EUROPEAN LANGUAGE EQUALITY

D2.14

Technology Deep Dive – Speech Technologies

Authors	Gerhard Backfried, Marcin Skowron, Eva Navas, Aivars Bērziņš, Joachim Van den Bogaert, Franciska de Jong, Andrea DeMarco, Inma Hernaez, Marek Kováč, Peter Polák, Johan Rohdin, Michael Rosner, Jon Sanchez, Ibon Saratxaga, Petr Schwarz

Dissemination level	Public
Date	28-02-2022

About this document

Project	European Language Equality (ELE)
Grant agreement no.	LC-01641480 – 101018166 ELE
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2021, 18 months
Deliverable number	D2.14
Deliverable title	Technology Deep Dive – Speech Technologies
Type	Report
Number of pages	72
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 28-02-2022
Work package	WP2: European Language Equality – The Future Situation in 2030
Task	Task 2.3 Science – Technology – Society: Language Technology in 2030
Authors	Gerhard Backfried, Marcin Skowron, Eva Navas, Aivars Bērziņš, Joachim Van den Bogaert, Franciska de Jong, Andrea DeMarco, Inma Hernaez, Marek Kováč, Peter Polák, Johan Rohdin, Michael Rosner, Jon Sanchez, Ibon Saratxaga, Petr Schwarz
Reviewers	Itziar Aldabe, Peter Polák
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland
	Prof. Dr. Andy Way – andy.way@adaptcentre.ie
	European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany
	Prof. Dr. Georg Rehm – georg.rehm@dfki.de
	http://www.european-language-equality.eu

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
5	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
6	CROSSLANG NV	CRSLNG	BE
7	European Federation of National Institutes for Language	EFNIL	LU
8	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR
9	European Civil Society Platform for Multilingualism	ECSPM	DK
10	CLARIN ERIC – Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium	CLARIN	NL
11	Universiteit Leiden (University of Leiden)	ULEI	NL
12	Eurescom (European Institute for Research and Strategic Studies in Telecommunications GmbH)	ERSCM	DE
13	Stichting LIBER (Association of European Research Libraries)	LIBER	NL
14	Wikimedia Deutschland (Gesellschaft zur Förderung freien Wissens e. V.)	WMD	DE
15	Tilde SIA	TILDE	LV
16	Evaluations and Language Resources Distribution Agency	ELDA	FR
17	Expert System Iberia SL	EXPSYS	ES
18	HENSOLDT Analytics GmbH	HENS	AT
19	Xcelerator Machine Translations Ltd. (KantanMT)	KNTN	IE
20	PANGEANIC-B. I. Europa SLU	PAN	ES
21	Semantic Web Company GmbH	SWC	AT
22	SIRMA AI EAD (Ontotext)	ONTO	BG
23	SAP SE	SAP	DE
24	Universität Wien (University of Vienna)	UVIE	AT
25	Universiteit Antwerpen (University of Antwerp)	UANTW	BE
26	Institute for Bulgarian Language "Prof. Lyubomir Andreychin"	IBL	BG
27	Sveučilište u Zagrebu Filozofski fakultet (Univ. of Zagreb, Faculty of Hum. and Social Sciences)	FFZG	HR
28	Københavns Universitet (University of Copenhagen)	UCPH	DK
29	Tartu Ulikool (University of Tartu)	UTART	EE
30	Helsingin Yliopisto (University of Helsinki)	UHEL	FI
31	Centre National de la Recherche Scientifique	CNRS	FR
32	Nyelvtudományi Kutatóközpont (Research Institute for Linguistics)	NYTK	HU
33	Stofnun Árna Magnússonar í íslenskum fræðum SAM (Árni Magnússon Inst. for Icelandic Studies)	SAM	IS
34	Fondazione Bruno Kessler	FBK	IT
35	Latvijas Universitātes Matemātikas un Informātikas institūts (Institute of Mathematics and Computer Science, University of Latvia)	IMCS	LV
36	Lietuvių Kalbos Institutas (Institute of the Lithuanian Language)	LKI	LT
37	Luxembourg Institute of Science and Technology	LIST	LU
38	Università ta Malta (University of Malta)	UM	MT
39	Stichting Instituut voor de Nederlandse Taal (Dutch Language Institute)	INT	NL
40	Språkrådet (Language Council of Norway)	LCNOR	NO
41	Instytut Podstaw Informatyki Polskiej Akademii Nauk (Polish Academy of Sciences)	IPIPAN	PL
42	Universidade de Lisboa, Faculdade de Ciências (University of Lisbon, Faculty of Science)	FCULisbon	PT
43	Institutul de Cercetări Pentru Inteligență Artificială (Romanian Academy)	ICIA	RO
44	University of Cyprus, French and European Studies	UCY	CY
45	Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied (Slovak Academy of Sciences)	JULS	SK
46	Institut Jožef Stefan (Jozef Stefan Institute)	JSI	SI
47	Centro Nacional de Supercomputación (Barcelona Supercomputing Center)	BSC	ES
48	Kungliga Tekniska högskolan (Royal Institute of Technology)	KTH	SE
49	Universität Zürich (University of Zurich)	UZH	CH
50	University of Sheffield	USFD	UK
51	Universidad de Vigo (University of Vigo)	UVIGO	ES
52	Bangor University	BNGR	UK

Contents

List of Acronyms

AAE	Adversarial Autoencoder
AD	Alzheimer's Disease
AI	Artificial Intelligence
AM	Acoustic Model
APAC	Asia Pacific Region
ASR	Automatic Speech Recognition
ASV	Automatic Speaker Verification
BERT	Bidirectional Encoder Representations from Transformers
BPE	Byte Pair Encoding
CA	Conversational Agent
CER	Character Error Rate
CLIP	Contrastive Language-Image Pre-training
CNN	Convolutional Neural Network
CTC	Connectionist Temporal Classification
CV	Curriculum Vitae
DARPA	Defense Advanced Research Projects Agency
DBN	Deep Belief Network
DCF	Detection Cost Function
DER	Diarisation Error Rate
DLE	Digital Language Equality
DNN	Deep Neural Network
DS	Data Science
DS-LTSM	Dual-Sequence Long Short-Term Memory
E2E	End-to-End
EER	Equal Error Rate
FA	False Accept
FR	False Reject
G2P	Grapheme-to-Phoneme
GAFA	Google, Apple, Facebook and Amazon
GDPR	General Data Protection Regulation
GMM	Gaussian Mixture Model
GPT-3	Generative Pre-trained Transformer 3
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
HCI	Human-Computer Interaction
HMM	Hidden Markov Model
IoT	Internet of Things
IPA	Intelligent Personal Assistant
JER	Jaccard Error Rate
JFA	Joint Factor Analysis
LID	Language Identification
LLM	Large Language Model
LLR	Log-Likelihood Ratio
LM	Language Model
LSTM	Long Short-Term Memory
LT	Language Technology
MFCC	Mel Frequency Cepstral Coefficients
ML	Machine Learning

MMLM	Multilingual Masked Language Modelling
MOS	Mean Opinion Score
MT	Machine Translation
NER	Named Entity Recognition
NIST	National Institute of Standards
NLP	Natural Language Processing
NLU	Natural Language Understanding
NN	Neural Network
OOV	Out-of-Vocabulary Rate
PER	Phoneme Error Rate
PLDA	Probabilistic Linear Discriminant Analysis
POS	Part-of-Speech
PR	Public Relations
RNN	Recurrent Neural Network
RWTH	Rheinisch-Westfälische Technische Hochschule Aachen (Aachen University)
S2S	Sequence-to-sequence
SA-ASR	Speaker-Attributed Automatic Speech Recognition
SdSV	Short-duration Speaker Verification
SER	Speech Emotion Recognition
SID	Speaker Identification
SOTA	State of the Art
SR	Speaker Recognition
SSH	Social Sciences and Humanities
ST	Speech Technologies
SV	Speaker Verification
SVM	Support Vector Machines
SW	Software
TDNN	Time delay neural network
TTS	Text to Speech
VA	Voice Assistant
WER	Word-Error-Rate
WFST	Weighted Finite-State Transducers

Abstract

D2.14 provides an overview and describes the state of the art and developments within the field of Speech Technologies (ST). This field is interpreted to comprise technologies aimed at the processing and production of the human voice, both, from a linguistic as well as paralinguistic angle. It provides an in-depth account of current research trends and applications in various ST sub-fields, details technical, scientific, commercial and societal aspects, relates ST to the wider fields of NLP and AI and provides an outlook of ST towards 2030. Chapters 3 and 4, presenting the main ST components and the state-of-the-are are divided according to the different sub-fields covered: Automatic Speech Recognition (ASR), Speaker Identification (SID), Language Identification (LID), technologies targeting paralinguistic phenomena and Text to Speech (TTS). Chapter 5 discusses the main gaps in speech technologies related to issues such as data requirements, ST performance, explainability of the critical methods, regulations influencing the pace of development in the field or specific requirements for less-resourced languages. The following chapter presents aspects of the wider impact of ST on society and describes the contributions of speech technologies to Digital Language Equality. Chapters 7-9 outline some breakthroughs needed, the main technology visions and present how ST may fit into and contribute to a wider vision of what may be termed Deep Natural Language Understanding. The deliverable integrates the views of companies and institutions involved in research, commercial exploitation and application of speech technologies.

1 Introduction

Speech – as the most natural manner for humans to interact with computers – has always attracted enormous interest in academia and the industry. Speech Technologies (ST) have consequently been the focus of a multitude of research and commercial activities over the past decades. From humble beginnings in the 1950'ies, they have come a long way to the current state-of-the-art, deep-neural-network (DNN) based approaches. Stimulated by a shift towards statistical methods, the 1980'ies witnessed an era of Hidden-Markov-Models (HMM), Gaussian-Mixture-Models (GMM) and word-based n-gram models combined into speech recognition engines employing ever more refined data-structures and search algorithms such as prefix-trees or Viterbi beam-search (Jelinek, 1998). The availability of data resources to train these systems was limited to only a few languages, often driven by security (and commercial) interest. Even then, work on Neural Networks (NN) was already being carried out and viewed by many as the most promising approach. However, it wasn't until later (2000's), when the availability of training data paired with advances in algorithms and computing power finally began to come together to unleash the full potential of NN-based ST.

As Artificial Intelligence (AI) was entering springtime again – following the so-called *AI-Winter* (Hendler, 2008; Floridi, 2020) – general interest, research activities, funding opportunities and investments witnessed dramatic growth. This has led to significant progress in many related fields, including those of Natural Language Processing (NLP), Machine Learning (ML) and Data Science (DS). Speech Technologies profited greatly from these advances and have become mainstream technologies, deployed in numerous domains and viewed as commodities in many instances.

Especially over the past couple of decades, ST have evolved dramatically and become omnipresent in many areas of human-machine interaction. Embedded into the wider fields of Artificial Intelligence (AI) and Natural Language Processing, the expansion and scope of ST and their applications have accelerated further and gained considerable momentum. In the recent years, these trends were paired with the undergoing, profound paradigm shift related to the rise of the foundation models (Bommasani et al., 2021), such as BERT (Devlin et al.,

2018), GPT-3 (Brown et al., 2020), CLIP (Radford et al., 2021) and DALL-E (Ramesh et al., 2021) – a class of models trained on broad data at scale, adaptable via natural language prompts and able to perform reasonably well on a wide range of tasks despite not being trained explicitly on those downstream tasks.

Changes, new requirements and restrictions introduced by the COVID-19 pandemic paired with substantial advances in algorithms and specialised, high-performance hardware as well as the wide availability of mobile devices have led to massively increased adoption and further technological improvements. With speech and natural language forming fundamental pillars of (human) communication, ST may now even be perceived as "speech-centric AI".

During these exceptional times of a global pandemic, businesses and administrations alike have been encouraged and urged to improve their virtual ties with customers and citizens. This has led to increased adoption and extension of the scope and application of virtual assistants, chatbots, and other voice-enabled technologies. With the emergence of intelligent virtual assistants ST have become ubiquitous, yet many ST systems can only cope with restricted settings and domains and can be used only with the most widely spoken of the world's many thousands of languages. For languages with fewer speakers and thus of lesser commercial interest, ST systems are still all but absent and/or severely limited in their scope. As a consequence, millions of individuals who speak these languages are virtually cut off from a wide range of speech-based services and applications or forced to communicate in languages other than their native one(s).

Current technologies often require the presence of large amounts of data to train systems and create corresponding models. Despite the lack of massive volumes of training material (e. g., transcribed speech in case of ASR or annotated audio for TTS), recent advances in ML and ST have begun to enable the creation of models also for less common languages. These approaches however are generally more complex, expensive and less suitable for wide adoption. While recently presented results indicate that novel approaches could indeed be applied to address some of the challenges related to the creation of models for low resourced languages, the scope of their application and inherent limitations are still the subject of ongoing research (Lai et al., 2021).

The democratisation of ST may thus be viewed as part of the democratisation of AI in general, not only in the sense of allowing the general public to participate in the generation of models and solutions, but also to equally participate in their use.

This report describes the state of the art in the field of speech technologies, shortcomings and challenges and outlines technical, scientific as well as commercial alleys towards the future. It provides an in-depth account of the current research activities and applications in various sub-fields and puts these activities into perspective and relation with each other.

Chapter 2 presents the scope of this deep dive and introduces the main fields covered. In chapter 3, the main components of the field of Speech Technologies are presented followed by a description of different sub-fields of ST: Automatic Speech Recognition (ASR), Speaker Identification (SID), Language Identification (LID), technologies addressing paralinguistic aspects of speech and Text to Speech (TTS). Chapter 4 provides an in-depth description of the current state-of-the-art methods of each of the ST-sub-fields covered. Chapter 5 discusses the main gaps in speech technologies such as the ones related to data needs, ST performance, explainability of the critical methods, regulations influencing the pace of development in the field, and specific challenges concerning low-resourced languages. In chapter 6, ST contributions to Digital Language Equality and the wider impact of the technologies on society are presented. These include the discussion about digital language inequalities, biases, fairness and ethical issues related to the use of ST. We also outline the impact of ST on users with special needs and present the relations between the development and application of ST in a wider landscape of business environments, its footprint on energy consumption and ramifications in the context of privacy and trust in technology. Chapter 7 outlines challenges and indicates several breakthroughs needed to overcome them. These include the access

and discoverability of training data and changes required in training paradigms. Furthermore, a range of challenges related to the performance, robustness, evaluation, integration of ST components beyond the field and requirements for reaching out to other communities and non-expert users is discussed. The chapter concludes with an overview of the requirements for the alignments with the existing EU policies and changes needed on the policy level. Chapter 8 presents the main technology visions and development goals until 2030. Finally, chapter 9, presents how ST fit into and contribute to a wider vision of Deep Natural Language Understanding.

2 Scope of this Deep Dive

The scope of this deep dive encapsulates a wide range of speech technologies including language identification and speaker recognition, automatic speech recognition, technologies to address paralinguistic phenomena as well as text to speech. It gathers and synthesises the perspectives of the European research and industry stakeholders on the current state of affairs, identifies several main gaps affecting the field, outlines a number of breakthroughs required and presents the technological vision and development goals in the next ten years.

The views expressed stem from a diverse set of groups and comprise elements of research as well as the industry.

In line with other deep dives of WP2, we adopt a multidimensional approach where both market/commercial as well research perspectives are considered and concentrate on these important aspects – technologies, models, data, applications and the impact of speech technologies on society.

The tendency for the combination of technologies into more powerful systems, encompassing several individual technologies and models has become apparent and is reflected in numerous occasions within this document. We expect this trend to continue and even get stronger over time.

ST can be investigated and researched in their own right and much effort has been invested in this direction (and continues to be invested). However, their full potential often only becomes evident when they are combined with further technologies forming intelligent systems capable of complex interaction and dialogues. This kind of interaction may encompass a diverse set of contexts, history and span multiple modalities. To the casual user, individual components then become blurred and almost invisible with one overall application acting as the partner within an activity which may otherwise be carried out together with a fellow human being. In this setting, the conglomerate and aggregation of technologies form a step away from narrow and highly specialised systems towards combined and complex systems, providing a notion of a more general and broader kind of intelligence. Speech and language, as the most natural and appropriate vehicle for humans to communicate with machines in many instances, thus becomes the gatekeeper to and core of a broader kind of AI.

3 Speech Technologies: Main Components

3.1 Automatic Speech Recognition

General introduction

The goal of an automatic speech recognition system is to convert speech into sequences of units such as words or characters. In the process, several steps are performed, involving a

variety of models and algorithms. These convert the raw audio input into increasingly more abstract units (such as features, phonemes, phones, morphemes, words) and eventually to the desired unit of transcription. Each of these steps involves different knowledge sources and can be modelled by individual components or by combined models spanning several steps. Traditionally, these knowledge sources were each modelled and optimised individually. More recently, models have been combined and eventually resulted in so-called end-to-end (E2E) models[1] aiming to represent (and optimise) the complete transcription process within a single model.

Conceptually, ASR systems consist of an acoustic model (AM), a lexicon and a language model (LM). The AM's goal is to model speech sounds and their variations. The lexicon defines the inventory of units (words) to be recognised. Each of these units has one or more pronunciations linking it to the AM and the lower-level audio-units (e. g., phonemes, via a pronunciation-lexicon). Special components mapping spellings to sounds (grapheme-to-phoneme, G2P) are employed to create these pronunciations in a consistent manner. Finally, the LM's task is to model how the units of the lexicon interact, typically by assigning probabilities to sequences of these units.

Processing

ASR systems typically perform some initial pre-processing to produce a sequence of features in regular time-intervals (e. g., every few milliseconds). Traditionally these features were modelled after the human-ear (e. g., mel-cepstrum coefficients) but potentially they may also be created by the initial stages of NNs. AM and LM are then employed within a search algorithm to produce a transcript or set of alternative transcripts. The basic process was introduced already in the 1980'ies by Bahl and Jellinek (Jelinek, 1998). Some post-processing, e. g., for handling numbers or acronyms, is sometimes applied to this initial transcript to produce a final transcript. The ASR-process can be carried out in a causal fashion (sometimes also referred to as online) and in (near-) real-time or running from pre-existing audio (sometimes referred to as offline), in which case more context is available for the process and faster than real-time processing can be achieved. During the past decade, hybrid models (Bourlard and Morgan, 1993; Seide et al., 2011), combining traditional elements such as HMM with NNs as well as E2E models, combining various levels of processing, have witnessed considerable progress. This tendency continues, also allowing for novel methods involving different modalities and the integration of research results from other fields such as machine translation (MT, e. g., transformers (Vaswani et al., 2017)) or visual processing (e. g., the use of convolutional NN's (Gu et al., 2018)).

Performance measures

The performance of ASR systems is typically measured in terms of Word-Error-Rate (WER) using dynamic alignment between reference and transcript. The WER depends on a variety of factors concerning the acoustic conditions under which the speech was produced, speaker-specific traits such as nativeness, emotional state, etc. and a series of linguistic factors such as dialects, social register, specific domains or the spontaneity of speech. As models are typically of a statistical nature, the availability of adequate training corpora is a key factor in determining the performance of ASR systems as are the algorithms applied to combine the various models during the search.

[1] The exact definition of E2E varies according to context and authors, but for practical purposes, it can be assumed to mean a direct conversion from input (audio) to output (transcript).

Current trends

Recent years have witnessed substantial progress requiring less and less labelled speech data for training and the adoption of semi-supervised or unsupervised strategies for model creation. However, there is still a big disparity between ASR performance in languages and domains for which there are copious amounts of labelled data and those without such rich resources.

The methods prevalent in ASR systems today are based on statistical methods and ML, particularly on E2E DNN models. These are active areas of research and vary widely in complexity and scope. Pre-trained models are often fine-tuned specifically for the intended use and also serve as the starting point of ASR for specific languages. Methods such as HMM, hybrid HMM-NN models or weighted finite-state-transducers (WFST) are likewise still in use. Prior approaches are slowly being phased out as progress and the applicability of NN-based models takes precedence in many cases. A set of frameworks and toolkits such as Kaldi (Povey et al., 2011) have allowed a broader audience to develop and use models for ASR. The utilisation of language models – also in combination with NN-based methods – is still favoured for ASR system performance. In this area, several standards exist, e. g., ARPALM or KenLM style language models can be supplied and plugged into the major ASR frameworks available. Whereas ASR systems typically output a single transcript, the output of richer structures such as n-best or lattices can be beneficial for information retrieval contexts.

Data needs

There has been a strong dependence on the availability of large corpora to train the current state of the art systems. This has left many institutions unable to compete or innovate to a meaningful extent. As a consequence, the rise of ASR systems by *super-institutions*, mostly industrial, has dominated ASR development. Data requirements have exploded, with increasingly tighter and more restricted returns on investment. For instance, it is not unheard of to see 1000 hours or more of language-specific transcribed data being used for the creation of an ASR model. The positive flip-side to this is that the architectures of the latest models have already been built with transfer learning and fine-tuning in mind, requiring far less aligned training data. The degree to which these pre-trained models perform when ported to other languages also depends on the linguistic proximity (phonetic, phonological, etc.) of the language to the base model. The smaller this proximity is, the more fine-tuning data is typically required (e. g., 10 hours vs. 100 hours). Data requirements for language model generation are far easier to come by for most languages, as written text is generally much more readily available. Specific instances, such as languages with strong dialectal influences and non-standard spelling (e. g., the various dialects of spoken Arabic) form notable exceptions here. Domain-specificity of texts as well as regional particularities and lexical shift over time need to be taken into account.

Target uses

ASR serves as a key technology for turning unstructured data into structured content, thus making it accessible for downstream processing (enrichment, information retrieval, dialogue systems, etc.). It is an essential ingredient in supporting speech input as the most natural way for humans to interact with computer systems.

Speech technologies have come to be regarded more and more as a commodity, with billions of mobile devices offering interactivity through voice and voice-appliances entering the homes of users (smart homes, IoT). Spurred by an increased need for sanitation, touchless interaction is becoming the preferred manner of interaction in times of a pandemic like the current COVID-19 one. From the perspective of businesses, speech technologies promise

the possibility to extend services with new capabilities, making interaction more intuitive and natural and allowing at the same time to scale services to levels that were previously impossible (due to lack of resources) or not cost-effective. As such, speech technologies are viewed to play a major part in the process of digital transformation.

