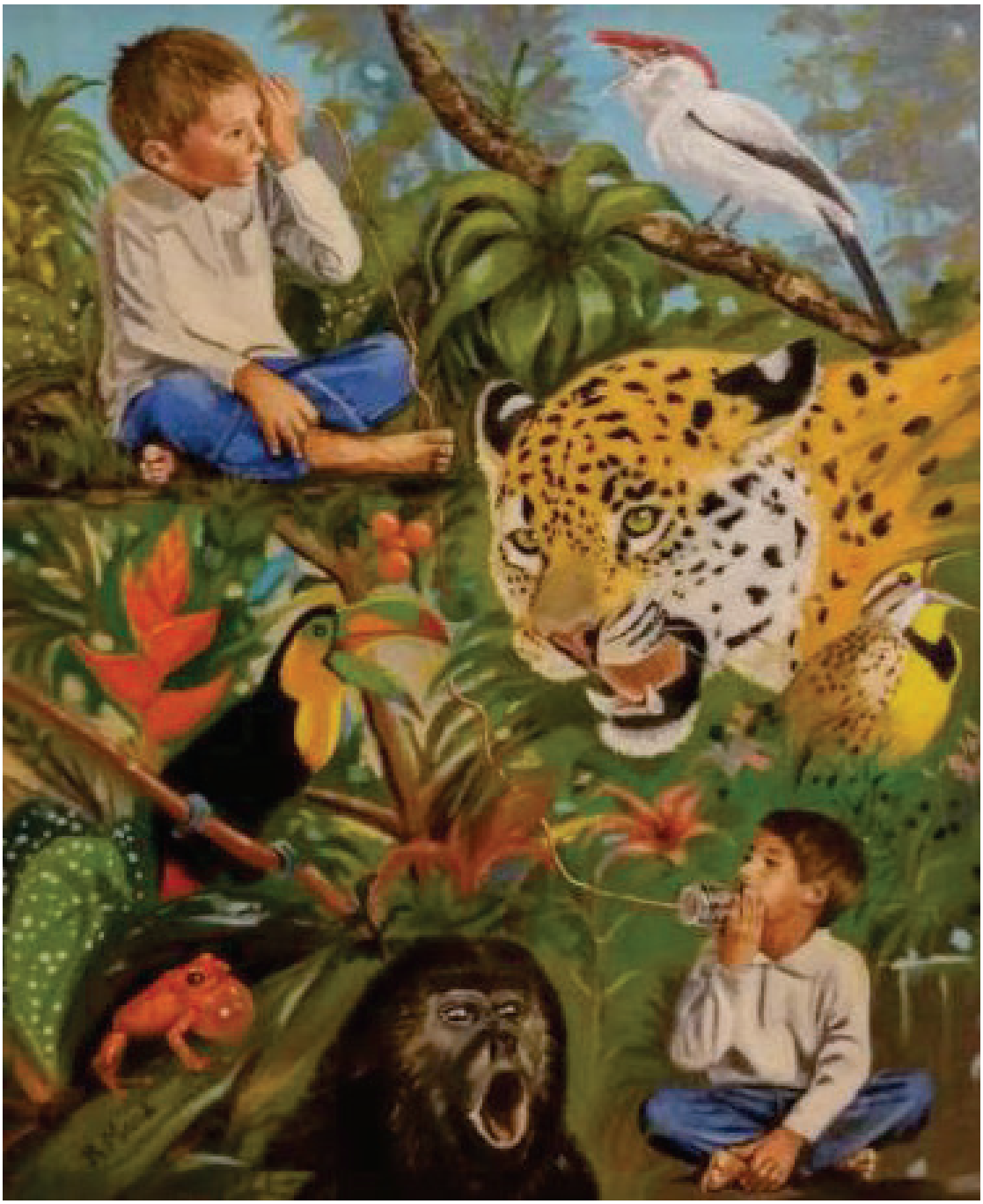
ACOUSTIC COMMUNICATION:
AN INTERDISCIPLINARY APPROACH



Organized by
Emma Otta & Patrícia Ferreira Monticelli
Universidade de São Paulo
Pró-Reitoria de Pesquisa da USP

DOI: 10.11606/9786587596198

Organizing Committee of the ACOUSTIC COMMUNICATION: AN INTERDISCIPLINARY APPROACH

Profa. Dra. Emma Otta (IPUSP)

Profa. Dra. Patrícia Ferreira Monticelli (FFCLRP)

Prof. Dr. Claudio Possani (IME-USP)

Dra. Tania Kiehl Lucci (IPUSP)

Dr. Ricardo Prist (IPUSP)

Cover Photo: Regina Macedo

Book formatting: Dra. Lilian Cristina Luchesi and Dra Aline Domingues Carneiro Gasco

English Editing Services: Michael Germain and Lisa Burger, MC TRADUÇÕES S/S LTDA

Catalogação na publicação
Serviço de Biblioteca e Documentação
Instituto de Psicologia da Universidade de São Paulo

Acoustic communication: an interdisciplinary approach / Organized por Emma Otta e Patrícia Ferreira Monticelli. -- São Paulo, Instituto de Psicologia da Universidade de São Paulo, 2021.

210 p.

E-book.

ISBN: 978-65-87596-19-8

DOI: 10.11606/9786587596198

1. Animal vocalization 2. Animal communication 3. Ethology I. Title

QL765

Ficha elaborada por: Elaine Cristina Domingues CRB5984/08

Funding Acknowledgements

Contents

Dedication

This book is dedicated to the memory of Professor César Ades who passed away on March 14, 2012. César was Professor of the Postgraduate Program of Experimental Psychology at the Institute of Psychology, University of São Paulo. He started the study of sound communication as part of animal behavior, from the perspective of ethology at the Department of Experimental Psychology. An analysis based on the Fonoteca Cesar Ades (FOCA) is presented in chapter 2. The author, Patricia Monticelli, did her master's and doctorate under his supervision studying the vocal repertoire of *Cavia aperea* and *Cavia porcellus*. The book is also dedicated to the memory of Edila Aparecida de Souza who worked for 23 years in the Ethology Lab. She was a motivated professional and gave her best to our University. Edila was one of thousands of people who have died from coronavirus. She passed away on June 3rd, 2020, at age 62. We remember her positive outlook on life, spontaneity, and willingness to work for a common goal and will continue working with this same attitude, in the face of enormous challenges. The Covid-19 pandemic has turned the world upside down. Vaccines are in development thanks to the efforts of scientists around the world. Science gives us hope for the future in our turbulent world.

