Multi-speaker TTS with Deep Learning

Ivan Carapinha
Department of Electrical and Computer Engineering
Instituto Superior Técnico, Universidade de Lisboa
Lisboa, Portugal
ivan.carapinha@tecnico.ulisboa.pt

***Abstract*—Recent advancements in technology have allowed for great development in the field of Speech Synthesis. As such, present-day speech synthesis applications are expected to function for multiple voices, and ensure a fast generation of natural–sounding synthetic speech for enhanced feasibility. This study suggests a multi-speaker text-to-speech (TTS) system for European Portuguese that enables the addition of new speakers without requiring extensive training and data. The proposed model framework comprises two systems: a sequence-to-sequence (Seq2Seq) regressive stage for acoustic feature prediction, followed by a neural vocoder for waveform generation. The model employs a universal vocoder which does not require fine-tuning for new voices.**

The Seq2Seq regressive model predicts acoustic features in the form of Mel-spectrograms by decoding the combination of linguistic embeddings — extracted from the text input —, and speaker embeddings conveying the target speaker identity. The model operates in a multi-speaker setting and can be fine-tuned simultaneously to multiple unseen speakers.

Subjective tests have shown that the proposed model registered comparable performance to another state-of-the-art TTS system, while employing less than half of training data. Furthermore, the proposed model was capable of producing meaningful results when trained with reduced data — under three minutes of speech. At last, the universal vocoder performed, on average, 11 times faster than the speaker–dependent neural vocoder of the state-of-the-art TTS approach used for comparison.

***Index Terms*—Speech Synthesis, Multi-Speaker TTS, Voice Conversion, Voice Cloning**

I. Introduction

Speech is the most natural and immediate form of communication. Nowadays, Speech Synthesis is broadly used in many applications, ranging from voice assistants to speaking aid systems for vocally handicapped people. The demand for speech synthesis in a wide variety of applications propels the development of new approaches that revolutionize the way we use and perceive technology.

The dynamic nature of human speech due to characteristics like language, intonation, vocabulary or accent, poses a demanding challenge for Speech Synthesis systems. Retaining all this information and transforming it such that it can be interpreted by a machine, motivated the conception of many methods that seek to address this subject. The new ability to gather a vast amount of data enabled great progress in speech synthesis applications.

The latest research on Speech Synthesis has developed systems that can generate high-quality natural–sounding speech, and address several voice conversion tasks. However, for the sake of synthetic speech quality, these systems are often trained for very particular configurations, posing several limitations. Most of these applications only operate for a restricted variety of voices and speaking styles. Moreover, these systems are usually configured for data that must attend to very specific requirements in terms of recording conditions and audio quality. Naturally, the more restrictions are raised, the harder is to find data meeting such conditions. For the Portuguese language this problem is even more accentuated as the diversity of speech corpora is considerably smaller than for the English language.

Present synthesis systems are set apart from its predecessors for two essential reasons: 1) they generate synthetic speech of distinctively superior quality; 2) they comprise a great amount of parameters, and employ very large datasets for training. Besides the drawbacks identified in the previous paragraphs, other downsides arise from the complexity of these models: training can be an intricate and exhaustive process, involving the tuning of multiple parameters; synthesis speed at inference time can be slow, often requiring high computational power.

Speech Synthesis is a broad scientific field in constant evolution. Thus, it would not be reasonable to address all the challenges regarding this subject, namely the problems identified in the previous two paragraphs. Within Speech Synthesis, this study focuses on solving the limitations that are pivotal to the performance of text-to-speech (TTS) techniques — that is, systems that generate a synthetic utterance from a given text — in a real-world context: the limited variety of voices for synthesis, and the often slow inference speed of present-day systems.

A. Objectives

The fundamental objective of this study consists in developing a TTS system for European Portuguese (EP), based on existing state-of-the-art implementations of Speech Synthesis systems. To attain this goal, it is first necessary to address the characteristics inherent to EP, in order to ensure correct pronunciation, specially in exceptional cases such as homographs. Moreover, one must identify the main components within the TTS framework and arrange them so the system can adapt to new voices, and ensure reasonable inference speed.

More specifically, this study aims to:

- Adapt a TTS framework to EP, particularly regarding the pronunciation in exceptional cases, namely homographs;

- Incorporate new speaker identities without requiring extensive training;
- Ensure faster synthesis of speech than previously proposed models for EP.

II. Background

A. Neural networks

Neural networks are systems capable of performing classification and regression tasks when given input data. This concept, inspired by the structure and function of the brain, is composed of neurons, connections, weights, and activation functions. Neurons are nodes that generate an output from a combination of received inputs. Each neuron has an activation function. The output of a neuron corresponds to the value of its activation function $f(s)$, where s denotes the combination of its inputs. Connections are weighted links between the output of one neuron and the input of another neuron. Neurons may have multiple input and output connections. The activation function computes the output of a neuron from its inputs, which are combined as a weighted sum of each input. The weight assigned to each input is the weight of the corresponding connection. A bias term can be added to the sum. As such, for n inputs, the combination of inputs s of a neuron is given by equation (1),

$$s = \sum_{i=1}^{n} (x_i w_i) + b \tag{1}$$

where x_i, w_i and b denote input i, weight of connection i and the bias term, respectively. Neural networks group neurons in three types of layers: the input layer, which receives the input data of the neural network, hidden layers, which perform intermediate processing, and the output layer, that produces the output of the neural network according to the desired format. A network should be adapted to the characteristics of input data and the type of problem to be solved. This implies adjusting parameters such as learning rate, and connection weights. The latter can be adjusted through back-propagation [1].

Over the years, the increase of computational power motivated the emergence of deep neural networks (DNNs) and other architectures, namely recurrent neural networks (RNNs) and convolutional neural networks (CNNs), that can detect more meaningful dependencies in the input data than the traditional multilayer perceptron.

B. Speech Synthesis

Speech synthesis aims to generate synthetic speech acceptable to human listeners. It can take in either textual or conceptual input to reproduce the characteristics of the typical human speaking process [2]. Synthesis from text, also known as text-to-speech (TTS), converts written text to a speech signal. TTS essentially consists of three stages, illustrated in figure 1: 1) text analysis; 2) regression; and 3) waveform generation [3]. Text analysis, also known as "frontend", is responsible for processing text inputs, and extracting the corresponding linguistic representations. The regression stage performs linguistic to acoustic feature mapping. Finally, the waveform generation stage produces a speech signal from the acoustic features previously generated. This stage defines the synthesis technique employed by the system (TTS techniques are described in the following paragraphs). From all the types of acoustic features that exist, Mel-spectrograms are the preferred one for present TTS systems.