In most contexts – other than personal dictation systems – ASR systems are meant to work in a speaker-independent manner. Due to the statistical nature of models and the datasets employed to train these models, certain kinds of bias have been observed in the past, e. g., early ASR systems used to work better for male speakers than for female speakers. As in all other ML-based systems, the specificities of the training data are reflected within the resulting models and special care needs to be taken in order to mitigate and minimise this effect. More recently, the use of language concerning gender is receiving higher attention. Morphological particularities, pronouns etc. need to be reflected properly in ASR models in order to adequately transcribe utterances containing such elements.

3.2 Speaker Recognition

General introduction

Speaker recognition (SR) refers to the process where a machine infers the identity of a speaker by analysing his/her voice/speech. The basis of SR is the task of speaker verification (SV). In this task, one or more enrolment (registration) utterances from a speaker are compared with a test utterance. The system then provides a score that indicates how likely it is that the test utterance is spoken by the same person as the enrolment utterances. Speaker verification is often referred to as an open set problem since neither the enrolled speakers nor the test speaker is used for building the system. A system that can perform the speaker verification task can also be used for speaker identification, where many speakers are enrolled (registered). In testing, an utterance is assigned to one of the enrolled speakers or possibly also to none of them. Speaker clustering is based on a set of unlabelled recordings, with the aim to infer the number of speakers and attribute them to the respective recordings. Many applications also require combinations of the above tasks. For example, a set of utterances could be subjected to identification after which all utterances which did not match any of the enrolled speakers are clustered.

Scope

Age and gender recognition are two tasks closely related to SR since in typical databases the property of “age” usually does not change significantly between different recordings of the same speaker and “gender” does not change at all. Utterance representations (embeddings) produced by SR can therefore be used as an input feature to age and gender recognition systems. Age detection is frequently divided into age ranges rather than being targeted at the exact age of a speaker.

Emotion recognition from the speech is another highly relevant topic. It is used in a wide variety of applications from businesses to governmental bodies. For example, in call centres, it supports monitoring of client support quality and is employed to study clients’ reactions to certain emotional triggers. Multiple studies have been conducted on emotion recognition from the speech signal.

Data needs

State-of-the-art speaker recognition systems are trained on data from several thousands of speakers, each providing many utterances. In addition, this training data is often augmented with versions of the utterances with additive noise or reverberation (data augmentation).

Most of the research papers investigate machine learning model performance on public artificially created datasets such as EMODB (Burkhardt et al., 2005), IEMOCAP (Busso et al., 2008), TESS (Pichora-Fuller and Dupuis, 2020), and RAVDESS (Livingstone and Russo, 2018). Although this approach ensures a common benchmark, it ignores the fact that in the real world speech data is not as clear or well-defined. A few papers such as (Kostoulas et al., 2008), (Dhall et al., 2013) and (Tawari and Trivedi, 2010) aim to address this issue.

Interpretability, explainability, transparency

Speaker recognition systems are typically designed to output the log-likelihood ratio (LLR) for the hypothesis that the speakers are the same vs. the (alternative) hypothesis that the speakers are different. In the mathematical sense, an LLR is completely interpretable for anyone with a sufficient level of expertise and training. The need for explainability, i. e., for the system to be able to explain how it reached its conclusion, naturally depends on the application. If the result of speaker recognition is to be used as forensic evidence at court, this may indeed be considered important. So far, methods for explainability of speaker recognition systems have received relatively little attention from researchers. Some work with gradient-based methods has been presented in Muckenhirn et al. (2019).

3.3 Language Identification

General introduction

Language identification aims to recognise which language is spoken in an utterance. Typically, it is treated as a closed set classification problem, i. e., the languages to be recognised are fixed. Accordingly, language recognition systems can be built using supervised machine learning techniques. In this sense, it is a simpler problem than e.g., ASR where the output space is a structured combination of fixed symbols (words or graphemes), or SR where the typical applications require the system to compare voices from speakers not seen in training. Partly for this reason, language recognition gains less attention in the research community than ASR or SR. For example, the most recent large evaluation LID took place in 2017 (organised by the National Institute of Standards and Statistics, NIST) whereas there are several recurring evaluations per year in ASR and SR. It should be noted that LID is generally performed before ASR. An alternative would be to process the audio with ASR systems for many different languages and then analyse their text output and confidence scores to determine the languages. However, this approach is not practical and generally not effective (even though it was allegedly applied by Amazon Alexa in the early stages).

Data needs

Building state-of-the-art LID systems typically require 15 or more hours of speech per language. Compared to other speech processing tasks, acquiring labelled data (i. e., the audio and the language label) for language identification training is fairly easy. For example, the spoken language in video recordings on the internet can often be inferred from metadata. Therefore, companies and research laboratories typically have LID data for more than 50 different languages. Dependencies on domain and audio conditions, however, still apply to some extent. Voxlingua107 (Valk and Alumäe, 2021) is a dataset for spoken language recognition of 6628 hours (62 hours per language on the average) and it is accompanied by an evaluation set of 1609 verified utterances.

3.4 Assessment of emotions, cognitive conditions and personality traits

General introduction

Recognition of emotions, mood and personality traits from the speech is an active research area with a wide range of potential applications such as, e. g., intelligent human-computer interaction, call centres, onboard vehicle driving systems, financial security and smart environments. As emotions play a crucial role in interpersonal communication, perceiving the emotional states of interlocutors is also an essential component for holistic modelling of users, enabling an instantaneous reception of feedback related to the system's actions and better-informed action selection. The emotional cues as well as the ability to detect users personality traits and cognitive conditions are also useful for a system adaptation to a user in the longer term. Automatic voice analysis techniques are also used for the detection of cognitive conditions, monitoring of patients suffering from a neurodegenerative disorder, such as Alzheimer's disease (AD).

Scope

Speech and the human voice have been investigated for a considerable amount of time with the goal to infer information about the speaker's mood and emotions (Schuller et al., 2011; Batliner et al., 2011; Koolagudi and Rao, 2012; Dasgupta, 2017; Albanie et al., 2018; Rouast et al., 2019; Akçay and Oğuz, 2020), mood disorders (Huang et al., 2019) and signs of depression (Cummins et al., 2015; Toto et al., 2021). A similar strand of research aims at the detection of specific cognitive states and conditions (Tóth et al., 2018; Konig et al., 2018; Pulido et al., 2020), certain diseases, such as COVID-19 (Dash et al., 2021; Schuller et al., 2021) and other health states (Sertolli et al., 2021), as well as personality traits (Mairesse et al., 2007; Polzehl et al., 2010; Mohammadi and Vinciarelli, 2012; Guidi et al., 2019; Lee et al., 2021). In Speech Emotion Recognition (SER) and in speech synthesis, both the discrete and dimensional emotional models are used (Kwon et al., 2003; Giannakopoulos et al., 2009; Schröder, 2004). In the dimensional models, two or three continuous spaces for arousal, valence, and potency are frequently considered. The discrete models' taxonomies are more diverse, with differences in the number of emotions included.

Work in the areas that target the paralinguistic aspects of speech is highly interdisciplinary in nature. It brings together fields such as psychology, medicine or computational linguistics. Due to its inter-disciplinary heritage, it has brought with it a wide diversity in terminology and approaches which can only be described at a superficial level within this document.

Data needs, interpretability, explainability, transparency

The amount, quality and diversity of data employed are as diverse as the efforts and teams active in this domain. In many cases, experiments are conducted in a qualitative manner. Ethical and legal issues are of high importance, especially in clinical settings. Regarding the automatic processing of speech with the focus on paralinguistic aspects, the Interspeech Computational Paralinguistics Challenge (ComParE[2]) has been taking place for the past 12 years, introducing new tasks every year and addressing important but as of yet under-explored paralinguistic phenomena (Schuller et al., 2020).

In SER a variety of data sets is being used, including the acted (simulated), natural and elicited (induced) speech. The examples of prominent data sets include Berlin Emotional Database (German) (Borchert and Dusterhoft, 2005), eNTERFACE'05 Audio-Visual Emotion

[2] http://www.compare.openaudio.eu

Database (English) (Martin et al., 2006), SEMAINE Database (English, Greek, Hebrew) (McKeown et al., 2011), RECOLA Speech Database (French) (Ringeval et al., 2013), Vera Am Mittag Database (German) (Grimm et al., 2008), AFEW Database (English) (Kossaifi et al., 2017), Turkish Emotional Speech Database (Turkish) (Oflazoglu and Yildirim, 2013). The size of the datasets used in this task varies significantly both in the number of speakers (from 2 to a few hundred), their profile, and the number of utterances included in a set. Similarly, different emotion taxonomies are being used. See (Swain et al., 2018; Khalil et al., 2019; Akçay and Oğuz, 2020) for an extensive review of databases, models, resources as well as approaches applied in the SER field.

Recently, dedicated datasets and challenges aimed at the detection of cognitive disorders, especially Alzheimer's dementia were created. For example, the ADReSS Challenge at INTERSPEECH 2020[3] introduced a task for evaluating different approaches to the automated recognition of Alzheimer's dementia from spontaneous speech. The challenge provided the researcher community with a benchmark speech dataset that has been acoustically preprocessed and balanced in terms of age and gender, defining two cognitive assessment tasks: detection of the Alzheimer's speech, and the neuropsychological score regression task (Luz et al., 2020).

3.5 Text to Speech

General introduction

The task of generating speech from some other modality like text, muscle movement information, lip-reading, etc. is called speech synthesis. In most applications, the text is used as the preferred form for the input and this particular case is called text to speech (TTS) conversion. The goal of a TTS system is to generate speech from written natural language. In the ideal case, TTS systems should produce natural voices that can communicate in a certain style, are able to reflect the accent, mood and other characteristics of the speaker and are indistinguishable from those of humans. Many factors affect the quality of the synthetic voices, such as the synthesis technique applied and the size and quality of the available speech databases used for model creation/adaptation.

Machine learning vs. symbolic methods

Current TTS systems are based on deep learning. Neural networks have all but replaced traditional synthesis technologies such as HMM-based statistical parametric synthesis and concatenative synthesis achieving a better voice quality while demanding less preparation of training data. Speech and the corresponding aligned text (or phonetic transcription) is usually the requirement to properly train these DNN based TTS systems.

DNN based systems are typically split into two parts – a neural acoustic model which generates acoustic features of the speech given linguistic features, such as graphemes or phonemes, and a neural audio generation model (also known as a vocoder) which generates an audio waveform given these acoustic features, e.g., Mel-spectrogram frames. However, the text is only able to describe certain aspects of the content that is to be produced as speech. Thus such models generally produce only neutral speech. Recently, efforts have been made towards synthesising expressive speech. This can be achieved either by directly controlling the rhythm, pitch, and energy of the speech (Valle et al., 2020a; Ren et al., 2020), or indirectly by passing an embedding corresponding to a certain emotional style (Wang et al., 2018).

[3] http://www.interspeech2020.org/index.php?m=content&c=index&a=show&catid=315&id=755, acessed 17.1.2022

Data needs

Very large corpora are needed to obtain good synthetic voices with this technology, often as big as hundreds of hours of very high-quality recordings. Big companies have taken the lead in the development of this kind of system. Using around 20 hours of carefully selected speech from a professional speaker and recorded in a professional studio is a common requirement fulfilled by most of the current commercial systems. Smaller research groups and companies have difficulties to compete taking into account the quantity of good quality speech recordings required. Also, small languages (in the sense of commercial interest and/or a number of speakers) suffer from this issue as the main companies develop their systems almost exclusively for major languages and there are no available speech corpora of the required size to train the proposed architectures for these languages.

Instead of recording dedicated datasets, audiobooks provide a viable alternative and can be used as a potential source of speech data. Audiobooks usually contain several hours of high-quality audio from a single speaker, thus adhering to the requirements of DNN based TTS system training. Coupling the audio with the text version of the book, e.g, with the help of forced alignment by ASR, can significantly decrease the effort required to obtain audio transcriptions. However, such an approach requires some additional processing steps, such as filtering the audio from any background noises, text normalisation, checking the audio and text for any discrepancies, etc.

4 Speech Technologies: Current State of the Art

Deliverable *D1.2 Report on the state of the art in LT and language-centric AI* provides an overview of previous and present approaches for the technologies covered by this document. In the following, further aspects and developments since the creation of the beforementioned document are discussed.

4.1 Automatic Speech Recognition

SOTA of the current methods and algorithms

The traditional (and by now *classical*) pipeline of ASR consists of components for audio preprocessing, an acoustic model, a pronunciation model as well as a language model defined over units of a lexicon. Within the scope of a search algorithm, these elements are combined to produce the most likely transcript given the input audio. In this scheme, models generally are of a generative kind (such as GMMs, HMMs and n-gram models for the LM) and optimised individually. This setup was considered standard in the first decade of this century.

However, already starting in the early 2000s, more and more of these components were being replaced with DNNs, hybrid DNN-HMMs, LSTM-HMMs or RNNs. This change was made possible by advances in algorithms and models as well as the massive increase in available training data and computing power (in particular of GPUs). As a result, WERs could be reduced by more than 50% in many domains and languages (Schlüter, 2019). However, the performance of ASR systems still varies dramatically depending on the domain and language, with low-resource languages still exhibiting WERs resembling those of English many years ago.

For applications in practice (*ASR in the Wild*), hybrid systems combining traditional elements such as HMMs and DNNs still dominate the state of play. As such, they can be regarded as state-of-the-art outside of research labs. Toolkits like Kaldi (Povey et al., 2011)

provide a sound basis for the development of systems for research as well as commercial environments. Kaldi is currently undergoing a redesign process and will be named K2[4].

The initial phases of the introduction of NNs concerned the pre-processing, e. g., tandem features and bottleneck features (Hermansky et al., 2000; Grézl et al., 2007) and AM, with models taking into account increasingly larger context (recurrent models like RNNs, LSTMs, GRUs). Approaches such as LSTM (Sundermeyer et al., 2015) augmented this by allowing novel manners to represent the LM.

The introduction of sequence-to-sequence (S2S) approaches such as Connectionist Temporal Classification, CTC (Graves et al., 2006), or "Listen, Attend and Spell" (Chan et al., 2015) took this process to the extreme. They introduced one global model that maps acoustic features directly to the text. This model is optimised with only one objective – as opposed to before, where different sub-models were optimised independently and using different objectives.

These end-to-end models typically consist of an encoder (DNN) generating a deep and rich representation of the input (audio) followed by a decoder (DNN) paying attention (Bahdanau et al., 2014) to the (encoded) input as well as its internal states and the last emitted outputs.

State-of-the-art approaches usually utilise RNNs and Transformers (Vaswani et al., 2017), though recent research suggest that the latter is better (Karita et al., 2019; Zeyer et al., 2019). Novel research tries to overcome some shortcomings of the Transformers for speech by combining them with, e.g., convolutional NNs (CNNs) (Dong et al., 2018; Gulati et al., 2020).

Novel approaches, such as Wav2Vec 2.0 by Facebook (Baevski et al., 2020) focus on leveraging vast amounts of unlabelled speech data. In this approach, latent representations of audio are produced which represent speech sounds similar to (sub-)phonemes which are then fed into a Transformer network. The approach has been shown to outperform other typical paths of semi-supervised methods, while also being conceptually simpler to implement and execute. The possibility to employ smaller amounts of labelled data as well as being able to train multilingual models provide strong arguments for this approach.

Current trends regarding the SOTA

Several trends concerning the SOTA can be discerned and can be expected to also continue in the foreseeable future:

- Manual configuration or customisation will be minimised or eliminated altogether.
- Several standard-evaluations of ASR exist, leading to a systematic push in frontiers and performance on a continuous basis. Existing evaluations are likely to be complemented by further, more complex setups (also non-English and/or multilingual!).
- As text is abundant and LMs can be trained from text-only, the incorporation of strong LMs (to bias NNs) will remain an active topic of research. Shallow and deep fusion (Le et al., 2021) to blend different models, such as specialised LMs and generic LMs, provide current approaches addressing this problem
- The integration of further knowledge sources into E2E systems.
- Reinforcement learning has gained popularity in a number of areas. The adoption also for certain tasks within ASR is pending.
- A lot of attention has been paid to single microphone settings (see, e. g., (Kanda et al., 2021a,b) for examples of recent works on E2E multi-speaker ASR for meeting transcription). Multi-speaker, multi-channel, multi-microphone setups may provide further angles and lead to improvements.

[4] https://www.kaldi.dev/industry_overview.html

- The combination of ASR with further NLP technologies (such as MT) in a single model may produce even more powerful combined, E2E models.
- As model complexity has become prohibitive in many cases for all but the most potent participants, the trade-off between system-complexity and performance is a promising target area for future work.
- A multi-pass approach to recognition, as it was popular in the pre-DNN days may see a revival, due to work on fusion and combination with further knowledge sources.
- Language-agnostic or multilingual models, cross-lingual and multi-lingual training and models (also in the guise of "co-training" of models) are receiving increased attention.
- Further influence from other fields such as MT, vision and ML, in general, may carry over to the field of ASR, as all of these fields tend to share methods and models in an increasing manner (e. g., the adoption of CNNs from vision to ASR).
- Novel manners to define the units of textual elements for vocabulary design emerge, to mitigate out-of-vocabulary (OOV)[5] effects e. g., via byte-pair-encoding (BPE) and the inclusion of single characters (Sennrich et al., 2016).
- Hyper-parameter tuning may receive an increase in interest, which may currently be low due to prohibitive costs for many participants in the ASR market.
- Further advances in search algorithms may emerge as methods like beam-searching do not guarantee optimal results.

Data use vs. other resources

The scarcity of training data (aligned data of audio and text) is a well-known problem for most languages. Whereas for commercially important languages such as English or Mandarin Chinese an abundance of data is available, this is not the case for many other languages. While companies like Google train models on 125.000h of speech, with a resultant model size of up to 87GB, this is unthinkable due to lack of data as well as resources for most other actors.

Several trends can be observed:

- The increased use of pre-trained models and fine-tuning/adaptation. Several platforms (like Huggingface[6]) provide a growing set of pre-trained models for a variety of languages and domains.
- Work on data augmentation and pooling of resources is receiving more attention. For example, there is some ongoing work in evaluating the best data augmentation and pooling methods, and their effect on ASR performance. This has been done extensively for Maltese speech data, where only around 7 hours of high-quality transcribed speech data is available (which is arguably low even for fine-tuning a system such as Wav2Vec). This has been documented extensively and could serve as a guide to other similar efforts for other languages (Mena et al., 2021). In fact, an absolute word error rate reduction of 15% is reported, just through careful augmentation alone – and without the help of a language model.
- Co-training of models: the combination of training data for several, related languages and domains to create multi-language, multi-domain models (or base-models for fine-tuning).

[5] OOVs are words that occur in the audio but which do not form part of the vocabulary
[6] https://huggingface.co

- Multilingual models: the creation of truly multi-lingual or language-agnostic models.
- Various methods to address the out-of-vocabulary problem. Words may thus be decomposed into smaller units (e. g., morphs), they may be reconstructed from intermediate search results (by extending lattices) or re-training of models may be carried out to include current vocabulary.
- Ageing of models: models are frequently outdated once they are deployed and need to be re-trained continuously. Through this, the shift in language may also be addressed (in addition to shifts in topics).
- Weakly- and semi-supervised training: There is also a strong interest in weakly- or semi-supervised training methods, that enable the application of and un-transcribed and un-annotated data for ASR training. In semi-supervised training, a series of models are trained where a given model in the series serves as a teacher to the succeeding model by generating labels on the unlabelled dataset. The student to this teacher model is trained on the dataset obtained by combining the supervised set with the teacher-labelled dataset. This idea has been shown to work and provide good improvements in recognition quality in multiple research papers, both in low-resource and high-resource scenarios (Wallington et al., 2021; Zhang et al., 2020; Synnaeve et al., 2019).

Accuracies, measures used, human vs. auto evaluation

The main measures of determining the performance of ASR systems are Word Error Rate (WER), Character Error Rate (CER) and Phoneme Error Rate (PER). WER is derived from the Levenshtein distance, performing a dynamic alignment of reference and output at the word level. WER, though mostly standard, does not always correlate with the quality of speech recognition systems, as some word-level errors can be qualitatively more acceptable than others. WER does not take into account this qualitative difference and treats all words equal – which for many purposes, such as information retrieval, is clearly not optimal. CER is similar to WER but typically applies to languages using character-based scripts. PER looks at an error rate which consists of the number of all phoneme errors. Given the nature of some of the ASR architectures, frequently utilising a flavour of Connectionist Temporal Classification (CTC), this is an increasingly important performance metric. In terms of performance over n-best results or lattices, measures such as precision, recall and their harmonic mean, the F1-measure are commonplace.

Downstream task accuracy, efficiency, thresholds

Typically, ASR outputs unstructured and normalised text without any punctuation marks. This is not an issue in use-cases, where the user input is short and concise, e.g., when asking a question to a virtual assistant. However, when generating transcripts for longer speech, it is crucial to restoring punctuation marks to improve readability and provide structure to the transcript. Moreover, punctuation marks are often used in further downstream tasks such as NER, POS tagging and MT. ASR systems can introduce errors that a standard MT system has not seen during training and thus cannot handle. In such instances, the translation quality may suffer, even to an extent where the translations are effectively incomprehensible (Ruiz et al., 2019).

4.2 Speaker Recognition

SOTA of current methods and algorithms

As for language recognition, state-of-the-art SR systems use neural networks to extract a representation (usually referred to as an embedding) for the speaker in an utterance. The input to the network is usually given by features extracted from frames of 20-30ms, although there are also ongoing efforts to take the raw waveform as input to the network. Embeddings are then compared with a backend in order to decide whether they are from the same person or not. Typical neural network architectures for embedding extraction are TDNN, ResNet, LSTM and versions thereof. The standard choice of backend is Probabilistic Linear Discriminant Analysis (PLDA) which is a generative model. In the recent few years, using cosine similarity plus an affine transformation have proven to give a competitive performance, especially for audio with a 16kHz sampling rate. An advantage of generative backends however is, that scoring with different numbers of enrolment utterances becomes trivial. In addition to variations of the embedding extractor architecture, many recent research efforts have focused on the training objective. If the task at hand is verification, the most intuitive manner would be to train the extractor for this task. However, in practice, it often works better to train the extractor for classification. That is, for a training utterance the network should classify who among the speakers in the training set speaks in the utterance.

