[1]This paper involves listening to audio. The crucial links are marked by 🔊 symbol. If you don't have audio and can't play mp3 over a url there's no need to read any further.



# Siri, Echo and Performance: You have to Suffer Darling

**Matthew P. Aylett**
CereProc Ltd.
Edinburgh, UK
matthewa@cereproc.com

**Benjamin R. Cowan**
University College Dublin.
Dublin, Ireland
benjamin.cowan@ucd.ie

**Leigh Clark**
University College Dublin.
Dublin, Ireland
leigh.clark@ucd.ie

## ABSTRACT

Don't ignore this because its about speech technology. VUIs (voice user interfaces) won a best paper in CHI 2018. Did that get your attention? Good. Siri, Ivona, Google Home, and most speech synthesis systems have voices which are based on imitating a neutral citation style of speech and making it sound natural. But, in the real world, darling, people have to act, to perform! In this paper we will talk about speech synthesis as performance, why the uncanny valley is a bankrupt concept, and how academics can escape from studying corporate speech technology as if it's been bestowed by God.

## CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**;


*CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland Uk*

ACM ISBN 978-1-4503-5971-9/19/05…$15.00
https://doi.org/10.1145/3290607.3310422


**Figure 1: Work by Hiroshi Ishiguro, such as Repliee Q2 can mimic such human functions as blinking, breathing and speaking, with the ability to recognize and process speech and touch, and then respond in kind. His work is regularly cited along with the concept of the *uncanny valley*. "Repliee Q2. Taken at Index Osaka" by BradBreattie is licensed under CC BY SA 3.0**

**KEYWORDS**

speech synthesis; personal assistants; voice interaction; emotion; personality

**ACM Reference Format:**
Matthew P. Aylett, Benjamin R. Cowan, and Leigh Clark. 2019. Siri, Echo and Performance: You have to Suffer Darling. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI'19 Extended Abstracts), May 4–9, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3290607.3310422

## ACT 1: INTRODUCTION

*"Imitation is the sincerest form of flattery that mediocrity can pay to greatness."* - Oscar Wilde

Technology and innovation is often inspired by nature. However, when technology enters the social domain, such as creating human-like robots, or having human-like conversations, mimicry can become an objective rather than an inspiration. The mimicry objective allows engineers and researchers to dispense with complicated issues concerning why they are building something, and to focus instead on trying to copy something. Speech technology has, for many years, been dominated by this mimicry objective. It has especially pervaded evaluation techniques in speech technology where automatic systems are compared to human performance in terms of speech recognition, and directly compared with natural human speech to measure *naturalness*.

In some cases the mimicry objective is a specific obsession. For example, the renowned roboticist, Hiroshi Ishiguro (See Figure 1), has based his career on it, and much language based chatbot work, by focusing on the so called *Turing Test* has also regarded mimicry as an ultimate objective. In general, with speech technology, mimicry has become the norm because, firstly, engineers are too lazy to think of other ways of evaluating their systems (see [4, 9]), and secondly, designers and HCI professionals have largely ignored this interactive modality for years [1] and are only beginning to challenge this approach e.g[3, 7, 11].

For designers and human computer interaction (HCI) professionals, the mimicry objective has led to the endless inappropriate use of the phrase *uncanny valley*, which has taken on the role of some sort of physical law. This has been an especially useful tool for explaining away things that users don't like much and for generating poorly grounded speculation as to why that might be so. On careful examination the term *uncanny valley* doesn't mean much that can't be summed up by the comment that *mimicry is creepy*. For a detailed recent examination of the term, both in terms of trying to move it to a more useful context of Bayesian analysis of perceptual cues, and for a rigorous evaluation of the replicability of the uncanny valley effect (spoiler: not very replicable) see [10, 12].

Perhaps the reason the term *uncanny valley* has gathered such momentum is because the whole idea makes engineers and researchers look *stoopid* which will always get a laugh. The harder you

work to make your system better the more people hate it, hilarious. Unfortunately the whole thing is pre-disposed on 1. the desire to mimic, 2. the meaning of the word *familiarity* (originally *shinwa-kan* which is a neologism in Japanese), and 3. that there is a valley.

Here we will argue that the mimicry objective has serious limitations in speech technology, that systems can be experienced in many other ways and with mimicry, it is less of a valley than an abyss because the creepiest mimicry of all is the one you can't detect. Imagine, a new partner takes you to meet their mother and later you find it was an actor pretending to be their mother. The acted mother, for the sake of argument, was undetectable as a mimic at the time, but the creepiness is not really minimised by that.