Fig. 1. The stages in a TTS pipeline. Adapted from [3].

The main TTS synthesis techniques developed in the past are the following: articulatory synthesis, formant synthesis, concatenative synthesis, and statistical parametric synthesis (SPSS). More recently, Deep Learning (DL) has also made a profound impact on speech synthesis, being the current state-of-the-art approach.

Contrarily to previous approaches, DL synthesis techniques process large amounts of data, allowing to extract more intricate features from raw inputs. This is particularly useful to tackle the limitations of the previous models [4], such as the lack of naturalness in speech produced by conventional SPSS systems. DL synthesis is mostly based on DNNs, CNNs, and sequence-to-sequence (Seq2Seq) neural networks.

In the scope of TTS, Seq2Seq neural networks (also known as encoder-decoder neural networks) are currently one the most effective approaches for linguistic to acoustic feature sequence mapping. Based on recurrent mechanisms, these networks suit well the sequential nature of speech signals, by converting variable-length inputs into fixed-length outputs, while retaining the meaningful temporal dependencies [5]. Besides TTS, Seq2Seq networks have been used for other tasks involving sequential data, such as machine translation and speech recognition [4]. In DL synthesis, the TTS pipeline stages comprise two essential blocks: 1) a Seq2Seq system, which implements the text analysis[1] and regression stages; and 2) a neural vocoder for waveform generation.

C. Evaluation metrics

1) AB/ABX preference test: Preference tests are frequently used to assess speech synthesis systems. In an AB preference test, as the name states, listeners are presented with two speech samples and are asked to select their preferred one according to a specific property, such as naturalness or similarity. A "no-preference" answer slot may be included. ABX tests differ from the traditional AB test with the inclusion of an "X" speech sample to be used as reference. In this case, samples "A" and "B" are evaluated using "X" as reference.

2) MUSHRA: In the scope of Speech Synthesis, the MUlti Stimuli with Hidden Reference and Anchor (MUSHRA) enables the subjective assessment of synthesized utterances and is mentioned in several studies, such as [6] and [7]. According

[1]Text preprocessing is excluded from this stage, as it is performed beforehand. Only the extraction of linguistic features from raw text is considered.

to this method, listeners rate audio samples, together with a low-quality anchor, and a hidden reference sample, in comparison to a high-quality reference sample. The low-quality anchor corresponds to a low-pass filtered sample, and its purpose is to ensure minor artifacts are not improperly penalized. Samples are rated regarding similarity or perceived quality on a scale of 0 to 100, where 0 and 100 are the worst and best scores, respectively [8].

D. Universal vocoding

Universal vocoders aim to improve the generalization capabilities of neural vocoders. The dependency on extensive datasets and computational power motivated the search for new solutions that could incorporate new speaker-styles, without further training [9].

Lorenzo-Trueba and colleagues (2019) proposed a Speaker Independent universal vocoder based on WaveRNN [10], capable of generalizing to unseen speakers. Authors considered four different training settings for the universal vocoder: a single-speaker configuration, two multi-speaker configurations, consisting of three and seven speakers respectively, and a universal vocoding configuration, comprising 74 speakers and 17 languages. The multi-speaker configurations aimed to assess how reducing the dataset size (number of utterances) and increasing the number of speakers would influence the output quality.

Evaluation and experiments included several scenarios: 1) in-domain speakers and speaking style; 2) out-of-domain speakers but similar speaking style; 3) out-of-domain speakers and speaking style. Additionally, various unseen scenarios were tested. MUSHRA tests were used to evaluate each case [9].

No significant difference was registered for scenario 1) since all training settings produced similar results at inference time. The universal vocoder setting registered a 98.5% relative[2] MUSHRA score. For scenario 2), results have shown that the more speakers are included during training, the better is the output quality. Although the 3-speaker setting comprised more data than the 7-speaker setting, the output quality was better for the latter, which indicates that speaker variability is more important than quantity of data regarding the universal vocoding task [9]. For scenario 3), the universal vocoder setting still provided a stable output, reaching a 98% relative MUSHRA score. Single-speaker and 7-speaker settings revealed poor results, unlike the remaining (3-speaker and universal vocoder). This contrast was mainly due to speaker dissimilarity among train and test speakers. The authors used Kullback-Leibler divergence (KLD) to measure speaker similarity. KLD was measured between the Gaussian Mixture Models of the training data of each vocoding approach and the speaker. The KLD between the test speaker and each one of the settings was 2.64 for the universal vocoder, 5.42 for 3-speaker, 14.45 for 7-speaker, and 14.62 for single-speaker, proving that dissimilarity between train and test speakers severely degraded the output quality for unseen speaking style scenarios.

Regarding robustness to voice quality, the universal vocoder still generalized well, achieving 91.6% and 89.5% relative MUSHRA scores for breathy and pressed voices, respectively. In terms of signal quality, the model's performance further worsened, achieving 79.4% and 76.4% relative MUSHRA scores for noisy, and reverberating signals, respectively. The most significant drop occurs for simultaneously noisy and reverberating signals, for which the relative MUSHRA score drops to 57.8%.