Accuracies, measures used, human vs. auto evaluation

The evaluation metric in SR depends on the task studied. The most common SR task in academic research is speaker verification. Due to the many evaluations in this task, there is a large consensus on which metrics to use. Two types of errors can occur: false accept (FA) to recognise the speaker in the enrolment- and test utterance as being the same when they are different, and false reject (FR) to recognise the speaker in the enrolment and test utterances as being different although they are identical. It is important to distinguish whether an evaluation metric is calibration sensitive or calibration insensitive. Put simply, calibration sensitive metrics care about whether the decision threshold is correctly specified whereas calibration insensitive metrics do not. The most common metrics are Equal error rate (EER), detection cost function (DCF) and log-likelihood ratio cost (C_{LLR}). The EER is defined as the error rate when the threshold is adjusted so that FA and FR are equal. Thus this metric disregards for the threshold used by the system and accordingly is a calibration insensitive metric. The detection cost function is based on user-specified costs of FA and FR as well as the prior probability for the speakers in the enrol and test utterances being identical. For the detection cost function, there is both a calibration sensitive (actual DCF) for which the system's threshold is used and a calibration insensitive variant (minimumDCF) for which the threshold that minimised the DCF on the test set is used. Finally, C_{LLR}, is designed to be an application-independent evaluation metric. It can be viewed as an average of DCFs. The performance of speaker verification systems varies greatly in particular depending on the duration of the utterance but also on the acoustic conditions such as the noise level and sampling rate. Mismatch in the languages spoken in the training data and test data may also degrade the performance. For 16kHz data with low noise conditions, a few seconds of speech is usually sufficient to produce an equal error rate of around 1%. For 8kHz telephone data in noisy environments and with a mismatch in training and testing languages, EER can be 5-10% or even worse.

4.3 Language Identification

State-of-the-art LID systems are, similarly to SR systems, based on DNNs (e. g., TDNN or ResNet) that ingest sequences of frame-level speech features as input, after some processing apply a pooling mechanism to these features to obtain an utterance level representation, and then finally try to classify the utterance level representations. In training, this whole chain is trained in an E2E fashion. In testing, either the trained DNN is used directly for classification or, the utterance level representations can be extracted and used in a simple backend for classification, e.g., a Gaussian linear classifier.

Several trends can be identified and be expected to continue:

- The use of fine-tuning on ASR models such as Wav2Vec to extract embeddings for LID (this also applies to SID/Emotion-ID).
- The performance of these systems is rapidly outperforming more specific methods e. g., the i-Vector approaches to LID/SID have now been mostly superseded. More recently, reports show the same tendency for emotion detection.
- Parallel attempts are still very much being given importance – e. g., i-Vector methods with SVMs, LSTM-DNN, attention-based and ResNet-based classifiers.
- Several similar categories e. g., SID, LID, Accent-ID, emotion detection and other speaker-profile type classifications fall under the same family of techniques.
- Access to corpora is generally not problematic, given the language-independent nature of the developed algorithms.
- The performance (standard metrics such as F1 scores) correlates with the length of utterance under test. The shorter the utterance, the more difficult the task.
- As in other sub-fields of speech processing challenges like the VoxCeleb Speaker Recognition Challenge 2021[7] and the Short-duration Speaker Verification 2021 (SdSV)[8] have been propagated to boost the development of technologies.
- The task of Speaker Diarisation – segmenting audio containing multiple speakers (and/or speaking conditions) is closely related to this field. Performance measures for diarisation include the Diarisation Error Rate (DER) and the Jaccard Error Rate (JER).

4.4 Assessment of emotions, cognitive conditions and personality traits

Due to the wide scope and different disciplines contributing to the field, no single state-of-the-art can easily be described which would address the complete field. We can therefore provide an overview of selected aspects only.

Evaluation typically takes place in a qualitative manner (i. e., by human-raters and inter-rater-agreement) and with datasets, which are specific to the particular task. Efforts such as the Computational Paralinguistics Challenge[9] aim to introduce further tasks on a yearly basis.

[7] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/competition2021.html
[8] https://sdsvc.github.io
[9] http://www.compare.openaudio.eu

Speech Emotion Recognition

In Speech Emotion Recognition (SER), a wide range of methods have been used to extract emotions from signals. Similar to other ST fields, Deep Learning is rapidly becoming a method of choice and several E2E models have recently been proposed (Tang et al., 2018; Kumar et al., 2021). However, in the SER field, unlike in ASR, despite the fact that there are SER systems and realisations of real-time emotion recognition, these have not yet become part of our everyday lives. To achieve this goal, SER systems require more accurately labelled data to improve training accuracy, more powerful hardware to speed up processing, and more powerful algorithms to improve the recognition rates. In addition, further insights from fields such as psychology or neurology may be required.

Examples of the SOTA methods applied for the detection of the emotion features to recognise emotion speech include the application of two CNN and LSTM networks to learn local and global emotion-related features from speech and log-mel spectrogram respectively (Zhao et al., 2019). The results demonstrated that the combination of networks achieve excellent performance on the task of recognising speech emotion, outperforming traditional approaches, such as DBN and CNN.

In another SOTA approach, a dual-level model that predicts emotions based on both MFCC features and mel-spectrograms produced from raw audio signals was explored. In this approach, each utterance was preprocessed into MFCC features and two mel-spectrograms at different time-frequency resolutions. A standard LSTM was applied to process the MFCC features, while a novel LSTM architecture, denoted as Dual-Sequence LSTM (DS-LSTM), processed the two mel-spectrograms simultaneously. The proposed model surpassed the state-of-the-art (2019) unimodal models (Wang et al., 2020).

A different line of work, motivated by the challenges in the development of robust SER systems related to the scarcity of emotion datasets, a multi-task learning framework that uses auxiliary tasks for which data is abundantly available was proposed by (Latif et al., 2020). The approach explored the benefits of the use of additional data to improve the primary task of SER for which only limited labelled data was available. Specifically, gender identifications and speaker recognition were targeted as auxiliary tasks, which allowed the use of very large datasets. To maximise the benefit of multi-task learning, Adversarial Autoencoders (AAE) were used within the framework along with the unsupervised AAE in combination with the supervised classification networks. The proposed semi-supervised learning helped to improve the generalisation of the framework and led to improvements in SER performance, demonstrated for categorical and dimensional emotion recognition as well as cross-corpus scenarios.

With the growing popularity of ambient intelligence technology that uses a variety of low-power, resource-constrained devices, the development of methods that effectively use computational resources has gained the increasing interest of the research community. Among others, these include applications in health and elderly care technologies, where interventions can be triggered by the detection of emotional states. Examples of recent, SOTA approaches to SER in such settings include Haider et al. (2021). The study demonstrated that similar or better accuracy could be achieved with subsets of features substantially smaller than the entire feature set.

Another recent work in this line of research describes a lightweight SER model using a CNN approach to learn the deep frequency features by using a plain rectangular filter with a modified pooling strategy (Anvarjon et al., 2020). The proposed model outperformed the state-of-the-art while lowering the computational costs.

Developments in IoT and edge computing have also motivated research in which the compact speech recognition network with spatio-temporal features for edge computing, EdgeRNN was described (Yang et al., 2020). It uses CNN to process the overall spatial information, RNN to process the temporal information and a simplified attention mechanism to enhance

the portion of the network that contributes to the final identification. On Raspberry Pi 3B+ (notably, a small low resource computer), the method improved both, speech emotion and keywords recognition.

Cognitive disorders, health conditions and personality traits

Detecting the cognitive states and reactions of a user is a step towards designing proactive systems capable of adapting to the user's needs, preferences and abilities. As other related ST-fields, the detection of personality traits, mood disorders, signs of depression and other medical conditions have found their application in recent years.

One of the first signs of neurodegenerative disorders is deterioration in language and speech production. In recent years, techniques based on automatic processing of the voice signal have been used for language and cognitive assessments. These approaches provide the means for quantifying signal properties relevant for the detection of specific pathologies. Due to the development of automatic methods facilitating the evolving control of a wide population suffering from AD, a number of industry applications aimed at the detection of neurodegenerative disorders, developed by companies such as IBM Watson[10], Cantab – Cambridge Cognition[11] or Winterlights Lab[12] were introduced.

The SOTA approaches applied in the AD detection from speech include methods that combine the automatically extracted acoustic markers from spontaneous speech with semantic linguistic features. In the task focused on the detection of subjects from patients and the healthy control group, and in distinguishing AD patients from those with mild cognitive impairment, the accuracy of the presented approach was in a range of 80-86%, and corresponding F1 values between 78-86% (Gosztolya et al., 2019). The detailed presentation of the SOTA in these and relevant subfields extends beyond the scope of this report (for the review of the recent works focused on the detection of AD from speech, see (Pulido et al., 2020; de la Fuente Garcia et al., 2020)).

Further SOTA works in the relevant subfields include the detection of mild-cognitive impairments (Tóth et al., 2018), assessment of cognitive impairment in elderly people (Konig et al., 2018; Schuller et al., 2021), the application of representation transfer learning from deep E2E speech recognition networks for the detection of speaker intoxication (Sertolli et al., 2021), detection of COVID-19 from speech signal using bio-inspired based cepstral features (Dash et al., 2021), mood disorders (Huang et al., 2019), signs of depression (Toto et al., 2021) and the detection of personality traits from speech (Guidi et al., 2019; Lee et al., 2021).

4.5 Text to Speech

General introduction

Neural networks have greatly impacted the speech synthesis field by improving the quality and naturalness of synthetic voices with respect to the traditional systems. Another contribution made by neural networks is the possibility of training and designing the systems in an E2E fashion. While traditional multi-stage pipelines are complex and require extensive domain expertise, E2E systems reduce the complexity by extracting the audio directly from the input text without requiring separated models. Although E2E TTS systems have shown excellent results in terms of audio quality and naturalness, there are still some issues to be faced. On the one hand, these systems usually suffer from low training efficiency, requiring a large set of audio recordings together with the corresponding text to train properly. On

[10] https://www.ibm.com/blogs/research/2020/10/ai-predict-alzheimers/, accessed 17.1.2022
[11] https://www.cambridgecognition.com/news/entry/speech-recognition-to-improve-clinical-trial-efficiency, accessed 17.1.2022
[12] https://www.veritone.com/press-releases/voice-analysis-detects-alzheimers-disease/, accessed 17.1.2022

the other hand, synthesised speech is usually not robust, due to alignment failures between input text and speech during the generation.

SOTA of the current methods and algorithms

Lately, the most favoured approach to speech synthesis systems is to substitute the whole chain in the TTS system with DNNs (Ning et al., 2019). Deep Voice (Arık et al., 2017) was the first system where all the steps in the TTS process were implemented by means of DNNs. The quality of the generated voices was inferior to that obtained with WaveNet (van den Oord et al., 2016), so several improvements were proposed, such as Deep Voice 2 (Gibiansky et al., 2017) and 3 (Ping et al., 2018b), where WaveNet could be used as a neural vocoder to analyse and synthesise the acoustic signal. Another approach that can be considered more E2E is Char2Wav (Sotelo et al., 2017), although it still concatenates two modules: the first predict acoustic parameters from text and the second, a neural vocoder, generates a waveform from these parameters. Full E2E architectures have also been proposed, including Tacotron (Wang et al., 2017), Tacotron2 (Shen et al., 2018), FastSpeech (Ren et al., 2019a), FastSpeech 2 (Ren et al., 2020) and ClariNet (Ping et al., 2018a). These systems are able to produce spectrograms from text, applying an encoder-decoder architecture that produces a latent representation of the input text (or phonetic transcription) that is subsequently transformed via convolutional neural networks associated with attention mechanisms into spectrograms, which are then converted into speech using the Griffin-Lim algorithm (Griffin and Lim, 1984), WaveNet or other neural vocoders such as WaveGlow (Prenger et al., 2019), WaveRNN (Kalchbrenner et al., 2018) and MelGAN (Kumar et al., 2019). The systems provide outstanding results in terms of the quality of the generated voices but require large amounts of high-quality recordings to be trained properly. Currently, efforts are being made to deploy these systems for low-resource languages by improving data efficiency (Chung et al., 2019), applying transfer learning (Chen et al., 2019) or training multilingual models (Zhang et al., 2019c). Other areas of intense research activity are style transfer (Zhang et al., 2019a; Li et al., 2021), new efficient neural vocoders (chun Hsu and yi Lee, 2020; Paul et al., 2020; Jang et al., 2021) and speaker adaptation with a reduced amount of data (Xin et al., 2021; Maniati et al., 2021).

Regarding expressive speech synthesis, Global Style Tokens (Wang et al., 2018) can be named as one of the most common approaches. It consists of a reference encoder, which encodes the reference speech Mel-spectrogram, and a style token layer, which learns different prosodic aspects in a set of trainable embeddings. The reference embedding is compared with each style token with the help of a sequence-to-sequence multi-head attention module, forming a weighted sum of the style tokens called style embedding. The style embedding is then concatenated to the text encoder output, thus conditioning the Mel-spectrogram synthesis on both text and encoded prosody of the speech. Other methods include Flowtron (Valle et al., 2020b), which uses a generative flow-based model for learning invertible transformations from data to a controlled latent space that can be sampled during inference to achieve the desired prosody. Mellotron (Valle et al., 2020a), Fastspeech 2 (Ren et al., 2020), and CtrlP (Mohan et al., 2021) control prosody by concatenating the text encoder output with more traditional acoustic features, such as F0 contour or energy.

Data use vs. other resources

Developing high-quality synthetic voices with DNN based techniques requires large amounts of good quality recordings from one single speaker. This requirement is often difficult to fulfil, especially for minority languages and dialectal speech. The generation of new synthetic voices is also hindered by this extensive data requirement. Efforts are being made to share data among languages and speakers in order to train the common aspects more robustly.

Multi-speaker and multi-language modelling is a usual strategy in DNN based TTS synthesis to achieve improved voice quality with a reduced amount of data from a single speaker (Yang et al., 2021; Casanova et al., 2021; Shang et al., 2021). The quality of these voices however is not yet comparable to the one obtained with large databases.

Accuracies, measures used, human vs. auto evaluation

The most popular measure of quality in TTS is the Mean Opinion Score (MOS) where people express their opinion about several aspects of the synthetic utterances on a 1 to 5 scale (Goldstein, 1995; Rec, 2006). Considerable time and effort must be devoted to the development of this subjective evaluation, as a large number of individuals is needed to reliably rate the TTS systems. Although MOS tests are still the most frequently used option to assess TTS, they have been criticised as they offer only a general measure of the overall quality and may not be suitable for evaluating long synthetic speech passages (Clark et al., 2019b; Wagner et al., 2019). Moreover, they are often produced using too few evaluators to be reliable (Wester et al., 2015). Other TTS performance measures focus on intelligibility. The main strategy for evaluating this aspect is to ask people to transcribe semantically unpredictable sentences and measure the WER of the transcriptions (Benoît et al., 1996). As this evaluation also calls for the participation of human evaluators and this is a time-consuming process, ASR is increasingly being used to evaluate intelligibility Taylor and Richmond (2021). New dimensions to be evaluated are also arising such as measuring listening effort while listening to a synthetic speech by means of pupilometry (Simantiraki et al., 2018) and electroencephalography and measuring cognitive load related to the process of listening to this kind of speech (Govender and King, 2018).

(Botinhao and King, 2021) propose a method for automatic error detection and analysis based on the attention alignment between the encoder and decoder. The attention alignment for a correctly synthesised speech should be uninterrupted and monotonic. Any deviations or artefacts in the alignment can indicate that the model failed to correctly synthesise the audio.

5 Speech Technologies: Main Gaps

5.1 Background and overview of the main challenges

While speech technologies have found their way into a series of application fields, several important issues have not been addressed thoroughly and remain active areas of research. In the following, we overview the main gaps in ST and present them in a wider context of the global and regional business activities, requirements related to the availability of qualified personnel, privacy and trust concerns, as well as technical and end-user perspectives.

Effects of scale

Beyond the progress made within academic institutions, such as the University of Toronto, the University of Cambridge, Johns Hopkins, RWTH and many more, much of the advances made during the past decade has been driven by the research labs of companies such as Google, Facebook, Apple, Amazon, and their Chinese counterparts. Understandably, the driving factors behind the activities of these companies is to generate business – and not to perform fundamental research. Hence, advances are motivated by a commercial perspective and thus some of them are not shared as they provide market advantages over the competition.

The shift of actors coincides with the rise of massive progress in ML, based on access to huge, and previously unthinkable, amounts of data and processing power. It is no surprise that the companies with the largest pools of data and the most extensive infrastructure are now the leading actors in their respective fields, leaving only niche markets and domains to smaller, but highly specialised players. These niches, of course, may also provide ample opportunities for success if targeted properly.

As outlined above, a trend towards increasingly complex E2E systems can be observed for all sectors of ST. Due to the extreme demand on resources such as data, computing power, energy, infrastructure, the generic construction of such models is in many cases limited to a handful of actors. The activities to make pre-trained models available for transfer learning and fine-tuning settings and thus to allow others to also participate from major advances are certainly beneficial. However, the extent of this transfer and the level of control in the hands of a few institutions poses a serious risk to other actors, to the market and potentially even to innovation in the ASR sector as a whole. Moreover, commercial interest may prompt institutions to not make their best-performing models available but rather only more limited, smaller versions of these models.

Further issues relate to the interest and capabilities of European entities vis-a-vis to dedicate the resources required for developing state-of-the-art ST systems. When compared to the resources allocated by the GAFA and their Chinese counterparts, (for example Google allegedly has more than 250 people working on ASR alone, training models on more than 125.000h of speech, with resultant model-sizes of up to 87GB) the resources available to European companies and institutions are very limited. A similar situation exists on the hardware side: companies like NVIDIA dominate the GPU market, and team up with Microsoft to train the largest language models, while there is no realistic European counterpart insight.

Lacking the necessary funding environment (venture capital as well as mindset) a European strategy cannot compete on the same terms but rather has to investigate and follow innovative paths that require fewer resources. This may also be promising with regard to sustainability goals.

Trained personnel and expertise

A further gap, concerning all areas of speech processing, can be identified in the scarcity of trained personnel and expertise as well as the risk of losing emerging talent to innovative power-players outside of Europe (with possibilities and salaries which can generally not be matched by European players). Even in light of the democratisation of technology and auto-ML, allowing a much broader audience to create models and deploy these for use, respective educational programs in speech (and NLP) technologies form the foundation for future European success in these areas and may hinder it if not appropriately established and strengthened.

Privacy and trust

Data leaks and scandals in recent years have spurred the interest on part of individuals as well as of policy-makers. Concerns have arisen regarding trust, privacy, intrusion, eavesdropping, or the hidden collection and use of data. These concerns have been recognised by many actors but are only addressed to a limited amount (clearly so, as long as they counteract commercial interests). Further work and investigation into these topics may be beneficial commercially, academically as well as for policy-making. In addition, processing for ST in commercial contexts often relies on cloud-based infrastructures with few (if any) guarantees regarding how data stored in the cloud is eventually used or will be used in the future by service providers.

Technical perspectives

On a technical level, the focus in the ASR subfield on rather constrained conditions has left gaps in more diverse settings such as: distant speech recognition instead of single microphones; noisy environments; accented speech, non-native speech, dialectal speech and sociolinguistic factors affecting speech; spontaneous, unplanned speech; emotional speech (including speech during stressful or dangerous situations) and connected aspects concerning sentiments expressed (empathy); the integration of speech technologies into collaborative environments, multiple, simultaneous speakers engaged in discussions; as well as the integration of paralinguistic aspects and technologies addressing them. All of these issues warrant future attention and research.

Modern TTS systems can produce high-quality speech provided they are trained with a sufficient amount of good quality data. However, they are usually prepared to synthesise isolated sentences as they are built using only this kind of recordings. Therefore, when trying to synthesise paragraphs, speech for dialogue or audiobooks, the generated speech is much less expressive and natural (Cambre et al., 2020). This issue limits the practical application of E2E TTS systems and their adoption by the public.

While most research focuses on a single user's interactions, speech technologies embodied in virtual assistants are becoming increasingly popular in social spaces. This highlights a gap in our understanding of the opportunities and constraints unique to multiple user scenarios. These include detecting if users address the system or other participants, speaker diarization (see Park et al. (2022) for a review of recent advances in speaker diarization with deep learning methods), understanding aspects of social dynamics, and finding interaction barriers are some of the factors that restrict the usefulness of voice interfaces in group settings. The connection to the field of digital humanities and computational social sciences is not yet firmly established but it could be beneficial to set up collaborative links with a range of disciplines and domains working with spoken data in the domain of social sciences and humanities (SSH). In particular, the insights and requirements stemming from the needs for transcription workflows and audio mining tools of communities producing and (re)using oral history data and interview recordings may help identify gaps in language resources for model training and domain adaptation (Draxler et al., 2020). Integration with methodology for the automated annotation of spoken interview data with paralinguistic features is gaining attention (see Akbari et al. (2021) for the role of silence), and can widen the basis for use cases in multidisciplinary setting, including the study of mental health conditions and therapeutic interventions (Catala et al., 2020). It could be beneficial to identify any unbalance in language-specific support for the recognition, annotation and retrieval of the types of structured conversational speech that are used in interview settings, both in SSH and beyond. Expertise from the humanities can also provide relevant insights for addressing the challenges in the digital archiving of interview data (F Pessanha and Salah, 2022).

The increase in modelling power and performance achieved over the last years also comes with some drawbacks and challenges. These include a need for even more data of aligned text/audio pairs, respectively a lack of interest and work on the creation of new paradigms using fewer data. Current approaches include shallow and deep fusion, but the question of how to optimally combine LMs and DNN structures has still not been addressed comprehensively. Models requiring the complete input sequence for processing do not match well with requirements to perform causal processing. Several attempts to enable causal processing are being explored, among them the use of neural transducers running processing at regular intervals. The extent of context may also incur additional processing costs which need to be balanced and mitigated.

Models are not transparent and thus hard to interpret. This is partly due to the fact that previously individual components have been combined into single models. The complex process of hyper-parameter tuning is often too resource-intensive and thus has not been

addressed in many instances. Elements of input/output like byte-pair-encodings (BPE) have been suggested but these contradict the idea of genuine E2E processing as this decision is taken beforehand and outside of the model itself.

