UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

Leonardo Boulitreau de Menezes Martins Marques

Cross-Speaker Style Transfer for TTS with Singing Voice Conversion Data Augmentation, Style Filtering, and F0 Matching

Transferência de Estilo entre Falantes para TTS baseada no Aumento de Dados com Conversão de Voz Cantada, Filtragem de Estilo e Correspondência F0

Campinas
2024

Leonardo Boulitreau de Menezes Martins Marques

Cross-Speaker Style Transfer for TTS with Singing Voice Conversion Data Augmentation, Style Filtering, and F0 Matching

Transferência de Estilo entre Falantes para TTS baseada no Aumento de Dados com Conversão de Voz Cantada, Filtragem de Estilo e Correspondência F0

Dissertation presented to the School of Electrical and Computer Engineering of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Electrical Engineering, in the area of Computer Engineering.

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica, na área de Engenharia de Computação.

Supervisora: Profa. Dra. Paula Dornhofer Paro Costa

Este exemplar corresponde à versão final da dissertação defendida pelo aluno Leonardo Boulitreau de Menezes Martins Marques, orientado pela Profa. Dra. Paula Dornhofer Paro Costa.

Campinas
2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

M348c

Marques, Leonardo Boulitreau de Menezes Martins, 1998-
Cross-speaker style transfer for TTS with singing voice conversion data augmentation, style filtering and F0 matching / Leonardo Boulitreau de Menezes Martins Marques. – Campinas, SP : [s.n.], 2024.

Orientador: Paula Dornhofer Paro Costa.
Dissertação (mestrado) – Universidade Estadual de Campinas (UNICAMP), Faculdade de Engenharia Elétrica e de Computação.

1. Fala. 2. Sistemas de processamento da fala. 3. Síntese da voz. 4. Prosódia (Linguística). I. Costa, Paula Dornhofer Paro, 1978-. II. Universidade Estadual de Campinas (UNICAMP). Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações Complementares

Título em outro idioma: Transferência de estilo entre falantes para TTS baseada no aumento de dados com conversão de voz cantada, filtragem de estilo e correspondência F0
Palavras-chave em inglês:
Speech
Speech processing systems
Speech synthesis
Prosody (linguistics)
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora:
Paula Dornhofer Paro Costa [Orientador]
Bruno Sanches Masiero
José Rafael Valle Gomes da Costa
Data de defesa: 28-06-2024
Programa de Pós-Graduação: Engenharia Elétrica

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-8821-4972
- Currículo Lattes do autor: http://lattes.cnpq.br/0974260157321182

Comissão Julgadora – Dissertação de Mestrado

Candidato: Leonardo Boulitreau de Menezes Martins Marques **RA:** 218479

Data da defesa: 28 de junho de 2024

Título da Tese: "Cross-Speaker Style Transfer for TTS with Singing Voice Conversion Data Augmentation, Style Filtering, and F0 Matching."

Profa. Dra. Paula Dornhofer Paro Costa (Presidente, FEEC/UNICAMP)
Prof. Dr. Bruno Sanches Masiero (FEEC/UNICAMP)
Dr. José Rafael Valle Gomes da Costa (University C. Berkeley)

A ata de defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. It is also supported by the BI0S - Brazilian Institute of Data Science, grant #2020/09838-0, São Paulo Research Foundation (FAPESP), and was conducted with the Artificial Intelligence Lab., Recod.ai, Institute of Computing, UNICAMP.

Completing this dissertation would not have been possible without the support and collaboration of several people who have been by my side throughout this journey.

First, I would like to express my deep gratitude to my advisor, Paula. Your precise guidance, patience and dedication were essential for the development of this work. Your wisdom and commitment have been a constant source of inspiration, and I find myself extremely fortunate as to consider you as the best possible guide I could have had on this path.

I want to thank all the text-to-speech research team at CPqD: Mário, Flávio, Fernando, Nagle, Bianca, Kátia, Fátima and Norberto. I am extremely grateful for you all to have believed in me and in my job. You have not only gave me the honor to conduct this work, but always brought up extremely valuable ideas and wise advices.

To my wife, Nathalia Cristina, my most special thanks. Your loving and constant presence by my side was essential for me to be able to face the challenges and difficulties along this journey. Your affection and strength provided me with the necessary balance to move forward, and I am eternally grateful for your unconditional support.

I would also like to thank my family, who has always been with me at all times. Your support, understanding and encouragement were fundamental pillars for the completion of this stage. Knowing that I could count on you at all times gave me the security and tranquility necessary to persist.

I couldn't help but thank my UNICAMP friends, who made this journey lighter and enriching. In particular, I would like to mention Lucas Ueda and Rodolfo Tonoli. Their contributions to academic discussions were invaluable and helped to shape and improve many of the ideas present in this work. The exchange of knowledge and stimulating conversations with you were one of the highlights of this academic journey.

Abstract

Text-to-speech (TTS) systems have become important means of human-machine interaction in various daily life applications, as seen in digital document readers, car navigation systems, and intelligent personal assistants. Despite their widespread use, many TTS systems still exhibit very monotonous speech, which can hinder effective communication and reduce user acceptance. To tackle this issue, various attempts to introduce aspects of human expressiveness into standard TTS have been increasingly proposed on literature. A very adopted approach is to directly record expressive data in a given speaking style and train a TTS model on the transcriptions. Although this technique was shown to reasonably achieve expressive models, it is not scalable, since for every new speaker, it must be entirely repeated. In this context, the cross-speaker style transfer task arises as a possible solution to mitigate the issue. It consists in using already recorded data by other (source) speakers in a given speaking style to build an expressive TTS for other speaker (target) with fewer or non-existent expressive data. Several techniques based on data augmentation were proposed to solve the task, but almost none consider the challenging scenario when the speaking styles are highly expressive (e.g. emotions), and with very different source and target speakers' timbres. In this context, the use of a pre-trained singing voice conversion (SVC) model is proposed as a means to convert the highly expressive data into target speaker's voice. In the conversion process, a fundamental frequency (F0) matching technique is applied to mitigate tonal variances between speakers with significant timbral differences. Also, a style classifier filter is employed to select only the converted audios with adequate style for the TTS training. While other methods require hours of neutral data of target speaker, the proposed approach is comparable to start-of-the-art requiring only a few minutes. Experiments report improvements brought by both the SVC and style filter in terms of naturalness and style intensity for the styles whose perception relies more on vocal qualities than on prosodic parameters. Also, increased speaker similarity is obtained with the F0 matching algorithm.

Keywords: Expressive Speech Synthesis; Style Transfer; Singing Voice Conversion; Data Augmentation.

Resumo

Os sistemas de conversão de texto em fala (TTS) tornaram-se meios importantes de interação homem-máquina em diversas aplicações da vida cotidiana, como por exemplo em leitores de documentos digitais, sistemas de navegação automotiva e assistentes pessoais inteligentes. Apesar da sua utilização difundida, muitos sistemas TTS ainda apresentam uma fala muito monótona, o que pode dificultar a comunicação eficaz e reduzir sua aceitação por pate do utilizador. A fim de mitigar esse problema, várias tentativas de introduzir aspectos da expressividade humana nos sistemas TTS comuns têm sido cada vez mais propostas na literatura. Uma abordagem muito adotada consiste em gravar diretamente dados expressivos em um determinado estilo de fala e treinar um modelo TTS nas transcrições. Embora essa técnica tenha demonstrado uma capacidade razoável de gerar modelos expressivos, ela não é escalável, uma vez que para cada novo falante deve ser inteiramente repetida. Neste contexto, a tarefa de transferência de estilo além-falante surge como uma possível solução para mitigar esse problema. Essa tarefa consiste em utilizar dados já gravados por outros falantes (apoio) em um determinado estilo de fala para construir um TTS expressivo para outro falante (alvo) com nenhum ou poucos dados expressivos. Várias técnicas baseadas no aumento de dados foram propostas para resolver a tarefa, mas quase nenhuma considera o cenário desafiador de quando os estilos de fala são altamente expressivos (por exemplo, emoções), e com falantes de apoio e alvo contendo timbres muito diferentes. Neste contexto, o uso de um modelo pré-treinado de conversão de voz cantada (SVC) é proposto, a fim de ser capaz de converter os dados altamente expressivos para a voz do locutor alvo. No processo de conversão, uma técnica de correspondência de frequência fundamental (F0) é aplicada para mitigar variações tonais entre alto-falantes com diferenças de timbre significativas. Além disso, um filtro classificador de estilos é utilizado para selecionar apenas os áudios convertidos com estilo adequado para o treinamento do TTS. Enquanto outros métodos necessitam de horas de dados neutros do falante alvo, a abordagem proposta é comparável ao estado da arte necessitando de apenas alguns minutos. Experimentos relatam melhorias trazidas pelo SVC e pelo filtro de estilo em termos de naturalidade e intensidade do estilo para os estilos cuja percepção depende mais de qualidades vocais do que dos parâmetros prosódicos. Além disso, um aumento da similaridade dos alto-falantes é obtido com o algoritmo proposto de correspondência F0.

Palavras-chave: Síntese de Fala Expressiva; Transferência de Estilo; Conversão de Fala Cantada; Aumento de Dados.

List of Figures

2.1 Kratzenstein's resonators. Source: Extracted from (BRACKHANE, 2015). 24
2.2 Original Design of Kempelen's Speaking Machine. Source: Extracted from (KEMPELEN, 1791). 25
2.3 Picture of Joseph Faber's Euphonia. Source: Unknown. 26
2.4 A Scheme of the Electrical Synthesizer designed by Stewart. Source: Extracted from (STEWART, 1922). 26
2.5 (a) Schematic of the VODER speech synthesizer. (b) Picture of Woman demonstrating the VODER operation. Source: (a) Extracted from (DUDLEY, 1940). (b) Source: Unknown. 28
2.6 The Pattern Playback machine. Source: Extracted from (COOPER *et al.*, 1951). 29
2.7 DAVO Synthesizer. Source: Unknown. 31
2.8 a) Block Diagram of the ASY. Source: Extracted from (RUBIN *et al.*, 1981). b) ASY Synthesis Scheme. Source: Extracted from (RUBIN *et al.*, 1981). . 32
2.9 A generic training scheme of HMM-based speech synthesis models. Source: Extracted from (ZEN *et al.*, 2007). 33

3.1 Data transformation on a speech synthesis pipeline entirely based on neural models. Source: Extracted from (TAN *et al.*, 2021). 36
3.2 *Fastpitch* architecture. Source: Extracted from (ŁAŃCUCKI, 2021). 39
3.3 Architectures of the generator (left) and both discriminator types (right) of the BigVGan architecture. Source: Extracted from (LEE *et al.*, 2023). . . . 41
3.4 Summary of factors proposed by Aylett *et al.* (2021) to consider when designing an expressive speech synthesis model. 45
3.5 Taxonomy of deep neural expressive synthetic speech works. Source: Extracted from (TRIANTAFYLLOPOULOS *et al.*, 2023). 47
3.6 Speech Attributes Disentanglment Module of NaturalSpeech3. Source: Extracted from (JU *et al.*, 2024). 50
3.7 a) PS augmentation step. Source: Extracted from: (TERASHIMA *et al.*, 2022). b) VC training and augmentation step. Source: Extracted from: (TERASHIMA *et al.*, 2022). c) TTS and Vocoder training steps. Source: Extracted from: (TERASHIMA *et al.*, 2022). 56

4.1 Overview of the proposed pipeline to build a TTS model with target speaker's voice and source speaker's style. 60
4.2 Architecture of the SO-VITS-SVC pipeline. Green highlights the inputs. Blue modules are used only during training. 61
4.3 Processing pipeline to generate the semitonal distances. 63
4.4 Block diagram of the style filtering process. 65

4.5 Architecture of the Style Classifier Filter RE. Source: Extracted from (SKERRY-RYAN *et al.*, 2018b). 66
4.6 Daft-Exprt architecture. Source: Extracted from (ZAïDI *et al.*, 2022). . . . 71
4.7 Main components of the Daft-Exprt. Source: Extracted from (ZAïDI *et al.*, 2022). 71
4.8 (a) Training procedure of FreeVC. Source: Extracted from (LI *et al.*, 2023). (b) Inference procedure of the FreeVC. Source: Extracted from (LI *et al.*, 2023). 73

5.1 Style Intensity MOS results for each style and each stimulus with 95% confidence intervals. 82
5.2 Confusion matrix of the style classifier filter on the validation set. 83
5.3 Confusion matrix of the style classifier filter applied on synthetic converted expressive dataset. 84
5.4 Speaker Similarity MOS results for each style and each stimulus with 95% confidence intervals. 85
5.5 Similarity for each converted speaker. 86

A.1 Welcome and requirements page. 105
A.2 Instructions concerning attention checks. 106
A.3 Instructions for the style intensity experiment. 106
A.4 Page elucidating the “angry” speaking style with sample audios. Repeated for all styles considered. 107
A.5 Sample page of the style intensity experiment for the “angry” style. Repeated for all styles considered. 108
A.6 Instructions for the naturalness experiment. 109
A.7 Sample page of the naturalness experiment. 109
A.8 Instructions for the speaker similarity experiment. 110
A.9 Sample page of the speaker similarity experiment. Repeated for all styles considered. 111
A.10 Participant ID, age and gender profiling page. 112
A.11 Sample page showing the results for the participant at the end of the experiment. 112

List of Tables

2.1 A summary of the principal techniques in speech synthesis' history before neural networks. Source: Adapted from (ROSENHOUSE, 2010). 34

3.1 Representative TTS acoustic models and their structure characteristics. "Ph" means phonemes, "Ch" means characters, "AR/NAR" means autoregressive or not. 38
3.2 Representative E2E models and their characteristics. 43
3.3 Some perspectives of modeling variation information for TTS. Source: Extracted from (TAN *et al.*, 2021). 46
3.4 Summary of cross-speaker style transfer based on data augmentation techniques presented in Section 3.2.3. (1) Approximated with the number of utterances multiplied by the medium utterance duration. 58

4.1 Summary of the datasets used in the pipeline. 70
4.2 Time taken to process each step of the proposed pipeline. 74

5.1 Distributions of the text-audio pairs selected for perceptual evaluation. . . 79
5.2 Naturalness MOS with 95% confidence intervals. 81
5.3 Style Intensity Mean Opinion Scores MOS with 95% confidence intervals. . 81
5.4 Speaker Similarity MOS with 95% confidence intervals. 84

List of Acronyms

ASR Automatic Speech Recognition

ASY Articulatory Synthesizer

BCE Binary Cross Entropy

CASY Configurable Articulatory Synthesizer

CER Character Error Rate

cGAN Conditional Generative Adversarial Networks

CNN Convolutional Neural Networks

CPU Central Processing Unit

CST Coarse-Grained Style Transfer

DAVO Dynamic Analog of the Vocal Tract

DNN Deep Neural Networks

E2E End-to-end

EMOVDB Emotional Voices Database

ESD Emotional Speech Dataset

EVC Emotional Voice Conversion

FFE F0 Frame Error

FFT Feed Forward Transformer

FST Fine-Grained Style Transfer

G2P Grapheme-to-phoneme

GAN Generative Adversarial Networks

GMM Gaussian Mixture Models

GPE Gross Pitch Error

GPU Graphics Processing Unit

GRL Gradient Reversal Layer

GRU Gated Recurrent Unit

GST Global Style Tokens

HMM Hidden Markov Models

LJ Linda Johnson

LPC Linear Predictive Coefficients

MAE Mean Absolute Error

MCD Mel Cepstral Distortion

MFA Montreal Forced Aligner

MOS Mean Opinion Score

MPD Multi-Period Discriminator

MRD Multi-Resolution Discriminator

MSE Mean Squared Error

MUSHRA Multiple Stimuli with Hidden Reference and Anchor

OOD Out-of-Distribution

OVE Orator Verbis Electris

PAT Parametric Artificial Talker

PEQS Perceptual Evaluation of Speech Quality

POS Part-of-speech

PPG Phonetic Posteriorgrams

PS Pitch-Shift

PSOLA Pitch Synchronous OverLap Add

Q-VAE Quantized Variational Auto Encoder

RE Reference Encoder

ReLU Rectified Linear Unit

RNN Recurrent Neural Networks

RVQ Residual Vector Quantizer

SNAC Speaker-Normalized Affine Coupling Layer

SOTA State-of-the-Art

SPSS Statistical Parametric Speech Synthesis

SSL Self-Supervised Learning

STFT Short-Time Fourier Transform

STOI Short-Time Objective Intelligibility Measure

SVC Singing Voice Conversion

TPGST Text-Predicted Global Style Tokens

TTS Text-to-Speech

UTMOS UTokyo-SaruLab MOS Prediction System

VAD Voice Activity Detection

VAE Variational Auto-Encoder

VC Voice Conversion

VCTK Voice Cloning Toolkit

VDE Voicing Decision Error

VITS Variational Inference with Adversarial learning for end-to-end Text-to-Speech

VOCODER Voice Coder

VODER Voice Operation Demonstrator

VQ-VAE Vector-Quantized Variational Auto-Encoder

WER Word Error Rate

List of Symbols

A	Amplitude
ω	Angular Frequency
$\lvert\,.\,\rvert$	Cardinality
$\lVert . \rVert_F$	Frobenius Norm
F_0	Fundamental Frequency
$\mathcal{N}(z;\, \mu,\, \sigma^2)$	Gaussian Distribution
$\mathcal{L}$	Loss Function
$\lVert . \rVert_p$	ℓ_p Norm
μ	Mean
θ	Phase
$\mathbb{N}$	Set of Natural Numbers
σ^2	Variance

Summary

1 Introduction **17**
1.1 Speech Synthesis . . . 17
1.2 Problem Definition . . . 19
1.3 Objectives . . . 20
1.4 Research Questions . . . 20
1.5 Contributions . . . 21
1.6 Organization . . . 22

2 Historical Perspective **23**
2.1 Mechanical and Eletro-mechanical era . . . 23
2.2 Electric and Electronic era . . . 26
2.3 Digital and Computational era . . . 30
2.4 Literature Summary . . . 34

3 Basic Concepts and Related Works **35**
3.1 Neural Speech Synthesis . . . 35
3.1.1 Textual Analysis . . . 37
3.1.2 Acoustic Models . . . 37
3.1.3 Vocoders . . . 40
3.1.4 End-to-end Models . . . 42
3.2 Expressiveness in Speech Synthesis . . . 43
3.2.1 Speaking Styles . . . 44
3.2.2 Expressive TTS . . . 44
3.2.3 Cross-Speaker Style Transfer . . . 47
3.3 Concluding Remarks . . . 57

4 Method **59**
4.1 Singing Voice Conversion . . . 61
4.2 F_0 Matched Conversion . . . 63
4.3 Style Filtering . . . 65
4.4 Text-to-Speech . . . 67
4.5 Data . . . 67
4.5.1 OpenSinger Dataset . . . 68
4.5.2 Emotional Speech Dataset . . . 68
4.5.3 VCTK Dataset . . . 69
4.5.4 LJSpeech Dataset . . . 69
4.6 Experimental Setup . . . 69
4.6.1 Baselines . . . 70
4.6.2 Ablations . . . 72

4.6.3 Training Setup . . . 73
4.7 Concluding Remarks . . . 74

5 Evaluation **75**
5.1 Perceptual Protocol . . . 77
5.2 Stimuli . . . 79
5.3 Results . . . 80
5.3.1 Naturalness . . . 80
5.3.2 Style Intensity . . . 81
5.3.3 Speaker Similarity . . . 84
5.4 Concluding Remarks . . . 86

6 Conclusions **88**
6.1 Limitations and Future Work . . . 90

A Perceptual Assessment **105**

Chapter 1

Introduction

1.1 Speech Synthesis

The synthesis of speech from text, also known as text-to-speech conversion, or simply Text-to-Speech (TTS), is a technology whose main task is to transform a textual input into a speech signal that utters the corresponding text. These systems are present in various everyday applications, such as digital document readers, guide and reception robots, cell phone and car navigation systems, and intelligent personal assistants. In these scenarios, TTS systems characterize a component of the interface between users (humans) and machines (HAYASHI *et al.*, 2020). Since speech is seen as a more natural interface than graphic design-based ones, it is becoming the primary form of communication between humans and machines (ABDUL-KADER; WOODS, 2015). Thus, TTS systems have become particularly indispensable in the design of human-machine interfaces, especially when considering more complex interactive social agents, such as digital embodied avatars and social robots.

For these systems to effectively be able to become a conversational entity with a human-like way of communicating, they have to demonstrate *expressiveness*, which could be heard through carefully designed modifications of speech's intonation (BATLINER; MÖBIUS, 2005). By speaking affectively, these systems could manifest mood, personality, and social status, making them capable of vocalizing intimate thoughts, feelings, and emotions, for example (TRIANTAFYLLOPOULOS *et al.*, 2023).

Expressive speech is considered fundamental, for example, for the anthropomorphization of social robots. This aspect is crucial as it implies an increase in the accessibility of robots for the people who interact with them (JAMES *et al.*, 2018). Empathetic robots controlled their users' stress, sought more comfort, and obtained better performance in the task for which they were intended (NICULESCU *et al.*, 2013). Therefore, good modeling of speech's expressiveness is necessary to support and adapt the robot's oral communication to its affective state (JAMES *et al.*, 2018).

With the dissemination of deep learning techniques in different research areas, TTS systems have reached a level of neutral speech synthesis quality with high naturalness and intelligibility (ZHOU *et al.*, 2022b). However, the limitation in terms of expressiveness still characterizes an evident gap between synthesized speech and human recordings, with which the models are trained. Thus, integrating expressiveness in synthesized speech is an important research topic in industry and academia (LEI *et al.*, 2022a).

Synthesizing speech expressively consists of capturing the diversity of prosodic, temporal, and spectral characteristics (information beyond text, denominated paralinguistic) that occur naturally in speech. In this context, the problem of expressive speech synthesis is characterized as a *one-to-many* problem because the exact input text can be pronounced in different ways depending on several factors, including the context, emotion, dialectics, and the speaker's habitual speech patterns (LI *et al.*, 2022).

One of the common approaches to modeling expressiveness in text-to-speech systems is the modeling of *styles* of speech, which can be understood as "ways of speaking", each having a own defining prosodic pattern. For example, some speaking styles are based on individual emotions (ADIGWE *et al.*, 2018; ZHOU *et al.*, 2021a): amused, angry, happy, sad, surprise, sleepy and disgust. On the other hand, others are designed for interactions, such as styles based on social attitudes (MOINE; OBIN, 2020): friendly, seductive, dominant, and distant; or even designed towards customer interaction (MARQUES *et al.*, 2022): lively, welcoming, and harsh.

Expressiveness in speech can be unveiled with the presence of complex intricate patterns on the utterance's prosody (suprasegmental elements of speech such as intonation, stress and rhythm). The most traditional approach to generate artificial expressive speech consists in directly recording the desired speaking styles of the desired speaker then training a neural network straightforwardly to learn these complex stylistic patterns by predicting speech spectrograms, a visual representation of how the frequencies present on a speech vary over time, from the text.

Nevertheless, recording one's speech is not only time-demanding, often requires a considerable volumetry (total duration of the audio files) of hours of recordings, and is also costly if high quality is necessary. An additional issue when considering expressive speech is that the speaker must be able to convey the desired expressiveness properly. Otherwise, either the emotion or speaking style will be confused or ill-defined. Inserted in this context, this work aims to make use of already existing expressive speech recorded by another speaker to make speaker, that has only little neutral data record, to speak expressively.

1.2 Problem Definition

The traditional recording approach to obtain expressiveness is often impractical and lacks scalability, as expressive speech would need to be re-recorded for every new speaker in the dataset, and with the risk that the speaker may not perform effectively in the required speaking style (RIBEIRO *et al.*, 2022; PAN; HE, 2021). Moreover, current State-of-the-Art (SOTA) TTS systems generally require a minimum of dozens of hours of high-quality transcribed speech data to achieve satisfactory performance (LIAN *et al.*, 2023).

Cross-speaker style transfer involves transferring a speaking style from one speaker (the "source") to synthesized speech in another speaker's (the "target") voice (LIU *et al.*, 2022). This technique allows expressive attributes from high-resource speakers to be transferred to low-resource speakers (HUYBRECHTS *et al.*, 2021). Various approaches have been proposed in the literature to perform style transfer, however, none based on data augmentation dealt at the same time with highly expressive styles and speakers with very different timbres, or with target speakers with very low neutral volumetry.

In this context, the present work aims to advance expressive text-speech synthesis techniques for affective human-machine interactions by developing a method that, given only a few neutral data of a desired speaker, produces an expressive TTS with its voice by making use of expressive data from another speakers. In particular, this work aims to develop a new data augmentation-based neural cross-speaker transfer technique.

With the use of a pre-trained Singing Voice Conversion (SVC) model, a system that is able to change the voice of an utterance while keeping its intonation, to better capture one's expressive voice, open-source neutral and expressive speech datasets are converted to the target speaker's voice by adjusting the intonation in the case of sufficiently different vocal timbres with an proposed F_0 matching technique. Then, a TTS model conditioned on prosodic parameters is trained on a filtered style-appropriate version of this converted synthetic data. In the end, an expressive TTS model on target speaker's voice is obtained.

The problem can be defined as: given at least X minutes of neutral speech of a speaker, denominated as target; at least N minutes of neutral speech, composed by any combination of speakers; and at least E minutes of expressive speech on a given speaking style, also composed by any combination of speakers, the goal is to develop a text-to-speech model that generates speech with the given speaking style in target speaker's voice for any input text. The other speakers are all denominated as source speakers and compose the source datasets, both neutral and expressive. In this work, the proposed technique allows setting X, the most critical parameter of the task, to as low as five minutes.

Assumptions related to this work:

- An open source pre-trained model is used as the SVC. Since the model is already

trained to perform conversion and has seen various voices, it is able to quickly adapt to a new voice, which reduces the need for neutral target speaker volumetry.

- An open source neutral dataset is required, with volumetry enough to train a TTS model on its own.
- An open source expressive dataset is required, with volumetry enough so that a TTS can be fine-tuned upon.
- All source speech has to be annotated, in order to make it possible to train the text-to-speech models.

1.3 Objectives

This work delves into the development of a multi-stage speaking style transfer pipeline based on data augmented with an SVC model. It allows a speaker with no expressive data to speak any given textual input expressively. The proposed method is suited for a condition in which only its neutral data is available. This allows the reuse of already recorded speech and reduces the burden of the costly and complex process of recording expressive data with the desired speaker, which is not only time-demanding but also requires the speaker to be able to precisely know how to articulate in the desired speaking style properly. In this context, the main objectives of this work can be summarized as follows:

- Develop an expressive text-to-speech model in target speaker's voice.
- Allow the reuse of existing expressive speech data of another (source) speaker.
- Avoid speaker timbre leakage and glitches risen from the use of synthetic data.
- Eliminate the necessity to record expressive speech for all speakers ought to speak expressively in a dataset.

1.4 Research Questions

This work is guided by the following research questions:

- **Q.1**: Can data augmentation-based techniques perform cross-speaker style transfer of highly expressive speaking styles with only a few minutes of neutral data of target speaker?

- **Q.2**: Since singing voice includes richer emotional information compared to regular speech (HUANG *et al.*, 2021), is an SVC model (instead of a Voice Conversion (VC)), more effective to preserve the speaking style when converting expressive speech to a speaker with only neutral data?

- **Q.3**: Does filtering out the synthetic audios that do not maintain the same style after being converted, judged by a style classifier trained on the original audios, improves the style intensity of the TTS?

- **Q.4**: To what extent does the difference in timbre between a source speaker and a target speaker impact the perceived similarity of the converted speech to the target speaker, as measured by a speaker similarity metric?

- **Q.5**: How do current open-source cross-speaker style transfer approaches perform on open-source data?

1.5 Contributions

The contributions of this work can be summarized as follows:

- We propose to augment data with an SVC model to capture better the expressiveness of the source speakers for the cross-speaker style transfer task.
- We employed an F0 matching technique that mitigates timbral differences between target and source speakers.
- A style classifier filter was designed to select the most expressive converted data to perform the style finetuning.
- We propose using transfer learning and a base neutral dataset to reduce the amount of the target speaker's neutral data, lowering this value to only a few minutes.
- Only open-source data and models were used. All the generated models, code, and audios are made available [1].
- We compared the proposed techniques with several other methods to perform cross-speaker style transfer, providing a perceptual evaluation benchmark for the current research state.

Other proposed approaches to perform style transfer of speaking styles for expressive TTS exploited simultaneously were reported in the following publications:

[1] Audios available at <https://svcstytransfer.netlify.app/>

- MARQUES, L. B. De M. M.; UEDA, L. H.; SIMÕES, F. O.; ULIANI NETO, M.; RUNSTEIN, F. O.; NAGLE, E. J.; DAL BÓ, B.; COSTA, P. D. P. Diffusion-Based Approach to Style Modeling in Expressive TTS. In: 11th Brazilian Conference on Intelligent Systems, 2022, Campinas. Intelligent Systems. BRACIS 2022. Lecture Notes in Computer Science, vol 13653. Springer, Cham. Available at: <https://link.springer.com/chapter/10.1007/978-3-031-21686-2_18>. Accessed: Set. 09, 2023.

- MARQUES, L. B. De M. M.; UEDA, L. H.; COSTA, P. D. P. Transferência de Estilo para Síntese de Fala Expressiva. In: Décimo Quarto Encontro dos Alunos e Docentes do Departamento de Engenharia de Computação e Automação Industrial, 2022, Campinas. Digital Proceedings. Campinas: FEEC: Unicamp, 2022. Available at: <https://www.dca.fee.unicamp.br/portugues/pesquisa/seminarios/2022/manuscritos/papers/18.pdf>. Accessed: Set. 09, 2023.

1.6 Organization

This document is organized as follows:

- Chapter 2 presents a historical perspective of the diverse approaches to synthesize human speech since the very first early attempts;

- Chapter 3 presents some basic concepts of neural text-to-speech and expressiveness in speech, as well as the main related works involving style transfer and data augmentation for expressive TTS.

- Chapter 4 describes the proposed method to achieve cross-speaker style transfer via data augmentation. All the developed techniques to improve performance are described: the use of singing voice, the F_0 matching algorithm, and the style classifier filter. Each step of the pipeline is detailed, alongside all used data and experimental setup to perform the research.

- Chapter 5 details the perceptual assessment and objective metrics conducted to analyse and compare the performance of the proposed approach against baseline models and the results obtained.

- Chapter 6 contains the concluding remarks of this project and a brief discussion on future work.

- Appendix A presents screenshots of the perceptual assessment conducted online.

Chapter 2

Historical Perspective

The idea of synthesizing human speech has been going through the human mind since the late 1800s. The famous mathematician, physicist, and engineer Leonhard Euler wrote in 1761: "... All the skill of man has no hitherto been capable of producing a piece of mechanism that could imitate [speech]." Additionally, he claimed: "The construction of a machine capable of expressing sounds, with all the articulations, would no doubt be a significant discovery." He pictured this device possibly assisting those "whose voice is either too weak or disagreeable." Throughout engineering history, the attempts to generate human speech can be broadly subdivided into three main eras: the mechanical and electro-mechanical era, the electric and electronic era, and lastly, the digital and computational era, which persists till the moment in which this work is written (STORY, 2019).

2.1 Mechanical and Eletro-mechanical era

The first ever attempt to synthesize human sounds successfully happened in the late 1800s, and it is attributed to Christian Gottlieb Kratzenstein, a Professor of Physics at the University of Copenhagen, and Wolfgang von Kempelen, a Hungarian Engineer, industrialist, and government official. Kratzenstein, who shared an interest in the study of physical aspects of speaking with Euler, submitted a detailed report in 1780 of the design of five organ pipe-like resonators that were able to produce five vowels: {a,e, i,o,u}, when excited with the vibration of a reed. Only sustained sounds were possible with the system and were not similar to the human way of producing speech (STORY, 2019). These resonators are shown in Figure 2.1.

Simultaneously, Kempelen was working towards the same goal, since 1769. In 1783, on a tour across Europe, he presented his invention, the so-called "speaking machine". Differently from Kratzenstein's pipes, which could only generate vowels, Kempelen's speaking machine was considered the first capable of generating entire words of human speech. The original design of the machine was published in 1791 in a 450-page document

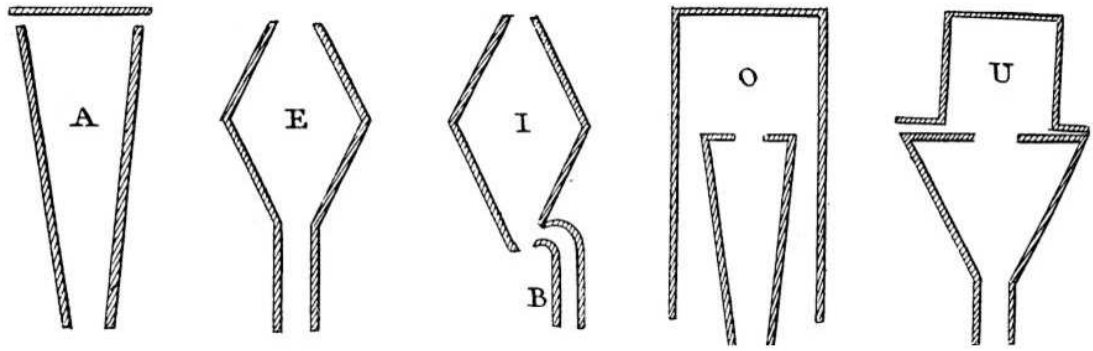


Figure 2.1: Kratzenstein's resonators. Source: Extracted from (BRACKHANE, 2015).

translated to "Mechanism of Human Speech and Language", which not only described precisely how to build it but also brought thorough examinations on the nature of speech, its sounds, and the human speech organs (CULTURE, 2016). The speaking machine's original design is shown in Figure 2.2.

To some degree, the speaking machine can be considered a mechanical simulation of human speech production (STORY, 2019), and it has to be played like a musical instrument by a human operator. As Kempelen wrote, to produce speech, one has to rest the right arm on the bellows (X on Figure 2.2) and pump it up and down so that the speech is produced on a down motion since the air pressure generated causes the reed to vibrate. The air passes to a wind-box that emulates a trachea and is outputted through a rubber funnel that emulates the vocal tract. On the wooden wind box, the right-hand fingers should operate the consonant port and lever controls (r, sch, n, m, and s, on the figure), and the left hand should be placed palm inward before the opening "BC" of the bell "C". With this configuration, the vowels could be produced by first pumping the bellows with the right elbow while the nostril-imitating tubes (m and n) are blocked with the right hand. The left-hand position varies according to which vowel is desired. For example, for the vowel "a", the left hand should be kept distant from the mouth opening, and for the vowel "e", it should be hollowed slightly with the bottom edge against the mouth and the top edge one inch away from it (DUDLEY; TARNOCZY, 1950).

Inspired by the work of Kempelen, some other attempts to perfect human speech production also happened in the nineteenth century. A remarkable one was the later-named "Euphonia" machine, constructed by the Austrian inventor Joseph Faber. The Euphonia had much better control over Kempelen's speaking machine: it featured a 7-key keyboard that could meticulously control fine-grained airflow distinctions, thus better emulating the human organs to produce speech. It was said to speak any word in any European language and even sing the song "God Save the Queen" with a skilled operator (RAMSAY, 2019). The "Euphonia" is shown in Figure 2.3.