E. Multi-speaker Seq2Seq regressive model

Zhang and co-authors (2019) proposed a non-parallel Seq2Seq voice conversion model, that operates in one of two configurations depending on the type of input: voice conversion (VC), or TTS. For the VC configuration, the acoustic feature sequence (in the form of a Mel-spectrogram) is extracted from a source utterance and is fed to the Seq2Seq regressive model. The VC configuration follows the traditional framework, which takes a source utterance as input. This setting preserves the input's linguistic content and embeds the target speaker's identity into the output. For the TTS configuration, text inputs are converted to phonetic transcriptions before being fed to the model. The TTS process unfolds similarly to the VC procedure since the output is also a combination of linguistic content and speaker identity. The most notable difference is the textual input, as opposed to a source utterance. Since TTS is the main focus of this study, we will focus on the system's TTS configuration. The model incorporates 5 components: a text encoder $\mathbf{E}^t$, a speaker encoder $\mathbf{E}^s$, a Seq2Seq decoder $\mathbf{D}^a$, a recognition encoder $\mathbf{E}^r$, and an auxiliary classifier $\mathbf{C}^s$. The text encoder extracts linguistic embeddings H^t from phoneme transcriptions T of input text sequences. The speaker encoder takes spectrograms as input and generates speaker embeddings h^s, capable of identifying a speaker. The decoder generates an acoustic feature sequence $\hat{A}$ by combining previously extracted linguistic and speaker embeddings. Its structure is analogous to Tacotron [11], [12]. Authors used a WaveNet vocoder [13] to recover speech waveforms from acoustic features [14]. The recognition encoder $\mathbf{E}^r$ and the auxiliary classifier $\mathbf{C}^s$ are only employed during the training process, thus are not used by the TTS configuration at inference time. The recognition encoder extracts linguistic representations from audio signals. For that, it takes acoustic feature sequences A as input, and and outputs a linguistic embedding H^r. Since H^t and H^r are expected to be similar, a contrastive loss is introduced to increase similarity between both linguistic representations. The auxiliary classifier predicts the speaker identity from a linguistic H^r. It is used for adversarial training to remove the remaining speaker-related content within the linguistic embedding [14].

Training the model consists of two stages: pre-training, and fine-tuning. In the pre-training phase, the system is trained with a large multi-speaker dataset, comprising utterances, text transcriptions, and the corresponding speaker identity tag.

[2]The relative MUSHRA score is the ratio between mean MUSHRA scores of a system, and of natural speech.

Fine-tuning was performed in a 2-speaker setting, nevertheless, it is possible to fine-tune the model to more than two speakers. The fine-tuning stage introduces unseen speakers during pre-training and converges faster than the first stage. The authors evaluated the effect of training data reduction on the model's performance at the fine-tuning stage. For this, the number of training utterances per speaker was gradually reduced from 500 to 100. Results have shown that the model has similar performance regardless of the amount of training data, suggesting that this implementation is suitable for scenarios where the amount of data is scarce [14].

III. Proposed Model

A. Model overview

The present section describes the proposed multi-speaker TTS system. This approach (depicted in figure 2) follows the prevailing state-of-the-art premise, combining a Seq2Seq system for acoustic feature prediction, with a neural vocoder for speech waveform recovery. The models detailed in previous

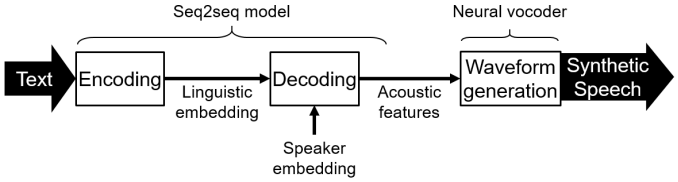


Fig. 2. Inference-time pipeline of the proposed model.

sections II-E and II-D were employed for acoustic feature prediction, and waveform recovery, respectively.

The WaveRNN-based universal vocoding architecture allowed for faster synthesis (as opposed to WaveNet), and more importantly, for better generalization to new speaker identities. This aspect was crucial to prevent training the neural vocoder at the speaker adaptation stage, thereby simplifying that process. The Seq2Seq system was adapted from the non-parallel Seq2Seq VC implementation [14], given its similarities with multi-speaker TTS. The Seq2Seq VC model was originally designed for the English language (EN), therefore, several adjustments were introduced to adapt this system to European Portuguese (EP). These adaptations consisted in text and audio preprocessing stages (described in sections III-C and III-D, respectively). The text encoder of this model relies on phonetic transcriptions, which require changing the grapheme-to-phoneme (GtoP) conversion stage from EN to EP.

B. Spoken language corpora

Selecting suitable voice banks to train the two main blocks of the proposed model (Seq2Seq model + neural vocoder) was an important step during implementation because it greatly influenced synthetic speech quality. Table I indicates the specifications of the EP corpora employed for pre-train and fine-tune stages.

*Corresponds to the number of speakers used for fine-tuning. Originally, the corpus contains more speakers.

TABLE I
List of EP voice corpora used for pre-train and fine-tune stages.

Specification		Pre-train BD-PUBLICO	Fine-tune BDFALA
Sampling frequency [kHz]		16	16
Number of speakers		100	2*
Utterances per speaker		79-83	300
Total duration [hh:mm:ss]		21:48:39	00:51:36
Utterance duration [s]	Mean	9.7	5.2
	Median	9.5	5.1

The neural vocoder, which did not require fine-tuning, was only trained with the BD-PUBLICO corpus.

C. Text preprocessing

Regarding text preprocessing, three different GtoP approaches were employed: 1) the Phonemizer [15] using the eSpeak backend; 2) a Sequence-to-Sequence GtoP toolkit, developed by CMUSphinx; and 3) a Festival-based approach, which is employed in DIXI+ [16]. Besides assessing the ability to generate correct phonetic transcriptions, other factors were considered in order to choose which approach is best, in particular, text normalization, and the ability to correctly transcribe homographs and acronyms. Text normalization is an important step in GtoP methods because it allows to obtain correct phonetic transcriptions for non-standard text, such as abbreviations or digits.

1) Punctuation: Although punctuation marks are not phonetic symbols, the main reason for including punctuation in phonetic transcriptions is for the model to learn pauses correctly, which are most frequently represented by commas or full-stops. From all punctuation marks, only commas, full stops, and question marks were included in phonetic transcriptions. Including all possible punctuation would contribute to unnecessarily large phoneme lists, that is, an excessively large number of different phonemes in the dataset. A phoneme list with many instances would not only hinder the training task, but could potentially cause out-of-memory (OOM) errors, since the number of instances in the phoneme list corresponds to the output size of one of the layers employed in the decoding process. In total, BD-PUBLICO comprises 13 002 commas, 273 colons and 38 semicolons. Given that commas are substantially more frequent, colons and semicolons were replaced with commas. Likewise, ellipses were replaced with full stops. Given that the sentences of BD-PUBLICO were all gathered from newspaper text, the number of exclamatory and interrogative sentences is very scarce. As such, exclamatory sentences were assumed as declarative ones, meaning that exclamation marks were replaced with full stops at the end of sentences.