Integration of several components into one model prompts the question of whether further downstream technologies, relying on ASR output to perform various NLP tasks will also become part of such integrated models. The combination in turn raises questions about the interpretability and transparency of such black box systems as well as concerning the modalities for the integration of further knowledge sources.

End-users perspective

Overall, speech technologies have made a leap in getting adopted in many commercial settings, with easy accessibility of technologies and powerful models for commercially attractive languages. Especially the proliferation of intelligent Voice Assistants (VAs) has made speech a common mode of interaction for a wide range of users. While providing several useful features, issues limiting the further adoption and widespread use of speech technologies have been identified. Concerning the users' perspective, among others, these include problems in accurately recognising accented speech (Cowan et al., 2017), a lack of trust in VAs to execute more complex or socially sensitive tasks (Porcheron et al., 2018), and concerns related to privacy as well as (clandestine) data collection and its use (Clark et al., 2019a; Ammari et al., 2019). This issue is further exacerbated by the fact that systems often operate "in the cloud" rather than on-premise.

Many VAs may now be utilised in languages other than English, but coverage and supported functionality vary greatly. The gaps in the support of different languages create barriers for users whose primary language is not fully supported, or supported only to a limited extent, forcing them to communicate in a non-native language or risk being excluded from using the ever more popular systems and services based on speech technologies. Thereby, non-native users are pushed to develop different strategies and modes of interaction, including a reduced level of language production in interaction and more frequent use of visual feedback (Wu et al., 2020).

5.2 Data: alignment, labelling, anonymisation, diversity

As outlined above, the main challenge/gap related to data concerns its availability – of adequate datasets for low-resource languages, of an appropriate amount and quality. Various efforts aim to mitigate this fact by focusing on transfer learning and fine-tuning of models. However, whereas this approach is certainly beneficial, it generally does not yield models of equal performance (as for languages exhibiting large amounts of training data). For a few, commercially highly interesting languages, an abundance of training data (corpora with aligned audio and transcripts) is available. However, for many (the majority) of languages, this is not the case and only corpora which are minuscule in comparison to English are available. Not only does this lack of aligned data mean that the resulting performance of ST will be substantially worse than for English, it effectively excludes certain approaches from being applied – as these depend exactly on the availability of large amounts of training data.

With regard to the textual contents required (e. g., for LM training), the situation is more balanced. However, certain languages and dialects do not have one defined way of spelling nor adequate amounts of textual data due to low levels of general digitalisation. In addition, certain markets are dominated by individual players with control over the resources required by potential competitors to build models. This strategy to protect one's own market further hinders progress and development for specific regions and languages. As a consequence and due to the high cost of voice data collection and labelling, current voice interac-

tion technologies have a strong bias in favour of languages with a wider user base (such as English), thus potentially excluding many users.

Compared to ASR, obtaining training data for speaker recognition and language identification display different challenges. In the case of SID and LID, the situation is more favourable since the only annotation needed is the identity of the speaker or language. On the other hand, in the case of SID this annotation cannot be created by simply listening to the utterance because humans are not good enough to recognise speakers by their voice. Further, it is crucial that the training data for SID and LID contain many recordings of the same speaker or of the same language, whereas for ASR training data it is preferable to have as many (different) speakers as possible. While there has been much progress in collecting data from videos on the internet, progress on telephony data is still limited by lack of data, in particular for less common languages.

Activities and literature regarding the detection of emotions from audio in less-resourced language are very limited. For example, neither datasets nor well-recognised research on the topic exists for the Latvian language.

Although there are some public databases available to train DNN based TTS systems, these are in general only useful for building monolingual neutral voices in a reduced number of major languages (Park and Mulc, 2019; Zen et al., 2019). The availability of open data free of restrictions such as copyright and limitations due to GDPR regulations in the remaining major languages and all minority languages would allow the development of TTS systems for these languages too. In addition, databases with more expressive and spontaneous recordings are needed to be able to build TTS systems suitable for more emotion-demanding applications like audiobook reading, movie dubbing and human-computer interaction that aims to be similar to interactions between humans. Moreover, the vast majority of datasets correspond to adult voices and there is a lack of data to generate child and elderly voices. Taking into account that the voice is an important component of our identity, more diverse datasets are needed in order to generate personalised voices that can suit any user.

The diversity of contexts and speakers represented by popular ASR benchmarks like Librispeech (Panayotov et al., 2015; Garnerin et al., 2021) (read speech), and Switchboard (Godfrey et al., 1992) (spontaneous speech) is limited. Recent works attempt to address this problem by introducing benchmarks that mimic real-world settings, with the goal of detecting model biases and flaws (Riviere et al., 2021). The results obtained on this set show that while contemporary models do not appear to have a gender bias, they often reveal significant performance differences by accent, and much greater differences depending on the socio-economic background of the speakers. When tested on conversational speech, all models exhibit a significant performance drop, and even a language model trained on a dataset as large as Common Crawl does not appear to have a significant positive effect, highlighting the importance of developing conversational language models. Other recent works in this area discuss the next generation of ASR benchmarks and frameworks designed to describe interactions between linguistic variation and ASR performance metrics (Aksënova et al., 2021). Among others, Apple and Google utilise distributed and anonymised learning e. g., privacy-oriented federated learning. For example, in the methodology applied by Google, voice queries are kept for a limited period of time for continuous semi-supervised learning. An assistant query like 'What is the tallest building in the world?' returns a reply and links to a Wiki article. If a user clicks on the article, it is an indication that the question was understood correctly. A re-query means that the ASR system was wrong. These soft labels are used for further training, spanning more voices/accents and a wider array of contexts.

5.3 Accuracy: reaching usable thresholds for applications

The single most frequently mentioned hindering factor for the broad adoption of speech technology is one that has been mentioned for the past 40 years: accuracy. The perceived accuracy and its exact meaning have changed dramatically – from individual words being mis-recognised to intentions that are not correctly interpreted in complex situations, with accuracy reaching well beyond the actual accuracy of ASR only, regarding it in a more comprehensive and embedded manner. Whereas WER as an evaluation measure has had its merits to measure progress in ASR (and still does so), more comprehensive approaches to measuring the impact of ASR performance on downstream tasks and actual deployments may require novel approaches. WER alone clearly does not provide the full picture when it comes to the perceived performance and usability of complete systems comprising several kinds of speech and language technologies.

WER still provides the standard measure for the evaluation of ASR systems. However, as noted above, it falls short of capturing certain qualitative aspects of language. Depending on the task and use of downstream technology, WERs may not have to be extremely low and still allow the application of ASR within a particular field (*it does not have to be perfect to make perfect sense*). Performing evaluations also beyond pure WERs may then be helpful in such instances.

Current applications of speech processing, especially including smart-home systems that make use of speech interfaces, are heavily biased towards major speech-technology enabled languages e. g., English, Mandarin. Inferior performance may render them less usable and less popular in Europe.

One issue that affects current TTS systems is the lack of robustness in the synthesised speech: some input sentences may lead to skipping or repeating words or to babbling, especially when the kind of sentences seen in the training is very different from the ones synthesised (He et al., 2019). This problem mostly occurs in attention-based systems, where the output frames are related to specific parts of the input sentence by means of an attention mechanism relating and aligning text and voice (Zhu et al., 2019; Ren et al., 2019b). In order to address this problem, different approaches have been taken, some of them focused on designing more robust attention mechanisms (He et al., 2019; Battenberg et al., 2020), others including alignment information at the input of the system Zhu et al. (2019). Some researchers have proposed to substitute the attention mechanism with networks that can predict the estimated duration of the input phonemes (Shen et al., 2020; Yu et al., 2020). However, the problem has not been solved completely yet and keeps hindering the practical application of TTS systems in many instances.

Speaker recognition technologies have already reached acceptable performance for many applications. In many situations, it may be acceptable if the system does not take any decision immediately when it is not confident enough. Such situations can then be treated accordingly in an application. For example, in dialogue systems, one could wait for the decision until more speech is available. However, this does not mean that there is no need or opportunity for further research. All applications of speaker recognition would benefit from better performance of the core system as well as better robustness to acoustic conditions, utterance duration and other variables that occur in speech data.

Likewise, regarding the expressiveness of TTS systems, ample room for improvement remains. Modelling prosody with the help of learned latent embeddings, such as Global Style Tokens, allows synthesising speech in a particular emotional style, which can be difficult to define by explicit acoustic features, such as F0, duration, and energy. However, these embeddings are often ineffective, entangled, and difficult to interpret. Efforts are made to improve embedding robustness and efficiency, for example, (Dai et al., 2021) propose adding a style embedding down-sampling and up-sampling layer, in order to reduce overfitting towards training data and force the model to focus on more general prosody features.

5.4 Dialectal speech and multilingual training

Most TTS systems produce speech in the main variety of languages. To date, little attention has been devoted to synthesising dialectal speech with the latest technology. Attempts to multilingual TTS have been made, using multilingual speakers if available (Maiti et al., 2020) and more commonly using monolingual datasets recorded by different speakers and then applying voice conversion to generate synthetic signals in several languages with the same voice (Zhang et al., 2019b; Nachmani and Wolf, 2019). The quality of the voices generated with these techniques is still worse than the one obtained using monolingual databases. To be truly multilingual the TTS system should also be able to cope with code switched text and although some efforts have been made in this regard (Cao et al., 2019; Zhou et al., 2020), there is still room for improvement.

Contrary to other ST such as ASR, a speaker recognition system can be used in languages different from the one that it was originally trained for. The performance of the system may however deteriorate in this case. Some progress has been made to make systems more language-independent for example by multilingual training or by adversarial adaptation. However, the effectiveness of this is not well understood for languages that are too different from the languages used in training.

5.5 Explainability and transparency for critical methods and technologies

While in the last decade ST research has made much progress in terms of performance of the systems as well as in applications of the technologies, progress in terms of understanding of the used architectures (why some architectures work better than others etc.) as well as the nature of the data and task (for example to understand to what extent it is possible to obtain domain invariant representations) has been much more limited. This is partly due to the fact that the neural networks used in modern systems are harder to understand than the generative models (GMMs, i-vectors, etc.) of the previous generation speaker recognition systems. But partly it is also due to a lack of interest from the industry and funding agencies to support that type of research. Students are also generally inclined to work on topics that mainly aim at improving performance since this increases their chances of obtaining a well-paid job in the industry after graduation. Historically, a good understanding of the methods has been crucial for technological breakthroughs though, e. g., for the transition into subspace-based methods for SR such as JFA (Joint Factor Analysis) and i-vectors. It is possible that a more *trial and error* based research methodology which is currently popular is indeed the most effective for the very complex models that are currently state of the art.

Technology adopters and end-users prompting for more insight into the capabilities of systems and the generation of results – potentially wanting to intervene in this process or influence it – may warrant further research. e. g., for ASR, this may concern the inventory of recognisable units, for LID the inventory of languages and language varieties which can be processed. In all cases, insights into how a particular result was reached may be beneficial for explanatory purposes.

6 Speech Technologies: Contribution to Digital Language Equality and Impact on Society

Purely technological systems alone do not exist – they are always embedded in a social context and should thus always rather be viewed as socio-technical systems. The applications of ST have diverse and multifaceted impacts on several key aspects for societies. Improved

technologies reaching performance levels resembling those of humans may in many aspects lead to a humanisation of technology, ascribing human attributes to system behaviour. Patterns of human-to-human interaction may be applied to human-machine interaction leading to heightened expectations and subsequent disillusion.

6.1 Digital language inequalities

The unbalanced availability and quality of ST resources, (e. g., data-sets, annotations, models) strongly impact the performance of ST for different groups of languages. This lack of parity in ST resources for different languages translates directly to digital language inequalities. For languages supported to a (much) lesser extent, performance and accuracy are typically significantly lower compared to resource-rich languages. In extreme cases, selected functionalities and/or support for minor languages may not be available at all. In addition to the support of a language per se, language varieties, dialects or accents may not be supported or only supported on very limited levels. ST are thus not accessible nor available to everyone on an equal level, i. e., functions, performance, robustness may be dramatically different from case to case.

While new advances in ST contribute to the reduction of this division between resource-rich and resource-poor languages, the lack of commercial interest in the long tail of *small languages*[13] translates to a significantly slower pace of ST improvements and commercial adoption for the latter group of languages. For native speakers of these languages, these imbalances lead to wider usage of the better-supported, major languages, such as English, French, German or Spanish.

Motivating speakers to use these major languages more frequently creates a new set of challenges related to handling accented and non-native speech. Compared to the level of service and the support provided for native speakers, this results in lower performance, weakened experience and reduced usability for this group of speakers, rendering ST less useful or even useless in the extreme case. For TTS, the limited performance may translate into synthesised speech which is not perceived as natural nor pleasant and consequently leads to lower acceptance and adoption.

In the longer term, children who are more exposed and flexible in terms of adoption of new technologies may end up speaking more of a foreign language or a mix of their native language(s) and a major language, causing issues of social friction with parents and relatives who may not possess the same command of that language or who may not be able to understand it at all. Generational issues may further arise by the fact that adults may be somewhat limited in their willingness, openness or capabilities to adopt new technologies, including ST, whereas younger generations may be much more flexible to the adoption and exploration of new types of voice-based interaction.

When it comes to peers, the widespread adoption of ST, including voice assistants may also influence how users might communicate with and address each other – potentially this may result in communication styles more similar to issuing commands to devices and smart-speakers at home.

Further, (children's) personification of voice assistants and smart devices may become a double-edged sword: it assists parents in encouraging young children to use the devices for knowledge seeking and learning, but it can also frustrate them as children can develop an attachment to the devices and come to rely on the technology to an extent that communication with parents and peers is reduced dramatically (Garg and Sengupta, 2020).

This raises the following questions:

- Will the commercially important languages stay ahead of the majority of languages also in the long run?

[13] Languages spoken by a relatively small number of speakers

- What effect will this have on speakers of such smaller (less frequently spoken) languages?
- Will a lack of commercial interest in such “small languages”, also translate to a lack of improvements and innovation in these languages?
- Will the imbalance between language support motivate speakers to use English (or another large language) more often as this might provide a better experience instead(but at the cost of non-native language use)?
- Will the digital footprint of minor languages be reduced to a minimum and eventually be marginalised?

6.2 Biases, fairness and ethical issues

The development, application and adoption of ST are also connected to a range of issues relating to fairness, biases, ethical and legal aspects that have to be accounted for and to be properly addressed to support adoption.

As technologies are entering the homes and offices of users on a broad scale, an enhanced level of attention to privacy concerns, ethics and policy is essential. Policymakers, policy watchdogs, the media and consumers alike need to assume the role of gatekeepers to the introduction of ST into many corners of society. Trust is viewed as the main currency and key to the adoption and acceptance of technologies as well as to the perception of market participants and their role in this process. Scandals and opaque behaviour on part of ST providers may have detrimental effects.

Voice assistants frequently utilise female voices. Some of them offer the possibility of using male voices, but the default voice is usually female. This fact has been extensively criticised as it can contribute to the outdated view of women as the gender that must help and take care of others. Moreover, nowadays the generation of gender-neutral voices is gaining importance, as many people do not identify themselves with the classic binary genders. More effort is needed in the development of modern TTS systems to include gender-sensitive practices and options for adaptation.

Similar to gender-related biases, race-related biases may be present in many kinds of ST models. Due to the fact that models depend on the amount and composition of training data, ethical aspects of language and language use present in these data may also be present in the resulting models. Systems capable of self-learning may adapt into directions completely unplanned and undesired by the developers or be gamed (attacked) by users into doing so[14]. Due to these inherent conditions, systems may subsequently perform at different levels of accuracy for particular sections of the population. Furthermore, disabilities related to language production may not be accounted for and exclude sections of the population from using ST systems at all.

Speaker recognition systems are usually less accurate for female voices than for males. This is not because women are underrepresented in the training data but more likely due to the properties of female and male voices. Various ethnic groups may however be underrepresented in the training data and thus less accurately recognised. It should also be noted that being in the group for which the system performs worse can be either an advantage or a disadvantage depending on the application and the type of error the system tends to commit

[14] After Microsoft’s release of its chatbot Tay in 2016, the chatbot began to post racist, sexually-charged, inflammatory and offensive tweets prompting Microsoft to shut down the service again within 16 hours of its launch (https://en.wikipedia.org/wiki/Tay_(bot)). Allegedly this behaviour was provoked by users gaming the service. This episode also prompt questions about proper evaluation and testing of such (self-learning) services before releasing them on a large scale.

more often (false positives or false negatives). Another ethical concern pertaining to SR are possible privacy breaches through mass surveillance of phone calls.

Current DNN based TTS systems have reached a quality level and a degree of similarity with the voice of real people that could be used to generate deepfake voices. Many of the possible applications of high-quality voices indistinguishable from those of humans are positive and people with speech disorders, visual impairment and other disabilities could greatly benefit from them. However, speech deepfakes could also be used as a tool for illegal activities such as committing fraud or discrediting people. New regulations and the development of ad-hoc legislation is critical to mitigating this pernicious effect of the TTS technology. Some new tools able to detect speech deepfakes must be produced, and anti-spoofing techniques that discriminate synthesised from natural speech must be developed in close collaboration with teams working in TTS.

6.3 Users with special needs

While state-of-the-art ASR systems achieve great accuracy on typical speech, they perform poorly on disordered speech and other atypical speech patterns. Personalisation of ASR models, a commonly applied solution to this problem, is usually performed on servers. That in turn poses problems related to data privacy, delayed model-update cycles, and communication cost for copying data and models between servers and mobile devices. While on-device personalisation of ASR recently showed promising, preliminary results in a home automation domain for users with disordered speech (Tomanek et al., 2021), more research is required to further increase the ASR performance for these groups of users and provide support for open conversations with longer phrases.

TTS technologies have a wide range of applications, some of them of great social impact. TTS is considered assistive technology and as such, it may contribute to the integration of people with visual impairments and learning disabilities like dyslexia. By developing robust systems capable of reading any text from any source including books, websites and social media, these people would be able to enjoy the same advantages as any person without a disability. It also facilitates equal access to education for people with visual and learning disabilities as well as for foreigners who may struggle with the language. This technology may help these students use computers in the classroom as the rest of their fellow students. In addition, it can contribute to the integration of immigrants by making it easier the learning of the local language as TTS allows listening to words and sentences when reading them. In this same line of applications, TTS can help people with literacy issues and pre-literate children learn to speak for the first time accessing any content presented in written form. Finally, TTS may prove helpful in times of ageing populations with degrading eyesight.

Another contribution of TTS to society relates to orally impaired people. Voice is an essential component of our identity that we usually take for granted. However, losing it can affect how others perceive us as well as our own sense of who we are. We communicate with other people mainly through our voices that help us make social connections. TTS technology is able to provide a voice for those who have lost their own. Synthetic voices can be personalised so they suit the characteristics desired by each user, by applying speaker adaptation techniques. Even generating synthetic voices that can reproduce the sound of the voice the person had before they lost it is possible, provided recordings are available. This way individuals can speak with synthetic voices that match their personality and character instead of using the standard voices provided by default by companies.

In another vein, ST can make our lives easier as it allows us to receive information while our eyes are engaged in other activities than reading. Thanks to this technology, we can access information on the go by means of sound or when we are involved in physical interaction, e. g., work, sports. Integrated with virtual assistants, TTS systems are able to provide

support to elderly people, assisting them with reminders of appointments and medication needs, providing them access to online information and improving both their ability to live by themselves and strengthen their autonomy. Eventually, this technology can also benefit any individual living alone by allowing them to have conversations and being a kind of social companion, helping to reduce loneliness.

Inclusion

Voice technologies and subsequent automation and multiplication of services could be beneficial for underrepresented minorities from an inclusion perspective. Parts of the population may not have access to smart devices or not be media-literate (see below). Technologies may not exist for particular languages or dialects or not function at the same level of performance. Language conveyed by means other than audio – sign languages – may be at a disadvantage and technically require different processing channels (visual processing). For speech output, powerful TTS technology ready to be used in many languages (any language) and equipped with efficient interfaces is imperative to achieve an inclusive society where everybody has equal access to information, education and communication.

In all of the above cases, situations may benefit from advances in ST and NLP technologies (such as mechanisms not requiring huge amounts of annotated training data) but equally, need to be considered on the policy and societal level.

Media-literacy

As human-computer interaction is being facilitated by the use of voice, a vast portion of online searches is already being performed via voice. The omnipresence of hand-held devices and smartphones paired with the presence of the Internet as part of daily routines has created an "information at your fingertips" world, where information is a mere search (type/click) or voice-command away. Will the increased use of voice technologies, in particular for search, accelerate this "don't need to know because I can query (typically google) it" attitude? Is the information so obtained reliable and can it be trusted? (because a presumably near-perfect technology produces it). And what effect will this have on media literacy in the mid- and long-term? A close watch needs to be kept on such effects on media literacy and more research directed towards these phenomena.

Politics and Democracy

It has been pointed out that language strongly influences the manner we think and argue about political issues and topics. Language causes mental frames to be activated and form our portfolio of ideas (Wehling, 2018). Politicians and influencers have long discovered these mechanisms and are applying them actively on a daily basis to push their respective agendas. Having this central and immediate effect on cognitive mechanisms, linguistic plurality also forms the basis of cognitive plurality and as such plays a fundamental role in securing diverse and democratic values. Limitation to a few individual languages – such as may happen due to limited digital support for certain languages – impoverishes and reduces this variety, the flexibility and spectrum for expression of thoughts and (political) ideas.

Regional differences

Speech technologies will probably exhibit the highest impact in the APAC region. This is largely due to population- and economic growth as well as the fact that character-based

scripts and keyboards are not an optimal combination for interaction. In addition, the penetration of smart devices is expected to increase even further in this region. On the other end of the growth spectrum, Africa can be identified, as the region with the lowest impact of ST. Will such developments broaden the digital divide even more? The design of ST typically reflects the social and economic background of the experts behind their creation and design. Potentially, this may result in a large gap between intended usability and actual adoption in the field concerning certain regions (e. g., parts of Africa and Asia).