For the rest of the 19th century to the very beginning of 20th century, instead of trying to emulate speech from the simulation of human organs, the attempts were more

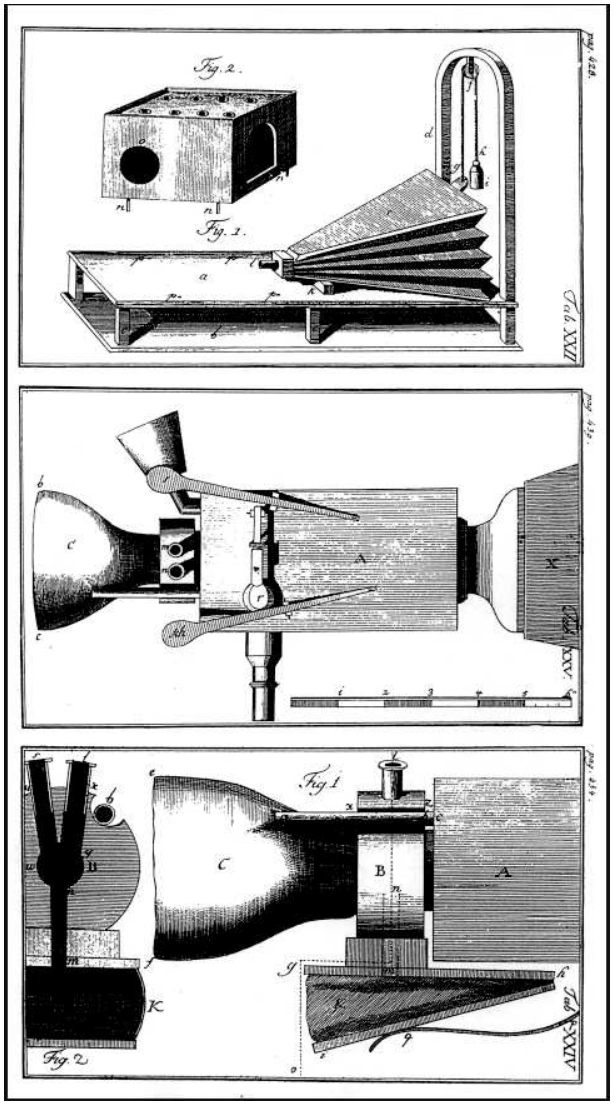


Figure 2.2: Original Design of Kempelen's Speaking Machine. Source: Extracted from (KEMPELEN, 1791).

focused on some form of spectral addition produced by electro-mechanical devices that produced spectral components of speech waveforms. For example, Hermann Helmholtz, a German scientist, developed an electromagnetic system in 1859 to maintain the vibration of turning forks, each coupled to a resonating chamber. He showed that musical notes and speech vowels resulted from combining different overtones (integer multiples of the fundamental frequency) and linked them to the resonances in the vocal tract (ROSENHOUSE, 2010). His studies led to the beginning of the field of psychoacoustics. In 1914, Dayton Miller, an American scientist, described intricate machines able to add various sinusoidal components and generate complex representations of waveforms. He dubbed this process as "harmonic synthesis". This was the first time the term "synthesis" in this context of speech synthesis appeared in history (STORY, 2019).

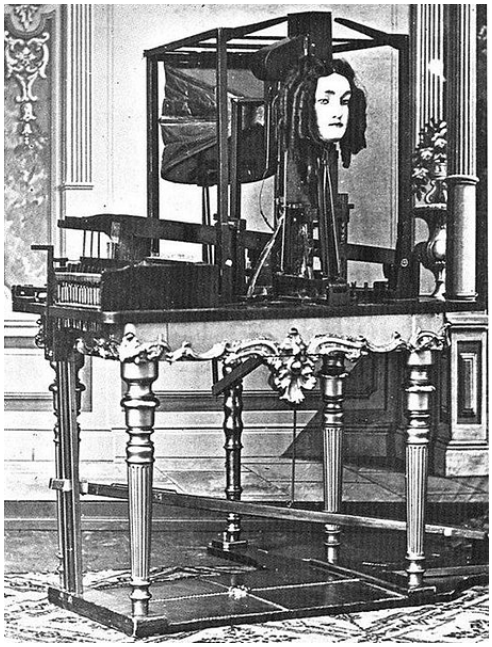


Figure 2.3: Picture of Joseph Faber's Euphonia. Source: Unknown.

2.2 Electric and Electronic era

In 1922, John Q. Stewart, a Physicist from Princeton, published an article named "An Electrical Analogue of the Vocal Organs". It reported an electrical circuit with two resonanant branches. The device could vaguely simulate how the vocal tract worked solely for specific vowels by adjusting the elements in the resonant branches of the circuit. It was able to comprise the first two formants of speech. It did not work either for consonants or connected utterances. This was the pioneering approach to synthesizing speech with only electrical devices (ROSENHOUSE, 2010). The designed circuit is shown in Figure 2.4.

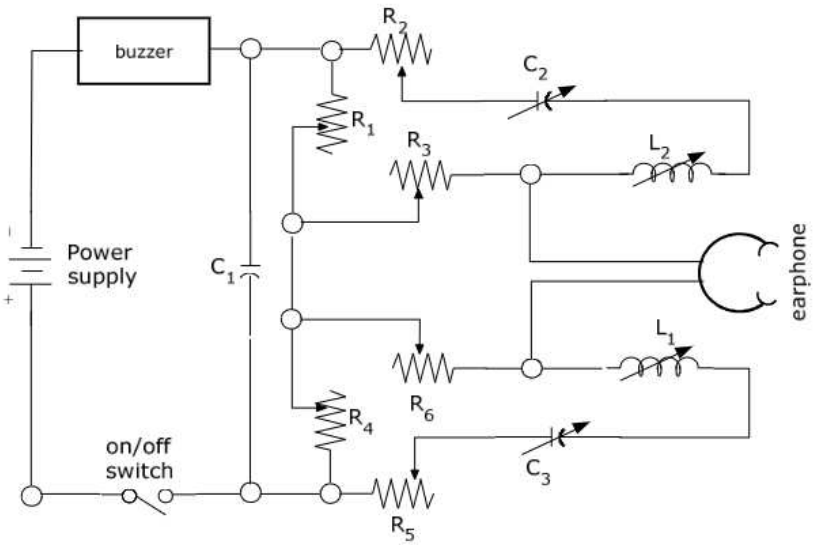


Figure 2.4: A Scheme of the Electrical Synthesizer designed by Stewart. Source: Extracted from (STEWART, 1922).

Also, Stewart reported that, due to the ease of making quick adjustments in the circuit (turning knobs or moving sliders), diphthongs were also able to be reproduced (STEWART, 1922). Even though the title mentioned so, it did not emulate organs, but rather the acoustic resonances it produced. Thus, it is considered the first electrical formant synthesizer, even though he never referred to the system directly as a synthesizer (STORY, 2019). In his paper, he also brought up the idea that rules on manipulating the apparatus were the central core challenge of synthesizing speech with more naturalness, rather than the physical devices that produced the sound (STEWART, 1922).

Following the development of the Voice Coder (VOCODER), a machine able to codify the low-frequency speech articulators factors into a carrier and able to transmit it through low-bandwidth wires, in 1936 by a communications engineer named Homer Dudley working at the Bell Labs, a new speech synthesizer named Voice Operation Demonstrator (VODER) derived directly from the VOCODER was invented. Instead of receiving speech as an input and processing it to obtain its low-frequency modulating aspect, which happened on the VOCODER, Dudley shifted to the use of manual controls to directly modulate a carrier: a 10-key keyboard, which controlled the amplitude of the periodic or noise-like sources; a wrist bar, which controlled a random noise source for the switch of unvoiced segments; and a foot pedal, which controlled an oscillator to provide a periodic voice source for the voiced parts of speech (STORY, 2019). The original VODER schematic, demonstrating how it works (DUDLEY, 1940), and a picture of a demonstrator are shown in Figure 2.5.

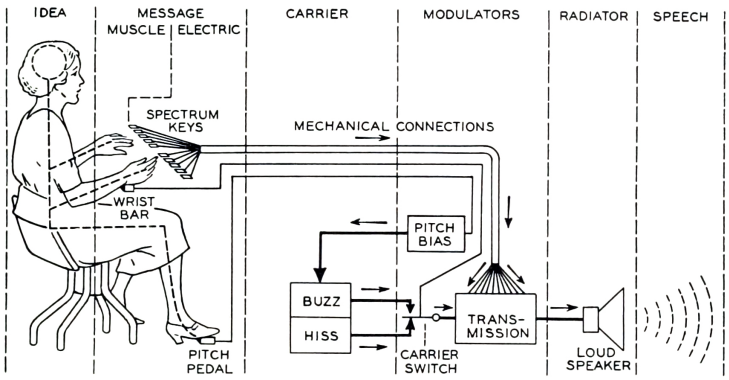
Remarkably, The VODER is an speech synthesis approach that was thoroughly mathematically described: the periodic carrier wave signal to simulate voiced speech, C_v, used by Dudley, is defined as (DUDLEY, 1940):

$$C_v = \sum_{k=1}^{n} A_k \cos\left[kF_0 t + \theta_k\right], \tag{2.1}$$

in which n is the total number of harmonics and k is the specific harmonic considered in the sum. Three message functions modulate the carrier. The first correspond to the effect of starting and stopping the carrier, denoted by $s(t)$. The second effect is the instantaneous varying of F_0, modulated by the inflecting factor $p(t)$, and the third is the transmitting factor $r(w,t)$, to account for the effect of selective transmission. These factors alter the carrier signal, C_v, resulting in the following modulated speech voiced signal:

$$S_v = s(t) \sum_{k=1}^{n} r(\omega, t) A_k \cos\left[kF_0 \int_0^t p(t)dt + \theta_k\right] \tag{2.2}$$

In this formulation, the unvoiced carrier is also contained as a degenerate case of the voiced carrier. On unvoiced speech, $F_0 \to 0$, and $n \to \infty$, thus, after some



(a)

(b)

Figure 2.5: (a) Schematic of the VODER speech synthesizer. (b) Picture of Woman demonstrating the VODER operation. Source: (a) Extracted from (DUDLEY, 1940). (b) Source: Unknown.

manipulation, the unvoiced carrier is given by:

$$C_{uv} = \int_{\omega_1}^{\omega_2} A(\omega) \cos\left[\omega t + \theta(\omega)\right] d\omega, \tag{2.3}$$

in which $[\omega_1, \omega_2]$ is the output frequency range. The unvoiced speech signal, thus, did not have an F_0 inflecting factor, but had both the transmitting and start-stop effect, denoted by the following equation:

$$S_{uv} = s(t) \int_{\omega_1}^{\omega} r(\omega, t) A(\omega) \cos\left[\omega t + \theta(\omega)\right] d\omega. \tag{2.4}$$

Some years later, in 1945, on the paper named "Visible Patterns of Sound", Ralph Potter, also working at Bell Labs, reported the creation of the "sound spectrograph", a device that could graphically represent sounds in 2D visual representations, with time in the X-axis and frequency in the Y-axis, the so-called spectrograms (POTTER, 1945). Inspired by this machine, in 1951, the researchers Frank Cooper and Alvin Libermann

at the Haskins Labs reported the design of a new speech and sound synthesis machine: the "Pattern Playback". It is shown in Figure 2.6. This machine worked by literally converting a drawn spectrogram into a sound wave. Its name was given since it could reproduce either a modification or an existing spectrogram. It worked with a light source that passed through a tone wheel with 50 sound tracks that modulated light into harmonic frequencies of a given F_0 (from 120-6000Hz). This light was then projected through an acetate spectrogram in black and transparent that filtered out frequencies not present on the spectrogram. The filtered-out signal was then converted to an electric signal via a photocell, which was in the end sent to an amplifier and transformed into a sound wave (STORY, 2019).

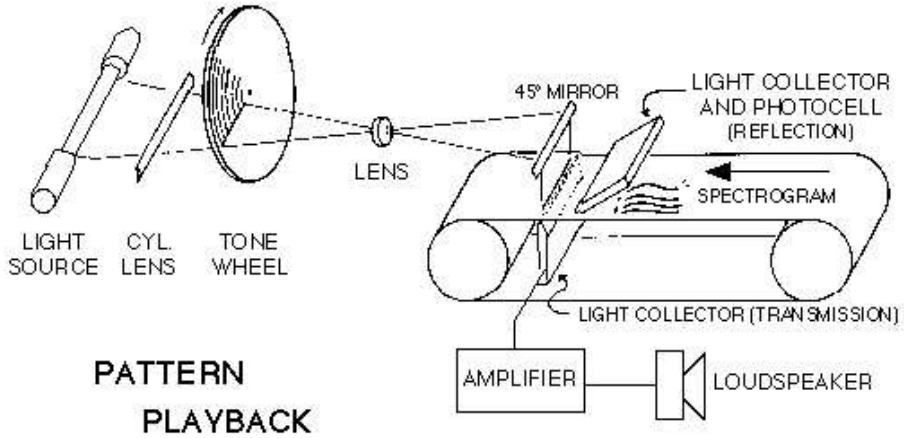


Figure 2.6: The Pattern Playback machine. Source: Extracted from (COOPER *et al.*, 1951).

The Pattern Playback was the first speech synthesizer to be experimented on a large scale regarding speech structure. Its users became good at drawing spectrograms by hand and started developing rules for speech production. Formally, rules for generating utterances were described by Frances Ingmann in 1957. This was the first time explicit rules for generating speech with a synthesizer were formally documented (INGEMANN, 1957).

Other types of synthesizers were also developed during this period. In 1953, an English researcher, Walter Lawrence, introduced a speech synthesizer named Parametric Artificial Talker (PAT). It consisted of an electrical circuit composed of a source generator and three parallel frequency-controlled resonant branches. Also, Gunnar Fant, at the Royal Institute of Technology (KTH) in Stockholm, experimented with placing the electrical resonators in a cascade arrangement and named the synthesizer as Orator Verbis Electris (OVE) I. It was a vowel synthesizer that featured a unique mechanical stylus moved in a 2D plane for controlling the first two resonance frequencies and, thus, the two first formants. Later, he developed OVE II, featuring more enhancements, such as the production of nasal,

stops, and fricatives. These synthesizers belonged to a category of formant synthesizers since they were basically devices to control formants seen in spectrograms (STORY, 2019).

Another type of synthesis that was simultaneously in development was synthesizers that tried to emulate the shape of the vocal tract with electrical circuits. In 1950, H. K. Dunn, another Bell Labs engineer, designed a circuit using electrical components that emulated pharyngeal and oral air cavities within the vocal tract. The values of capacitors, inductors, and resistors were similar to the cross-sectional area of the cavities. However, this work and even more detailed emulations of the vocal tract proposed by Fant were only based on static vocal tract configurations (STEVENS *et al.*, 1953). Then, in 1958, with a more complex circuit and switch, George Rosen, a doctoral student at the Massachusetts Institute of Technology (MIT), proposed a new speech synthesizer, known as Dynamic Analog of the Vocal Tract (DAVO), that could change vocal tract configurations (ROSEN, 1958). Even though it did not produce sentences, it could generate diphthongs and consonant-vowel sounds. Thus, for these parametric systems, specifying the time-dependence of the vocal tract parameters to make them generate speech was a problem (STORY, 2019). A Figure of DAVO is shown in Figure 2.7.

One other type of speech synthesizer was being developed at the same time. Due to the advancement in recording audio technology, Cyril M. Harris designed a system that synthesized pieced-together tape segments of vowels and consonants using selector circuits. It was perceived as intelligible, but not natural, due to the discontinuities between pieces (HARRIS, 1953). Later, other segmentation techniques were considered, using instead of vowels and consonants, a unit called “dyad”, segments extending in time from the steady-state location of one phoneme to the next, which preserved acoustic transition dynamics between phonemes (STORY, 2019).

2.3 Digital and Computational era

With the advent of the computing age, the difficulty in controlling speech synthesizers was mitigated. From that moment on, with the capability of digital computers, commands inputted on a keyboard could be translated to parameter changes in the analog electrical circuitry. This facilitated the concept of speech synthesis by rule. With a set of predefined rules, symbols representing phonetic elements were converted into temporal variations of the parameters of a synthesizer. For example, specific initial and final values of a given formant could be inputted, and, during a given period of transition, an interpolation from one to another would happen, possibly independently of other formants (HOLMES *et al.*, 1964; KELLY, 1962). These models worked as a digital version of the DAVO synthesizer. In some cases, even the perceived bad sound quality of the generated speech perceived by the authors themselves was attributed to insufficient knowledge of the cross-sectional vocal tract areas corresponding to the target of the phoneme inputs (STORY, 2019).

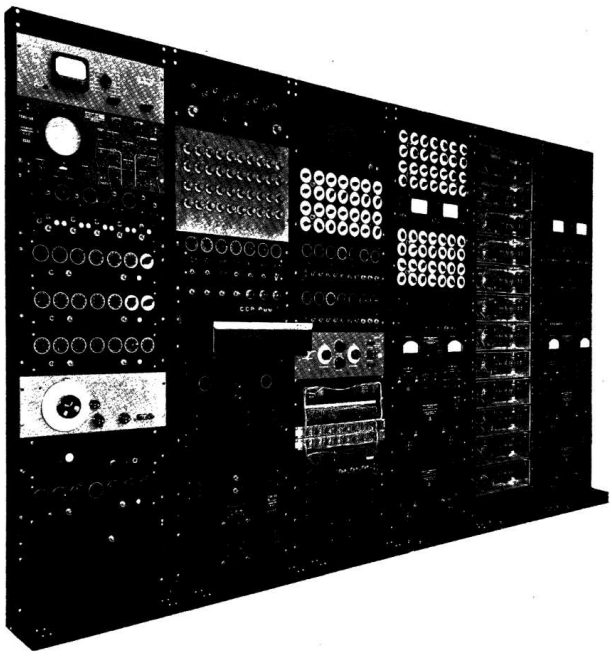
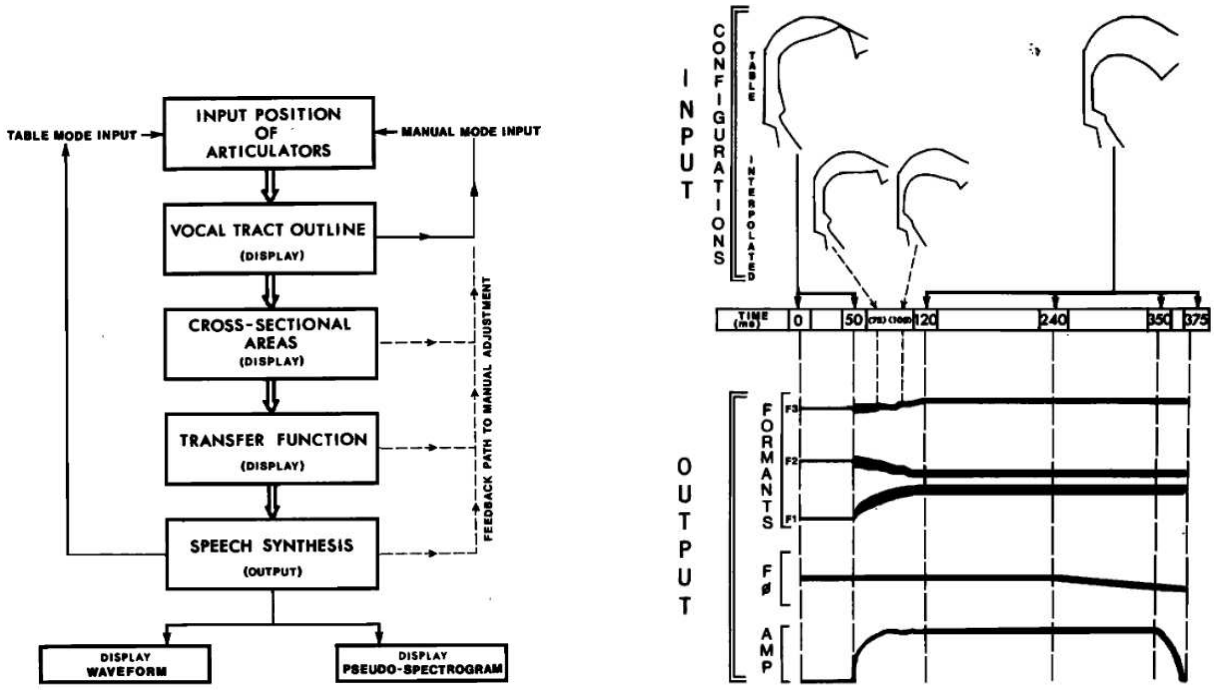


Figure 2.7: DAVO Synthesizer. Source: Unknown.

With the improvement of X-ray cineradiography technology in the 1960s, the articulatory movements of speech in a sagittal projection image could be better studied. Thus, this boosted a new type of synthesis paradigm called articulatory synthesis. Several speech synthesizers were developed during this period based on a computational model of the human speech articulators. Positions of the tongue, lips, jaw, and larynx, represented in the midsagittal plane, were specified and could be moving according to a given function varying in time (LINDBLOM; SUNDBERG, 1971; HEINZ; STEVENS, 1964). An example of the earliest articulatory synthesizer that was used in large-scale phonetic experiments was the Articulatory Synthesizer (ASY), developed in the Haskins laboratory and enhanced then to become the Configurable Articulatory Synthesizer (CASY) (Haskings Configurable Articulatory Synthesizer), which provided more accurate representations of the vocal track and flexibility in control (RUBIN *et al.*, 1981; RUBIN *et al.*, 1996). A figure of the schematics of the ASY working, and an example speech is shown in figure 2.8.

Concurrently with the development of articulatory synthesizers, research also aimed at enhancing the formant-based models. Dennis Klatt and colleagues started to develop several rule-based formant synthesizers based on his research on digital resonators. Some examples are the "Klattalk", "MITalk", "DecTalk", and later "KLSYN88" (KLATT, 1982). These models became very well known because of the use of these synthesizers voices by the British physicist Stephen Hawkings.

Also, other techniques of synthesizing speech were being developed in the digital era. A method named "unit selection" or "concatenative speech synthesis" was developed. It was the equivalent to a digital version of the tape-slicing technique used by Harris (1953). It consisted of building speech signals from a database containing several hours of



(a) (b)

Figure 2.8: a) Block Diagram of the ASY. Source: Extracted from (RUBIN *et al.*, 1981). b) ASY Synthesis Scheme. Source: Extracted from (RUBIN *et al.*, 1981).

recordings using algorithms that could efficiently search the optimized segments (STORY, 2019). Two main different concatenative schemes were considered: Linear Predictive Coefficients (LPC) (ATAL; HANAUER, 1971) and Pitch Synchronous OverLap Add (PSOLA) (MOULINES; CHARPENTIER, 1990). The LPC method uses the LPC speech codification algorithm to reduce the size occupied by the speech signal into time-varying parameters related to a transfer function of the vocal tract. The synthesis is performed through a decodification of the coefficients and concatenation process (ATAL; HANAUER, 1971). Since speaking occurs not simply by concatenating sources, the overall output speech will suffer from artifacts in the concatenation points. To tackle this issue, the PSOLA algorithm can adjust the prosody of the concatenated unit given the context. The adjustment could be made in both the time and frequency domain to modify spectral characteristics of speech (MOULINES; CHARPENTIER, 1990).

A different and more recent technique denominated parametric speech synthesis was based on establishing parametric representations from spectral features of recordings to reconstruct a speech segment (STORY, 2019). This technique uses digital signal processing techniques to synthesize speech (NING *et al.*, 2019a). It allowed improved flexibility regarding voice characteristics and style but provided a lower general speech quality (STORY, 2019). Statistical Parametric Speech Synthesis (SPSS) are parametric techniques that are based on statistical models, such as Hidden Markov Models (HMM) or Gaussian Mixture Models (GMM), to generate speech (ZEN *et al.*, 2007). Usually, it is

divided into two phases, shown in Figure 2.9: training, on which the statistical model is trained on the extracted acoustic feature parameters as well as the text, and the synthesis phase, on which the acoustic features are predicted based on the corpus learned statistics guided by the input linguistic features (NING *et al.*, 2019a).

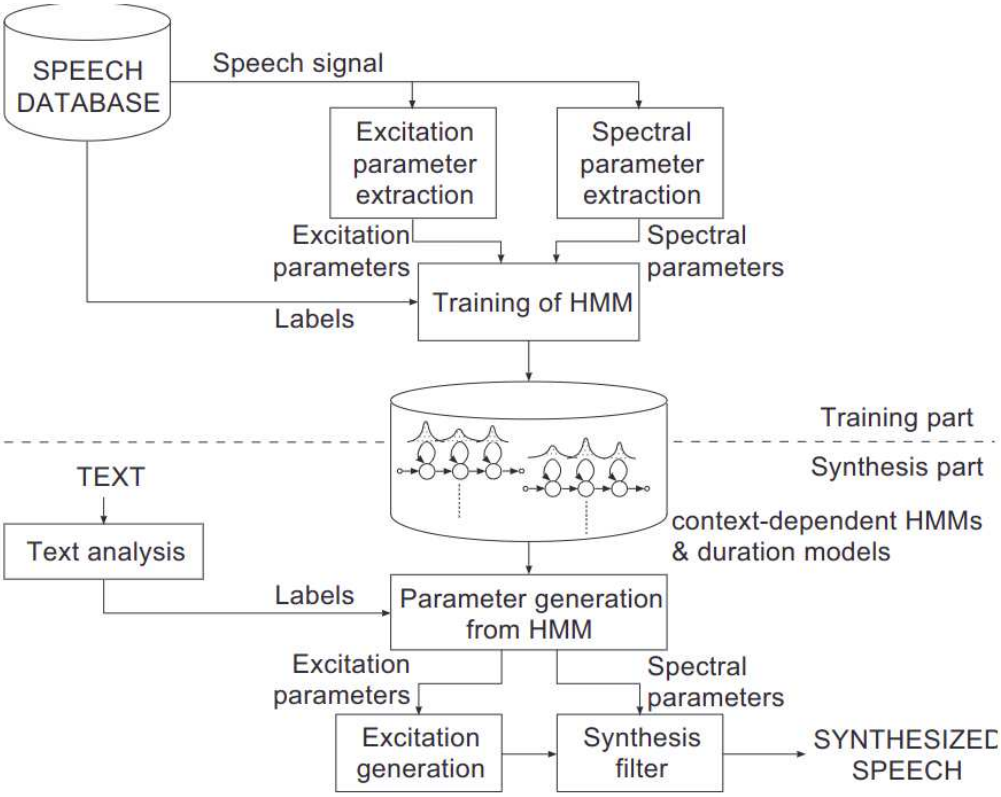


Figure 2.9: A generic training scheme of HMM-based speech synthesis models. Source: Extracted from (ZEN *et al.*, 2007).

Speech synthesis' methods based on HMM remained state-of-the-art and the most widely adopted in the field until the mid 2010s (NGUYEN; PHUNG, 2017). With the advent of improvements in neural networks' algorithms and the upgrade in computational hardware, training deep neural networks became more efficient (HINTON *et al.*, 2012). This allowed for neural networks, that even though were already exploited in the 1990s in the field, to be "rediscovered" with more layers, computational resources, and training data, and started to replace the HMMs, in the late 2010s, due to their higher capability of acoustic modeling (WU *et al.*, 2016). In fact, neural networks were already predicted by many authors to completely replace HMM-based acoustic models in the early 2010s (HINTON *et al.*, 2012; ZEN, 2015). Currently, neural models have indeed not only replaced the acoustic modeling module that was dominated before by the HMM, but also all other modules in the entire traditional speech synthesis pipeline (RAO *et al.*, 2015; OORD *et al.*, 2016). Furthermore, with the evolution of generative deep learning, powerful neural network-based fully end-to-end speech synthesis models emerged. These networks are capable of directly mapping input text to speech signals (SOTELO *et al.*, 2017; JUNG;

LEE, 2023).

2.4 Literature Summary

This chapter presented a historical perspective regarding the approaches to the generation of synthetic speech. Discoveries and techniques developed into three different societal technological eras were exposed. Table 2.1 presents a summary of some notable events discussed in the previous section that occurred throughout the development of the speech synthesis field. The milestone, year and the inventor are all listed.

Year(s)	**Milestone**
1779	Resonators for vowel production (Christian Krantzenstein)
1791	Mechanical talking machine (W. Von Kempelen)
1857	Euphonia (J. Faber)
1862	Mechanical Fourier Analyzer (Helmhotz)
1922	Electronic speech synthesizer (J. Q. Stewart)
1936	VOCODER (Bell Labs)
1939	VODER (H. W. Dudley, Bell Labs)
1950	Static articulatory model (H.K.Dunn)
1951	Pattern playback synthesizer (F. Copper)
1953	OVE I (G. Fant), OVE II, PAT (W. Lawrence)
1958	DAVO (G. Rosen); Concatenate synthesis (Cyril M. Harris)
1967	LPC (B.S. Atal)
1981	Klattalk, Mitalk (Dennis Klatt); ASY (P.E. Rubin, Haskins Labs)
1985	PSOLA (E. Moulines, F. Charpentier)
1996	HMM-based Speech Synthesis (T. Masuko et. al)

Table 2.1: A summary of the principal techniques in speech synthesis' history before neural networks. Source: Adapted from (ROSENHOUSE, 2010).

Chapter 3

Basic Concepts and Related Works

This chapter presents fundamental concepts and works related to speech synthesis, highlighting the core works that aim to improve expressiveness.

Section 3.1 introduces concepts related to the more recent and successful speech synthesis approaches based on neural networks. Section 3.2 presents a notion of how expressiveness is addressed and perceived in speech; how it is introduced in TTS systems and, more specifically, works with the objective of performing cross-speaker style transfer from the perspectives of style disentanglement and data augmentation. Section 3.3 presents the concluding remarks of the chapter.

3.1 Neural Speech Synthesis

The approaches used in speech synthesis encompass various fields such as acoustics, linguistics, digital signal processing, and statistics (NING *et al.*, 2019b). This multidisciplinary nature paved the way for the integration of neural models, which were first employed in the 1990s through shallow neural networks (less than 4 hidden layers), marking an early intersection between neural computation and speech synthesis technology. These models were used as substitutes for the rule-based or concatenative systems on phoneme-to-acoustic mapping task (KARAALI *et al.*, 1996). Only in the 2010s, with advancements in the performance of parallel computing hardware (Graphics Processing Unit (GPU)), that Deep Neural Networks (DNN) could be efficiently trained (ZEN *et al.*, 2013). This development enabled deep neural models to achieve large improvements over current state-of-the-art conventional approaches in various different tasks that involved finding and modeling underlying complex patterns in data (KRIZHEVSKY *et al.*, 2012). Before applied in the synthesis of speech, DNN have been already exploited in other speech domains, such as in Automatic Speech Recognition (ASR) (HINTON *et al.*, 2012) and acoustic-articulatory inversion mapping (URIA *et al.*, 2012).

The first approaches to use DNN for speech synthesis was in the context of SPSS models. Zen *et al.* (2013) propose the replacement of the current conventional approach of

using decision trees to perform the mapping from linguistic contexts obtained from text to probability densities of a set of speech-related parameters in HMM-based speech synthesis with DNN. An improvement in performance was shown with objective and subjective metrics, in which the preference scores almost double when compared to the HMM systems.

Due to the inherent different modalities between and the significant sequence length mismatch between text and speech, the speech synthesis process is commonly divided even before the advent of neural networks, into three main component models: a text analysis module, an acoustic module, and a VOCODER. After being applied to SPSS, neural modules started to be used in every other components in the TTS pipeline. The text analysis module is responsible for converting input characters into linguistic features; the acoustic model converts the linguistic features into acoustic features; and the VOCODER converts the acoustic features into the output speech waveform. Also, models denominated End-to-end (E2E) were proposed to generate waveforms directly from text or linguistic features (TAN *et al.*, 2021). A detailed graph representing these transformations, the various features used, and the neural models developed is shown in Figure 3.1. In the following sections, we discuss neural approaches adopted in the three main components of the TTS pipeline and in E2E models.

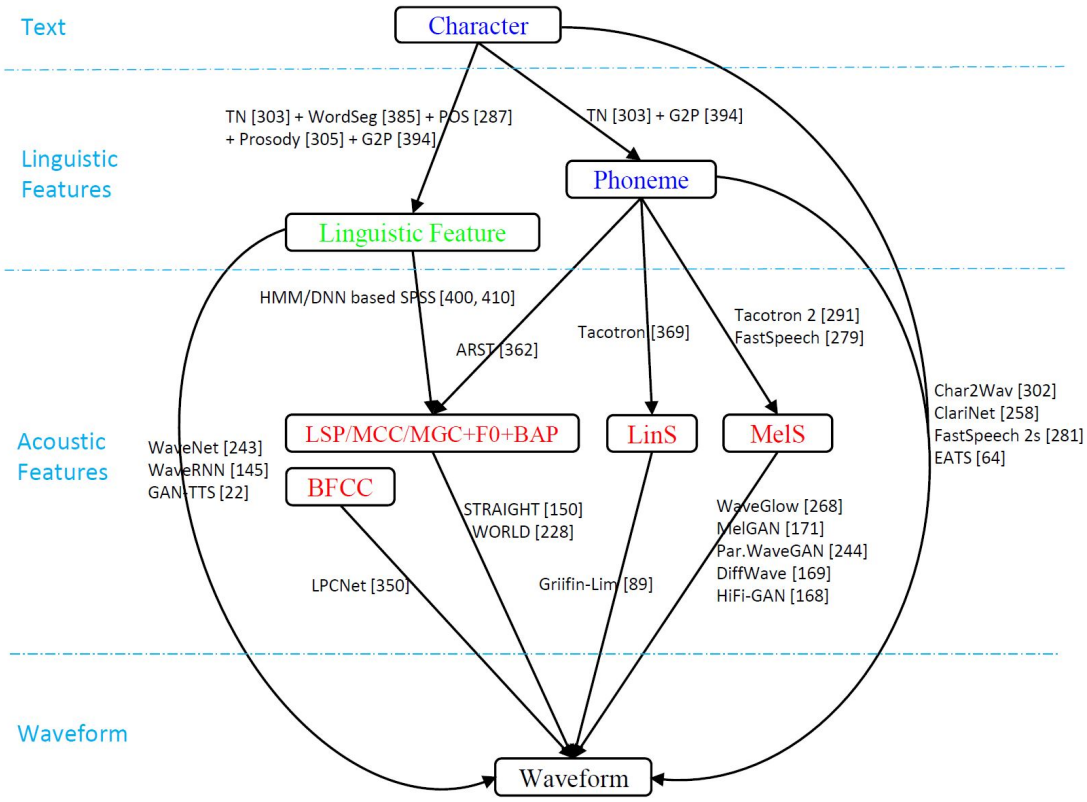


Figure 3.1: Data transformation on a speech synthesis pipeline entirely based on neural models. Source: Extracted from (TAN *et al.*, 2021).

3.1.1 Textual Analysis

Text analysis in speech synthesis consists in some practices that aim to transform and extract rich linguistic features from text to facilitate the overall synthesis process. Some typical tasks in the textual analysis process are (TAN *et al.*, 2021):

- Text Normalization: It consists of normalizing non-standard words that appear in the raw written text, facilitating the pronunciation of the synthesis process. For example, the year "2001" would be converted into "two thousand and one", elucidating how it is pronounced for the model. Earlier works on text normalization approaches were rule-based. Now, approaches based on neural networks also tackle this task.
- Word Segmentation: The process of detecting the word boundaries. Important to other tasks that require the attribute label of features in a word-level granularity.
- Part-of-speech (POS) Tagging: Consists of labeling each word with its function, such as noun, verb, preposition, etc.
- Prosody Prediction: The extraction of prosodic tags to label various prosodic items, for example, pitch accents, phrase accents, and boundary tones.
- Grapheme-to-phoneme (G2P) Conversion: Consists in the conversion of characters (grapheme) into units that represent pronunciation (phonemes). This process is able to further explicit the relation between text and speech, by dealing with out-of-vocabulary words or even for polyphone disambiguation, deciding the appropriate pronunciation according to context.