The mimicry objective was recently highlighted by Google, who used their speech synthesis engine and dialogue system to book hairdresser's appointments over the phone without the human at the other end realising they were talking to an artificial system. Apart from raising significant ethical issues concerning deception, this highlighted the question *why mimicry?*

One reason to pursue mimicry is specifically so you *can* deceive a listener. There are many possible applications based on deception as we can already see with robot chatbots pretending to be hire car support staff. There is a tacit suspicion that the person you are communicating over text chat is a robot, but voice normally belies this suspicion. There are hundreds of evil, or moderately evil applications for this technology. For example, cold calling could be done without having to pay minimum wage to a human being, making it resource effective to increase cold calling by a factor of a hundred. That would certainly be progress.[1]

[1]This is a sarcastic use of the word *progress* as in not really progress at all

However, if we are not competing for the evil genius award in speech technology, mimicry has severe limitations. It's engagement, it's immersion, it's user experience that are key to future applications. It's, *dramatic pause*, performance darling.

## ACT 2: WHY DOES SPEECH SYNTHESIS NEED TO PERFORM?

*"We are all just actors trying to control and manage our public image. We act based on how others might see us"* Goffman, [5]

As Goffman points out the concept of performance can be extended to many social interactions. As computers enter the social domain this concept of performance also extends to the way an artificial system presents itself. Human beings have spent decades learning how to control the ways they speak, and to control the way ideas are presented in speech in order to facilitate story-telling. These same techniques are widely used to give speeches and communicate ideas outside of a traditional story-telling environment. Although the ability to mimic voices can be important in producing a refined performance (especially for satirical purposes), it is this sense of performance that creates the

experience. It is, in fact, crucial for satire to be discernibly not real, otherwise it cannot function as satire.

When we see an actor taking on the role of a character we know it's not real. It isn't presented as real. That's the difference between acting and impersonating. Take stage acting, there are many different techniques that are used to produce a performance. Most are not natural in human dialogue, voice projection, waiting for a cue, scripted filled pauses etc. etc. At heart it is all story telling. The art is often to appear natural without being natural, to convey the truth in a story within a cultural invention of a stage play. We don't call 911 when Mercutio gets stabbed, we don't scream, *"get him medical attention"* when he does a 5 minute passionate rebuke of the vendetta. We know it's not real, not natural, but also that it has artistic truth. The closest most speech technology gets to artistic truth is a decorating colour chart.

Control of speech in order to convey character and story are key skills for an actor. A sound-a-like is not a simple mimic, rather they are trying to convey an artistic representation of a person. When Alec Baldwin acts as Donald Trump we know he isn't really Donald Trump. He is very far from being a convincing copy of Donald Trump, he is better, he is a caricature.

Vocal mimicry is similar in many ways to the use of resemblance in pictorial art. Scott McCloud [8] discusses this as a "Big Triangle" with resemblance in one corner, abstraction in another and meaning in the third. *"The continuum from realistic to cartoony images represents increasing levels of iconic abstraction; removing (abstracting) an image from its retinal source, but still retaining its basic meaning."* Scott McCloud (http://scottmccloud.com/4-inventions/triangle/04.html)

We can regard the performance element of speech as an abstraction from real day-to-day speech, an abstraction that is conveying a deeper meaning in the way a cartoon conveys a deeper meaning of an image without the constraint of naturalness.

Speech synthesis research has avoided the complex issues of vocal performance. Rather, the focus has been to mimic natural speech well enough to be indistinguishable from human speech. All modern speech synthesis techniques are now corpus based. A large corpora of speech from single or multiple speakers is used to generate new speech with new content. To test naturalness you play this speech to a human listener and ask them if they can tell it's synthetic. To do this correctly you must have some held out speech data from the corpus which you can compare against. The default evaluation process is the mimicry objective.

For many speech synthesis applications this has proved a successful strategy. Many artificial systems that will speak are giving simple updates or well defined information. The default voice style has tended to be a news reader style of speech, neutral, warm, clear, avoiding any emotional engagement with the content being produced.

Synthesis has therefore been engineered purely to meet the mimicry objective. Mimicry of a limited vocal style, neutral, warm, citation speech. This is not a *fait accompli*. It is an implicit and almost

**Figure 2: Wax work of Jack Nicholson. Looks like Jack but doesn't perform like Jack and is very, very creepy. "Jack Nicholson figure at Madame Tussauds Hollywood" by lorenjavier is licensed under CC BY NC ND 2.0**

[2]Curses, see how pernicious this term is. As a reviewer pointed out, even we used it after spending half the paper saying no one should

[3]https://www.youtube.com/watch?v=5j2F4VcBmeo

universally adopted design decision, rarely tested, explored, or challenged. From Ivona to Siri to Google Home, speech synthesis is all like this.