Emma Otta and Patrícia Ferreira Monticelli

Acknowledgments

The organizers of ACOUSTIC COMMUNICATION: AN INTERDISCIPLINARY APPROACH are grateful to Professor Sylvio Canuto, Dean of Research at the University of São Paulo for supporting the online event, the starting point for the book. We would also like to express our sincerest gratitude to the Organizing Committee, Professor Claudio Possani, Ricardo Prist, Ph.D., Tania Kiehl Lucci, Ph.D., and doctoral students Vinicius Frayze David and Bruna Campos Paula, in addition to all the contributors. Our colleague Regina Macedo, from the University of Brasília, painted *All the Voice*s [*Todas as Vozes*], which we also call *Brazilian Polyphony* [Polifonia Brasiliana] and became the cover of our book. At the closing session of the online session, Professor Regis Rossi Faria, colleague from USP, delighted the participants with SAPOS (2014, 6'30"), a guided auditory tour to an amphibious environment in concert, performed live. We are thankful for this ending experience. We thank the IPUSP Publishing Center [Núcleo de Publicações do Instituto de Psicologia da USP] for their help in editing this book.

About the Contributors

Aline D. Carneiro Gasco

Ph.D. in Science, honorary research fellow of the Ethology and Bioacoustics Laboratory (EBAC), Department of Psychology, FFCLRP, University of São Paulo, Ribeirão Preto, SP, Brazil.

Arnaldo Candido-Jr

Professor in the Department of Computer Science and Computational Mathematics, São Carlos Institute of Mathematical and Computer Sciences (ICMC), São Carlos, SP, Brazil.

Bruna Campos Paula

Ph.D. student at the Laboratory of Acoustics and Environment, Department of Mechanical Engineering, Polytechnic School of the University of São Paulo, São Paulo. Collaborating researcher of the Ethology and Bioacoustics Laboratory (EBAC), FFCLRP, University of São Paulo, Ribeirão Preto, SP, Brazil.

Bruna Lima Ferreira

Master Dissertation student of the Ethology and Bioacoustic Lab. (EBAC). Department of Psychology, FFCLRP, University of São Paulo, Ribeirão Preto, SP, Brazil.

Claudio Possani

Senior Professor at the Department of Mathematics, Mathematics and Statistics Institute (IME) of the University of São Paulo, São Paulo, SP, Brazil.

Daniel Bowling

Ph.D. Instructor at the Social Neurosciences Research Program, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, USA.

Domingos Hiroshi Tsuji

Professor at the Department of Otolaryngology of the Medical School Clinical Hospital (FMUSP/HC) of the University of São Paulo, São Paulo, SP, Brazil.

Edresson Casanova

Ph.D. student at the Department of Computer Science and Computational Mathematics, São Carlos Institute of Mathematical and Computer Sciences (ICMC), São Carlos, SP, Brazil.

Elodie Floriane Mandel-Briefer

Professor at the Behavioural Ecology group, Department of Biology, University of Copenhagen, Copenhagen, Denmark.

Emma Otta

Professor at the Department of Experimental Psychology, Institute of Psychology (IPUSP), of the University of São Paulo, São Paulo, SP, Brazil.

Gabriel Francescoli

Professor at the Ethology Section, Faculty of Sciences, University of the Republic, Montevideo, Uruguay.

Lilian Cristina Luchesi

Ph.D. in Science, honorary research fellow of the Ethology and Bioacoustics Laboratory (EBAC), Department of Psychology, FFCLRP, University of São Paulo, Ribeirão Preto, SP, Brazil.

Marcelo Finger

Professor at the Department of Computer Science, Mathematics and Statistics Institute (IME) of the University of São Paulo, São Paulo, SP, Brazil.

Maria Luisa da Silva

Professor of the Institute of Biological Sciences (ICB), Laboratório de Ornitologia e Bioacústica. Federal University of Pará (UFPA), Belém, PA, Brazil.

Patrícia Ferreira Monticelli

Professor at the Department of Psychology (FFCLRP), head of the Ethology and Bioacoustics Laboratory (EBAC), University of São Paulo, Ribeirão Preto, SP, Brazil.

Paulo do Canto Hubert Junior

Instructor at the School of Business Administration, Fundação Getúlio Vargas (FGV), São Paulo, SP, Brazil

Plínio Almeida Barbosa

Professor at the Department of Linguistics, Institute for Language Studies, University of Campinas, Campinas, SP, Brazil.

Regis Rossi Faria

Professor at the School of Arts, Sciences and Humanities (EACH), of the University of São Paulo, São Paulo, SP, Brazil.

Rogério Grassetto Teixeira da Cunha

Professor at the Institute of Natural Sciences (ICN), Federal University of Alfenas (UNIFAL), Alfenas, MG, Brazil.

Silvia Orlandi

Post-Doctoral Fellow at the Bloorview Research Institute, Holland Bloorview Kids Rehabilitation Hospital, Toronto, Canada.

Sylvia Corte

Professor at the Ethology Section, Faculty of Sciences, University of the Republic, Montevideo, Uruguay.

Vinicius Frayze David

Researcher at the Department of Experimental Psychology, Institute of Psychology (IPUSP), of the University of São Paulo, São Paulo, SP, Brazil.

About the book

The chapters in this eBook provide an overview of the scientific topics discussed at the online scientific meeting that inspired the book. The meeting *Acoustic Communication: An Interdisciplinary Approach* took place online on November 19-20, 2020, with the support of the Dean's Office for Research of the University of São Paulo, Brazil. We were experiencing a Covid-19 pandemic declared by the World Health Organization on March 11, 2020. Quarantine was declared in the State of São Paulo on March 13, 2020. In this context, we stress the importance of creating and preserving opportunities for the communication and exchange of ideas and research experiences among researchers, in addition to the University's initiative in fostering new ways of holding scientific meetings.

The meeting and the book were jointly organized by us, Professor Emma Otta, from the Department of Experimental Psychology of the University of São Paulo's Institute of Psychology (IPUSP), and Professor Patrícia Ferreira Monticelli, from the University of São Paulo's Department of Psychology at the Ribeirão Preto School of Philosophy, Science and Languages (FFCLRP-USP), as an extension of our research collaboration. Professor Monticelli coordinates the Laboratory of Ethology and Bioacoustics where research is carried out on reproductive, parental and communication behavioral aspects in terrestrial mammals. Professor Otta coordinates the Laboratory of Psychoethology, where research projects on Human Ethology are conducted. While studying nonverbal communication, she became interested in paralanguage, the non-verbal dimension of speech that contributes to its emotional quality.

