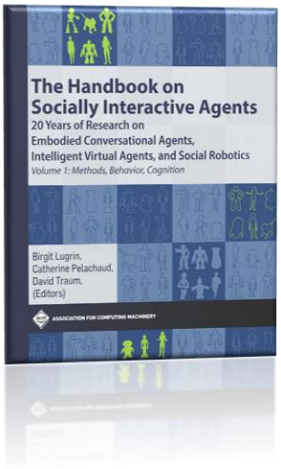
Building and Designing Expressive Speech Synthesis

Matthew P. Aylett, Leigh Clark, Benjamin R. Cowan, and Ilaria Torre



Author note:

This is a preprint. The final article is published in "The Handbook on Socially Interactive Agents" by ACM books.

Citation information:

Aylett, M.P., Clark, L., Cowan, B.R., and Torre, I., (2021). Building and Designing Expressive Speech Synthesis. In B. Lugrin, C. Pelachaud, D. Traum (Eds.), *Handbook on Socially Interactive Agents – 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics*, Volume 1: Methods, Behavior, Cognition (pp. 173-211). ACM.

DOI of the final chapter: 10.1145/3477322.3477329

DOI of volume 1 of the handbook: 10.1145/3477322

Correspondence concerning this chapter should be addressed to Matthew P. Aylett (matthewaylett@gmail.com), Leigh Clark (leigh.clark@ucd.ie), Benjamin R. Cowan (benjamin.cowan@ucd.ie), or Ilaria Torre (ilariat@kth.se)

6 Building and Designing Expressive Speech Synthesis

Matthew P. Aylett, Leigh Clark, Benjamin R. Cowan and Ilaria Torre

"You all know the test for artificial intelligence – the Turing test. A human judge has a conversation with a human and a computer. If the judge can't tell the machine apart from the human, the machine has passed the test. I now propose a test for computer voices – the Ebert test. If a computer voice can successfully tell a joke and do the timing and delivery as well as Henny Youngman, then that's the voice I want." — Roger Ebert.

6.1 Introduction & Motivation

We know there is something special about speech. Our voices are not just a means of communicating. They also give a deep impression of who we are and what we might know. They can betray our upbringing, our emotional state, our state of health. They can be used to persuade and convince, to calm and to excite.

As speech systems enter the social domain they are required to interact, support and mediate our social relationships with 1) each other, 2) with digital information, and, increasingly, 3) with AI-based algorithms and processes. Socially Interactive Agents (SIAs) are at the forefront of research and innovation in this area. There is an assumption that in the future *"spoken language will provide a natural conversational interface between human beings and so-called intelligent systems."* [Moore 2017, p. 283]. A considerable amount of previous research work has tested this assumption with mixed results. However, as pointed out *"voice interfaces have become notorious for fostering frustration and failure"* [Nass and Brave 2005, p.6].

It is within this context, between our exceptional and intelligent human use of speech to communicate and interact with other humans, and our desire to leverage this means of communication for artificial systems, that the technology, often termed *expressive speech synthesis* uncomfortably falls. Uncomfortably, because it is often overshadowed by issues in interactivity and the underlying intelligence of the system which is something that emerges from the interaction of many of the components in a SIA. This is especially true of what we might term conversational speech, where decoupling *how* things are spoken, from *when* and

to *whom* they are spoken, can seem an impossible task. This is an even greater challenge in evaluation and in characterising full systems which have made use of expressive speech.

Furthermore when designing an interaction with a SIA, we must not only consider how SIAs should speak but how much, and whether they should even speak at all. These considerations cannot be ignored. Any speech synthesis that is used in the context of an artificial agent will have a perceived accent, a vocal style, an underlying emotion and an intonational model. Dimensions like accent and personality (cross speaker parameters) as well as vocal style, emotion and intonation during an interaction (within-speaker parameters) need to be built in the design of a synthetic voice. Even a *default* or *neutral* voice has to consider these same expressive speech synthesis components. Such design parameters have a strong influence on how effectively a system will interact, how it is perceived and its assumed ability to perform a task or function. To ignore these is to blindly accept a set of design decisions that ignores the complex effect speech has on the user's successful interaction with a system. Thus expressive speech synthesis is a key design component in SIAs.

This chapter explores the world of expressive speech synthesis, aiming to act as a starting point for those interested in the design, building and evaluation of such artificial speech. The debates and literature within this topic are vast and are fundamentally multidisciplinary in focus, covering a wide range of disciplines such as linguistics, pragmatics, psychology, speech and language technology, robotics and human-computer interaction (HCI), to name a few. It is not our aim to synthesise these areas but to give a scaffold and a starting point for the reader by exploring the critical dimensions and decisions they may need to consider when choosing to use expressive speech. To do this, the chapter explores the building of expressive synthesis, highlighting key decisions and parameters as well as emphasising future challenges in expressive speech research and development. Yet, before these are expanded upon we must first try and define what we actually mean by expressive speech.

6.2 Expressive Speech – a working definition

The term expressive speech can be used for many features that we see in human speech. It is a problematic term because it can rely on the concept of neutral speech which assumes there is such a thing as *un-expressive* speech. This view is often one used in speech technology because implicitly, from an engineering perspective, it is an issue of control. The concept of neutral speech helps to frame synthesis development parameters and challenges. For instance, if a speech synthesis system produces so called *default speech* for a series of words, how might we alter the emotion or the emphasis perceived in this speech? What controls would we need? What changes to the synthesis system would be required? This is echoed in the definition used by Govind et al. *"In expressive speech synthesis, along with text, the desired expression also forms an additional input to the text processing stage"* [Govind and Prasanna 2013, p. 237].

Pitrelli et al. [Pitrelli et al. 2006, p. 1099] define expressive speech as a system's ability to *"...reinforce the message with paralinguistic cues, to communicate moods and other content beyond the text itself"*. However, the term *other content* is open-ended, making it hard to pin down what exactly the term expressive speech should or should not cover. They go on to use a more concrete working definition *"say the same words differently and appropriately for different situations"* (p.1099).

Influenced by early work on emotional speech synthesis [e.g. Schröder 2001], the term expressive speech has also been used synonymously with the term emotional speech. For example, according to Schröder, the purpose of expressive synthesis is in*"...making a voice sound happy or subdued, friendly or empathic, authoritative or uncertain..."* [Schröder 2009, p. 111]. This emotional emphasis is further echoed by Govind and Prassana: *"we have considered different emotions as the expressions and hence emotions and expressions are interchangeably used."* [Govind and Prasanna 2013, p. 237]. This viewpoint on expressive synthesis focused on developing and evaluating a distinct set of between three and nine, extreme, sometimes termed *primitive* emotional states within an artificial voice, such as disgust, fear, anger, joy, sadness, and surprise [Schröder 2001]. Yet, as computer applications began to take on roles such as of a trainer or tutor, or attempted to offer emotional support for users, a more nuanced view of emotion in speech was required [Aylett et al. 2013, Theune et al. 2006].

As applications such as personal digital assistants and other forms of SIAs become more commercially deployed, we see the definition of expressive speech also becoming an important part of defining the perceived personality of the system [Aylett et al. 2017], becoming more dependent on interaction. Expressive speech *"has the potential to provide the user with the choice to select a nuanced tone of voice suited to their intent and to the communicative setting"* but that *"In an interactive situation however, this does not become a real possibility, until a functional interaction model is available to control aspects of the expressive synthetic speech to ensure timely and effortless delivery"* [Székely 2015, p. 4]. This perspective echoes advice from Pitrelli [2006]. We would add that to this, the design of a voice to convey an intended character and emotional state, and to widen *"different situations"* to include interaction situations such as conversation, monologue and even dramatic performance [see Aylett et al. 2019a].

The definition of expressive synthesis is also hard to decouple from the aim of improving the naturalness of synthetic speech. Recent work in *neural* or *Wavenet-style* TTS [e.g. Oord et al. 2016] has dramatically improved the perceived naturalness of synthetic speech. Improved to the extent that Google were able create a system to book appointments (i.e. the hairdresser) that, in an interactive phone call, could not be perceived as artificial[1]. Although this is important to consider when conceptualising expressive synthesis, a blind drive for nat-

[1] https://www.youtube.com/watch?v=D5VN56jQMWM

uralness can eclipse a requirement for appropriateness needed for artificial systems [Aylett et al. 2019a,b]. Both the drive for naturalness and the appropriateness of such naturalness must therefore be considered.

6.3 Building Expressive Synthesis

Models and approaches in expressive speech synthesis have been heavily constrained by the historical techniques used to generate speech. Recently there has been a paradigm shift in speech synthesis technology where recurrent neural networks have successfully been applied to model speech output with unprecedented quality [e.g. Lorenzo-Trueba et al. 2018, Oord et al. 2016, Ping et al. 2017].

There is often a lag between the state-of-the-art synthesis systems available and systems used by SIAs. Availability, cost, performance, in-house technical experience and language coverage are important factors when choosing a system. However, new paradigm neural Text-To-Speech (TTS) systems are available commercially and there are extensive open source repositories. Despite this, many systems will still be operating with unit selection techniques. These systems can also offer very good performance and quality and in many cases could be appropriate for SIAs. Finally, although the speech quality of older types of systems is significantly worse, research systems built for low resourced languages or low cost commercial systems such as toys may still use diphone and formant synthesis.

6.3.1 Overview of Major Approaches to Speech Synthesis

Most systems assume text is the input to the system, possibly with XML markup (see section 6.3.3), and that an audio waveform is the final output, together with information on phone, word and mark-up timings that can be used to drive a SIA's speech. Conventionally, the text is processed by a series of modules. The first set (often termed the front-end) extracts linguistic and pronunciation information. The second set (often termed the back-end) then converts this information into audio. Of particular importance for the control of expressive speech are the linguistic specifications of the front-end that determines stress, emphasis and phrasing – the modules which convert these specifications into duration and pitch parameters in the back-end. Below, we describe the different major approaches to speech synthesis.

Formant Synthesis: Rule-based formant synthesis is one of the oldest approaches to speech synthesis (e.g. [Klatt 1980]). A set of rules dictate parameters for 3 or 4 formants. Formants are the key resonant frequencies of the vocal tract and contain information on whether a pulse (voiced) or white noise (un-voiced) source is fed through the resulting filter. Stephen Hawking used a formant synthesiser and its sound became synonymous with his vocal identity. Although the quality is poor, such systems can easily be pre-built onto a chip and be used in toys and games.