With powerful machine learning algorithms now being applied to speech synthesis with considerable success, the mimicry objective dominates the field as it never has before. You can mimic anything, you just need enough coded data to do it. You want expressive speech, get a corpora of coded expressive speech, you want emotional speech, get a corpora of emotional speech. With deep neural nets speech synthesis can remain the design free zone it has always been.

None of this will produce systems that can perform. This is because performance requires creative artistry and mimicry cannot produce that on its own. This mimicry is also anchored to a perception of humanness as naturalness and as we have argued, performance in fact diverges from this naturalness.

Without the ability to perform, speech technology will be forever trapped in its *uncanny valley* of creepy mimicry[2]. It will forever be unable to engage an audience, to convey a deeper emotional meaning to the words it utters, or enter the social domain as anything more than a Q&A fact system that can turn on the lights.

How do we get speech synthesis to perform rather than just mimic? It requires design input, the artistic direction of a creative mind. It requires designers to engage with speech technology and suffer for their art. Given the lack of enthusiasm in the HCI and design community for getting involved in speech engineering research (unless forced to by a large salary paying corporation) we are a long way form realising a sense of performance from speech systems. In the next section we will give concrete examples of performance generated from such systems, what the technology is behind it, and how you might do it.

## ACT 3: THE SPEECH SYNTHESIS PERFORMANCE

Col Jessep: *I'll answer the question. You want answers?*
LTJG Kaffee: *I think I'm entitled to them.*
Col Jessep: *You want answers?!*
LTJG Kaffee: *I want the truth!*
Col Jessep: *You can't handle the truth!*
- A Few Good Men[3]

The performance of Jack Nicholson, together with inspired dialogue and film story telling that accentuated the tension in the famous court room scene resulted in the phrase *"You can't handle the truth!"* becoming an Internet meme. Satirised on Seinfeld, a Burger King advert, and many more (see https://tvtropes.org/pmwiki/pmwiki.php/Main/YouCantHandleTheParody for a more definitive list) the power of a performance launched this meme.

Standard speech synthesis technology is currently rubbish at acting. Lets take a scene from the film *Her* by Spike Jonze. In this scene Scarlett Johannson (whose voice pretty much defies standard analysis as it is so breathy) communicates vulnerability, anger, and despair. Here is an example performed by *Siri* 🔊 **https://tinyurl.com/y9vd29ov**. So although we have not even begun to explore how we might simulate Jack Nicholson at his scene chewing best we can certainly do better than that.

**Act 3 Scene 1: Content**

User: *Hey Siri: Are you a robot?*
Siri: *Well, I don't want to brag, but I got a B+ on the Turing test!*

The first element of performance comes from dramatic or engaging content. Given engaging the user is regarded as a *good thing* you may think speech interfaces would have a lot of engaging content to communicate. Sadly, perhaps because they are so bad at communicating engaging content, this is far from the case. Perhaps the closest we get is *Siri's* snarky rather smug techy jokes in response to stupid questions. As MacWorld puts it *"Bored? Need a good laugh? Here are some Funny things to ask Siri."* https://www.macworld.co.uk/feature/iphone/funny-things-siri-3656639/. Most of the staged responses are so sad and unfunny it is debatable whether they are funny or indeed result in anyone having a *"good laugh"*. But hey, they can fill up some of those endless lonely minutes of a sad pointless existence, and at least there was an attempt to use content to generate a sense of character.

Although we argue against the mimicry objective, the ability to mimic is at times very useful for generating a performance of a well known voice, given content. Perhaps the most challenging subject for this type of performance is Donald Trump, where getting him to say anything outrageous or humorous is eclipsed by his awful reality. In this clip we tried to generate content that we thought was funny for Donald Trump. However, we felt this clip only worked if it was discernibly not real and thus we added a little singing to make the content so completely surreal it might be funny. This clip is also a good example of how good speech synthesis mimicry has become 🔊 **https://tinyurl.com/y9ej7hkb**.

If we start with the belief that technology should be engaging and delightful then one of the first design questions that should be asked when faced with the problem of designing a speech system is *What is it going to Say? And Why?*.