2) Phonemizer with eSpeak backend: The open-source repository used for the Seq2seq regressive model already employs a GtoP toolkit, the Phonemizer [15], which supports different languages and frameworks, namely Festival and eSpeak. The rule-based eSpeak backend was used since it was

the only one available for EP in this toolkit. Unlike the Festival backend, eSpeak employs the International Phonetic Alphabet (IPA) to represent phonemes. In comparison to other phonetic notations, such as SAMPA, IPA is more complex, comprising over 100 different symbols representing vowels and consonants only. Hence, using the eSpeak backend generated very large phoneme lists. For EP, this backend generated a list comprising 73 symbols, excluding punctuation. To overcome the problem of extensive phoneme lists, 22 symbols were merged, and one symbol was excluded.

3) Seq2Seq GtoP toolkit: This GtoP approach relies on the Transformer architecture [17], thus only using attention mechanisms to establish the mapping between text inputs and phonetic sequence outputs. The training process of the model relies on a dictionary of words and their phonetic transcriptions. The model for EP was trained with a dictionary comprising 60 700 entries . Phonetic transcriptions followed the SAMPA notation, including primary stress. With this GtoP approach, the phoneme list comprised 50 different symbols, excluding punctuation. Despite being easily trainable and efficient with individual words, this toolkit had several limitations. Words were processed separately, therefore, sentences were incorrectly converted at word boundaries. To mitigate this issue, the most common sandhi rules were manually implemented on top of the output phonetic sequences. Also, the model did not distinguish homographs, neither included any text normalization stage.

4) Festival-based GtoP: This technique follows the Festival framework, relying on a set of classification trees for GtoP conversion [16]. Phonetic transcriptions followed the SAMPA notation, and additional symbols were employed to represent foreign phones — the symbol "H", for example, was used to represent the sound of letter H in foreign words like "**H**ollywood", "Man**h**attan" or "**H**ezbollah"). In total, the phoneme list obtained with this approach was composed of 44 symbols excluding punctuation. This approach did not discern between different punctuation signs, hence converting every sign to the symbol "#" in phonetic transcriptions. To include punctuation in the phonetic transcriptions of each sentence, the signs and their order of appearance were saved to a list. Then, each sign in the list replaced each occurrence of the symbol "#" in the phonetic transcription in the same order.

From the three GtoP approaches that were experimented, the Festival-based is which deals best with non-standard words. The existence of an addenda allowed to specify the pronunciation of such words and override the output the system would normally produce [16]. Regarding homographs, upon phonetic symbol prediction, PoS tagging is employed both in the eSpeak backend as well as in the Festival-based approach. The usage of PoS tagging allows to correctly predict most of homograph cases. The only type of homographs the GtoP modules fail to transcribe are homographs with the same PoS category (e.g. *Tenho s**e**de. Hoje estive na s**e**de.*). When the PoS categories differ (e.g. verb/noun as in *Eu alm**o**ço depois. Está na hora do alm**o**ço.*), both modules produce the correct transcription.

D. Audio preprocessing

Before Mel-spectrogram extraction, audio signals were normalized and underwent a pre-emphasis filter. Mel-spectrogram extraction unfolded in two steps: 1) computing the magnitude of the short-time Fourier transform (STFT) of the audio signal, which corresponds to a linear spectrogram; 2) converting the linear spectrogram to a Mel-spectrogram. The following STFT parameters were employed: a 50 ms window length, 12.5 ms window shift, and 2048-point Fourier transform, as in the Tacotron original implementation [11]. Mel-spectrograms comprised 80 Mel-bands, a minimum frequency of 50 Hz, and a default maximum frequency of 8 kHz, that is, half of the sampling frequency.

1) Leading and trailing silence removal: To ease convergence during the training process, leading and trailing silence was removed from audio signals. This procedure was carried out with a toolkit for speech diarisation [18]. Voice activity detection (VAD) was employed solely to capture the audio samples at which speech started and ended, and audio files were trimmed accordingly.

E. Model training

1) Seq2Seq regressive model — pre-train: The Adam optimizer [19] was employed during training. L2 regularization with weight 10^{-6} was used. The model was trained for 8250 iterations (approximately 58 epochs), at a constant learning rate (10^{-3}), with a batch size of 32. The training process was assessed based on encoder-decoder alignments plots and Mel-spectrogram visualization.

2) Neural vocoder: The model was trained for 100 000 iterations, starting at a learning rate of 4×10^{-4} that decayed by 50% every 20 000 iterations. As the training process unfolded, synthesis quality was perceptually evaluated from generated audio samples.

IV. Experiments

A. Experimental setup

The main reason for employing a universal vocoder is to reduce the amount of training stages during the speaker adaptation stage. As such, speaker adaptation only involves fine-tuning the Seq2Seq regressive model, instead of the whole system. The model was fine-tuned for three distinct settings: 1) two adult speakers; 2) two adolescent speakers; and 3) two child speakers. For each setting, the pair of speakers comprised one male and one female voice. All configurations employed non-parallel data.

The adult–speaker fine-tune setting is the standard implementation of the proposed model, since the characteristics of data (namely the prosody, and type of sentences) and the speakers' traits (namely pitch values) are in line with the speech corpus employed during the pre-train stage. Fine-tuning employed 300 utterances, of which 33 for validation, for each each speaker *sp_01* and *sp_02* of the BDFALA corpus. Synthetic speech was assessed in terms of naturalness, voice similarity, and intelligibility for both speakers. Additionally, the same setting with reduced data — 20 utterances

per speaker, of which four for validation — was tested for naturalness and similarity.

The fine-tune configurations for child and adolescent voices were addressed separately and did not take part in subjective assessments, given that synthetic speech was perceptually worse than for the adult–speaker setting. Characteristics of both the data and the speakers were substantially different from those used in pre-training, hence, these configurations were analyzed as out-of-domain (OOD) scenarios (see Section IV-B). In each of these settings, the amount of fine-tuning data was considerably lower than in the standard adult–speaker setting.

TABLE II
AMOUNT OF FINE-TUNING DATA PER CONFIGURATION.