During our collaborative research we noticed the need to consult specialists, given the interdisciplinary nature of the research topics under investigation. We systematize this experience here and share it with the readers through chapters that present the innovative research discussed in the talks and subsequent discussions, in the form of Peer Comments or Q&A transcription, prepared by the moderators of the presentations. We have divided the book into four parts: Animal Bioacoustics, Human Bioacoustics, Methods used in Bioacoustical Research and Analysis used in Bioacoustical Research.

Emma Otta & Patricia Ferreira Monticelli

Foreword

This book covers a fascinating topic: bioacoustics. After the course of events that guided me from an electronic engineer to a speech scientist, I thought my early dream of becoming a zoologist was over. This dream came true when I first met Patricia Monticelli and her lovely EBAC students back in 2016. Patricia's collaborations with psychologist Emma Otta and other colleagues were very fruitful, culminating recently in the Acoustic Communication: An Interdisciplinary Approach workshop. The book I have the honor of introducing is an outcome of this memorable event.

In fourteen chapters, the 22 contributors offer not only an inherently interdisciplinary approach to bioacoustics but give examples of several decades of scientific research developed for (human) speech. This is presented in the four parts: Animal Bioacoustics, Human Bioacoustics, Methods used in Bioacoustic Research, and Analysis used in Bioacoustic Research.

Part A opens this volume with a presentation of Bioacoustics as a subfield of Animal Communication, the latter a subfield of Ethology. Communication through sound is shown to be pervasive in both humans and non-humans, characterizing a social behavior that is crucial for each species. This is demonstrated when the authors investigate different intra- and inter-species behavior in primates, birds, guinea pigs, domestic and wild pigs, and domestic and wild horses, related or not to human interaction.

Part B describes what is known about speech production since Gunnar Fant’s work on Source-Filter theory, as well as what the study of the prosody of emotions in both verbal and non-verbal behaviors can offer to Bioacoustics, including the possibility of emotion recognition.

Part C presents software and algorithms developed for acoustic analysis in both human and non-human species, including infant cries. Praat, R, and BioVoice were used by the contributors of Part C, including presentations and examples of Machine Learning techniques for recognizing differences across bird species, different infant needs from their cries, and event detection.

Part D closes the book by presenting acoustic analysis applications to highlight the commonalities and differences in twins' speech, identify breathing issues in the case of voice disorders and build devices for speech synthesis and speaker identification, which is relevant for Forensic Phonetics in Speaker Comparison.

The researchers from Brazil, Canada, Denmark, and Uruguay that contributed to this wonderful book are prominent figures in the area of human and non-human sound communication. One major advantage of the 14 chapters is that they are written in a language that can be understood by both experts and people new to the area.

This book will be an important tool not only for students of Biology but also those in areas such as Computer Science, Electronic Engineering (including Telecommunications), Linguistics (including Phonetics), and Psychology. Experts from the same disciplines will also find a valuable resource for deepening their understanding of communication in all its shades and meanings by opening a window to a world where social networks must include non-human social networks aimed at a time where harmonic co-existence between species and nature will be a reality.

Plínio A. Barbosa
Department of Linguistics, University of Campinas
January 27th 2021

Chapter 14

Deep Learning approaches for Speech Synthesis and Speaker Verification

Edresson Casanova[26], Christopher Shulby[27] and Sandra Maria Aluísio[28]

Abstract

Speech synthesis is the artificial production of human speech, which can be used in applications such as text-to-speech, music generation, navigation systems and accessibility for visually-impaired people. As for the speaker recognition task, we can define it as the process of recognizing the speaker of a speech segment by processing speech signals, which can be broadly classified as speaker identification and verification. This chapter summarizes the Deep Learning practices applied in the field of speech synthesis and speaker verification. Speech synthesis and speaker verification have been widely investigated in speech technology applications, especially due to the popularity of virtual assistants. Considerable research has been conducted and significant progress has been made in the last 5-6 years. As Deep Learning techniques advance in most fields of machine learning, older state-of-the-art methods are also being replaced by Deep Learning methods in both speech synthesis and speaker verification areas. Thus, Deep Learning has apparently become the next generation solution for the synthesis and verification of speakers.

Keywords: Speech Technologies; Speech Synthesis; Speaker Verification; Deep Learning approaches.

Speech synthesis systems, also known as Text-To-Speech (TTS), have received considerable attention in recent years due to the popularization of virtual assistants, such as Amazon Echo (Purington et al., 2017), Google Home (Dempsey, 2017) and Apple Siri (Gruber, 2009). However, according to Tachibana et al. (2017), traditional Speech Synthesis systems are not easy to develop, since they are typically composed of many specific modules, such as a text analyzer, grapheme-to-phoneme

[26] PhD student at the Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, SP, Brazil
[27] Defined Crowd, Lisbon, Portugal.
28 Professor at the Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, SP, Brazil

converter, duration estimator, F0 generator, spectrum generator and vocoder. Figure 14.1 presents the main components of a traditional speech synthesis system. In summary, given an input text, the text analyzer module converts dates, currency symbols, abbreviations, acronyms, and numbers into their standard formats to be pronounced or read by the system, i.e., carries out text normalization and tackles problems such as homographs, then with the normalized text, the phonetic analyzer converts the grapheme into phonemes. In turn, the duration estimator estimates the duration of each phoneme. The acoustic model is used to generate acoustic characteristics such as F0 and a spectral envelope that corresponds to linguistic characteristics. Finally, the vocoder converts the spectrum into a waveform (Ze et al., 2013).

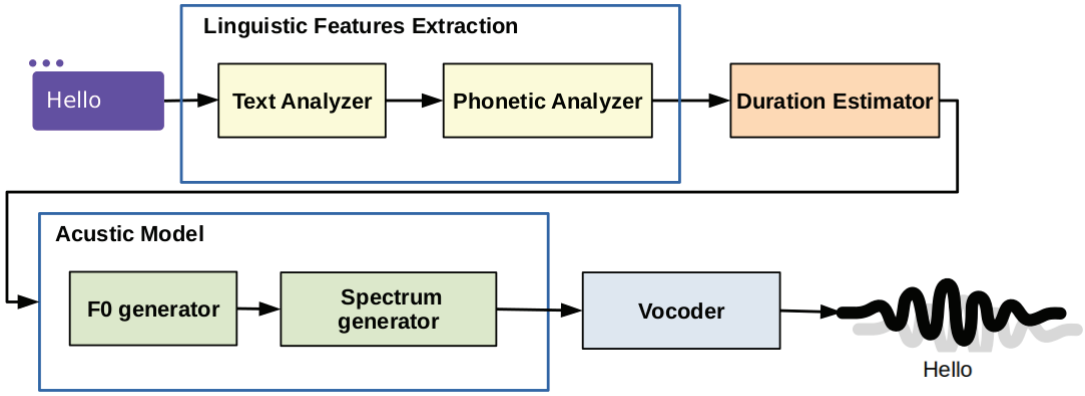


Figure 14.1. The main components of a traditional speech synthesis system.