6.4 Businesses and effects of scale

Speech technologies have a significant impact in the wider context of economical and business activities. Among others, these include the differences in the legal and financial frameworks in which the global technology companies operate, and to which they can flexibly adjust depending on the most suitable set of conditions offered in the different countries. These favours or even enable specific types of activities related to data collection, its processing and use, that can either be more difficult, cost-intense or not possible at all in the other regions. In this scope, the far-reaching consequences of the regional difference in the development of competitive business environments cannot be overlooked as they significantly impact society at large. By influencing the pace of economical, technological and societal development they create opportunities, effects of scale and influence the decisions of the individuals and enterprises about the regions in which they engage, invest and operate.

Bearing in mind the Matthew Effect (Rigney, 2010) the question remains if the current dominance of a handful of super-actors will increase even further in the future. And if so, if a certain kind of monopoly of speech technologies will ensue.

A further economical aspect concerns the impact of ST on automation and as a consequence on the job market as a whole. As technologies such as chatbots are being adopted in pursuit of efficiency, they also perform an increasing number of tasks previously reserved for humans. ST and AI thus blur the boundary between humans and technology leading to shifts in jobs and entire industries. Clearly, a message of cooperation and support rather than of rivalry and replacement needs to be communicated and acted upon.

6.5 Energy consumption and sustainability

The growing energy consumption required for the ever-expanding amount of data being processed and the tendency towards continuously more complex ST models has become evident since the race for the largest models has been going on. A trend towards increasingly complex E2E systems can be observed in many areas of AI, NLP and ST. Due to the extreme demand on resources (data, computing power, energy, infrastructure) the generic construction of such models, in many cases is now limited to a few actors. The motion to make pre-trained models available for transfer learning and fine-tuning thus allowing others to also participate from major advances is certainly beneficial. However, the extent of this transfer and the level of control in the hands of a few institutions poses a serious risk to other actors, to the market and potentially to innovation in the ST sector as a whole. Surging interest in sustainability and ethics may cause actors to reconsider the massive increase in energy consumption that currently accompanies progress in ST. An opportunity (and marketing advantage) may arise from directing efforts specifically towards the creation of high-performance/low energy-consumption ST – green ST.

6.6 Privacy, surveillance and trust

Whenever ST is linked to a person's identity and this link is used for access control or authorisation, the issue of trust becomes especially important. The main applications of Automatic Speaker Verification (ASV) are exactly the areas of access control, surveillance, forensics or voice assistants. ASV is used to authorise access to resources such as a bank account or building. In surveillance applications, it is used for detecting and identifying a wanted criminal in a collection of telephone recordings. In forensics, ASV is used for comparing a voice recording from a crime scene with the voice of a suspect or a victim. For voice assistants, speaker recognition can be essential to make sure that certain requests are fulfilled only if made by the owner of the respective device or commodity (e. g., computer, phone, house or vehicle). All of the above applications rely on high-performance and trusted ST and can benefit tremendously in commercial terms if applied within these contexts.

Another ST which is effective in intelligence and surveillance tasks is the identification of the language(s) spoken in an audio file or stream. Language ID is a prerequisite step in settings when downstream processing (e. g., ASR) is to be applied and models are available for particular languages only.

As pointed out, many STs require huge amounts of speech data to reach state-of-the-art performance. The standard today is to store audio (the voices of persons) in the cloud and label them manually. There are no guarantees regarding how data stored in the cloud is used or will be used in the future by cloud service providers (or whether it may leak).

This general approach raises critical privacy concerns and it has led to market and data concentration in the hands of a few, big corporations. Dramatic improvements in speech synthesis (Székely et al., 2019), voice cloning (Vestman et al., 2020) and speaker recognition (Snyder et al., 2018) pose severe privacy and security threats to the users. This resulted in a growth of interest in new voice privacy-preserving transformations and voice privacy evaluations (Srivastava et al., 2019, 2020; Ribaric et al., 2016; Qian et al., 2018). Recently the VoicePrivacy initiative was started to spearhead the effort to develop privacy preservation solutions for speech technology and create a new community (Tomashenko et al., 2020).

In the long run, the question will be whether any possible breaches, leaks or scandals involving ST will erode trust to a level that users will no longer volunteer to provide their data for training purposes (deep fakes may pose a particular risk). Of course, the distrust will be weighed against the commodity of using certain devices and platforms whose terms of use may simply require the user to do so.

A further area of concern is the extent of unlawful surveillance by governments, state agencies or (large) corporations, infringing citizens' rights, liberties, adversely affecting public discourse, democratic values and influencing the political powers (Stahl, 2016). The Snowden revelations sparked a global discussion about the general nature of mass surveillance and its consequences for state and corporate intelligence services. The concerns about the extent of privacy invasion, accountability of intelligence and security services, the (non-)conformity of mass surveillance activities with fundamental rights (Garrido, 2021), their effects on the social fabric of nations can only be considered and analysed jointly with the rapidly extending technological capacities, including ST, and the pervasiveness of devices able to capture, process and transmit relevant data. Regardless of the form of current government, the growing extent of mass surveillance and especially its unlawful application may lead to erosion of public trust in governments and state agencies (see Lora Anne and Laidler, 2021 (Westerlund et al., 2021) for a recent, in-depth presentation of theoretical and empirical relationships between transparency and trust in the context of surveillance). In addition, data leaks caused by such agencies may inadvertently lead to further and cascaded infringements and illegal use of data.

A very different kind of risk is posed by overly eager salespersons overselling ST dramatically and the following – inevitable – chasm into which users will fall in disappointment.

Proper responsibility and management of expectations need to be carried out in order to avoid this detrimental situation and a situation similar to the Winter which AI went through.

7 Speech Technologies: Main Breakthroughs Needed

The list of the main breakthroughs needed stems from the limitations identified in chapter 5, the recognition of a wider-reaching impact of Speech Technologies on society at large, and their contributions to Digital Language Equality.

At the technological level, these relate to accuracy, reaching acceptable thresholds for applications and in the creation of the datasets required for the continuous improvement of the core ST components. In the context of the DLE, an important challenge and breakthrough required relate to the resources available for the development of less common languages; improving the performance and extending the capacities of the ST components for these languages in parallel with the SOTA systems. The extended proliferation of ST, including to the areas with a high potential impact on individuals and large groups of users, also has to be considered in a wider context of policies governing ST and relevant fields and calls for major breakthroughs in terms of explainability for the critical methods and technologies. Policies and governance concerning the use of ST and data – in particular personal data – need to be kept up to date and on par with rapidly developing technologies and applications. In order to democratise voice technologies and to strengthen their position within LT and the even wider field of AI, the base of users – on all levels of expertise – should be widened. An increase in educational programs, including in general AI, ML, NLP, and inter-disciplinary activities, projects and programs are deemed beneficial for the generation of experts in these fields able to draw upon expertise in voice technologies but at the same time also in domain-specific fields thus forming the links between them.

7.1 Access to and discoverability of training data

To build a DNN based TTS system nowadays tens of hours of high-quality speech recordings must be at hand and considerable computing capacity is required. This severely limits the possibilities for small companies to compete and be able to develop their own custom voices. Optimising the architectures to make them less intensive from the computational point of view would allow for companies with limited resources to create their own TTS services and voices and boost the competence and competition in a field that is being more and more dominated by a few very big companies like Google, Amazon or Baidu.

For ASR, the same limitation regarding the availability of large amounts of annotated data applies. Only that in this case, the order of magnitude of training material is typically even higher. Whereas in the early 2000s, several dozens of hours of audio were regarded as a sufficient base for training AM for languages, this amount has rapidly increased to hundreds or (tens of) thousands of hours of annotated speech. The problem is aggravated by the fact that training data needs to be available in a particular language or dialect, as sharing of acoustic data between languages is often not deemed possible. Semi-supervised methods have allowed extending datasets, however, the amount of data available to industry giants exceeds that of common market players by orders of magnitude. Datasets for languages of lesser commercial interest are scarce and in some cases, individual players have achieved a quasi-monopoly on datasets for particular languages and domains.

For SID, the situation is slightly different in that the amount of training material may not be as much of a factor as for other speech technologies. Here, the availability of the right kind of data paired with mechanisms for effective and rapid model adaptation may be key. Privacy plays a particularly important role for this type of data.

A plethora of different licensing agreements and mechanisms pose further obstacles to access to datasets and resources. Simplification and harmonisation of these mechanisms would be highly beneficial.

While some of these issues fit into a larger theme of open data sharing and bringing digital technology to businesses, citizens and public administrations which are in the focus of, for example, the Digital Europe Programme – DIGITAL[15], it is important to consider in such frameworks the specific requirements and challenges related to the acquisition and use of the datasets typical for training and evaluation of ST and LT models.

7.2 New training paradigms

For approaches requiring large amounts of properly annotated data, strategies and frameworks for joint (potentially distributed) data collection, improved data annotation (potentially automated), as well as joint provision, may be needed. This not only concerns the collection itself but equally the storage and provision of such resources. A lack of data for particular domains and languages due to a lack of commercial interest needs to be countered by public efforts to jump-start and boost efforts in these languages and not to risk certain languages becoming effectively extinct in the digital realm.

From the perspective of data augmentation, the generation and use of synthetic data may provide a complementary alley in the creation or extension of datasets. Likewise, the application of methods modifying the audio signals themselves may provide a viable manner to extend datasets and make resulting models more robust.

Work on advancing algorithms and methods to require less data or to yield more robust models using smaller amounts of data, more effective use of transfer learning and fine-tuning likewise provide promising approaches to alleviate the lack-of-data dilemma. For specific fields of speech technologies, improved use of unlabelled data in an unsupervised or semi-supervised manner (pre-training, self-supervised training) may provide further possibilities (Lai et al., 2021).

Novel strategies like MMLM (Multilingual Masked Language Modelling) which have successfully been applied to learn multi-lingual (or cross-lingual) representations of language may provide further angles (Goyal et al., 2021). While some preliminary works exist, e. g., (van der Goot et al., 2021), extensive studies are required to assess and evaluate the extent to which such progress can be transferred and applied to voice technologies.

In addition, experiments indicate that MMLM may also aid in the cross-lingual transfer of deep representations due to the learned shared latent properties of language, linking this advance to the tendency of including broader and deeper context with speech technologies to arrive at applications combining technologies and allowing for many comprehensive applications and user experiences.

In the area of SID, the transfer of knowledge learned from languages with a lot of training data to model speakers of languages for which only little training data is available has not been examined thoroughly. The interaction between the (front-end) extractor and the back-end likewise need further research (E2E training, novel training objectives that encourage the embeddings to match the assumptions made in the back-end etc.).

For several technologies, making better use of the hierarchical structure and relatedness of languages may be beneficial. A system that has not seen data of a particular language (or dialect) in training should still be able to benefit from data from similar (close[16]) languages which may provide more data. Eventually, even with very limited training data for a particular language, it should then be possible to train a model using data from the specific language

[15] https://digital-strategy.ec.europa.eu/en/activities/digital-programme
[16] the appropriate definition of closeness may depend on the specific technology and application

as well as from these closely related languages. Methods like one-shot learning or few-shot learning likewise provide promising approaches.

To complement technological/algorithmic advances it may be beneficial to develop new schemes involving users more actively in the generation of datasets for training and evaluation purposes. Such approaches need to have safeguards implemented to prevent pollution and/or biasing (intentional or unintentional) and be transparent in the preparation and curation of the data. Furthermore, incentives for participation must be present and the scope of use of the resulting datasets be managed such that their use also benefits the general public (the European citizen) and not only a selected crowd of commercial actors.

7.3 Confluence and context information integration

Whereas previously the focus of activities was often placed on the advancement of individual technologies and specific capabilities, a tendency towards confluence – of the combination of technologies and inclusion of a larger context as well as the history of events and interactions – can be observed already to some extent and can also be assumed to play a more pronounced role in the future.

For example, the recently presented E2E model for speaker-attributed automatic speech recognition (SA-ASR) was proposed as a joint model of speaker counting, speech recognition and speaker identification for monaural overlapped speech (Kanda et al., 2020). It produced encouraging results for simulated speech mixtures consisting of various numbers of speakers. However, in order to conduct speaker identification, the model required prior knowledge of speaker profiles, which severely limited the model's applicability. The follow-up work addressed the issue where no speaker profile is available by performing speaker counting and clustering with the internal speaker representations of the E2E SA-ASR model to diarise the utterances of the speakers whose profiles were missing from the speaker inventory (Kanda et al., 2021a).

The increased presence of conversational interfaces, a proliferation of chatbots combining ASR, NLP and TTS with an ever-increasing presence of AI, in general, has modified not only the technical and commercial landscape but also the expectations of users when interacting with such systems have grown dramatically. Users tend to view such systems as a kind of digital assistant, a personal concierge more than a mere block of interconnected components (and really, as casual users, they should also not be concerned that in reality, this might be the case). This rising expectation and perceived user need have been accelerated by increased periods of home-office setups and virtual meetings which are likely to also continue in the future.

More powerful tools and greater capabilities also prompt the inclusion of upstream technologies such as summarisation or sentiment analysis to be integrated with voice technologies. Speech synthesis is bound to become as emotional and persuasive as the human voice itself. The automatic translation may be used within the loop to bridge language boundaries. Furthermore, technologies will need to be integrated in a manner allowing for feedback loops and adaptation in a seamless way. Models need to be dynamic and methods allowing for dynamic adaptation – learning and unlearning certain features – will need to be developed to account for flexible and continuously changing conditions.

The integration of technologies and the inclusion of a much broader type of context may allow capturing intentions and real user needs, creating an overall experience of real conversational AI-powered by speech and language technologies, fully interconnected with business applications and private data sources.

Thus a required step relates to the transfer of voice technology performance improvements into downstream technologies and then to improved overall user experience.

Subsequently, such setups may need to be interconnected between groups of persons, e. g.,

family, friends, and team-members to also include further context not limited to a single individual. Examples are business settings involving multiple speakers speaking different languages during meetings or group activities where cultural factors and background information of several persons involved will need to be taken into account.

Areas of linguistics such as pragmatics as well as paralinguistics will need to be considered and integrated to a much higher extent than currently to allow for more natural and human-like interaction. Adding emotions and affections into the recipes for human-machine interaction, recognising intent and taking into account a broad variety of contexts holds the potential to turn these interactions into truly human-like experiences. The components related to emotional understanding and empathy, while relevant to all Intelligent Personal Assistants (IPAs) and Conversational Agents (CAs), are especially relevant for systems functioning in social domains, such as healthcare, education, and customer service. Combining emotional awareness with CA technologies and approaches also necessitates incorporating insights from multiple domains, including psychology, artificial intelligence, human-computer interaction, sociology, educational research (Andre et al., 2004; Vinciarelli et al., 2011; Skowron et al., 2013; Zhou et al., 2018; Belainine et al., 2020; D'Mello and Graesser, 2012).

7.4 Explainability, transparency and privacy concerns

The above-outlined increase in the complexity and combination of technologies and models requires a careful balance with regard to privacy and ethics. Scandals and data-leaks like the one caused by Cambridge Analytica (Hu, 2020) or Facebook Files[17], the often intransparent manner of how personal data are handled by many companies and growing conscience of the value of personal data has lead to increased interest and level of anxiety across societies. Activities like the EU's GDPR regulation (Regulation, 2016) are aiming to pave the way towards a higher level of data sovereignty. Attitudes towards such motions are certainly different depending on region, culture, political system etc. but may be seen to play a more important role on a global scale in the future.

Trust in speech technologies and in the use of data obtained by interaction with these technologies may become a decisive factor in the adoption of technologies and of the success of individual market players. An increased interest in "what happens under the hood" and in providing more transparency of data use and system functionality can be observed across the board in many areas of ML and AI. This is certainly also true for ST and will become more pronounced if these technologies are to be coupled with other sources of data (as described above).

A fundamental question to be answered transparently by providers will be where exactly processing is performed and to what extent and purpose data is used to modify (retrain, adapt) models.

One end of the spectrum of processing is large, anonymous data-centres spread around the globe. The other end of the spectrum is formed by strictly local processing on personal devices (on the edge). Private, on-premise solutions provided by companies or institutions form an intermediate setting. In all of these setups, the balance between capabilities and the requirements to achieve these capabilities will need to be determined and balanced against ethical concerns and personal and privacy-preserving arguments. The extent and amount of end-user control and transparency on part of the application providers will be a crucial factor in this equation. Methods to allow for flexible and transparent ways to allow for such control may be promising areas not only for voice technologies and models. Approaches like privacy-by-design accompanied by high ethical and legal standards may be determining factors in enabling trust, fostering adoption and leading to economic success.

[17] https://www.washingtonpost.com/technology/2021/10/25/what-are-the-facebook-papers/, accessed 17.1.2022

7.5 Support for less-resourced languages

To be able to provide first-rate ST in any language, additional high-quality datasets are essential. Ideally, they should be open and available without usage rights limitations for all the languages and include recordings with a variety of conditions and representative settings. These include a variety of speakers, language varieties, dialects, sociolects, data including spontaneous speech, varied prosodic patterns, diverse sentence lengths and a wide range of emotions. Creating this wide set may not be feasible in general, but could be achieved at least for several major European languages. New techniques for transfer learning and model adaptation from systems trained for one resource-rich language to systems able to function in languages with more reduced quantities of available data should be developed. These techniques would allow the development of cutting-edge ST systems also for less-resourced languages. Also, new architectures allow using resources from several languages in such a way that commonalities among languages are learned in a more robust way by cross-lingual knowledge-sharing or methods for the creation of multilingual or language-agnostic models which can be applied to a number of different languages are of utmost importance.

7.6 Performance, robustness and evaluation paradigms

Driven by various national and international (e. g., DARPA-sponsored) evaluations standard performance measures have been defined and measured on standard test sets during concerted evaluations. Current measures like the standard WER only take certain performance aspects into account and may need to be reconsidered, resp. be extended or complemented. Robustness and generalizability of ST components and models as well as standard-evaluation sets for multiple languages and evaluation sets allowing the "parallel evaluation" of several technologies (e. g., LID, followed by SID and ASR all on the same dataset) should be devised. The topic of ageing and recency of data for evaluation sets (e. g., ASR talking about George Bush as the US president in a dataset) need to be taken into consideration. Likewise, changing technology standards regarding audio quality should be revisited (e. g., for SID where the target speakers furthermore pose a similar problem as ageing vocabulary does for ASR). In general, evaluation (as well as training) datasets should be viewed more as *work in progress* than static artefacts.

In certain instances, the current state of the art TTS systems suffer from a lack of robustness in the generated speech, mainly when the kind of sentences seen during training is different from the one used during inference. Differences in length and syntactic structure make the underlying attention mechanism lose track and word-skipping, long silences, word repetition or even babbling may arise in the generated signal. However, these malfunctions are mostly scarce. Therefore, even if the system suffers from this problem it is difficult to observe it in a limited set of sentences like the one usually included in subjective evaluations. This lack of robustness, even if rare, limits the application of TTS technology and degrades the user experience. Guaranteeing robustness in modern TTS systems is paramount to ensure their ubiquitous presence and adoption in real-life products.

Being able to measure performance on several dimensions simultaneously, e. g., by measuring WER for ASR but under specific runtime and memory, constraints might be beneficial when investigating different setups and balances between performance, model-scale and resource consumption.

Even regarding established technologies such as LID, evaluations should be updated in order to allow for such multidimensional evaluations. Extension to further languages and language varieties, dialects and speaking conditions likewise should receive further attention to ensure broad availability and adoption.

Another very needed innovation is a method to objectively measure TTS results. TTS systems are evaluated by means of subjective evaluations campaigns which makes the measure-

ment of any advance time consuming and laborious. Several attempts to develop a robust objective measure that correlates well with people's judgements have been made, but no reliable algorithm has been found yet. Such an algorithm would make it easier to evaluate the development of new TTS techniques and would boost the advancements in the field.

Standard evaluations regarding privacy issues and bias of models are largely missing for many areas of ST. Evaluations in the realm of paralinguistics are still only scarce (a notable effort in this direction are yearly Paralinguistic Challenges at Interspeech conferences).

7.7 Outreach – communities, non-experts

Recent years have witnessed an increase in interest in the democratisation of AI. This concerns many fields of AI and ML; among them also the fields of voice and text technologies as well as the larger areas of NLP and NLU. The widespread application of ML and the well-known fact that experts in ML and AI have become scarce resources has led to the desire to empower a wider set of individuals to participate in the creation and use of these technologies. Toolkits and *do-it-yourself modelling* form part of the trend to democratise voice technologies. Approaches like Auto-ML aim to provide access to ML also for non-experts and as such align with strategies to allow a wider audience to participate in the process. As language technologies are aggregated and applied to more complex settings, inter-disciplinary research and activities e. g., from fields in the social sciences are becoming more relevant and synergies become apparent. Programs and funding schemes to actively engage these communities and foster inter-disciplinary research would further boost developments.

7.8 Alignments with EU policies and breakthroughs needed on the policy level

In terms of copyright, rules in Europe are more restrictive than in other economic regions and countries such as the United States. For example, utilising closed captions from TV broadcasts or subtitles from a copyrighted film to train and evaluate ST models could enable access to high-quality language data if lawmakers could agree that training of models on copyrighted data constitutes fair use, as long as it does not diminish the value of the assets or reduce the profits reasonably expected by the owner.

Similarly like in other LT, the pace of the development of Speech Technologies in Europe could be further increased by introducing changes that enable the re-use of existing data, while at the same time ensuring that the value of the copyright owners is not impaired.

The GDPR introduced a new global standard that places an emphasis on individual rights and reflects European values, and as such contributes to building trust in AI technologies. Regrettably, the GDPR has had a negative impact on the majority of Europe's LT business and research activities (Smal et al., 2020). For example, many stakeholders in data management, publication, and collection have come to wrongly believe that all data is personal by default. As a result, costly legal counsel and anonymisation methods are used in circumstances when they may be avoided or are not required at all. Furthermore, non-European AI firms have been able to operate free of GDPR constraints (in cases where neither data storage nor the identity of citizens concerns Europe) since then, giving them an economic advantage over EU firms. One of the required breakthroughs relates thus to ensure that while the individual rights are protected, the extent of these, in particular, in the practical settings and day-to-day operations, does not extend beyond the intended scope. Automatic, efficient and free anonymisation tools like the ones offered by the Mapa project[18] are required for all European languages.