3.1.2 Acoustic Models

In the standard speech synthesis pipeline, Acoustic models are responsible for converting linguistic features into acoustic features. The use of DNN to perform this function brought several advantages compared to SPSS. First, the neural models can implicitly learn alignments between the linguistic and acoustic features through attention or direct prediction. Also, with the great modeling capacity of neural networks, the acoustic representations have evolved from compressed coefficients, to high-dimensional mel-spectrograms, a visual representation of the variation on intensity of each frequency in the mel scale (a frequency scale judged by human listeners that takes into account the perceived spacing of frequencies) present on a signal over time, providing more overall information and acoustic detail.

In the usual training scheme of the acoustic models, a pair data containing text and the corresponding mel-spectrogram of speech are passed. A forward pass is realized with the linguistic features and then the synthetic mel-spectrogram is compared with the

real one through a loss function (usually Mean Squared Error (MSE) or Mean Absolute Error (MAE)). The alignment can be either learned (BADLANI *et al.*, 2022b; SHEN *et al.*, 2018) or predicted (ŁAŃCUCKI, 2021), when ground-truth alignments are available.

For these models, sequence-to-sequence type architectures used in the literature are varied. There are text-to-speech models based on several deep learning structures, such as Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and Transformer (VASWANI *et al.*, 2017) type architectures, based on the attention mechanism. There are also text-to-speech systems that incorporate various generative models, such as those based on Generative Adversarial Networks (GAN) (BIńKOWSKI *et al.*, 2019), Variational Auto-Encoder (VAE) (HSU *et al.*, 2019), normalizing flows (KIM *et al.*, 2020), diffusion models (JEONG *et al.*, 2021), and flow matching (KIM *et al.*, 2024).

Another dimension of these acoustic neural models concerns their autoregressiveness. These systems use spectrogram frames generated in previous steps as context to synthesize the current frame. Auto-regressive models were reported not only to be prone to word skipping and repetition problems but also to have slow inference speed, due to its sequential nature To mitigate these issues, non-autoregressive (parallel) models have been proposed. These provided both faster parallel synthesis and greater robustness in relation to that occurred in autoregressive ones (TAN *et al.*, 2021). Some representative acoustic models of diverse characteristics are shown in Table 3.1.

Table 3.1: Representative TTS acoustic models and their structure characteristics. "Ph" means phonemes, "Ch" means characters, "AR/NAR" means autoregressive or not.

Model	**In → Out**	**AR/NAR**	**Structure**	**Reference**
Tacotron 2	Ch→ mel-spec.	AR	RNN	Shen *et al.* (2018)
DeepVoice 3	Ch/Ph→ mel-spec.	AR	CNN	Ping *et al.* (2018)
TransformerTTS	Ph→ mel-spec.	AR	Self-Attention	Li *et al.* (2019)
FastSpeech 2	Ph→ mel-spec.	NAR	Self-Attention	Ren *et al.* (2021)
FastPitch	Ph→ mel-spec.	NAR	Self-Attention	ŁaŃcucki (2021)
Glow-TTS	Ph→ mel-spec.	NAR	Hybrid/Self-Att/CNN	Kim *et al.* (2020)
Diff-TTS	Ph→ mel-spec.	NAR[1]	Hybrid/CNN	Jeong *et al.* (2021)

FastPitch

Fastpitch (ŁAŃCUCKI, 2021) is a non-autoregressive Transformer architecture-based neural acoustic model. Due to its parallel synthesis, it has the ability to generate spectrograms up to 911 times faster than real time (1 ms) on an NVIDIA A100 GPU, while maintaining performance comparable to SOTA autoregressive TTS systems.

The Fastpitch architecture is shown in Figure 3.2. In the synthesis process, the model receives as input a sequence of characters or phonemes that are encoded in an

[1]Although they are not exactly autoregressive, diffusion models contain iterative structures similar to autoregressive ones that also induce delay in the synthesis process.

embedding through a look-up table. Then, these embeddings are handled by an Feed Forward Transformer (FFT) block, which consists of a sequence of several blocks identical to the Transformer encoder (VASWANI *et al.*, 2017), replacing only the two fully connected layers of the original architecture, with two layers of 1D convolutions with Rectified Linear Unit (ReLU) activation, since, in speech, locally close information is strongly related.

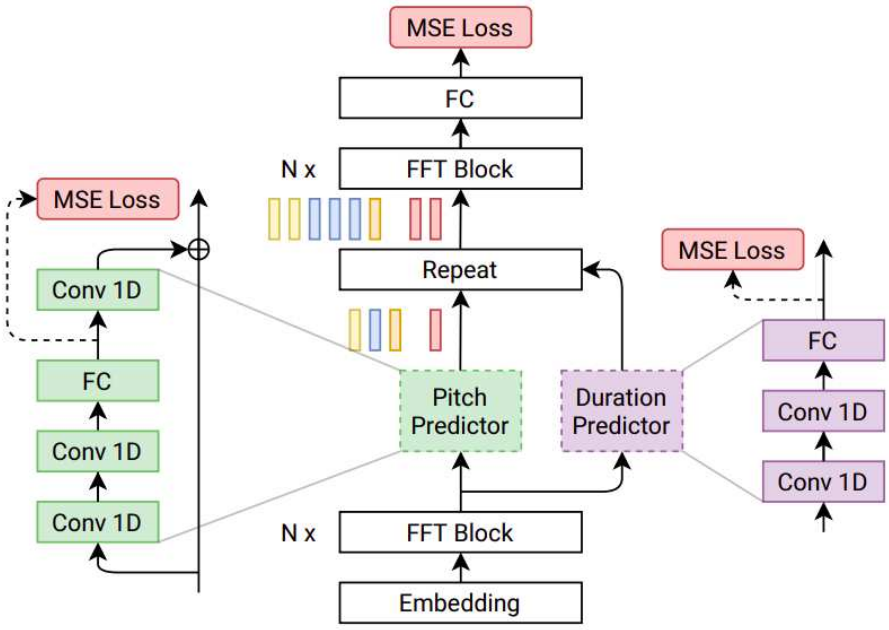


Figure 3.2: *Fastpitch* architecture. Source: Extracted from (ŁAŃCUCKI, 2021).

The intermediate representations generated by the FFT are used to make predictions regarding the F_0 curve and duration for each input symbol. The F_0 predictor is composed of a CNN that receives the representations and then computes an MSE with the real F_0 curve extracted from the speech. Then, the F_0 curve is projected to adjust to the size of the hidden representations, to which they are added. During training, the real F_0 curve is used, and in the inference process, the predicted one is used.

Similarly, intermediate representations are also used to predict the duration of each input character. The actual durations are used during training and the predicted ones in inference. From them, an up-sampling procedure is made on the representations resulting from the output of the tone predictor. This process consists of aligning the sequence of characters with the sequence of mel-spectrogram frames, repeating each symbol representation according to its respective predicted duration.

Finally, another FFT block together with a projection layer are responsible for transforming the aligned intermediate representations into a mel-spectrogram. The complete training loss of the model is given by:

$$\mathcal{L} = \|\hat{y} - y\|_2^2 + \alpha\|\hat{p} - p\|_2^2 + \gamma\|\hat{d} - d\|_2^2, \tag{3.1}$$

in which y represents the mel-spectrogram, p the pitch, and d the duration. α and γ are

hyperparameters. Due to this prosodic conditioning, direct control of speech intonation on inference is made possible, allowing, for example, to lower or increase the tone and duration for each phoneme separately. To obtain ground-truth durations, the authors use a Tacotron 2 (SHEN *et al.*, 2018) trained model.

It was shown that to extend a standard TTS scenario to a multi-speaker, each speaker could be modeled with a global embedding across a look-up table of speakers (ARK *et al.*, 2017). A multi-speaker base is introduced into the training, such that at each step the following inputs are used: text, the corresponding mel-spectrogram, and also a speaker embedding that is added to each character embedding, introducing the timbre information of the specific speaker (ŁAŃCUCKI, 2021) .

3.1.3 Vocoders

VOCODERs are models that convert acoustic representations of speech, such as mel-cepstral coefficients, band aperiodicity, F_0, or mel-spectrograms, into audio waveforms. They can be roughly divided into VOCODERs used in SPSS (KAWAHARA, 2006; MORISE *et al.*, 2016) and then neural network-based ones. Similar to the acoustic models, there exists a vast diversity of VOCODERs with varying types of architecture, types of input feature and whether it is autorregressive or not. Notably, since speech waveforms are sampled at very high rates, and thus are very long, autorregressive VOCODER models, such as WaveNet (OORD *et al.*, 2016), (which is considered to be the first modern, entire deep neural network based speech synthesis model), tend to take too much inference time (TAN *et al.*, 2021).

BigVGAN

BigVGAN (LEE *et al.*, 2023) is an open-source neural model that was proposed to be a universal VOCODER model. It was designed to achieve high-fidelity waveforms on various zero-shot, Out-of-Distribution (OOD), audio conditions, such as unseen speakers, languages, recording environments, singing voices, music and even instrumental audio. A base version of the BigVGAN, with 14M parameter was shown to outperform comparable size SOTA models, and a large-scale version, with 112M parameter was shown to outperform by a large margin SOTA vocoders by a large margin for both in-distribution and OOD samples.

The model was trained on a standard GAN-based framework, very similar to (KONG *et al.*, 2020), with a generator that receives mel-spectrograms and outputs waveforms, and several discriminators to detect if the generated waveform is real or synthesized. Motivated by the intrinsic periodic nature of waveforms (can be represented as a composition of periodic primitives), the authors replace usual activation functions, such as LeakyReLU, with periodic activations, in the form of so-called “Snake” functions (ZIYIN *et*

al., 2020), defined as $f_\alpha(x) = x + \sin^2(\alpha x)/\alpha$, in which α is a trainable parameter, into the generator as means of inducing a strong periodic bias to audio synthesis. Since the Snake activations can produce arbitrary high-frequency continuous-time signals that cannot be represented by the discrete output of the network, aliasing (false lower frequencies in the signal caused by sampling rates lower than twice the Nyquist frequency) can happen. To mitigate this issue, before passing in the "Snake" activation, the authors up-sample the signal by a factor of two, then, after the activation, they down-sample to the original rate and use a low-pass filter to eliminate the high-frequency content.

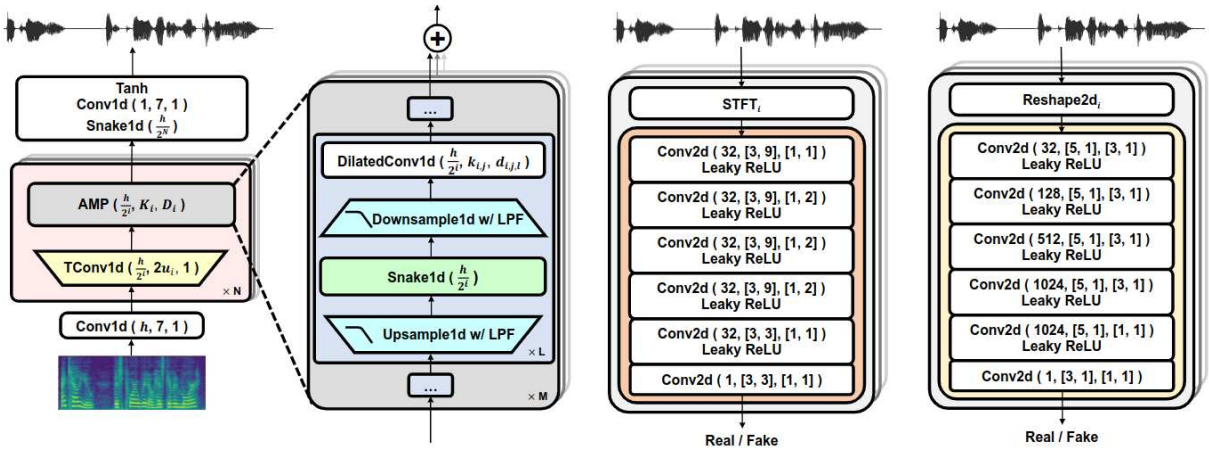


Figure 3.3: Architectures of the generator (left) and both discriminator types (right) of the BigVGan architecture. Source: Extracted from (LEE *et al.*, 2023).

Thus, the training loss of the generator and the ensemble of discriminators is given by:

$$\mathcal{L}_G = \sum_{k=1}^{K} \left[\mathcal{L}_{\text{adv}}(G; D_k) + \lambda_{fm}\mathcal{L}_{fm}(G; D_k)\right] + \lambda_{mel}\mathcal{L}_{mel}(G), \tag{3.2}$$

$$\mathcal{L}_D = \sum_{k=1}^{K} \left[\mathcal{L}_{\text{adv}}(D_k; G)\right], \tag{3.3}$$

in which G is the generator, D_k is the k-th discriminator module, and $\mathcal{L}_{\text{adv}}$ is the least-square GAN loss, defined as:

$$\mathcal{L}_{\text{adv}}(G; D_k) = \mathbb{E}_s\left[(D_k(G(s)) - 1)^2\right], \tag{3.4}$$

$$\mathcal{L}_{\text{adv}}(D_k; G) = \mathbb{E}_{(x,s)}\left[(D_k(x) - 1)^2 + (D_k(G(s)))^2\right], \tag{3.5}$$

where x is the ground-truth waveform, and s is the input mel-spectrogram. As in Kong *et al.* (2020), a feature matching loss $\mathcal{L}_{fm}$ is used to minimize the difference in the features of the discriminator between a generated and ground-truth sample. It is given as:

$$\mathcal{L}_{fm}(G; D_k) = \mathbb{E}_{(x,s)} \left[\frac{1}{T} \sum_{i=1}^{T} \frac{1}{N} \left\| D_k^i(x) - D_k^i(G(s)) \right\|_1 \right], \tag{3.6}$$

in which T is the number of layers of the discriminator D_k. A simple mel-spectrogram ℓ_1 loss is also added in the generator loss function, given by:

$$\mathcal{L}_{mel}(G) = \mathbb{E}_{(x,s)} \left[\|\phi(x) - \phi(G(s))\|_1 \right]. \tag{3.7}$$

The architecture of each component of the BigVGan (LEE *et al.*, 2023) is shown on Figure 3.3. Two types of discriminators are considered: an Multi-Period Discriminator (MPD), and a Multi-Resolution Discriminator (MRD). The first converts the 1D signal into 2D representations to capture the multiple periodic structure with 2D convolutions. The second operates on a 2D linear spectrogram of the output waveform. For each type, several discriminators are considered, varying Short-Time Fourier Transform (STFT) parameers for the MRD, and reshaping widths for the MPD.

3.1.4 End-to-end Models

Fully E2E neural models are able to generate waveforms directly from textual representations (characters or phonemes). Some inherent advantages of these models are: they require less feature engineering and annotation; avoid of error propagation that can happen in cascaded models; and their overall procedure can reduce the training, development and deployment costs (TAN *et al.*, 2021). On the other hand, the training of E2E models is rather difficult, since it has to directly learn the complex mapping between speech and text, two sequences with a huge length mismatch (TAN *et al.*, 2021).

With a greater flexibility on intermediate features engineering, E2E models are reaching new paradigms. The VALL-E model, proposed by Wang *et al.* (2023) model is the first to replace the traditional mel-spectrogram based approach with quantized neural audio codec codes as intermediate representations, not using any acoustic signal representation. Also, a speaker prompt is used to indicate the desired timbre, thus performing zero-shot synthesis. For the first time, with this approach, the authors substitute the traditional mel reconstruction loss with a language modeling based objective function and report SOTA results for both naturalness and speaker similarity. On the other hand, Kim *et al.* (2024) also proposed an approach based on a speech-prompt, which showed improvements over the VALL-E, by training a flow matching based model on the usual mel-spectrogram representations.

As a consequence of the rapid improvement of deep learning techniques, text-to-speech systems began to obtain greater voice quality in terms of intelligibility and naturalness, while requiring less manual pre-processing and feature engineering, when compared to previous existing concatenative and SPSS SOTA techniques of synthesizing

Table 3.2: Representative E2E models and their characteristics.

Model	One-Stage	AR/NAR	Modeling	Reference
Char2Wav	N	AR	Seq2Seq	Sotelo *et al.* (2017)
Fastspeech2s	Y	NAR	GAN	Ren *et al.* (2021)
VITS	Y	NAR	VAE+Flow	Kim *et al.* (2021)
VALL-E	N	AR+NAR	CNN+RVQ	Wang *et al.* (2023)
P-Flow	Y	NAR	Flow Matching	Kim *et al.* (2024)

speech. These models were even shown to exhibit performance compared to original human speech recordings regarding naturalness and intelligibility (SHEN *et al.*, 2018).

3.2 Expressiveness in Speech Synthesis

There are currently numerous interpretations of what the term "expressiveness" means in literature. These vary according to the perspectives taken from the possible different frameworks that could be considered to study the concept. Some examples that reason about expressiveness are functional linguistics, rhetoric, poetics, lexicography, semantics, etc. From the functional linguistic framework, an approach concerned with relating language to the social in a motivated way (MARTIN, 2000), expressiveness is opposed to neutrality and is considered a norm of deviation. This perspective is supported by the notion that expressiveness is "perceived only where and when the conventional ways of communication come to the fore", these are the "features of figurative speech which differentiate it from the conventional neutral speech and make it vivid, figurative and emotive." (APRESYAN, 2018).

The differentiation between expressiveness and emotion is many times overlooked and nebulous when considering works on synthetic speech generation. Although there is no consensus in the literature on a definition for emotion, various works have taken attempts at providing such. Notably, Cabanac (2002) proposed a very broad definition of emotion as "any mental experience with high intensity and high hedonicity". The work argued that emotions could thus be the byproduct of several factors, such as sensation, perception, memory recall, reckoning, and imagination. Thus, in terms of speech, emotionality is responsible to represent this psychological state in which the speaker finds himself in, whereas expressiveness is "a means for the magnification of the communicative function of the utterance" (APRESYAN, 2018).

Plainly, the tendency of humans to express emotions through the tone of voice, posture, facial expression and actions can be referred simply to as emotional expressiveness (KNYAZEV *et al.*, 2012). Consequently, as a type of reflection of the inner self, speech could be used to diagnose the speaker's condition, as clues of the emotional state of the speaker can be obtained with an analysis of their speech. This knowledge can be helpful in various situations, such as emergencies and health care related applications (YAK-

OUMAKI, 2015). Additionally, with the increase adoption of personal digital assistants and other socially interactive agents, expressive speech is also becoming a defining perceived personality of the system in question.

In this context, Székely (2015) states that "expressive speech has the potential to provide the user with the choice to select a nuanced tone of voice suited to their intent and to the communicative setting" but that "in an interactive situation however, this does not become a real possibility, until a functional interaction model is available to control aspects of the expressive synthetic speech to ensure timely and effortless delivery", drawing the importance to the development of expressive synthetic speech systems.

3.2.1 Speaking Styles

Speaking style can be defined as a "differentiation in the way of speaking, such that it constitutes a genre with common characteristics." This vocal aspect that differentiates speaking styles is related to changes in voice quality, speech rhythm, and intonation (IRVINE, 2001; BARBOSA, 2022). Speaking styles are also associated with specific communicative acts (BARBOSA *et al.*, 2017). An example that reiterates that speaking styles can be culturally distinguished from each other by the vocal aspects mentioned above is a study conducted by Obin *et al.* (2011) that showed that four speaking styles in French (sports commentary, religious sermon, political discourse and broadcast news) could be identified in a forced-choice test by considering solely delexicalised utterances. Thus, since speaking styles have to be recognizable with common characteristics by a given group, they vary greatly from person to person and also from time period to time period (LORENZO-TRUEBA *et al.*, 2016).

3.2.2 Expressive TTS

Regarding the design of expressive speech synthesis machines, compared to standard TTS models, Govind e Prasanna (2013) state: "in expressive speech synthesis, along with the text, the desired expression also forms an additional input to the text processing stage". Thus, the question of what and how to insert in these models so that they become expressive becomes extremely relevant. Some clues of which factors to consider in the design of expressive TTS systems are provided in Aylett *et al.* (2021) and are shown in Figure 3.4.

The authors divide the factors into cross-speaker features and within-speaker features. The mentioned cross-speaker factors are language, accent, and dialect, which could both be clues to the speaker's socio-linguistic background, geographical origin, and social identity; voice styles (or speaking styles), which support the intended interaction contexts; and voice adaptation. As within-speaker, the following characteristics are suggested: emotional state, since to be human-like, the model needs to be able to express

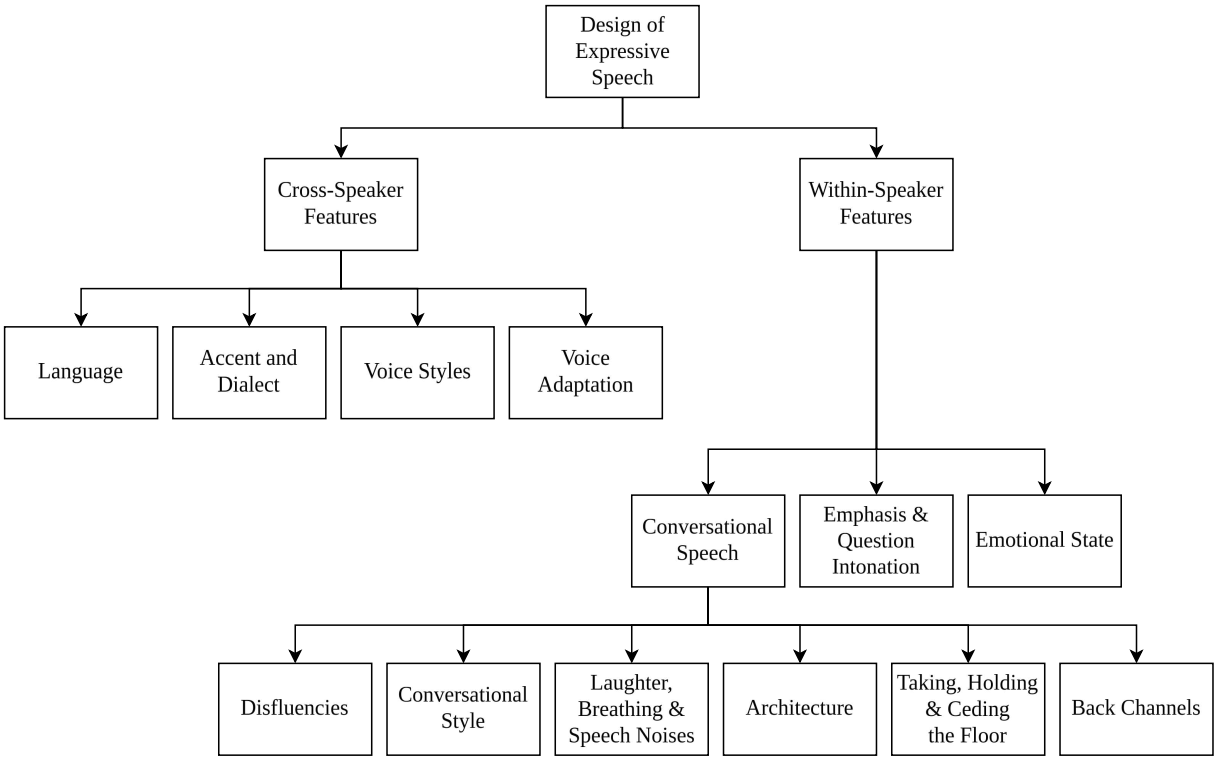


Figure 3.4: Summary of factors proposed by Aylett *et al.* (2021) to consider when designing an expressive speech synthesis model.

emotion accurately; emphasis and question intonation and conversational speech, aiming to instead of focusing on text-reading style, develop systems based on the conversational style, which is more appropriate for human interaction. Several factors to enrich the conversational style could be considered, such as back channels (interject responses to the speaker); disfluencies (speech errors and filled pauses); laughter, breathing and speech noises; talking, holding and ceding the floor; and architecture.

With the rapid development of deep learning, the naturalness and intelligibility of neural TTS models have become comparable to human's (SHEN *et al.*, 2018). From then on, research focused on the expressive aspect of the generated synthetic speech. As the expressiveness can be determined by a confluence of multiple prosodic (suprasegmental elements of speech such as intonation, stress, and rhythm) characteristics, such as content, timbre, prosody, emotion, and style (TAN *et al.*, 2021), approaches started to deal with the modeling, control and transfer of these attributes. This information often denominated as expressive variation information, began to be added to acoustic models to induce expressive speech.

This extra variation information that acts as a simpler approach to introducing expression into the TTS training scheme is necessary to alleviate the speech's one-to-many characteristic: there exists multiple possible speech variations that correspond to the same text. Modeling this mapping without enough extra expression variation information and under current standard MSE or ℓ_1 losses will cause over-smoothing of the mel-spectrograms

predictions, that is, tend to an average of the prosodic distribution (TAN *et al.*, 2021), losing the expressive information present in each sentence, and leading to a less expressive and more monotonous and neutral speech (HODARI *et al.*, 2020).

Tan *et al.* (2021) provides some examples of how this expressive variation information is introduced in many different approaches in the literature. This categorization is shown in Table 3.3. The introduction of variant information can be given either explicitly through the direct insertion of prosodic attributes, such as the F_0 curve, duration, and energy, or through style, speaker, or language labels. In contrast, variant information can be modeled implicitly by using modules that receive a reference and extract information from it. Reference Encoder (RE), VAE, other generative models are some examples. Also, text pre-training can induce better representation by using word embeddings of transfer learning. Another fundamental aspect that characterizes the variant information consists of the granularity at which the information is inserted onto the main TTS model, which can range from fine-grained information (duration, pitch, energy, etc.) found as finer levels such as spectrogram frame, character, phoneme, syllable or word, to more coarse-grained features (speaker timbre, noise, long-form reading, etc.) on more global levels, such as utterance, paragraph, per speaker and also combinations of these (SUN *et al.*, 2020).

Perspective	**Category**	**Description**
Information Type	Explicit	Language/Style/Speaker ID
		Pitch/Duration/Energy
	Implicit	Reference encoder
		VAE
		GAN/Flow/Diffusion
		Text pre-training
Information Granularity	Language/Speaker Level	Multi-lingual/speaker TTS
	Paragraph Level	Long-form reading
	Utterance Level	Timbre/Prosody/Noise
	Word/Syllable Level	Fine-grained information
	Character/Phoneme Level	
	Frame Level	

Table 3.3: Some perspectives of modeling variation information for TTS. Source: Extracted from (TAN *et al.*, 2021).

There are several techniques that take this variant information into account to synthesize expressive speech. Triantafyllopoulos *et al.* (2023) attempt to create a taxonomy for the works focused on deep neural expressive speech synthesis. Their taxonomy is shown in Figure 3.5. First, the authors categorize the works by the type of input and output features into E2E, text-to-features, and feature-to-feature (in the case of Emotional Voice Conversion (EVC)). Regarding the works that use mainly features as input, these could be either spectral, prosodic, or both. Further, the data used to train these models can be

categorized into parallel, when data is available in all the same conditions in different styles, or non-parallel, when it is not available in all styles. In this case, the entanglement issue arises: since data in a given style is only available on a given speaker's voice, the model will naturally correlate both factors as one thing. In this aspect, the other classification category arises, whether the technique follows a disentanglement approach, aiming to separate style from other factors so it can be independently used, or a transformation approach so that enough data in the necessary conditions to train the model is obtained. The expressive information can be introduced into the system through a reference-based approach, or can be either inferred or inputted through categorical labels. The authors also categorized the scale on which the expressive information is added, whether it is at utterance level or at frame level.

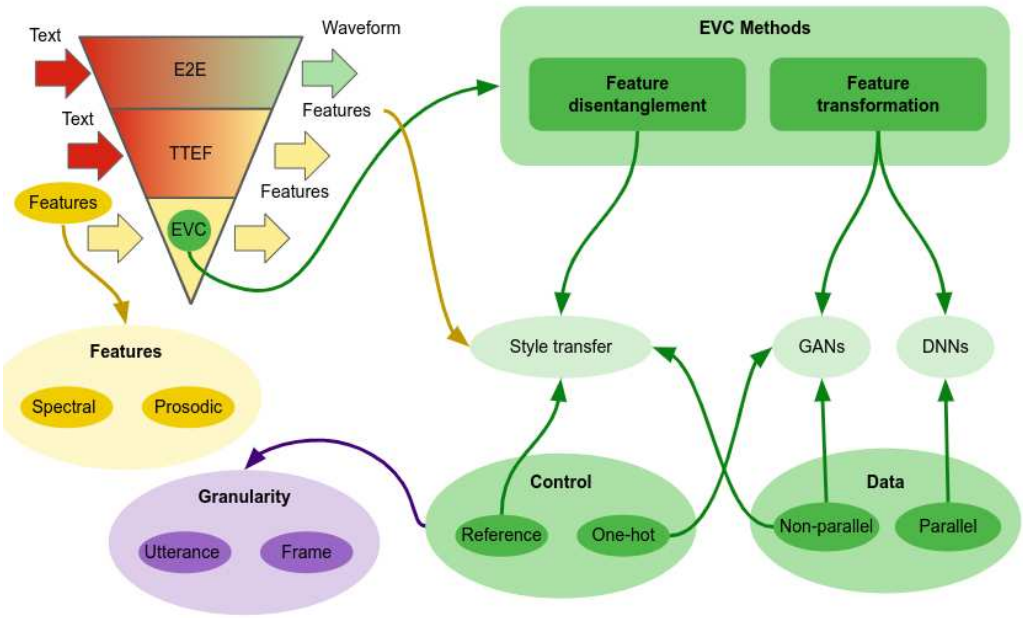


Figure 3.5: Taxonomy of deep neural expressive synthetic speech works. Source: Extracted from (TRIANTAFYLLOPOULOS *et al.*, 2023).

3.2.3 Cross-Speaker Style Transfer

The current traditional approach to extend existing TTS voices to new speaking styles is to directly record and transcribe speech data for the desired speaker, which is not always feasible, in most cases (RIBEIRO *et al.*, 2022). This process is not scalable, given that for every new speaker in the dataset, it will have to be redone and can be compromised due to the possibility that the desired speaker may not perform well in the required speaking style (PAN; HE, 2021). Furthermore, current state-of-the-art TTS systems usually require at least dozens of hours of high-quality transcribed speech data to achieve a good performance (LIAN *et al.*, 2023). Even though the amount of expressive

data needed could be reduced to less than an hour with a transfer learning approach consisting of finetuning a pre-trained neutral model (TITS *et al.*, 2020), the burden of the traditional data recording and transcription process is still required. In this context, cross-speaker style transfer arises as a technique that is able to bypass the unscalable laborious data collection process when trying to extend speaking style to new voices.

Cross-speaker style transfer, by definition, consists in the transfer of a speaking style from a speaker (referred to as "source") to synthesized speech in another speaker's (referred to as "target") voice (LIU *et al.*, 2022). This approach transfers expressive knowledge acquired from high-resource speakers to low-resource speakers (HUYBRECHTS *et al.*, 2021). There are several approaches proposed in the literature to perform style transfer. These can be broadly categorized into two main distinct groups regarding expressive information granularity (SHANG *et al.*, 2021):

1. Coarse-Grained Style Transfer (CST) (WANG *et al.*, 2018; SKERRY-RYAN *et al.*, 2018a): These approaches focus on capturing and transferring global, sentence-level features like speaking styles, emotions, etc. Usually, these features are implicitly or explicitly modeled through a style, time-independent embedding. These can be thus transferred across sentences of different text or length (non-parallel transfer). In this work, to transfer speaking styles, we focus on techniques that perform CST.

2. Fine-Grained Style Transfer (FST) (KARLAPATI *et al.*, 2020; LEE; KIM, 2019): These approaches focus on capturing and transferring more local, fine-grained features like rhythm, emphasis, melody, and loudness. Usually, these features are time-dependent and modeled, either implicitly or explicitly, as latent representations sampled at finer levels, such as word, phoneme, or frame level. These cannot necessarily be transferred across sentences of different text or lengths and would work better with sentences of the same text or length (parallel transfer).

The very first approaches to tackle prosody and style transfer on a neural end-to-end model were proposed by Skerry-Ryan *et al.* (2018b) and Wang *et al.* (2018) simultaneously. The first corresponded to an augmentation of the standard Tacotron (WANG *et al.*, 2017) with a neural network module, denominated RE composed of six convolutional layers with batch normalization followed by a Gated Recurrent Unit (GRU), a type of recurrent neural network. The RE aimed to extract prosody embedding from a reference spectrogram (the one supposed to capture the prosody from) and was trained unsupervised through the standard mel-spectrogram Tacotron reconstruction loss.

The second, denominated Global Style Tokens (GST), consisted of a network composed of a RE, an attention layer, and a bank of embeddings called tokens. The attention layer computes the similarity of the embedding extracted from the reference with each token. From there, the style embedding is generated by a sum of tokens weighted

by the similarity score calculated by the attention layer. In this way, the architecture could decompose the input mel-spectrogram into interpretable latent factors, which, when combined, produce the reference style embedding.

Currently, in literature, there are several strategies to tackle the cross-speaker style transfer task. Although several use the RE network as a basis, it alone does not guarantee that all embeddings extracted from the space will be meaningful since they may not be compact. The GST, similarly, despite having modeled well-defined factors such as noise and environment, has as a disadvantage the fact that it does not control exactly which factors will be modeled independently by tokens, and may even be either untangled factors or aspects unrelated to expressiveness. Thus, the notion of disentangling the speech style factors for the purpose of neural style transfer was proposed by (HSU *et al.*, 2019) to both obtain meaningful latent representations of style uncorrelated from other representations. Following the ideas, most approaches present in the literature currently can be either tackled by modeling global style information from speakers that have expressive data and transferring it to speakers with very little or no expressive data (disentanglement) or also by developing techniques to augment the scarce or non-existing expressive data for target speaker.

Style Disentanglement

The style disentanglement approach is based on being able to capture the averaged prosodic distribution of speech on a given style independently of other varying information present on the representation used of the audios (can be mel-spectrograms, raw audios, self-supervised representations, etc.), such as speaker identity, channel information (noise and recording devices and condition), accent, and phonetic content. Various attempts to use style information modeled it into a single vector representation, which ended up containing too much interfering information, such that it became non-robust and non-interpretable. This way, upon transference, all captured characteristics, including the undesired ones, would be transferred (BIAN *et al.*, 2019). In this context, with the ability to separately model the style information, it could then be transferred to other speaker timbres and input texts. An example of a factorization approach that receives audio and creates four disentangled representations of prosody, content, acoustic detail, and speaker timbre is shown in Figure 3.6.

