2	3.33 ± 0.22	3.44 ± 0.24
Full-sentence	4.15 ± 0.20	4.17 ± 0.19

Table 6.5: Speech naturalness MOS of cross-lingual samples with 95% CI.

Cross-lingual	English	Spanish
k=0	2.85 ± 0.22	2.26 ± 0.22
k=1	3.23 ± 0.20	2.77 ± 0.23
k=2	3.18 ± 0.20	2.97 ± 0.21
Full-sentence	4.04 ± 0.19	3.99 ± 0.20

Speaker similarity

In the speaker similarity test, participants are asked to rate from 1 to 5 how close the speaker voice of cross-lingual synthetic samples in the target language sounds compared to ground truth recordings of the same speaker in the source language. Table 6.6 shows 5-scale speaker similarity MOS with 95% CI. This evaluation is limited to the full-sentence scenario as the incremental settings should have no impact regarding voice cloning capabilities.

Table 6.6: Speaker similarity MOS with 95% CI of cross-lingual samples compared with a reference utterance.

Model	English	Spanish
Full-sentence	3.84 ± 0.22	3.67 ± 0.21

The encouraging results shown in Table 6.6 regarding speaker similarity assess the cross-lingual voice cloning capabilities of the TTS component.

6.6 S2S latency evaluation

The accumulated latency of the resulting simultaneous English ⇄ Spanish S2S systems is measured similarly as in [Li+20; Ira+20b]. Accumulative chunk-level latencies are defined at five successive points in the S2S pipeline, as the time elapsed between the last word of a chunk being spoken and:

1) the consolidated hypothesis for that chunk is provided by the ASR system;
2) the segmenter defines that chunk on the ASR consolidated hypothesis;
3) the MT system translates the chunk defined by the segmenter;
4) the TTS finishes synthesizing the translated chunk;
5) the synthesized audio playback is finished.

The latter (5) is closely related to the Ear-Voice Span (EVS) metric commonly used in simultaneous interpretation (i.e. the time lag between the speaker and the interpreter, usually measured in seconds). All five latency figures are reported in Table 6.7 for both translation directions. Latency tests were run on a single-GPU machine equipped with a NVIDIA GeForce GTX 2080 Ti.

Table 6.7: Accumulative latency mean and standard deviation in seconds for the successive points in the S2S pipeline.

Model	English-Spanish	Spanish-English
1) ASR	2.7 ± 1.5	1.8 ± 1.2
2) Segmenter	3.5 ± 1.8	2.8 ± 1.3
3) MT ($k = 4$)	5.2 ± 2.4	4.3 ± 2.0
4) TTS		
$k = 0$	5.6 ± 2.5	4.6 ± 2.0
$k = 1$	5.8 ± 2.6	4.8 ± 2.2
$k = 2$	6.0 ± 2.7	5.1 ± 2.3
5) Playback		
$k = 0$	8.1 ± 1.7	6.4 ± 1.6
$k = 1$	8.7 ± 1.7	7.1 ± 1.7
$k = 2$	9.1 ± 1.7	7.6 ± 1.7

The latency added by the segmentation is mostly explained by the consolidation of the ASR output, that also depends on the long-range dependencies of the neural language models behind. The wait-k ($k = 4$) policy of MT and the

lookahead-k policy of TTS define the corresponding latencies introduced by these components of the pipeline.

The playback delay is explained by the impossibility to recover from the delays introduced in previous chunks when the synthesized speech in the target language is longer or equal to the original speech in the source language. This could be addressed by explicitly controlling the speaking rate of the synthetic speech. The EVS latency of our S2S system configuration ranges approximately from 7 to 9 seconds. This is an acceptable range, not far from that of human interpreters, which varies between 3 to 6 seconds [Led78]. Nevertheless, the S2S pipeline components can be tuned for a different trade-off between response time and translation performance, depending on the application needs.

6.7 Conclusions

In this chapter we presented a novel cascaded simultaneous S2S system built from state-of-the-art streaming ASR, simultaneous MT, and incremental TTS components. The emphasis has been put on the incremental TTS component, designed according to the latest developments in the field as an end-to-end deep learning architecture including a number of refinements for fast, incremental multilingual and multi-speaker TTS.

The assessment of the proposed system has been conducted on a realistic simultaneous machine interpretation task for European Parliament debates, from the newly introduced Europarl-ST dataset [Ira+20a]. This choice posed additional challenges, such as training TTS models out of speech data presenting unfavorable conditions (reverberation, room noise, spontaneous speech, non-verbatim transcripts, etc). Each individual component has been first assessed in terms of quality, using appropriate, conventional criteria in each case, and then the full S2S pipeline has been evaluated in terms of latency on a single, standard GPU-enabled PC. Although quality and speed are mutually conflicting objectives, and thus different trade-offs between them can be chosen, we showed that results close to state-of-the-art quality can be achieved by the proposed S2S system running on a standard PC with an *ear-voice span* latency roughly doubling that of human interpreters (3–6 secs).

The figures presented for the streaming ASR and simultaneous MT, as well as the subjective speech naturalness MOS achieved by the proposed TTS models makes us very optimistic about the utilization of these technologies in real-life applications. As future work, individual components in the S2S pipeline can be improved by incorporating long-span dependency neural models that naturally

take advantage of a streaming scenario to improve system accuracy. Also, the speaking rate of the TTS component can be adjusted dynamically according to the accumulated delay at a given time to recover from the successively added delays in the playback. Finally, language models can be used to predict the future context as an alternative to the lookahead-k policy.

The following work was done in collaboration with others:

- The implementation of the streaming ASR models, as well as the system training and evaluation was carried out by Adrià Giménez and Javier Jorge.
- The simultaneous MT systems were trained and evaluated by Javier Iranzo.
- The following international conference article, which was derived from this work, was prepared in collaboration with other members of the MLLP:
  - Pérez-González-de-Martos, A., Iranzo-Sánchez, J., Pastor, A. G., Jorge, J., Silvestre-Cerdà, J. A., Civera, J., Sanchis, A. & Juan, A. (2021). *Towards Simultaneous Machine Interpretation.* Proc. Interspeech 2021, 2277-2281.

Chapter 7

Zero-Shot Speaker Adaptation

7.1 Introduction

Text-to-speech models are generally built to produce synthetic speech in a predefined set of voices, corresponding to those speakers contributing with their speech recordings to the training dataset. Adapting existing models to arbitrary new speakers using a small amount of data (speaker adaptation) remains a challenge in the field.

An effective approach is to fine-tune all or part of the model (e.g. only decoder weights) on the available data from the target speaker. This is usually referred to as few-shot learning or few-shot adaptation, and can provide very satisfactory results given just a few minutes of speech data is available [Che+19; Kon+19]. However, this approach requires, on the one hand, of transcripts to be available and, on the other hand, to go through a time-consuming training process which requires of a proper GPU-enabled infrastructure.

For some applications, the fine-tuning approach becomes unfeasible for a number of reasons: there is not enough adaptation data (e.g. just a few seconds of speech), transcripts are not available, the source and target languages are different, there are tight response-time constraints (e.g. simultaneous S2S), or the computational requirements are too high. Under such constraints, an alternative approach is to build TTS systems capable of *mimicking* or adapting the voice characteristics to that from a given reference utterance, usually comprising no more than a few seconds of speech. This approach is known as zero-shot learning or, more specifically, zero-shot speaker adaptation. Usually, an auxiliary pre-trained speaker encoder network is leveraged to extract speaker characteristics (i.e. speaker embeddings) from the audio recordings, which

are then used to condition the TTS model on the speaker identity [DBM17; Jia+18; Coo+20; Cas+21].

In this chapter, zero-shot speaker adaptation of TTS models in the context of an S2S pipeline is addressed. The main objective is to develop TTS models with cross-lingual mimicking capabilities, that is, models capable of producing speech from unseen speakers in a target language given a reduced number of reference utterances in the source language, usually comprising no more than just a few seconds of speech, are provided. This aligns with the fourth goal raised in this thesis. As in Chapter 4, UPV[Media] is used as a case study and, in particular, the speech-to-speech translation of UPV[Media] online learning materials from Spanish/Catalan into English for unseen speakers. The proposed zero-shot speaker adaptive models are assessed in terms of speech naturalness and speaker similarity compared to ground truth utterances from the DeX-TTS dataset introduced in Section 4.3.

The rest of this chapter is organized as follows. First, the transfer learning approach to zero-shot speaker adaptation is introduced in Section 7.2. Section 7.3 introduces the LibriTTS dataset used for training the zero-shot speaker adaptive English TTS models. Section 7.4 describes the proposed acoustic model architecture. Section 7.5 introduces a novel generative adversarial acoustic model discriminator (GAN) architecture that allows for the prediction of more realistic spectrograms, alleviating the oversmoothing problem commonly present in the spectrogram generation task. Then, Section 7.6 details the training procedure of the proposed models (baseline and GAN) on the LibriTTS dataset. The subjective evaluation of the resulting zero-shot multi-speaker models is addressed in Section 7.7, where both the baseline and GAN approaches are compared in terms of speech naturalness and speaker similarity between reference and synthetic utterances. The integration of the resulting zero-shot TTS systems into the UPV[Media] automatic transcription and translation pipeline is addressed in Section 7.8. Finally, some concluding remarks are given in Section 7.9.

7.2 Speaker conditioning via transfer learning

The usual approach to zero-shot speaker adaptation is to leverage an auxiliary pre-trained speaker encoder network to generate a fixed-dimensional embedding vector for each utterance [Jia+18; Cas+21]. This embedding should summarize the unique attributes in the voice characteristics of a speaker. To obtain best generalization performance, the speaker encoder network is trained on a

speaker classification or verification task using a large and diverse independent dataset comprising thousands of speakers. Then, the TTS network is trained under a transfer learning configuration, where the pre-trained speaker encoder model (whose parameters are frozen) is used to extract speaker embeddings from reference recordings.