Diphone Synthesis: Diphone synthesis depends on a voice talent recording examples of every diphone (one phone sound transitioning into another phone sound) within a carrier phrase and this set forming the basis of output speech (e.g. [Lenzo and Black 2000]). Here, a complex prosodic model based on front-end features would predict the duration and pitch parameters of each phone. Digital signal processing would alter the pitch and duration of each diphone to achieve these targets and then concatenate the diphones together (often with a smoothing algorithm) to create the output speech.

Unit Selection: Unit selection extends diphone synthesis by using hundreds of thousands of diphones (or other speech units) extracted from a very large corpus recorded from a single speaker [Hunt and Black 1996]. Linguistic features and prosodic targets are used in a dynamic programming algorithm to select a good sequence of units. These are concatenated together with both smoothing and some processing to control duration and pitch. This produced a major jump in quality and many systems still use this approach.

Parametric Synthesis with a Traditional Vocoder: Within this approach, machine learning in the form of Hidden Markov Models (HMMs) [Tokuda et al. 2000] or neural networks (e.g. [Zen et al. 2013]) trained on linguistics features are used instead of rules to control the spectral and prosodic parameters of the output speech. These parameters are then converted into a waveform using a vocoder – a digital signal processing algorithm which can change spectral specifications into a waveform. The use of machine learning allows voices to be modelled, and these models can be adapted and manipulated. This in turn means that the systems offer significant flexibility not present in unit selection synthesis. However, the modelling and vocoding process reduces the perceived quality and naturalness of the speech output.

Neural TTS: By replacing the vocoding algorithm with a recurrent multi-level neural net, these systems are able to significantly improve the quality of parametric-based systems to the extent that the quality exceeds output from unit selection systems (e.g. [Oord et al. 2016]). In some cases the linguistic front-end can also be released by sequence-to-sequence neural net models [Wang et al. 2017].

Work on expressive speech exists for all these types of systems, though diphone, formant, and traditionally vocoded systems are now generally regarded as having unacceptably low perceived naturalness for most purposes. How much expressive speech control is available will also depend on whether the system is a research system, an off-the-shelf commercial system, or a bespoke system (depending on voice recordings made specifically for the system). Open source speech synthesis systems are notoriously labour intensive in terms of building bespoke voices, ensuring sufficient quality, and integrating with a full system (e.g. SIAs). They can also present significant challenges in terms of performance. In contrast, commercial systems vary extensively in terms of price, flexibility, and amount of expressiveness control they offer.

6.3.2 The Importance of Corpora

In general, the speech corpus used to create a voice will have a direct impact on the expressive variation available in the voice. Both unit selection and neural TTS systems are corpus-based. For unit selection the corpus underlying the voice has a rather rigid effect on the voice produced, as only fragments of the recorded corpus are used to generate new speech output. For neural TTS it is easier to generalise across the corpus and to also generalise across speaker, accent, and even language. Open source speech synthesis systems will typically be based on corpora that are available to the academic community. Open source corpora include Arctic [Kominek and Black 2004], LJ Speech[2] VCTK [Yamagishi et al. 2019], LibriTTS [Zen et al. 2019] and LADS [Braude et al. 2019]. Other corpora are also available but may not be licensed for commercial use. Services such as ELDA[3] and LDC[4] also have corpora but currently most are captured to support automatic speech recognition (ASR) research and development. Note that corpora captured for ASR are not typically appropriate for synthesis. Having background noise, speech noises, disfluencies and very casual speech are an expected challenge in processing speech for recognition. Therefore ASR corpora tend to include these effects. For speech synthesis, where clear articulate speech is normally desired, such effects can cause significant problems for any synthesis generation technique.

Recording a bespoke corpus for a voice is resource intensive and time consuming. Unit selection systems will require upwards of five hours of data which could take over 30 hours to record. For neural TTS approaches adaptive voice modelling can be used to reduce the data required. The amount required is changing as modelling approaches become more sophisticated, but at least one hour of recorded data is a sensible starting point. The make up of a corpus can affect the amount of data required and will also affect the flexibility of any voice based on it. However, when designing for a SIA, a bespoke corpus can have significant benefits. Custom voice styles and accents can be chosen to fit the form and functionality of the target system, adding the required expressive speech to the corpus. For a commercial system, creating a custom voice also has the benefit of unique branding.

As we discuss the models, approaches and parameters to expressive speech synthesis below, we will reflect on the corpus requirements for specific functionalities. Often using a custom corpus will be determined by the resources and access to speech synthesis expertise available to a SIA project. Off-the-shelf systems may not offer the desired functionality, and an informed decision is required as to how to proceed with any given project.

6.3.3 Controlling Expressiveness – Speech Synthesis Mark-up

The dominant method for controlling expressive functionality in speech synthesis is through the use of XML mark-up. The advantage of XML-based systems is that different synthesis

[2] https://keithito.com/LJ-Speech-Dataset/

[3] http://catalogue.elra.info/en-us/

[4] https://www.ldc.upenn.edu/

systems can implement a sub-set of functionality, or extend functionality, without damaging output (i.e. reading out the mark-up instead of the content). SSML[5] [Taylor and Isard 1997] is the dominant standard, although Microsoft offer their own XML mark-up. Apple's *embedded speech commands* do not use XML and are not compatible across non-Apple platforms. Below is an example of SSML markup applied to alter the word emphasis of some synthetic speech.

```
<speak>
     I already told you I <emphasis level="strong">really
     like</emphasis> that person.
</speak>
```

SSML tags can control prosodic elements of speech such as speech rate, pitch and phrasing. It can also control lexical elements such as pronunciation of words using the International Phonetic Alphabet (IPA). It can even be used to insert pre-recorded audio into speech, control the interpretation of symbols and digits, and be used to mark word boundaries so timing can be extracted after speech is generated to synchronise animation.

However, there are a number of issues regarding SSML mark-up when used to control expressive synthetic speech for SIAs:

1. The standard is 10 years old and was designed without any knowledge of neural TTS approaches; also, it was historically designed before unit selection became mainstream. This leads to some problems in interpreting the functionality of some tags in speech synthesis systems.
2. The prosodic control is intentionally simplistic and does not specify how these controls should be implemented. For example, if you change emphasis, what should happen to previously emphasised words? How should emphasis be created in terms of pitch, duration, energy and spectral characteristics? What sort of intonation should occur at prosodic breaks?
3. Tags can be easily misused. For example the tag:

   ```
   <prosody duration="2s">hello there</prosody>
   ```

   will attempt to say "hello there" in exactly two seconds. If implemented, this will produce very unnatural slow speech. To change speech rate the markup below would be more appropriate:

   ```
   <prosody rate="slow">hello there</prosody>
   ```

[5] https://www.w3.org/TR/speech-synthesis11

4. When switching to a different synthesis system, the way a tag is implemented, and if it is implemented, may vary widely, meaning that the text may not be synthesised correctly.

Nonetheless, SSML is a very useful standard and, given the very big changes that have occurred in speech synthesis technology, it has remained extremely useful both as a practical mark-up system and as a reference for the sort of control that would be available. Part of the strength of SSML is its lack of concrete specification, giving engineers the flexibility to implement tags appropriately in different synthesis systems.

Commercial systems have long developed internal tag sets to offer a more specified set of controls, as well as controls not covered by SSML, such as variation in emotion and speech style. For example, Alexa offers a set of *amazon:* tags that can be used with different voices.

To conclude, understanding standard mark-up is important when designing the voice for a SIA, but making a design decision to use a specific synthesis system and its internal mark-up, given the under-specification of SSML, is also a valid strategy.

6.3.4 Cross and within speaker variation

When considering models, approaches and design choices in expressive speech synthesis, it is useful to distinguish between cross speaker features and within speaker features that can be given to the voice. In general, cross speaker features are the basis of speaker identity and perceived personality and can be grouped in terms of language, accent, dialect and vocal style. Within speaker features are changes across a sentence such as emotion, emphasis and conversational control.

In modern corpus-based speech synthesis, cross speaker features are very heavily dependent on the corpora used to build the voice. For single speaker corpora, the recorded material will completely dictate these features. For multiple speaker corpora, they can potentially be constructed drawing features from different sources. In general, cross speaker features would remain constant throughout an interaction, and normally across whole utterances.

In contrast, within speaker features are primarily used to convey the inner state of the speaking agent by altering the way an utterance is spoken. Together with text, movement or animation, expressive speech synthesis can convey a change in emotion, empathy, focus, and appropriately match the interactional context. It is here that speech synthesis XML mark-up (described in section 6.3.3) becomes extremely important. Ideally, a language system could automatically insert such mark-up to control the speech synthesis.

There is no clear-cut separation between cross and within speaker features. For example, a multilingual system may switch languages within a sentence, or different voice styles may be used across an utterance to render complex emotions. It is rather the pragmatic design of the system that creates the distinction. Many voices will only have one language, one accent, one dialect and one voice style. After corpora are recorded and voices are built, these features cannot be altered. Many systems will offer a limited set of word by word XML controls that in

turn dictate the within speaker features that can be applied. When designing a speech system for a SIA it is important to determine what flexibility will be required, whether the voice can produce such flexibility and whether the TTS system can support it.

Over the next few years, as interactive systems make greater use of the expressive functionality on speech synthesis, we will see a further blurring of the cross and within speaker feature boundary. Just as human speakers can mimic other speakers, alter voice style at will, and code switch between multiple languages, we will see future TTS systems offering such functionalities. However, this is an important distinction to understand the challenges of the current design process, and we will discuss expressive speech design within this context.

6.3.5 Designing Expressive Speech – Cross Speaker Features

6.3.5.1 Language

The language a synthetic voice speaks (or is required to speak) has wide ranging effects on both the quality and expressive functionality of the system. Despite many years of linguistic research exploring cross-linguistic features of expressive speech [e.g. Ohala 1983], within speech synthesis the vast majority of systems are built on language specific data. Although this is changing with recent work exploring the use of neural networks to generalise across languages [e.g. Zhang et al. 2019], from a practical perspective the language(s) a SIA is required to speak will have a key impact on the expressive speech functionality that is available. Research in expressive speech is historically Anglo-centric. For example, in emotion recognition research itr has been pointed out that collected databases are *"dominated by English language"* and *"very few databases are collected in languages such as: Russian, Dutch, Slovenian, Swedish, Japanese and Spanish. There is no reported reference of an emotional speech database in any of the Indian languages"* [Koolagudi and Rao 2012, pp. 103-104]. Speech synthesis mark-up such as SSML is also Anglo-centric and not all tags will transfer appropriately across languages. In addition, research exploring and controlling expressive speech synthesis across languages is rare.