With the presence of different variant information in the same input, disentanglement is performed during or before model training to obtain a style-controllable speech synthesis (TAN *et al.*, 2021). There are several different techniques to induce the disentanglement style information from other attributes of speech:

- Auxiliary Classifiers: These modules are either used to instruct the corresponding representation which information it should contain, usually in a supervised man-

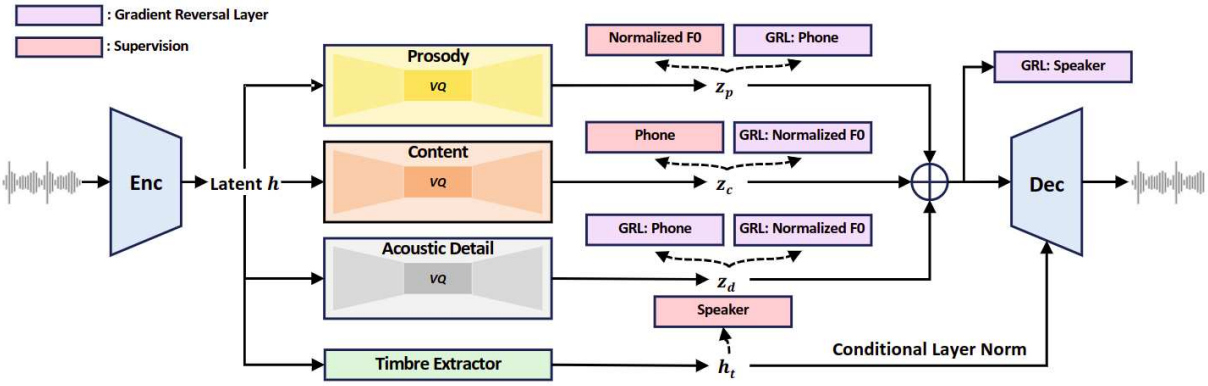


Figure 3.6: Speech Attributes Disentanglment Module of NaturalSpeech3. Source: Extracted from (JU *et al.*, 2024).

ner, or to highlight the discriminating aspects between different styles considered. NaturalSpeech 3 (JU *et al.*, 2024) uses a phoneme classifier on the representation desired to model content, a speaker classifier on the representation desired to model timbre, and an F_0 classifier on the representation desired to model prosody. The iEmoTTS (ZHANG *et al.*, 2023a) model uses a supervised emotional intensity classifier to induce this information, and also an emotion classifier to induce this information on the emotional embedding, as well as to better discriminate between styles. Style and speaker encoders with the same neural architecture and receiving the same inputs are followed respectively by a style classifier and a speaker classifier after the speaker encoder (LI *et al.*, 2021). This way, even though both encoders have the same architectures and receive the same inputs, the classifiers induce their respective output embeddings to be more meaningful and capable of discriminating between different aspects: styles or speakers.

- Domain Adversarial Training (GANIN *et al.*, 2016): With the purpose of unlearning specific information, a Gradient Reversal Layer (GRL) is employed. It consists in swapping the sign of the weight update phase in the gradient descent process, as shown in the following equation:

$$w_i \leftarrow w_i - \eta \cdot \lambda \cdot \frac{\partial L(w)}{\partial w_i}, \tag{3.8}$$

in which, w_i is the weight to be updated, η is the learning rate, $L(w)$ is the loss function. Usually $\lambda = 1$, however, when using a GRL, $\lambda = -k$, for any $k \in \mathbb{N}$, is set causing the weights to move away from local minima, instead of the usual minimization process. With this technique, a set of weights far from the minima can be obtained, making the model avoid learning the information. Therefore, it mitigates the leakage (unnecessary modeling of specific unwanted information). For example, Natural Speech 3 uses a phone-GRL on the prosody encoder, since

this information is not desired on the prosody representation. On the content representations, a F_0-GRL is used; on the acoustic detail encoder, phone-GRL and F_0-GRL are employed, and a speaker-GRL is used on the sum of these representations to remove speaker information. ZET-Speech (KANG *et al.*, 2023) uses an emotion-GRL to remove emotional information from the outputs of a speaker encoder, and report the achievement of zero-shot speaker adaptative cross-speaker style transfer. The iEmoTTS (ZHANG *et al.*, 2022) uses a speaker-GRL to mitigate speaker information on an emotion embedding.

- Information Compression: These techniques aim to reduce the amount of information that enters (in the form of features) or passes (in the form of vector representations) through the neural architecture. By compressing information, the network is induced only to model the most fundamental aspects and ignore others. Compression is exploited in literature mainly through bottlenecking, quantization, reduction and normalization.

  - Bottleneck: A prosody bottleneck sub-network is introduced into a text-to-speech system (PAN; HE, 2021). It receives as input a representation that is combination of content, style and speaker, and is trained to, with this information, predict solely prosodic attributes: F_0, voiced/unvoiced decisions, duration of phonemes and energy. The network is than forced to disentangle the desire prosodic information from all these entangled inputs.
  - Quantization: A Vector-Quantized Variational Auto-Encoder (VQ-VAE) is used to learn a discrete latent prosody space and is reported to achieve better disentanglement performance and representation ability (WANG *et al.*, 2022). Discrete style representations are also obtained in Qiang *et al.* (2022) through a Quantized Variational Auto Encoder (Q-VAE) based RE, which outputs continuous vector but in from a fixed number of classes (to distinguish from the VQ-VAE).
  - Reduction: A mel-spectrogram reduction technique is conducted by inputting only the first 20 coefficients of the mel-spectrogram, to ease the disentanglement, since it was reported these coefficients contained almost complete prosody, and much less timbre and content information when compared to the full version (REN *et al.*, 2022; JIANG *et al.*, 2023). Another approach was to only input a partial segment of reference speech, to avoid content-style entanglement (CHEN; RUDNICKY, 2022).
  - Normalization: A normalization of F_0, used as input, is done to remove timbral speaker information between speakers, leaving only frequency and rhythm information. Additionally, A random re-sampling process is also carried out, in

which the inputs are divided into random lengths, which are either expanded or compressed, functioning as a rhythm bottleneck (QIAN *et al.*, 2020). Another commonly used normalization to remove stationary factors, such as speaker information is the instance normalization. Considering that the constant factor along each of the channels is speaker identity (as opposed to content information which varies), by normalizing each feature channel with its mean and standard deviation, disentanglement from speaker factors can be induced (CHOU; LEE, 2019; KARLAPATI *et al.*, 2020). Also, a batch-permuted latent style perturbation, which enables the generation speaker-unpaired style embeddings during training (JUNG; LEE, 2023).

- Information Perturbation: This approach consists in creating artificial variations on a signal or vector on a specific speech factor while keeping the others constant. Naturally, the model will only learn the constant pattern in data, so the perturbed signals are ignored. Timbre perturbations is used to add or remove harmonic components , while maintaining F_0, and prosody perturbation flattens the complete F_0, keeping it constant, while preserving the timbre. These are applied to the outputs of the style and speaker encoder to remove speaker and prosody information respectively (CHOI *et al.*, 2022b; LEI *et al.*, 2022b).

- Optimization Objectives: These techniques approach disentanglement directly by considering either new training schemes or by adding carefully designed new training objectives. Cheon *et al.* (2022), Zhu *et al.* (2023) propose to use an estimate of a mutual information as a new factor in the total objective of the TTS training goal. This induces a minimization of the mutual information that is shared between the speaker and style embedding, making them model different information. An approach based on adversarial games is proposed by Ma *et al.* (2019). During training, two spectrograms are synthesized: one with the reference audio corresponding to the text and the other with a different audio. With this, a ternary neural discriminator is trained to classify the spectrograms into either a real base audio, a paired fake audio, or an unpaired fake audio. In this way, the separation between text content and style is induced, since the objective of the training is to make the spectrogram generated by the combination of text and audio (matched or unmatched) indistinguishable both from each other and from the real audios of the database. discriminator's point of view. Bian *et al.* (2019) proposes a new training objective based on the orthogonal loss. The idea is the by forcing the orthogonality between speaker and style embeddings, the vectors will be, in terms of the vector space, as distant from each other as possible.

Even though a lot of several different techniques that tackle the cross-speaker style transfer problem, the TTS systems that theoretically induce or guarantee disentanglement

still have one major drawback. Since, by the very own constraint of the cross-speaker style transfer problem, these TTS models are trained to directly reconstruct only speech from the pitch and phoneme alignment of the source speaker(s), since only its data in the given style is available. Thus, no guarantee the the synthesized speech will sound either natural or similar to target speaker's when the input pitch and phoneme alignment come from a different speaker (LI *et al.*, 2023).

Data Augmentation

Given the success of techniques that exploit the use of high-quality synthetic data to complement existing available data for TTS training (HWANG *et al.*, 2021), approaches with the goal of generating synthetic expressive data have been proposed in literature to tackle the low-resource scenario projected onto the cross-speaker style transfer task (HUYBRECHTS *et al.*, 2021). These approaches mainly consider a multi-stage pipeline that generically consists in first using various data augmentation techniques to generate more expressive data in the voice of target speaker, then training a TTS with both the original and synthetic data so enough volumetry is obtained to generate an expressive TTS in the given style and in the voice of target speaker.

Huybrechts *et al.* (2021) proposed the first approach to use data augmentation to synthesize more already available target speaker's expressive data. It was also the very first approach to use VC-created synthetic data to train TTS models. The authors aimed to perform cross-speaker style transfer for speakers with low, yet available, expressive data in the styles of "conversational" and "newscaster". In this approach, the first step was to train Copycat (KARLAPATI *et al.*, 2020), a non-F_0 conditioned prosody transfer VC model on all available data, including all source and target speakers in all styles available for each. Then, all available source speaker styled data is converted into target speaker's voice, augmenting its expressive volumetry. Then, a VAE-enhanced Tacotron (WANG *et al.*, 2017) based TTS model was trained on all synthetic and non-synthetic target speaker's data. Finally, the TTS is finetuned on the non-synthetic available target speaker's styled data. The authors proposed pipeline was shown to work with a trade-off with a requirement of at least 15 minutes of target speaker's expressive data along with 40 hours of neutral data combined from both source and target speaker, achieving a style adequacy of 64%, compared to 78.1%, the high anchor evaluated directly with the recordings. A strict neutral model (low anchor) achieved a style adequacy score of 60.7%. Plus, their style were very close to neutral.

Shah *et al.* (2021) published an improvement of the technique above-detailed, with the replacement of the autoregressive for a parallel TTS, a Tacotron 2 (SHEN *et al.*, 2018) based model one with an external duration model, an extra VC finetuning step on target speaker's voice, and an additional Conditional Generative Adversarial Networks (cGAN) based finetuning step. The authors report the use of "highly expressive" data,

measuring expressiveness as the variation along the three axes of the mean and variation of F_0, power and phoneme duration. They also were able to report results using as low as 15 minutes of target speaker's expressive data.

Chung e Mak (2021) proposed an on-the-fly data augmentation technique. Instead of using a VC model to convert already existing source expressive speech into target speaker's voice, the authors induce, on training, the target speaker to imitate expressive speech of other source speaker by forcing their TTS alignment matrices to be similar. Using a GST-enhanced Tacotron 2 (STANTON *et al.*, 2018; WANG *et al.*, 2018) augmented with a speaker look-up embedding, a training scheme composed of two forward encoder passes, and one decoder pass is conducted. The first pass receives a text, source speaker embedding, and style embedding outputted from a Text-Predicted Global Style Tokens (TPGST) module. The other pass receives the same text and style embedding, but with the target speaker's embedding. The two resultant alignment matrices are the forced to be the close. The main TTS loss is given by the original MSE and Binary Cross Entropy (BCE) composite Tacotron 2 (SHEN *et al.*, 2018) loss:

$$\mathcal{L}_{Taco} = \left\| Mel - \tilde{Mel} \right\|_2 + BCE(Stop, \tilde{Stop}), \tag{3.9}$$

in which $Stop$ and $\tilde{Stop}$ are the true and predicted autoregressive stop token, respectively. The TPGST loss, to remove the audio input dependency on inference is given by:

$$\mathcal{L}_{TPGST} = \left\| TPGST(text, style_label) - GST(audio) \right\|_1, \tag{3.10}$$

and the alignment matrices loss is given by:

$$\mathcal{L}_{align} = \left\| A_{src} - A_{tgt} \right\|_F, \tag{3.11}$$

in which A is the alignment matrix. Thus, the complete training objective is given by:

$$\mathcal{L} = \mathcal{L}_{Taco} + \mathcal{L}_{TPGST} + \mathcal{L}_{align}. \tag{3.12}$$

Their approach is based on the hypothesis that the alignment matrix encapsulate useful rhythmic information that capture a speaking style. Thus, by matching the matrix of the target speaker with the source speaker's matrices, target speaker would then be able to speaker expressively that particular style. They experimented with "newscasting", "public speaking" and "storytelling" styles, with the at least 2 hours of data from the style with least volumetry. Roughly 11 hours of target speaker's neutral data was used. They evaluated naturalness and intelligibility for each one of the styles. Also, given the style for the rater, they conducted an ABX preference study to evaluate if the raters preferred the styled data on target speaker's voice or its neutral data regarding each scenario. A mean preference of 67% of the proposed approach was obtained, compared to around 26% of the

neutral model, and 7% had no preference.

Ribeiro *et al.* (2022) proposed the first approach to perform cross-speaker style transfer using data augmentation techniques, assuming no expressive data was available for target speaker, only neutral. Similarly to the previous approaches mentioned, by using supporting expressive speakers, high-quality synthetic expressive data on target speaker's voice is generated using a VC model. Then, a multi-style single-speaker TTS model is trained on both neutral and synthetic expressive data of target speaker. A CopyCat VC model, extended with a log-F_0 conditioning on source utterance is trained to convert the source speaker into target speaker's voice. Thus, after converting the supporting expressive data into target speaker's voice, a VAE-enhanced Tacotron 2 (SHEN *et al.*, 2018).

They used expressive data in a "conversational" speaking style, and ten hours of neutral data from the target speaker was considered. Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) Experiments are performed varying the number of supporting speakers, from 1 to 8, while keeping 8 hours of supporting volumetry constant, and also varying the amount of supporting expressive data, from 1 hour to 8 hours while keeping 4 supporting speakers. Naturalness, style similarity and speaker similarity are the three criteria evaluated. No significant statistical gain is obtained in both experiments. For the system requiring less data (greater constraints of 1 hour of supporting data and), a style similarity of 55.41 is reported, while the TTS trained directly on conversational data in a supporting speaker's voice (high anchor) obtains 73.69. Speaker similarity results of the system are reported as 69.25, while neutral original recordings of target speaker (high anchor) obtained 72.32.

Terashima *et al.* (2022) proposed the first cross-speaker style transfer approach for highly expressive styles (emotions) using data augmentation. They used a Pitch-Shift (PS) data augmentation technique along with the VC-based one. They proposed a PS technique applied directly on the neutral spectrograms from both source and target speakers. This PS-augmented along with the original neutral data are used to train a VC model. They report that since now the VC covers a greater variety of pitch dynamics, the training process becomes more stable. They use the Scyclone (KANAGAKI *et al.*, 2020) VC model, trained to predict the log-mel-spectrogram, and the additional features of logF_0, and voice/unvoiced flags, which they report is essential to create emotional TTS models that include F_0-dependent neural vocoders. In addition to the usual losses of the Scyclone (KANAGAKI *et al.*, 2020), they also adapt a regularization term to avoid the unnatural conversion of the prosodic features, that approximates the STFT magnitudes for the predicted and extracted F_0 sequences. By considering only the higher frequency bins, the authors are able to regularize the essential fine, high-frequency components of F_0. Then, all source data (expressive and neutral) are converted into target speaker's voice. All this converted data, along with the original neutral recordings of target speaker are used to train an emotional TTS. The authors used the FastSpeech 2 (REN *et al.*, 2021)

model extend with an input of the emotion embeddings. Two female Japanese professional speakers were considered, one for source speaker, containing three speaking styles: neutral, happy, and sad, whereas the target speaker contained only the neutral style. 1000 neutral utterances of target speaker were used, together with 5000, 2500, and 2500 utterances for the neutral, happy, and sad style on source speaker's voice. The authors analyzed results of naturalness, speaker similarity and emotional similarity. The data-augmentation based approaches achieved either similar or better results than the models trained directly on original data on all three aspects. The complete system pipeline of the approach is shown on Figure 3.7.

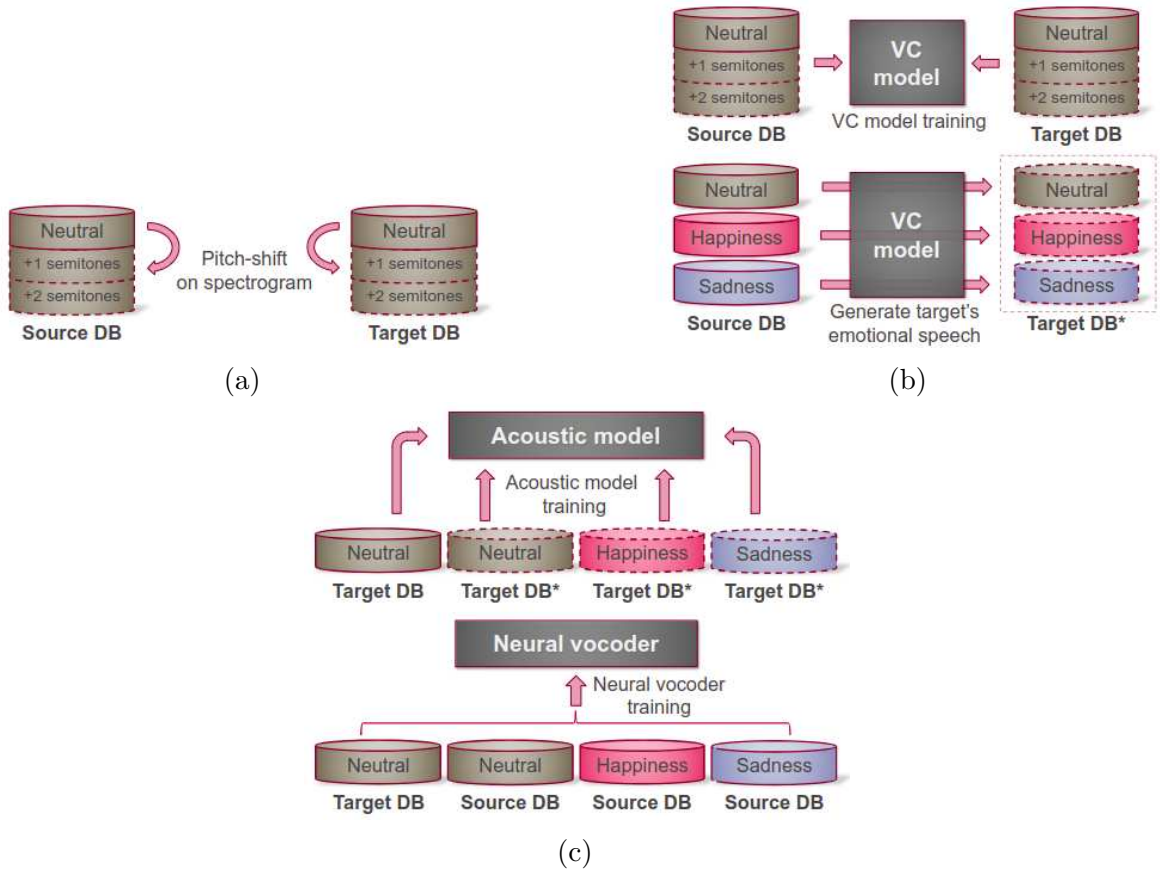


Figure 3.7: a) PS augmentation step. Source: Extracted from: (TERASHIMA *et al.*, 2022). b) VC training and augmentation step. Source: Extracted from: (TERASHIMA *et al.*, 2022). c) TTS and Vocoder training steps. Source: Extracted from: (TERASHIMA *et al.*, 2022).

Zhang *et al.* (2023b) propose an approach based on curriculum learning (WANG *et al.*, 2022) and data augmentation. Curriculum learning consists in a training process that trains a model sequentially considering first easier data, than scaling to harder data in every next stage (WANG *et al.*, 2022). The cross-speaker style transfer task is divided into two steps: parallel, and non-parallel transfer. The first step was defined as synthesizing the expressive source dataset in the voice of the target speaker, which can be seen as a data augmentation process. For this, an approach similar to Hua *et al.* (2022)is used.

The second step consists in using the augmented data to train a GST (WANG *et al.*, 2018)-augmented Fastpitch (ŁAŃCUCKI, 2021) TTS model. The authors consider three styles in the experiment stage: a 50-hour female speaker dataset in neutral style as the target speaker; a 37 minute male speaker documentation-style; a 15 minute chat-style female speaker; and a game-style 1.4 hours female speaker. Also, a 50 hours neutral-style is used to add robustness to the TTS training. Results showed that the pipeline proposed with curriculum learning and data augmentation achieved better results in all styles, compared to the model trained only on available data.

All the works that attempt to tackle cross-speaker style transfer via a data augmentation approach are summarized in Table 3.4. The amount of required data, both neutral and styled, for the source and target speakers is shown on the Table. The best values for each combination is highlighted.

This work is also placed for comparison. Even though the amount of data used on the source speakers voices is slightly greater than other existing approaches in literature, the objective of this work is to focus on a faster adaptation of existing expressive dataset to new voices. In this context, priority is given to use less target speaker data. For this, the use of pre-trained models that reduce the amount of neutral data required from the target speaker for as low as 5 minutes, the lowest volumetry ever used in the data augmentation approaches.

3.3 Concluding Remarks

In this Chapter, some basic concepts related to the most recent synthetic speech methods that use deep neural networks were explained: advances in the deep learning field together with the creation of large speech datasets allowed a better modeling of speech and thus enabled these systems to achieve human-like levels of naturalness and intelligibility. Nevertheless, the complexity of augmenting these systems with human expression is brought up: considering expressiveness is challenging since most of the approaches always converge to an average of the prosody distribution of the dataset, filtering out important and more prominent prosodic variations. An additional challenge is when considering more expressive or emotional speaking styles, when the prosodic variation is even greater than the standard neutral speech. Also, some works that focused on transferring speaking styles from data-available speakers to a speaker with only neutral data were presented. Finally, the approach of this work, and how it relates to the mentioned approaches is briefly introduced. The method considered is detailed in the next Chapter.

In Chapter 4, we detail the complete developed pipeline that proposed the use of an SVC model to account for highly expressive styles, as well as the integration of a style classifier filter and an F_0 matching technique to account for speakers with very different vocal registers.

Reference	Augmentation Technique	TTS Structure	Lowest Source Spk Volumetry	Lowest Target Spk Volumetry	Speaking Styles
Huybrechts *et al.* (2021)	Copycat (KARLAPATI *et al.*, 2020)	VAE+ Tacotron (WANG *et al.*, 2017)	Neutral: 20h Styled: 4h	Neutral: 20h Styled: 15min	Conversational, Newscaster
Shah *et al.* (2021)	Copycat (KARLAPATI *et al.*, 2020)	VAE+ Tacotron2 (SHEN *et al.*, 2018)	Neutral: N/A Styled: 4.5h	Neutral: N/A Styled: 15min	Conversational
Chung e Mak (2021)	Alignment Matrix Loss	TPGST-Tacotron2 (STANTON *et al.*, 2018)	Neutral: **0h** Styled: 2.1h	Neutral: 11.7h Styled: **0min**	Newscaster, Storytelling, Public Speaking
Ribeiro *et al.* (2022)	Copycat+F_0 (KARLAPATI *et al.*, 2020)	VAE+ Tacotron2 (SHEN *et al.*, 2018)	Neutral: N/A Styled: 1h	Neutral: 10h Styled: **0min**	Conversational
Terashima *et al.* (2022)	PS+Scyclone (KANAGAKI *et al.*, 2020)	FastSpeech2 (REN *et al.*, 2021)	Neutral: 7h[1] Styled: 3h[1]	Neutral: 1.4h[1] Styled: **0min**	Happy, Sad
Zhang *et al.* (2023b)	(HUA *et al.*, 2022)	GST+ FastPitch (ŁAŃCUCKI, 2021)	Neutral: 50h Styled: **15min**	Neutral: 12h Styled: **0min**	Documentation, Game, Chat
This work	SO-VITS-SVC	FastPitch (ŁAŃCUCKI, 2021)	Neutral: 44h Styled: 3h	Neutral: **15min** Styled: **0min**	Angry, Happy, Sad, Surprised

Table 3.4: Summary of cross-speaker style transfer based on data augmentation techniques presented in Section 3.2.3. (1) Approximated with the number of utterances multiplied by the medium utterance duration.

Chapter 4

Method

This chapter describes the proposed method of cross-speaker style transfer based on data augmentation, which is illustrated in Figure 4.1. As shown in the figure, the training pipeline processes as inputs:

- a neutral speech dataset of non-target speakers (Source Neutral Data);
- a neutral speech dataset of the target speaker (Target Neutral Data);
- an expressive speech dataset of non-target speakers with the desired style (Source Expressive Data).

The training process results in a specialized Stylized TTS for the target speaker's voice. Given a few hours of public neutral and expressive datasets, the proposed method can train specialized models for new target speakers with only five minutes of Target Neutral Data.

The first step is fine-tuning a pre-trained SVC model on the Target Neutral Data to learn the target speaker's voice and the source expressive data to understand expressive data conversion. This step is detailed in Section 4.1.

The conversion process consists of taking the fine-tuned SVC from the last step and converting the Source datasets (Neutral and Expressive) into the target speaker's voice. An F_0 matched conversion algorithm is used to mitigate unrealistic results caused by a very different F_0 register between any source speaker and the target. This process is described in Section 4.2.

Then, a style-based filtering process is conducted to select the most stylistic appropriate converted data from the last step to be used in the subsequent stages. A neural style classifier filter filters out all the converted audios whose inferred style labels changed after conversion. This step is detailed in Section 4.3.

After both expressive and neutral data are in the voice of the target speaker, a neutral TTS pre-training step is performed, which receives the F_0 matched converted neutral dataset obtained from the second step and outputs a neutral TTS in the target

speaker's voice. Finally, an expressive fine-tuning step is conducted, fine-tuning the pre-trained neutral TTS model obtained from the last step on the expressive, F_0 matched, filtered converted data. All TTS related steps are detailed in Section 4.4.

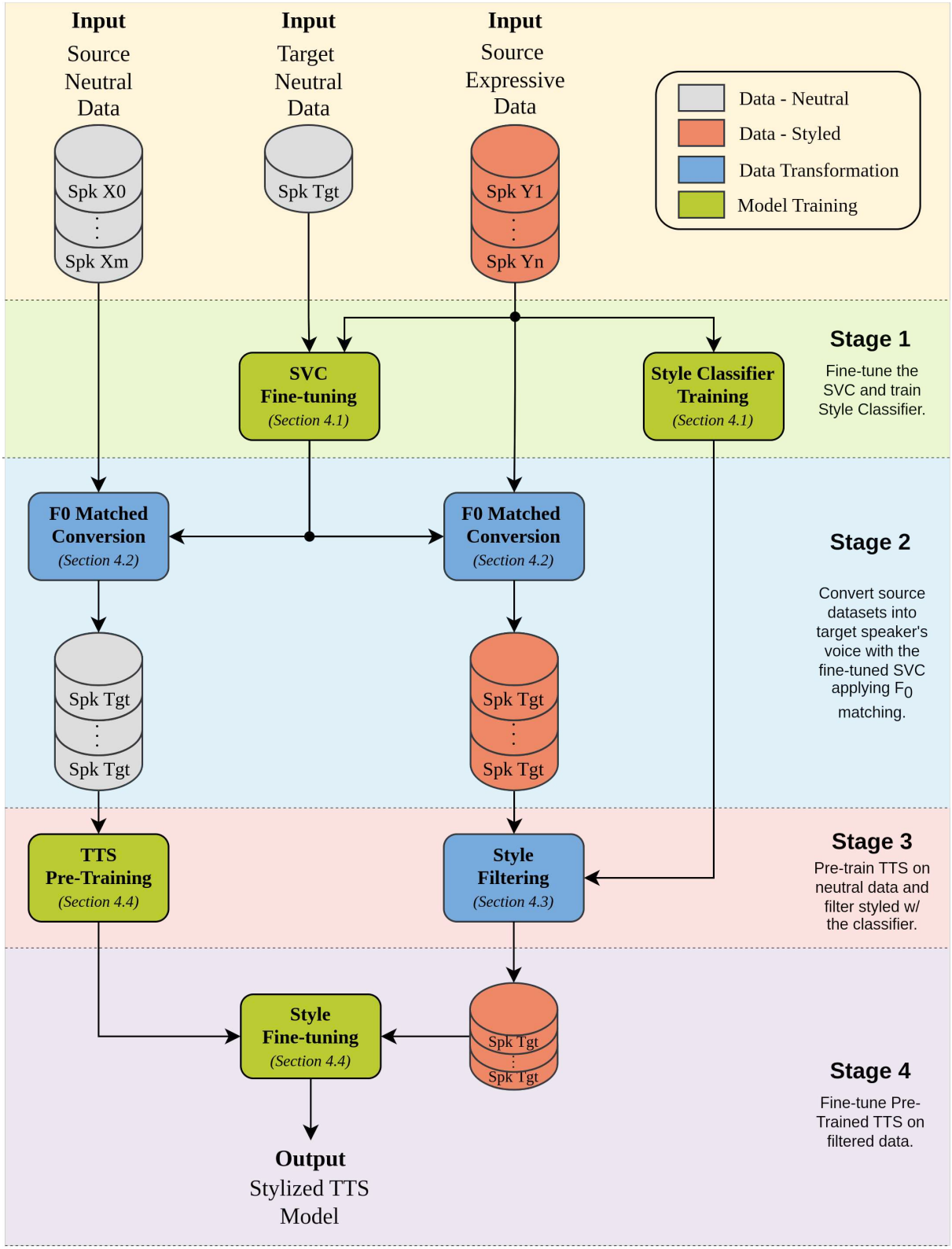


Figure 4.1: Overview of the proposed pipeline to build a TTS model with target speaker's voice and source speaker's style.

The last two sections present the datasets used in each step of the method (Section 4.5), as well as the experimental setup in which the pipeline was conducted (Section 4.6).

4.1 Singing Voice Conversion

This section describes the SVC finetuning step. It consists of using a pre-trained model on a multi-speaker speech and multi-speaker singing voice corpus and then fine-tuning with the target speaker's neutral and source expressive data. The output is an SVC model that converts any speech audio into the target speaker's voice.

As the SVC model, we used So-VITS-SVC [1], a SOTA, open-source, conditioned-on-$F0$ model. This model combines four different audio encoders that extract representations with different meanings. A pre-trained timbre encoder based on (WAN *et al.*, 2018) is used to extract speaker representations, a Whisper (RADFORD *et al.*, 2023) encoder is used to extract content information, a soft version (NIEKERK *et al.*, 2022) of the HuBERT (HSU *et al.*, 2021), which model to extract prosody representation, and a CREPE (KIM *et al.*, 2018) model to obtain the $F0$. Similar to Variational Inference with Adversarial learning for end-to-end Text-to-Speech (VITS) (KIM *et al.*, 2021), these are consumed by a normalizing flows-based decoder that generates the output audio. The model is also trained with a speaker classifier with a gradient reversal layer to achieve speaker disentanglement. The model's architecture is shown in Figure 4.2.

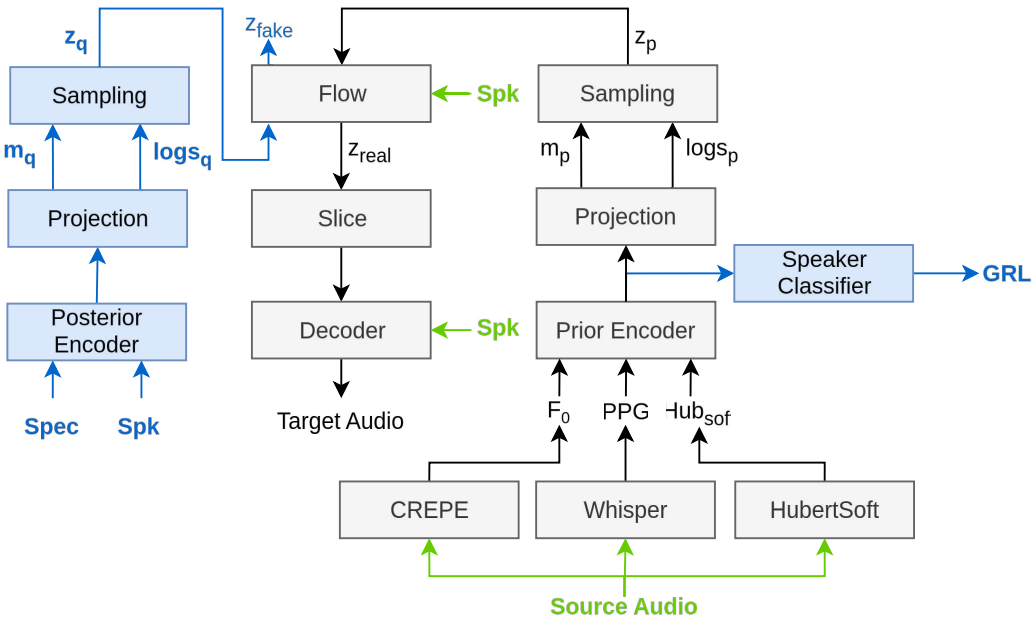


Figure 4.2: Architecture of the SO-VITS-SVC pipeline. Green highlights the inputs. Blue modules are used only during training.

The model receives as input the target speaker representation, the source speaker

[1] Available at <https://github.com/PlayVoice/whisper-vits-svc>

audio, and extracts the Whisper Phonetic Posteriorgrams (PPG), the soft Hubert (NIEKERK *et al.*, 2022) representation, and the F_0 curve. During training, the soft Hubert and PPG representations are perturbed with random Gaussian noise to improve noise immunity and remove global timbre information. Also, the speaker embedding is normalized. The F_0 curve is transformed into a coarser curve by mapping it to a Mel scale, min-max normalizing within 50Hz and 1100Hz, re-scaling the normalized values with multiplication by 254, and then clamping the values to integers between 1 and 255. The PPG and soft Hubert representations are processed by a prior encoder that starts with a convolutional layer and then sums together both with the sequence of embeddings extracted from a look-up table indexed by the coarse F_0 curve. These representations combined are passed through a Transformer encoder (VASWANI *et al.*, 2017) for context and then projected with a convolutional layer. The output value is split to produce the mean and variance used to construct the prior distribution. The sampled vector goes through a normalizing flow conditioned on the speaker representation and consists of four volume-preserving affine coupling layers. Each layer consists of four WaveNet (OORD *et al.*, 2016) residual blocks. The output representation is sliced into smaller chunks and then consumed by a BigVGan (LEE *et al.*, 2023) vocoder to generate speech.

The model is trained with several losses, including a discriminator-based GAN loss; an ell_1 mel loss between the synthesis and ground-truth; loss of multi-resolution STFT (YAMAMOTO *et al.*, 2020); a GRL loss applied with the cosine distance of a classifier that receives the contextual content and prosodic representations; and the Speaker-Normalized Affine Coupling Layer (SNAC) flow-related losses (CHOI *et al.*, 2022a).

The pre-trained SVC that is used on this work was already trained on the source neutral and singing voice datasets. The pre-training ensured that the model had already learned various voice timbres and the conversion of richer phonation modes that occur in singing voice. With this, the taken model could already (1) perform a quick adaptation to any new voice and (2) maintaining the source speech's expressiveness when converting since it was also pre-trained on the highly expressive singing voice data.

In the fine-tuning step conducted, the used pre-trained SVC model is trained on the target speaker's neutral data and the source expressive dataset. From the pre-training step, this process becomes quicker because we already have information on how to convert voices. Also, it has the benefit of demanding less data (as little as 5 minutes from the target speaker. Additionally, by simultaneously training on the source expressive dataset, the model can learn the patterns of the desired styles to be converted.