Transfer learning is crucial to achieve good performance in the zero-shot adaptation task. By separating the training of the speaker encoder and the TTS models, the requirements for multi-speaker TTS training data are significantly reduced. On the one hand, the speaker encoder does not require of high quality clean speech or transcripts for training, and thus can be trained on a wider variety of speech datasets such as VoxCeleb [CNZ18] or LibriSpeech [Pan+15] accounting for thousands of speakers. On the other hand, though not as important, it avoids the requirement for speaker labels on the TTS training data.

In this work, the speaker encoder network follows the LSTM architecture from [Wan+18a]. It comprises a stack of 3 LSTM layers of 768 units followed by a linear projection layer to the embedding dimension (256). Different than the original work, the model is optimized to minimize the recently proposed Angular Prototypical loss [Chu+20] as in [Cas+21].

7.3 The LibriTTS multi-speaker English corpus

As it was just outlined, generalization is crucial for achieving good performance in the zero-shot adaptation task. So far, it has been discussed how the speaker encoder should be trained on a large and diverse dataset comprising speech recordings from thousands of different speakers for best performance [Jia+18]. However, generalization is no less important when it comes to the TTS task. In this regard, the TTS model should also be trained on a large (yet smaller, due to data availability constraints) variety of speakers for best performance.

To that end, the LibriTTS multi-speaker English corpus [Zen+19] is considered for training the TTS models. The LibriTTS corpus is derived from the original materials (MP3 audio files from the LibriVox project [Lib] and texts from Project Gutenberg[1]) of the LibriSpeech corpus and is distributed under the same non-restrictive license. The main differences with respect to the original LibriSpeech corpus, which was originally conceived for the ASR task, are the following: the audio files are at 24kHz sampling rate; the speech is split at sen-

[1] `https://www.gutenberg.org/`

tence breaks; both original and normalized texts are included; and utterances with significant background noise are excluded.

Table 7.1 summarizes the different data subsets of the LibriTTS corpus. It comprises a total of 2,456 different speakers accounting for 585.8 hours of human-read speech recordings. LibriTTS is, so far, the largest multi-speaker English TTS open-sourced corpus available, and thus a very appropiate choice for the task of zero-shot speaker adaptation.

Table 7.1: Data subsets in LibriTTS

Subset	Hours	Female speakers	Male speakers	Total speakers
dev-clean	9.0	20	20	40
test-clean	8.6	19	20	39
dev-other	6.4	16	17	33
test-other	6.7	17	16	33
train-clean-100	53.8	123	124	247
train-clean-360	191.3	430	474	904
train-other-500	310.1	560	600	1,160
Total	585.8	1,185	1,271	2,456

However, the LibriTTS corpus poses a challenge when it comes to train high-quality natural sounding TTS models due, mainly, to the heterogeneous audio acquisition (acoustic) conditions of the speech recordings. This variability, which is not explained by text, speaker embeddings, pitch or energy inputs, can result in oversmoothing issues on the spectrogram generation task, particularly (but not only) regarding higher frequencies. When paired with neural vocoders, oversmoothed spectrograms tend to produce more noisy and metallic synthetic speech, directly affecting the perceived speech quality.

7.4 Proposed zero-shot multi-speaker architecture

According to the latest developments in the field, the proposed system uses a stack of Conformer blocks [Gul+20] both as the encoder and decoder modules. Conformer is a Transformer [Vas+17] variant integrating both convolutional and Transformer components. A Conformer block is composed of four modules stacked together: a feed-forward module, a self-attention module, a convolution module, and a second feed-forward module. However, this original architecture is modified following the recommendations from [Liu+21]. First, the Swish

activation function is replaced with ReLU for better generalization, particularly on long sentences. Second, the depthwise convolution is placed before the self-attention module for faster convergency. Finally, the linear layers in feed-forward modules are replaced by convolution layers. Figure 7.1 shows the improved Conformer block.

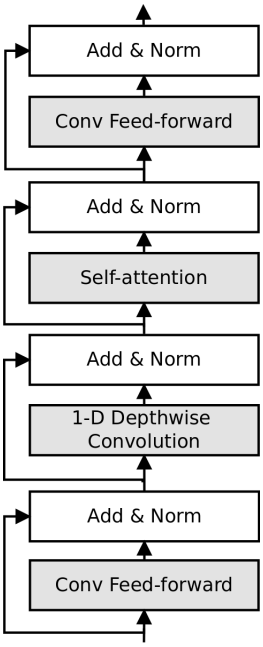


Figure 7.1: Modified Conformer block as in [Liu+21].

LibriTTS and similarly large multi-speaker datasets coming from heterogeneous sources usually present very different recording/acoustic conditions among speakers (microphone and recording settings, room reverberations, etc). Although this certainly helps generalization in classification tasks such as ASR, it poses additional challenges for building good performing TTS models. In the transfer learning configuration, the speaker encoder is built to be robust against different acoustic conditions, where it is common to perform data augmentation by adding different types of noises (e.g. background music, reverberations, Gaussian noise, etc.) to the original utterances. Thus, ideally, the resulting speaker embeddings should not account for the acoustic environment information. Without explicit modeling of the acoustic conditions, the TTS is forced to jointly model speaker and acoustic information, which can result in suboptimal performance.

To model variability in acoustic conditions, a Variational Auto-Encoder [KW14] (VAE) reference encoder similar to [Zha+19b] is introduced, which constructs a relationship between unobserved continuous random latent variables $\mathbf{z}$ and observed dataset $\mathbf{x}$. The VAE encodes a reference audio into a fixed-length short vector (reference embedding) of latent representation by introducing a recognition model $q_\phi(\mathbf{z}|\mathbf{x})$ as an approximation to the intractable posterior $p_\theta(\mathbf{z}|\mathbf{x})$. The reader is referred to [Zha+19b] for further details. This reference embedding is broadcast-concatenated to the encoder hidden states to condition the spectrogram reconstruction on latent representations $\mathbf{z}$ extracted from the reference audio. However, different from similar works on unsupervised style modeling in TTS [Wan+18c; Zha+19b], we propose to limit the reference encoder input to a reduced fixed-length random segment from each training utterance (e.g. 1 second) to prevent it from modeling other long-term speech attributes such as speaking style or prosody. During inference, a fixed reference embedding obtained empirically (e.g. feeding a high quality clean recording to the reference encoder) is used.

Figure 7.2 depicts the overall model architecture. The Conformer encoder and decoder modules consist of 6 Conformer blocks with attention dimension 384 and a kernel size of 1536 for convolutional feed-forward modules. The speaker encoder network follows the configuration described in Section 7.2. The variance adaptor modules (duration, pitch and energy predictors) follow the convolutional architecture in [Ren+21] with 3, 5 and 2 layers, respectively. The pitch prediction is done similarly as in [Łań20], where frame-wise F_0 values are first converted to the logarithmic domain and averaged over every input symbol using phoneme durations. Then, predicted (or ground truth) phoneme-level pitch values are projected to the encoder hidden states ($\mathbf{h}$) dimensionality by means of a 1-D convolution and added to $\mathbf{h}$.

7.5 Least Squares Generative Adversarial Networks for TTS acoustic modeling

Even though the introduced VAE framework can help better modeling acoustic information of speech recordings, preliminary results show predicted spectrograms still suffer from oversmoothing, as the comparatively large LibriTTS dataset size hinders the acoustic model from predicting fine-grained spectrogram details using regular $\ell 1$ or $\ell 2$ reconstruction losses. This results in unpleasant noisy and metallic synthetic speech as a consequence of the significant mismatch between vocoding training and inference conditions. Although the fine-tuning of the vocoder on the ground truth aligned (GTA)

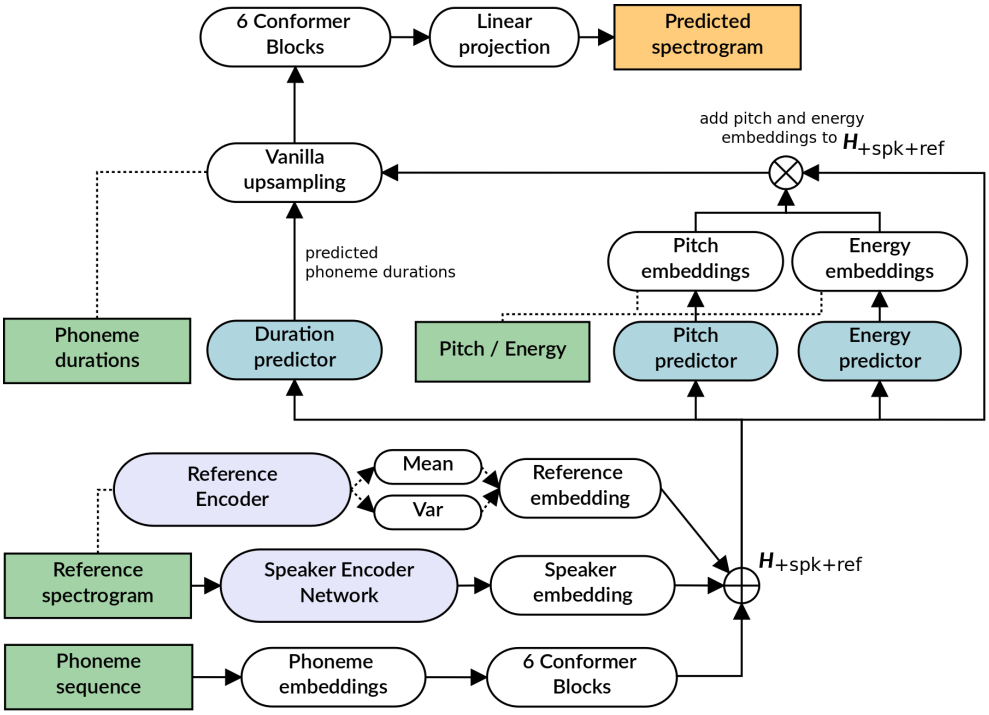


Figure 7.2: Conformer (baseline) zero-shot TTS model based on Conformer Encoder/Decoder blocks (dashed lines correspond to training-only connections).

predicted spectrograms has been shown to reduce this gap and improve audio quality [She+18; KKB20], non-autoregressive acoustic models cannot leverage teacher forcing to predict GTA features. This results in a larger mismatch between predicted (GTA) and ground truth training set spectrograms, particularly on large datasets such as LibriTTS as a consequence of the improved generalization, hindering the fine-tuning process.