Configuration	Speaker	# Utterances/files	Total Duration [hh:mm:ss]
Adults-standard	*sp_01*	300	00:51:36
	sp_02	300	
Adolescents	*sp_03*	34	00:03:10
	sp_04	33	
Children	*sp_11*	19	00:02:44
	sp_36	21	
Adults-reduced	*sp_01*	20	00:03:18
	sp_02	20	

B. Differences in training data: pitch and prosody

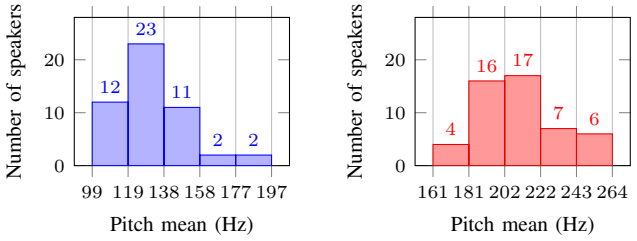
The most notable differences among the data of each fine-tuning configuration in comparison to pre-training data are the mean pitch values of the speakers and the prosodic contours of the utterances they recorded. The differences in pitch were, unsurprisingly, most prominent in the voices of children, since these had the highest pitch. Prosodic contours were analyzed according to two factors: 1) the standard deviation (SD) of pitch for each speaker; and 2) the type of sentences — declarative, interrogative, or exclamatory — employed in each training setting. Table III specifies the pitch values of pre-train speakers and each fine-tune speaker.

TABLE III
PITCH VALUES FOR EP SPEAKERS IN DIFFERENT AGE GROUPS.

Speech corpus	Speaker	Age	Pitch [Hz]	
			Mean	SD
BD-PUBLICO	*o000-o049* (M)	19-28	99-197	6-39
	o050-o099 (F)		161-264	15-38
BDFALA	*sp_02* (M)	35	114	15
	sp_01 (F)	41	200	33
BDFALA02	*sp_03* (M)	14	165	38
	sp_04 (F)	13	254	49
EUROM.1	*sp_11* (M)	10	291	52
	sp_36 (F)	9	272	56

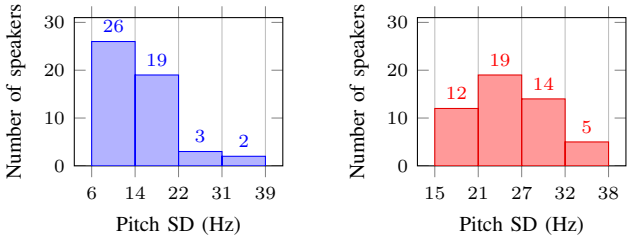
Regarding mean pitch values, both children (*sp_11* and *sp_36*) standout with higher values than the remaining fine-tune speakers, and more importantly, pre-train speakers. Children voices are clearly an OOD scenario in terms of mean pitch as these yield a mean pitch outside the range of values of pre-train data.

Histograms displayed in figure 3 complement the data of table III by illustrating how mean pitch values are distributed across the pre-train corpus. It is clear that in terms of mean pitch, adolescent speakers (*sp_03* and *sp_04*) only were similar to a small minority of voices employed during pre-training. Likewise, histograms in figure 4 specify the SD values across the BD-PUBLICO corpus to complement table III. Only two male and five female speakers from BD-PUBLICO achieved a pitch SD higher than 31 Hz and 32 Hz, respectively.



(a) Male speakers (b) Female speakers

Fig. 3. Histograms of pitch mean values of speakers in the BD-PUBLICO corpus.



(a) Male speakers (b) Female speakers

Fig. 4. Histograms of pitch SD values of all speakers in the BD-PUBLICO corpus.

Table IV specifies the number of sentences of each type in the datasets employed in each training configuration. In terms of sentence type, the adolescent–speaker fine-tune setting differs the most from the pre-train configuration, since over 80% of the sentences are either exclamatory or interrogative. Given these discrepancies, the adolescent–speaker fine-tune setting was also considered as an OOD scenario. Although child and adult fine-tune settings employed similar data in terms of sentence type, the former setting presented more expressive prosody as it achieved the highest pitch SD values, as seen in table III. From tables III and IV, one may infer that both the adolescent and child fine-tune configurations comprised expressive prosody as opposed to the remaining. Of all fine-tune configurations, both adult–speaker settings (standard and reduced) are the most similar to pre-train data both in terms of pitch and sentence type.

C. Seq2Seq regressive model — fine-tune

Overall, the fine-tuning process was very similar to the pre-train stage. Fine-tuning contemplated the whole Seq2Seq regressive model except for the speaker encoder, as stated

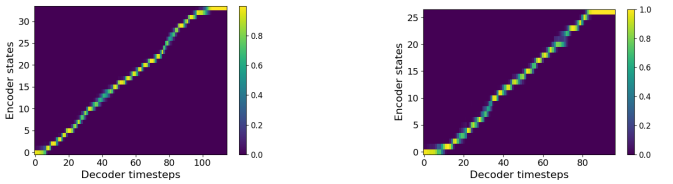
TABLE IV
NUMBER OF DECLARATIVE, INTERROGATIVE AND EXCLAMATORY SENTENCES FOR EACH TRAINING SETTING.

Training setting	# Declarative	# Interrogative	# Exclamatory
Pre-train	8021 (99.16%)	47 (0.58%)	21 (0.26%)
Adults-standard	527 (87.83%)	70 (11.67%)	3 (0.50%)
Adults-reduced	34 (85.00%)	6 (15.00%)	—
Adolescents	22 (19.64%)	25 (22.32%)	65 (58.04%)
Children	34 (85.00%)	6 (15.00%)	—

in the original paper [14]. For the fine-tune stage, the same hyperparameters were used as in the original implementation proposed by the authors. In comparison to the pre-train stage, only the batch size was changed from 32 to 8, the number of training epochs at constant LR was changed from 70 to 7, and weighting factors of speaker encoder and speaker adversarial losses were changed to 0 and 0.2, respectively.

Prior to fine-tuning the model, all data underwent the same text and audio preprocessing steps as before pre-training, except for the silence removal stage, as audio files did not include long sections of leading or trailing silence. During the training process, trainable speaker embeddings were employed for each speaker. Each embedding was initialized with the average of speaker encoder outputs for all training utterances of the speaker in question. At inference, contrarily to the default fine-tune implementation, the proposed model directly employs the average of speaker encoder outputs. For all fine-tune configurations, no difference was perceived in synthetic speech when employing the trained embeddings instead of the average speaker encoder outputs. As such, for the sake of simplicity, we chose to solely use the pre-trained speaker encoder for embedding extraction at fine-tuning. Similarly to the pre-train stage, fine-tuning was assessed based on encoder-decoder alignment plots and Mel-spectrograms. For each voice, the alignments and spectral representations were generated from test sentences (unseen during training). To minimize prosodic differences that otherwise could standout in Mel-spectrograms, we selected declarative sentences without pauses.