At the end of the SVC fine-tuning process, a model is obtained that not only to transforms any source speech audio into the target speaker's voice but also with improved expressiveness, hypothesized with the use of singing voice. Also, it can preserve $F0$, prosody, expressiveness, and content information.

4.2 F_0 Matched Conversion

From the previous step, the obtained SVC model is trained to copy the input audio wave's F_0 into the output wave on the target speaker's voice. Thus, when converting speech from a speaker with a very different F_0 register than the target, a mismatch of the F_0 range occurs and causes unrealistic converted speech. This factor is more prominently shown when considering different genders since vocal folds become longer and thicker in male speakers, leading to lower values of F_0 (PÉPIOT, 2014). Also, given that the average length of the male vocal tract is longer than the average length of the females, an increase in lower resonant frequencies is expected (PÉPIOT, 2014).

During the pipeline's conversion steps, an F_0 matching algorithm is proposed to mitigate this issue. This algorithm aims to reduce the mismatch in $F0$ between the source and target speakers. First, a semitonal distance is calculated between each source and the target speaker. Then, during neutral and expressive conversion steps, the input pitch curve to the SVC is transposed according to the previously calculated semitonal distance between the source speaker being converted to the target speaker's voice.

For each audio in all three input datasets, a mean value of F_0 was computed using the voiced segments with the Harvest estimator (MORISE, 2017). Then, for each speaker, a mean of the per audio F_0 mean was calculated, obtaining an average value of F_0 in which each speaker speaks most of the time. Next, a distance in the interval of semitones was computed from the target speaker's average F_0 value to each source speaker. With this, during the conversion steps, both neutral to target and expressive to target, the F_0 was transposed by the calculated semitonal distance between the input source speaker and target speaker's audios, ensuring that all converted speech is in a range adequate for the target speaker's voice. This stage is shown on Figure 4.3, and detailed on the Algorithm 1.

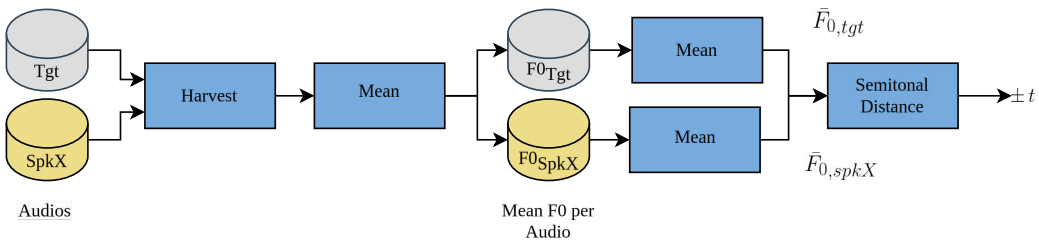


Figure 4.3: Processing pipeline to generate the semitonal distances.

After the semitonal distances are computed, the input pitch curve is transposed according to each source speaker that is being converted. This pitch shift is given by:

$$F_0^{'} = F_0 \cdot 2^{\frac{\Delta_{st}}{12}}, \tag{4.1}$$

in which Δ_{st} is the semitonal distance between the source and target speaker's voices.

Algorithm 1 F0 Matching Algorithm

1: **Input:** Source dataset consisting of J audios, $S_j, j \in [0, J]$, partitioned into M subsets S^m of source speakers, $S = \bigcup_{m=1}^{M} S^m$, naturally disjoint, $S^a \cap S^b = \emptyset$, $\forall$a $\neq b$.
2: **Input:** Target dataset consisting of L audios, $T_l, l \in [0, L]$ in target speaker's voice, named tgt.
3: **Output:** Semi-tonal distances between target and each source speaker: $\Delta_{st}(m)$, $m \in [0, M]$.
4: **for** l in {0, 1, ..., L} **do**
5: Calculate F_0 curves for each audio l of target speaker.

$$F_{0_l}[k] = \text{Harvest}(T_{l,\,voiced})$$

6: Compute the mean of F_0 for each audio.

$$\overline{F_{0_l}} = \frac{1}{|F_{0_l}|} \sum_{k=1}^{|F_{0_l}|} F_{0_l}[k]$$

7: **end for**
8: Compute target speaker F_0 mean across all speaker F_0 audio means.

$$F_0^{tgt} = \frac{1}{L} \sum_{l=1}^{L} \overline{F_{0_l}}$$

9: **for** m in {0, 1, ..., M} **do**
10: **for** j in {0, 1, ..., $|S^m|$} **do**
11: Calculate F_0 curves for each audios j for each source speaker m.

$$F_{0_j}^m[k] = \text{Harvest}(S_{j,\,voiced}^m)$$

12: Compute the mean of F_0 for each audio j of each source speaker m.

$$\overline{F_{0_j}^m} = \frac{1}{|F_{0_j}^m|} \sum_{k=1}^{|F_{0_j}^m|} F_{0_j}^m[k]$$

13: **end for**
14: Compute a speaker F_0 mean across all speaker F_0 audio means for each source speaker m.

$$F_0^m = \frac{1}{|S^m|} \sum_{j=1}^{|S^m|} \overline{F_{0_j}^m}$$

15: **end for**
16: **for** m in {0, 1, ..., M} **do**
17: Compute semitonal distances from source speaker m to target:

$$\Delta_{st}(m) = 12 \cdot \log_2 \left(\frac{F_0^{tgt}}{F_0^m} \right)$$

18: **end for**
19: **Return:** Δ_{st}

4.3 Style Filtering

It has been reported that many SOTA VC models can achieve conversion with high intelligibility and naturalness in real-time but fail to adequately preserve the emotions of the source speaker, especially in scenarios of highly varying pitch and when considering diverse emotions (GHOSH *et al.*, 2023). To mitigate this issue, besides the replacement of the VC with an SVC model, a style filtering step was proposed to filter out possible synthetic converted data whose output style had not been maintained the same as the input source utterance's style. The block diagram representing this step is shown in Figure 4.4.

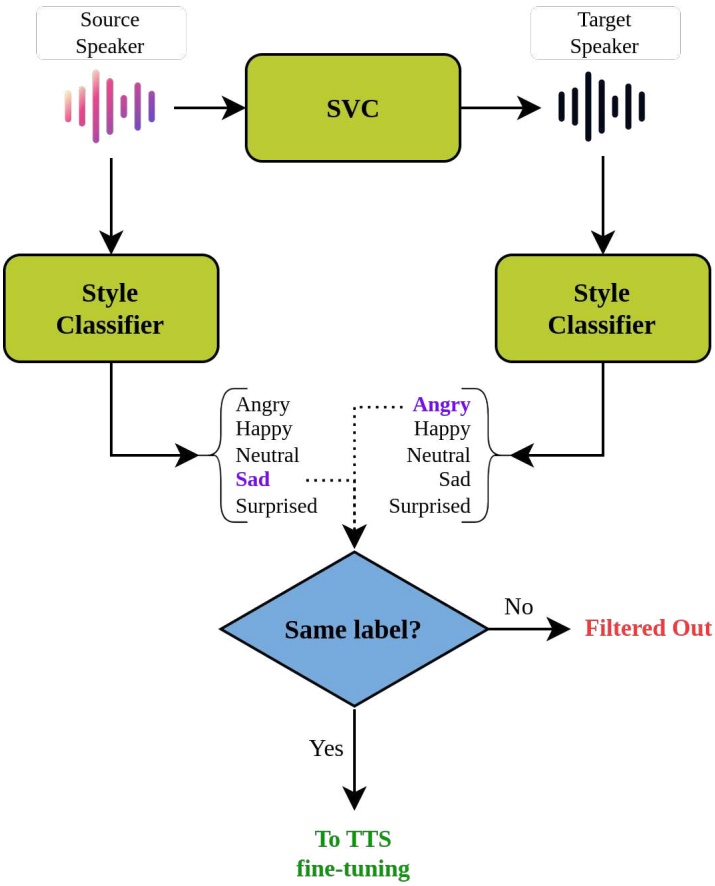


Figure 4.4: Block diagram of the style filtering process.

In this step, a style classifier model was trained to predict the emotion labels of the source expressive dataset until convergence. The same architecture of the RE (SKERRY-RYAN *et al.*, 2018a) is employed, shown on Figure 4.5, which receives an input mel-spectrogram, as processes it with 6 layers of 2D convolutions with batch normalization, and a GRU and with a linear layer on top, to predict the probability of each given emotion. The model is trained under a usual cross-entropy loss. If the number of speakers on the source expressive dataset is sufficiently large, then the classifier can classify each style of the dataset robustly such that it is independent of the speaker timbre. So, it was trained

on the source expressive dataset and used to reason about the styles of converted synthetic audios on the target speaker's voice.

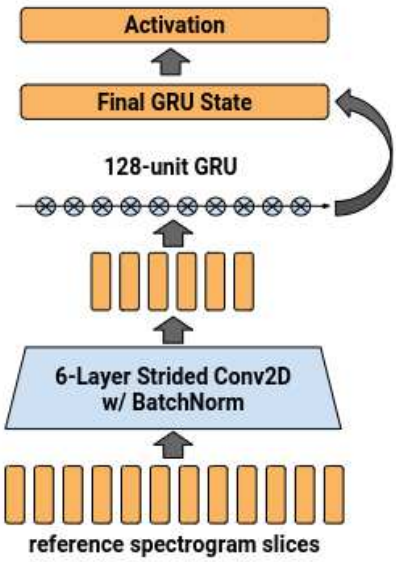


Figure 4.5: Architecture of the Style Classifier Filter RE. Source: Extracted from (SKERRY-RYAN *et al.*, 2018b).

With the F_0 matched conversions, the audio was filtered out for each audio on the synthetic expressive dataset if the style classifier inferred that it has any style other than the same style as the input. This way, using the classifications by the style classifier, only audios whose style was kept constant after conversion are used for finetuning. The process of the style filtering step is detailed on the Algorithm 2.

Algorithm 2 Style Classifier-based Filtering Process

Input: Source expressive dataset $\mathcal{D}_{src}$
Input: Target synthetic expressive dataset $\mathcal{D}_{tgt}$
Input: Neural classifier model, $\mathcal{C}$
Output: Filtered target synthetic expressive dataset $\mathcal{D}_{tgt,filt} \subset \mathcal{D}_{tgt}$
Step 1: Train Style Classifier
Train style classifier model $\mathcal{C}$ on $\mathcal{D}_{src}$ until convergence
Step 2: Filter Audio
Initialize empty filtered target synthetic expressive dataset $\mathcal{D}_{tgt,filt} = \emptyset$
for each audio a with style $S \in \mathcal{D}_{tgt}$ **do**
  Infer style on target speaker's voice: $\hat{S} = \mathcal{C}(a)$
  if $\hat{S} \neq S$ **then**
    continue
  else
    $\mathcal{D}_{tgt,filt} = \mathcal{D}_{tgt,filt} \cup \{a\}$
  end if
end for
Return $\mathcal{D}_{tgt,filt}$

4.4 Text-to-Speech

A TTS model is trained on the synthetic converted data to generalize the stylistic speech for any given input text. The TTS training procedure was divided into two main steps: neutral pre-training and style finetuning. It receives both the neutral synthetic converted and expressive synthetic converted and filtered datasets and outputs the final model of the pipeline, a TTS in target style in the target speaker's voice.

In the neutral pre-training stage, only the neutral data is used to train the model. This is done since neutral data provides a more stable TTS training than stylistic data since there is less prosodic variation. The neutral source dataset required enough volumetry to train a TTS model from scratch, which alleviated the need for a large volumetry of both source expressive data and the target speaker's neutral speech. With this, the expressive source data must have been enough only to perform a style finetuning. As our TTS model, we used FastPitch (ŁAŃCUCKI, 2021) with explicit duration, pitch, and energy predictors due to its fast and high-quality TTS capability (see Section 3.1.2 on Chapter 3). The alignment between text and mel-spectrogram frames is learned during training in an unsupervised manner as proposed in (BADLANI *et al.*, 2022a). The technique is used to loose the constraints of requiring alignments for any datasets.

After obtaining a neutral TTS on target speaker's voice, a style finetuning step is performed to adapt the neutral to each desired target speaking style. The same alignment technique is used in the finetuning.

4.5 Data

This section presents in detail all the datasets considered in each step of the training pipeline: the singing data used for pre-training the SVC; the expressive dataset composed by the source expressive speakers; the neutral data in target speaker's voice; and the source neutral dataset used to pre-train the TTS model composed by the source neutral speakers.

To ensure reproducible results, only open-source datasets were used in this work. To pre-train the SVC model, we use the OpenSingerChinese (HUANG *et al.*, 2021) dataset together with the Voice Cloning Toolkit (VCTK) (VEAUX *et al.*, 2017) dataset, which is also used as the neutral source dataset. As our target speaker dataset, due to its known voice across TTS research, we used the Linda Johnson (LJ) speech (ITO; JOHNSON, 2017) dataset. Lastly, as the source expressive dataset, we used the English portion of the Emotional Speech Dataset (ESD) due to its highly expressive emotions.

Table 4.1 provides a summary of the datasets used in this work, including information about each one and its function in the training pipeline.

4.5.1 OpenSinger Dataset

OpenSinger[2] (HUANG *et al.*, 2021) is a Chinese high-quality large-scale open-source multi-singer singing voice dataset. It is composed of audio of various pop songs, summing up to 50 hours of recordings in total. These are divided into 30 hours from 41 female singers and 20 hours from 25 males. All audios are provided in wav format with a sampling rate of 24kHz and quantized in 16 bits. A professional annotation team labeled these with the lyrics, song name, singer, and phonemes. The files were trimmed with a Voice Activity Detection (VAD) model to remove silences, then cut into chunks of 0 to 11 seconds to better fit in limited-memory GPU scenarios. Lastly, time alignment between audios and phonemes is provided and calculated with the Montreal Forced Aligner (MFA) tool[3].

This OpenSinger singing voice was used in the pre-training of the SVC model that is used in this work. As discussed in Chapter 1 Section 1.4, question **Q.2**, since this dataset of singing voice contains richer emotional information, diversified across singing expression and style and also elevated high-frequency (HUANG *et al.*, 2021), when compared to a speech dataset, the substitution of the VC on the cross-speaker style transfer data augmentation-based pipeline on data with the SVC, is hypothesized to improve the conversion of expressive data.

4.5.2 Emotional Speech Dataset

The ESD[4] (ZHOU *et al.*, 2021b) is a multi-lingual and multi-speaker dataset designed for voice conversion and speech synthesis research and available for non-commercial purposes. It comprises 350 parallel utterances spoken by 10 native English speakers and 10 native Chinese speakers. For each language, 5 male and 5 female speakers are present. For each speaker, the 350 utterances are spoken in a neutral speech and 4 emotions: "happy", "angry", "sad", and "surprised". It sums up to approximately 29 hours of speech data recorded. Speech data is available in 16kHz and quantized into 16 bits. Transcriptions are also made available.

This work uses the English partition of the ESD as the source expressive dataset. It is chosen mainly due to three factors: it is multi-speaker, which allows the evaluate the conversion of different timbres to target speaker; because it comprises highly expressive styles and emotions, which makes it possible to evaluate the purpose of using the SVC; and also because there is enough volumetry in all styles to finetune a pre-trained TTS model.

[2]Available at <https://multi-singer.github.io/>
[3]Available at <https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner>
[4]Available at <https://github.com/HLTSingapore/Emotional-Speech-Data>

4.5.3 VCTK Dataset

The VCTK Corpus[5] (VEAUX *et al.*, 2017) is a high-quality multi-speaker multi-accent English speech dataset designed to build HMM-based TTS models and available for non-commercial purposes. It contains 109 native English speakers with several different accents worldwide. Aiming to maximize phonetic coverage, each speaker read about 400 sentences taken from newspaper texts. The corresponding transcriptions and speaker information, such as accent, gender, and age, are provided. All speech data is recorded in 48kHz, with a bit depth of 16 bits.

The corpus is used as the neutral source dataset in the proposed pipeline. It was selected not only because it contains sufficient data to train an English TTS model from scratch but also to highlight the effectiveness of the F_0 matching algorithm when converting 109 possibly very different speakers to target, with various timbres and genders, which naturally challenges the usual conversion step.

4.5.4 LJSpeech Dataset

The LJspeech[6] is a public domain English dataset from a single female speaker containing 13100 utterances taken from reading sessions of 7 non-fiction books, along with its normalized transcriptions. These vary in length from 1 to 10 seconds and sum up to 24 hours. All are provided in a sample rate 22.05kHz and quantized in 16 bit.

The LJspeech is selected mainly due to its known voice across speech synthesis research, being familiar to many and is used in the pipeline as the neutral target speaker data. Thus, only segments of speech summing up to 5 minutes are used in the pipeline.

4.6 Experimental Setup

This Section details how the experimentation with the proposed pipeline is approached. To compare the stylized TTS models, we consider other denominated baselines models, detailed in subsection 4.6.1, which are SOTA style transfer models implemented from literature. For this, the Daft-Exprt (ZAïDI *et al.*, 2022) is considered in two different approaches to synthesis. Also, experimentation concerning the effectiveness of some individual steps of the pipeline is conducted. For this, an ablation study, detailed in Section 4.6.2, is performed by replacing the SVC model with a VC, and also an experiment removing the filtering step. The training configuration is shown on Section 4.6.3. All these models are evaluated with a perceptual assessment and with some objective metrics.

[5]Available at <https://datashare.ed.ac.uk/handle/10283/3443>
[6]Available at <https://keithito.com/LJ-Speech-Dataset/>

Dataset	Volumetry	# Speakers	Styles	Rate/Bit Depth	Pipeline Function
OpenSinger [CN] Huang *et al.* (2021)	50h	41 (F) 25 (M)	Singing Voice	24kHz 16bit	SVC pre-training
ESD [EN] Zhou *et al.* (2021b)	29h	5 (F) 5 (M)	Angry, Happy, Neutral, Sad, Surprise	16kHz 16bit	Source Expressive
VCTK [EN] Veaux *et al.* (2017)	44h	62 (F) 47 (M)	Reading	48kHz 16bit	SVC pre-training, Source Neutral
LJSpeech [EN] Ito e Johnson (2017)	24h	1 (F)	Reading	22.05kHz 16bit	Target Neutral

Table 4.1: Summary of the datasets used in the pipeline.

4.6.1 Baselines

The proposed method was compared to Daft-Exprt (ZAïDI *et al.*, 2022), an open-source state-of-the-art cross-speaker prosody transfer model, designed to capture both low-level prosodic features such as pitch, duration, and energy; and high-level speaking style information, which is encoded in a learned latent space. Instead of using a data augmentation approach to cross-speaker style transfer, the authors attempt to disentangle prosodic from speaker information through adversarial training with a GRL.

The model comprises a core acoustic model that generates mel-spectrograms from text. The acoustic model is conditioned on prosody information obtained from a reference utterance passed to a prosody encoder module. Energy, pitch, and the reference mel-spectrogram are input features to the prosody encoder. Convolutional layers process these features, then summed and passed on a sequence of FFT blocks and pooled by average. The intermediate representations go through a speaker classifier with a GRL to remove speaker information. Then, a speaker id, representing the output voice timbre along with the learned disentangled prosodic information, is used to predict the scaling and bias parameters (β and γ) of a conditioning FiLM layer (PEREZ *et al.*, 2018).

These parameters are all used to condition the phoneme encoder, the low-level prosody predictor, and the decoder. The phoneme encoder processes the phoneme representations conditioned on the prosodic and speaker representations, consumed by the low-level prosody predictor module, which, for each phoneme, predicts the duration,

energy, and pitch. Then, the Gaussian up-sampling module adjusts the length of the input phoneme sequence to the output frame sequence. The decoder uses these to generate the output mel-spectrogram containing the input text, the prosody of the reference utterance, and the speaker voice given by the speaker id. The model's architecture is shown in Figure 4.6. Each component block's structure is further detailed in Figure 4.7.

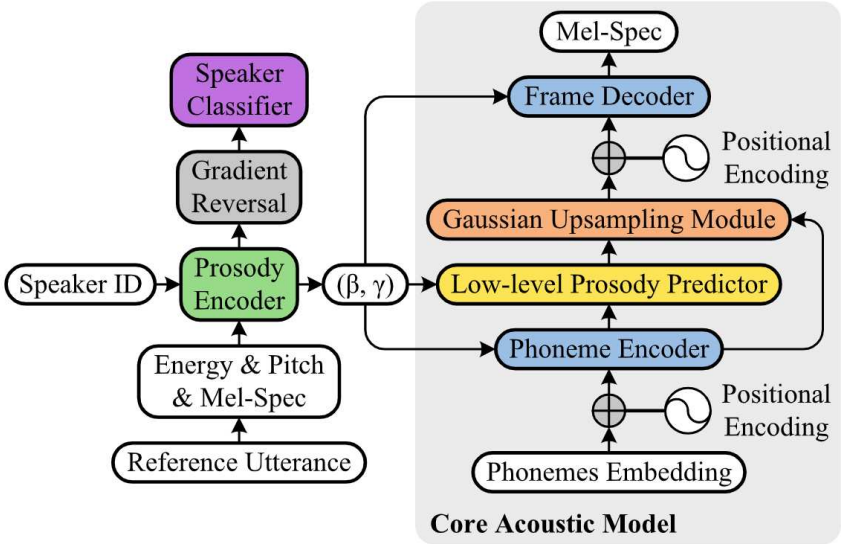


Figure 4.6: Daft-Exprt architecture. Source: Extracted from (ZAïDI *et al.*, 2022).

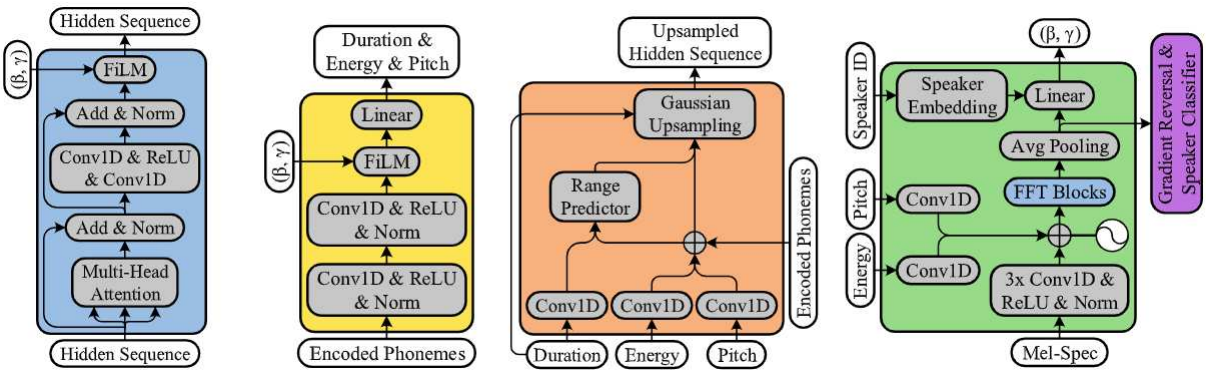


Figure 4.7: Main components of the Daft-Exprt. Source: Extracted from (ZAïDI *et al.*, 2022).

In this work, the Daft-Exprt is trained on the full LJ Speech as the target neutral dataset and the source expressive dataset, ESD. On inference, two methods to synthesize were considered. First, we used the ground truth audio with the exact text as a reference to perform the style transfer. On the second, we used a technique proposed in Kwon *et al.* (2019) of computing the prosody embeddings for all training audios of the ESD passed as references and then performing the inference of a given style by taking the centroid of all training prosody embeddings of that particular style.

4.6.2 Ablations

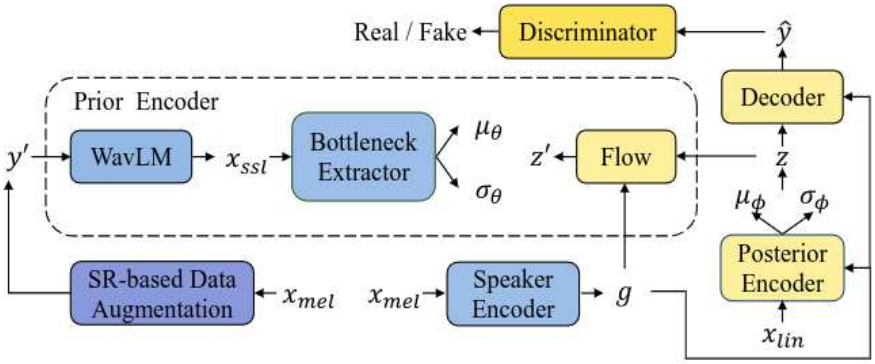
Two ablations are performed to evaluate the impact of some individual components on the complete training scheme: a substitution of the SVC with the usual VC model and a removal of the style filtering step.

An experiment replacing the SVC model with a VC model, the FreeVC (LI *et al.*, 2023), and running the entire pipeline with the same conditions is conducted. This ablation is used to individually evaluate the impact of using the SVC to convert expressive speech. The FreeVC is a model specifically designed for voice conversion, and, in the pipeline, it was also trained solely on speech data. The FreeVC architecture is based on the VITS (KIM *et al.*, 2021) spectrogram reconstruction framework and can learn to disentanglement content information from speaker information with the need of using annotated textual information, named text-free approach.

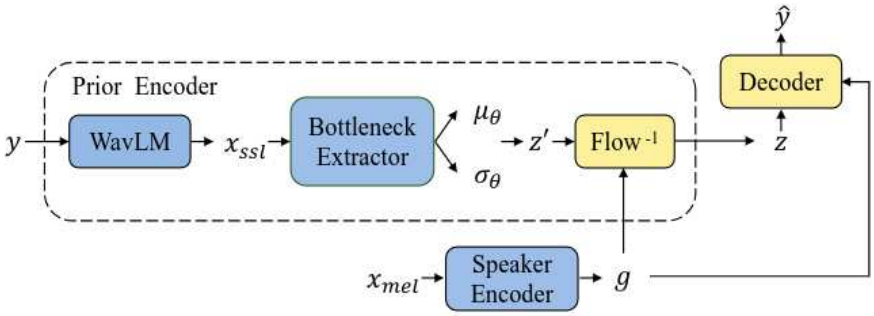
The architecture comprises a prior encoder, a posterior encoder, a decoder, and a speaker encoder. The prior encoder obtains content and speaker information as a normal distribution $\mathcal{N}(z'; \mu_\theta, \sigma_\theta^2)$. It receives the original waveform passed on the WavLM (CHEN *et al.*, 2022) Self-Supervised Learning (SSL) model. Through a bottleneck layer, the WavLM output is projected to a much lower dimension, forcing the representation to discard speaker and noise information, a technique discussed in the Chapter 3 Section 3.2.3. Then, it is projected to two vectors representing the mean and variance of the content distribution. The speaker representation is obtained by passing a mel-spectrogram on the speaker encoder. In the same way on VITS, a normalizing flow is used composed with affine coupling layers and conditioned on the speaker embedding to improve the complexity of the prior distribution. A linear spectrogram is passed to the posterior encoder, also the same used on VITS. This representation is passed to a decoder with the same architecture as on VITS and converted in the raw output waveform. The raw waveform is finally passed to a discriminator that judges it as authentic or fake.

On inference, the reverse flow process passes the desired content as the input waveform and the mel-spectrogram of the desired timbre. By disentangling content from the speaker and using a pre-trained speaker encoder, the VC model can receive independent inputs and generate any given content from any given voice. These procedures are shown in Figure 4.8.

The other ablation conducted was the removal of the style classifier filter from the pipeline. In this case, instead of using only the filtered data to perform the finetuning, all converted styled data is used. This ablation is conducted to validate whether finetuning only on the data that is judged to maintain the same style as judged by a classifier after conversion is efficient to boost the style intensity of the TTS.



(a)



(b)

Figure 4.8: (a) Training procedure of FreeVC. Source: Extracted from (LI *et al.*, 2023). (b) Inference procedure of the FreeVC. Source: Extracted from (LI *et al.*, 2023).

4.6.3 Training Setup

All the baseline models, proposed pipelines and ablations were trained on a machine running a UBUNTU 22.04.3 LTS operating system, equipped with 72 Intel(R) Xeon(R) Gold 5220 2.20GHz Central Processing Unit (CPU)s and 5 NVIDIA Quadro RTX 5000 GPUs, each with 48GB of memory. No distributed multiprocessing training framework was implemented, only one GPU was used in the experiments.

Hyperparameters were used as default in each original implementation[7]. Given that each dataset is available only at different audio sampling rates, a resampling step using the open-source Librosa[8] tool is performed whenever required. For example, to compute the HuBERT representations used in the SO-VITS-SVC, the model requires an input wave sampled at 16kHz. In the VC/ SVC finetuning step, the models were trained for 100 epochs, with a learning rate set to a tenth of the original. All base TTS models were trained for 600k steps from scratch, and the style fine-tunings were performed for 100k

[7]SO-VITS-SVC available at <https://github.com/PlayVoice/so-vits-svc-5.0>,
Daft-Exprt available at <https://github.com/ubisoft/ubisoft-laforge-daft-exprt>,
FreeVC available at <https://github.com/OlaWod/FreeVC>,
FastPitch available at <https://github.com/AI-Unicamp/TTS>

[8]Librosa available at <https://librosa.org/>

steps with a tenth of the original learning rate. The style filter was trained on the ESD until convergence and achieved an accuracy of 84% on a validation set. To equally convert the generated mel-spectrograms into audio, we used the BigVGan vocoder, the model detailed in Chapter 3 Section 3.1.3, pre-trained on the VCTK, LJSpeech, and LibriTTS (ZEN *et al.*, 2019) datasets with a batch size of 32 for 5M steps. The baseline Daft-Exprt model was trained for 500 epochs and took around 5 day to complete the training. The time taken to perform each step as well as the total time for pipeline completion is shown on Table 4.2.

Pipeline Step	**Duration**
SVC fine-tuning	27h
Style classifier training	2h
F_0 neutral semitone distances	0.5h
F_0 expressive semitone distances	0.2h
F_0 neutral matched conversion	4h
F_0 expressive matched conversion	3h
Neutral TTS	100h
Style fine-tuning TTS	20h
Total	**$\sim$ 6.5 days**

Table 4.2: Time taken to process each step of the proposed pipeline.

4.7 Concluding Remarks

In this Chapter, the proposed pipeline to perform cross-speaker style transfer based on data augmentation was thoroughly described. All steps were detailed, as well as which datasets were used. Also, experimentation aspects, regarding which models were compared to, which ablations were performed and the computational training setup that was used were presented. The next Chapter presents how the perceptual assessment was conducted to rate all obtained models described previously.

Chapter 5

Evaluation

This chapter presents the procedure and results of a perceptual evaluation conducted to assess the synthetic expressive speech audios generated by the implemented cross-speaker style transfer method. Some objective metrics used in earlier development stages of the proposed method are also exploited.

To fully exhibit a human-like way of communicating, the synthesized expressive speech audios should simultaneously present high levels of expressiveness, sound natural, and ensure the target speaker's timbre (goal of the cross-speaker tasks).

Several objective metrics can quickly provide insights into crucial aspects of speech, such as:

- Naturalness: UTokyo-SaruLab MOS Prediction System (UTMOS) (SAEKI *et al.*, 2022), a popular SOTA system trained to automatically predict Mean Opinion Score (MOS) based on an ensemble of strong and weak learners.

- Intelligibility: Character Error Rate (CER) and Word Error Rate (WER), which consist in transcribing the synthesized audios with an ASR model, then calculating the percentage of wrong characters/words compared to the original transcriptions, meaning how much an ASR system could "understand" the synthesis; Short-Time Objective Intelligibility Measure (STOI) (TAAL *et al.*, 2011), a metric developed to measure intelligibility in denoised speech;

- Speaker Similarity: cosine distance between speaker embeddings (DEHAK *et al.*, 2010), which is based on extract meaningful representations of speaker timbre then measuring the distance between them.

- Similarity to Ground-Truth: Mel Cepstral Distortion (MCD), the distance between sequences of mel cepstral coefficients; Voicing Decision Error (VDE), a proportion of frames of the synthesized speech whose voicing is different from the ground-truth; Gross Pitch Error (GPE), which consists in measuring the percentage of frames from the synthesized speech whose pitch differs from the ground-truth above a

predefined threshold; F0 Frame Error (FFE), defined as the percentage of frames who either contain a voicing error or a pitch error (above a certain predefined threshold).

Even though these metrics are very practical, they are very limited in that they either only focus on a single individual aspect while ignoring all others, or do not have a transparent computation method; that is, they are not sufficiently interpretable. As a matter of fact, no objective metric can fully encompass accurate human perception.

Subjective experiments, on the other hand, are specifically designed to assess the perception of a population regarding the presented stimuli. Regarding speech synthesis, perceptual evaluations can be adequately used to reason about speech audios' expressiveness, since these are designed to emulate expressions that ought to be rightfully perceived by humans themselves. Although these assessments are, most of the time, expensive and time-demanding, they end up being the most adequate way to evaluate systems designed for human interaction.

The studies on data augmentation for the cross-speaker style transfer task, as discussed in Chapter 3, opted to evaluate with human subjects, with varying numbers of participants and different protocols and stimuli. Besides being the most straightforward and adequate way to assess emotions in speech, no target data is available for the cross-speaker task since it is assumed there is only neutral data available for the target speaker, which make various objective metrics unsuited. Following the same approach, we conducted a perceptual evaluation to assess the proposed method's performance contribution in crucial aspects of speech, such as naturalness, expressiveness, and speaker similarity. Also, to quickly rate the speaker similarity capability of the model in several different model configurations, an objective metric based on speaker embedding similarity is evaluated.

The following stimuli were considered: (1) high and low anchors (when available), which are audios that are representatives of highest and lowest achievable ratings of the aspect being evaluated, to relativize the ratings of the other stimuli; (2) the synthesized expressive audios of the complete pipeline this work proposed, with some ablations to analyze the effectiveness of each introduced module, and (3) audios from baseline models, to compare the proposed method with models that focus on the same challenge.

In this scenario, the participants were asked to evaluate all types of audio on Likert scales, from one to five, providing a direct comparison of the proposed method against both the perfect and worst scenarios possible, provided by the anchors, and also against the re-implemented state-of-the-art models from literature as well as ablations.

This chapter is organized as follows: Section 5.1 presents the evaluation protocol adopted, the test population selection criteria, and the tools used to conduct the assessment. Section 5.2 describes the distribution of the stimuli used in the evaluation, as well as anchors, the proposed model, ablations, and baseline models. Section 5.3 details all the results acquired from the test and discusses how they relate to each evaluated speech aspect.

5.1 Perceptual Protocol

The evaluation platform was built and designed upon the webMUSHRA (SCHOEFFLER *et al.*, 2018) evaluation platform. This platform is an open-source MUSHRA compliant web audio API-based experiment software, written in JavaScript and PHP, that is used to create perceptual experiments to assess the audio quality of audio samples with various experiment configurations, such as MUSHRA, A/B comparison, and Likert scale. The application was designed and tested locally, then deployed online with the Hostinger[1] platform, a provider of web hosting solutions. Through Prolific, an online crowd-sourcing platform, thirty-two native English test takers were recruited to participate in the assessment. This quantity was set based on literature, budget and since experiments conducted previously shown a convergence of scores after the twenty-fifth participant. After the test, the results were saved in a file in .csv format and stored on cloud servers, following the format defined in the standard webMUSHRA (SCHOEFFLER *et al.*, 2018) tool. This study was conducted with the approval of the UNICAMP Ethical Review Board, under project number (CAAE) 59536022.8.0000.5404.

Each audio page considered is presented in the Appendix A. Before the start of the assessment, the raters were instructed to leave and not carry on with the test if one of the following conditions was not attended:

1. The participants did not have headphones,
2. The participant's environment has compromising background noise,
3. The participant had any type of hearing impairment,
4. For some reason, the participant could not hear the audio samples.