Multilingualism, where a speaker can use more than one language, is believed to be more common than monolingualism [Tucker 1999]. The use of code switching, where a speaker switches language during speech, is often cultural, expressive, and intimately connected to social relationships [e.g. Paugh 2005]. Speech synthesis systems normally have very limited multi-language support. This can present a serious problem in systems where foreign words are often used. How an English word should be pronounced in a Spanish system is different from how it should be pronounced in a German system.

Traditional monolingual speech synthesis systems use what is often termed a *front-end* or *G2P system* to take a series of marked up words and convert them into a series of phones (the basic sounds that make up a language), and other linguistic features such as syllable structure, stress, and phrasing. Such front-ends depend on a large body of manually constructed data with the largest component comprising of a pronunciation dictionary. Such a dictionary will

have an entry for every word in all forms, describing its pronunciation in terms of phones and often stress,(e.g. the CMU pronunciation dictionary[6]). Depending on language there may also be modules for, amongst others, letter to sound generation for out of vocabulary words, normalisation of symbols and digits, homograph disambiguation, stress and phrasing modelling, part of speech tagging, word disambiguation, archiphones (where phone sounds alter in sentence context), and more.

More recently, significant work has explored generating synthesis using sequence-to-sequence neural network models which go directly from text to speech without any front-end processing. The most notable example of this work is Tacotron [Wang et al. 2017]. Although such end-to-end synthesis is attractive (especially for less resourced languages), from the perspective of expressive speech it raises the question of how expression can be controlled. For example, if the end-to-end system has no overt model of stress, how might we control emphasis? Recent work [Hodari et al. 2019, Skerry-Ryan et al. 2018] explores this issue and proposes various solutions. However, from a practical perspective, when using an end-to-end speech synthesis system for a SIA it is important to be aware of what expressive speech control is available, if any, as well as the ability to correct errors in the synthesis produced.

6.3.5.2 Accent and Dialect

Accent can be defined as *"The cumulative auditory effect of those features of pronunciation which identify where a person is from, regionally or socially."* [Crystal 1997, p. 2]. It acts as a clue to a speaker's social identity, socio-linguistic background and geographical origin [Crystal 2011, Ikeno and Hansen 2007], strongly influencing perceptions of a speaker, eliciting specific stereotypes and assumptions associated with a particular accent [Coupland and Bishop 2007, Ryan and Giles 1982]. Accent differs from dialect in that it is concerned with pronunciation. Conversely, dialect refers to language patterns, such as grammar and vocabulary that we associate with particular geographical regions or social groups [Hughes et al. 2013]. Like accent, dialect terms used by people can signal cues to their identity.

It is possible to model multi-accent voices and use accent shift as an expressive synthesis technique. If the accents can be broadly described by the same phone set then this can be implemented using a voice style (see 1.3.5.3). If not, techniques used for multi-language systems are required (see section 6.3.5.4).

The decision to use accents in synthesis needs to be made with full awareness of the social context in which the voice will be used. For instance, research has shown that people tend to prefer interacting with agents that use standard-accented speech due to their supposed prestigious status [Bishop et al. 2005, Torre and Maguer 2020]. Singaporean users rated British-accented speech in a helpdesk agent more positively compared to Singaporean-accented English [Niculescu et al. 2008], in part due to the perception of prestige given by a

[6] http://www.speech.cs.cmu.edu.

British accent within formal interactions within Singaporean society [Niculescu et al. 2008]. Similarly, within the UK, artificial agents speaking with accents traditionally deemed to be prestigious were trusted more than other accents, even when the agents were equally trustworthy [Torre et al. 2015, 2018a]. Andrist et al. [2015] have also found that social status and credibility interact, in a study involving social robots speaking with different Arabic accents. People have also shown a tendency to prefer interacting with agents that speak with similar accents to their own [Cargile and Giles 1997, Kinzler et al. 2011], congruent with in-group membership preferences and *similarity attraction* effects [Dahlbäck et al. 2001].

A system's accent also has ramifications for the dialect choices people are likely to use in interaction. A recent study found that accent significantly impacted the likelihood of using lexical alternatives [Cowan et al. 2019]. When playing a referential communication game where a number of objects could be described using either US or Hiberno-English dialect alternatives (e.g. wrench and spanner), people were more likely to use US English dialect terms in their descriptions when interacting with US-accented system compared to an Irish-accented system [Cowan et al. 2019]. This parameter, as well as influencing perceptions, may therefore also play a significant role in shaping the dialect that users use in interaction.

6.3.5.3 Voice Styles

Sub-corpora, where a speaker is recorded speaking in a certain style, can be used to control synthetic voice styles. Often such sub-corpora will be acted, for example asking a speaker to speak sadly when reading material for recording. But they can also be elicited by asking speakers to read appropriate material, for example upbeat enthusiastic sentences to encourage an upbeat and enthusiastic voice style. In unit selection approaches, voice styles are most effective when recorded with appropriate target material. This is because unit selection systems require very high coverage of phonetic and prosodic contexts, and without recording a very large amount of material the sub-corpora will tend to produce less natural results the more it strays for the text used to record it. For neural TTS systems, by using voice adaptation (see section 6.3.5.4), this problem is minimised and allows the construction of voices with many sub-corpora.

Initially, work on voice styles was carried out in order to synthesise emotions [e.g. Hofer et al. 2005], where Happy, Angry and Neutral sub corpora were used. Later work explored sub corpora based on different voice qualities so that that they could be coupled with emotional synthesis techniques based on changing rate and pitch (see section 6.3.6.1). A stressed (tense) voice quality tends to be rated negatively, while a lax (calm) voice quality tends to be rated positively [Aylett et al. 2013, Potard et al. 2016]. Other examples of work include using sub corpora of conversational speech [Andersson et al. 2010] and recording voice styles specifically for the use cases of an Intelligent Virtual Agent (IVA), such as a motivating voice style for developing a virtual sports coach [Aylett and Braude 2018].

If resources are available to record bespoke audio for a SIA then considering a set of voice styles that will support the intended interaction contexts is a fundamental design step. When choosing an off-the-shelf TTS system, it is important to be aware of what voice styles may be available and how they may be used in a SIA.

6.3.5.4 Voice Adaptation

Recent methods for voice cloning (the creation of a synthetic voice that sounds like a specific source speaker) and voice style depend on the use of voice adaptation techniques. Voice adaptation is a process where speech data collected from other speakers is used to improve a model for a new speaker or speaker with limited data. The technique was initially applied to speech recognition but generalised to synthesis models in pioneering work on parametric synthesis based on Hidden Markov models [Yamagishi et al. 2006]. Hidden Markov models were later replaced with deep neural networks (DNNs). Until the vocoder, the system that converted model parameters into speech, was also replaced by DNNs in neural TTS, the output quality was much lower than unit selection systems and not widely used.

However, neural TTS open up an entire world of possibilities by allowing voice adaptation to be applied to high quality speech synthesis as exampled by recent work [Arik et al. 2018, Bollepalli et al. 2019, Luong and Yamagishi 2020, Prateek et al. 2019, Zhu and Xue 2020]. It is important to note that this field is changing very rapidly. Whether voice adaptation could be used to rapidly generate voice styles, and clone voices for SIAs would depend very much on what neural TTS system is used and to what extent there is collaboration with speech synthesis experts.

6.3.6 Designing Expressive Speech – Within Speaker Parameters

6.3.6.1 Emotional state

The ability to develop speech that can emote has been one of the core pillars of expressive synthesis development. The notion here is that for a voice to be truly natural or human-like, it needs to be able to accurately express emotion, with much work dedicated to understanding what parameters in speech are related to emotion and how these can be applied in speech synthesis [see Gangamohan et al. 2016, Govind and Prasanna 2013, Kamiloğlu et al. 2020, Schröder 2001, Schröder 2009, for reviews]. The manipulation of prosodic features (e.g. Fundamental frequency (F0) contour, level and range, speech tempo, loudness, and voice quality) are key to generating specific emotions [see Govind and Prasanna 2013, Schröder 2001, Schröder 2009, for detailed breakdown across emotional states]. The majority of previous work exploring emotional and expressive speech synthesis has focused on a distinct set of between three and nine, extreme, sometimes termed *Darwinian* emotional states, such as disgust, anger, joy, sadness, and surprise [Schröder 2001]. However as Schröder [2004, p. 211] points out:

> *"In a dialogue, an emotional state may build up rather gradually, and may change over time as the interaction moves on. Consequently, a speech synthesis system should be able to gradually modify the voice in a series of steps towards an emotional state. In addition, it seems reasonable to assume that most human-machine dialogues will require the machine to express only mild, non-extreme emotional states. Therefore, the need to express full blown emotions is a marginal rather than a central requirement, while the main focus should be on the system's capability to express a large variety of emotional states of low to medium intensity."*

Thus, dimensional models of emotion, such as the circumplex model, might be more appropriate to convey emotion in artificial speech [Rubin and Talarico 2009].

More recently there has been a growing understanding that emotional content of synthesis should be dependent on the task and requirements of the system. Given this we would expect our emotional categorisation to be task-dependent and that synthesis researchers and dialogue researchers would work closely together to both specify, design and evaluate the resulting system.

Although task-based schemes for emotional response do exist (for example the framework proposed by Ortony, Clore and Collins [Ortony et al. 1988] or OCC model), the emotional categorisation used in most emotional synthesis research is typically not task-dependent.

Pitch and speech rate contribute strongly to the perception of emotion [Ramakrishnan 2012]. For most commercial speech synthesisers, controlling speech rate and pitch is normally possible using SSML or bespoke XML markup. For example, raising the pitch and increasing the speaking rate to convey cheerful enthusiasm and lowering both to convey a sense of sadness. However pitch and rate features are strongly related to other speech features such as voice quality, spectral tilt and prosodic and phonetic context. Thus, although some variation can be generated using pitch and rate change, it may often sound unnatural especially if modified by more than 10%.

An alternative is to record a sub-corpus of speech acted with a specific emotion. This is similar to using the cross speaker approach of voice styles (see sections 6.3.5.3 and 6.3.2). This sub-corpus can then be used either as a data for unit selection or to model the emotion. Speech output based on the sub-corpus can then be inserted into the speech stream when a change is desired. However, being able to transition between such voice styles is a challenge. Voice adaptation techniques (see section 6.3.5.4) can potentially be used to produce a graded effect [i.e. Zhu and Xue 2020] but how to move gracefully between emotions is still a key issue.