1) Adult–speaker configurations: Standard and reduced adult–speaker configurations were trained for 1600 iterations (approximately 23 epochs), and 200 iterations (50 epochs), respectively. Figure 5 illustrates the encoder-decoder alignments obtained for both configurations. For each speaker, the same test sentences were employed in both configurations. For these sentences, despite the difference in amount of fine-tuning data, encoder-decoder alignments remained very similar. Figure 6 illustrates the predicted and original Mel-spectrograms obtained from these sentences. Predicted Mel-spectrograms were similar for both configurations regardless of the amount of training data, suggesting that the Seq2Seq regressive model can produce reasonable results in a reduced data setting.



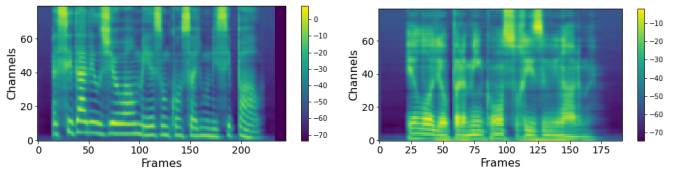
(a) Standard setting, speaker *sp_01*

(b) Standard setting, speaker *sp_02*

(c) Reduced setting, speaker *sp_01*

(d) Reduced setting, speaker *sp_02*

Fig. 5. Encoder-decoder alignments for fine-tuning test sentences — adult speakers.



(a) Predicted Mel-spectrogram for *sp_01*, standard setting

(b) Predicted Mel-spectrogram for *sp_02*, standard setting

(c) Predicted Mel-spectrogram for *sp_01*, reduced setting

(d) Predicted Mel-spectrogram for *sp_02*, reduced setting

(e) Original Mel-spectrogram for *sp_01*

(f) Original Mel-spectrogram for *sp_02*

Fig. 6. Mel-spectrograms for fine-tuning test sentences — adult speakers.

D. Speaker identity discrimination

To demonstrate that the model can effectively discriminate among different voices, we extracted speaker embeddings from different speakers and analyzed these with t-SNE, a technique that allows to visualize high–dimensional data in a two or three–dimensional space [20].

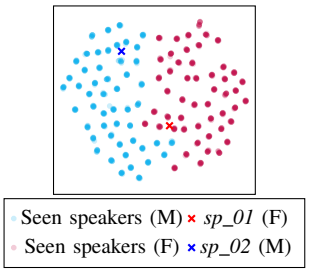
When visualizing a wide range of speaker identities, illustrated in figures 7(a), (b), and (c), one can notice that the placement of the embeddings relatively to one another allows to identify the speakers' gender. Furthermore, the embeddings of fine-tune speakers are also represented.

Pitch is known to be one of the most common features in speaker identification tasks. In fact, the presence of pitch–related information is noticeable in the speaker embeddings employed in the proposed model, since it is possible to correlate the relative location of an embedding in the plot, with the mean pitch of the speaker associated to that embedding.

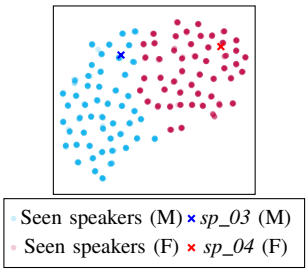
1) Adult speakers — sp_01 and sp_02: Besides all the embeddings of pre-train speakers, figure 7(a) depicts the embeddings of adult fine-tune speakers. From a gender viewpoint, the fine-tune embeddings are placed in accordance with the pre-train speakers' embedding distribution. In comparison to the range of pitch values of pre-train speakers, the male speaker *sp_02* registered a relatively low mean pitch value — 114 Hz — quite far from the lowest mean pitch registered for pre-train female speakers — 161 Hz. Female speaker *sp_01* displayed a mean pitch of 200 Hz which is close to the highest value recorded for pre-train male speakers — 197 Hz. This suggests that the embedding of *sp_01* is closer to male–speaker embeddings than the embedding of *sp_02* is to female–speaker embeddings, and can be observed in figure 7(a).

2) Adolescent speakers — sp_03 and sp_04: The embeddings of adolescent fine-tune speakers are illustrated in figure 7(b). Regarding gender, fine-tune embeddings are placed in conformity with the distribution of pre-train speakers. Both fine-tune speakers registered relatively large values of mean pitch in comparison to pre-train speakers of the same gender. Male speaker *sp_03* recorded a mean pitch of 165 Hz, among the highest for male pre-train speakers. Moreover, this value is close to the lowest pitch value for pre-train female speakers. This suggests that the embedding of *sp_03* is one of the closest to embeddings of female speakers. On the other hand, given that female speaker *sp_04* recorded a mean pitch of 254 Hz, which is one of the highest among female pre-train speakers, it is expected that the corresponding embedding is one of the farthest from male–speaker embeddings. As seen in the figure, the embedding of speaker *sp_03* is significantly closer to female speaker embeddings than the embedding of speaker *sp_04* is to male speaker embeddings.

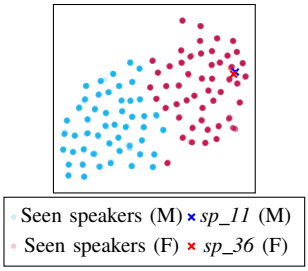
3) Child speakers — sp_11 and sp_36: Embeddings of child fine-tune speakers are depicted in figure 7(c). Both child voices registered the highest mean pitch values among all speakers. The placement of these embeddings relative to pre-train embeddings suggests that both are female speakers, even though only speaker *sp_36* is. The misplacement of of the male speaker embedding is due to pitch similarity between the voices of children. Furthermore, the fact that both children displayed a very high mean pitch suggests that their embeddings are among the farthest from male–speaker embeddings, which in fact occurs.



(a) Pre-train and adult fine-tune speakers



(b) Pre-train and adolescent fine-tune speakers



(c) Pre-train and child fine-tune speakers

Fig. 7. Visualization of speaker embeddings.