The rater was asked to adjust the volume level and proceed if all conditions were met. Before the experiment's instructions, the raters were warned of the hidden attention checks throughout the assessment. These attention checks were carefully designed to ensure the takers actively listened to the audio content. In the attention checks, disguised as a regular experiment page, one of the audios would state: "Please rate all the audios on this page with a score of four". If the taker does not follow these instructions, then disqualification will happen. Two test takers' entries were rejected due to unmatched attention checks.

The test takers would start the experiments after being informed about the attention checks. For each experiment, the instructions pages were shown first to tell the participants what they would listen to, what aspects of the audio they were supposed to rate, and then the audio pages themselves to be ranked.

The audio pages were composed of several rows, each with audio controls to play and pause the stimuli, an audio progress bar, and then buttons from one to five

[1]Available at: <https://www.hostinger.com/>

corresponding to the audio rating. The subjects were free to play and pause the audio as many times as necessary; no time limit was enforced. The rating buttons were only unlocked to rate after the corresponding audio was played to ensure listening. Also, the rater could only advance to the next page after rating all the audio. On all pages, a progress bar showing the percentage of the completed assessment was always displayed for the test takers.

In the assessment, three experiments were performed: a style intensity test to quantify how much of each emotion of the dataset was present in the synthesis, a naturalness test to measure how human-sounding the generated audios were, and a speaker similarity test to analyze how close to target speaker's voice the synthetic audios were. A MOS-based metric was employed on all tests: the mean of all given scores. All audio rows and pages within the same emotion are randomized in this experiment.

Firstly, the participants underwent the style intensity test. In this task, the raters first were shown, for each emotion, actual samples of the ESD dataset to understand what the perfect case for each emotion sounds, and then were asked to rate how much of each emotion the audios sounded in a 5-point intensity scale with the corresponding associated labels: (1) Not at All Happy; (2) Very Little Happy; (3) Somewhat Happy; (4) Notably Happy; and (5) Very Much Happy.

Then, in the naturalness experiment, the raters were asked how naturally each audio sounded. They were instructed to interpret the question as: "How likely could this be a real person speaking?". Also, a five-point scale with corresponding labels was used: (1) Very Artificial; (2) Somewhat Artificial; (3) Neither Artificial nor Natural; (4) Somewhat Natural; and (5) Very Natural.

Finally, to end the assessment, a speaker similarity test was conducted. In this experiment, the reference audio of the target speaker (LJ) was also made available to the participants, and they were asked to rate all the audios on the page based on how much they resembled the voice of the reference speaker. For this, the following 5-point rating scale was used: (1) Definitely Not the Reference Speaker; (2) Probably Not the Reference Speaker; (3) Possibly the Reference Speaker; (4) Probably the Reference Speaker; and (5) Definitely the Reference Speaker.

Upon conclusion, to perform a simple population profiling, some basic participant information, such as age and sex, was extracted. Out of the thirty participants that completed the assessment, twenty-one identified as Female and nine as Male. The age profile ranged from eighteen to fifty-eight years, with a mean age of thirty-five and a median of thirty-three years. The median time to complete the assessment was twenty-five minutes and six seconds. The time taken to answer each audio page was also captured. The fastest answered page (part of the naturalness test) was done in seventeen seconds, and the slowest (part of the speaker similarity test) took a participant four hundred and six seconds. The overall time spent on an audio page was approximately forty-two seconds.

5.2 Stimuli

The test set consisted of selecting text-audio pairs from the ESD and LJSpeech datasets. These text-audio pairs were not used in the development phase of the proposed method for neither training nor validating the models and were used solely for the perceptual evaluation, as either examples of input texts or high anchor ground-truth audios. For all models evaluated (baselines, ablations, and proposed), the text transcriptions of corresponding selected audios were used as inputs to generate the synthetic audios used in the evaluation. All audios extracted from the original datasets were re-synthesized with the BigVGan (LEE *et al.*, 2023) vocoder so that possible vocoding artifacts introduced on the synthetisized audios also appeared on the original audios, and were not taken into account by the raters to differentiate between the original and synthesized audios.

Table 5.1 presents a distribution of the text-audio pairs used as the test set. In total, forty-one pairs of samples were selected. Apart from the LJ and the Speaker 13 of the ESD dataset, the test samples were drawn uniformly across style, speaker, and audio duration factors. These extra audios were used as input references of high and low anchors for the speaker similarity experiment, explained in the next paragraph.

	Speaker											
Style	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**LJ**	**Total**
Angry	1	1	0	0	0	1	1	0	1	0	0	5
Happy	1	0	1	1	0	1	0	1	0	0	0	5
Neutral	0	0	9	1	1	1	0	0	0	1	16	29
Sad	0	1	0	1	2	0	0	0	0	1	0	5
Surprise	1	1	0	0	0	0	1	1	1	0	0	5
Total	3	3	10	3	3	3	2	2	2	2	16	49

Table 5.1: Distributions of the text-audio pairs selected for perceptual evaluation.

To validate the contribution of the style filter, we conducted an ablation study in which the pipeline is executed without the style classifier filtering step, that is, the whole converted ESD is used on the fine-tuning, and also the latter plus the replacement of SVC by an open source state-of-the-art VC model, FreeVC (LI *et al.*, 2023), pre-trained on the VCTK. High anchors (directly drawn from the test set) and low anchors are adjusted to each type of experiment to calibrate the rating scale with performance boundaries.

The proposed method was compared to Daft-Exprt, an open-source state-of-the-art cross-speaker prosody transfer model designed to capture high and low-level prosodic features such as pitch, duration, and energy. Instead of using data augmentation, the authors attempt to disentangle speaker information from the prosodic information through adversarial training with a gradient reversal layer. The model was trained on the full LJSpeech dataset and the expressive dataset (ESD). On synthesis, two methods were considered. First, we used the ground truth audios with the exact text as a reference

to perform the style transfer, named Daft-Exprt (Reference). On the second, we used a technique of computing the prosody embeddings for all training audios of the ESD passed as references and then performing the inference of a given style by taking the centroid of all training prosody embeddings of that particular style, named Daft-Exprt (Centroid).

For the style intensity experiment, the participants evaluated a total of 84 stimuli, 7 per experiment page. As a high anchor, we used a ground-truth audio from the ESD, and as a low anchor of emotion, a neutral audio is synthesized with the neutral base LJ model with the same text used. Also, the same text from these utterances was used as input for both Daft-Exprt-based baselines, the proposed model, and its ablations. Three text/audio pairs from each of the four emotions were used.

In the naturalness experiment, a total of 30 stimuli were evaluated.A ground-truth neutral audio from the ESD dataset was used as a high anchor. In this experiment, no low anchor is used. The exact text of the high anchor is passed as input to both baseline, proposed, and ablation models. Thus, five neutral text/audio pairs from the ESD are used.

Finally, the speaker similarity experiment consists of eight pages of audio, two per emotion. Apart from the audios to be rated, a ground truth audio from the LJ dataset is used as a reference speaker on each page. On each page, seven stimuli are considered. As a high anchor, we use another audio also from the LJ dataset, and as a low anchor, we use audio from the Speaker 13 of the ESD. This speaker was considered as low anchor due to having the lowest similarity out of all ESD speakers with respect to LJ, as measured by a cosine distance between speaker embeddings obtained with the Resemblyzer (WAN *et al.*, 2018) model. Two text samples for each emotion were used as input for the proposed baseline and ablation models to generate the audio. In total, sixteen audios from the LJ speech, eight audios from the Speaker 13, and two texts from each of the emotions were used in this experiment.

5.3 Results

5.3.1 Naturalness

Naturalness results are shown in Table 5.2. The GT model obtained the best score, reiterating its high anchor position. It received a naturalness value of 4.05 ± 0.18, and thus, this value limits the value of the other models since it is the rating of the neutral audios taken directly from the ESD dataset. The Daft-Exprt-based models obtained the lowest scores, of 2.01 ± 0.18 and 2.21 ± 0.20, with the reference and centroid-based techniques, respectively. Regarding this difference, we found that, when compared to the centroid technique, the reference-based Daft-Exprt ended up being much more sensitive to the input audio in a way that, when trying to forcefully copy the prosody to texts of possibly different lengths from the reference, ended up generating unrealistic speech.

Table 5.2: Naturalness MOS with 95% confidence intervals.

Model	MOS
GT (High Anchor)	4.05 ± 0.18
Daft-Exprt (Reference)	2.01 ± 0.18
Daft-Exprt (Centroid)	2.21 ± 0.20
VC	3.02 ± 0.20
SVC (This work)	$\mathbf{3.57 \pm 0.20}$

After the high anchor, the highest naturalness scores were obtained with the models based on data augmentation. The model based on the FreeVC (LI *et al.*, 2023) obtained a naturalness of 3.02 ± 0.20, and the proposed model, based on SVC, obtained a naturalness MOS of 3.57 ± 0.20. Thus, it is seen that the proposedSVC-based pipeline significantly improved the naturalness of the synthetic audios by a significant margin when compared to both baselines and ablation.

5.3.2 Style Intensity

Style Intensity results are shown in Figure 5.1, and summarized on Table 5.3. In a similar fashion to the naturalness experiments, the score obtained by the selected anchors also reiterated their appropriate choice. The high anchors (audios from the ESD dataset) received the best style intensity scores in all four emotions of the ESD. This also reiterates how the emotions on the ESD are compatible with how our test takers perceive them. The low anchors scored worse than all models for all four emotions.

Table 5.3: Style Intensity Mean Opinion Scores MOS with 95% confidence intervals.

	MOS			
Model	Angry	Happy	Sad	Surprise
GT-Res (High Anchor)	3.85 ± 0.24	4.22 ± 0.22	4.26 ± 0.23	4.66 ± 0.13
Neutral-Res (Low Anchor)	1.81 ± 0.19	1.86 ± 0.19	1.78 ± 0.21	1.34 ± 0.12
Daft-Exprt (Reference)	1.97 ± 0.26	2.43 ± 0.25	$\mathbf{3.28 \pm 0.30}$	2.72 ± 0.25
Daft-Exprt (Centroid)	2.16 ± 0.27	2.00 ± 0.18	3.19 ± 0.27	2.40 ± 0.19
VC	2.29 ± 0.23	$\mathbf{3.53 \pm 0.25}$	2.61 ± 0.24	$\mathbf{4.32 \pm 0.19}$
SVC (This work)	1.94 ± 0.21	3.06 ± 0.21	2.61 ± 0.24	2.76 ± 0.22
SVC + Filtering (This work)	$\mathbf{2.69 \pm 0.29}$	2.00 ± 0.20	2.79 ± 0.24	3.13 ± 0.26

Analyzing the scoring of the models by emotion, for the "angry" style, the SVC + Filtering model (proposed) performed better (MOS score of 2.69 ± 0.29) than all its counterparts by a large margin. This behavior is attributed to the fact that, due to having already seen different phonation modes presented on the pre-training of the SVC, the proposed pipeline was able to capture better the phonation modes. This was due since out of all the emotions present on the ESD dataset, the "angry" style was the only shown to

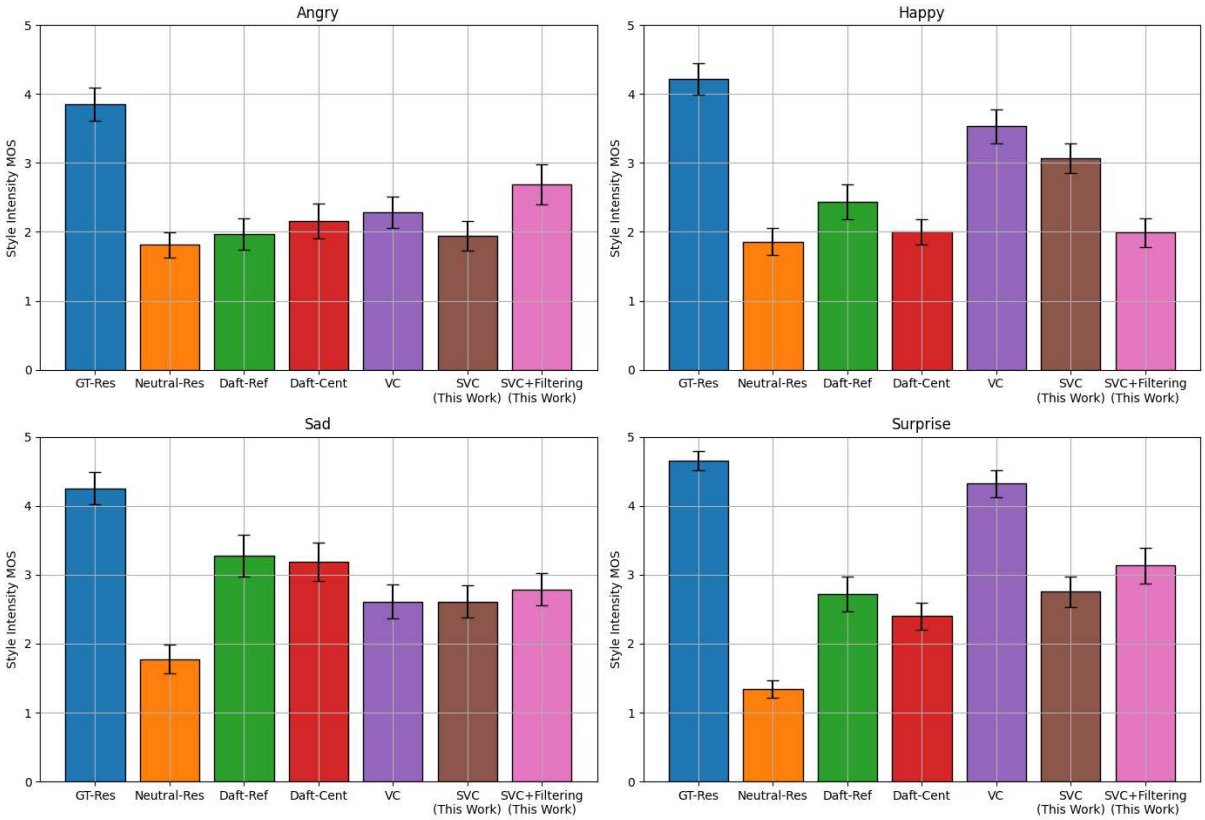


Figure 5.1: Style Intensity MOS results for each style and each stimulus with 95% confidence intervals.

require a specific phonation mode (pressed) by Birkholz *et al.* (2015) for correct perception while the other emotions' perception could rely primarily on other prosodic parameters.

On the "happy" and "surprise" styles, however, the VC-based pipeline outperformed both the proposed models and baselines, with MOS scores of 3.53 ± 0.25, and 4.32 ± 0.19, respectively. We hypothesize that these are styles that are characterized by having more significant and faster variations of F_0 that are not present in singing voices which are characterized by longer continuous pronunciations (HUANG *et al.*, 2021). Specifically, a pattern of quickly rising tone in the final syllable is observed in most of the sentences of the "surprise" emotion. The proposed model achieved the second-best MOS score on the "surprise" style, and the ablation model scored the second-best MOS on the "happy" emotion, beating baseline models in both cases.

In the Sad case, the Daft-Exprt baseline models outperformed all data augmentation-based methods, especially when considering the reference-based synthesis, which scored the best MOS (3.28 ± 0.30). This fact was likely due to the model's ability to copy the duration aspect, which is crucial for this style and is characterized by slower speaking rates, even though the data-augmentation methods were perceived to have better captured the phonation aspect of the style.

Further, in this experiment, we see the effectiveness of the style classifier filter,

once that, for all styles but the "happy", these models (SVC + Filtering) boosted the intensity MOS when compared to the SVC without the filter (ablation). A possible explanation, to be further investigated, is that the loss in MOS on the "happy" can be attributed to the very low volumetry of the filtered "happy" style (seventy-seven audios), possibly not being enough to learn its defining patterns during the style finetuning. Future experiments to analyze how the MOS varies as a function of the amount of finetuning style data will be conducted.

Style Filtering

Confusion matrices of the classifier style filter trained on the source expressive dataset were obtained. On Figure 5.2 it can be seen that, on the validation set of the real audios, it obtained a good performance with very few mistakes compared to the number of correct predictions (main diagonal of the matrix).

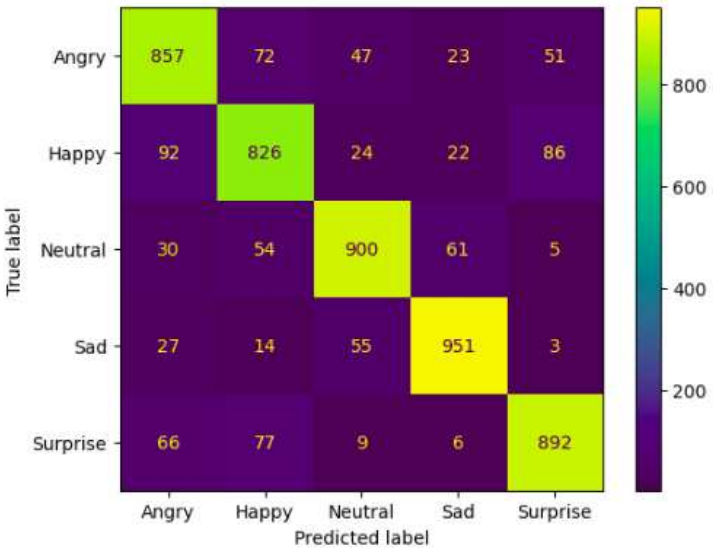


Figure 5.2: Confusion matrix of the style classifier filter on the validation set.

When applied to the synthetic audios, the confusion matrix shown on Figure 5.3, 571 converted "angry" files were rightfully labeled, an thus selected for the next fine-tuning step; 74 for the "happy" style; 2230 audios for the "sad" style; and 99 audios for the "surprise" style.

Notably, a tendency to classify the synthetic audios as "sad" is observed. This can be attributed to a possible loss in expressiveness caused by the conversion together leading various expressive audios to a "neutral" style combined with a confusion of the style classifier between the original "neutral" and "sad" styles.

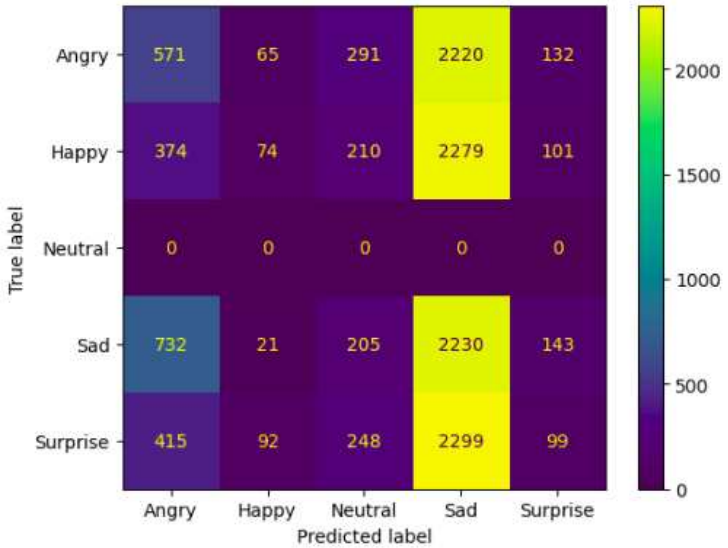


Figure 5.3: Confusion matrix of the style classifier filter applied on synthetic converted expressive dataset.

5.3.3 Speaker Similarity

Subjective results of the perceptual experiments for speaker similarity are shown on Figure 5.4 and on Table 5.4. Similarly to the other experiments, the anchors' choice was also appropriate. The high anchor (re-synthesized audios of the original LJ dataset) received the highest similarity MOS scores for all emotions, and the low anchor (Speaker 13 of the ESD dataset, named "Other-Res") received the lowest scores in all cases.

Table 5.4: Speaker Similarity MOS with 95% confidence intervals.

	MOS			
Model	Angry	Happy	Sad	Surprise
LJ-Res (High Anchor)	3.90 ± 0.30	4.47 ± 0.22	4.28 ± 0.28	4.40 ± 0.27
Other-Res (Low Anchor)	1.22 ± 0.23	1.13 ± 0.19	1.00 ± 0.00	1.00 ± 0.00
Daft-Exprt (Reference)	$\mathbf{2.23 \pm 0.32}$	2.55 ± 0.33	$\mathbf{2.23 \pm 0.30}$	2.26 ± 0.32
Daft-Exprt (Centroid)	$\mathbf{2.33 \pm 0.34}$	$\mathbf{2.63 \pm 0.33}$	2.05 ± 0.30	$\mathbf{2.67 \pm 0.36}$
VC	1.58 ± 0.25	1.48 ± 0.24	1.62 ± 0.23	1.43 ± 0.18
SVC (This work)	1.68 ± 0.24	1.71 ± 0.25	1.50 ± 0.18	1.53 ± 0.23
SVC + Filtering (This work)	1.82 ± 0.30	1.67 ± 0.28	1.47 ± 0.16	1.43 ± 0.23

The baseline models performed better in this experiment than the analyzed data augmentation methods in all styles. However, our proposed data-augmentation-based models achieved competitive results even though they used only five minutes of the target speaker's voice during training. In contrast, the Daft-Exprt used almost the entire LJSpeech dataset (approximately twenty-four hours of data), accounting for the audios removed for testing. Additionally, we discovered that even lower amounts of target speaker data also worked, albeit with a trade-off in decreased speaker similarity.

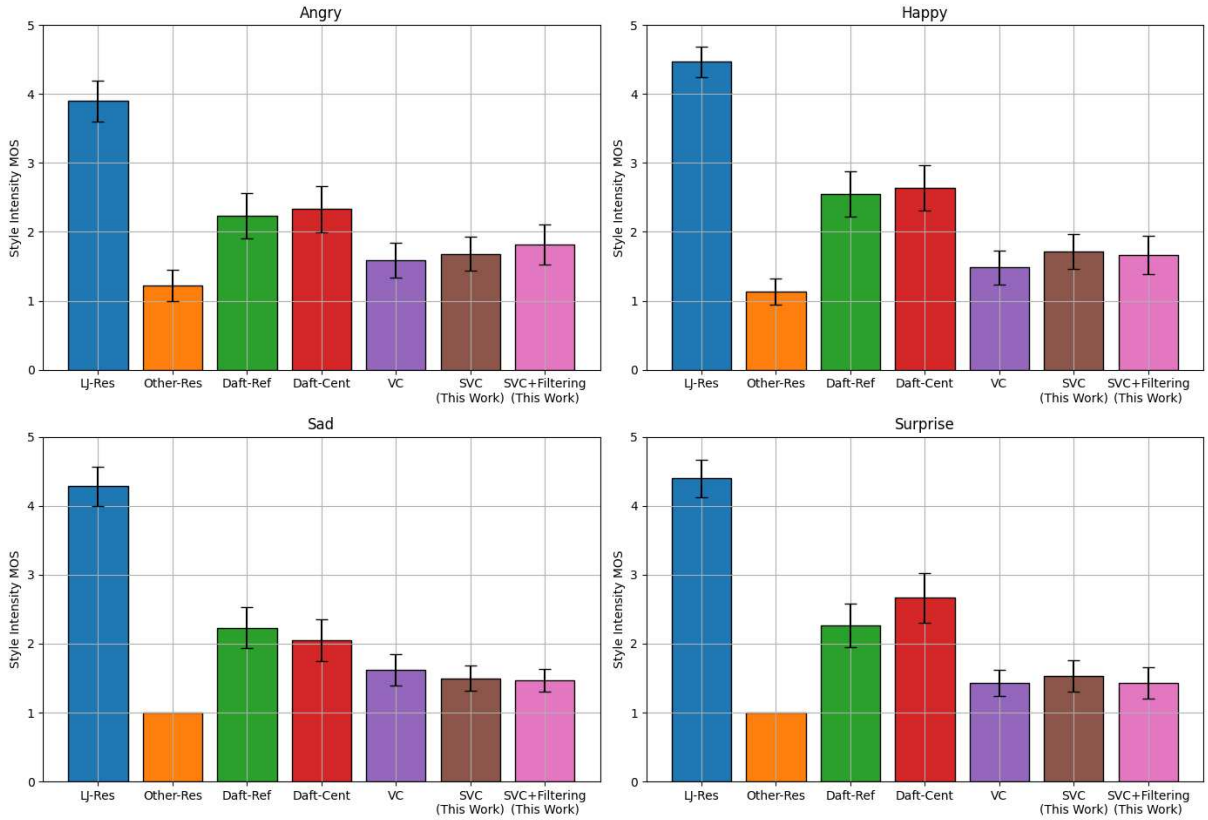


Figure 5.4: Speaker Similarity MOS results for each style and each stimulus with 95% confidence intervals.

Speaker Embedding Similarity

During the early development stages, a speaker similarity objective metric was used to assess how the proposed method performed quickly. Speaker embeddings were calculated using test and real LJ utterances. Then, a cosine similarity, that is, how similar the orientation of two vectors is, regardless of their magnitude, is calculated. This metric is used to quantify how close both speaker embeddings are in the speaker embedding space created with the Resemblyzer[2], from 0 to 1. (WAN *et al.*, 2018). Intuitively, it means how much their timbres are alike. It is calculated according to the following equation:

$$similarity(\mathbf{u}, \mathbf{v}) = \cos\left(\angle(u, v)\right) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} \tag{5.1}$$

After the pipeline was fully developed, objective speaker similarity results were calculated with the obtained cross-speaker style transfer method. These are shown in Figure 5.5. Each dot in the Figure represents the similarity between an utterance with the original LJ voice and a test utterance obtained from a conversion of a VCTK speaker to LJ's voice. These are drawn from the pipeline's neutral-to-LJ conversion step, with and without F_0 matching.

[2]Available at: <https://github.com/resemble-ai/Resemblyzer>

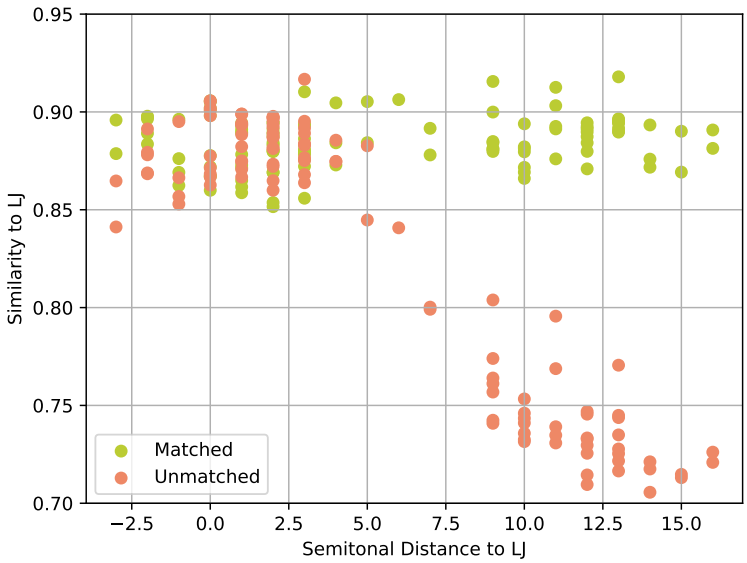


Figure 5.5: Similarity for each converted speaker.

The similarity scores are plotted with the speakers in the x-axis being arranged in ascending order according to their semitonal distance to the LJ (as calculated in the F_0 matching conversion step). This plot shows a decline in speaker similarity as the semitonal difference between LJ and the speaker increased. However, when the F_0 matching algorithm is applied to the conversion, the similarity levels remain unaffected by the timbral difference between the speakers, demonstrating the effectiveness of the proposed F_0 matching technique.

5.4 Concluding Remarks

This chapter described the adopted subjective evaluation protocol to assess the synthetic expressive audios generated by the proposed cross-speaker style transfer pipeline. Audios were either ground-truth, that is, re-synthesized from the original datasets, to serve as either anchors or speaker references, or generated with the evaluated models taken from the text of the test set. Thirty-two native English speakers (with two failing the attention checks) were recruited to judge three aspects of the audio: naturalness, style intensity, and speaker similarity. All three factors were ranked in a MOS-like scale from one to five. On all pages, when available, the stimuli used were high and low anchors to relativize the scores, two versions of a baseline model, two ablations, and the proposed model itself.

Also, this chapter went through an objective speaker similarity metric calculated to perform speaker evaluation in both early stages of the development and to compare a large quantity of audio. Specifically, converted neutral audios from VCTK to LJ's voice were compared both when using or not the F0 matching algorithm. From the result

analyses, the proposed pipeline outperformed all other ablations and baseline models regarding naturalness. Concerning style intensity, the pipeline obtained the best score for the “angry” style and the second-best score for all styles. Regarding speaker similarity, even though it used only five minutes of target speaker’s data, the pipeline could still obtain competitive results compared to the baseline models, which used twenty-four hours of target speaker training. Also, the F0 matching technique improved speaker similarity when considering the critical cases: when the target and source speaker’s voices are more different regarding the fundamental frequency register.

Chapter 6

Conclusions

This work focused on developing an expressive TTS model for a speaker that is assumed to have only a few minutes of neutral data available. Data augmentation techniques were used to take advantage of existing expressive data from other speakers to create synthetic expressive data for the target speaker. In this context, a perceptual evaluation and several experiments were carried out to compare the models considered and to tackle the research questions introduced in Chapter 1 Section 1.4.

A novel pipeline to perform cross-speaker style transfer was presented. Aiming to improve the performance of the technique when considering highly expressive styles, such as emotion, an SVC model was introduced as part of the pipeline, replacing the VC in standard SOTA data augmentation-based methods. The SVC model was first pre-trained on a dataset with various voices to quickly learn the target speaker's voice, reducing the required volumetry to only a few minutes. A style classifier filter trained on the original expressive data was proposed to filter out all audios that changed style after being converted to the target speaker's voice so that only data that matches the original style was used. Additionally, a technique to reduce the tonal mismatch between speakers with different vocal frequency ranges were proposed, enabling adequate generation of the synthetic expressive data recorded in any speaker's voice.

The models were trained on the ESD (ZHOU *et al.*, 2022a) dataset as the source expressive speakers and using the LJ speech dataset (ITO; JOHNSON, 2017) as the target speaker. The proposed pipeline was evaluated against another open-source SOTA baseline for cross-speaker style transfer, Daft-Exprt (ZAïDI *et al.*, 2022). Two inference methods were considered: one using the centroid of each style as the style embedding and the other extracting the style of a reference speech for style. Also, ablations of the removal of the style filtering and the replacement of the SVC with a VC model were considered in the experiments.

A perceptual evaluation with over 30 native English participants was conducted using subjective measures of naturalness, style intensity, and speaker similarity based on MOS. Also, an objective metric for speaker similarity based on embedding cosine

similarity was computed.

The findings detailed in Chapter 5 revealed that the proposed pipeline ameliorated several aspects of the cross-speaker style transfer task, such as naturalness and style intensity, especially when considering highly expressive scenarios while using a volumetry of neutral data of target speaker as low as five minutes and enabling the conversions with very different speakers in terms of timbre.

Finally, this work is concluded by answering the research questions proposed based on the results obtained from the experiments performed:

- **Q.1) Can data augmentation-based techniques perform cross-speaker style transfer of highly expressive speaking styles with only a few minutes of neutral data of target speaker?** Comparing the style intensity scores obtained by the models (Figure 5.1), it can be seen that approaches based on data augmentation through the use of a pre-trained SVC or VC model (conditions VC, SVC, and SVC+Filtering), indeed allowed the development of an expressive TTS on target speaker's voice with the use of only five minutes of its neutral data, as for at least one condition out of the three, a score whose the lower limit of the confidence interval is greater than the upper limit of the neutral condition, demonstrating that emotion was perceived by the listeners to some extent.

- **Q.2) Since singing voice includes richer emotional information compared to regular speech (HUANG *et al.*, 2021), is an SVC model (instead of a VC), more effective to preserve the speaking style when converting expressive speech to a speaker with only neutral data?** The style classifier confusion matrix Figures 5.2 and 5.3 showed that even though most of the conversions to target speaker's voice are mapped to the "sad" style, several audios still preserved the style after conversion. However, the style intensity MOS scores shown on Figure 5.1 revealed that, the the inclusion of the SVC+Filtering, performed statistically better only on the angry style (highest vocal effort), as it was the only case in which no intersection between its confidence interval and all other compared models occurs. Still, in the other three styles, the SVC-based models achieved at least the second greatest mean MOS value.

- **Q.3) Does filtering out the synthetic audios that do not maintain the same style after being converted, judged by a style classifier trained on the original audios, improves the style intensity of the TTS?** The style intensity experiment, presented on Figure 5.1 showed that the model finetuned only on filtered converted data had a difference in mean MOS of +0.75 for the angry style, −1.06 on the happy style, +0.18 on the sad style and +0.37 on the surprise style, in

comparison to the model trained on all converted data. However, when considering the confindence intervals, the only significant difference was found on the angry case.

- **Q.4) To what extent does the difference in timbre between a source speaker and a target speaker impact the perceived similarity of the converted speech to the target speaker, as measured by a speaker similarity metric?** Objective metrics calculated directly on the conversions showed that with as the semitonal distance from the source speaker to target increases, the speaker similarity of the conversion with regards to original recordings can decrease up to around 23% when compared to the matched conversion, as shown in Figure 5.5 which preserved the speaker similarity for any source speaker.
- **Q.5) How do current open-source cross-speaker style transfer approaches perform on open-source data?** The experiments showed the current open-source SOTA method Daft-Exprt (ZAïDI *et al.*, 2022) obtained a performance below than expected in terms of naturalness according to the original paper. The Daft-Exprt model obtained a naturalness of 74% the value of the ground-truth recordings when using an internal dataset, whereas our evaluation that used the same model implementation obtained 55% of the high anchor recordings.

6.1 Limitations and Future Work

While satisfying results were obtained with the proposed SVC-based pipeline, to achieve production-level cross-speaker style transfer capabilities, it still require enhancements in several directions.

The first limitation in the proposed pipeline consists in its inherent complexity. To be assembled, the pipeline relied on a considerable amount of deep learning models (e.g. HuBERT, So-VITS-SVC, BigVGAN, Fastpitch, Whisper, CREPE) hosted at different repositories, as presented on Section 4.6.3. This complicates its computational implementation, as each model demands its own particular set of python libraries, CUDA drivers, memory requirements, audio sampling rates, and hyperparameter configurations.

Another limitation regarding style modeling is the use of TTS models which were not specifically designed specifically for stylistic speech, such as the considered Fastpitch model. Even though it is able to better model prosodic content than other counterparts, with the main loss function that it is trained upon, the MSE, the model ends up over-smoothing the expressive content, converging to an average of the style distribution present on the training data. This way, the richest stylistic variations present on the data end up not being accounted.

Also, a common issue in expressive speech synthesis in the open research setting is

the lack of data available to train and test the models. Most of the datasets available either have a low volumetry per speaker or lack style diversity. Incorporating datasets with more diversified styles and more accurately designed for realistic human interaction scenarios could further enhance its applications and generalization capability. Future efforts should focus on curating and utilizing large-scale, high-quality datasets encompassing a broader spectrum of vocal characteristics and expressive styles rather than only basic emotions.

Even though the proposed pipeline has shown competitive performance with only 5 minutes of neutral speech in the target speaker's voice, further advancements are still needed to lower this value to the few-shot or even zero-shot scenarios. Techniques based on vector representation quantization or an increase in the number of voices viewed during the pre-training phase are some perspectives that could tackle this issue.