[18] https://mapa-project.eu

8 Speech Technologies: Main Technology Visions and Development Goals

8.1 Speech technologies – the interface of the future

In many settings, voice provides the most natural way to interact with devices and appliances. The Internet of Things (IoT) and the tendency for computation to take place "at the edge" is turning into a key enabler of voice and speech technologies in many fields and application areas.

The coming years will witness an increased advance in voice technologies to the point that interacting with automated systems will be virtually indistinguishable from communication with human beings in many cases (ideally such systems should make it clear from the start that they are indeed not human). Interfaces predominately relying on typing, clicking and swiping will gradually transform into multimodal (or even fully virtual) interfaces including voice, shifting the task of adaptation from human users to computer systems.

At the same time, compared to the other modalities currently dominating the Human-Computer Interaction (HCI) landscape, communication will encompass richer kinds of (linguistic and paralinguistic) information, including gender, age, emotional or cognitive state, health conditions or speaker specific traits allowing for a more sophisticated and accurate speaker identification, modelling, adaptation and personalisation. These factors and their integration into HCI – as beneficial and powerful as they may be – also give rise to privacy and ethical concerns. They prompt questions of control, user understanding and intent when it comes to sharing information and the extent to which different kinds of information are transmitted and used in the future. Ensuing risks and the potential impact need to be carefully met and balanced with measures to increase security and trust, both, by technical means as well as policy- legislative measures. This formation of this balance will affect the adoption of a wide range of devices and services: from intelligent voice assistants in homes and on smartphones, navigation and control systems in cars to cooperative office and work environments and systems supporting a wide range of business and leisure activities.

The heavy increase in the use of virtual communication technology experienced during the COVID pandemic is viewed to be a trend that is also likely to continue into the future. A variety of platforms were able to increase their presence and extend their functionality during that period. Further extension and the inclusion of speech and language technologies into existing and emerging offers are likely to seamlessly link project management and communication tools with natural language processing.

TTS technologies will advance until they are able to generate natural speech with any desired voice, speaking style or emotion. ASR technologies can be expected to advance to performance levels on par with human operators (or exceeding it). In general, our interaction with machines will increasingly be carried out through speech and natural language. Our home appliances, cars, electronic gadgets and digital assistants will communicate with us to inform us about malfunctions, to help us program and use them, to advise us about their *needs*, to entertain us and keep us company and act as virtual colleagues at work. Technologies for input (like ASR) can be expected to increase their capacity to handle different expressions of language as much as output technologies (like TTS) can be expected to gain expressiveness and be able to generate voices with diverse speaking styles and personalities.

Progress in ASR will allow tapping into the large sea of multimedia data (including large existing archives), SID will act for authentication and personalisation, TTS systems integrated into our devices will allow converting any digital content into a multimedia experience.

8.2 Capabilities and technology shifts

User and application contexts A trend towards the integration of richer context is to be expected, regardless of the sub-field of voice processing. This concerns the individual technologies as well as their combination.

For TTS, to have a truly interactive experience when dealing with our devices, the integration of context will play a major role. E.g., the correct way to pronounce a message should be inferred from the text context or the previous steps of a dialogue. In this way, TTS systems would be able to generate the response with the correct inflexion so paralinguistic factors are correctly conveyed in addition to the purely linguistic information.

Technologies will need to be sensitive to the user's character, state, mood and needs and adapt themselves accordingly. Potentially, they will also need to take into account other members' states in case of group activities such as business meetings. Topics of pragmatics will be reflected by all technologies. Rather than individual communication turns, complete conversations with history and context will be the norm.

Addressing the existing technological gaps In the area of ASR, continues efforts towards better understanding and modelling human speech perception might result in sophisticated speech recognition addressing several of the technical limitations and gaps identified in current approaches. Improved handling of audio conditions currently perceived as difficult (e. g., multiple simultaneous speakers in noisy environments speaking spontaneously and highly emotionally in a mix of languages) will be possible by such advances. At the same time, a wider deployment and further popularisation of ST will also require solutions that offer high robustness, low latency, efficient customisation and the ability to provide possible equal support for a diverse set of speakers.

Speech technologies integration An intimate relation of ASR, SID and TTS with downstream NLP and NLU technologies is needed to allow the correct interpretation of the input so that recognition, meaning and output can be produced in a natural and correct manner. A combination of technologies to interact in multimodal ways (including visuals) and the efficient combination of inter-linked models will be able to guarantee the best experience possible. In turn, the successful combination will result in an enhanced easiness and naturalness of use, hiding individual components and allowing to perceive systems as assistants using natural language much in the way that human assistants would.

Multimodal models Recently introduced neural net architectures, e. g., Perceiver IO (Jaegle et al., 2021), support encoding and decoding schemes of various modalities. They can directly work with BERT-style masked language modelling using bytes instead of tokenised inputs. Another advantage of this type of architecture is that the computation and memory requirements of the self-attention mechanism don't depend on the size of the inputs and outputs, as the bulk of computing happens in a common Transformer-amenable latent space. Although being a task-agnostic architecture, the model provides competitive results on modalities such as language, vision, multimodal data, and point clouds. In the near future, this type of architecture will be commonly used in a range of applications where multimodal content needs to be jointly analysed. Further, the future line of work relates to the training of a single, shared neural net encoder on several modalities at the same time, and only using modality-specific pre- and post-processors. The computational complexity of Perceiver IO[19] is linear and the bulk of the processing

[19] https://huggingface.co/docs/transformers/model_doc/perceiver

occurs in the latent space, allowing to process larger inputs and outputs when compared to the standard Transformers. This enables large and heterogeneous models to become more scalable, and available to wider groups of users and application scenarios. Moreover, connections to areas such as knowledge-representation and knowledge-graphs may provide further alleys for research. In the longer-term perspective, such multimodal, *plug and play* architectures and models, will provide strong baselines in many areas, potentially also supporting less technical users with visual design tools, tractable hyper-parameter search, automated architecture, popularising the access to high performance, multimodal analysis and inferences.

Development pace The pace of development in voice-based technologies is driven by general advances in ML and associated hardware as well as domain-specific advances in the modelling of speech perception and production. The former can be expected to accelerate even more due to general interest in ML and AI from a wide portfolio of domains. Advances in transfer learning, reinforcement learning, fine-tuning, the use of pre-trained models and components as well as the arrival of platforms such a Hugging Face have created additional momentum. GPU support and extension of GPU capabilities can likewise be expected to continue at a fast pace, which might also have effects on the availability of hardware resources. The latter topics have been receiving increased attention as voice and language technologies entered the mainstream. Voice, being the most natural way to interact with systems can surely be assumed to attract even more commercial and academic interest in the future.

Training and evaluation Simultaneously, there will be further improvements introduced in the process of creation and distribution of ever-growing, ever more coherent (labelling quality), and diverse datasets. These will also include the creation of and increase in a number of large, multilingual, multi-domain and multimodal datasets, that will become de facto standard sets for the training and evaluation of the ST methods and systems that include ST components. In the next years, we will also witness an increase in labelling efficiency, a wider adaptation of continuous learning, self-adaptation and self-modification paradigms. While the number of languages available in the datasets will continue to grow, the quality and amount of data available for the most common, currently rich-resourced and the less common, currently low-resourced languages are unlikely to converge in a shorter term.

This development in the creation of more complex and multifaceted datasets calls for a more comprehensive evaluation and quality criteria; a shift that would change a focus from an individual speech technology to an end-user assessment of a complete experience when conducting a specific task in a given, non-laboratory environment and in a given operational and personalised contexts.

Whereas current learning paradigms focus predominately on training models on massive amounts of data in *one go (even though this itself may comprise many iterations)*, human learning takes place in complex steps over time, refining itself constantly along the way. New paradigms incorporating complex sequence learning may not only provide further insight into human language acquisition but likewise lead to even more powerful ST (NLP, NLU) models.

Infrastructure, hardware Extrapolating from the current trends a further rapid increase in the capacities of the ST related hardware and infrastructure can be foreseen. These include, e. g., faster communication networks and higher bandwidths, development and wider deployment of the specific hardware solutions dedicated to efficiently support specific ST components, e. g., audio level for ASR. Also, further popularisation of the ST solutions in particular in a form of IoT, and a new set of voice-enabled devices that will

be available to users in work, leisure and commerce settings can be foreseen. These developments create in turn additional challenges related to load and scalability of the underlying infrastructure, hardware and networks used. Moving computation to edge devices will certainly also continue to be a trend in the near future.

8.3 Privacy, accountability and regulations

The future of ST and a wider LT field development will be strongly influenced by the regulations governing the collection, storage, transmission, and use of personal data. These relate to the users' concerns and expectations, the influence of the groups of interest, both at the national and trans-national levels, and the future developments of the ST, their growing scope of application, functionalities and performance improvements. In the context of European AI companies and research institutes, the development pace appears to be particularly strongly influenced by the current and future regulation schemes. Lawmakers' decisions will thus have to consider the wide and profound impact of their regulations – on the protection of citizens personal data and privacy on the one hand, and on the wider field of AI technologies (research, development and application) and the comparative advantages and disadvantages vis-a-vis other geopolitical regions on the other hand.

Extrapolating from the current regulations concerning user privacy, the differences in data collection and use, the divide between the EU and non-EU countries will continue to grow. As AI technologies in the future will play a crucial role in defining competitive advantages across the different fields of human activities, including the commercial, social, military and intelligence, it is unlikely that a wider and far-reaching consensus between the competing countries and regions will be found, which would lead to a standardising set of regulations across the regions.

With the growing presence of speech technologies, ML and AI in general, growing concerns about the hidden flaws, shortcomings and baked-in biases of such systems is gaining momentum. This is certainly true from citizens perspectives, but also from academia as well as industry perspectives. Whereas citizens and academia may work towards enhancing transparency and creating mechanisms that may be able to avoid certain phenomena, the industry may work towards obfuscation and hindrance of the very same mechanisms.

In the US, laws requiring audits of algorithms used by employers for hiring and promotion are being installed and bills are drafted by Congress about the evaluation of decision-making systems used in areas such as healthcare, housing, employment or education[20]. At the same time, EU policy-makers are considering legislation requiring inspection of high-risk AI and a public registry of such systems.

Hiring software is known to already assign certain personality scores based on the SW used to create CVs or whether a bookshelf is visible in the background during an interview. The use of voice and speech technologies can easily be envisioned to extend such scenarios, e. g., by measuring anxiety in an applicant's voice using emotion detection technology.

Users will neither be able nor want to distinguish between AI, NLP or NLU, between a platform and a particular application or part thereof. To them, the overall system will be what they interact with and potentially what they will perceive as being biased, unfair or harming them in any way.

Disclosure of the use of AI/speech technologies Due to the ever more human-like nature of speech technologies, the use of AI technologies should be disclosed at the earliest stage possible for all transactions and applications. Making users aware of what they interact with is regarded as a fundamental step in the creation of more transparency.

[20] https://www.wired.com/story/movement-hold-ai-accountable-gains-steam/, Algorithmic Accountability Act, accessed on 12/13/2021

This will not prevent humans to attribute personhood to machines (think of toys and pet animals) or hinder human-like communication, but present an ethical and transparent frame around such settings.

Mandatory audits of algorithms and models Auditors will have to be independent for this to make sense and not open the door to even more secretive and evasive behaviour by companies. Federal agencies or boards may be required to preside over such activities. Standard test sets and tests may have to be created and applied.

Mandatory impact assessments of the introduction of such technologies The concept of measuring impact and potential harm is firmly established in fields such as environmental impact. Similarly, algorithmic impact assessments would need to cover a broad range of factors, with speech technologies and NLP focussing on language and language-use related aspects.

Public repositories of incidents where AI/NLP caused harm Public repositories and ways to report problematic uses of AI would allow to identify of repeat offenders and fine them in case of recurring problems or the unwillingness to act. Furthermore, making such cases known publicly may serve as an incentive to correct or prevent such cases.

Policy and law making Methods for assessments and measuring impacts need to transform into law so that they can properly be addressed and employed by courts and judges.

However, as the scope and impact of harm produced by AI/NLP are only known to a small extent at this stage, further research into all of the above areas is needed to create a sound foundation for the proper management of such risks.

Effects on society, workplace The current discussion about which jobs or areas within domains are likely candidates to be replaced by AI also carries over directly to the domain of speech processing – as well as to NLP more generally – as they can be seen to form a core element of AI in this context. Issues concerning automation and job replacement and the ensuing policy-making and social ramifications thus also directly concern speech technologies and their perception.

Pervasiveness A further spread and ubiquitous presence of voice-based technologies, and wider deployment of speech technologies across a multitude of services and devices due to reduction in size and integration into wearable and virtual environments can be expected. This may also concern further persons being in the vicinity of such deployments who may be involved indirectly by someone else's use of ST.

Sectors The most likely commercial areas where the ST will see further dynamic growth include banking, finance and insurance, consumer and electronics, education, healthcare as well as the automotive sector. These sectors will take advantage of the hardware directly supporting speech and voice technologies while reducing the operating costs related to processing the customers' requests and needs. The growth in these sectors will in turn require large language models (LLMs) trained on massive amounts of domain-specific data for different industries and verticals, as well as the *all-terrain NLP*. The processing will be handled by large data centres, proprietary systems on the premises, as well as on end-user devices at the edge. The distribution of this load across different processing facilities and devices will be driven by the policies and regulations governing the data collection, storage and use, the capacities of the portable devices and networks used, effects of scale and new training paradigms developed as well as by requirements and preferences of users and businesses using the technology.

A range of new applications will emerge, e. g., in-car assistance, combined with the self-driving cars, converting commuting into moving offices; self-service restaurants, which

combine ST, NLP and recommender technology, supporting fully automated order taking; or intelligent business meetings or travel assistance (see section 8.6 outlining further examples of Intelligent Personal Assistants). The *usual drivers* of innovation are also porn and crime. A wider adaptation of ST by these sectors gives a raise to a question on the (negative) effects, beyond the previously discussed challenges related to the data protection, hacking, fakes, mass and unlawful surveillance by a growing number of actors, and the ever-present privacy concerns.

In the environment where "your voice becomes your identity" ensuring the security of this sensitive data is of prime importance, e. g., zero trust approaches for voice requiring stringent authentication and monitoring, combined with advanced encryption.

8.4 Future applications

ST and in particular their combination with other NLP and AI technologies to form intelligent applications with human-like capabilities have the potential for disruptive innovation in a variety of sectors. Intelligent assistants and chat-bots currently provide the leading paths towards general and broad adoption. Future applications will be expected to understand users intents over sequences of interactions, blurring or completely eliminating perceived boundaries between individual technologies.

8.4.1 Customer contact centres / call centres

ST is already being used by multiple industries for customer contact centres to increase self-service functionalities, reduce average handling time, increase availability and reduce employee costs. ST within this scope – in particular when able to interact in a variety of languages and taking into account context – has the potential to increase customer acceptance and satisfaction.

8.4.2 Media and entertainment

The gaming and entertainment industries have traditionally been at the forefront of the adoption of new and emerging technologies. Whereas for gaming, the immersive experiences including interaction by voice have become a reality already, the further adoption of ST in the media industry may provide decisive mechanisms to reach global audiences (or local audiences speaking different languages). Automatic ASR, MT and TTS to produce subtitles or closed-captions on the fly, as well as the same array of technologies for interaction with smart home devices, provide alleys to deliver products to multinational audiences with minimal additional costs. A booming podcast industry, producing high volumes of multimedia content may likewise benefit dramatically from ST by allowing it to expand to previously inaccessible regions and audiences.

8.4.3 Marketing and PR

The integration of ST and in particular of paralinguistic factors into the feedback cycle for customer reviews and comments may provide a promising field for extension of applications beyond current capabilities. These mechanisms may also be combined with other settings such as in customer contact centres (e. g., to indicate fluctuations of sentiment or phases of anxiety or anger during conversations) or media and entertainment applications (e. g., to allow for personalisation of recommendations based on customer mood). Furthermore, ST may be adopted in the area of reputation management and news screening by allowing to

gain insights not only on textual information but also on multimedia information concerning institutions and individuals.

8.4.4 Healthcare

ST have been actively used in the healthcare sector for a considerable period of time already. Whereas companies such as Philips and Nuance have been successful in the transcription of radiology services, Microsoft is aiming to accelerate its presence in the medical sector by acquiring Nuance as of 2021.[21] The combination of technologies to advance conversational agents and NLU is expected to increase Microsoft's footprint in the area of ambient clinical intelligence solutions. ST may be helpful in situations where people cannot use language, do not know how to read or write (e.g be too young) or may suffer from cognitive disabilities.

8.4.5 Fraud detection and security

The use of ST on publicly available content as well as on phone conversations may aid in detecting financial crimes and support compliance and risk-management efforts.

8.4.6 Personalised ST

Voices for TTS will be generated for any language and be fully customisable. In the same way, as we can now personalise the avatars in video games, we will be able to set every aspect of the synthetic voice as we please to suit the characteristics we prefer for each situation. Capabilities of the voices will be increased and the systems will be able to sing in any musical style. They will also have the ability to adapt to the acoustic environment and produce speech that is easy to listen to even in unfavourable conditions.

Moreover, TTS technology will extend and speech will be generated not only from the text but from other input information that could be more convenient for some users who do not have easy access to text or for some situations like the ones requiring privacy. Multi-modal systems will allow generating speech from lip-reading, articulatory data acquired by diverse technologies such as electromyography, permanent magnet articulography and other silent speech interfaces, and even cerebral activity with brain-computer interfaces.

ASR technologies will be customisable depending on speaker preferences and traits, e. g., adapting themselves to speaking style, jargon, preferences in formulations etc. Furthermore, they may change dynamically on social and professional context.

8.5 Possible future directions and visions

8.5.1 Actors and markets

In order to counteract the Matthew Effect (whereby the GAFAs would be getting into an even stronger position over time), it is imperative to boost efforts leading to more data effective use of resources. Trends supporting the home market, leading to increased innovation focussing on local consumers and values, should be assisted and facilitated.

8.5.2 Customisation

Technologies may have reached an advanced level of maturity for many languages and domains. However, numerous further niches remain which require expertise and adaptation of base models to cover the last mile to the customer. In all areas of ST, the opportunity to

[21] https://techcrunch.com/2021/04/12/microsoft-is-acquiring-nuance-communications-for-19-7b/

capitalise on efforts and tasks which fall into this category exists and can be taken up by local champions.

8.5.3 Privacy and ethics

A sequence of scandals and growing interest in issues of ethics and privacy have led to an increased awareness in society. Trust in technology is a key ingredient for the adoption of technologies by a large portion of the population. Transparency in how privacy is integrated into technologies – algorithms and models – is expected to be a crucial ingredient to earn customers' trust. Despite the disparity of legislation across national and economic conglomerates, privacy-by-design beyond mere statements may become a decisive factor for technology uptake and market success. It is worth pointing out that privacy does not end within/around one's own device or sphere, but may also include neighbours and bystanders as was made explicit by a recent court ruling in the UK.[22]

8.5.4 Ambient Intelligence

The confluence of individual technologies to form an entity that is larger than the sum of the individual technologies is a recurrent theme within this document. This is especially important when combining human-like modalities for input and output with knowledge representation and reasoning, potentially in an augmented or virtual environment. Viewing ST as a means for intelligent interaction, integrating nuanced and fine-grained context and input from multiple modalities can be expected to lead to more human-like systems where the perception of individual components will blur into an overall experience for end-users. Such combinations may be a step towards a broader kind of AI as opposed to the narrow, highly-specialised versions in use today.

8.5.5 Augmented Intelligence

The current wave of AI holds the potential to impact and change a wide field of businesses and workplaces. Whereas experts predict that putting AI (including NLP and ST) to work at a larger scale will add more than 15 trillion USD to the global economy by 2030[23], this outlook also creates fears and anxieties about the replacement of a large portion of the workforce by machines. Efforts to position AI as intelligent partners in collaborative environments of humans and machines – augmenting human capabilities with AI – with each side capitalising on their respective strengths may not only lead to greater productivity but also lead to higher societal acceptance of the (disruptive) introduction of novel technology into areas previously perceived to be the domain of humans only.

8.5.6 On the road to winter again?

The hype about AI and the accompanying over-selling by sales organisations has caused waves of deception and eventually ended in the so-called AI Winter before (Hendler, 2008; Floridi, 2020; Muehlhauser, 2016). Currently, AI is experiencing another wave of hype. Investments, in particular in the USA and China, are exploding[24], with start-ups and companies receiving hundreds of millions of dollars[25]. ST (in the guise of Language AI, Voice Technologies or Conversational AI) is at the centre of many such investments and activities. Based on

[22] https://www.theguardian.com/uk-news/2021/oct/14/amazon-asks-ring-owners-to-respect-privacy-after-court-rules-usage-broke-law
[23] https://hbr.org/2021/03/ai-should-augment-human-intelligence-not-replace-it
[24] https://www.forbes.com/sites/robtoews/2021/12/22/10-ai-predictions-for-2022/amp/
[25] https://towardsdatascience.com/nlp-how-to-spend-a-billion-dollars-e0dcdf82ea9f

the assumption that technologies have matured and are now ready for deployment on the market, large sums are being invested and high expectations raised. It remains to be seen whether and to what extent these expectations can be fulfilled or whether over-selling will send AI – and with it NLP and ST – into another phase of hibernation.

8.5.7 Supermodels

Recent years have witnessed a fierce race between renowned institutions and research labs on who can build the largest model for NLP. It has become customary that only actors with enormous resources at their disposal can participate in this race: Facebook, Google, Microsoft and their Chinese counterparts. Recently, Microsoft teamed up with Nvidia to create a language model with 530 billion parameters (MT-NLG[26]) and DeepMind created Retro[27]. Whereas these huge foundation models suffer from the same shortcomings as their predecessors in terms of bias, the integration of toxic language, the lack of explainability, etc., performance on many tasks is still improving with the number of parameters and no end of this race is currently in sight. As is the case for search technologies, the US and Chinese giants are leading these activities. European efforts like the German OpenGPT-X project[28] aim to mitigate this imbalance. Text-creation in a human-like manner, for a multitude of domains, different languages and including a variety of stylistic elements are already possible today and can only be expected to improve further. With regard to ST and E2E models or as part of aggregated models comprising several technologies, similar activities can be expected. Whereas in the past access to sufficient amounts of data has been the determining factor, this tendency turns access to computing resources into the next crucial bottleneck. As for the large language models, again GAFAs with access to data (and users feeding more data to them every day) are in the pole position for market dominance. In the recently published work, Bommasani et al. 2021 (Bommasani et al., 2021), provides a thorough account of the opportunities and risks of such foundation models, ranging from their capabilities, technical principles, applications and societal impacts.