In the recent years, generative adversarial networks (GANs) [Goo+14] have shown powerful performance in a wide variety of generative tasks, particularly in the image domain [Wan+18b; Zha+19a]. GAN-based models have also been successfully adopted in the speech domain, especially in neural vocoder models to improve inference speed with respect to their autoregressive counterparts [YSK20; Kum+19; KKB20]. However, although there have been some prelimiary works on applying GANs for the spectrogram reconstruction task, this is not yet a widely adopted approach for improving TTS acoustic modeling.

The authors in [SHP19] propose using GANs to train a separate spectrogram enhancer model that reconstructs fine-grained details from oversmoothed predicted spectrograms. A GAN-based end-to-end TTS approach is proposed in [Biń+20] with relative success, yet showing worse performance in terms of speech naturalness and audio quality than regular two-stage autoregressive models. More recently, the authors in [Yan+21] showed slight improvements on speech naturalness over a baseline FastSpeech2 [Ren+21] non-autoregressive model by adding a joint conditional and unconditional (JCU) adversarial loss [Zha+19a] similar to [Yan+20b], where the JCU discriminator is conditioned on speaker embeddings.

In this work, following [Yan+21], we propose to introduce a spectrogram discriminator to enable adversarial training of the zero-shot multi-speaker Conformer model described in Section 7.4. However, we propose a novel simplified unconditional discriminator architecture capable of producing artifact-free realistic spectrograms using a regular frame-wise least-squares adversarial loss [Mao+17]. The discriminator comprises 2 convolutional layers with 256 5×1 filters (i.e. each filter spans 5 frames), followed by batch normalization [IS15] and ReLU activations. The output of the last convolutional layer is passed into a single bi-directional LSTM layer containing 128 units per direction. Finally, LSTM outputs are linearly projected down to a scalar predicting the frame-wise least-squares score values.

In Least Squares Generative Adversarial Networks (LSGANs), the generator (G) and discriminator (D) are generally optimized to minimize the following losses:

$$\mathcal{L}_D(G, D) = \sum_{n=1}^{N} \frac{1}{2}(D(y_n) - r)^2 + \frac{1}{2}(D(G(z_n)) - f)^2 \tag{7.1}$$

$$\mathcal{L}_{G_{adv}}(G, D) = \sum_{n=1}^{N} (D(G(z_n)) - r)^2 \tag{7.2}$$

where in this context y_n and z_n denote, respectively, the target spectrogram and the generator inputs (text, pitch, energy, etc.) corresponding to sample n; and r and f are the labels for the *real* and *fake* data, which are commonly set to 1.0 and 0.0.

Figure 7.3 shows the log-scale mel spectrogram of the same LibriTTS test utterance predicted by the baseline and the LSGAN model compared to ground

truth. As it can be seen, the adversarially trained model is able to predict more sharpen and realistic spectrograms that are richer in details compared to the baseline model, particularly regarding higher frequencies. This results in overall better audio quality of final audio waveforms, as a consequence of the reduced gap between training and inference conditions of the neural vocoding step.

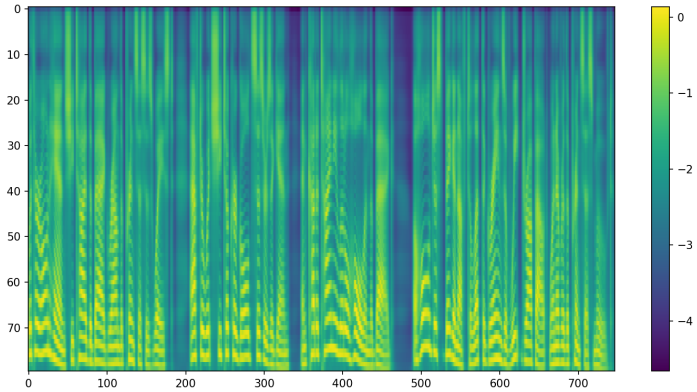
7.6 Experiments

This section describes the zero-shot speaker adaptive models trained for the speech-to-speech translation of UPV[Media] online learning materials into English. Both the baseline zero-shot TTS model described in Section 7.4 and the extended LSGAN model described in Section 7.5 are trained on the full LibriTTS multi-speaker dataset, comprising a total of 585 hours of read speech recordings from more than 2400 different speakers.

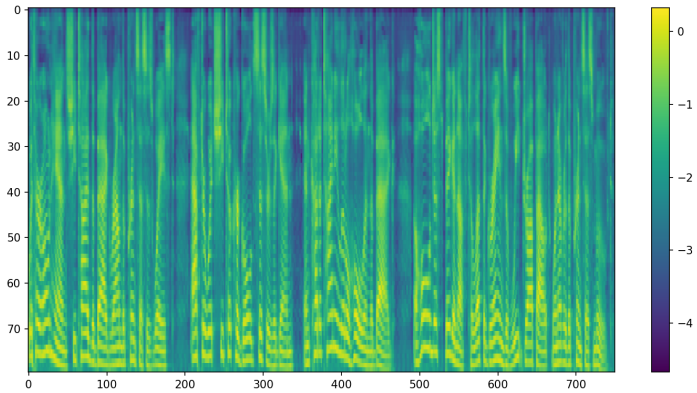
The training procedure is as follows. First, leading and trailing silence from all LibriTTS audio recordings is removed. Then, audios are downsampled to 16kHz for acoustic model training, while 24kHz audios are used for training a vocoder model that generates 24kHz audio conditioned on the predicted 16kHz spectrograms. This design choice is similar to [Liu+21], where the vocoder generates 48kHz audio from 16kHz spectrograms, which can better trade off training efficiency, modeling stability and voice quality. Then, 80-bin log magnitude Mel-scale spectrograms with Hann windowing, 50ms window length, 12.5ms hop size and 1024 point Fourier transform are extracted from the 16kHz audio samples. Phoneme sequences are extracted from normalized text transcriptions using the eSpeak NG[2] tool. Frame-wise pitch (F_0) values are estimated using the WORLD [MYO16] vocoder toolkit [MKK09]. To extract phoneme durations, the forced-aligner autoencoder model described in Chapter 5 (Section 5.4.3) is trained on the same LibriTTS task.

A pre-trained speaker encoder model trained on a speaker classification task including LibriTTS, VCTK [VYM17], VoxCeleb [CNZ18] and CommonVoice [Ard+19] datasets is leveraged for computing speaker embeddings from recorded utterances [Cas+21][3]. The reference encoder follows the same architecture as [Wan+18c], consisting of six 2-D convolutional layers with consecutive filter sizes of 32, 32, 64, 64, 128 and 128 followed by batch normalization and ReLU activations. Then, the output of the last convolutional layer is fed into a GRU

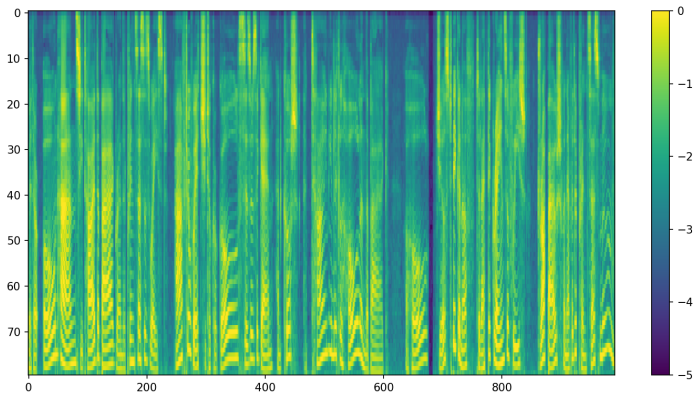
[2] `http://espeak.sourceforge.net`
[3] `https://github.com/Edresson/SC-GlowTTS`



(a) Conformer (baseline)



(b) Conformer LSGAN



(c) Ground truth

Figure 7.3: Log-scale mel spectrogram of the same LibriTTS test utterance predicted with and without adversarial training compared to ground truth.

layer with a hidden size of 128 units. The last hidden state of the GRU, which denotes some 128-dim embedding of the reference audio, is then passed through two separate 32-dim fully connected layers with linear activation to generate the mean and standard deviation of latent variables.

The baseline model is optimized to minimize a combination of the $\ell 1$ loss and the SSIM (Structural SIMilarity index measure) [Wan+04] between reference and predicted spectrograms. Additionally, auxiliar $\ell 1$ losses are used also for the duration, pitch and energy variance prediction modules between reference and predicted values. An auxiliar $\ell 1$ loss between standard deviation values of target and predicted pitch contours (F_0 values) is used to encourage the pitch predictor produce less flattened prosody as the result of training on a huge variety of speakers. Finally, a Kullback-Leibler (KL) divergence regularization term is included to account for the VAE framework as in [Zha+19b].

The GAN model extends the baseline model including the adversarial LSGAN discriminator loss described in Section 7.5 after 500K steps with weight $\lambda_{adv} = 0.1$ for 500K additional steps. The discriminator learning rate is set to 0.0001.