As mentioned, when considering the use of emotional synthesis, it is important to note that emotional expression is not just relevant to the speech signal itself. The perception of emotion is hugely influenced by the context, speaker and speech content (e.g. syntax and lexicon) [Erickson 2005]. When creating expressive speech, as much consideration needs to be paid

to linguistic content, sociolinguistic factors and their interplay with prosodic factors within the speech signal. Expression of emotion through speech is also multimodal in nature. For a very simple example, consider smiling: the act of pulling the muscles around our lips upwards modifies our vocal tract, which in turn will give a different quality to speech when it is smiled, than when it is 'neutral'. This is reflected in the fact that we can 'hear' smiles, for example when we are on the phone with someone [e.g. Tartter and Braun 1994]. Although they are yet to make their way into commercial systems (along with much of expressive synthetic voices), work is already underway to create TTS systems that can switch from neutral to smiling [El Haddad et al. 2015a,b] and laughing [Sundaram and Narayanan 2007]. While researchers have not agreed on an acoustic definition of a smile yet, it seems that smiling is reflected in our voice by means of prosodic and spectral changes (namely, increased fundamental and formant frequency, and increased spectral centroid) [Arias et al. 2018, El Haddad et al. 2017]. Synthesising smiling and laughing speech is not trivial, because of the lack of one-to-one mappings between acoustic and perceptual features, but it is a promising avenue for research that could have a real impact on agent technologies, as currently there are very few "emotional" voices for artificial agents.

6.3.6.2 Emphasis & Question Intonation

Changing the intonation of an utterance with the desire to alter the speech style or emotion, is typically connected with information content and dialogue context. For example to make an utterance sound like a question, or to change the emphasis on a word.

Many speech synthesis systems will support question intonation to some extent. In unit selection systems, effective question intonation synthesis is hampered by a requirement for sound coverage. Typically pre-recorded question tags, and the use of non final intonation present in non-sentence final utterances can be used to generate question intonation. This has mixed success across languages but is often acceptable especially as the perception of a question is often driven more strongly by lexical content than question prosody.

In neural TTS systems the long term relationship between question prosody can be modelled without requiring full sound coverage. However, question intonation across sentences and speakers is very far from homogeneous making learning such forms from a corpus a challenging task. Research in this area is still at an early stage, with the exploration of different neural net models and feature architectures to improve the variation expected in spoken intonation a source of current research [Kenter et al. 2019, Marelli et al. 2019, Sun et al. 2020].

For emphasis, the issue of a lack of homogeneity in the way speakers emphasise words caused by sound and sentence context and individual speaker styles make modelling emphasis even more challenging that question intonation. Most current work in speech synthesis assumes that if the intonation modelling is good, appropriate emphasis will be created automatically (e.g. research work looking at paragraph intonation to model changes in emphasis across utterances [Aubin et al. 2019]).

Although many commercial systems implement the SSML emphasis tag, the implementation can vary and can have various levels of success. For non-commercial systems, mark-up may not be present at all. For designers and engineers working with SIAs, this common lack of an explicit control of emphasis can be frustrating.

However, if we return to the original phonetics literature [e.g. Cruttenden et al. 1997], we can see that emphasis is typically manifested by extended phonetic duration and pitch excursions. Therefore, if pitch and rate control is available, manipulating duration and pitch directly can often be used to create the perception of emphasis as well as rising question intonation.

6.3.6.3 Conversational Speech

Kawahara [2019] presents the challenges and research objectives for a spoken dialogue system for a human-like conversational Robot ERICA. He points out that *"Speech synthesis should be designed for the conversational style rather than text-reading applications, which are conventional targets of text-to-speech. Moreover, a variety of non-lexical utterances such as back channels, fillers and laughter are needed with a variety of prosody."* [Kawahara 2019, p. 5]. Thus, as well as requiring expressive speech techniques described in earlier sections, interactive conversational speech also requires specialised expressive speech functionality.

Most commercial systems are not designed for conversational speech, expecting rigid turn-by-turn interaction. Although this is rapidly changing, with Google's Duplex having a major impact in the ability of systems to engage in conversation (in this case in a limited appointment/reservation domain) and use conversational features which meant, over the telephone, the artificial system is indistinguishable from a human dialogue partner.

There is great scope for SIA research to move forward the state of the art in speech interaction, but to do so it is important to be aware of techniques that are available for creating synthetic conversational styles and interaction.

Conversational Speech Style As we discussed in section 6.3.2, modern synthesis systems are based on speech corpora. Most corpora are created from read speech. Corpora of spontaneous and conversational speech are rare and present significant challenges for analysis. However, a conversational speech style is a requirement for an interactive system. Whereas unit selection systems face several challenges in creating conversational speech [Andersson et al. 2010], neural TTS systems can potentially model conversational styles of speech [e.g. Székely et al. 2019]. Such systems, with much lower requirements in terms of corpora size, especially with voice adaptation, could also successfully model acted conversational speech. Another feature of conversational speech is prosodic accommodation, where human dialogue partners tend to match each other's speech rate and style of speech [De Looze et al. 2014]. Using standard markup that implements global pitch and rate change this can be achieved by most synthetic speech systems.

Back Channels Giving speech feedback while listening is a fundamental part of conversational interaction. Such speech is termed back-channeling and has significant prosodic differences to the same words spoken in different contexts. There is nothing that prevents most speech synthesis systems generating back channels or stock phrases to support a dialogue partner [e.g. DeVault et al. 2014, Schröder et al. 2011].

Disfluencies Disfluency, often defined as a combination of speech errors and filled pauses, is a normal part of conversational speech. Filled pauses in particular can play an important role in conversational dynamics. Previous research has explored the use of filled pauses in speech synthesis [Adell et al. 2007, Andersson et al. 2010, Wester et al. 2015], however standard speech synthesis systems will not generally offer such functionality.

Laughter, Breathing & Speech Noises The modelling of non-speech noises is common in speech recognition but rarely used in speech synthesis. Google Duplex did use some of these techniques, but we are not aware of any published work. Synthesising laughter is extremely challenging [Trouvain and Schröder 2004]. As with sobbing, laughter often merges with speech, making modelling difficult.

Taking, Holding & Ceding the Floor In dialogue the act of speaking is mediated by the interaction with other speakers. In order to take the floor and be able to speak, prosodic techniques are used (i.e. abrupt in-breaths to show a desire to speak, raising your voice to prevent interruption, using very clear prosodic drop to show a readiness for another person to speak, etc.). Some of these effects can be modelled using current expressive synthesis functionality, but many remain unexplored in the current state of the art.

Architecture An interactive use of speech synthesis requires a major change in the assumed architecture of a dialogue system. Being able to interrupt speech output in response to an outside event is a fundamental requirement. Work on reactive [Wester et al. 2017] and incremental [Baumann and Schlangen 2012] speech synthesis architectures explore how speech output might change dynamically as external events occur. In complex artificial systems there is a tendency to play audio and move on, meaning that it cannot be interrupted, and if it is interrupted the system does not know what part of the message has already been played. Furthermore, bearing in mind that a response is required within 200ms, integrating speech recognition, dialogue planning, natural language generation and speech synthesis to respond effectively is a major challenge. For example, speech synthesis from a cloud service may just not be fast enough.

6.3.7 Summary

The technology underpinning expressive speech synthesis is rapidly changing. For much SIA work the state-of-the-art is not required, but a good understanding of the control and functionality of any proposed speech synthesis system is crucial for the design, implementation and

integration of synthetic speech into an agent or robot that is required to speak. In the previous sections we have tried to cover the engineering techniques and types of control that may be typically available or make sensible part of a development project. But we note that key for any technology is how it is used and integrated into the overall system.

6.4 Fundamental considerations when designing expressive agents

6.4.1 Should you use speech at all?

Although synthetic expressive voices can be built, whether to use expressive synthetic speech is fundamentally a design decision. This may not always be necessary and those who wish to use these voices may need to weigh up the benefits and drawbacks of speech as a modality before committing to such a design path.

6.4.1.1 The benefits of using speech

Speech has been touted as a more natural modality for interface interaction [Moore 2017]. Despite the stark differences between human-human and human-machine dialogue and their underlying purposes [Clark et al. 2019b], introducing speech in SIAs can have practical benefits. In *hands-busy and eyes-busy scenarios*, where people are otherwise occupied on other tasks, incorporating speech in SIAs can make a great deal of sense. A wealth of tasks require our attention. These include more mundane tasks like reading a newspaper or book, through to safety-critical tasks like driving or performing surgery [Heo et al. 2017, Large et al. 2017]. If systems wish to interact with us (or us with them) when we are otherwise busy, using speech can allow for minimal interference on our primary task – although this might depend on the difficulty of the primary task being conducted [Edwards et al. 2019].

Speech can also play an essential role in making an interface more accessible. For some users, using speech is not only preferable but an essential way to interact with a device [Corbett and Weber 2016]. Speech input and output can be critical in supporting people with limited motor capabilities [Corbett and Weber 2016], those who have visual impairments [Abdolrahmani et al. 2018, Reyes-Cruz et al. 2020], as well as having the potential to support interaction for older adult users [Sayago et al. 2019]. For diverse demographics, speech can provide accessibility interactions where other modalities (e.g. through GUIs and tactile interfaces) are difficult, if not impossible [Corbett and Weber 2016]. This allows users to conduct common tasks like web browsing [Sato et al. 2011, Williams et al. 2020]. Speech can also help technology to be more socioeconomically inclusive, supporting people whose levels of literacy may be low [Medhi et al. 2009]. Compared to other modalities, speech can also allow designers to more easily give the interface a personality. In human-human interaction, speech is a powerful indicator of identity and personality [Cameron 2001, Goffman 2005] and is one of the primary means of social identification [Barthes 1977]. These perceptions are also made in HCI [Nass and Lee 2000, 2001]. Through manipulating accent or speech rate,

a specific identity can be created for an interface, which can have a notable impact on user performance, learning, trust and even purchasing habits [Nass and Brave 2005].

6.4.1.2 The drawbacks of using speech

At times, speech can be a cognitively demanding modality [Aylett et al. 2014]. As such, it may not be appropriate for delivering large amounts of information [Doyle et al. 2019]. SIAs delivering long-form information such as lists could make interactions cognitively taxing for the user [Jung et al. 2020]. Recent research has also found that synthetic speech can impose a higher cognitive load to process than non-synthetic speech, particularly for lower quality synthesis [Govender and King 2018]. The cognitive demand of interaction is particularly acute for non-native (L2) speakers. In a recent study, users who had to use their non-native language to interact with a voice agent experienced significantly higher levels of cognitive load than those who could interact with their native language [Wu et al. 2020a]. For these users, pairing speech with visual feedback may be critical in supporting the interaction [Wu et al. 2020b].