The advent of Deep Learning (Goodfellow et al., 2016) has made it possible to integrate all processing steps into a single model and connect them directly from the input text to the synthesized audio output, which is known as end-to-end learning. Although neural models are sometimes criticized as being difficult to interpret, several end-to-end trained speech synthesis systems (e.g., Sotelo et al., 2017, Wang, Skerry-Ryan et al., 2017, Shen et al., 2018, Tachibana et al., 2018, Ping et al.,2018, Kim et al.,2020, and Valle et al., 2020) have been able to estimate spectrograms from text entries with promising performances.

Due to the sequential characteristic of text and audio data, the recurring units were the standard building blocks for speech synthesis, as in Tacotron 1 and 2 (Wang,Skerry-Ryan et al., 2017; Shen et al., 2018). In addition, the convolutional

layers showed good performance while reducing computational costs, as observed in the DeepVoice 3 (Ping et al., 2018) and Deep Convolutional Text To Speech (DCTTS) (Tachibana et al., 2018) models. On the other hand, with the recent popularization of Transformers (Vaswani et al., 2017), some transformer-based synthesis models have emerged, such as that proposed by Li et al. (2019), which performed similarly to Tacotron 2 (Shen et al., 2018), and trained 4.25 times faster. Finally, the flow-based models (Kingma et al., 2016; Hoogeboom et al., 2019; Durkan et al., 2019) attracted attention in the speech synthesis area, where the Flowtron (Valle et al., 2020) model surpassed the results reported by Tacotron 2 for enabling the manipulation of the latent space, allowing a change in characteristics such as speech speed and prosody. On the other hand, Kim et al. (2020) proposed GlowTTS, whose performance resembled that of the Tacotron 2, synthesizing speech 15.7 times faster.

The advent of Deep Learning has also enabled significant advances in speaker recognition. Speaker Recognition can be divided into three different subtasks: Speaker Verification (SV), Speaker Identification and Speaker Diarization. The objective of SV is to determine if two distinct audios contain the voice of the same speaker. On the other hand, speaker identification seeks to ascertain which speaker produced the voice on the audio file. Finally, Speaker Diarization splits an input audio stream into homogeneous segments according to the speaker's identity. In this study, we will only address Speaker Verification because it can be used in both of the other tasks cited above (Sztahó et al., 2019).

Currently, state-of-the-art (SOTA) Speaker Verification systems (Wang, Wang, Law et al., 2019; Deng et al., 2019; Chung, Huh et al., 2020; Casanova, Candido Junior, Shulby et al., 2020) allow the identification of new speakers without the need to retrain the model. This feature is very useful for different applications, such as meeting loggers, telephone-banking systems (Bowater & Porter, 2001) and automatic question answering (Ferrucci et al., 2010).

The objective of this study was to review the SOTA methods using Deep Learning that are applied in the speech synthesis area, focusing on Sequence-to-Sequence (seq2seq) models and speaker verification tasks. This text is subdivided in Speech datasets (main datasets employed in speech synthesis and speaker verification tasks); Deep Learning for the speech synthesis task; Deep Learning for the speaker verification task and conclusions and reflections.

.

Speech datasets

As with many tasks related to machine learning, the issue of the dataset used is fundamental. The methods developed can be evaluated and compared only if the same test circumstances are used. It is difficult to say whether an approach performs better if it is evaluated on a different dataset (or corpus) (Sztahó et al., 2019). Some datasets are used for speaker recognition and speech synthesis. Section *Speech synthesis datasets* presents the most commonly used datasets for speech synthesis in the English language, as well as the unique dataset publicly available for Brazilian Portuguese. Section *Speaker verification datasets* presents the main datasets used in the training and evaluation of speaker recognition models.

Speech Synthesis datasets

For the speech synthesis task, high quality datasets recorded in controlled environments are required. Since the purpose of speech synthesis is to synthesize high quality voice, if the training dataset contains noise, the model can synthesize it, which is not desired. The most widely used for training single-speaker speech synthesis models is the LJ Speech (Ito, 2017) dataset, which consists of 24 hours of speech by an English-language speaker. On the other hand, for multi-speaker synthesis, the LibriTTS (Zen et al., 2019) and VCTK (Veaux et al., 2016) datasets are the most commonly used. Although the most popular datasets are for English, other languages also have open datasets. With Portuguese, for example, the only publicly available dataset is the TTS-Portuguese Corpus (Casanova, Candido Junior, de Oliveira et al., 2020). Table 14.1 shows the approximate number of hours and total number of speakers of the main publicly available datasets for speech synthesis in English and the only dataset available for Portuguese.

Speaker Verification Datasets

For the SV task, the datasets created for the development of Automatic Speech Recognition (ASR) systems are commonly used due to their characteristics. Unlike speech synthesis datasets, their ASR counterparts generally have several speakers and few samples for each speaker; this feature is desired, since for Speaker Verification we want as many speakers as possible during model training (Sztahó et al., 2019). Thus, the datasets built for ASR models can be used to train and evaluate SV models. However, some datasets are made specifically for Speaker Verification. For example, VoxCeleb 2 (Chung et al., 2018) is currently the largest dataset built for SV. It consists of samples from more than 6,000 speakers downloaded from YouTube. Table 14.2 shows the approximate number of hours and total number of speakers in the main publicly available datasets for ASR and provides information about the VoxCeleb 2 dataset.