A 24kHz 4-band UnivNet vocoder [Jan+21; Yan+20a] is trained on the LibriTTS recordings to reconstruct the final waveforms conditioned on the 16kHz predicted spectrograms. UnivNet is a recent GAN-based vocoder that has been shown to produce high quality speech of comparable quality to best performing GAN vocoders (HiFi-GAN) while bringing improved inference speed ($\sim 1.5\times$). The model is trained with a batch size of 64 distributed along 4 GPUs for 1M steps.

7.7 Evaluation

To evaluate the performance of the proposed zero-shot multi-speaker systems in cross-lingual settings, a subjective evaluation is carried out to assess different aspects of the resulting synthetic speech. To that end, a number of recordings from the DeX-TTS dataset introduced in Section 4.3 is used as a test set. Since models are trained on the LibriTTS dataset, all test samples correspond to unseen speakers. A total of 20 (10 female and 10 male) speakers from DeX-TTS having recorded samples both in English and either Spanish or Catalan are chosen at random. For each of them, a total of 10 recorded utterances in the source language (either Spanish or Catalan) are randomly selected and used to compute speaker embeddings, accounting for 58 seconds of speech in average per speaker. Additionally, 10 English utterances per speaker are chosen at random among the DeX-TTS recordings, accounting for a total of

200 ground truth samples. Finally, the text corresponding to the selected ground truth utterances is used to generate the synthetic test samples from each model. For completeness, 50 random ground truth recordings from the LibriTTS *test-clean* set are also included in the naturalness test as a native reference, as DeX-TTS recordings are performed by Spanish/Catalan speakers with different proficiency levels of the English language.

A total of 10 participants with high proficiency of the English language participated in the subjective listening tests. The subjective evaluation consisted in evaluating, on the one hand, the resulting speech naturalness of the synthetic and ground truth utterances and, on the other hand, the speaker similarity between human (ground truth) and synthetic recordings from the same speaker in five-point (star) opinion scales. Additionally, an ABX preference test between baseline and LSGAN models is performed.

For naturalness evaluation, participants listened to synthetic and ground truth utterances at random, accompanied by their corresponding texts, and rated speech naturalness from 1 to 5. Table 7.2 shows the naturalness MOS with 95% confidence intervals for both synthetic and control samples, as well as the number of evaluated samples for each case.

Table 7.2: Naturalness MOS with 95% confidence intervals.

	Naturalness MOS	Evaluated
Conformer (baseline)	2.8 ± 0.06	1291
Conformer LSGAN	4.1 ± 0.05	1261
Ground truth (DeX-TTS)	4.9 ± 0.04	387
Ground truth (LibriTTS)	4.9 ± 0.05	306

As can be seen in Table 7.2, the LSGAN model clearly outperforms the baseline Conformer model in terms of speech naturalness. However, we believe the prosody modeling performance of both models (i.e. intensity, vocal pitch, rhythm) is very similar, while the audio quality obtained by the LSGAN model is significantly better due to its ability to reduce the gap between vocoder train and inference conditions.

Another important aspect of the zero-shot multi-speaker models is their ability to *mimick* or produce speech in the voice of any (unseen) speaker given a short reference audio (usually comprising no more than a few seconds). In order to assess this particular aspect, speaker similarity between original recordings and synthetic samples from the same speaker is evaluated. Speaker similarity is an ill-defined similarity measure depending on diverse perceptual speaker

features such as rate, tone, texture or intonation. In the speaker similarity test, participants listen both to an original recording and a synthetic utterance from the same speaker and rate the speaker similarity in a five-point opinion scale. Table 7.3 shows the speaker similarity MOS with 95% confidence intervals for both baseline and LSGAN utterances, as well as the number of evaluated samples.

Table 7.3: Speaker similarity MOS with 95% confidence intervals.

	Speaker similarity MOS	Evaluated
Conformer (baseline)	2.3 ± 0.06	997
Conformer LSGAN	3.0 ± 0.06	1008

Although the LSGAN model obtained significantly better results in terms of cross-lingual voice cloning capabilities when compared to the baseline model, the performance obtained by both models still leaves room for improvement.

Finally, in the ABX preference test participants were shown 200 pairs of synthetic samples, corresponding to the same utterance synthesized by each model, and were asked for their overall preference among the two. Table 7.4 shows the ABX preference test results. As it could be expected from the naturalness results previously shown in Table 7.2, there is a strong preference for the LSGAN synthetic utterances over the baseline ones.

Table 7.4: ABX preference test results.

ABX (%)		
Conformer (baseline)	No preference	Conformer LSGAN
5.1	23.4	71.5

7.8 Integration into UPV[Media] transcription and translation pipeline

As it has been previously mentioned, UPV[Media] makes use of state-of-the-art ASR and MT systems to generate automatic multilingual subtitles of its online educational resources in different languages (see Section 4.2 for more details). The transLectures-UPV Platform (TLP) is used for this purpose [Sil+13b; Pér+15a]. TLP is an open source set of software tools that allows for the integration of automated and assisted transcription and translation technologies into media repositories. In brief, all UPV[Media] contents are automatically ingested into the TLP platform, which is responsible for generating accurate

multilingual subtitles by means of pretrained ASR and MT systems [del+14; Baq+22]. It also provides a web interface for post-editing the automatic transcriptions and translations.

To enable the speech-to-speech translation of UPV[Media] contents into English, the zero-shot Conformer LSGAN model described in previous sections is integrated into the existing TLP ASR+MT pipeline. This allows for the generation of separate (synthetic) audio tracks for all contents including English subtitles (either automatic or reviewed). Users can then select the alternate audio tracks when convinient.

Figure 7.4 depicts the full speech-to-speech translation pipeline. To enhance the performance of the zero-shot TTS model in this context, music, reverberations, ambient sounds and other non-speech events are removed from the original audio tracks before extracting speaker embeddings, which brings improved performance both in terms of resulting synthetic speech and speaker adaptation quality. To that end, the open-source speech enhancement library Asteroid [Par+20] is used along with an existing pretrained speech enhancement model[4]. In this process, the residual audio resulting of substracting the recovered clean speech from the original audio (music, ambient sounds, etc.) is mixed with the synthesized speech for better user experience.

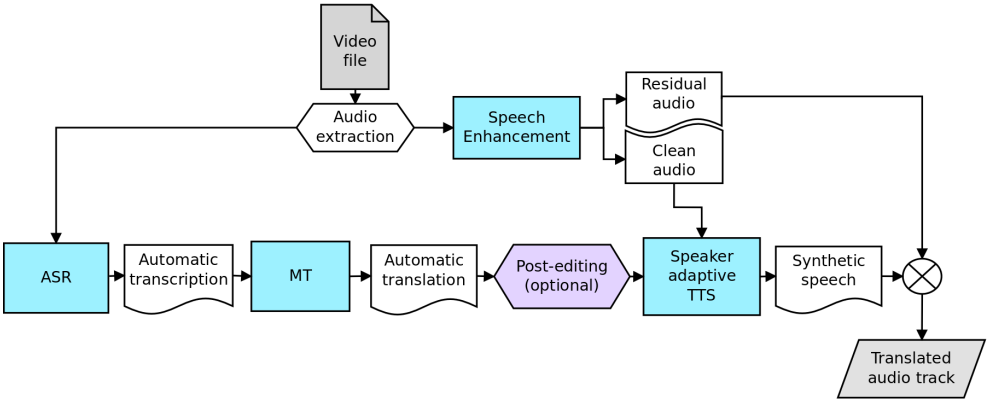


Figure 7.4: TLP speech-to-speech translation pipeline.

[4] https://huggingface.co/JorisCos/DPRNNTasNet-ks2_Libri1Mix_enhsingle_16k

7.9 Conclusions

In this chapter we have addressed the zero-shot adaptation of TTS models to new (unseen) speakers in cross-lingual settings. We have shown how an speaker encoder network trained on a speaker classification task can be leveraged to extract speaker information from reference utterances in a transfer learning configuration. We have proposed a zero-shot TTS model architecture based on Conformer blocks and the VAE framework, capable of achieving remarkably good performance both in terms of resulting speech naturalness and speaker similarity. Additionally, we have proposed a novel spectrogram discriminator architecture that has been proven to be very successful, allowing for the generation of fine-grained realistic spectrograms, and alleviating the oversmoothing issues commonly present in the spectrogram reconstruction task. This resulted in improved overall audio quality and higher naturalness MOS with respect to the baseline model.

The subjective evaluation results endorse the satisfactory performance of the proposed zero-shot model configuration. The Conformer LSGAN model described in this chapter has been successfully integrated in the UPV[Media] production environment, allowing for the speech-to-speech translation (provided of existing good performing ASR and MT systems) of online teaching materials into English while keeping reasonably similar speaker voice characteristics.

Chapter 8

Conclusions and future work

8.1 Scientific and technological achievements

In this section, we discuss the achievement of the scientific and technological goals formulated in Section 1.2. The underlying goal of this thesis has been to enhance the performance and widen the applicability of modern neural TTS systems in real-life settings, both in offline and streaming conditions, in the context of the speech-to-speech translation task.

The first scientific goal of this thesis has been covered in Chapter 3, where state-of-the-art neural TTS architectures have been adapted to multilingual and multi-speaker settings, paying particular attention to cross-lingual voice cloning capabilities. In Chapter 4, the performance of the proposed system has been assessed in the context of its applicability in online learning, where 47 UPV lecturers endorsed the application of this technology for video lecture dubbing.

In regards with the second goal, in Chapter 5 we have proposed a novel non-attentive neural TTS architecture for improved inference robustness, efficiency and controllability. This model has been assessed by participating in the Blizzard Challenge 2021, where it achieved very satisfactory results in terms of speech naturalness and intelligibility, only outperformed by one out of 12 participating teams.

As for the third goal, in Chapter 6 we have assessed the performance of an incremental TTS system with cross-lingual voice cloning capabilities in the context of a simultaneous S2S cascaded system, where we analyzed the degradation caused by limiting both past and future context in terms of speech naturalness.