A consistent challenge when using speech also lies in how people determine what they can actually do through speech with the interface. Facilitating discoverability – the ability for users to discover and use interface commands – needs major consideration when designing a speech interaction. Recent work identified that including a strategy to help users discover functions (either explicitly prompting users or through user initiating a help request) leads to significantly better usability scores than not designing for this at all [Kirschthaler et al. 2020]. Methods can also be used to aid discoverability when using more visual and multimodal forms of interaction, for example via screens to display common options or functions. Whether through vision- or speech-based means, thought needs to be put into how users will actually know what to do in a speech-enabled interaction.

Critically, we also need to consider whether the context within which the interaction occurs is appropriate for speech. The user experience of a speech system is highly dependent not only on the purpose of a system, but also on where it will be deployed, and the type of interaction the users are planned to have with it in that context. For instance, noisy environments can significantly impact the intelligibility of speech output [Cooke et al. 2013], cancelling out any multitasking benefits of using speech. Using speech in public environments may pose additional drawbacks. Research has shown that people prefer to use speech in private settings [Begany et al. 2016, Cowan et al. 2017b, Luger and Sellen 2016b], partly due to fear of social embarrassment. Not all settings may afford this potential embarrassment, however. Multiparty interactions with IPAs are common in home settings [Porcheron et al. 2018], and when designed for the context, they could lead to the feeling of social speech interface interaction being more appropriate [Porcheron et al. 2017].

6.4.1.3 Synthetic or pre-recorded speech

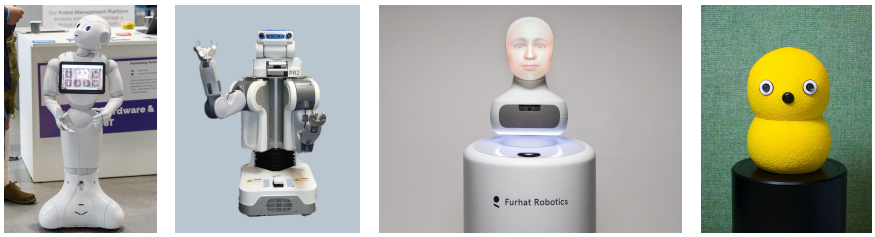
After deciding to use speech, it is important to consider whether to use synthetic or pre-recorded speech. Synthetic speech may be accessed relatively cheaply in contrast to pre-recorded speech, both in terms of open source synthesizers and commercially available voices. Additionally, using synthetic speech allows designers to make rapid changes to the spoken output, for example in changing entire utterances, ordering of words or expressivity. While creating synthetic speech is not without its cost (see 1.3.1), pre-recorded speech can incur in additional equipment, service and time costs that are not always feasible or within monetary constraints [Pincus et al. 2015]. These can further be increased if aiming for the highest quality pre-recordings, such as through the use of voice actors.

With both speech options, there are differences in quality to be considered. A prior systematic evaluation of both speech types in comparing utterances found that voice actor recordings were rated as more likeable, conversational, and natural than both amateur human recordings and synthesised voices [Georgila et al. 2012]. However, synthetic speech emitted from either a high-quality general purpose or a ”good limited domain voice” can outperform amateur human recordings [Georgila et al. 2012, p. 3525]. Designers must also consider what the speech in a system's use is intended to be. Pincus et al. [2015] discuss that a human voice is more appropriate for evoking an intended reaction from a listener, though only if evoking specific perceptions of a voice can warrant the additional costs that human recordings bring. If designing a system to have very precise expressive characteristics and quality, a human voice may be more suitable. Additionally, the perception of linguistic content spoken by a system can be related to the voice used to speak it. Clark et al. [2016] argue that phenomena like being polite and using vague language may be better received by a human voice than a synthetic voice. The purpose and use of a system should be considered when deciding on the type of voice to use for expressive speech. This also requires an understanding of how best to evaluate expressive speech and the context a system is used in (see 1.5.3).

Finally, many speech synthesis systems will support a library of pre-recorded prompts that can be accessed with XML markup. Thus, if the resources are available for building a custom voice, both the dynamic benefits of speech synthesis and the subtle quality enhancements from pre-recorded prompts can be used to support expressive output for a SIA.

6.4.2 The decision to embody - Intelligent Virtual Agents and Social Robots

Another clear decision point when designing SIAs that use expressive speech, is whether the agent being developed will be embodied or not. This can interplay significantly with the voice, influencing perceptions. When considering embodiment, there are two common forms: through either a virtual representation of an agent (e.g. an Intelligent Virtual Agent, IVA) or through a physical robot (e.g. a Social Robot, SR). Due to their embodied nature, both have the ability to use natural cues (gaze, gestures, etc.) to express emotion and exhibit personality,



(a) (b) (c) (d)

Figure 6.1 Social robots with different levels of embodiment: (a) Pepper robot by SoftBank Robotics © 2019 Marco Verch, (b) PR2 robot by Willow Garage © 2019 Ilaria Torre/Clearpath Robotics, (c) Furhat robot by Furhat Robotics © 2020 Furhat Robotics, (d) Keepon robot by BeatBots © 2007 Hideki Kozima, Marek Michalowski/BeatBots LLC.

with their ability to emulate social characteristics being important for acceptance [Fong et al. 2003].

The most obvious difference between the two is that SRs have a physical body that can interact with the environment, while IVAs only inhabit a virtual world. This has implications for their ability to be expressive, as well as realistically interact with the environment around them. This can be driven not only by their perceived physical form but also by their built-in sensors and algorithms. For example, some SRs such as Pepper or PR2 have cameras for vision, speakers and microphones for speaking and hearing, and manipulators for touch; the Furhat robot can 'see' and 'speak', but cannot touch; the Keepon robot can only 'hear' (Figure 6.1).

These affordances drive their ability to be expressive in interaction such as their gestures [Bremner et al. 2011, de Wit et al. 2020, Kose-Bagci et al. 2009, Salem et al. 2013] or gaze behaviours [Mumm and Mutlu 2011, Srinivasan and Murphy 2011].

When designing speech output for IVAs and SRs, it is important to be aware that these entities elicit different psychological and behavioural responses from users. People tend to rate robots as more similar to humans, and significantly more engaging than virtual characters [Kidd and Breazeal 2004], because of the perception that the agent is a real entity, as opposed to a virtual one [Kidd and Breazeal 2004]. A recent meta-review of studies comparing robots and virtual agents [Li 2015] also found that the majority of participants reported more positive attitudes – e.g. trust, enjoyment, attraction – towards robot than virtual agents. These differences may be rooted in the differing physical presence afforded by SRs over IVAs. This may not only lead to more realistic expressiveness, but presence augments their ability to generate rich communication, which in turn leads to a more successful interaction and higher

acceptance from the user [Wainer et al. 2006]. Physical co-location has also been shown to increase compliance and influence decision-making [Bainbridge et al. 2011].

The discussion on embodiment is integral to the development of expressive speech in artificial agents, as embodiment may significantly impact how expressive speech is perceived. This further ties to the concept of multimodality, as expressivity is multimodal by nature. For example, if we want to highlight a word in an utterance, we use prosodic stress and will often accompany it by so-called 'visual beats', such as hand gestures or eyebrow movements [Krahmer and Swerts 2007, Swerts and Krahmer 2010]. Expressive phenomena (such as highlighting a word, smiling, or showing empathy) can still be conveyed through a single modality, for example in the case of disembodied artificial agents, and they generally increase positive perceptions [Bretan et al. 2015] and decision-making [Torre et al. 2020]. However, expressing the same phenomenon through multiple modalities can increase communication success, such as in the case of lip-reading [Campanella and Belin 2007]. Not only this, but accidentally – or not – pairing modalities that do not match actually interferes with communication, as exemplified by the McGurk Effect [cf.t McGurk and MacDonald 1976]. In human-agent interaction, it has been shown that agents expressing different emotions in the face and voice elicit different behavioural responses than when the emotion is expressed in only one channel [Antos et al. 2011, Torre et al. 2018b]. This suggests that expressivity enhances communication when there is congruence between modalities, but interferes with it when there the modalities do not match. This could be an issue when designing an expressive voice for a social robot that cannot make the corresponding facial and bodily gestures, for example because it lacks the appropriate degrees of freedom. In sum, while expressivity at large can facilitate communication and increase person perception, designers must ensure that all the available modalities on an artificial agent have the same expressive capabilities, otherwise the resulting expression mismatch could be detrimental for the interaction.

6.5 Current Challenges & Future Directions in Expressive Synthesis

Recent work has highlighted a number of challenges and future directions related to expressive speech that need to be addressed. Clark et al. [2019a] provide an overview of speech systems within HCI, noting a number of research challenges alongside methodological and evaluation challenges. Cambre and Kulkarnia [2019] highlight the social implications of designing voices for smart devices and provide a research framework for designers to utilise to help shape user's experiences. Finally, Wagner et al. [2019] discuss the future of evaluating speech synthesis, suggesting a move towards HCI-focused approaches of evaluating speech in appropriate contexts with users. Here, we build upon this existing work and present three challenges for those working in different areas of expressive synthesis.

6.5.1 Considering why and where we need expressive synthesis

One of the aims of developing and using expressive synthesis is to emulate human-like qualities within the voice [Akuzawa et al. 2018]. Although a seemingly innocuous goal, recent work emphasises that this may have profound consequences for interaction from the user perspective. The voice of a speech system is likely one of the key reference points when thinking and reasoning about what a system using speech can (and cannot) do. Design decisions in the voice may be important drivers of people's perceptions of partner competence and ability [Luger and Sellen 2016a] (i.e. a user's partner model) [Cowan et al. 2017a]. Work (highlighted in the sections above) demonstrates that expressiveness (e.g through accent choices) can significantly impact user perceptions [Dahlbäck et al. 2007] and user language choices [Cowan et al. 2019] in interaction. Crucially, making voices more human-like through using expressive synthesis may over-inflate users' perceptions of what the system being used can achieve in interaction [Luger and Sellen 2016a, Moore 2017], significantly affecting the quality of interaction. Research with both interviews and focus group exploring the user experience of speech-based IPAs [Cowan et al. 2017b, Luger and Sellen 2016b] found – unsurprisingly – that users see conversation as the key metaphor for interaction, with the human-like nature of speech output and synthesis being major cues to support this assumption. Yet these clearly do not accurately map to current system capabilities. This leads to a problem in inaccurate expectation setting [Leahu et al. 2013, Luger and Sellen 2016b, Moore et al. 2016], resulting in potential communication breakdown and unsuccessful engagement with systems. These perceptions of humanness also appear to be multidimensional in nature rather than monolithic, and designers may need to reconsider them as such [Doyle et al. 2019], for example by matching levels of human-likeness both in visual appearance and voice quality [McGinn and Torre 2019]. Identifying when and whether to use expressive synthesis as well as evaluating its effect on user interaction is a significant challenge for future research in this domain.