Table 14.1. Speech Synthesis datasets

Corpus	**Hours (~)**	**Total Speakers (~)**
LibriTTS (Zen et al. 2019)	586	2,456
M-AILAB	75	2
VCTK (Veaux et al. 2016)	44	109
LJ Speech (Ito 2017)	24	1
TTS-Portuguese Corpus (Casanova, Candido-Jr, de Oliveira, et al. 2020)	10.5	1

Table 14.2. Speaker Verification datasets

Corpus	**Hours (~)**	**Total Speakers (~)**
LibriSpeech (Panayotov et al. 2015)	986	2,848
Common Voice (Ardila et al. 2019)	2,508	58,250
TED-LIUM V3 (Hernandez et al. 2018)	452	2,028
VoxCeleb (J. S. Chung et al. 2018)	2,000	6,112

Sequence-to-Sequence Voice Synthesis Approaches

With the advent of Deep Learning, speech synthesis systems have evolved considerably and are still being studied intensively. Models based on Recurrent Neural Networks such as Tacotron Wang, Wang, Skerry-Ryan et al., 2017), Tacotron 2 (Shen et al., 2018), Deep Voice 1 (Arik, Chrzanowski et al., 2017) and Deep Voice 2 (Arık, Diamos et al., 2017) have gained prominence, but have high computational costs because they use recurring layers. This led to the development of fully convolutional models, such as DCTTS (Tachibana et al., 2018) and Deep Voice 3 (Ping et al., 2018), which sought to reduce the computational costs while maintaining good synthesis quality. On the other hand, more recently with the popularization of the Transformers, new Transformer-based models (Li et al., 2019; Kim et al., 2020) have emerged, and due to the parallelization of this architecture, the models achieved results similar to those of recurrent architectures with lower computing costs. Finally, the flow-based models (Kingma et al., 2016; Hoogeboom et al., 2019; Durkan et al., 2019) attracted attention in the synthesis area, allowing the training of simpler models with reduced computing costs. For example, the quality of the GlowTTS (Kim et al., 2020) model is similar to that of the recurrent Tacotron 2 model, but it can synthesize speech 15.7 times faster. The speech synthesis models are trained by receiving a text as input and a spectrogram as an expected output that represents the speech of the respective text input.

The model must learn to generate a spectrogram given the input text; the spectrogram is then transformed into a waveform using a vocoder. Neural vocoders have better quality speech synthesis, while phase reconstruction methods such as Griffin-Lim (GLA) (Griffin & Lim, 1984) and RTISI-LA (Real-Time Iterative Spectrogram Inversion with Look-Ahead) (Zhu et al., 2007) are based on Short Fast Fourier Transform (SFFT) redundancy (Sorensen & Burrus, 1988) and have higher synthesis speed and reduced quality. Figure 14.2 presents a general flow diagram of a TTS system based on Deep Learning. Briefly, given an input text, it is passed to the TTS model, which returns a spectrogram. Finally, this spectrogram is converted into a waveform by the vocoder.

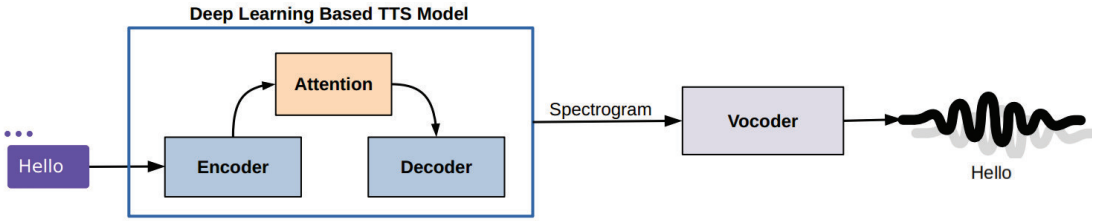


Figure 14.2. General flow diagram of a TTS system based on Deep Learning.

The most popular neural vocoders today are Wavenet (Tamamori et al., 2017), WaveRNN (Kalchbrenner et al., 2018), Waveglow (Prenger et al., 2019), GAN-TTS (Bínkowski et al., 2019), MelGan (Kumar et al., 2019) and more recently WaveGrad (Chen et al., 2020). Each of these vocoders has its advantages; some focus on higher quality and others on faster synthesis. In this study, we will not discuss vocoders, but they play a very important role in speech synthesis, converting a spectrogram into a waveform. In this chapter, we will only focus on models that convert text into spectrograms.

As mentioned above, a large amount of data is required to train speech synthesis models. For the English language, the most popular single speaker dataset for speech synthesis is called LJ Speech (Ito, 2017) and contains 24 hours of speech. On the other hand, in Brazilian Portuguese, the only available dataset is TTS-Portuguese Corpus (Casanova, Candido Junior, de Oliveira, et al., 2020) and contains 10 hours of speech. The speech synthesis models are subjectively evaluated using the Mean Opinion Score (MOS). Ribeiro et al. (2011) proposed a methodology for calculating MOS in speech synthesis and the vast majority of studies follow this technique. To calculate the MOS, the evaluators are asked to assess the naturalness of the statements generated on a five-point scale (from 1 = Bad to 5 = Excellent). Each participant evaluates the audio and the average MOS of the participant is calculated.

Tacotron 1 (Wang, Wang, Skerry-Ryan et al., 2017) was one of the first speech synthesis models to use only neural networks to transform text into a spectrogram. The authors proposed the use of a single deep neural network trained from end-to-end. Tacotron 1 includes an encoder, decoder and post-processing module, in addition to using an attention mechanism (Bahdanau et al., 2014) and convolutional filters, skipping connections (Srivastava et al., 2015) and Gated Recurrent Unit (GRU) neurons (Chung, Gulcehre et al., 2014). Tacotron also uses the Griffin-Lim algorithm

to convert the STFT spectrogram into the waveform (Griffin & Lim, 1984). Simultaneously, the Deep Voice 1 (Arik, Chrzanowski et al., 2017) model also emerged, which uses several neural submodels to synthesize speech into text. The Deep Voice 2 (Arık, Diamos et al., 2017) model was then proposed. This model is based on Deep Voice 1; however, the authors proposed improvements to surpass the results obtained by Tacotron 1. In addition, the authors proposed improvements in Tacotron 1 and changed the Griffin-Lim vocoder in favor of the WaveNet neural vocoder, thereby increasing the quality of the synthesized speech.

On the other hand, Shen et al. (2018) proposed an improvement on the Tacotron 1 model. They simplified the architecture and combined the new model with a modified version of the WaveNet (Tamamori et al., 2017) vocoder. Tacotron 2 is composed of a recurrent network of sequence prediction features that maps the incorporation of characters to Mel spectrograms, followed by a modified WaveNet model acting as a vocoder to synthesize waveforms in the time domain from these spectrograms. They also demonstrated that the use of Mel spectrograms as a conditioning input for WaveNet, instead of linguistic characteristics, allows for a significant reduction in the size of the WaveNet architecture, and consequently faster speech synthesis.

Furthermore, with the popularization of Transformers (Vaswani et al., 2017) in the Natural Language Processing (NLP) area, and the use of several language models such as BERT (Devlin et al., 2018), some transformer-based synthesis models have emerged. We can cite the work proposed by Li et al. (2019) which achieved quality comparable to that of Tacotron 2 (Shen et al., 2018), but trained 4.25 times faster.