V. Evaluation

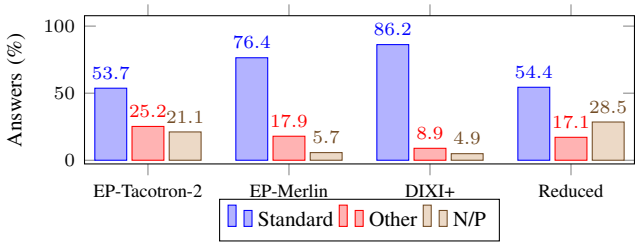
This section focuses on assessing the performance of the proposed model in comparison to other systems. Performance tests were divided into three groups: 1) naturalness and similarity; 2) intelligibility; and 3) synthesis speed. The proposed model was tested for the adult–speaker standard configuration. Each test is described in the subsections that follow.

A. Naturalness and similarity

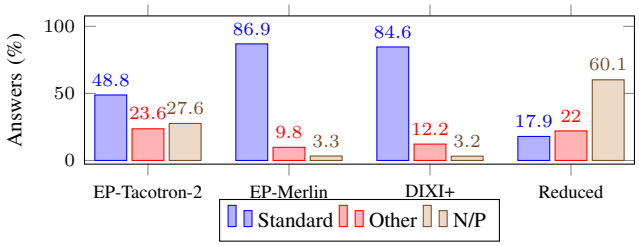
Naturalness and similarity were rated using AB and ABX preference tests, respectively. "A" and "B" refer to synthetic utterances, and X to a sentence uttered by the target speaker, used as reference. Three unseen sentences were randomly selected from the test set of each speaker. Ideally, one should use a larger number of utterances per speaker for testing, but that would have made the tests very lengthy and time-consuming. To provide a context of the model's performance in the scope of TTS, naturalness and similarity tests were extended to four different TTS approaches besides the proposed model: 1) EP-Tacotron-2[3]: a DL-based model inspired by the original Tacotron 2 system; 2) EP-Merlin: an SPSS-based model; 3) DIXI+: a concatenative synthesis system; and 4) the reduced adult–speaker configuration of the proposed

[3]We refer to the Tacotron 2 implementation of a previous Master Thesis as "EP-Tacotron-2". Not to be confused with the original implementation, Tacotron 2.

model. Approaches 1) and 2) were developed within the scope of previous Master Theses [21] and [22], respectively. DIXI+ [16] was not fine-tuned for speakers *sp_01* and *sp_02*. Instead, a readily available configuration was used, which comprised different voices from *sp_01* and *sp_02*. Thus, DIXI+ only took part in naturalness assessments. Results for naturalness and similarity tests are displayed in figures 8 and 9, respectively. In total, 41 listeners participated in these tests. The standard configuration of the proposed model is referred to as "Standard" in the legend of the figures, and "Other" refers to each of the four alternative approaches that were tested. Each approach is specified across the horizontal axis of the bar charts. Regarding the naturalness test, the proposed model produced distinctively better results for both voices than EP-Merlin, and DIXI+, as expected. Surprisingly, the standard



(a) Female voice *sp_01*

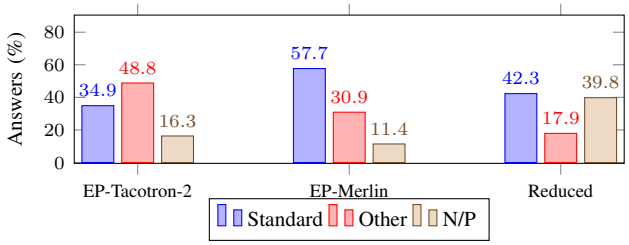


(b) Male voice *sp_02*

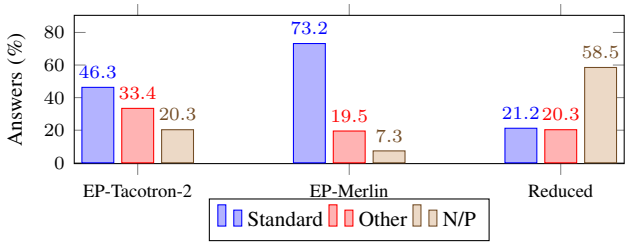
Fig. 8. Naturalness AB preference test results.

configuration of the proposed model was also chosen over the DL-based EP-Tacotron-2 more times, for both voices. One must notice, however, that in the latter case the percentage of no preference (N/P) answers was considerably larger than for EP-Merlin and DIXI+. Regarding similarity, results showed that the proposed model clearly performed better than EP-Merlin for both voices, although not as blatantly as in the naturalness test. This could be due to the fact that EP-Merlin produced samples with noticeably better audio quality than the proposed model. Listeners may have been influenced by voice–unrelated features during assessments. In comparison with EP-Tacotron-2, the results for the female voice *sp_01* were as expected. Regarding the male voice *sp_02*, the proposed model achieved better results, even though over 20% of listeners did not prefer one model over the other.

Synthetic samples generated by EP-Tacotron-2 occasionally contained bursts of subtle noise and/or loudness. The latter phenomenon was particularly noticeable in fricative consonants — synthetic speech samples suddenly sounded louder from a specific timestep onward. Some participants



(a) Female voice *sp_01*



(b) Male voice *sp_02*

Fig. 9. Similarity ABX preference test results.

may have been influenced by these factors when assessing the naturalness and similarity of EP-Tacotron-2 speech samples, thus preferring the proposed model.

In regard to the reduced configuration of the proposed model, results show that synthetic speech quality decreased for the female voice *sp_01* upon fine-tune data reduction. Nevertheless, the high percentage of (N/P) answers, especially in the similarity test, mitigates this phenomenon. For the male voice *sp_02*, both in naturalness and similarity assessments, no noticeable difference was registered in comparison to the standard configuration since the majority of participants expressed no preference. Also, the percentage of listeners that selected one configuration over the other is very balanced for both configurations.

B. *Intelligibility*

Intelligibility was assessed from listeners' textual transcriptions of semantically unpredictable sentences. Participants were asked to transcribe 10 sentences for each test speaker while listening to each sentence only once. The first test sentence of each speaker served as a dry run and was assessed separately. The word error rate (WER) was determined from the transcriptions to measure how accurately the sentences were perceived by the listeners. The test was destined to native or fluent Portuguese speakers. In total, 33 participants took the intelligibility test. Table V displays the WER obtained from the transcriptions of participants. In some sentences several participants misheard specific words, which increased the WER. This was the case for the dry run sentence for speaker *sp_01* — one third of the participants could not perceive the word *colunas* —, hence the abnormally large WER for this sentence. Table VI displays the most frequently misheard words.