8.6 Examples for Intelligent Personal Assistants

Public transport While waiting for a train at the train station, a commuter notices that the train seems to be late. As is often the case, no announcement about this state was made and the display still shows the original time of departure. On the platform teeming with people, the user, upset about this recurrent situation asks their device “what’s the matter with the stupid train?” This question is uttered in a noisy environment (a busy train station) and in a highly emotional manner. Potentially, dialect or slang is used when expressing discontent. In spite of the challenging conditions, the system recognises the commuters question. By searching the train company’s database, it retrieves the reason for the delay and the expected time of departure. Taking into account the time of day and week-day (and hence the commuting conditions), it suggests an alternative way to travel. Output is provided in a way taking into account the noisy conditions (volume, speed) as well as the annoying state of the commuter (appropriate phrasing, tone and voice).

Business meeting During a business meeting with several members, several issues remain to be resolved and follow-up data and time is being searched. While a couple of col-

[26] https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/

[27] https://deepmind.com/research/publications/2021/improving-language-models-by-retrieving-from-trillions-of-tokens

[28] https://www.iais.fraunhofer.de/de/presse/news/news-210701.html

leagues discuss a possible slot for a meeting on the morning of the following Thursday, one of them addresses the user with the question "any chance you can make it?" The system recognises that the colleague who asked is John and that its owner is the one being addressed by John, translates time and day into actual calendar days, searches John's and the owner's calendar and finds, that a slot after 10 AM would be best as the owner has a private (doctor's appointment at 8 AM) and that John usually does not come to work before 10 AM. As others are listening and the doctor's appointment is deemed to be a personal matter, the system decides not to use voice but rather display the information on the smart device's screen, signalling the owner of the potential availability at the intended time and day.

Business Assistant While preparing a presentation for an upcoming meeting, a user realises that they have created a slide on a similar topic not too long ago. By asking the IPA "Can you find that slide with the figure on the different kinds of media and the role they play with regard to hate-speech for me?" a search is triggered. The IPA responds with "You mean the one which you used at the speech you gave at the University of Vienna in December?" turning speech into an interactive tool for search. Upon confirmation, the correct information is provided.

Travel Trying to book a flight, the user tells their IPA "I'd like to fly to Boston next Tuesday or Wednesday". Recognising the intent to find a suitable connection, the IPA takes into account that the user prefers direct flights, is an aficionado of new aircraft (so would prefer e. g., an Airbus A350 or a Boeing 787), prefers airlines within the Miles and More group, hates Paris' Charles-de-Gaulle airport and prefers window seat upfront. Furthermore, the IPA knows that the user's home base in Munich, so the departure would likely be from there. Searching the respective databases, the system finds a small number of flights from MUC to BOS and presents the top-ranking one (Lufthansa which happens to be using the new A350-900 on this route) in a verbose manner whereas only pointing out some key facts about the remaining flights. Receiving confirmation the dialogue then proceeds to the actual booking.

Multimodal search While searching online for a new computer, using smart glasses and voice, the user is presented with visual feedback about several models fitting the bill. The user scrolls through the list and at the end of it finds that the 3rd model on the first list of results suited them best. The user enters into a dialogue like the following "show me the details again of that one with the new Intel quantum XYZ CPU and with the 100TB drive, yes that one and the one below as well for comparison". The IPA translates from descriptive language into the actual items, taking into account the history and visual positioning of previously shown results.

Personal Coach During a long day of tense and stressful business meetings, the IPA notices an increased level of tension in the user's voice. Combining information from wearables, it determines that the user should definitely take a deep breath and maybe take a short break before continuing. Taking into account the meeting's state it sends a buzz to alert the user and then uses a smooth and reconfirming voice to prompt her to open the window and have a cup of coffee – taking into account that she has had one cup so far and usually consumes up to five cups of coffee a day.

9 Speech Technologies: Towards Deep Natural Language Understanding

In this document, Natural Language Understanding is viewed and treated as a subset of the field of Natural Language Processing which itself is a subfield of Artificial Intelligence. Furthermore, as outlined in previous chapters, Speech Technologies form a subfield of NLP. The term ***Deep*** in this context is interpreted as a means to distinguish this kind of NLU from previous approaches in two fundamental ways: the inclusion of a variety of knowledge sources allowing for a richer and more complex manner of processing and interaction as well as the use of DNNs (with multiple heterogeneous layers) to encompass several models which traditionally formed separate units into one overall model (E2E). Furthermore, these systems and models may involve different modalities, receiving their input as a mix of audio, video and text such as described by (Akbari et al., 2021).

As stated in the chapters above, in many instances the most natural manner for humans to interact with machines is through voice. This entails using voice to issue commands or queries as well as the use of voice for the generation of answers and statements[29].

Certain types of scenarios (e. g., ones limiting the interaction to small, hand-held devices) may call for voice-only interaction whereas other scenarios (e. g., allowing for feedback via large screens, augmented- or virtual reality environments) may favour multimedia settings, permitting the flow of information across different modalities in parallel. Yet other scenarios may ask for communication completely without the use of audio, in particular when considering special needs and inclusive communication.

Speech technologies play a role in the ingestion of information – by acting as a kind of sensor conveying linguistic as well as paralinguistic inputs and converting them into structured information. Equally, their use concerns the output of information in auditive (speech but also non-speech, such as confirming "uh-huh") form to communicate with human users. Both directions of the flow of information apply to human-computer interaction as well as human-to-human interaction in the case of groups of human users interacting with each other or with computers, e. g., during meetings with intelligent assistants for transcription, translation and summarization.

Speech technologies thus form an intermediate interface layer between humans and machines. Inbound (auditive) information is captured and enriched by ST before being passed on to downstream Natural Language Understanding (NLU) processing. Outbound information is enriched, transformed and eventually realised as audio based on content, structure and meta-information provided by semantic components.

The semantics and interpretation of utterances as well as the generation of appropriate responses based on a logical representation and state of a conversation fully resides within the scope and components of NLU and technologies such as dialogue-managers (to carry on conversations) or knowledge graphs (networks for semantic representation). As such, ST provide essential contributions to the functioning of NLU both, in the input as well as the output directions. However, they do not perform any semantic processing (understanding) themselves.

As indicated above, visual cues such as gestures or manual articulation (sign-language) may replace the audio-element of ST when operating in noisy environments or involving hard hearing impaired or deaf people. Technologies from the field of visual processing assume the roles of ST in these cases. The combination of modalities is also possible and may be appropriate/imperative depending on the actual environment, such as working environ-

[29] In saying so, we silently assume that "we expect to communicate with the computer in the same way we would with another human" (Winograd, 2006). However, we acknowledge that there might be practical and even philosophical objections to encouraging people to attribute human qualities and abilities to their computers (idem). Whether and to what extent this is possible is a matter of discussion.

ments requiring a hands-free operation.

The contribution of ST towards achieving deep NLU may thus lie in the improvement and extension of the individual technologies (both from accuracy as well as a language-/domain-coverage perspective), from their integration into E2E systems allowing for joint operation and optimisation, including different kinds of knowledge sources and from their flexible and dynamic configuration depending on the state and context of an application or user. Approaches including the combination of several modalities both, for input and for output may likewise provide beneficial in the context of achieving deep NLU.

In many cases, the real power of NLU will become perceptible when it features as part of a complex system functioning as a human-like counterpart in communication – exhibiting context, history and elements of general intelligence. However, it may also be then, that NLU is overshadowed by the cognitive downstream processing and eventually perceived as a mere commodity. The element of admiration and awe on part of the user will then concern the complete system performance, with NLU itself disappearing in importance as a small part of a much larger and complex intelligent system.

10 Summary and Conclusions

The substantial advances made in the field of ST over the past decades hold the potential for disruptive innovation in many areas and application domains. Combined with the progress of related fields such as AI, NLU, NLP and ML, they provide the basis for broad adoption of speech and voice as the primary modality to interact with computer systems as part of larger and more complex systems modelling human-like communication and interaction.

This report has identified several fields and business domains that provide promising areas for the use of ST and their inclusion into larger solutions providing a more natural means of communication. However, at the same time, several issues and challenges have also been identified which need to be addressed and resolved in order to make this promise materialise. The following list summarises the key elements identified within this report and provides a list of directions and recommendations for possible future actions.

In general, ST are expected to become part of larger systems, interacting with users in a human-like manner and thus allowing a wider adoption of ST, NLP and NLU. In parallel, the individual technologies and components will continue to be improved, both in terms of accuracy as well as of coverage (of language). All these strands of advancements can aid in supporting the overarching goal of achieving digital language equality by providing services made possible by these technologies to larger audiences or equal (similar) levels of scope and performance.

Pandemic changes As people are now more used to online, collaborative environments due to lockdowns and limited access to offices, they will likely want to keep using them. This creates further demand for better and more tightly integrated ST and downstream processing.

The importance of data The availability/scarcity of training data is still a key factor in the creation of ST as long as supervised paradigms prevail. Accessibility is often limited, or even locked, with individual actors amassing massive amounts of data, effectively creating monopolies for certain markets. Licenses and data related regulations as well as operability and compatibility of different data resources and providers remains an obstacle that needs to be addressed. Methods not relying on vast amounts of data (e. g., fine-tuning) form an active area of research. Furthermore, language-agnostic (or multilingual) models may provide answers and workarounds. Approaches like PARP (Lai et al., 2021) already provide promising results but need to be extended further. Even if

training paradigms might change to resemble human learning (focussing on complex sequences rather than on single avalanches of data), this challenge will remain.

Multi-lingual, language-agnostic models "Take any language, for example, English ..." – this has been a running gag in many institutions dedicated to ST, but it equally applies to NLP and NLU. Even though the scope of languages supported by ST has increased dramatically over the past decades, English still holds a unique position when it comes to resources. On the one hand, the creation of resources for further languages and dialects (some may only be spoken) is an ongoing activity. On the other hand, the investigation of phenomena that are only present in other language families forms an active area of research around the globe. The creation of multi-lingual or language-agnostic models provides further alleys for improvement.

Complex E2E systems, combined knowledge sources Substantial growth in the availability of data for some languages paired with a boost of processing capabilities created a trend of integrating models, which previously existed in isolation, into a combined overall, model. Training and optimisation take place in a single framework rather than individually, better capitalising on joint factors. Considerable progress in performance has been made through this approach which can be expected to continue also in the future. The integration of semantic components such as NLU or knowledge-graphs, into these frameworks, may provide additional elements required for truly intelligent interaction. However, an increased lack of explainability may ensue from such integration and prompt additional activities and parallel efforts to address this significant limitation. Progress and collaboration with fields such as neuroscience and psychology may lead to deeper and more human-like approaches to learning and modeling of cognitive capabilities.

Diversity of context In current setups and applications, different components, including ST, largely operate in an independent and isolated manner. For example, ASR recognises speech without any specific context concerning dialogue-state or user-preferences or intentions. The dynamic inclusion and integration of such further context would potentially allow for ST to operate on a significantly higher level of accuracy, eliminating errors and narrowing down alternatives based on the increased context and/or boosting more sensible alternatives. Various ways for the fusion of information have been investigated such as early and late fusion, but have not effectively come to fruition in many circumstances. Novel ways employing systems, parallel systems for multiparty conversation settings and multimodal approaches may provide a way forward.

Multimodality ST predominantly address the modality of using voice for interaction with computers. This encompasses linguistic as well as paralinguistic elements and may extend to sign-language to some extent. The combination of ST with multimodal inputs and outputs may provide a basis for next-generation HCIs. Inclusion of gestures, facial expression, emotions or haptics as well as the generation of multimodal outputs reflecting these elements may result in a much richer and more natural user experience and lead to wider adoption and acceptance of ST.

Measuring performance, benchmarks and robustness WER has been the standard measure of performance for ASR for decades. Although this established measure allows quantifying progress in ASR, it only tells part of the truth when it comes to the real application of ASR and its combination with downstream processing. The evaluation of TTS is still largely subjective and relies on human subjects and lacks a truly objective approach. In many fields of ST, performance has reached (near-) human levels under controlled conditions with academic progress being significant in theory but often only marginal when translated into reality. A shift towards increasing robustness

and generality of results may prove beneficial at this stage. Several standard datasets for evaluation exist for a variety of ST and languages. However, for several areas, e. g., ones concerning paralinguistic phenomena or certain languages and dialects, no such standard datasets exist and remain to be established.

ST as commodities Recent progress and an abundance of ST in chatbots like Amazon's Alexa or Apple's Siri may evoke expectations of ST being a mere commodity and raise unrealistic expectations on the part of casual users. On the one hand, like other technologies and models, ST perform considerably worse when applied to conditions, unlike the ones for which they were originally created. Adaptation and customisation to special domains thus provide an opportunity and market niche for specialists. In addition, the management of expectations and open communication about the possibilities but also limitations on part of the ST community may help set expectations to realistic and practical levels.

Digital language equality If an equal level of support across languages is the long-term goal, it can be fully addressed neither in the current non-free market, digital oligopolies dominated environment, nor by purely free-market mechanisms. Development and progress in ST has been driven by commercial entities over the past years. These, understandably operate in terms of markets and shareholder value and thus will only be willing to invest efforts on economically less appealing languages under certain limited conditions, e. g., PR, influential individuals, policy-demands or when the profits foreseen for providing support for a less common language outweighs the marginal profits that can be realised by providing an incremental improvement to one of the already supported languages or dialects.

Biases and ethical issues The interest and concern about fairness and biases baked into models and ethical issues relating to models and their use have been receiving increased attention. Methods for detecting biases and de-biasing need to be improved and are expected to become a more active area of development. Furthermore, access to ST for people with disabilities and impairments, e. g., by the inclusion of visual processing, needs to be extended. Cultural factors of language and its use (e. g., levels of politeness etc.) should be configurable and adaptive to the situation at hand.

Another related ethical issue arises when considering influential agents and bots (Allouch et al., 2021). With the current and near-future state of the technology, many businesses, political parties and ideological movements may develop conversational agents as a representative to convey their agenda and sway public opinion to get support for their cause. Situations, where the agents' identity is known or hidden, should be clearly distinguished. Cases where a company or party is represented by a single conversational agent or by several, hundreds, or even thousands to create a representation of mass support, should be clarified and marked. While many applications that integrate ST and LT are useful and even necessary, these application scenarios should be closely monitored for ethical and privacy aspects.

Transparency and interpretability Triggered by an increased interest in the fairness of the application of AI systems, e. g., filtering and preliminary assessments of job applications, prison-parole, credits, sectors like NLP and ST are and justifiably will in future continue to be confronted with similar questioning and scrutiny. Users are likely to demand explanations on the capabilities and functioning of ST. Results are likely to be questioned with some application areas demanding audits of models and algorithms. Technical issues will need to be addressed and accompanied on a policy-making and legislative level. Standardisation of evaluations and publication of results may function as motivating factors for providers to address these issues more thoroughly.

Balance between convenience and privacy Scandals, data leaks and an increase in cybercrime have brought issues of security and privacy onto the table. On the one hand, devices are ever more pervasive, taking ST into people's offices and homes. IoT and wearables will continue and further accelerate this trend. On the other hand, users are becoming increasingly wary of the risks and undesired effects related to the introduction of ST. Clandestine manners of data collection and eavesdropping infringing privacy are published and castigated by the media. Actors risk suffering dire consequences if they do not respond and put corrective measures into position. The balance between convenience and privacy will remain a fluid one to be negotiated repeatedly and on multiple levels.

Policies and regulations The legislation governing the acquisition, storage, transmission, and use of personal data will have a significant impact on the future of ST and the wider LT area. These in turn stem from user concerns and expectations, the influence of interest groups both at the national and transnational levels as well as the rapid developments in the relevant fields. Extrapolating from the current trends, the gap between the regulations used in different regions will continue to widen. As AI technologies play a critical role in creating competitive advantages across a wide range of human activities, including commercial, social, military, and intelligence, it is unlikely that competing countries and regions will be able to reach a broad, far-reaching agreement, resulting in one standardised set of regulations, respected and followed in practical settings in different sectors. The lawmakers' decisions will thus have to consider a wide and profound impact of their regulations, on the protection of citizens personal data and privacy on the one hand, and on the pace of development in a wider field of AI technologies: research, development and application and the comparative advantages and disadvantages vis-a-vis other regions and the global centres of the AI technologies development.

ST impact on society As technologies are never socially neutral and need to be accepted by society in order to be adopted, technological advancements as described in this document are not exclusively technical ones, but need to be accompanied by progress from the humanities. Multi-disciplinary approaches, as demonstrated by the rise of the digital humanities may prove advantageous also in these scenarios. As systems become natural companions, the fields of psychology, neuroscience and philosophy will bring new aspects and visions to the agenda and inspire novel approaches. Fear and anxieties generated by overly aggressive marketing, science-fiction and disinformation need to be met with prudent transparency, adequate management of expectations and accompanying policy measures. An inclusive approach in the sense of making ST (and AI) visible, transparent and understandable to a larger public – a kind of AI-literacy in the sense of media-literacy – may be a strong supporting topic for all above-mentioned domains. An increase in transparency may be expected to lead to changes in what is perceived as NLU (or AI): a deeper knowledge of algorithms and models made change the notion of what intelligence per se means – much as when viewing a mosaic up close and stating that it is "a mere collection of small tiles and some mud inbetween", rather than marvelling at the byzantine mosaics of the Hagia Sophia from below.

People have always tended to humanise machines. More powerful systems formed by the combination and integration of technologies and components described above may effectively be attributed human-like qualities and personhood by their users. It is imperative that ethical aspects of such interaction also be addressed in parallel with technological progress. Transparency, e. g., by chatbots introducing themselves and stating clearly that they are a machine, and openness is among the key factors to be considered when leaving users a freedom of choice rather than imposing technology on them. This

certainly reaches far beyond ST but rather concerns AI in general.

Lastly, in trying to address the goals of establishing DLE and aiming to measure its importance, we should maybe also ask ourselves what the consequences and effects would be of NOT doing so.

References

Hassan Akbari, Linagzhe Yuan, Rui Quian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Gong Boquinq. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104:11178*, 2021.

Mehmet Berkehan Akçay and Kaya Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116: 56–76, 2020.

Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. How might we create better benchmarks for speech recognition? In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 22–34, 2021.

Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 292–301, 2018.

Merav Allouch, Amos Azaria, and Rina Azoulay. Conversational agents: Goals, technologies, vision and challenges. *Sensors*, 21(24):8448, 2021.

Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley. Music, search, and iot: How people (really) use voice assistants. *ACM Transactions on Computer-Human Interaction*, 2019.

Elisabeth Andre, Matthias Rehm, Wolfgang Minker, and Dirk Bühler. Endowing spoken language dialogue systems with emotional intelligence. In *Tutorial and Research Workshop on Affective Dialogue Systems*, pages 178–187. Springer, 2004.

Tursunov Anvarjon, Soonil Kwon, et al. Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features. *Sensors*, 20(18):5212, 2020.

Sercan Ö Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. Deep voice: Real-time neural text-to-speech. In *International Conference on Machine Learning*, pages 195–204. PMLR, 2017.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 2020.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Anton Batliner, Björn Schuller, Dino Seppi, Stefan Steidl, Laurence Devillers, Laurence Vidrascu, Thurid Vogt, Vered Aharonson, and Noam Amir. The automatic recognition of emotions in speech. In *Emotion-Oriented Systems*, pages 71–99. Springer, 2011.

Eric Battenberg, RJ Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, and Tom Bagby. Location-relative attention mechanisms for robust long-form speech synthesis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6194–6198. IEEE, 2020.

Billal Belainine, Fatiha Sadat, and Hakim Lounis. Modelling a conversational agent with complex emotional intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13710–13711, 2020.

Christian Benoît, Martine Grice, and Valerie Hazan. The sus test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech communication*, 18(4):381–392, 1996.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Martin Borchert and Antje Dusterhoft. Emotions in speech-experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. In *2005 International Conference on Natural Language Processing and Knowledge Engineering*, pages 147–151. IEEE, 2005.

Cassia Valentini Botinhao and Simon King. Detection and analysis of attention errors in sequence-to-sequence text-to-speech. In *Interspeech 2021: The 22nd Annual Conference of the International Speech Communication Association*, 2021.

Herve Bourlard and Nelson Morgan. Continuous speech recognition by connectionist statistical methods. *IEEE Transactions on Neural Networks*, 4(6):893–909, 1993.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

Yuewen Cao, Xixin Wu, Songxiang Liu, Jianwei Yu, Xu Li, Zhiyong Wu, Xunying Liu, and Helen Meng. End-to-end code-switched tts with mix of monolingual recordings. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6935–6939, 2019.

Edresson Casanova, Christopher Shulby, Eren Gölge, Nicolas Michael Müller, Frederico Santos de Oliveira, Arnaldo Candido Jr., Anderson da Silva Soares, Sandra Maria Aluisio, and Moacir Antonelli Ponti. SC-GlowTTS: An Efficient Zero-Shot Multi-Speaker Text-To-Speech Model. In *Proc. Interspeech 2021*, pages 3645–3649, 2021.

Alejandro Catala, Deniece S. Nazareth, Paulo Félix, Khiet P. Truong, and Gerben J. Westerhof. Emobook: A multimedia life story book app for reminiscence intervention. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '20. Association for Computing Machinery, 2020. ISBN 9781450380522. doi: 10.1145/3406324.3410717.

William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*, 2015.

Yuan-Jui Chen, Tao Tu, Cheng chieh Yeh, and Hung-Yi Lee. End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning. In *Proc. Interspeech 2019*, pages 2075–2079, 2019.

Po chun Hsu and Hung yi Lee. WG-WaveNet: Real-Time High-Fidelity Speech Synthesis Without GPU. In *Proc. Interspeech 2020*, pages 210–214, 2020.

Yu-An Chung, Yuxuan Wang, Wei-Ning Hsu, Yu Zhang, and RJ Skerry-Ryan. Semi-supervised training for improving data efficiency in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6940–6944. IEEE, 2019.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019a.