Finally, more recent flow-based models (Kingma et al., 2016; Hoogeboom et al., 2019; Durkan et al., 2019) attracted attention in the synthesis area. Valle et al. (2020) proposed the Flowtron model, which reformulates from Tacotron 2 to provide high-quality and significant Mel spectrogram synthesis. Flowtron is optimized to maximize the likelihood of training data, which makes training simple and more stable. It allows the manipulation of several aspects of speech synthesis, such as pitch, tone, speech rate, cadence and accent. It achieved MOS scores slightly higher than those of Tacotron 2 and also allows for speech manipulation. On the other hand, Kim et al. (2020) proposed GlowTTS, whose quality is similar to that of Tacotron 2, but synthesizes speech 15.7 times faster. It uses transformers in its architecture and also

allows one to manipulate the velocity of speech. Both Flowtron and GlowTTS use the Waveglow neural vocoder.

Speaker Verification approaches

In the last decade, the area of speaker recognition has undergone major changes. In the past, speaker identification models could only identify speakers seen during training, and required a reasonable amount of speaker data to be able to learn to identify that speaker. Currently, speaker recognition models are able to identify speakers not seen in training using just a few seconds of the speaker's voice; this is known as the open-set scenario. This advance was possible due to the evolution of the machine learning area and the introduction of new cost functions applied to the training of these models.

Current speaker verification methods are trained using acoustic features, such as Mel-frequency cepstral coefficients (MFCCs) (Davis & Mermelstein, 1990) or Mel spectrograms, as inputs and use speaker IDs to calculate the loss. The models aim to learn a representation (speaker embedding), which is a vector of fixed size, to which the distance of the vectors of two different speakers is the greatest possible, while the distance of vectors of two samples of the same speaker are as close as possible. After training, the distance between these embeddings is usually calculated, allowing speakers to be identified. The performance of SV systems is commonly evaluated by the Equal Error Rate (EER) (Cheng & Wang, 2004). EER is a biometric security system algorithm used to predetermine threshold values due to its false acceptance index and false rejection rate (Cheng & Wang, 2004). EER indicates that the proportion of false acceptances is equal to that of false rejections, and the lower the EER, the more accurate the biometric system (Sztahó et al., 2019).

An SV system can be evaluated in two scenarios. In the closed-set scenario, where samples of speakers seen in the training of the SV model are used, the model recognizes these speakers. In the Open-set scenario, where speaker samples never seen in the training of the model are used, the model does not recognize these speakers. The models usually report only EER results for the Open-set scenario, since the goal of SV systems is to learn to differentiate speakers never seen in training, eliminating the need to retrain the neural model (Casanova, Candido Junior, Shulby et al., 2020).

The first studies to use deep neural networks in speaker recognition in an open-set scenario used speaker embeddings learned via the Softmax loss. Although the Softmax classifier can learn different embeddings for different speakers (Snyder et al. 2017, 2018), it is not non-discriminatory enough (Chung et al., 2020). To overcome this problem, the models trained with Softmax were combined with backends built in Probabilistic Linear Discriminant Analysis (PLDA) (Ioffe, 2006) to generate scoring functions (Ramoji et al., 2020; Snyder et al., 2018). On the other hand, Liu et al. (2017) proposed Softmax Angular, where the cosine similarity is used as logit input for the Softmax layer, showing its superiority over Softmax alone. Subsequently, Wang et al. (2018) proposed the use of Additive Margins in Softmax (AM-Softmax) to increase inter-class variance by introducing a cosine margin penalty to the target logit. However, according to Chung, Hu et al. (2020), training with AM-Softmax and AAM-Softmax (Deng et al., 2019) proved to be a challenge, since they are sensitive to scale and margin value in the loss function.

The use of contrastive (Chopra et al., 2005) and triple loss (Schroff et al., 2015; Bredin, 2017) has also achieved promising results in speaker recognition, but these methods require a careful choice of pairs or triplets, which is time-consuming and can interfere with performance (Chung, Hu et al., 2020).

Wang, Wang et al. (2019) proposed the use of prototypical networks (Snell et al., 2017) in speaker recognition. Prototypical networks seek to learn a metric space in which the classification of open sets of speakers can be performed by calculating distances for prototypical representations of each class. Generalized end-to-end loss (GE2E) (Wan et al., 2018) and Prototypical Angular (Chung, Hu et al., 2020) follow the same principle and recently achieved SOTA results in speaker recognition. Chung et al. (2020) compared the different loss functions mentioned above in the training of two convolutional models proposed by the authors. They showed that the Prototypical Angular loss function performs better than the others, demonstrating that it is more suitable for training SV models.

Finally, Casanova, Candido Junior, Shulby, et al. (2020) proposed a new training approach consisting of reconstructing the 1-second pronunciation of the phoneme /a/ in the voice of the speakers. After training, the model is able to approximate the pronunciation of /a/ in the voice of any speaker and an embedding of this reconstruction is extracted from an intermediate layer of the neural network. Given that the reconstruction of /a/ from the same speaker is always closer to their own than

to others, the model is applied in open-set scenarios. In addition, the method surpassed a model trained in a 500x larger dataset with the GE2E loss function. It also surpassed the result of the best model proposed by Chung, Hu et al. (2020) and trained with the Angular Prototypical loss function in one of the four datasets used to compare the models. Therefore, the method requires fewer data points to achieve competitive results.

Concluding remarks

In this chapter, we aimed to list the main Deep Learning approaches applied in the fields of Speech Synthesis and Speaker Verification. In the era of Deep Learning, as in most tasks involving machine learning, significant improvements in performance have been achieved when compared to classic/traditional methods. As Deep Learning techniques advance in most fields of machine learning, older, state-of-the-art methods are also being replaced by those using Deep Learning in both speech synthesis and speaker verification. Thus, Deep Learning has apparently become the next generation solution for speech synthesis and speaker verification (Sztahó et al., 2019). In some cases, Deep Learning opened up new research fronts, allowing us to meet demands that were not previously possible. In addition, speaker verification and speech synthesis systems are still evolving. In the Speech Synthesis field, the current goal is to reduce the computing cost of the models and improve speech manipulation mechanisms, with a view to synthesizing more expressive speech (Valle et al., 2020; Kim et al., 2020). On the other hand, in Speaker Verification, researchers still seek to advance the current results and focus more on new training methods for modeling (Chung, Hu et al., 2020; Casanova, Candido Junior, Shulby et al., 2020).