**Act 3 Scene 2: Voice Style**

The underlying vocal style of a speech synthesis system is created by mimicking the vocal style in a collected corpus. Here mimicry, if not regarded as the final objective, and providing the obsession with naturalness is dropped, is very useful for creating a performance style voice. Voice talents have enormous experience using vocal styles to encourage a sense of performance and these can be harnessed. The main guideline is to steer away from the neutral new reader style voice, and explore

the collection of corpora with more radical vocal styles. Here are two examples of speech synthesis using the voice style to create a performance:

- **https://tinyurl.com/ybjhn7vl** : was a voice built for Kentucky Fried Chicken and used in the Harlan robot, a light hearted project which demonstrates very nicely that you can create something fun and engaging with mimicry but without *naturalness*. The robot doesn't look natural, the voice is a caricature to go with it. The short tongue in cheek video that went with this project (https://tinyurl.com/yau4nant), also does a great job making fun of the uncanny valley and we recommend watching this to get a feeling of how a final product might create a sense of performance.
- **https://tinyurl.com/ybpu4las** : Is an example of an acted child voice. It is the voice of a young boy created by a female voice talent. There is also an example of using over the top prosody manipulation to give a story telling feel to the clip.

Neither of these clips sound *natural*. Neither sound creepy. The first was carefully designed for the KFC robot, the second especially chosen for a child-based embodied conversational agent.

### Act 3 Scene 3: Expressive speech

XML is a common markup system for text which allows instructions to be placed around text. Control of prosody, speech rate etc. has been available both through industry standard XML such as Voice XML and speech synthesis markup language (SSML)[4], as well as using bespoke tag sets available from many providers. The speech synthesiser then renders the text and uses the XML tags to control a whole set of intonation an speech style functionality. Some examples are:

```
<sig rate="1.1" f0="0.9">Be gone from this
house.</sig>
```
Increase the speech rate by 10% and lower the pitch by 10%

```
<usel variant="1">Sorry, could you repeat
that?</usel>
```
Say this sentence again but slightly differently.

```
The <usel speed="0.8">cat</usel> sat on
the mat.
```
Choose longer units for the word cat to give more emphasis to that word and produce *"The CAT sat on the mat"*

```
<usel genre="stressed">You are getting
sleepy, please stop for a break!</usel>
```
In an in-car application forcefully instruct the driver to do something.

To give the reader a flavour of how a combination of appropriate voice style and expressive control can be at improving a performance listen to **https://tinyurl.com/y92rvve5** where an attempt is

[4]For a full list of tags available for speech synthesis markup language (SSML) or Voice XML which are tag sets that work across synthesis providers see: https://www.w3.org/TR/voicexml20/, https://www.w3.org/TR/speech-synthesis11/

made to improve the performance of the *Her* text. Its not exactly ScarJo but it does show we can do much better than something out of the box.

### Act 3 Scene 4: Post Processing

Sound designers play a fundamental part in creating a soundscape for audio adverts and radio as well as post-producing audio for video and film. Sound design should be a fundamental element of any speech synthesis output system. Generally, it is not. There is no reason that speech synthesis can not be mixed with background music, sound effects and filtered to come from a specific sound space. Moreover, voices can be post-processed to sound non-natural but still retain their communicative power. The following example uses post processing to create a demon and evil robot voice from standard speech synthesis voices[5] but also with emotional variation and other effects. See 🔊 **https://tinyurl.com/ya65bk2b**.

[5]The demand for non-natural voices that could be effectively used in games and other applications was so large CereProc even sells these types of voices.

## ACT 4: LET'S BUILD OUR OWN STUFF!

It is 5 years since Aylett et al [1] at Alt.Chi presented a critique of the HCI communities relationship with speech technology (and a critique of the speech technologists also). During that time we have seen a massive increase in the use of speech technology in interactive systems. Perhaps most notable being the appearance of smart speakers such as Echo and Google Home. Yet we have seen comparatively limited engagement from the mainstream HCI community in speech related topics.

The truth is that speech has always been a minority pursuit within the HCI community. A recent review of leading HCI venues [2](Under Review) found a total of 68 papers researching speech based issues spanning from 1990 to 2017. That is an average of 2.4 papers per year. That wouldn't even fill a session at CHI, and that is across 14 publication venues! Things are getting better, but from a limited base. Take MobileHCI, in 2010 there were no speech related papers, in 2018 there were 3. At CHI, searching for key words such as *speech synthesis* or *speech recognition* lead to few results, but seem more positive if you search for the term *voice user interface* or VUI. We saw approximately 20 papers on this subject in CHI 2018, one of which won a best paper award [13]. However, this gradual change may have a big impact in the long term because the key to overcoming to the mimicry objective is to use evaluations which do not produce it. HCI has the experience and potential to make this change as these early studies have begun to show.