Rob Clark, Hanna Silen, Tom Kenter, and Ralph Leith. Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs. In *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, pages 99–104, 2019b.

Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. " what can i help you with?" infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–12, 2017.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

Xudong Dai, Cheng Gong, Longbiao Wang, and Kaili Zhang. Information sieve: Content leakage reduction in end-to-end prosody for expressive speech synthesis. *arXiv preprint arXiv:2108.01831*, 2021.

Poorna Banerjee Dasgupta. Detection and analysis of human emotions through voice and speech pattern processing. *arXiv preprint arXiv:1710.10198*, 2017.

Tusar Kanti Dash, Soumya Mishra, Ganapati Panda, and Suresh Chandra Satapathy. Detection of covid-19 from speech signal using bio-inspired based cepstral features. *Pattern Recognition*, 117, 2021. doi: https://doi.org/10.1016/j.patcog.2021.107999.

Sofia de la Fuente Garcia, Craig Ritchie, and Saturnino Luz. Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: a systematic review. *Journal of Alzheimer's Disease*, pages 1–27, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Association for Computational Linguistics*, 2018.

Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 509–516, 2013.

Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE, 2018.

Christoph Draxler, Henk van den Heuvel, Arjan van Hessen, Silvia Calamai, and Louise Corti. A CLARIN transcription portal for interview data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3353–3359, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.411.

Sidney D'Mello and Art Graesser. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157, 2012.

Francisca F Pessanha and Almila Akdag Salah. A computational look at oral history archives. *Journal on Computing and Cultural Heritage*, 15(1), 2022. doi: https://doi.org/10.1145/3477605.

Luciano Floridi. Ai and its new winter: from myths to realities. *Philosophy & Technology*, 33(1):1–3, 2020.

Radhika Garg and Subhasree Sengupta. He is just like me: a study of the long-term use of smart speakers by parents and children. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–24, 2020.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. Investigating the impact of gender representation in asr training data: a case study on librispeech. In *3rd Workshop on Gender Bias in Natural Language Processing*, pages 86–92. Association for Computational Linguistics, 2021.

Miguelángel Verde Garrido. Why a militantly democratic lack of trust in state surveillance can enable better and more democratic security. In *Trust and Transparency in an Age of Surveillance*, pages 221–240. Routledge, 2021.

Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis. A dimensional approach to emotion recognition of speech from movies. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 65–68. IEEE, 2009.

Andrew Gibiansky, Sercan Ömer Arik, Gregory Frederick Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech. In *Proceedings of NIPS*, 2017.

John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society, 1992.

M Goldstein. Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener. *Speech communication*, 16(3):225–244, 1995.

Gábor Gosztolya, Veronika Vincze, László Tóth, Magdolna Pákáski, János Kálmán, and Ildikó Hoffmann. Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using asr and linguistic features. *Computer Speech & Language*, 53:181–197, 2019.

Avashna Govender and Simon King. Measuring the Cognitive Load of Synthetic Speech Using a Dual Task Paradigm. In *Proc. Interspeech 2018*, pages 2843–2847, 2018.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. Larger-scale transformers for multilingual masked language modeling. *arXiv preprint arXiv:2105.00572*, 2021.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

Frantisek Grézl, Martin Karafiát, Stanislav Kontár, and Jan Cernocky. Probabilistic and bottle-neck features for lvcsr of meetings. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–757. IEEE, 2007.

Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. The vera am mittag german audio-visual emotional speech database. In *2008 IEEE international conference on multimedia and expo*, pages 865–868. IEEE, 2008.

Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.

Andrea Guidi, Claudio Gentili, Enzo Pasquale Scilingo, and Nicola Vanello. Analysis of speech features and personality traits. *Biomedical Signal Processing and Control*, 51:1–7, 2019. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2019.01.027. URL https://www.sciencedirect.com/science/article/pii/S1746809419300230.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040, 2020. doi: 10.21437/Interspeech.2020-3015. URL http://dx.doi.org/10.21437/Interspeech.2020-3015.

Fasih Haider, Senja Pollak, Pierre Albert, and Saturnino Luz. Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods. *Computer Speech & Language*, 65: 101119, 2021.

Mutian He, Yan Deng, and Lei He. Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS. In *Proc. Interspeech 2019*, pages 1293–1297, 2019.

James Hendler. Avoiding another ai winter. *IEEE Intelligent Systems*, 23(02):2–4, 2008.

Hynek Hermansky, Daniel PW Ellis, and Sangita Sharma. Tandem connectionist feature extraction for conventional hmm systems. In *2000 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1635–1638. IEEE, 2000.

Margaret Hu. Cambridge analytica's black box. *Big Data & Society*, 7(2):2053951720938091, 2020.

Kun-Yi Huang, Chung-Hsien Wu, and Ming-Hsiang Su. Attention-based convolutional neural network and long short-term memory for short-term detection of mood disorders based on elicited speech responses. *Pattern Recognition*, 88:668–678, 2019.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation. In *Proc. Interspeech 2021*, pages 2207–2211, 2021.

Frederick Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, USA, 1998. ISBN 0262100665.

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pages 2410–2419. PMLR, 2018.

Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, Tianyan Zhou, and Takuya Yoshioka. Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers. *arXiv preprint arXiv:2006.10930*, 2020.

Naoyuki Kanda, Xuankai Chang, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka. Investigation of end-to-end speaker-attributed asr for continuous multi-talker recordings. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 809–816. IEEE, 2021a.

Naoyuki Kanda, Guoli Ye, Yu Wu, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka. Large-scale pre-training of end-to-end multi-talker asr for meeting transcription with single distant microphone. *arXiv preprint arXiv:2103.16776*, 2021b.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE, 2019.

Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345, 2019.

Alexandra Konig, Aharon Satt, Alex Sorin, Ran Hoory, Alexandre Derreumaux, Renaud David, and Phillippe H Robert. Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. *Current Alzheimer Research*, 15(2):120–129, 2018.

Shashidhar G Koolagudi and K Sreenivasa Rao. Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117, 2012.

Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017.

Theodoros Kostoulas, Todor Ganchev, and Nikos Fakotakis. Study on speaker-independent emotion recognition from speech on real-world data. In *Verbal and nonverbal features of human-human and human-machine interaction*, pages 235–242. Springer, 2008.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Puneet Kumar, Sidharth Jain, Balasubramanian Raman, Partha Pratim Roy, and Masakazu Iwamura. End-to-end triplet loss based emotion embedding system for speech emotion recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8766–8773. IEEE, 2021.

Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. Emotion recognition by speech signals. In *Eighth European conference on speech communication and technology*, 2003.

Cheng-I Jeff Lai, Yang Zhang, Alexander H Liu, Shiyu Chang, Yi-Lun Liao, Yung-Sung Chuang, Kaizhi Qian, Sameer Khurana, David Cox, and James Glass. Parp: Prune, adjust and re-prune for self-supervised speech recognition. *arXiv preprint arXiv:2106.05933*, 2021.

Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Julien Epps, and Bjórn Wolfgang Schuller. Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. *IEEE Transactions on Affective Computing*, 2020.

Duc Le, Gil Keren, Julian Chan, Jay Mahadeokar, Christian Fuegen, and Michael L Seltzer. Deep shallow fusion for rnn-t personalization. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 251–257. IEEE, 2021.

Sinae Lee, Jangwoon Park, and Dugan Um. Speech characteristics as indicators of personality traits. *Applied Sciences*, 11(18):8776, 2021.

Xiang Li, Changhe Song, Jingbei Li, Zhiyong Wu, Jia Jia, and Helen Meng. Towards Multi-Scale Style Control for Expressive Speech Synthesis. In *Proc. Interspeech 2021*, pages 4673–4677, 2021.

Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Alzheimer's dementia recognition through spontaneous speech: the adress challenge. *arXiv preprint arXiv:2004.06833*, 2020.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500, 2007.

Soumi Maiti, Erik Marchi, and Alistair Conkie. Generating multilingual voices using speaker space translation based on bilingual speaker data. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7624–7628, 2020.

Georgia Maniati, Nikolaos Ellinas, Konstantinos Markopoulos, Georgios Vamvoukakis, June Sig Sung, Hyoungmin Park, Aimilios Chalamandaris, and Pirros Tsiakoulis. Cross-Lingual Low Resource Speaker Adaptation Using Phonological Features. In *Proc. Interspeech 2021*, pages 1594–1598, 2021.

Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. The enterface'05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 8–8. IEEE, 2006.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17, 2011.

Carlos Mena, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, and Albert Gatt. Data augmentation for speech recognition in maltese: A low-resource perspective. *arXiv e-prints*, pages arXiv–2111, 2021.

Gelareh Mohammadi and Alessandro Vinciarelli. Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing*, 3(3):273–284, 2012.

Devang S Ram Mohan, Vivian Hu, Tian Huey Teh, Alexandra Torresquintero, Christopher GR Wallis, Marlene Staib, Lorenzo Foglianti, Jiameng Gao, and Simon King. Ctrl-p: Temporal control of prosodic variation for speech synthesis. *arXiv preprint arXiv:2106.08352*, 2021.

Hannah Muckenhirn, Vinayak Abrol, Mathew Magimai-Doss, and Sébastien Marcel. Understanding and visualizing raw waveform-based cnns. In *Interspeech*, pages 2345–2349, 2019.

Luke Muehlhauser. What should we learn from past ai forecasts. *Open Philanthropy Project*, 2016.

Eliya Nachmani and Lior Wolf. Unsupervised polyglot text-to-speech. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7055–7059, 2019.

Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. A review of deep learning based speech synthesis. *Applied Sciences*, 9(19):4050, 2019.

Caglar Oflazoglu and Serdar Yildirim. Recognizing emotion from turkish speech using acoustic features. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1):1–11, 2013.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

Kyubyong Park and Thomas Mulc. Css10: A collection of single speaker speech datasets for 10 languages. *arXiv preprint arXiv:1903.11269*, 2019.

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J Han, Shinji Watanabe, and Shrikanth Narayanan. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317, 2022.

Dipjyoti Paul, Yannis Pantazis, and Yannis Stylianou. Speaker Conditional WaveRNN: Towards Universal Neural Vocoder for Unseen Speaker and Recording Conditions. In *Proc. Interspeech 2020*, pages 235–239, 2020.

M. Kathleen Pichora-Fuller and Kate Dupuis. Toronto emotional speech set (TESS), 2020. URL https://doi.org/10.5683/SP2/E8H2MF.

Wei Ping, Kainan Peng, and Jitong Chen. Clarinet: Parallel wave generation in end-to-end text-to-speech. In *International Conference on Learning Representations*, 2018a.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: 2000-speaker neural text-to-speech. In *Proceedings of ICLR*, pages 214–217, 2018b.

Tim Polzehl, Sebastian Möller, and Florian Metze. Automatically assessing personality from speech. In *2010 IEEE Fourth International Conference on Semantic Computing*, pages 134–140. IEEE, 2010.

Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

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

María Luisa Barragán Pulido, Jesús Bernardino Alonso Hernández, Miguel Ángel Ferrer Ballester, Carlos Manuel Travieso González, Jiří Mekyska, and Zdeněk Smékal. Alzheimer's disease and automatic speech analysis: a review. *Expert systems with applications*, 150:113213, 2020.

Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, and Xiang-Yang Li. Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, pages 82–94, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

ITUT Rec. P. 800.1, mean opinion score (mos) terminology. *International Telecommunication Union, Geneva*, 2006.

General Data Protection Regulation. Regulation eu 2016/679 of the european parliament and of the council of 27 april 2016. *Official Journal of the European Union*, 2016.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *arXiv preprint arXiv:1905.09263*, 2019b.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

Slobodan Ribaric, Aladdin Ariyaeeinia, and Nikola Pavesic. De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication*, 47:131–151, 2016.

Daniel Rigney. *The Matthew effect: How advantage begets further advantage*. Columbia University Press, 2010.

Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013.

Morgane Riviere, Jade Copet, and Gabriel Synnaeve. Asr4real: An extended benchmark for speech models. *arXiv preprint arXiv:2110.08583*, 2021.

Philipp V Rouast, Marc Adam, and Raymond Chiong. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing*, 2019.

Nicholas Ruiz, Mattia Antonino Di Gangi, Nicola Bertoldi, and Marcello Federico. Assessing the tolerance of neural machine translation systems against speech recognition errors. *arXiv preprint arXiv:1904.10997*, 2019.

Ralf Schlüter. Survey Talk: Modeling in Automatic Speech Recognition: Beyond Hidden Markov Models. In *Proc. Interspeech 2019*, 2019.

Marc Schröder. Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. In *Tutorial and research workshop on affective dialogue systems*, pages 209–220. Springer, 2004.

Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech communication*, 53 (9-10):1062–1087, 2011.

Björn Schuller, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, Georgios Rizos, Maximilian Schmitt, Lukas Stappen, Harald Baumeister, Alexis Deighton MacIntyre, and Simone Hantke. The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks. In *INTERSPEECH*, 2020.

Björn W Schuller, Anton Batliner, Christian Bergler, Cecilia Mascolo, Jing Han, Iulia Lefter, Heysem Kaya, Shahin Amiriparian, Alice Baird, Lukas Stappen, et al. The interspeech 2021 computational paralinguistics challenge: Covid-19 cough, covid-19 speech, escalation & primates. *arXiv preprint arXiv:2102.13468*, 2021.

Frank Seide, Gang Li, and Dong Yu. Conversational speech transcription using context-dependent deep neural networks. In *Twelfth annual conference of the international speech communication association*, 2011.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, 2016.

Benjamin Sertolli, Zhao Ren, Björn W Schuller, and Nicholas Cummins. Representation transfer learning from deep end-to-end speech recognition networks for the classification of health states from speech. *Computer Speech & Language*, 68:101204, 2021.

Zengqiang Shang, Zhihua Huang, Haozhe Zhang, Pengyuan Zhang, and Yonghong Yan. Incorporating Cross-Speaker Style Transfer for Multi-Language Text-to-Speech. In *Proc. Interspeech 2021*, pages 1619–1623, 2021.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

Jonathan Shen, Ye Jia, Mike Chrzanowski, Yu Zhang, Isaac Elias, Heiga Zen, and Yonghui Wu. Non-attentive tacotron: Robust and controllable neural tts synthesis including unsupervised duration modeling. *arXiv preprint arXiv:2010.04301*, 2020.

Olympia Simantiraki, Martin Cooke, and Simon King. Impact of Different Speech Types on Listening Effort. In *Proc. Interspeech 2018*, pages 2267–2271, 2018.

Marcin Skowron, Mathias Theunis, Stefan Rank, and Arvid Kappas. Affect and social processes in online communication–experiments with an affective dialog system. *IEEE Transactions on Affective Computing*, 4(3):267–279, 2013.

Lilli Smal, Andrea Lösch, Josef van Genabith, Maria Giagkou, Thierry Declerck, and Stephan Busemann. Language data sharing in european public services–overcoming obstacles and creating sustainable data sharing infrastructures. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3443–3448, 2020.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. Char2wav: End-to-end speech synthesis. In *Proceedings of 5th International Conference on Learning Representations*, pages 1–6, 2017.

Brij Mohan Lal Srivastava, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent. Privacy-preserving adversarial representation learning in asr: Reality or illusion? In *INTERSPEECH 2019-20th Annual Conference of the International Speech Communication Association*, 2019.

Brij Mohan Lal Srivastava, Nathalie Vauquier, Md Sahidullah, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent. Evaluating voice conversion-based privacy protection against informed attackers. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2802–2806. IEEE, 2020.

Titus Stahl. Indiscriminate mass surveillance and the public sphere. *Ethics and Information Technology*, 18(1):33–39, 2016.

Martin Sundermeyer, Hermann Ney, and Ralf Schlüter. From feedforward to recurrent lstm neural networks for language modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):517–529, 2015.

Monorama Swain, Aurobinda Routray, and Prithviraj Kabisatpathy. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1):93–120, 2018.

Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. End-to-end asr: from supervised to semi-supervised learning with modern architectures. *arXiv preprint arXiv:1911.08460*, 2019.

Éva Székely, Gustav Eje Henter, Jonas Beskow, and Joakim Gustafson. Spontaneous conversational speech synthesis from found data. In *INTERSPEECH*, pages 4435–4439, 2019.

Dengke Tang, Junlin Zeng, and Ming Li. An end-to-end deep learning framework for speech emotion recognition of atypical individuals. In *Interspeech*, pages 162–166, 2018.

Ashish Tawari and Mohan M Trivedi. Speech emotion analysis in noisy real-world environment. In *2010 20th International Conference on Pattern Recognition*, pages 4605–4608. IEEE, 2010.

Jason Taylor and Korin Richmond. Confidence Intervals for ASR-Based TTS Evaluation. In *Proc. Interspeech 2021*, pages 2791–2795, 2021.

Katrin Tomanek, Françoise Beaufays, Julie Cattiau, Angad Chandorkar, and Khe Chai Sim. On-device personalization of automatic speech recognition models for disordered speech. *arXiv preprint arXiv:2106.10259*, 2021.

Natalia Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, et al. The voiceprivacy 2020 challenge evaluation plan, 2020.

László Tóth, Ildikó Hoffmann, Gábor Gosztolya, Veronika Vincze, Gréta Szatlóczki, Zoltán Bánréti, Magdolna Pákáski, and János Kálmán. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Current Alzheimer Research*, 15(2):130–138, 2018.

Ermal Toto, ML Tlachac, and Elke A Rundensteiner. Audibert: A deep transfer learning multimodal classification framework for depression screening. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4145–4154, 2021.

Jörgen Valk and Tanel Alumäe. Voxlingua107: a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658. IEEE, 2021.

Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro. Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6189–6193. IEEE, 2020a.

Rafael Valle, Kevin Shih, Ryan Prenger, and Bryan Catanzaro. Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis. *arXiv preprint arXiv:2005.05957*, 2020b.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. In *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, page 125, 2016.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. From masked language modeling to translation: Non-english auxiliary tasks improve zero-shot spoken language understanding. *arXiv preprint arXiv:2105.07316*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Ville Vestman, Tomi Kinnunen, Rosa González Hautamäki, and Md Sahidullah. Voice mimicry attacks assisted by automatic speaker verification. *Computer Speech & Language*, 59:36–54, 2020.

Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2011.

Petra Wagner, Jonas Beskow, Simon Betz, Jens Edlund, Joakim Gustafson, Gustav Eje Henter, Sébastien Le Maguer, Zofia Malisz, Éva Székely, Christina Tånnander, and Jana Voße. Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program. In *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, pages 105–110, 2019.

Electra Wallington, Benji Kershenbaum, Peter Bell, and Ondřej Klejch. On the learning dynamics of semi-supervised training for asr. In *Interspeech 2021: The 22nd Annual Conference of the International Speech Communication Association*, pages 716–720. International Speech Communication Association, 2021.

Jianyou Wang, Michael Xue, Ryan Culhane, Enmao Diao, Jie Ding, and Vahid Tarokh. Speech emotion recognition with dual-sequence lstm architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6474–6478. IEEE, 2020.

Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech 2017*, pages 4006–4010, 2017.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, pages 5180–5189. PMLR, 2018.

Elisabeth Wehling. *Politisches Framing: Wie eine Nation sich ihr Denken einredet - und daraus Politik macht*. Ullstein Ebooks, 2018. ISBN 9783843718578. URL https://books.google.at/books?id=tlFaDwAAQBAJ.

Mirjam Wester, Cassia Valentini-Botinhao, and Gustav Eje Henter. Are we using enough listeners? no! — an empirically-supported critique of interspeech 2014 TTS evaluations. In *Proc. Interspeech 2015*, pages 3476–3480, 2015.

Mika Westerlund, Diane A Isabelle, and Seppo Leminen. The acceptance of digital surveillance in an age of big data. *Technology Innovation Management Review*, 11(3), 2021.

Terry Winograd. Shifting viewpoints: Artificial intelligence and human-computer interaction. *Artif. Intell.*, 170:1256–1258, 2006.

Yunhan Wu, Daniel Rough, Anna Bleakley, Justin Edwards, Orla Cooney, Philip R Doyle, Leigh Clark, and Benjamin R Cowan. See what i'm saying? comparing intelligent personal assistant use for native and non-native language speakers. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–9, 2020.

Detai Xin, Yuki Saito, Shinnosuke Takamichi, Tomoki Koriyama, and Hiroshi Saruwatari. Cross-Lingual Speaker Adaptation Using Domain Adaptation and Speaker Consistency Loss for Text-To-Speech Synthesis. In *Proc. Interspeech 2021*, pages 1614–1618, 2021.

Jinhyeok Yang, Jae-Sung Bae, Taejun Bak, Young-Ik Kim, and Hoon-Young Cho. GANSpeech: Adversarial Training for High-Fidelity Multi-Speaker Speech Synthesis. In *Proc. Interspeech 2021*, pages 2202–2206, 2021.

Shunzhi Yang, Zheng Gong, Kai Ye, Yungen Wei, Zhenhua Huang, and Zheng Huang. Edgernn: a compact speech recognition network with spatio-temporal features for edge computing. *IEEE Access*, 8: 81468–81478, 2020.

Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi Lei, Dan Su, and Dong Yu. DurIAN: Duration Informed Attention Network for Speech Synthesis. In *Proc. Interspeech 2020*, pages 2027–2031, 2020.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A corpus derived from LibriSpeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

Albert Zeyer, Parnia Bahar, Kazuki Irie, Ralf Schlüter, and Hermann Ney. A comparison of transformer and lstm encoder decoder models for asr. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 8–15. IEEE, 2019.

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949. IEEE, 2019a.

Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, R.J. Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. In *Proc. Interspeech 2019*, pages 2080–2084, 2019b.

Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, R.J. Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. In *Proc. Interspeech 2019*, pages 2080–2084, 2019c.

Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*, 2020.

Jianfeng Zhao, Xia Mao, and Lijiang Chen. Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical Signal Processing and Control*, 47:312–323, 2019.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Xuehao Zhou, Xiaohai Tian, Grandee Lee, Rohan Kumar Das, and Haizhou Li. End-to-end code-switching tts with cross-lingual language model. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7614–7618, 2020.

Xiaolian Zhu, Yuchao Zhang, Shan Yang, Liumeng Xue, and Lei Xie. Pre-alignment guided attention for improving training efficiency and model stability in end-to-end speech synthesis. *IEEE Access*, 7: 65955–65964, 2019.