The fourth goal regarding the zero-shot adaptation of neural TTS models to new speakers has been addressed in Chapter 7. We have shown how a speaker encoder trained on a speaker classification task can be leveraged in a transfer learning configuration to condition the TTS model on the extracted speaker information from reference utterances. We have proposed a GAN model using a novel spectrogram discriminator architecture that has been proven to signficantly outperform a strong Conformer-based baseline in terms of speech naturalness and audio quality.

Finally, the last goal pursued in this thesis has been achieved with the successful integration of the resulting zero-shot models introduced in Chapter 7 into UPV[Media]'s production pipeline to enable the speech-to-speech translation of video lecture materials into English.

8.2 Publications

Most of the work of this thesis has directly yielded articles in international workshops, conferences and journals. In this section we enumerate these contributions to the scientific community, highlighting their relationship with the chapters of this thesis.

The modifications proposed to the original Tacotron 2 architecture to enable cross-lingual voice cloning capabilities along with its assessment in a higher education environment presented in Chapters 3 and 4 resulted in a publication in an international journal:

- Pérez-González-de-Martos, A., Díaz-Munío, G. G., Giménez, A., Silvestre-Cerdà, J. A., Sanchis, A., Civera, J., Jiménez, M., Turró, C. & Juan, A. (2021). *Towards cross-lingual voice cloning in higher education.* Engineering Applications of Artificial Intelligence, 105, 104413.

The robust, efficient and controllable TTS system proposed in Chapter 5 participated in the Blizzard Challenge 2021, achieving excellent results in terms of speech naturalness and intelligibility:

- Pérez-González-de-Martos, A., Sanchis, A., & Juan, A. (2021). *VRAIN-UPV MLLP's system for the Blizzard Challenge 2021.* arXiv preprint arXiv:2110.15792.

The cascaded approach to simultaneous S2S based on state-of-the-art streaming ASR, simultaneous MT and incremental TTS components presented in Chapter 6 resulted in a publication in an international conference:

- Pérez-González-de-Martos, A., Iranzo-Sánchez, J., Pastor, A. G., Jorge, J., Silvestre-Cerdà, J. A., Civera, J., Sanchis, A. & Juan, A. (2021). *Towards Simultaneous Machine Interpretation.* Proc. Interspeech 2021, 2277-2281.

Finally, needless to say that other publications not direcly related with the main topic of this thesis have been produced in collaboration:

- Baquero-Arnal, P.; Jorge, J.; Giménez, A.; Iranzo-Sánchez, J.; Pérez, A.; Garcés Díaz-Munío, G.V.; Silvestre-Cerdà, J.A.; Civera, J.; Sanchis, A.; Juan, A. (2022). *MLLP-VRAIN Spanish ASR Systems for the Albayzín-RTVE 2020 Speech-to-Text Challenge: Extension.* Applied Sciences. 2022, 12, 804.

- Garcés Díaz-Munío, Gonçal V; Silvestre-Cerdà, Joan Albert ; Jorge, Javier; Giménez, Adrià; Iranzo-Sánchez, Javier; Baquero-Arnal, Pau; Roselló, Nahuel; Pérez-González-de-Martos, Alejandro; Civera, Jorge; Sanchis, Albert; Juan, Alfons. (2021). *Europarl-ASR: A Large Corpus of Parliamentary Debates for Streaming ASR Benchmarking and Speech Data Filtering/Verbatimization.* Proc. Interspeech 2021, pp. 3695–3699, Brno (Czech Republic), 2021.

- Jorge, Javier; Giménez, Adrià; Baquero-Arnal, Pau; Iranzo Sánchez, Javier; Pérez-González-de-Martos, Alejandro; Garcés Díaz-Munío, Gonçal V; Silvestre Cerdà, Joan Albert; Civera, Jorge; Sanchis, Albert; Juan, Alfons. (2021). *MLLP-VRAIN Spanish ASR Systems for the Albayzin-RTVE 2020 Speech-To-Text Challenge.* Proc. of IberSPEECH 2021, pp. 118–122, Valladolid (Spain), 2021.

- Piqueras, Santiago; Pérez-González-de-Martos, Alejandro; Turró Ribalta, Carlos; Jiménez, Manuel; Sanchis, Albert; Civera, Jorge; Juan, Alfons. (2021). *Hacia la traducción integral de vídeo charlas educativas.* In Proc. of In-Red 2017 - III Congreso Nacional de Innovación Educativa y Docencia en Red, Valencia (Spain), 2017.

- Pérez González de Martos, A., Silvestre Cerdà, J. A., Rihtar, M., Juan Císcar, A., & Civera Saiz, J. (2014, November). *Using automatic speech transcriptions in lecture recommendation systems.* In Conference Proceedings iberSPEECH 2014: VIII Jornadas en Tecnologías del Habla and IV SLTech Workshop (pp. 149-158). Universidad de Las Palmas de Gran Canaria.

- J. D. Valor Miró, R. N. Spencer, A. Pérez González de Martos, G. Garcés Díaz-Munío, C. Turró, J. Civera, and A. Juan. *Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures.* Open Learning: The Journal of Open, Distance and e-Learning, vol. 29, iss. 1, pp. 72-85, 2014.
- J. D. Valor Miró, R. N. Spencer, A. Pérez González de Martos, G. Garcés Díaz-Munío, C. Turró, J. Civera, and A. Juan. *Evaluación del proceso de revisión de transcripciones automáticas para vídeos poliMedia.* In Proc. of I Jornadas de Innovación Educativa y Docencia en Red (IN-RED 2014), Valencia (Spain), 2014.
- J.A. Silvestre-Cerdà, A. Pérez-González-de-Martos, C. Turró. *A System Architecture to Support Cost-Effective Transcription and Translation of Large Video Lecture Repositories.* IEEE SMC 2013 Conference, October 2013, Manchester (England).

8.3 Future work

This work was focused on enhancing and adapting neural TTS technologies to the context of speech-to-speech translation, both in offline and streaming (real-time) settings. Some open research topics in the TTS field such as cross-lingual voice cloning, incremental TTS or zero-shot speaker adaptation have been addressed with relative success. However, these still remain a challenge in the field and future improvements are to be expected in the next years by using more powerful models and algorithms.

As future work, we intend to keep improving the performance of neural TTS models both in terms of naturalness and efficiency, focusing on their applicability in online teaching and education with the main objective of improving the accessibility and engagement both in live and offline environments. We will also explore the application of these technologies in the media industry (e.g. voiceover or automatic dubbing), where automatic speech recognition and machine translation are already rapidly becoming standardized tools to help professional translators in the subtitling process.

List of Figures

List of Tables

Bibliography

[Led78] Marianne Lederer. "Simultaneous interpretation—units of meaning and other features". In: *Language interpretation and communication*. Springer, 1978, pp. 323–332 (cit. on p. 82).

[GL83] D. Griffin and Jae Lim. "Signal estimation from modified short-time Fourier transform". In: *ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 8. 1983, pp. 804–807. DOI: 10.1109/ICASSP.1983.1172092 (cit. on p. 17).

[Lee88] Kai-Fu Lee. "On large-vocabulary speaker-independent continuous speech recognition". In: *Speech communication* 7.4 (1988), pp. 375–379 (cit. on p. 14).

[Hun90] Melvyn J. Hunt. "Figures of merit for assessing connected-word recognisers". In: *Speech Communication* 9.4 (1990), pp. 329–336 (cit. on p. 20).

[Lee90] K-F Lee. "Context-independent phonetic hidden Markov models for speaker-independent continuous speech recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 38.4 (1990), pp. 599–609 (cit. on p. 14).

[Kub93] Robert F. Kubichek. "Mel-cepstral distance measure for objective speech quality assessment". In: *Proc. of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*. Vol. 1. 1993, pp. 125–128 (cit. on p. 21).

[ITU94] ITU-T. *ITU-T Recommendation P.85: A method for subjective performance assessment of the quality of speech voice output devices*. https://www.itu.int/rec/T-REC-P.85-199406-I/en. Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T). 1994 (cit. on p. 20).

[KN95] Reinhard Kneser and Hermann Ney. "Improved backing-off for m-gram language modeling". In: *1995 international conference on acoustics, speech, and signal processing*. Vol. 1. IEEE. 1995, pp. 181–184 (cit. on p. 14).

[HB96] Andrew J Hunt and Alan W Black. "Unit selection in a concatenative speech synthesis system using a large speech database". In: *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Vol. 1. IEEE. 1996, pp. 373–376 (cit. on p. 16).

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780 (cit. on pp. 7, 26).

[Lav+97] Alon Lavie et al. "JANUS-III: Speech-to-speech translation in multiple languages". In: *Proc. of ICASSP 1997*. Vol. 1. 1997, pp. 99–102 (cit. on p. 22).

[SP97] M. Schuster and K.K. Paliwal. "Bidirectional recurrent neural networks". In: *IEEE Transactions on Signal Processing* 45.11 (1997), pp. 2673–2681. DOI: 10.1109/78.650093 (cit. on p. 26).

[Zar00] Juan Jesús Zaro. "Perspectiva social del doblaje y la subtitulación". In: *Traducción subordinada (I). El doblaje* (2000), pp. 127–138 (cit. on p. 2).

[Rix+01] A.W. Rix et al. "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs". In: *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*. Vol. 2. 2001, 749–752 vol.2. DOI: 10.1109/ICASSP.2001.941023 (cit. on p. 21).

[Pap+02a] Kishore Papineni et al. "BLEU: a Method for Automatic Evaluation of Machine Translation". In: *Proc. of ACL*. 2002, pp. 311–318 (cit. on p. 20).

[Pap+02b] Kishore Papineni et al. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proc. of ACL*. 2002, pp. 311–318 (cit. on p. 74).