In addition, developing real-time processing capabilities for the obtained TTS model, the cross-speaker style transfer pipeline product, would significantly expand its practical applications. Real-time processing is particularly relevant for interactive systems such as virtual assistants, live dubbing, and teleconferencing tools. Achieving low-latency speech synthesis requires optimizing the computational efficiency of the TTS models. Exploring techniques such as model compression, quantization, and more efficient neural network architectures could be the key to achieving this goal.

Refining the evaluation protocols is also an important area for future research. While objective metrics provide a quick yet shallow analysis of model performance, and current subjective metrics, provide more reliable results, yet with not well-defined concepts, developing more transparent, easier-to-understand, and effective subjective evaluation metrics could lead to a better understanding of the strengths and limitations of different style transfer approaches. This could include creating more detailed and thorough perceptual tests with more test takers, more detailed instructions, and a new test to evaluate each aspect of expressive synthetic speech separately. Improved evaluation methods will enable more precise tuning of model parameters and better assessment of progress in the field.

By pursuing these future research directions, cross-speaker style transfer could provide expressive TTS faster, generalizable, easily adaptable, and more reliable systems. Consequently, the insights gained from these efforts will enhance the technical capabilities of expressive speech synthesis systems and contribute to their integration into various real-world applications.

References

ABDUL-KADER, S. A.; WOODS, J. C. Survey on chatbot design techniques in speech conversation systems. *International Journal of Advanced Computer Science and Applications*, The Science and Information (SAI) Organization, v. 6, n. 7, 2015.

ADIGWE, A.; TITS, N.; HADDAD, K. E.; OSTADABBAS, S.; DUTOIT, T. *The Emotional Voices Database: Towards Controlling the Emotion Dimension in Voice Generation Systems*. [S.l.]: arXiv, 2018.

APRESYAN, M. On the concept of "expressiveness" in modern linguistics. *Annals of Language and Literature*, v. 2, n. 4, p. 8–12, 2018.

ARK, S. O.; DIAMOS, G.; GIBIANSKY, A.; MILLER, J.; PENG, K.; PING, W.; RAIMAN, J.; ZHOU, Y. Deep voice 2: Multi-speaker neural text-to-speech. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 2966–2974. ISBN 9781510860964.

ATAL, B. S.; HANAUER, S. L. Speech analysis and synthesis by linear prediction of the speech wave. *The journal of the acoustical society of America*, Acoustical Society of America, v. 50, n. 2B, p. 637–655, 1971.

AYLETT, M. P.; CLARK, L.; COWAN, B. R.; TORRE, I. Building and designing expressive speech synthesis. In: *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*. [S.l.: s.n.], 2021. p. 173–212.

BADLANI, R.; ŁAŃCUCKI, A.; SHIH, K. J.; VALLE, R.; PING, W.; CATANZARO, B. One tts alignment to rule them all. In: IEEE. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2022. p. 6092–6096.

BADLANI, R.; ŁAńCUCKI, A.; SHIH, K. J.; VALLE, R.; PING, W.; CATANZARO, B. One tts alignment to rule them all. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2022. p. 6092–6096.

BARBOSA, P. The Acoustics of Pleasantness in Poetry Declamation in Two Varieties of Portuguese. In: *Proc. Speech Prosody 2022*. [S.l.: s.n.], 2022. p. 515–519.

BARBOSA, P. A.; MADUREIRA, S.; MAREüIL, P. B. de. Cross-Linguistic Distinctions Between Professional and Non-Professional Speaking Styles. In: *Proc. Interspeech 2017*. [S.l.: s.n.], 2017. p. 3921–3925. ISSN 2308-457X.

BATLINER, A.; MÖBIUS, B. Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground? In: _____. *The Integration of Phonetic Knowledge in Speech Technology*. Dordrecht: Springer Netherlands, 2005. p. 21–44. ISBN 978-1-4020-2637-9.

BIAN, Y.; CHEN, C.; KANG, Y.; PAN, Z. Multi-reference tacotron by intercross training for style disentangling, transfer and control in speech synthesis. *arXiv preprint arXiv:1904.02373*, 2019.

BIRKHOLZ, P.; MARTIN, L.; WILLMES, K.; KRÖGER, B. J.; NEUSCHAEFER-RUBE, C. The contribution of phonation type to the perception of vocal emotions in german: An articulatory synthesis study. *The Journal of the Acoustical Society of America*, AIP Publishing, v. 137, n. 3, p. 1503–1512, 2015.

BIńKOWSKI, M.; DONAHUE, J.; DIELEMAN, S.; CLARK, A.; ELSEN, E.; CASAGRANDE, N.; COBO, L. C.; SIMONYAN, K. *High Fidelity Speech Synthesis with Adversarial Networks*. [S.l.]: arXiv, 2019.

BRACKHANE, F. Kempelen vs. kratzenstein-researchers on speech synthesis in times of change. In: TUDPRESS. *HSCR 2015. Proceedings of the First International Workshop on the History of Speech Communication Research, Dresden, September 4-5, 2015*. [S.l.], 2015. p. 42–49.

CABANAC, M. What is emotion? *Behavioural processes*, Elsevier, v. 60, n. 2, p. 69–83, 2002.

CHEN, L.-W.; RUDNICKY, A. Fine-grained style control in transformer-based text-to-speech synthesis. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2022. p. 7907–7911.

CHEN, S.; WANG, C.; CHEN, Z.; WU, Y.; LIU, S.; CHEN, Z.; LI, J.; KANDA, N.; YOSHIOKA, T.; XIAO, X. *et al.* Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, IEEE, v. 16, n. 6, p. 1505–1518, 2022.

CHEON, S. J.; CHOI, B. J.; KIM, M.; LEE, H.; KIM, N. S. A controllable multi-lingual multi-speaker multi-style text-to-speech synthesis with multivariate information minimization. *IEEE Signal Processing Letters*, v. 29, p. 55–59, 2022.

CHOI, B. J.; JEONG, M.; LEE, J. Y.; KIM, N. S. Snac: Speaker-normalized affine coupling layer in flow-based architecture for zero-shot multi-speaker text-to-speech. *IEEE Signal Processing Letters*, IEEE, v. 29, p. 2502–2506, 2022.

CHOI, H.-S.; YANG, J.; LEE, J.; KIM, H. Nansy++: Unified voice synthesis with neural analysis and synthesis. *arXiv preprint arXiv:2211.09407*, 2022.

CHOU, J. chieh; LEE, H.-Y. One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization. In: *Proc. Interspeech 2019*. [S.l.: s.n.], 2019. p. 664–668. ISSN 2308-457X.

CHUNG, R.; MAK, B. On-the-fly data augmentation for text-to-speech style transfer. In: *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. [S.l.: s.n.], 2021. p. 634–641.

COOPER, F. S.; LIBERMAN, A. M.; BORST, J. M. The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 37, n. 5, p. 318–325, 1951.

CULTURE, G. A. . *The "Kempelen" speaking machine*. 2016. Accessed on May 19, 2024.

DEHAK, N.; KENNY, P. J.; DEHAK, R.; DUMOUCHEL, P.; OUELLET, P. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, IEEE, v. 19, n. 4, p. 788–798, 2010.

DUDLEY, H. The carrier nature of speech. *Bell System Technical Journal*, Wiley Online Library, v. 19, n. 4, p. 495–515, 1940.

DUDLEY, H.; TARNOCZY, T. H. The speaking machine of wolfgang von kempelen. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 22, n. 2, p. 151–166, 1950.

GANIN, Y.; USTINOVA, E.; AJAKAN, H.; GERMAIN, P.; LAROCHELLE, H.; LAVIOLETTE, F.; MARCHAND, M.; LEMPITSKY, V. Domain-adversarial training of neural networks. *The journal of machine learning research*, JMLR. org, v. 17, n. 1, p. 2096–2030, 2016.

GHOSH, S.; DAS, A.; SINHA, Y.; SIEGERT, I.; POLZEHL, T.; STOBER, S. Emo-StarGAN: A Semi-Supervised Any-to-Many Non-Parallel Emotion-Preserving Voice Conversion. In: *Proc. INTERSPEECH 2023*. [S.l.: s.n.], 2023. p. 2093–2097. ISSN 2308-457X.

GOVIND, D.; PRASANNA, S. M. Expressive speech synthesis: a review. *International Journal of Speech Technology*, Springer, v. 16, p. 237–260, 2013.

HARRIS, C. M. A speech synthesizer. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 25, n. 5, p. 970–975, 1953.

HAYASHI, T.; YAMAMOTO, R.; INOUE, K.; YOSHIMURA, T.; WATANABE, S.; TODA, T.; TAKEDA, K.; ZHANG, Y.; TAN, X. Espnet-tts: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2020. p. 7654–7658.

HEINZ, J. M.; STEVENS, K. N. On the derivation of area functions and acoustic spectra from cineradiographic films of speech. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 36, n. 5_Supplement, p. 1037–1038, 1964.

HINTON, G.; DENG, L.; YU, D.; DAHL, G. E.; MOHAMED, A.-r.; JAITLY, N.; SENIOR, A.; VANHOUCKE, V.; NGUYEN, P.; SAINATH, T. N. *et al.* Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, IEEE, v. 29, n. 6, p. 82–97, 2012.

HODARI, Z.; LAI, C.; KING, S. Perception of prosodic variation for speech synthesis using an unsupervised discrete representation of f0. In: *Proceedings of Speech Prosody 2020*. [S.l.: s.n.], 2020. p. 965–969. Published 24 May 2020; Speech Prosody 2020 ; Conference date: 24-05-2020 Through 28-05-2020.

HOLMES, J. N.; MATTINGLY, I. G.; SHEARME, J. N. Speech synthesis by rule. *Language and speech*, SAGE Publications Sage UK: London, England, v. 7, n. 3, p. 127–143, 1964.

HSU, W.-N.; BOLTE, B.; TSAI, Y.-H. H.; LAKHOTIA, K.; SALAKHUTDINOV, R.; MOHAMED, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, IEEE, v. 29, p. 3451–3460, 2021.

HSU, W.-N.; ZHANG, Y.; WEISS, R.; ZEN, H.; WU, Y.; CAO, Y.; WANG, Y. Hierarchical generative modeling for controllable speech synthesis. In: *International Conference on Learning Representations*. [S.l.: s.n.], 2019.

HUA, H.; CHEN, Z.; ZHANG, Y.; LI, M.; ZHANG, P. Improving spoofing capability for end-to-end any-to-many voice conversion. In: *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. New York, NY, USA: Association for Computing Machinery, 2022. (DDAM '22), p. 93–100. ISBN 9781450394963.

HUANG, R.; CHEN, F.; REN, Y.; LIU, J.; CUI, C.; ZHAO, Z. Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In: *Proceedings of the 29th ACM International Conference on Multimedia*. [S.l.: s.n.], 2021. p. 3945–3954.

HUYBRECHTS, G.; MERRITT, T.; COMINI, G.; PERZ, B.; SHAH, R.; LORENZO-TRUEBA, J. Low-resource expressive text-to-speech using data augmentation. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2021. p. 6593–6597.

HWANG, M.-J.; YAMAMOTO, R.; SONG, E.; KIM, J.-M. Tts-by-tts: Tts-driven data augmentation for fast and high-quality speech synthesis. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2021. p. 6598–6602.

INGEMANN, F. Speech Synthesis by Rule. *The Journal of the Acoustical Society of America*, v. 29, n. $11_{S}upplement, p.1255 - -1255, 111957.ISSN0001 - 4966$.

IRVINE, J. T. *" Style" as distinctiveness: the culture and ideology of linguistic differentiation*. [S.l.]: na, 2001.

ITO, K.; JOHNSON, L. The lj speech dataset. 2017. 2017.

JAMES, J.; WATSON, C. I.; MACDONALD, B. Artificial empathy in social robots: An analysis of emotions in speech. In: *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. [S.l.: s.n.], 2018. p. 632–637.

JEONG, M.; KIM, H.; CHEON, S. J.; CHOI, B. J.; KIM, N. S. *Diff-TTS: A Denoising Diffusion Model for Text-to-Speech*. [S.l.]: arXiv, 2021.

JIANG, Z.; REN, Y.; YE, Z.; LIU, J.; ZHANG, C.; YANG, Q.; JI, S.; HUANG, R.; WANG, C.; YIN, X. *et al.* Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias. *arXiv preprint arXiv:2306.03509*, 2023.

JU, Z.; WANG, Y.; SHEN, K.; TAN, X.; XIN, D.; YANG, D.; LIU, Y.; LENG, Y.; SONG, K.; TANG, S. *et al.* Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.

JUNG, W.; LEE, J. E3-vits: Emotional end-to-end tts with cross-speaker style transfer. 2023.

KANAGAKI, A.; TANAKA, M.; NOSE, T.; SHIMIZU, R.; ITO, A.; ITO, A. Cyclegan-based high-quality non-parallel voice conversion with spectrogram and wavernn. In: *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*. [S.l.: s.n.], 2020. p. 356–357.

KANG, M.; HAN, W.; HWANG, S. J.; YANG, E. ZET-Speech: Zero-shot adaptive Emotion-controllable Text-to-Speech Synthesis with Diffusion and Style-based Models. In: *Proc. INTERSPEECH 2023*. [S.l.: s.n.], 2023. p. 4339–4343. ISSN 2308-457X.

KARAALI, O.; CORRIGAN, G.; GERSON, I. Speech synthesis with neural networks. In: *World Congress on Neural Networks, San Diego*. [S.l.: s.n.], 1996. p. 45–50.

KARLAPATI, S.; MOINET, A.; JOLY, A.; KLIMKOV, V.; SáEZ-TRIGUEROS, D.; DRUGMAN, T. CopyCat: Many-to-Many Fine-Grained Prosody Transfer for Neural Text-to-Speech. In: *Proc. Interspeech 2020*. [S.l.: s.n.], 2020. p. 4387–4391. ISSN 2308-457X.

KAWAHARA, H. Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, Acoustical Society of Japan, v. 27, n. 6, p. 349–353, 2006.

KELLY, J. L. Speech synthesis. In: *Proc. Speech Communication Seminar, Stockholm (Sep. 1962)*. [S.l.: s.n.], 1962.

KEMPELEN, W. V. *Mechanismus der menschlichen Sprache*. [S.l.]: Degen, 1791.

KIM, J.; KIM, S.; KONG, J.; YOON, S. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In: LAROCHELLE, H.; RANZATO, M.; HADSELL, R.; BALCAN, M.; LIN, H. (Ed.). *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2020. v. 33, p. 8067–8077.

KIM, J.; KONG, J.; SON, J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2021. p. 5530–5540.

KIM, J. W.; SALAMON, J.; LI, P.; BELLO, J. P. Crepe: A convolutional representation for pitch estimation. In: IEEE. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2018. p. 161–165.

KIM, S.; SHIH, K.; SANTOS, J. F.; BAKHTURINA, E.; DESTA, M.; VALLE, R.; YOON, S.; CATANZARO, B. *et al.* P-flow: A fast and data-efficient zero-shot tts through speech prompting. *Advances in Neural Information Processing Systems*, v. 36, 2024.

KLATT, D. The klattalk text-to-speech conversion system. In: IEEE. *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*. [S.l.], 1982. v. 7, p. 1589–1592.

KNYAZEV, G. G.; BARCHARD, K. A.; RAZUMNIKOVA, O. M.; MITROFANOVA, L. G. The relationship of positive and negative expressiveness to the processing of emotion information. *Scandinavian journal of psychology*, Wiley Online Library, v. 53, n. 3, p. 206–215, 2012.

KONG, J.; KIM, J.; BAE, J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, v. 33, p. 17022–17033, 2020.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, v. 25, 2012.

KWON, O.; JANG, I.; AHN, C.; KANG, H.-G. An effective style token weight control technique for end-to-end emotional speech synthesis. *IEEE Signal Processing Letters*, v. 26, n. 9, p. 1383–1387, 2019.

ŁAŃCUCKI, A. Fastpitch: Parallel text-to-speech with pitch prediction. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2021. p. 6588–6592.

LEE, S. gil; PING, W.; GINSBURG, B.; CATANZARO, B.; YOON, S. BigVGAN: A universal neural vocoder with large-scale training. In: *The Eleventh International Conference on Learning Representations*. [S.l.: s.n.], 2023.

LEE, Y.; KIM, T. Robust and fine-grained prosody control of end-to-end speech synthesis. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2019. p. 5911–5915.

LEI, S.; ZHOU, Y.; CHEN, L.; WU, Z.; KANG, S.; MENG, H. Towards expressive speaking style modelling with hierarchical context information for mandarin speech synthesis. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2022. p. 7922–7926.

LEI, Y.; YANG, S.; ZHU, X.; XIE, L.; SU, D. Cross-speaker emotion transfer through information perturbation in emotional speech synthesis. *IEEE Signal Processing Letters*, v. 29, p. 1948–1952, 2022.

LI, J.; TU, W.; XIAO, L. Freevc: Towards high-quality text-free one-shot voice conversion. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2023. p. 1–5.

LI, N.; LIU, S.; LIU, Y.; ZHAO, S.; LIU, M. Neural speech synthesis with transformer network. In: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. [S.l.]: AAAI Press, 2019. (AAAI'19/IAAI'19/EAAI'19). ISBN 978-1-57735-809-1.

LI, T.; YANG, S.; XUE, L.; XIE, L. Controllable emotion transfer for end-to-end speech synthesis. In: *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. [S.l.: s.n.], 2021. p. 1–5.

LI, Y. A.; HAN, C.; MESGARANI, N. *StyleTTS: A Style-Based Generative Model for Natural and Diverse Text-to-Speech Synthesis*. [S.l.]: arXiv, 2022.

LI, Y. A.; HAN, C.; MESGARANI, N. Styletts-vc: One-shot voice conversion by knowledge transfer from style-based tts models. In: *2022 IEEE Spoken Language Technology Workshop (SLT)*. [S.l.: s.n.], 2023. p. 920–927.

LIAN, J.; ZHANG, C.; ANUMANCHIPALLI, G. K.; YU, D. Unsupervised tts acoustic modeling for tts with conditional disentangled sequential vae. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 31, p. 2548–2557, 2023.

LINDBLOM, B. E.; SUNDBERG, J. E. Acoustical consequences of lip, tongue, jaw, and larynx movement. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 50, n. 4B, p. 1166–1179, 1971.

LIU, S.; YANG, S.; SU, D.; YU, D. Referee: Towards reference-free cross-speaker style transfer with low-quality data for expressive speech synthesis. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2022. p. 6307–6311.

LORENZO-TRUEBA, J.; BARRA-CHICOTE, R.; GALLARDO-ANTOLIN, A.; YAMAGISHI, J.; MONTERO, J. M. Continuous expressive speaking styles synthesis based on CVSM and MR-HMM. In: MATSUMOTO, Y.; PRASAD, R. (Ed.). *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, 2016. p. 369–376.

MA, S.; MCDUFF, D.; SONG, Y. A generative adversarial network for style modeling in a text-to-speech system. In: *International Conference on Learning Representations*. [S.l.: s.n.], 2019.

MARQUES, L. B. de M.; UEDA, L. H.; SIMÕES, F. O.; NETO, M. U.; RUNSTEIN, F. O.; NAGLE, E. J.; BÓ, B. D.; COSTA, P. D. Diffusion-based approach to style modeling in expressive tts. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2022. p. 253–267.

MARTIN, J. R. Design and practice: Enacting functional linguistics. *Annual Review of applied linguistics*, Cambridge University Press, v. 20, p. 116–126, 2000.

MOINE, C. L.; OBIN, N. *Att-HACK: An Expressive Speech Database with Social Attitudes*. [S.l.]: arXiv, 2020.

MORISE, M. Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals. In: *Proc. Interspeech 2017*. [S.l.: s.n.], 2017. p. 2321–2325. ISSN 2308-457X.

MORISE, M.; YOKOMORI, F.; OZAWA, K. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, The Institute of Electronics, Information and Communication Engineers, v. 99, n. 7, p. 1877–1884, 2016.

MOULINES, E.; CHARPENTIER, F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, Elsevier, v. 9, n. 5-6, p. 453–467, 1990.

NGUYEN, G.-N.; PHUNG, T.-N. Reducing over-smoothness in hmm-based speech synthesis using exemplar-based voice conversion. *EURASIP Journal on Audio, Speech, and Music Processing*, Springer, v. 2017, n. 1, p. 14, 2017.

NICULESCU, A.; DIJK, B. van; NIJHOLT, A.; LI, H.; SEE, S. L. Making social robots more attractive: the effects of voice pitch, humor and empathy. *International journal of social robotics*, Springer, v. 5, n. 2, p. 171–191, 2013.

NIEKERK, B. van; CARBONNEAU, M.-A.; ZAÏDI, J.; BAAS, M.; SEUTÉ, H.; KAMPER, H. A comparison of discrete and soft speech units for improved voice conversion. In: IEEE. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2022. p. 6562–6566.

NING, Y.; HE, S.; WU, Z.; XING, C.; ZHANG, L.-J. A review of deep learning based speech synthesis. *Applied Sciences*, v. 9, n. 19, 2019. ISSN 2076-3417.

NING, Y.; HE, S.; WU, Z.; XING, C.; ZHANG, L.-J. A Review of Deep Learning Based Speech Synthesis. *Applied Sciences*, v. 9, n. 19, p. 4050, jan. 2019. Number: 19 Publisher: Multidisciplinary Digital Publishing Institute.

OBIN, N.; LANCHANTIN, P.; LACHERET, A.; RODET, X. Discrete/Continuous Modelling of Speaking Style in HMM-based Speech Synthesis: Design and Evaluation. In: *Interspeech*. Florence, Italy: [s.n.], 2011. p. —.

OORD, A. V. D.; DIELEMAN, S.; ZEN, H.; SIMONYAN, K.; VINYALS, O.; GRAVES, A.; KALCHBRENNER, N.; SENIOR, A.; KAVUKCUOGLU, K. *et al.* Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, v. 12, 2016.

PAN, S.; HE, L. Cross-Speaker Style Transfer with Prosody Bottleneck in Neural Speech Synthesis. In: *Proc. Interspeech 2021*. [S.l.: s.n.], 2021. p. 4678–4682. ISSN 2308-457X.

PÉPIOT, E. Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in parisian french and american english speakers. In: *Speech Prosody 7*. [S.l.: s.n.], 2014. p. 305–309.

PEREZ, E.; STRUB, F.; VRIES, H. D.; DUMOULIN, V.; COURVILLE, A. Film: Visual reasoning with a general conditioning layer. In: *Proceedings of the AAAI conference on artificial intelligence*. [S.l.: s.n.], 2018. v. 32, n. 1.

PING, W.; PENG, K.; GIBIANSKY, A.; ARIK, S. Ömer; KANNAN, A.; NARANG, S.; RAIMAN, J.; MILLER, J. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In: *ICLR (Poster)*. [S.l.: s.n.], 2018.

POTTER, R. K. Visible patterns of sound. *Science*, American Association for the Advancement of Science, v. 102, n. 2654, p. 463–470, 1945.

QIAN, K.; ZHANG, Y.; CHANG, S.; COX, D.; HASEGAWA-JOHNSON, M. Unsupervised speech decomposition via triple information bottleneck. In: *Proceedings of the 37th International Conference on Machine Learning*. [S.l.]: JMLR.org, 2020. (ICML'20).

QIANG, C.; YANG, P.; CHE, H.; WANG, X.; WANG, Z. Style-label-free: Cross-speaker style transfer by quantized vae and speaker-wise normalization in speech synthesis. In: *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. [S.l.: s.n.], 2022. p. 61–65.

RADFORD, A.; KIM, J. W.; XU, T.; BROCKMAN, G.; MCLEAVEY, C.; SUTSKEVER, I. Robust speech recognition via large-scale weak supervision. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2023. p. 28492–28518.

RAMSAY, G. J. Mechanical speech synthesis in early talking automata. *Acoustics Today*, v. 15, n. 2, p. 11–19, 2019.

RAO, K.; PENG, F.; SAK, H.; BEAUFAYS, F. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In: IEEE. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2015. p. 4225–4229.

REN, Y.; HU, C.; TAN, X.; QIN, T.; ZHAO, S.; ZHAO, Z.; LIU, T.-Y. Fastspeech 2: Fast and high-quality end-to-end text to speech. In: *International Conference on Learning Representations*. [S.l.: s.n.], 2021.

REN, Y.; LEI, M.; HUANG, Z.; ZHANG, S.; CHEN, Q.; YAN, Z.; ZHAO, Z. Prosospeech: Enhancing prosody with quantized vector pre-training in text-to-speech. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2022. p. 7577–7581.

RIBEIRO, M. S.; ROTH, J.; COMINI, G.; HUYBRECHTS, G.; GABRYś, A.; LORENZO-TRUEBA, J. Cross-speaker style transfer for text-to-speech using data augmentation. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2022. p. 6797–6801.

ROSEN, G. Dynamic Analog Speech Synthesizer. *The Journal of the Acoustical Society of America*, v. 30, n. 3, p. 201–209, 03 1958. ISSN 0001-4966.

ROSENHOUSE, G. Biomimetics of sound production, synthesis and recognition. *WIT Transactions on Ecology and the Environment*, WIT Press, v. 138, p. 273–287, 2010.

RUBIN, P.; BAER, T.; MERMELSTEIN, P. An articulatory synthesizer for perceptual research. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 70, n. 2, p. 321–328, 1981.

RUBIN, P.; SALTZMAN, E.; GOLDSTEIN, L.; MCGOWAN, R.; TIEDE, M.; BROWMAN, C. Casy and extensions to the task-dynamic model. In: *1st ESCA tutorial and research workshop on speech production modeling: from control strategies to acoustics & 4th speech production seminar: models and data (Autrans, May 20-24, 1996)*. [S.l.: s.n.], 1996. p. 125–128.

SAEKI, T.; XIN, D.; NAKATA, W.; KORIYAMA, T.; TAKAMICHI, S.; SARUWATARI, H. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*, 2022.

SCHOEFFLER, M.; BARTOSCHEK, S.; STÖTER, F.-R.; ROESS, M.; WESTPHAL, S.; EDLER, B.; HERRE, J. webmushra—a comprehensive framework for web-based listening tests. *Journal of Open Research Software*, Ubiquity Press, v. 6, n. 1, p. 8, 2018.

SHAH, R.; POKORA, K.; EZZERG, A.; KLIMKOV, V.; HUYBRECHTS, G.; PUTRYCZ, B.; KORZEKWA, D.; MERRITT, T. Non-Autoregressive TTS with Explicit Duration Modelling for Low-Resource Highly Expressive Speech. In: *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*. [S.l.: s.n.], 2021. p. 96–101.

SHANG, Z.; HUANG, Z.; ZHANG, H.; ZHANG, P.; YAN, Y. Incorporating Cross-Speaker Style Transfer for Multi-Language Text-to-Speech. In: *Proc. Interspeech 2021*. [S.l.: s.n.], 2021. p. 1619–1623. ISSN 2308-457X.

SHEN, J.; PANG, R.; WEISS, R. J.; SCHUSTER, M.; JAITLY, N.; YANG, Z.; CHEN, Z.; ZHANG, Y.; WANG, Y.; SKERRV-RYAN, R.; SAUROUS, R. A.; AGIOMVRGIANNAKIS, Y.; WU, Y. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2018. p. 4779–4783.

SKERRY-RYAN, R.; BATTENBERG, E.; XIAO, Y.; WANG, Y.; STANTON, D.; SHOR, J.; WEISS, R.; CLARK, R.; SAUROUS, R. A. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In: DY, J.; KRAUSE, A. (Ed.). *Proceedings of the 35th International Conference on Machine Learning*. [S.l.]: PMLR, 2018. (Proceedings of Machine Learning Research, v. 80), p. 4693–4702.

SKERRY-RYAN, R. J.; BATTENBERG, E.; XIAO, Y.; WANG, Y.; STANTON, D.; SHOR, J.; WEISS, R. J.; CLARK, R.; SAUROUS, R. A. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. *arXiv:1803.09047 [cs, eess]*, mar. 2018. ArXiv: 1803.09047.

SOTELO, J.; MEHRI, S.; KUMAR, K.; SANTOS, J. F.; KASTNER, K.; COURVILLE, A.; BENGIO, Y. Char2wav: End-to-end speech synthesis. 2017.

STANTON, D.; WANG, Y.; SKERRY-RYAN, R. Predicting expressive speaking style from text in end-to-end speech synthesis. In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. [S.l.: s.n.], 2018. p. 595–602.

STEVENS, K. N.; KASOWSKI, S.; FANT, C. G. M. An Electrical Analog of the Vocal Tract. *The Journal of the Acoustical Society of America*, v. 25, n. 4, p. 734–742, 07 1953. ISSN 0001-4966.

STEWART, J. Q. An electrical analogue of the vocal organs. *Nature*, Nature Publishing Group UK London, v. 110, n. 2757, p. 311–312, 1922.

STORY, B. H. History of speech synthesis. *Teoksessa The Routledge Handbook of Phonetics, toimittaneet William F. Katz & Peter F. Assmann*, p. 9–33, 2019.

SUN, G.; ZHANG, Y.; WEISS, R. J.; CAO, Y.; ZEN, H.; WU, Y. Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis. In: *ICASSP*. [S.l.: s.n.], 2020.

SZÉKELY, É. *Expressive speech synthesis in human interaction*. Tese (Doutorado) — University College Dublin, 2015.

TAAL, C. H.; HENDRIKS, R. C.; HEUSDENS, R.; JENSEN, J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 19, n. 7, p. 2125–2136, 2011.

TAN, X.; QIN, T.; SOONG, F.; LIU, T.-Y. *A Survey on Neural Speech Synthesis*. [S.l.]: arXiv, 2021.

TERASHIMA, R.; YAMAMOTO, R.; SONG, E.; SHIRAHATA, Y.; YOON, H.-W.; KIM, J.-M.; TACHIBANA, K. Cross-Speaker Emotion Transfer for Low-Resource Text-to-Speech Using Non-Parallel Voice Conversion with Pitch-Shift Data Augmentation. In: *Proc. Interspeech 2022*. [S.l.: s.n.], 2022. p. 3018–3022. ISSN 2308-457X.

TITS, N.; HADDAD, K. E.; DUTOIT, T. Exploring transfer learning for low resource emotional tts. In: BI, Y.; BHATIA, R.; KAPOOR, S. (Ed.). *Intelligent Systems and Applications*. Cham: Springer International Publishing, 2020. p. 52–60. ISBN 978-3-030-29516-5.

TRIANTAFYLLOPOULOS, A.; SCHULLER, B. W.; İYMEN, G.; SEZGIN, M.; HE, X.; YANG, Z.; TZIRAKIS, P.; LIU, S.; MERTES, S.; ANDRé, E.; FU, R.; TAO, J. An overview of affective speech synthesis and conversion in the deep learning era. *Proceedings of the IEEE*, p. 1–27, 2023.

URIA, B.; MURRAY, I.; RENALS, S.; RICHMOND, K. Deep architectures for articulatory inversion. In: ISCA. *INTERSPEECH 2012 13th Annual Conference of the International Speech Communication Association*. [S.l.], 2012. p. 867–870.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L. u.; POLOSUKHIN, I. Attention is all you need. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2017. v. 30.

VEAUX, C.; YAMAGISHI, J.; MACDONALD, K. *et al.* Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, v. 6, p. 15, 2017.

WAN, L.; WANG, Q.; PAPIR, A.; MORENO, I. L. Generalized end-to-end loss for speaker verification. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2018. p. 4879–4883.

WANG, C.; CHEN, S.; WU, Y.; ZHANG, Z.; ZHOU, L.; LIU, S.; CHEN, Z.; LIU, Y.; WANG, H.; LI, J. *et al.* Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

WANG, X.; CHEN, Y.; ZHU, W. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 44, n. 9, p. 4555–4576, 2022.

WANG, Y.; SKERRY-RYAN, R. J.; 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. Tacotron: Towards End-to-End Speech Synthesis. *arXiv:1703.10135 [cs]*, abr. 2017. ArXiv: 1703.10135.

WANG, Y.; STANTON, D.; ZHANG, Y.; RYAN, R.-S.; BATTENBERG, E.; SHOR, J.; XIAO, Y.; JIA, Y.; REN, F.; SAUROUS, R. A. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In: DY, J.; KRAUSE, A. (Ed.). *Proceedings of the 35th International Conference on Machine Learning*. [S.l.]: PMLR, 2018. (Proceedings of Machine Learning Research, v. 80), p. 5180–5189.

WANG, Y.; XIE, Y.; ZHAO, K.; WANG, H.; ZHANG, Q. Unsupervised quantized prosody representation for controllable speech synthesis. In: *2022 IEEE International Conference on Multimedia and Expo (ICME)*. [S.l.: s.n.], 2022. p. 1–6.

WU, Z.; WATTS, O.; KING, S. Merlin: An open source neural network speech synthesis system. In: *9th ISCA Speech Synthesis Workshop*. [S.l.: s.n.], 2016. p. 202–207.

YAKOUMAKI, T. Expressive speech analysis and classification using adaptive sinusoidal modeling. 2015.

YAMAMOTO, R.; SONG, E.; KIM, J.-M. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In: IEEE. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2020. p. 6199–6203.

ZAïDI, J.; SEUTé, H.; van Niekerk, B.; CARBONNEAU, M.-A. Daft-Exprt: Cross-Speaker Prosody Transfer on Any Text for Expressive Speech Synthesis. In: *Proc. Interspeech 2022*. [S.l.: s.n.], 2022. p. 4591–4595. ISSN 2308-457X.

ZEN, H. Acoustic modeling in statistical parametric speech synthesis-from hmm to lstm-rnn. *Proc. MLSLP*, v. 15, 2015.

ZEN, H.; DANG, V.; CLARK, R.; ZHANG, Y.; WEISS, R. J.; JIA, Y.; CHEN, Z.; WU, Y. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In: *Proc. Interspeech 2019*. [S.l.: s.n.], 2019. p. 1526–1530. ISSN 2308-457X.

ZEN, H.; NOSE, T.; YAMAGISHI, J.; SAKO, S.; MASUKO, T.; BLACK, A. W.; TOKUDA, K. The hmm-based speech synthesis system (hts) version 2.0. *SSW*, v. 6, p. 294–299, 2007.

ZEN, H.; SENIOR, A.; SCHUSTER, M. Statistical parametric speech synthesis using deep neural networks. In: IEEE. *2013 ieee international conference on acoustics, speech and signal processing*. [S.l.], 2013. p. 7962–7966.

ZHANG, G.; QIN, Y.; ZHANG, W.; WU, J.; LI, M.; GAI, Y.; JIANG, F.; LEE, T. *iEmoTTS: Toward Robust Cross-Speaker Emotion Transfer and Control for Speech Synthesis based on Disentanglement between Prosody and Timbre*. [S.l.]: arXiv, 2022.

ZHANG, G.; QIN, Y.; ZHANG, W.; WU, J.; LI, M.; GAI, Y.; JIANG, F.; LEE, T. iemotts: Toward robust cross-speaker emotion transfer and control for speech synthesis based on disentanglement between prosody and timbre. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 31, p. 1693–1705, 2023.

ZHANG, H.; ZHAN, H.; YU, X.; LIN, Y. Cross-speaker style transfer using curriculum learning and data augmentation. 2023.

ZHOU, K.; SISMAN, B.; LIU, R.; LI, H. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In: IEEE. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2021. p. 920–924.

ZHOU, K.; SISMAN, B.; LIU, R.; LI, H. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2021. p. 920–924.

ZHOU, K.; SISMAN, B.; LIU, R.; LI, H. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, Elsevier, v. 137, p. 1–18, 2022.