References

Arik, S. O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., & Shoeybi, M. (2017). *Deep voice: Real-time neural text-to-speech.* arXiv preprint arXiv:1702.07825. http://proceedings.mlr.press/v70/arik17a/arik17a.pdf

Arık, S. O., Diamos, G., Gibiansky, A., Miller, J., Peng, K., Ping, W., Raiman, J., & Zhou, Y. (2017). *Deep voice 2: Multi-speaker neural text-to-speech*. arXiv preprint arXiv:1705.08947 https://arxiv.org/pdf/1705.08947.pdf

Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate.* arXiv preprint *a*rXiv:1409.0473. https://arxiv.org/pdf/1409.0473.pdf?utm_source=ColumnsChannel

Bínkowski, M., Donahue, J., Dieleman, S., Clark, A., Elsen, E., Casagrande, N., & Simonyan, K. (2019). *High fidelity speech synthesis with adversarial networks*. arXiv preprint arXiv:1909.11646. https://arxiv.org/pdf/1909.11646.pdf

Bowater, R. J., & Porter, L. L. (2001, August 21). Voice recognition of telephone conversations. *Google Patents.* (US Patent 6,278,772).

Bredin, H. (2017). Tristounet: triplet loss for speaker turns embedding. (2017, March) In: *2017 Acoustics, Speech and Signal Processing (ICASSP) International Conference on Acoustics, Speech and Signal Processing (IEEE)* (pp. 5430–5434).

Casanova, E., Candido Junior, A., de Oliveira, F. S., Shulby, C., Teixeira, J. P., Ponti, M. A., & Aluisio, S. M. (2020). *End-to-end speech synthesis applied to Brazilian Portuguese*. arXiv preprint arXiv:2005.05144. https://arxiv.org/pdf/2005.05144.pdf

Casanova, E., Candido Junior, A., Shulby, C., da Silva, H. P., Cordeiro, A. F., Guedes, V. d. O., & Aluisio, S. M. (2020). *Speech2phone: A multilingual and text independent speaker identification model.* arXiv preprint arXiv:2002.11213. https://arxiv.org/pdf/2002.11213.pdf

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., & Chan, W. (2020). *Wavegrad: Estimating gradients for waveform generation*. arXiv preprint arXiv:2009.00713. https://arxiv.org/pdf/2009.00713.pdf

Cheng, J.-M., & Wang, H.-C. (2004, December*).* A method of estimating the equal error rate for automatic speaker verification. In: *International Symposium on Chinese Spoken Language Processing* (pp. 285–288).

Chopra, S., Hadsell, R., & LeCun, Y. (2005, June). Learning a similarity metric discriminatively, with application to face verification. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition: 1* (pp. 539–546).

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). *Empirical evaluation of gated recurrent neural networks on sequence modeling.* https://arxiv.org/pdf/1412.3555.pdf?ref=hackernoon.com

Chung, J. S., Huh, J., Mun, S., Lee, M., Heo, H. S., Choe, S., Ham, C., Jung,S., Lee, B.-J., & Han, I. (2020). In defence of metric learning for speaker recognition. *In Interspeech*. https://arxiv.org/pdf/2003.11982.pdf

Chung, J. S., Nagrani, A., & Zisserman, A. (2018). *Voxceleb2: Deep speaker recognition*. https://arxiv.org/pdf/1806.05622.pdf

Davis, S. B., & Mermelstein, P. (1990). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, *28*(4), 357-366. 10.1109/TASSP.1980.1163420 Retrieved 2020 from: https://ieeexplore.ieee.org/document/1163420

Dempsey, P. (2017). The teardown: Google home personal assistant. *Engineering & Technology, 12*(3), 80–81. 10.1049/et.2017.0330

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4690–4699). https://openaccess.thecvf.com/content_CVPR_2019/papers/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.pdf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805. https://arxiv.org/pdf/1810.04805.pdf&usg=ALkJrhhzxlCL6yTht2BRmH9atgvKFxHsxQ

Durkan, C., Bekasov, A., Murray, I., & Papamakarios, G. (2019). Neural spline flows. In: *Advances in neural information processing systems* (pp. 7511–7522).

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, L., Murdock, J. M., Nyberg, E., Prager, J., Schlaefer, N., & Welty, C. (2010). Building watson: An overview of the deepqa project. *AI magazine, 31*(3), 59–79. https://doi.org/10.1609/aimag.v31i3.2303

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* Cambridge, MA: MIT Press. http://www.deeplearningbook.org

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(2), 236–243. 10.1109/TASSP.1984.1164317. Retrieved 2020, from: https://ieeexplore.ieee.org/abstract/document/1164317

Gruber, T. R. (2009). Siri, a virtual personal assistant-bringing intelligence to the interface. In: *Semantic Technologies Conference*.

Hoogeboom, E., Berg, R. V. D., & Welling, M. (2019). Emerging convolutions for generative normalizing flows. *arXiv preprint* arXiv:1901.11137. Retrieved 2020, from: http://proceedings.mlr.press/v97/hoogeboom19a/hoogeboom19a.pdf

Ioffe, S. (2006). Probabilistic linear discriminant analysis. In: *European Conference on Computer Vision* (pp. 531–542). Springer, Berlin, Heidelberg.

Ito, K. (2017). *The lj speech dataset*. Retrieved April, 29 2020 from: https://keithito.com/LJ-Speech-Dataset/

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, A. van den, Dieleman, S., & Kavukcuoglu, K. (2018). *Efficient neural audio synthesis.* arXiv preprint arXiv:1802.08435. http://proceedings.mlr.press/v80/kalchbrenner18a/kalchbrenner18a.pdf.

Kim, J., Kim, S., Kong, J., & Yoon, S. (2020). *Glow-tts: A generative flow for text-to-speech via monotonic alignment search*. arXiv preprint arXiv:2005.11129. https://arxiv.org/pdf/2005.11129.pdf.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., & Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In: *Advances in neural information processing systems* (pp. 4743–4751).

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., Brebisson, A. de, Bengio, Y., Courville, A. C. (2019). Melgan: Generative adversarial networks for conditional waveform synthesis. In: *Advances in neural information processing systems* (pp. 14910–14921). Retrieved 2020 from: https://arxiv.org/pdf/1910.06711.pdf

Li, N., Liu, S., Liu, Y., Zhao, S., & Liu, M. (2019). Neural speech synthesis with transformer network. In: *Proceedings of the AAAI Conference on Artificial Intelligence: 33* (pp. 6706–6713).

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 212–220).

Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J., & Miller, J. (2018). *Deep voice 3: 2000-speaker neural text-to-speech*. [Conference paper]. Proceedings of the International Conference on Learning Representations - ICLR (pp. 214–217). Retrieved 2, May 2021 from: https://openreview.net/references/pdf?id=SyD5g8sPM

Prenger, R., Valle, R., & Catanzaro, B. (2019). Waveglow: A flow-based generative network for speech synthesis. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3617–3621).

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017). "Alexa is my new BFF" social roles, user satisfaction, and personification of the amazon echo. In: *Proceedings of the 2017 Conference on Human Factors in Computing Systems (CHI)* (pp. 2853–2859).

Ramoji, S., Krishnan V, P., Singh, P., & Ganapathy, S. (2020). Pairwise discriminative neural plda for speaker verification. *arXiv preprint* arXiv:2001.07034. https://arxiv.org/pdf/2001.07034.pdf

Ribeiro, F., Florêncio, D., Zhang, C., & Seltzer, M. (2011). Crowdmos: An approach for crowd-sourcing mean opinion score studies. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2416–2419).

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 815–823).

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, Rj., Saurous, R. A., Agiomvrgiannakis, Y., & Wu, Y. (April, 2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 4779–4783). 10.1109/ICASSP.2018.8461368. Retrieved 2020, from https://ieeexplore.ieee.org/xpl/conhome/8450881/proceeding

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. In: *Advances in neural information processing systems* (pp. 4077–4087).

Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. In: *Interspeech* (pp. 999–1003).

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018, April). X-vectors: Robust dnn embeddings for speaker recognition. In: *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 5329–5333). 10.1109/ICASSP.2018.8461375. Retrieved 2020, from https://ieeexplore.ieee.org/xpl/conhome/8450881/proceeding

Sorensen, H. V., & Burrus, C. S. (1988*)*. Efficient computation of the short-time fast fourier transform. In: *International Conference on Acoustics, Speech, and Signal Processing* (pp. 1894–1895).

Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A., & Bengio, Y. (2017). Char2wav: End-to-end speech synthesis. *Proceedings of the International conference on learning representations.*

Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Training very deep networks. In: *Advances in Neural Information Processing Systems* (pp. 2377–2385).

Sztahó, D., Szasz´ak, G., & Beke, A. (2019). *Deep learning methods in speaker recognition: a review.* arXiv preprint arXiv:1911.06615. https://arxiv.org/ftp/arxiv/papers/1911/1911.06615.pdf.

Tachibana, H., Uenoyama, K., & Aihara, S. (2018). Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4784-4788).

Tamamori, A., Hayashi, T., Kobayashi, K., Takeda, K., & Toda, T. (2017). Speaker-dependent wavenet vocoder. In: *Proceedings of Interspeech* (pp. 1118–1122).

Valle, R., Shih, K., Prenger, R., & Catanzaro, B. (2020). Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis. arXiv preprint arXiv:2005.05957. https://arxiv.org/pdf/2005.05957.pdf.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. In: *Advances in neural information processing systems* (pp. 5998–6008).

Veaux, C., Yamagishi, J., MacDonald, K., et al. (2016). Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh*. The Centre for Speech Technology Research (CSTR).

Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2018, April). Generalized end-to-end loss for speaker verification. *In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4879–4883). 10.1109/ICASSP.2018.8462665. Retrieved 2020, from https://ieeexplore.ieee.org/xpl/conhome/8450881/proceeding

Wang, F., Cheng, J., Liu, W., & Liu, H. (2018). Additive margin softmax for face verification. *IEEE Signal Processing Letters*, *25*(7), 926–930.

Wang, J., Wang, K.-C., Law, M. T., Rudzicz, F., & Brudno, M. (2019). Centroid-based deep metric learning for speaker recognition. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3652–3656).

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A.(2017). *Tacotron: A fully end-to-end text-to-speech synthesis model.* arXiv preprint arXiv:1703.10135. https://arxiv.org/pdf/1703.10135.pdf%EF%BC%89

Ze, H., Senior, A., & Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In: *2013 IEEE. International Conference on Acoustics, Speech and Signal Processing* (pp. 7962–7966).

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., & Wu, Y. (2019). Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint* arXiv:1904.02882. https://arxiv.org/pdf/1904.02882.pdf

Zhu, X., Beauregard, G. T., & Wyse, L. L. (2007). Real-time signal estimation from modified short-time fourier transform magnitude spectra. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(5), 1645–1653.

Peer Commentary

by Claudio Possani

This chapter by Edresson Casanova, Christopher Shulby and Sandra Maria Aluísio presents the contents of the first author's lecture at the Bioacoustic Meeting Brazil 2020. Edresson is a young researcher (doctoral student at ICMC/USP/SC) trained in Computer Sciences with a focus on Neural Networks and Deep Learning.

In this chapter the focus is on Speech Synthesis and Speaker Verification. This area has been receiving increasing attention for several years, with the emergence of so-called virtual assistants. The early 21st century saw the birth of methods known as Deep Learning. There was a revolution in the scope and possibilities that emerged.

Speech Synthesis techniques obtain good reproductions of human voices. The voice quality obtained is fundamental. It is important to underscore that the English language has received the largest number of resources and hours of recording and therefore, the best results. The chapter presents the primary models used, recording times and general characteristics of this type of study. The neural network concepts play an important role in this area.

Under the general name of Speaker Recognition, recent decades have seen enormous progress in tasks involving: (SV) Speaker Verification, (SI) Speaker Identification, and (SD) Speaker Diarization. Deep Learning also caused a revolution in the field of studies.

The specific aim of SV is to decide whether two different audio recordings were produced by the same person/speaker. SI attempts to identify the speaker that produced a certain sound recording from previously collected recordings. This is what some bank security systems do. SD splits the audio input stream into homogeneous segments according to the speaker's identity.

In the present chapter the authors address only questions related to Speaker Verification. One of the significant recent advances obtained from Deep Learning techniques are the so-called Open-set scenarios in which the system recognizes a speaker with even just a few seconds of acoustic recording, even if the speaker's recordings were not used by the system in the "learning" phase.

The final part of the chapter includes an original contribution by the first author and collaborators, which consists of the reconstruction of 1 second of the pronunciation of the /a/ phoneme, constant in the speaker's voice and, after a training period, the model is capable of producing the /a/ sound of any speaker, even from a very short recording. This makes it possible to identify the speaker or determine whether two recordings are of the same speaker by comparing the sounds produced. The present chapter is an introduction to this type of study, which is becoming increasingly relevant.