[Wan+04] Zhou Wang et al. "Image Quality Assessment: From Error Visibility to Structural Similarity". In: *Image Processing, IEEE Transactions on* 13 (May 2004), pp. 600–612. DOI: 10.1109/TIP.2003.819861 (cit. on pp. 64, 95).

[GS05] A. Graves and J. Schmidhuber. "Framewise phoneme classification with bidirectional LSTM and other neural network architectures". In: *Neural networks* 18.5-6 (2005), pp. 602–610 (cit. on p. 22).

[Gra+06] Alex Graves et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks". In: *Proc. of ICML*. 2006, pp. 369–376 (cit. on p. 60).

[Sno+06] Matthew Snover et al. "A Study of Translation Edit Rate with Targeted Human Annotation". In: *Proc. of AMTA*. 2006, pp. 223–231 (cit. on p. 20).

[Mül07] Meinard Müller. "Dynamic Time Warping". In: *Information Retrieval for Music and Motion*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 69–84. ISBN: 978-3-540-74048-3. DOI: 10.1007/978-3-540-74048-3_4 (cit. on p. 21).

[MKK09] Masanori Morise, Hideki Kawahara, and Haruhiro Katayose. "Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech". In: *Audio Engineering Society Conference: 35th International Conference: Audio for Games*. Audio Engineering Society. 2009 (cit. on p. 93).

[Tur+09] Carlos Turró et al. "Polimedia: a system for successful video e-learning". In: *Proc. of the EUNIS Annual Congress*. 2009 (cit. on p. 38).

[Mik+11] Tomáš Mikolov et al. "Extensions of recurrent neural network language model". In: *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2011, pp. 5528–5531 (cit. on p. 14).

[Tie12] Jörg Tiedemann. "Parallel Data, Tools and Interfaces in OPUS". In: *Proc. of LREC*. 2012 (cit. on p. 74).

[KB13] Nal Kalchbrenner and Phil Blunsom. "Recurrent continuous translation models". In: *Proceedings of the 2013 conference on empirical methods in natural language processing*. 2013, pp. 1700–1709 (cit. on p. 15).

[Sil+13a] J. A. Silvestre-Cerdà et al. "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)*. 2013, pp. 3994–3999 (cit. on p. 41).

[Sil+13b] Joan Albert Silvestre-Cerdà et al. "A System Architecture to Support Cost-Effective Transcription and Translation of Large Video Lecture Repositories". In: *2013 IEEE International Conference on Systems, Man, and Cybernetics*. 2013, pp. 3994–3999. DOI: 10.1109/SMC.2013.682 (cit. on p. 97).

[Sun+13] Antti Suni et al. "Wavelets for intonation modeling in hmm speech synthesis". In: *8th ISCA Workshop on Speech Synthesis* (Jan. 2013), pp. 285–290 (cit. on p. 56).

[Wah13] Wolfgang Wahlster. *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media, 2013 (cit. on p. 22).

[ZSS13] Heiga Ze, Andrew Senior, and Mike Schuster. "Statistical parametric speech synthesis using deep neural networks". In: *2013 ieee international conference on acoustics, speech and signal processing*. IEEE. 2013, pp. 7962–7966 (cit. on p. 16).

[del+14] M. A. del-Agua et al. "The transLectures-UPV toolkit". In: *Proc. of VIII Jornadas en Tecnología del Habla and IV Iberian SLTech Workshop (IberSpeech 2014)*. Las Palmas de Gran Canaria (Spain), Jan. 1, 2014. published (cit. on pp. 73, 98).

[Fan+14] Yuchen Fan et al. "TTS synthesis with bidirectional LSTM based recurrent neural networks". In: *INTERSPEECH*. 2014 (cit. on p. 16).

[Goo+14] Ian J. Goodfellow et al. "Generative Adversarial Nets". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 2014, pp. 2672–2680 (cit. on pp. 13, 57, 91).

[GJ14] Alex Graves and Navdeep Jaitly. "Towards end-to-end speech recognition with recurrent neural networks". In: *International conference on machine learning*. PMLR. 2014, pp. 1764–1772 (cit. on p. 15).

[KW14] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. 2014. arXiv: `http://arxiv.org/abs/1312.6114v10 [stat.ML]` (cit. on p. 90).

[Shi+14] Y. Shi et al. "Efficient one-pass decoding with NNLM for speech recognition". In: *IEEE Signal Processing Letters* 21.4 (2014), pp. 377–381 (cit. on p. 23).

[SVL14] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. "Sequence to Sequence Learning with Neural Networks". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 2014, pp. 3104–3112 (cit. on pp. 7, 15).

[Tur+14] Carlos Turró et al. "Deployment and Analysis of Lecture Recording in Engineering Education". In: *Proc. of 2014 IEEE Frontiers in Education Conference (FIE)*. 2014, pp. 1–5 (cit. on p. 38).

[YD14] Dong Yu and Li Deng. *Automatic Speech Recognition: A Deep Learning Approach*. Springer Publishing Company, Incorporated, 2014. ISBN: 1447157788 (cit. on p. 22).

[Bah+15] Dzmitry Bahdanau et al. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *Proc. of ICLR*. Ed. by Yoshua Bengio and Yann LeCun. 2015 (cit. on pp. 8, 9, 26).

[Ben+15] Samy Bengio et al. "Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, 2015, pp. 1171–1179 (cit. on p. 32).

[CL15] William Chan and I. Lane. "Deep Recurrent Neural Networks for Acoustic Modelling". In: *ArXiv* abs/1504.01482 (2015) (cit. on p. 22).

[Cho+15] Jan Chorowski et al. "Attention-Based Models for Speech Recognition". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, 2015, pp. 577–585 (cit. on pp. 26, 29).

[IS15] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, 2015, pp. 448–456 (cit. on pp. 26, 92).

[KB15] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *Proc. of ICLR*. Ed. by Yoshua Bengio and Yann LeCun. 2015 (cit. on pp. 44, 78).

[LPM15] Thang Luong, Hieu Pham, and Christopher D. Manning. "Effective Approaches to Attention-based Neural Machine Translation". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 1412–1421. DOI: 10.18653/v1/D15-1166 (cit. on p. 9).

[Pan+15] Vassil Panayotov et al. "Librispeech: An ASR corpus based on public domain audio books". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964 (cit. on p. 87).

[Pér+15a] Alejandro Pérez González de Martos et al. *MLLP Transcription and Translation Platform*. Short paper for demo presentation accepted at 10th European Conf. on Technology Enhanced Learning (EC-TEL 2015), Toledo (Spain), 2015. Sept. 16, 2015. published (cit. on p. 97).

[Pér+15b] Alejandro Pérez et al. "MLLP Transcription and Translation Platform". In: *Proc. of the 10th European Conf. on Technology Enhanced Learning (EC-TEL)*. 2015 (cit. on p. 41).

[Val+15a] Juan D. Valor-Miró et al. "Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories". In: *Speech Communication* 74 (2015), pp. 65–75 (cit. on p. 41).

[Val+15b] Juan D. Valor-Miró et al. "Efficient Generation of High-Quality Multilingual Subtitles for Video Lecture Repositories". In: *Proc. of the 10th European Conf. on Technology Enhanced Learning (EC-TEL)*. 2015, pp. 485–490 (cit. on p. 41).

[Cha+16] William Chan et al. "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, pp. 4960–4964 (cit. on p. 15).

[CH16] Kai Chen and Qiang Huo. "Training deep bidirectional LSTM acoustic model for LVCSR by a Context-Sensitive-Chunk BPTT approach". In: *IEEE/ACM TASLP* 24.7 (2016), pp. 1185–1193. DOI: 10.1109/TASLP.2016.2539499 (cit. on p. 22).

[CPS16] Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. "Wav2Letter: an End-to-End ConvNet-based Speech Recognition System". In: *CoRR* abs/1609.03193 (2016). arXiv: 1609.03193 (cit. on p. 15).

[Gan+16] Yaroslav Ganin et al. "Domain-Adversarial Training of Neural Networks". In: *J. Mach. Learn. Res.* 17.1 (Jan. 2016), pp. 2096–2030. ISSN: 1532-4435 (cit. on p. 76).

[Joz+16] R. Jozefowicz et al. "Exploring the limits of language modeling". In: *arXiv preprint arXiv:1602.02410* (2016) (cit. on p. 22).

[MYO16] Masanori MORISE, Fumiya YOKOMORI, and Kenji OZAWA. "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications". In: *IEICE Transactions on Information and Systems* E99.D.7 (2016), pp. 1877–1884. DOI: 10.1587/transinf.2015EDP7457 (cit. on p. 93).

[Oor+16] Aäron van den Oord et al. *WaveNet: A Generative Model for Raw Audio*. arXiv preprint arXiv:1609.03499. 2016 (cit. on pp. 17–20, 25, 57).

[Ran+16] Marc'Aurelio Ranzato et al. "Sequence Level Training with Recurrent Neural Networks". In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2016 (cit. on p. 18).

[Wu+16] Yonghui Wu et al. "Google's neural machine translation system: Bridging the gap between human and machine translation". In: *arXiv preprint arXiv:1609.08144* (2016) (cit. on p. 16).

[ZSN16] Albert Zeyer, Ralf Schlüter, and Hermann Ney. "Towards Online-Recognition with Deep Bidirectional LSTM Acoustic Models". In: *Proc. Interspeech 2016*. 2016, pp. 3424–3428. DOI: 10.21437/Interspeech.2016-759 (cit. on p. 22).

[Ari+17] Sercan Arik et al. "Deep Voice 2: Multi-Speaker Neural Text-to-Speech". In: *Proc. of NIPS*. 2017, pp. 2962–2970 (cit. on pp. 31, 56).

[DBM17] Rama Doddipatla, Norbert Braunschweiler, and Ranniery Maia. "Speaker Adaptation in DNN-Based Speech Synthesis Using d-Vectors". In: *Proc. Interspeech 2017*. 2017, pp. 3404–3408. DOI: 10.21437/Interspeech.2017-1038 (cit. on p. 86).