TABLE V
INTELLIGIBILITY TEST RESULTS.

Speaker	WER (%)	
	Dry run	Remaining
sp_01 (F)	16.67	3.74
sp_02 (M)	1.01	4.22

TABLE VI
FREQUENTLY MISTAKEN WORDS.

Original word	Incorrect transcription(s)	Total Occurrences
colunas	*runas*, *com umas*	11
invocam	*evocam*	4
abstraída	*distraída*	24
do	*no*	31

C. Synthesis speed

The proposed model's neural vocoder was compared with its counterpart in the EP-Tacotron-2 implementation. For each model, synthetic speech was generated for the female voice *sp_01* from 10 sentences, and synthesis times were compared for each sentence. On average, the proposed model's neural vocoder synthesized speech 11 times faster than its counterpart in EP-Tacotron-2, with an NVIDIA GeForce GTX TITAN X GPU.

VI. CONCLUSIONS

This study proposed a multi-speaker TTS system for EP, based on state-of-the-art Speech Synthesis methodologies. The addition of the Festival–based text preprocessing module ensured coherent phonetic transcriptions, including a wide variety of exceptions. Results show that the proposed model was superior than previous–generation systems in terms on naturalness and similarity. The proposed model achieved results comparable to a different state-of-the-art TTS approach, while employing less than half of the data for fine-tuning. For one of the voices, the reduced and standard settings achieved very similar performance meaning that it is possible to preserve output quality while reducing the amount of training data.

The proposed model proved to be significantly faster than the implementation of EP-Tacotron-2. The WaveRNN–based vocoder employed in the proposed model was, on average, 11 times faster than the WaveNet employed in EP-Tacotron-2, thus being more practical.

REFERENCES

[1] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986. [Online]. Available: https://doi.org/10.1038/323533a0

[2] J. Holmes and W. Holmes, *Speech Synthesis and Recognition*, 2nd ed. USA: Taylor & Francis, Inc., 2002.

[3] S. King, "What is "end-to-end" text-to-speech synthesis?" University of Lancaster, 2019. [Online]. Available: http://media.speech.zone/images/Simon_King_Lancaster_2019_compressed_for_publication.pdf

[4] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, "A Review of Deep Learning Based Speech Synthesis," *Applied Sciences*, vol. 9, no. 19, p. 4050, Jan. 2019, number: 19 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2076-3417/9/19/4050

[5] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3104–3112. [Online]. Available: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[6] J. Latorre, J. Lachowicz, J. Lorenzo-Trueba, T. Merritt, T. Drugman, S. Ronanki, and K. Viacheslav, "Effect of data reduction on sequence-to-sequence neural TTS," *arXiv:1811.06315 [cs, eess]*, Nov. 2018, arXiv: 1811.06315. [Online]. Available: http://arxiv.org/abs/1811.06315

[7] T. Merritt, B. Putrycz, A. Nadolski, T. Ye, D. Korzekwa, W. Dolecki, T. Drugman, V. Klimkov, A. Moinet, A. Breen, R. Kuklinski, N. Strom, and R. Barra-Chicote, "Comprehensive evaluation of statistical speech waveform synthesis," *arXiv:1811.06296 [cs, eess]*, Dec. 2018, arXiv: 1811.06296. [Online]. Available: http://arxiv.org/abs/1811.06296

[8] C. Mendonça and S. Delikaris-Manias, "Statistical tests with MUSHRA data," 2018.

[9] J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, R. Barra-Chicote, A. Moinet, and V. Aggarwal, "Towards achieving robust universal neural vocoding," *arXiv:1811.06292 [cs, eess]*, Jul. 2019, arXiv: 1811.06292. [Online]. Available: http://arxiv.org/abs/1811.06292

[10] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient Neural Audio Synthesis," *arXiv:1802.08435 [cs, eess]*, Jun. 2018, arXiv: 1802.08435 version: 2. [Online]. Available: http://arxiv.org/abs/1802.08435

[11] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," *arXiv:1703.10135 [cs]*, Apr. 2017, arXiv: 1703.10135. [Online]. Available: http://arxiv.org/abs/1703.10135

[12] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," *arXiv:1712.05884 [cs]*, Feb. 2018, arXiv: 1712.05884. [Online]. Available: http://arxiv.org/abs/1712.05884

[13] A. Polyak and L. Wolf, "Attention-based Wavenet Autoencoder for Universal Voice Conversion," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom: IEEE, May 2019, pp. 6800–6804. [Online]. Available: https://ieeexplore.ieee.org/document/8682589/

[14] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-Parallel Sequence-to-Sequence Voice Conversion with Disentangled Linguistic and Speaker Representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 540–552, 2020, arXiv: 1906.10508. [Online]. Available: http://arxiv.org/abs/1906.10508

[15] M. Bernard, hadware, R. Riad, A. Greyber, J. Benjumea, Isn0gud, and S. Li, "bootphon/phonemizer: phonemizer-2.2.1," Jul. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3959035

[16] S. Paulo, L. C. Oliveira, C. Mendes, L. Figueira, R. Cassaca, C. Viana, and H. Moniz, "DIXI – A Generic Text-to-Speech System for European Portuguese," in *Computational Processing of the Portuguese Language*, A. Teixeira, V. L. S. de Lima, L. C. de Oliveira, and P. Quaresma, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, vol. 5190, pp. 91–100, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-540-85980-2_10

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv:1706.03762 [cs]*, Dec. 2017, arXiv: 1706.03762. [Online]. Available: http://arxiv.org/abs/1706.03762

[18] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, 2020.

[19] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Jan. 2017, arXiv: 1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[20] L. V. D. Maaten and G. E. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[21] S. Quintas, "Portuguese Speech Synthesis for Amyotrophic Lateral

Sclerosis," Master's thesis, Instituto Superior Técnico - Universidade de Lisboa, 2019.

[22] A. C. R. Gonçalves, "Text-to-Speech Synthesis in European Portuguese using Deep Learning," Master's thesis, Instituto Superior Técnico - Universidade de Lisboa, 2018.