6.5.2 Towards gender-neutral voices in expressive speech design?

A typical decision by most designers of speech systems is to opt for a female voice as the default [Danielescu 2020]. This has brought about a significant sense of users anthropomorphising speech agents as female, consistently referring to agents as "her" or "she" when describing their experiences [Cowan et al. 2017b, Luger and Sellen 2016b]. A recent UNESCO report identified that current design of speech agents is not gender-neutral, potentially amplifying gender stereotypes, with the need to consider how voices can be designed to be more gender-neutral [West et al. 2019]. A key challenge in expressive synthesis is in exploring how we can build expressive synthesis that minimises this bias. According to UNESCO, such a decision may have significant consequences: "Because the speech of most voice assistants is female, it sends a signal that women are obliging, docile and eager-to-please helpers, available at the touch of a button or with a blunt voice command like 'hey' or 'OK'." [West et al. 2019,

p. 150]. Consequently, it is clear we need to consider carefully the design rationale for gendered voices within speech and agent based systems. Seminal work by Nass, Steur & Tauber has highlighted that similar gender stereotypes in human-human interaction also appear when using male and female agent voices, with male voices being perceived as more dominant and assertive than female voices [Nass et al. 1994]. As highlighted by Sutton [2020] and Cambre & Kulkarni [2019], we must deal with gender issues with speech interfaces with nuance, avoiding the common conflation of biological sex (i.e. male or female) with perceptions of gender in VUI design, which are influenced by a number of variables.

6.5.3 More and better user evaluation needed

Evaluating speech is currently done using three key approaches [Wagner et al. 2019]: *objective assessments* classifying systems with particular scores or contrasting them with other speech (e.g. through mel-cepstral distortion (MCD) ratings); *subjective assessments* rating speech on concepts such as intelligibility and naturalness: and *behavioural assessments* examining user actions like task completion time or physiological arousal. Traditionally, synthesis is evaluated through listening tests [e.g. Black and Tokuda 2005], where speech samples are subjectively rated for perceptions of naturalness and intelligibility [Wu et al. 2019]. These assessments are often done in a state of quasi-isolation and contrasted against other samples of synthetic or human speech. Given that speech is rarely listened to in a vacuum, the ecological validity of this approach has been questioned [Mendelson and Aylett 2017]. Instead, speech forms part of a specific application in a particular interaction context [Clark et al. 2019a]. As such, it is critical that expressive speech synthesis (indeed all synthesis) is evaluated in a manner relevant to how, where and with whom it is deployed. We also need more evaluation of how expressive synthesis parameters impact user experience and user behaviour. Future expressive synthesis work should follow recent calls to adopt approaches often seen in HCI literature, where some form of user interaction with a system is crucial [Wagner et al. 2019]. These interactions may include simple prototypes or mock systems, or even Wizard of Oz scenarios, placed within an appropriate interaction context. Following these methods will not only improve ecological validity and remove the burden of focusing on incremental improvements to human-likeness [Aylett et al. 2019a], but will also shed light on how aspects of expressiveness influence the end user and their interaction experience.

6.6 Summary

As mentioned in the introduction, we know there is something special about speech, and as speech systems are being interacted with widely, in a varying set of contexts, the expressiveness of the voices they use is a timely topic. It is one at the forefront of research and development in speech systems. Through this chapter we hope that we have given the reader a flavour of some of the key definitions, methods and considerations when building expressive speech synthesis into any Socially Interactive Agent. We highlight the importance of aspects

such as corpora as well as the usefulness and limitations of SSML in the creation of expressive voices. Even though we emphasise the types of parameters that can currently be added, or are being explored in relation to synthetic speech, our chapter is not meant to suggest that these should be used at all times and in all situations. On the contrary. We wish to emphasise to the reader that it is important to think about the interaction being developed and designed for. Is expressive speech going to benefit this interaction? How will it affect user's perceptions and behaviours? Will embodiment be able to support the expressiveness of the speech appropriately? Do you even need speech at all? The focus on user-related concepts in expressive synthesis is a significant omission in the field currently, and as such it is a major challenge for future researchers looking to influence expressive synthesis research. We hope that this chapter will not only give readers a set of signposts to support them when exploring the world of expressive synthesis, but also act as a catalyst for readers to think critically about the role, nature and place of expressive synthesis in systems being designed.

Bibliography

A. Abdolrahmani, R. Kuber, and S. M. Branham. 2018. " siri talks at you" an empirical investigation of voice-activated personal assistant (vapa) usage by individuals who are blind. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 249–258.

J. Adell, A. Bonafonte, and D. Escudero. 2007. Filled pauses in speech synthesis: towards conversational speech. In *International Conference on Text, Speech and Dialogue*, pp. 358–365. Springer.

K. Akuzawa, Y. Iwasawa, and Y. Matsuo. Sept. 2018. Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder. In *Interspeech 2018*, pp. 3067–3071. ISCA. http://www.isca-speech.org/archive/Interspeech_2018/abstracts/1113.html. DOI: 10.21437/Interspeech.2018-1113.

S. Andersson, K. Georgila, D. Traum, M. Aylett, and R. A. Clark. 2010. Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection. In *Speech Prosody 2010-Fifth International Conference*.

S. Andrist, M. Ziadee, H. Boukaram, B. Mutlu, and M. Sakr. 2015. Effects of culture on the credibility of robot speech. In *Proceedings of the 10th Annual International Conference on Human-Robot Interaction*, HRI '15, pp. 157–164. ACM/IEEE. DOI: 10.1145/2696454.2696464.

D. Antos, C. M. De Melo, J. Gratch, and B. J. Grosz. 2011. The influence of emotion expression on perceptions of trustworthiness in negotiation. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*.

P. Arias, C. Soladie, O. Bouafif, A. Robel, R. Seguier, and J.-J. Aucouturier. 2018. Realistic transformation of facial and vocal smiles in real-time audiovisual streams. *IEEE Transactions on Affective Computing*.

S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou. 2018. Neural voice cloning with a few samples. In *Advances in Neural Information Processing Systems*, pp. 10019–10029.

A. Aubin, A. Cervone, O. Watts, and S. King. 2019. Improving speech synthesis with discourse relations. In *INTERSPEECH*, pp. 4470–4474.

M. P. Aylett and D. A. Braude. 2018. Designing speech interaction for the sony xperia ear and oakley radar pace smartglasses. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, pp. 379–384.

M. P. Aylett, B. Potard, and C. J. Pidcock. 2013. Expressive speech synthesis: Synthesising ambiguity. In *Eighth ISCA Workshop on Speech Synthesis*.

M. P. Aylett, P. O. Kristensson, S. Whittaker, and Y. Vazquez-Alvarez. 2014. None of a CHInd: relationship counselling for HCI and speech technology. In *Proceedings of the extended abstracts of the 32nd annual ACM conference on Human factors in computing systems - CHI EA '14*, pp. 749–760. ACM Press, Toronto, Ontario, Canada. ISBN 978-1-4503-2474-8. http://dl.acm.org/citation.cfm?doid=2559206.2578868. DOI: 10.1145/2559206.2578868.

M. P. Aylett, A. Vinciarelli, and M. Wester. 2017. Speech synthesis for the generation of artificial personality. *IEEE transactions on affective computing*.

M. P. Aylett, B. R. Cowan, and L. Clark. 2019a. Siri, echo and performance: You have to suffer darling. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, p. alt08. ACM.

M. P. Aylett, S. J. Sutton, and Y. Vazquez-Alvarez. 2019b. The right kind of unnatural: designing a robot voice. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, pp. 1–2.

W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati. 2011. The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, 3(1): 41–52.

R. Barthes. 1977. *Image-music-text*. Macmillan.

T. Baumann and D. Schlangen. 2012. Inpro_iss: A component for just-in-time incremental speech synthesis. In *Proceedings of the ACL 2012 System Demonstrations*, pp. 103–108.

G. M. Begany, N. Sa, and X. Yuan. 2016. Factors affecting user perception of a spoken language vs. textual search interface: a content analysis. *Interacting with computers*, 28(2): 170–180.

H. Bishop, N. Coupland, and P. Garrett. 2005. Conceptual accent evaluation: Thirty years of accent prejudice in the UK. *Acta Linguistica Hafniensia*, 37(1): 131–154. DOI: 10.1080/03740463.2005.10416087.

A. W. Black and K. Tokuda. 2005. The blizzard challenge-2005: Evaluating corpus-based speech synthesis on common datasets. In *Ninth European Conference on Speech Communication and Technology*.

B. Bollepalli, L. Juvela, P. Alku, et al. 2019. Lombard speech synthesis using transfer learning in a tacotron text-to-speech system. In *Interspeech*, pp. 2833–2837.

D. A. Braude, M. P. Aylett, C. Laoide-Kemp, S. Ashby, K. M. Scott, B. O. Raghallaigh, A. Braudo, A. Brouwer, and A. Stan. 2019. All together now: The living audio dataset. In *INTERSPEECH*, pp. 1521–1525.

P. Bremner, A. G. Pipe, C. Melhuish, M. Fraser, and S. Subramanian. 2011. The effects of robot-performed co-verbal gesture on listener behaviour. In *2011 11th IEEE-RAS International Conference on Humanoid Robots*, pp. 458–465. IEEE.

M. Bretan, G. Hoffman, and G. Weinberg. 2015. Emotionally expressive dynamic physical behaviors in robots. *International Journal of Human-Computer Studies*, 78: 1–16.

J. Cambre and C. Kulkarni. 2019. One voice fits all? social implications and research challenges of designing voices for smart devices. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–19.

D. Cameron. 2001. *Working with spoken discourse*. Sage.

S. Campanella and P. Belin. 2007. Integrating face and voice in person perception. *Trends in Cognitive Sciences*, 11(12): 535–543.

A. C. Cargile and H. Giles. 1997. Understanding language attitudes: Exploring listener affect and identity. *Language & Communication*, 17(3): 195–217.

L. Clark, A. Ofemile, S. Adolphs, and T. Rodden. 2016. A multimodal approach to assessing user experiences with agent helpers. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(4): 29.

L. Clark, P. Doyle, D. Garaialde, E. Gilmartin, S. Schlögl, J. Edlund, M. Aylett, J. Cabral, C. Munteanu, J. Edwards, et al. 2019a. The state of speech in hci: Trends, themes and challenges. *Interacting with Computers*, 31(4): 349–371.

L. Clark, N. Pantidi, O. Cooney, P. Doyle, D. Garaialde, J. Edwards, B. Spillane, E. Gilmartin, C. Murad, C. Munteanu, et al. 2019b. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12.