Yet a lot of prominent recent work analyses the usage of corporate products, rather than systems designed and built outside the commercial sphere. Although trying to build prototype systems is (unnecessarily)hard[6], HCI offers many ways of prototyping than creating actual working system (probes, Wizard of Oz experiments - *WoZ*). WoZ could be an especially good starting point[2]. So the reason for this lack of prototyping is really due to a lack of engagement by academic designers

[6]More and more toolkits and packages have become available over time such as Kaldi[15]/Idlak[14] for ASR and Speech Sythesis, OpenDial[6], IrisTK (www.iristk.net), Aspect Prophecy (www.aspect.com/), and Amazon's Skills Kit (https://developer.amazon.com/alexa-skills-kit).

and HCI professionals. This means that large corporations dictate how this technology is designed, implemented and deployed. This is a poor state of affairs and needs to be rectified.

Analysing the systems in place, possibly repeating commercial studies that have remained confidential, has a place in expanding our knowledge of what happens already, but it will not innovate and change what speech interaction could be. Be under no illusion; the engineers at Amazon or Apple will not read your paper and take notice of it. They already know what you have found.

This is especially problematic given speech technology is used primarily by companies like Amazon and Google to have controlled channel to the user, which can be used to exploit personal data, and act as a barrier to open data and open communication. The whole point of Echo is to get you buy stuff from Amazon, hardly the broadest objective and one which is naturally antagonistic to supporting communities and vulnerable groups.

## ACT 5: FINAL ACT

As designers and HCI professionals it is time for you to get involved with speech technology. There is scope for doing ground breaking work. Let's tear ourselves away from the *Siri* and *Echo* style of service provision and actually try to build novel and provocative prototypes. Access to decent speech recognition and synthesis has never been better. We are frustrated at how speech technology has not fully been embraced by designers. But it should be.

## ACKNOWLEDGEMENTS

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 780890 (Grassroot Wavelengths).

## REFERENCES

[1] Matthew P Aylett, Per Ola Kristensson, Steve Whittaker, and Yolanda Vazquez-Alvarez. 2014. None of a CHInd: relationship counselling for HCI and speech technology. In *CHI'14*. ACM, 749–760.

[2] Leigh Clark, Phillip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew P. Aylett, João P. Cabral, Cosmin Munteanu, and Benjamin R. Cowan. 2018. The State of Speech in HCI: Trends, Themes and Challenges. *CoRR* abs/1810.06828 (2018). arXiv:1810.06828 http://arxiv.org/abs/1810.06828

[3] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. What can I help you with?: infrequent users' experiences of intelligent personal assistants. In *Human-Computer Interaction with Mobile Devices and Services*. ACM, 43.

[4] Benoit Favre, Kyla Cheung, Siavash Kazemian, Adam Lee, Yang Liu, Cosmin Munteanu, Ani Nenkova, Dennis Ochei, Gerald Penn, Stephen Tratz, et al. 2013. Automatic human utility evaluation of ASR systems: Does WER really predict performance?. In *INTERSPEECH*. 3463–3467.

[5] Erving Goffman. 1959. The Presentation of Self in Everyday Life.

[6] Pierre Lison and Casey Kennington. 2016. OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules. *Proceedings of ACL-2016 System Demonstrations* (2016), 67–72.

[7] Ewa Luger and Abigail Sellen. 2016. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *CHI '16*. ACM, 5286–5297.
[8] Scott McCloud. 1993. Understanding comics: The invisible art. *Northampton, Mass* (1993).
[9] Joseph Mendelson and Matthew Aylett. 2017. Beyond the Listening Test: An Interactive Approach to TTS Evaluation. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech 2017), Stockholm, Sweden*. 20–24.
[10] Roger K Moore. 2012. A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. *Scientific reports* 2 (2012), 864.
[11] Roger K Moore. 2017. Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In *Dialogues with Social Robots*. Springer, 281–291.
[12] Jussi Palomäki, Anton Kunnari, Marianna Drosinou, Mika Koverola, Noora Lehtonen, Juho Halonen, Marko Repo, and Michael Laakasuo. 2018. Evaluating the replicability of the uncanny valley effect. *Heliyon* 4, 11 (2018), e00939.
[13] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *CHI '18*. ACM, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214
[14] Blaise Potard, Matthew P Aylett, David A Baude, and Petr Motlicek. 2016. Idlak Tangle: An Open Source Kaldi Based Parametric Speech Synthesiser Based on DNN.. In *INTERSPEECH*. 2293–2297.
[15] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. 2011. The Kaldi speech recognition toolkit. *Proc. IEEE ASRU* (2011).