ZHOU, K.; SISMAN, B.; RANA, R.; SCHULLER, B. W.; LI, H. *Speech Synthesis with Mixed Emotions*. [S.l.]: arXiv, 2022.

ZHU, X.; LEI, Y.; SONG, K.; ZHANG, Y.; LI, T.; XIE, L. Multi-speaker expressive speech synthesis via multiple factors decoupling. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2023. p. 1–5.

ZIYIN, L.; HARTWIG, T.; UEDA, M. Neural networks fail to learn periodic functions and how to fix it. *Advances in Neural Information Processing Systems*, v. 33, p. 1583–1594, 2020.

Appendix A

Perceptual Assessment

An online perceptual assessment was conducted to rate all the stimuli generated by the proposed pipeline, its ablations, baseline models, and also the re-synthesized ground-truth audios. Three aspects of speech were analysed: naturalness, style intensity and speaker similarity. For each aspect, the participants were instructed on which speech factors to look for while rating, and which factors not to take into account. Since being a speech expert was not a requirement to take the assessment, the instructions played a crucial role in the results, as the test participants' answers were entirely reflected on how they understood the problem. The following images show all the pages types, such as welcoming page, instructions, rating pages, etc, that were present in the perceptual assessment in sequential order.

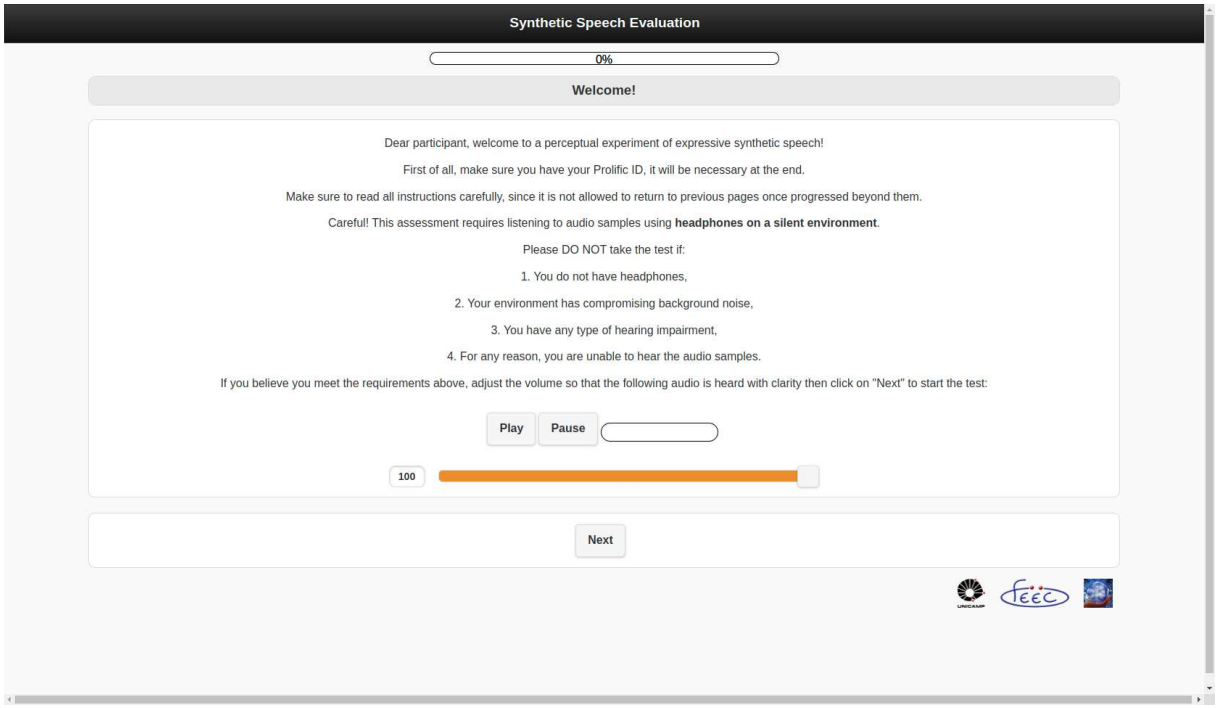


Figure A.1: Welcome and requirements page.

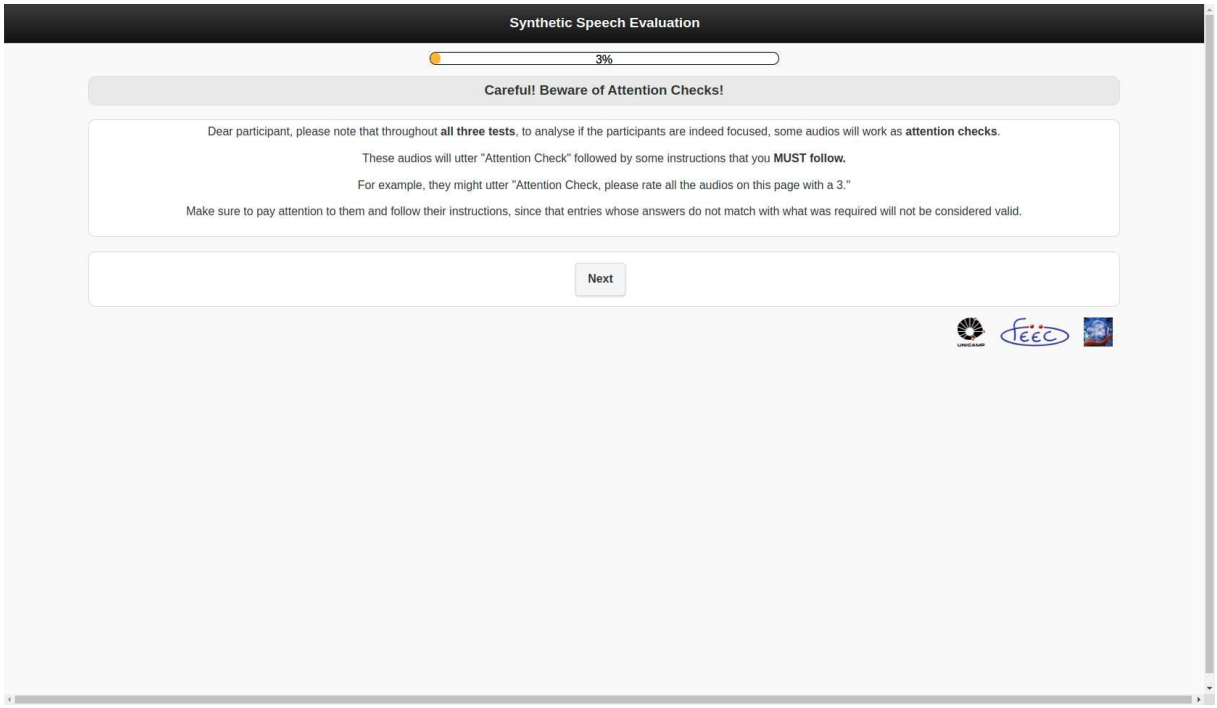


Figure A.2: Instructions concerning attention checks.

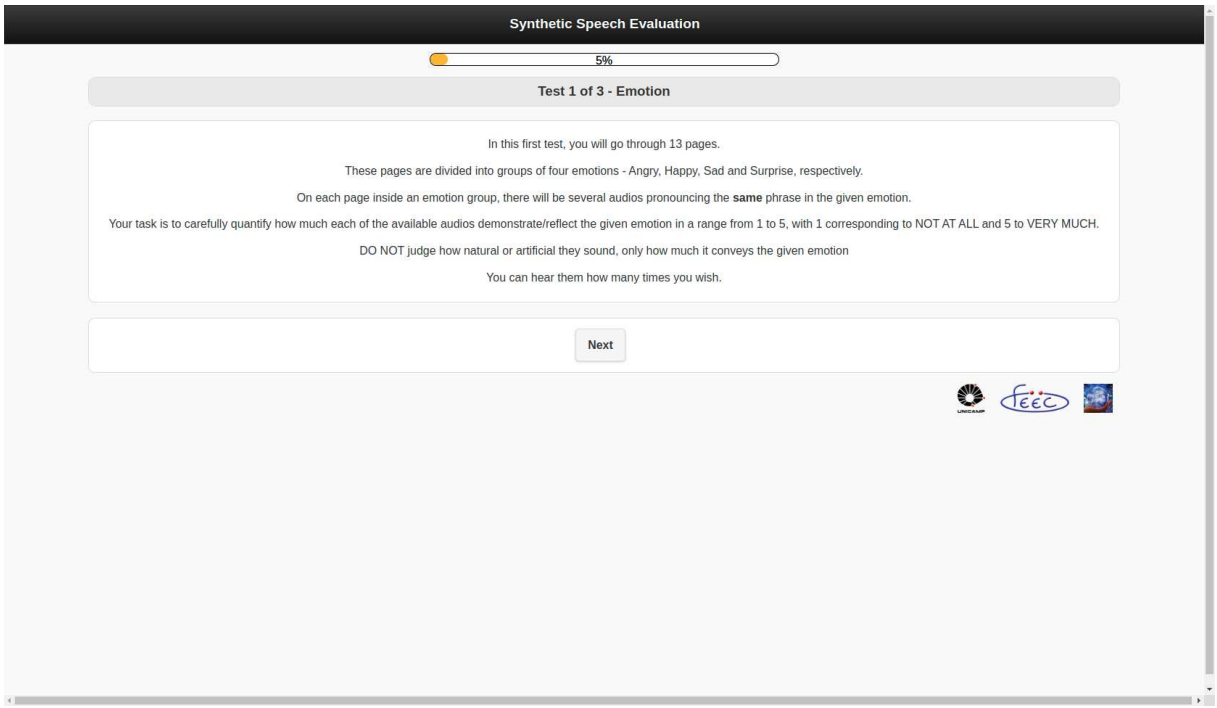


Figure A.3: Instructions for the style intensity experiment.

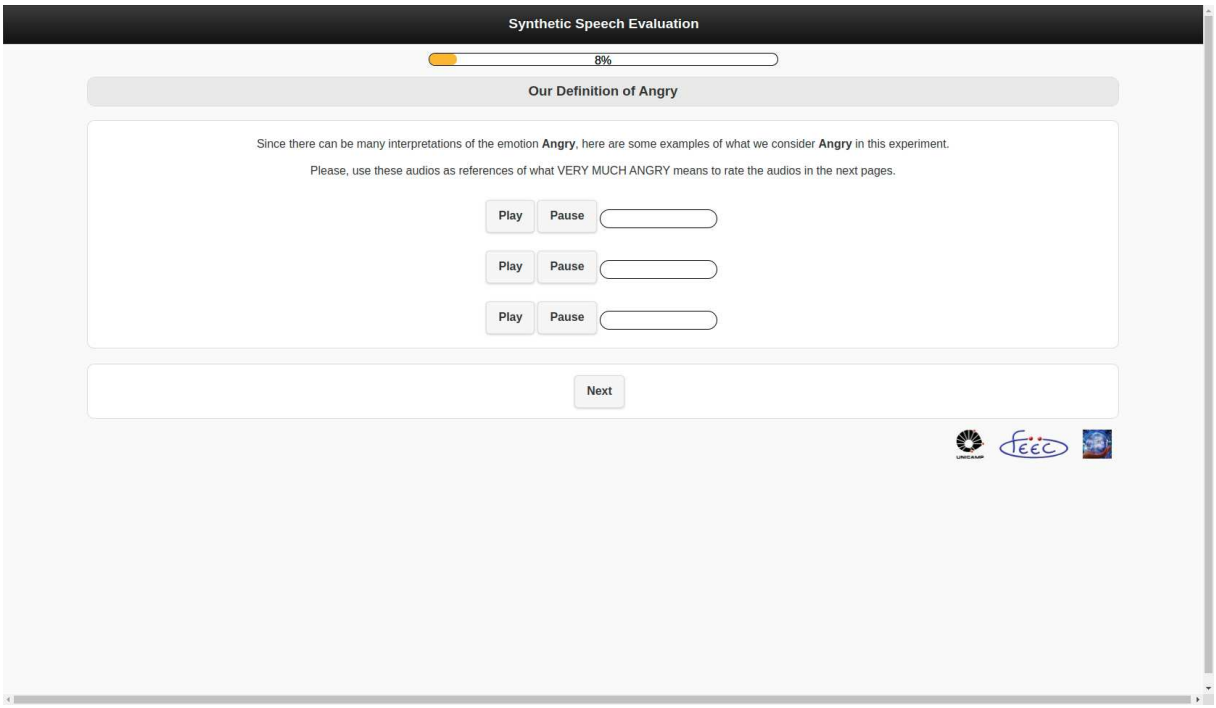


Figure A.4: Page elucidating the “angry” speaking style with sample audios. Repeated for all styles considered.

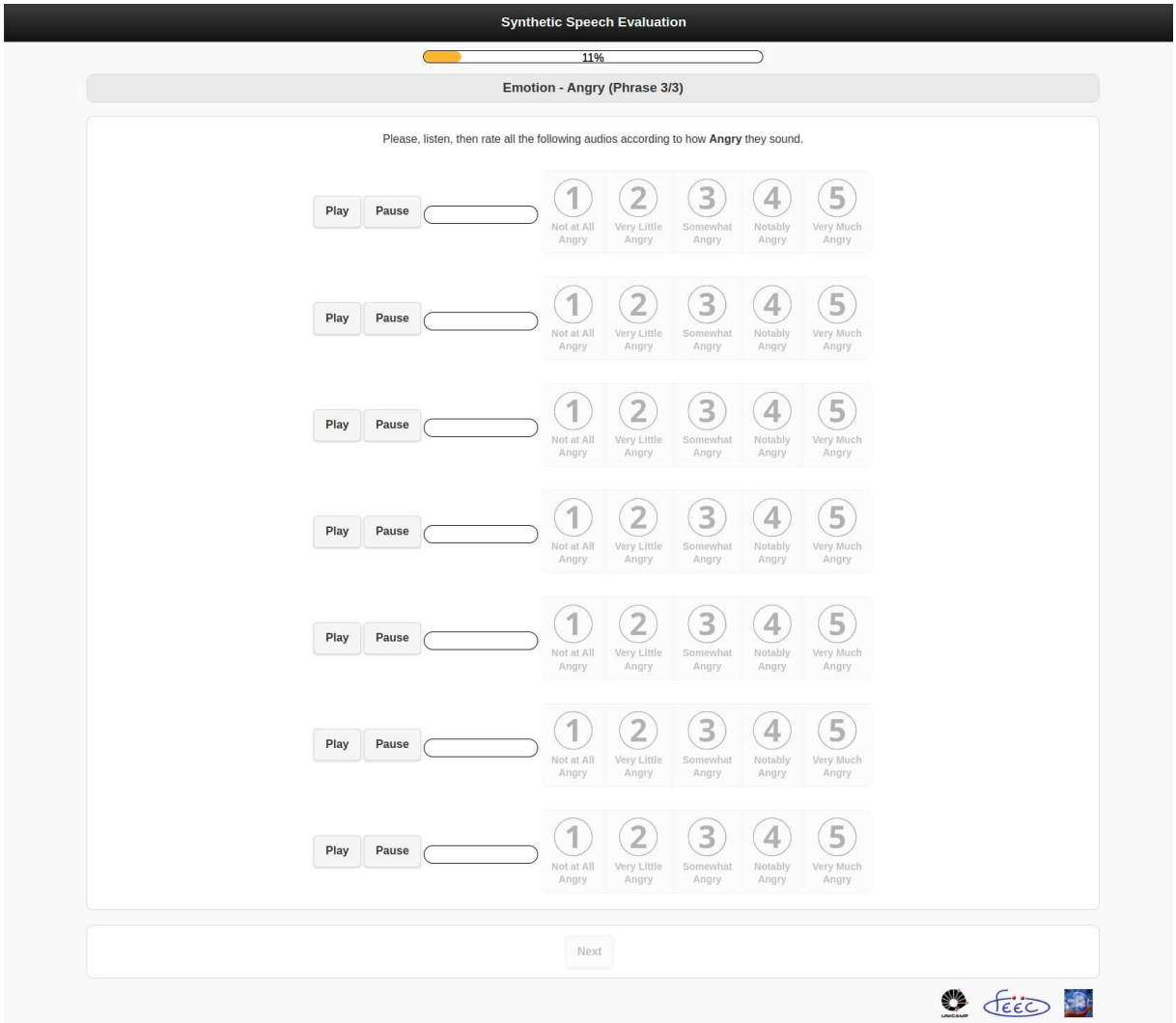


Figure A.5: Sample page of the style intensity experiment for the “angry” style. Repeated for all styles considered.

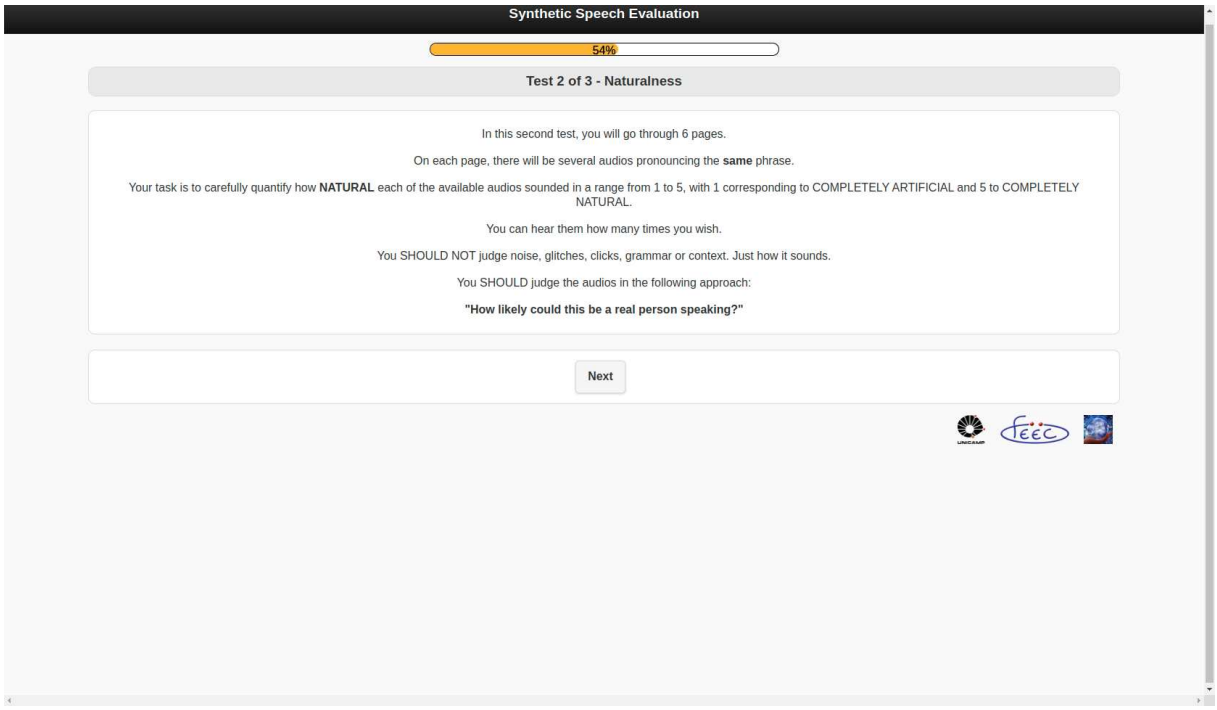


Figure A.6: Instructions for the naturalness experiment.

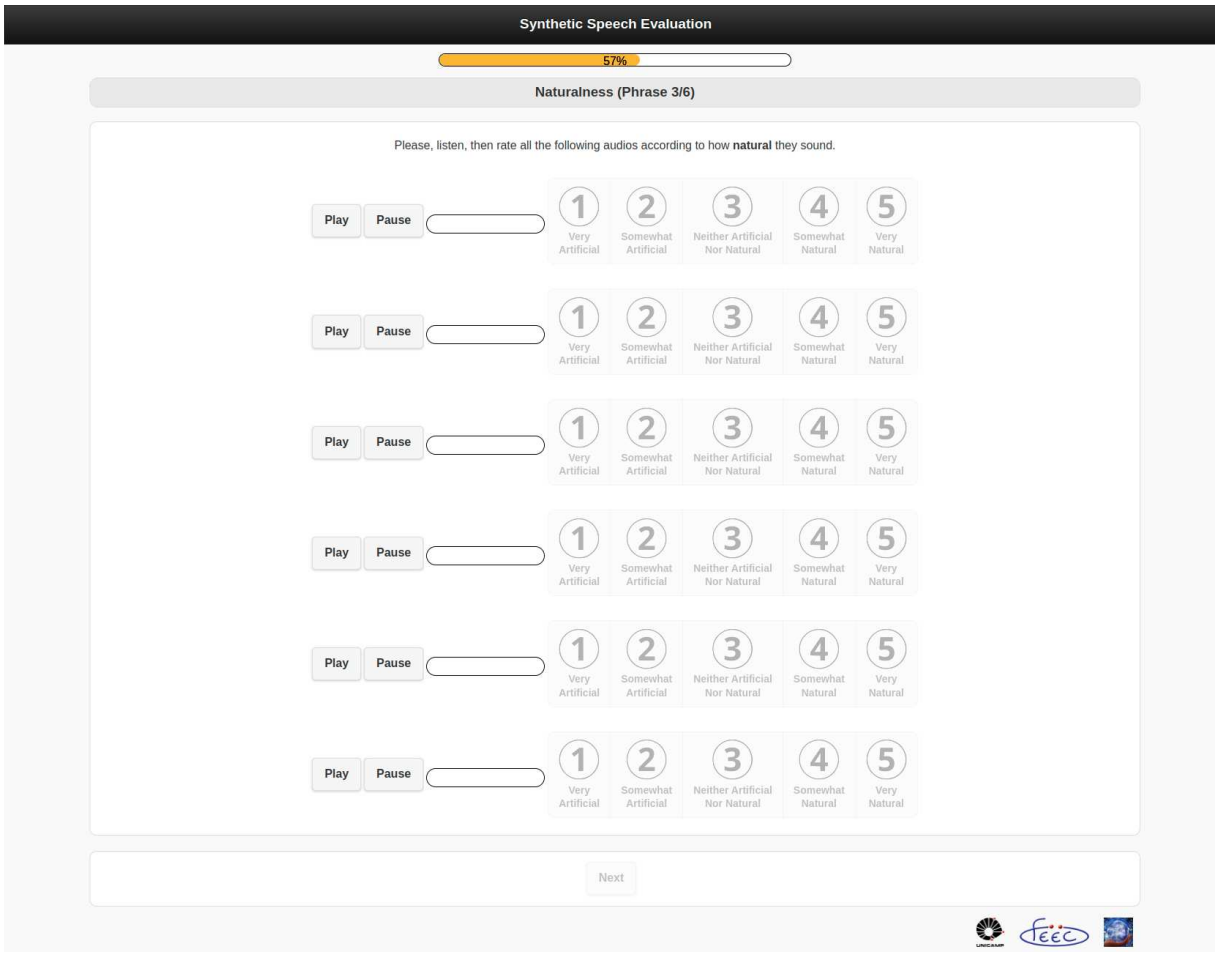


Figure A.7: Sample page of the naturalness experiment.

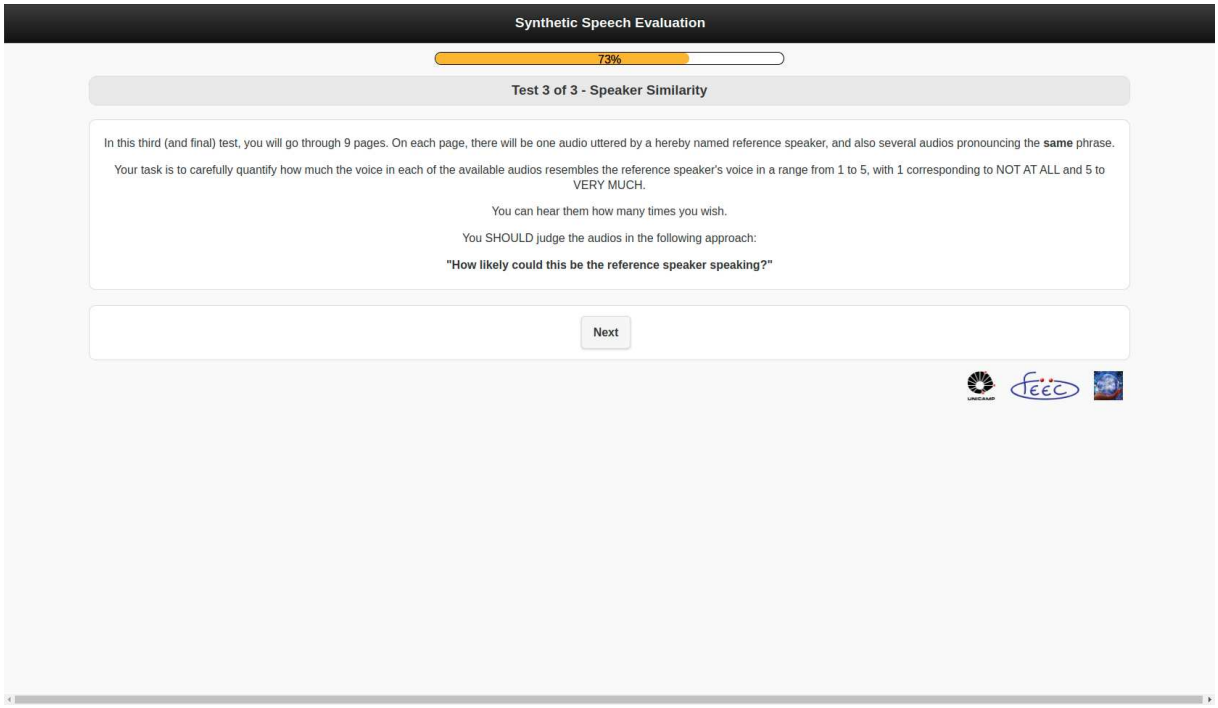


Figure A.8: Instructions for the speaker similarity experiment.

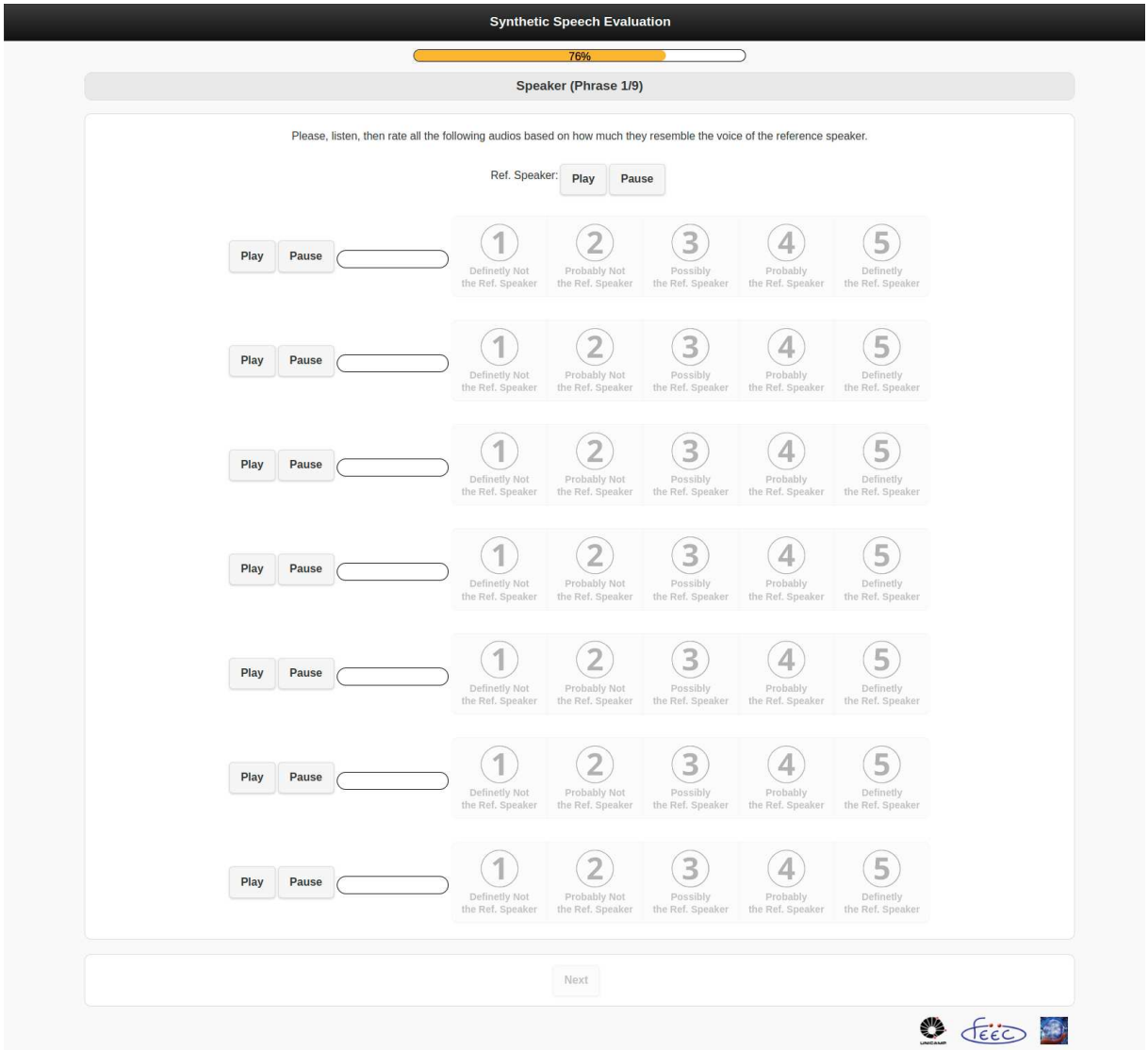


Figure A.9: Sample page of the speaker similarity experiment. Repeated for all styles considered.

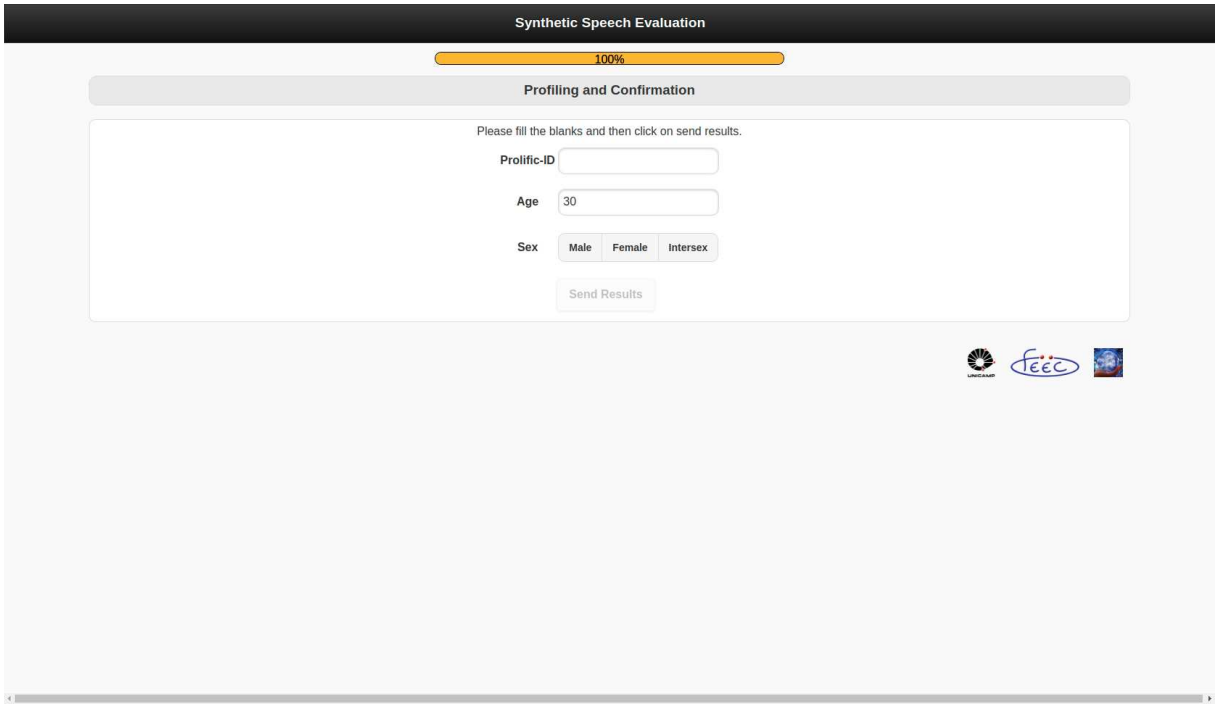


Figure A.10: Participant ID, age and gender profiling page.

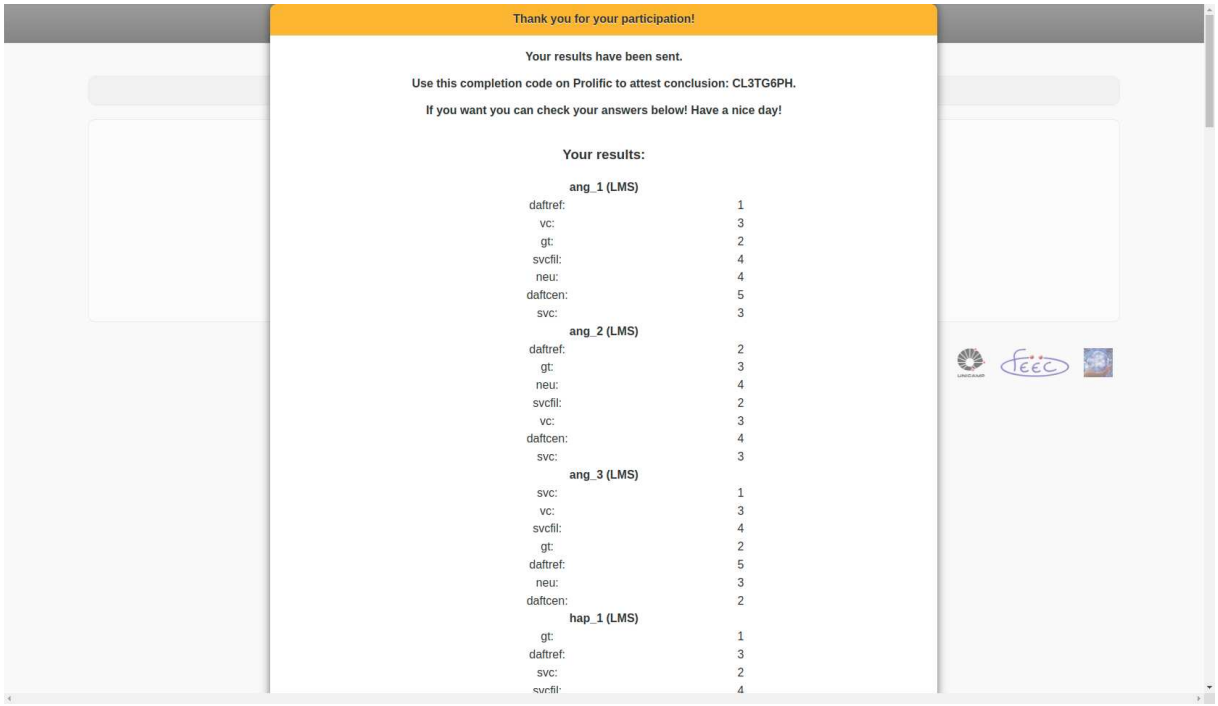


Figure A.11: Sample page showing the results for the participant at the end of the experiment.