[IJ17] Keith Ito and Linda Johnson. *The LJ Speech Dataset*. https://keithito.com/LJ-Speech-Dataset/. 2017 (cit. on p. 66).

[Mao+17] Xudong Mao et al. "Least Squares Generative Adversarial Networks". In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 2813–2821 (cit. on p. 92).

[McA+17] Michael McAuliffe et al. "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi." In: 2017 (cit. on p. 60).

[Vas+17] Ashish Vaswani et al. "Attention is All you Need". In: *Proc. of NIPS*. 2017, pp. 5998–6008 (cit. on pp. 10–12, 16, 23, 88).

[VYM17] C. Veaux, J. Yamagishi, and Kirsten MacDonald. "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit". In: 2017 (cit. on pp. 66, 93).

[Wan+17] Yuxuan Wang et al. "Tacotron: Towards End-to-End Speech Synthesis". In: *Proc. of Interspeech*. Aug. 2017, pp. 4006–4010. DOI: 10.21437/Interspeech.2017-1452 (cit. on pp. 17, 18, 44, 62).

[CNZ18] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. "VoxCeleb2: Deep Speaker Recognition". In: *Proc. Interspeech 2018*. 2018, pp. 1086–1090. DOI: 10.21437/Interspeech.2018-1929 (cit. on pp. 87, 93).

[Jia+18] Ye Jia et al. "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis". In: *Proc. of NIPS*. 2018, pp. 4485–4495 (cit. on pp. 86, 87).

[Jin+18] Zeyu Jin et al. "FFTNet: a Real-Time Speaker-Dependent Neural Vocoder". In: *The 43rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2018 (cit. on p. 19).

[Kal+18] Nal Kalchbrenner et al. "Efficient Neural Audio Synthesis". In: *Proc. of ICML*. Vol. PMLR 80. 2018, pp. 2410–2419 (cit. on pp. 19, 25, 35, 57).

[Ma+18] Mingbo Ma et al. "STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework". In: *arXiv preprint arXiv:1810.08398* (2018) (cit. on pp. 23, 73, 75).

[Mam18] Rayhane Mama. *Tacotron-2*. github.com/Rayhane-mamah/Tacotron-2. 2018 (cit. on p. 35).

[McC18] Ollie McCarthy. *WaveRNN*. github.com/fatchord/WaveRNN. 2018 (cit. on p. 44).

[Oor+18] Aaron van den Oord et al. "Parallel WaveNet: Fast High-Fidelity Speech Synthesis". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 3918–3926 (cit. on pp. 19, 57).

[Pin+18] Wei Ping et al. "Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning". In: *Proc. of ICLR*. 2018 (cit. on pp. 17, 18, 20, 54).

[She+18] Jonathan Shen et al. "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions". In: *Proc. of ICASSP*. 2018, pp. 4779–4783 (cit. on pp. 2, 17, 18, 20, 25, 42, 54, 62, 91).

[Wan+18a] Li Wan et al. "Generalized End-to-End Loss for Speaker Verification". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 4879–4883. DOI: 10.1109/ICASSP.2018.8462665 (cit. on p. 87).

[Wan+18b] Ting-Chun Wang et al. "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8798–8807. DOI: 10.1109/CVPR.2018.00917 (cit. on p. 91).

[Wan+18c] Yuxuan Wang et al. "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis". In: *Proc. of ICML*. Vol. PMLR 80. 2018, pp. 5180–5189 (cit. on pp. 54, 90, 93).

[ZLD18] Jing-Xuan Zhang, Zhen-Hua Ling, and Li-Rong Dai. "Forward Attention in Sequence-To-Sequence Acoustic Modeling for Speech Synthesis". In: Apr. 2018, pp. 4789–4793. DOI: 10.1109/ICASSP.2018.8462020 (cit. on pp. 32, 33, 53).

[Ard+19] Rosana Ardila et al. "Common voice: A massively-multilingual speech corpus". In: *arXiv preprint arXiv:1912.06670* (2019) (cit. on p. 93).

[Ari+19] Naveen Arivazhagan et al. "Monotonic Infinite Lookback Attention for Simultaneous Machine Translation". In: *Proc. of ACL*. 2019, pp. 1313–1323 (cit. on pp. 23, 24).

[Che+19] Yutian Chen et al. "Sample Efficient Adaptive Text-to-Speech". In: *International Conference on Learning Representations*. 2019 (cit. on p. 85).

[God19] Robert Godwin-Jones. "In a World of SMART Technology, Why Learn Another Language?" In: *Journal of Educational Technology & Society* 22.2 (2019), pp. 4–13 (cit. on p. 2).

[Gov+19] Prachi Govalkar et al. "A Comparison of Recent Neural Vocoders for Speech Signal Reconstruction". In: *10th ISCA Workshop on Speech Synthesis (SSW 10)* (2019) (cit. on p. 54).

[He+19] Mutian He et al. "Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS". In: *Proc. of Interspeech*. 2019, pp. 1293–1297 (cit. on pp. 32, 34, 53, 54, 60).

[Hsu+19] Wei-Ning Hsu et al. "Hierarchical Generative Modeling for Controllable Speech Synthesis". In: *International Conference on Learning Representations*. 2019 (cit. on p. 54).

[Iri+19] Kazuki Irie et al. "Language Modeling with Deep Transformers". In: *Proc. of Interspeech*. 2019, pp. 3905–3909 (cit. on p. 22).

[Jia+19] Ye Jia et al. "Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model". In: Sept. 2019, pp. 1123–1127. DOI: 10.21437/Interspeech.2019-1951 (cit. on p. 22).

[Jor+19] J. Jorge et al. "Real-Time one-pass decoder for speech recognition using LSTM language models". In: *Proc. of Interspeech*. 2019, pp. 3820–3824. DOI: 10.21437/Interspeech.2019-2798 (cit. on pp. 23, 73).

[Kon+19] Zvi Kons et al. "High Quality, Lightweight and Adaptable TTS Using LPCNet". In: *Proc. Interspeech 2019*. 2019, pp. 176–180. DOI: 10.21437/Interspeech.2019-1705 (cit. on p. 85).

[Kum+19] Kundan Kumar et al. "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019 (cit. on pp. 19, 57, 58, 91).

[Lat+19] Javier Latorre et al. "Effect of data reduction on sequence-to-sequence neural tts". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 7075–7079 (cit. on p. 34).

[Li+19] Naihan Li et al. "Neural speech synthesis with transformer network". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 6706–6713 (cit. on pp. 54, 60).

[Pra+19] Vineel Pratap et al. "Wav2letter++: A fast open-source speech recognition system". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 6460–6464 (cit. on p. 15).

[PVC19] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. "Waveglow: A Flow-based Generative Network for Speech Synthesis". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 3617–3621. DOI: 10.1109/ICASSP.2019.8683143 (cit. on pp. 19, 57).

[Ren+19] Yi Ren et al. "FastSpeech: Fast, Robust and Controllable Text to Speech". In: *Proc. of NIPS*. 2019 (cit. on pp. 20, 42, 54, 55, 60, 62, 63).

[SHP19] Leyuan Sheng, Dong-Yan Huang, and Evgeniy N. Pavlovskiy. *High-quality Speech Synthesis Using Super-resolution Mel-Spectrogram*. 2019. arXiv: 1912.01167 [eess.AS] (cit. on p. 92).

[Spe+19] Matthias Sperber et al. "Attention-passing models for robust and data-efficient end-to-end speech translation". In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 313–325 (cit. on p. 22).

[VS19] Jean-Marc Valin and Jan Skoglund. "LPCNET: Improving Neural Speech Synthesis through Linear Prediction". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 5891–5895. DOI: 10.1109/ICASSP.2019.8682804 (cit. on pp. 19, 57).

[Zen+19] Heiga Zen et al. "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech". In: *Proc. Interspeech 2019*. 2019, pp. 1526–1530. DOI: 10.21437/Interspeech.2019-2441 (cit. on p. 87).

[Zha+19a] Han Zhang et al. "StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.8 (2019), pp. 1947–1962. DOI: 10.1109/TPAMI.2018.2856256 (cit. on pp. 91, 92).

[Zha+19b] Ya-Jie Zhang et al. "Learning Latent Representations for Style Control and Transfer in End-to-end Speech Synthesis". In: May 2019, pp. 6945–6949. DOI: 10.1109/ICASSP.2019.8683623 (cit. on pp. 90, 95).

[Zha+19c] Yu Zhang et al. "Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning". In: *Proc. of Interspeech*. 2019, pp. 2080–2084 (cit. on pp. 26, 31, 76).

[Zhe+19] Yibin Zheng et al. "Forward–Backward Decoding Sequence for Regularizing End-to-End TTS". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.12 (2019), pp. 2067–2079. DOI: 10.1109/TASLP.2019.2935807 (cit. on pp. 33, 53, 54).

[Baq+20] Pau Baquero-Arnal et al. "Improved Hybrid Streaming ASR with Transformer Language Models". In: *Proc. of InterSpeech*. 2020, pp. 2127–2131 (cit. on pp. 14, 22, 73).

[Bar+20] Loïc Barrault et al. "Findings of the 2020 Conference on Machine Translation (WMT20)". In: *WMT*. 2020 (cit. on p. 23).

[Bat+20] Eric Battenberg et al. "Location-Relative Attention Mechanisms for Robust Long-Form Speech Synthesis". In: May 2020, pp. 6194–6198. DOI: 10.1109/ICASSP40776.2020.9054106 (cit. on pp. 53, 54, 60).

[Biń+20] Mikołaj Bińkowski et al. "High Fidelity Speech Synthesis with Adversarial Networks". In: *International Conference on Learning Representations*. 2020 (cit. on pp. 58, 92).