M. Cooke, C. Mayo, and C. Valentini-Botinhao. 2013. Intelligibility-enhancing speech modifications: the hurricane challenge.

E. Corbett and A. Weber. 2016. What can i say? addressing user experience challenges of a mobile voice user interface for accessibility. In *Proceedings of the 18th international conference on human-computer interaction with mobile devices and services*, pp. 72–82.

N. Coupland and H. Bishop. 2007. Ideologised values for british accents 1. *Journal of sociolinguistics*, 11(1): 74–93.

B. R. Cowan, N. Pantidi, D. Coyle, K. Morrissey, P. Clarke, S. Al-Shehri, D. Earley, and N. Bandeira. 2017a. 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*, p. 43. ACM.

B. R. Cowan, N. Pantidi, D. Coyle, K. Morrissey, P. Clarke, S. Al-Shehri, D. Earley, and N. Bandeira. 2017b. 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*, p. 43. ACM.

B. R. Cowan, P. Doyle, J. Edwards, D. Garaialde, A. Hayes-Brady, H. P. Branigan, J. a. Cabral, and L. Clark. 2019. What's in an accent? the impact of accented synthetic speech on lexical choice in human-machine dialogue. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, CUI '19. Association for Computing Machinery, New York, NY, USA. ISBN 9781450371872. https://doi-org.ucd.idm.oclc.org/10.1145/3342775.3342786. DOI: 10.1145/3342775.3342786.

A. Cruttenden et al. 1997. *Intonation*. Cambridge University Press.

D. Crystal. 1997. A dictionary of linguistics and phonetics . 1997. *UK: Blackwell*.

D. Crystal. 2011. *A dictionary of linguistics and phonetics*, volume 30. John Wiley & Sons.

N. Dahlbäck, S. Swamy, C. Nass, F. Arvidsson, and J. Skågeby. 2001. Spoken interaction with computers in a native or non-native language-same or different. In *Proceedings of INTERACT*, pp. 294–301.

N. Dahlbäck, Q. Wang, C. Nass, and J. Alwin. 2007. Similarity is more important than expertise: Accent effects in speech interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 1553–1556. ACM.

A. Danielescu. 2020. Eschewing gender stereotypes in voice assistants to promote inclusion. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, pp. 1–3.

C. De Looze, S. Scherer, B. Vaughan, and N. Campbell. 2014. Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication*, 58: 11–34.

J. de Wit, A. Brandse, E. Krahmer, and P. Vogt. 2020. Varied human-like gestures for social robots: Investigating the effects on children's engagement and language learning. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 359–367.

D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, et al. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 1061–1068.

P. R. Doyle, J. Edwards, O. Dumbleton, L. Clark, and B. R. Cowan. 2019. Mapping perceptions of humanness in intelligent personal assistant interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '19. Association for Computing Machinery, New York, NY, USA. ISBN 9781450368254. https://doi-org.ucd.idm.oclc.org/10.1145/3338286.3340116. DOI: 10.1145/3338286.3340116.

J. Edwards, H. Liu, T. Zhou, S. J. J. Gould, L. Clark, P. Doyle, and B. R. Cowan. 2019. Multitasking with alexa: How using intelligent personal assistants impacts language-based primary task performance. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, CUI '19. Association for Computing Machinery, New York, NY, USA. ISBN 9781450371872. https://doi-org.ucd.idm.oclc.org/10.1145/3342775.3342785. DOI: 10.1145/3342775.3342785.

K. El Haddad, H. Cakmak, A. Moinet, S. Dupont, and T. Dutoit. 2015a. An hmm approach for synthesizing amused speech with a controllable intensity of smile. In *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pp. 7–11. IEEE.

K. El Haddad, S. Dupont, N. d'Alessandro, and T. Dutoit. 2015b. An hmm-based speech-smile synthesis system: An approach for amusement synthesis. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 5, pp. 1–6. IEEE.

K. El Haddad, I. Torre, E. Gilmartin, H. Çakmak, S. Dupont, T. Dutoit, and N. Campbell. 2017. Introducing amus: The amused speech database. In N. Camelin, Y. Estève, and C. Martín-Vide, eds., *Proceedings of Statistical Language and Speech Processing Conference*, pp. 229–240. Springer International Publishing. ISBN 978-3-319-68456-7. DOI: 10.1007/978-3-319-68456-7.

D. Erickson. 2005. Expressive speech: Production, perception and application to speech synthesis. *Acoustical science and technology*, 26(4): 317–325.

T. Fong, I. Nourbakhsh, and K. Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4): 143–166.

P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana. 2016. Analysis of emotional speech—a review. In *Toward Robotic Socially Believable Behaving Systems-Volume I*, pp. 205–238. Springer.

K. Georgila, A. W. Black, K. Sagae, and D. R. Traum. 2012. Practical evaluation of human and synthesized speech for virtual human dialogue systems. In *LREC*, pp. 3519–3526.

E. Goffman. 2005. *Interaction ritual: Essays in face to face behavior*. AldineTransaction.

A. Govender and S. King. 2018. Using pupillometry to measure the cognitive load of synthetic speech. *System*, 50: 100.

D. Govind and S. M. Prasanna. 2013. Expressive speech synthesis: a review. *International Journal of Speech Technology*, 16(2): 237–260.

S. Heo, M. Annett, B. J. Lafreniere, T. Grossman, and G. W. Fitzmaurice. 2017. No need to stop what you're doing: Exploring no-handed smartwatch interaction. In *Graphics Interface*, pp. 107–114.

Z. Hodari, O. Watts, and S. King. 2019. Using generative modelling to produce varied intonation for speech synthesis. *arXiv preprint arXiv:1906.04233*.

G. Hofer, K. Richmond, and R. Clark. 2005. Informed blending of databases for emotional speech synthesis. In *Proc. Interspeech*.

A. Hughes, P. Trudgill, and D. Watt. 2013. *English accents and dialects: An introduction to social and regional varieties of English in the British Isles*. Routledge.

A. J. Hunt and A. W. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pp. 373–376. IEEE.

A. Ikeno and J. H. Hansen. Nov 2007. The effect of listener accent background on accent perception and comprehension. *EURASIP Journal on Audio, Speech, and Music Processing*, 2007(1): 076030. ISSN 1687-4722. https://doi.org/10.1155/2007/76030. DOI: 10.1155/2007/76030.

J. Jung, S. Lee, J. Hong, E. Youn, and G. Lee. 2020. Voice+ tactile: Augmenting in-vehicle voice user interface with tactile touchpad interaction. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12.

R. G. Kamiloğlu, A. H. Fischer, and D. A. Sauter. 2020. Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic Bulletin & Review*, pp. 1–29.

T. Kawahara. 2019. Spoken dialogue system for a human-like conversational robot erica. In *9th International Workshop on Spoken Dialogue System Technology*, pp. 65–75. Springer.

T. Kenter, V. Wan, C.-A. Chan, R. Clark, and J. Vit. 2019. Chive: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network. In *International Conference on Machine Learning*, pp. 3331–3340.

C. D. Kidd and C. Breazeal. 2004. Effect of a robot on user perceptions. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 4, pp. 3559–3564. IEEE.

K. D. Kinzler, K. H. Corriveau, and P. L. Harris. 2011. Children's selective trust in native-accented speakers. *Developmental Science*, 14(1): 106–111.

P. Kirschthaler, M. Porcheron, and J. E. Fischer. 2020. What can i say? effects of discoverability in vuis on task performance and user experience. In *Proceedings of the 2nd International Conference on Conversational User Interfaces*, CUI '20. Association for Computing Machinery, New York, NY, USA. ISBN 978-1-4503-7544-3/20/07. DOI: 10.1145/3405755.3406119.

D. H. Klatt. 1980. Software for a cascade/parallel formant synthesizer. *the Journal of the Acoustical Society of America*, 67(3): 971–995.

J. Kominek and A. W. Black. 2004. The cmu arctic speech databases. In *Fifth ISCA workshop on speech synthesis*.

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

H. Kose-Bagci, E. Ferrari, K. Dautenhahn, D. S. Syrdal, and C. L. Nehaniv. 2009. Effects of embodiment and gestures on social interaction in drumming games with a humanoid robot. *Advanced Robotics*,

23(14): 1951–1996.

E. Krahmer and M. Swerts. 2007. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3): 396–414.

D. R. Large, L. Clark, A. Quandt, G. Burnett, and L. Skrypchuk. Sept. 2017. Steering the conversation: A linguistic exploration of natural language interactions with a digital assistant during simulated driving. *Applied Ergonomics*, 63: 53–61. ISSN 00036870. https://linkinghub.elsevier.com/retrieve/pii/S0003687017300790. DOI: 10.1016/j.apergo.2017.04.003.

L. Leahu, M. Cohn, and W. March. 2013. How categories come to matter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, p. 3331. ACM Press, Paris, France. ISBN 978-1-4503-1899-0. http://dl.acm.org/citation.cfm?doid=2470654.2466455. DOI: 10.1145/2470654.2466455.

K. A. Lenzo and A. W. Black. 2000. Diphone collection and synthesis. In *Sixth International Conference on Spoken Language Processing*.

J. Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77: 23–37.

J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, and R. Barra-Chicote. 2018. Robust universal neural vocoding. *arXiv preprint arXiv:1811.06292*.

E. Luger and A. Sellen. 2016a. Like having a really bad pa: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5286–5297. ACM.

E. Luger and A. Sellen. 2016b. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5286–5297. ACM.

H.-T. Luong and J. Yamagishi. 2020. Nautilus: a versatile voice cloning system. *arXiv preprint arXiv:2005.11004*.

F. Marelli, B. Schnell, H. Bourlard, T. Dutoit, and P. N. Garner. 2019. An end-to-end network to synthesize intonation using a generalized command response model. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7040–7044. IEEE.

C. McGinn and I. Torre. 2019. Can you tell the robot by the voice? an exploratory study on the role of voice in the perception of robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 211–221. IEEE.

H. McGurk and J. MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264(5588): 746–748.

I. Medhi, S. N. Gautama, and K. Toyama. 2009. A comparison of mobile money-transfer uis for non-literate and semi-literate users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1741–1750.

J. Mendelson and M. P. Aylett. 2017. Beyond the listening test: An interactive approach to tts evaluation. In *INTERSPEECH*, pp. 249–253.

R. K. Moore. 2017. Is spoken language all-or-nothing? implications for future speech-based human-machine interaction. In *Dialogues with Social Robots*, pp. 281–291. Springer.