[BOU20] BOUPV20. *Official Bulletin of the UPV*. `http://hdl.handle.net/10251/145577`. Retrieved on June 2020 (in Catalan and Spanish). 2020 (cit. on p. 40).

[Cam+20] Carolien A.N. Knoop-van Campen et al. "Effects of audio support on multimedia learning processes and outcomes in students with dyslexia". In: *Computers & Education* 150 (2020) (cit. on p. 2).

[Chi+20] Erin K. Chiou et al. "How we trust, perceive, and learn from virtual humans: The influence of voice quality". In: *Computers & Education* 146 (2020) (cit. on p. 2).

[Chu+20] Joon Son Chung et al. "In Defence of Metric Learning for Speaker Recognition". In: *Proc. Interspeech 2020*. 2020, pp. 2977–2981. DOI: `10.21437/Interspeech.2020-1064` (cit. on p. 87).

[Cla20] ClassCentral. *The 100 Most Popular Online Courses of All Time (2020)*. `https://www.classcentral.com/report/most-popular-online-courses`. Retrieved on June 2020. 2020 (cit. on p. 38).

[Coo+20] Erica Cooper et al. "Zero-Shot Multi-Speaker Text-To-Speech with State-Of-The-Art Neural Speaker Embeddings". In: May 2020, pp. 6184–6188. DOI: `10.1109/ICASSP40776.2020.9054535` (cit. on p. 86).

[Elb+20] Maha Elbayad et al. "Efficient Wait-k Models for Simultaneous Machine Translation". In: *Proc. of Interspeech*. 2020, pp. 1461–1465 (cit. on pp. 23, 74).

[Eli+20] Isaac Elias et al. *Parallel Tacotron: Non-Autoregressive and Controllable TTS*. arXiv preprint arXiv:2010.11439. 2020 (cit. on pp. 54, 55, 63).

[Gul+20] Anmol Gulati et al. "Conformer: Convolution-augmented transformer for speech recognition". In: *arXiv preprint arXiv:2005.08100* (2020) (cit. on pp. 15, 88).

[Ira+20a] J. Iranzo-Sánchez et al. "Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates". In: *Proc. of ICASSP*. 2020, pp. 8229–8233 (cit. on pp. 71, 72, 82).

[Ira+20b] Javier Iranzo-Sánchez et al. "Direct Segmentation Models for Streaming Speech Translation". In: *Proc. of EMNLP*. 2020, pp. 2599–2611 (cit. on pp. 23, 73, 81).

[Jor+20] Javier Jorge et al. "LSTM-Based One-Pass Decoder for Low-Latency Streaming". In: *Proc. of ICASSP*. 2020, pp. 7814–7818 (cit. on pp. 23, 73).

[KKB20] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 17022–17033 (cit. on pp. 19, 57, 58, 91).

[Łań20] Adrian Łańcucki. *FastPitch: Parallel Text-to-speech with Pitch Prediction*. 2020. arXiv: `2006.06873 [eess.AS]` (cit. on pp. 54–56, 60, 90).

[Li+20] B. Li et al. "Towards Fast and Accurate Streaming End-To-End ASR". In: *Proc. of ICASSP*. 2020, pp. 6069–6073. DOI: `10.1109/ICASSP40776.2020.9054715` (cit. on p. 81).

[Liu+20] Rui Liu et al. "Teacher-student training for robust tacotron-based tts". In: *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2020, pp. 6274–6278 (cit. on pp. 18, 32, 33, 53, 54).

[Ma+20] Mingbo Ma et al. "Incremental Text-to-Speech Synthesis with Prefix-to-Prefix Framework". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Nov. 2020, pp. 3886–3896. DOI: `10.18653/v1/2020.findings-emnlp.346` (cit. on pp. 24, 75, 78).

[Med20] MediaUPV. *The MediaUPV repository*. `https://media.upv.es`. Retrieved on June 2020. 2020 (cit. on p. 38).

[Mia+20] Haoran Miao et al. "Transformer-based online CTC/attention end-to-end speech recognition architecture". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 6084–6088 (cit. on p. 15).

[Ope20] Opencast. *Opencast*. `https://opencast.org`. Retrieved on June 2020. 2020 (cit. on p. 38).

[Par+20] Manuel Pariente et al. "Asteroid: the PyTorch-based audio source separation toolkit for researchers". In: *Proc. Interspeech*. 2020 (cit. on p. 98).

[She+20] Jonathan Shen et al. *Non-Attentive Tacotron: Robust and Controllable Neural TTS Synthesis Including Unsupervised Duration Modeling*. arXiv preprint arXiv:2010.04301. 2020. eprint: `2010.04301` (cit. on pp. 60, 63).

[SJF20] Jiaqi Su, Zeyu Jin, and A. Finkelstein. "HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks". In: *INTERSPEECH*. 2020 (cit. on pp. 59, 64).

[Sud+20] Katsuhito Sudoh et al. "Simultaneous Speech-to-Speech Translation System with Neural Incremental ASR, MT, and TTS". In: *arXiv preprint arXiv:2011.04845* (2020) (cit. on p. 22).

[UPV20a] UPValenciaX. *UPValenciaX: UPV as an edX member*. `https://www.edx.org/school/upvalenciax`. Retrieved on June 2020. 2020 (cit. on p. 38).

[UPV20b] UPVX. *UPVX: The MOOC initiative at the UPV*. `https://www.upvx.es`. Retrieved on June 2020. 2020 (cit. on p. 38).

[VD20] Jan Vainer and Ondřej Dušek. "SpeedySpeech: Efficient Neural Speech Synthesis". In: *Proc. of Interspeech*. 2020, pp. 3575–3579. DOI: `10.21437/Interspeech.2020-2867` (cit. on pp. 64, 77).

[Wan+20] Yongqiang Wang et al. "Transformer-based acoustic modeling for hybrid speech recognition". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 6874–6878 (cit. on p. 15).

[YSK20] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. "Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 6199–6203. DOI: `10.1109/ICASSP40776.2020.9053795` (cit. on pp. 19, 57, 91).

[Yan+20a] Geng Yang et al. "Multi-band MelGAN: Faster Waveform Generation for High-Quality Text-to-Speech". In: *arXiv preprint arXiv:2005.05106* (2020) (cit. on pp. 57, 75, 77, 95).

[Yan+20b] Jinhyeok Yang et al. "VocGAN: A High-Fidelity Real-Time Vocoder with a Hierarchically-Nested Adversarial Network". In: *Proc. Interspeech 2020*. 2020, pp. 200–204. DOI: `10.21437/Interspeech.2020-1238` (cit. on p. 92).

[Yu+20] Chengzhu Yu et al. "DurIAN: Duration Informed Attention Network For Speech Synthesis". In: *Proc. of Interspeech* (2020), pp. 2027–2031 (cit. on pp. 54, 55, 60, 62).

[Zhe+20] Baigong Zheng et al. "Simultaneous Translation Policies: From Fixed to Adaptive". In: *Proc. of ACL*. Association for Computational Linguistics, 2020, pp. 2847–2853 (cit. on p. 23).

[Cas+21] Edresson Casanova et al. "SC-GlowTTS: An Efficient Zero-Shot Multi-Speaker Text-To-Speech Model". In: *Proc. Interspeech 2021*. 2021, pp. 3645–3649. DOI: `10.21437/Interspeech.2021-1774` (cit. on pp. 86, 87, 93).

[Jan+21] Won Jang et al. "UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation". In: *Proc. Interspeech 2021*. 2021, pp. 2207–2211. DOI: 10.21437/Interspeech.2021-1016 (cit. on p. 95).

[Liu+21] Yanqing Liu et al. *DelightfulTTS: The Microsoft Speech Synthesis System for Blizzard Challenge 2021*. 2021. arXiv: 2110.12612 [cs.SD] (cit. on pp. 65, 88, 89, 93).

[PSJ21] Alejandro Pérez-González-de-Martos, Albert Sanchis, and Alfons Juan. *VRAIN-UPV MLLP's system for the Blizzard Challenge 2021*. 2021. arXiv: 2110.15792 [cs.SD] (cit. on p. 58).

[Pér+21] Alejandro Pérez et al. "Towards cross-lingual voice cloning in higher education". In: *Engineering Applications of Artificial Intelligence* 105 (Oct. 1, 2021), p. 104413. published (cit. on p. 51).

[Ren+21] Yi Ren et al. "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech". In: *International Conference on Learning Representations*. 2021 (cit. on pp. 56, 60, 90, 92).

[VA21] Valentin Vielzeuf and Grigory Antipov. "Are E2E ASR models ready for an industrial usage?" In: *arXiv preprint arXiv:2112.12572* (2021) (cit. on p. 15).

[Yan+21] Jinhyeok Yang et al. "GANSpeech: Adversarial Training for High-Fidelity Multi-Speaker Speech Synthesis". In: *Proc. Interspeech 2021*. 2021, pp. 2202–2206. DOI: 10.21437/Interspeech.2021-971 (cit. on p. 92).

[Zha+21] Aston Zhang et al. "Dive into deep learning". In: *arXiv preprint arXiv:2106.11342* (2021) (cit. on p. 9).

[Baq+22] Pau Baquero-Arnal et al. "MLLP-VRAIN Spanish ASR Systems for the Albayzín-RTVE 2020 Speech-to-Text Challenge: Extension". In: *Applied Sciences* 12.2 (2022). ISSN: 2076-3417. DOI: 10.3390/app12020804 (cit. on p. 98).

[Mur22] Kevin P Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022 (cit. on pp. 5, 7, 10, 18).

[DD] Jonathan Duddington and Reece Dunn. *eSpeak NG Text-to-Speech*. https://github.com/espeak-ng/espeak-ng (cit. on p. 44).

[Lib] LibriVox. *LibriVox*. https://librivox.org/ (cit. on p. 87).