R. K. Moore, H. Li, and S.-H. Liao. Sept. 2016. Progress and Prospects for Spoken Language Technology: What Ordinary People Think. pp. 3007–3011. http://www.isca-speech.org/archive/Interspeech_2016/abstracts/0874.html. DOI: 10.21437/Interspeech.2016-874.

J. Mumm and B. Mutlu. 2011. Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 331–338.

C. Nass and K. M. Lee. 2000. Does Computer-generated Speech Manifest Personality? An Experimental Test of Similarity-attraction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pp. 329–336. ACM, New York, NY, USA. ISBN 978-1-58113-216-8. http://doi.acm.org/10.1145/332040.332452. DOI: 10.1145/332040.332452.

C. Nass and K. M. Lee. 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of experimental psychology: applied*, 7(3): 171.

C. Nass, J. Steuer, and E. R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 72–78.

C. I. Nass and S. Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge, MA.

A. Niculescu, G. M. White, S. S. Lan, R. U. Waloejo, and Y. Kawaguchi. 2008. Impact of english regional accents on user acceptance of voice user interfaces. In *Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges*, NordiCHI '08, p. 523–526. Association for Computing Machinery, New York, NY, USA. ISBN 9781595937049. https://doi-org.ucd.idm.oclc.org/10.1145/1463160.1463235. DOI: 10.1145/1463160.1463235.

J. J. Ohala. 1983. Cross-language use of pitch: an ethological view. *Phonetica*, 40(1): 1–18.

A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

A. Ortony, G. Clore, and A. Collins. 1988. *The Cognitive Structure of Emotion.* CUP, Cambridge.

A. L. Paugh. 2005. Multilingual play: Children's code-switching, role play, and agency in dominica, west indies. *Language in society*, 34(1): 63–86.

E. Pincus, K. Georgila, and D. Traum. 2015. Which synthetic voice should i choose for an evocative task? In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 105–113.

W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller. 2017. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654*.

J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny. 2006. The ibm expressive text-to-speech synthesis system for american english. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4): 1099–1108.

M. Porcheron, J. E. Fischer, and S. Sharples. 2017. "Do Animals Have Accents?": Talking with Agents in Multi-Party Conversation. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '17, pp. 207–219. ACM, New York, NY, USA. DOI: 10.1145/2998181.2998298.

M. Porcheron, J. E. Fischer, S. Reeves, and S. Sharples. 2018. Voice interfaces in everyday life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 640. ACM.

B. Potard, M. P. Aylett, and D. A. Braude. 2016. Cross modal evaluation of high quality emotional speech synthesis with the virtual human toolkit. In *International Conference on Intelligent Virtual Agents*, pp. 190–197. Springer.

N. Prateek, M. Łajszczak, R. Barra-Chicote, T. Drugman, J. Lorenzo-Trueba, T. Merritt, S. Ronanki, and T. Wood. 2019. In other news: A bi-style text-to-speech model for synthesizing newscaster voice with limited data. *arXiv preprint arXiv:1904.02790*.

S. Ramakrishnan. 2012. Recognition of emotion from speech: A review. *Speech Enhancement, Modeling and recognition–algorithms and Applications*, 7: 121–137.

G. Reyes-Cruz, J. E. Fischer, and S. Reeves. 2020. Reframing disability as competency: Unpacking everyday technology practices of people with visual impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1–13. ACM, New York, NY, USA. ISBN 9781450367080. https://doi.org/10.1145/3313831.3376767. DOI: 10.1145/3313831.3376767.

D. C. Rubin and J. M. Talarico. 2009. A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words. *Memory*, 17(8): 802–808.

E. B. Ryan and H. Giles. 1982. An Integrative Perspective for the Study of Attitudes Towards Language Variation. In *Attitudes towards language variation: Social and applied contexts*, pp. 1–19. Edward Arnold London.

M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, 5(3): 313–323.

D. Sato, S. Zhu, M. Kobayashi, H. Takagi, and C. Asakawa. 2011. Sasayaki: augmented voice web browsing experience. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 2769–2778.

S. Sayago, B. B. Neves, and B. R. Cowan. 2019. Voice assistants and older people: Some open issues. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, CUI '19. ACM, New York, NY, USA. ISBN 9781450371872. https://doi.org/10.1145/3342775.3342803. DOI: 10.1145/3342775.3342803.

M. Schröder. 2001. Emotional speech synthesis: A review. In *Proceedings Eurospeech 01*, pp. 561–4.

M. Schröder. 2004. Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. In *Proceedings Workshop on Affective Dialogue Systems*, pp. 209–220.

M. Schröder. 2009. Expressive speech synthesis: Past, present, and possible futures. In *Affective information processing*, pp. 111–126. Springer.

M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. Ter Maat, G. McKeown, S. Pammi, M. Pantic, et al. 2011. Building autonomous sensitive artificial listeners. *IEEE transactions on affective computing*, 3(2): 165–183.

R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. *arXiv preprint arXiv:1803.09047*.

V. Srinivasan and R. Murphy. 2011. A survey of social gaze. In *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 253–254.

G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu. 2020. Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6264–6268. IEEE.

S. Sundaram and S. Narayanan. 2007. Automatic acoustic synthesis of human-like laughter. *The Journal of the Acoustical Society of America*, 121(1): 527–535.

S. Sutton. 2020. Gender ambiguous, not genderless: Designing gender in voice user interfaces (vuis) with sensitivity. In *Proceedings of the 2nd International Conference on Conversational User Interfaces*, CUI '20. Association for Computing Machinery, New York, NY, USA.

M. Swerts and E. Krahmer. 2010. Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions. *Journal of Phonetics*, 38(2): 197–206.

E. Székely. 2015. Expressive speech synthesis in human interaction. *PhD, University College Dublin*.

É. Székely, G. E. Henter, J. Beskow, and J. Gustafson. 2019. Spontaneous conversational speech synthesis from found data. In *Interspeech*.

V. C. Tartter and D. Braun. 1994. Hearing smiles and frowns in normal and whisper registers. *Journal of the Acoustical Society of America*, 96(4): 2101–2107.

P. Taylor and A. Isard. 1997. Ssml: A speech synthesis markup language. *Speech communication*, 21(1-2): 123–133.

M. Theune, K. Meijs, D. Heylen, and R. Ordelman. 2006. Generating expressive speech for storytelling applications. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4): 1137–1144.

K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. 2000. Speech parameter generation algorithms for hmm-based speech synthesis. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pp. 1315–1318. IEEE.

I. Torre and S. L. Maguer. 2020. Should robots have accents? In *Proceedings of the 29th International Workshop on Robot and Human Interactive Communication*, RO-MAN '20. IEEE.

I. Torre, J. Goslin, and L. White. 2015. Investing in accents: How does experience mediate trust attributions to different voices? In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*.

I. Torre, E. Carrigan, K. McCabe, R. McDonnell, and N. Harte. 2018a. Survival at the museum: A cooperation experiment with emotionally expressive virtual characters. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pp. 423–427. ACM. DOI: https://doi.org/10.1145/3242969.3242984.

I. Torre, E. Carrigan, K. McCabe, R. McDonnell, and N. Harte. 2018b. Survival at the museum: A cooperation experiment with emotionally expressive virtual characters. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pp. 423–427. ACM. DOI: https://doi.org/10.1145/3242969.3242984.

I. Torre, J. Goslin, and L. White. 2020. If your device could smile: People trust happy-sounding artificial agents more. *Computers in Human Behavior*, 105: 106215. DOI:

https://doi.org/10.1016/j.chb.2019.106215.

J. Trouvain and M. Schröder. 2004. How (not) to add laughter to synthetic speech. In *Tutorial and Research Workshop on Affective Dialogue Systems*, pp. 229–232. Springer.

G. R. Tucker. 1999. A global perspective on bilingualism and bilingual education. eric digest.

P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, É. Székely, C. Tånnander, et al. 2019. Speech synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program. In *Proceedings of the 10th Speech Synthesis Workshop (SSW10)*.

J. Wainer, D. J. Feil-Seifer, D. A. Shell, and M. J. Mataric. 2006. The role of physical embodiment in human-robot interaction. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 117–122. IEEE.

Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

M. West, R. Kraut, and H. Chew, 2019. I'd blush if i could: closing gender divides in digital skills through education. https://unesdoc.unesco.org/ark:/48223/pf0000367416.

M. Wester, M. Aylett, M. Tomalin, and R. Dall. 2015. Artificial Personality and Disfluency. *Interspeech 2015*, p. 5.

M. Wester, D. A. Braude, B. Potard, M. P. Aylett, and F. Shaw. 2017. Real-time reactive speech synthesis: Incorporating interruptions. In *INTERSPEECH*, pp. 3996–4000.

A. Williams, J. Cambre, I. Bicking, A. Wallin, J. Tsai, and J. Kaye. 2020. Toward voice-assisted browsers: A preliminary study with firefox voice. In *Proceedings of the 2nd International Conference on Conversational User Interfaces*, CUI '20. Association for Computing Machinery, New York, NY, USA. ISBN 978-1-4503-7544-3/20/07. DOI: 10.1145/3405755.3406154.

Y. Wu, J. Edwards, O. Cooney, A. Bleakley, P. R. Doyle, L. Clark, D. Rough, and B. R. Cowan. 2020a. Mental workload and language production in non-native speaker ipa interaction. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, CUI '20. Association for Computing Machinery, New York, NY, USA. ISBN 9781450375443. https://doi-org.ucd.idm.oclc.org/10.1145/3405755.3406118. DOI: 10.1145/3405755.3406118.

Y. Wu, D. Rough, A. Bleakley, J. Edwards, O. Cooney, P. R. Doyle, L. Clark, and B. R. Cowan. 2020b. See what i'm saying? comparing intelligent personal assistant use for native and non-native language speakers. In *Proceedings of the 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, Mobile HCI '20. Association for Computing Machinery, New York, NY, USA.

Z. Wu, S. Le Maguer, J. Cabral, and S. King. 2019. The blizzard challenge 2019.

J. Yamagishi, K. Ogata, Y. Nakano, J. Isogai, and T. Kobayashi. 2006. Hsmm-based model adaptation algorithms for average-voice-based speech synthesis. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pp. I–I. IEEE.

J. Yamagishi, C. Veaux, K. MacDonald, et al. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92).

H. Zen, A. Senior, and M. Schuster. 2013. Statistical parametric speech synthesis using deep neural networks. In *2013 ieee international conference on acoustics, speech and signal processing*, pp.

7962–7966. IEEE.

H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran. 2019. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. *arXiv preprint arXiv:1907.04448*.

X. Zhu and L. Xue. 2020. Building a controllable expressive speech synthesis system with multiple emotion strengths. *Cognitive Systems Research*, 59: 151–159.