Robot, Alien and Cartoon Voices: Implications for Speech-Enabled Systems

Sarah Wilson[1], *Roger K. Moore*[2]

[1]Scribestar Ltd., Central Point, 45 Beech Street, London, UK
[2]Speech & Hearing Research Group, Dept. Computer Science, University of Sheffield, UK

sarahwilson608@gmail.com, r.k.moore@sheffield.ac.uk

Abstract

Since the early days of cinema and television, fictional characters such as 'robots' and 'aliens' have almost always been portrayed with correspondingly robotic or alien voices. Likewise, animated cartoon characters are often given quirky or unusual vocal characteristics. A wide variety of different techniques are used to create these imaginary voices, and the precise properties of each are usually carefully selected to fit the narrative context. In marked contrast, the voices of speech-enabled artefacts in the non-fictional world (such as Apple's *Siri* or Amazon Echo's *Alexa*) invariably sound humanlike, despite the risk that users might be misled about the capabilities of the underlying technology. The research reported here attempts to bridge the gap by collating and analysing a large corpus of robot, alien and cartoon voices with a view to understanding the relationship between particular vocal characteristics and the perceived 'persona' of the different characters portrayed. The results show that voice quality, delay/echo/reverberation and voice breaks are major factors, and it is concluded that a more in-depth understanding could lead to guidelines and tools that would allow designers of speech synthesis systems to create more *appropriate* voices in line with the 'affordances' of the target persona.

Index Terms: robot voice, alien voice, cartoon voice, vocal affordances, speech synthesis

1. Introduction

Since the early days of cinema and television, fictional characters such as 'robots' and 'aliens' have almost always been portrayed with correspondingly robotic or alien voices. Perhaps one of the most famous examples (certainly in the UK) is the harsh metallic (and terrifying) voice of the 'Daleks' - a race of hostile alien machine-like organisms which appeared in the BBC television science-fiction series *Doctor Who* in 1963. The Dalek's voice was produced using a technique known as 'ring modulation', and the catchphrase "*Exterminate!*" in a suitably monotonic tone has subsequently become an icon of evil.

In a similar manner, animated characters are often given quirky or unusual voices. For example, cartoon series made by Warner Brothers such as *Looney Tunes* and *Merrie Melodies* featured the popular character 'Daffy Duck' - an anthropomorphic black duck who spoke with a heavily exaggerated (and much imitated) lisp. Daffy was given this particular speech impediment specifically in order to reflect the possible consequences of a duck having an extended mandible.

A wide variety of different approaches are used to create these imaginary voices, from skilled voice actors to technology-based vocal manipulation. In each case, the aim is to select the vocal characteristics that fit the narrative context. In other words, such voices are specifically tailored to be *appropriate* to the character being portrayed, and this is regarded as a part-objective part-subjective highly-skilled activity.

> "*I usually first think, if these objects, places, robots or machines really existed what would they sound like? How would they be powered? What would be the actual physics of how they work? But if I find a sound isn't working within a scene, I'll abandon the science and go with what works emotionally.*"
>
> Ben Burtt [1]
> (sound designer for *R2-D2*, *ET* and *Wall-E*)

In marked contrast, the voices of speech-enabled artefacts in the non-fictional world (such as Apple's *Siri* or Amazon Echo's *Alexa*) are invariably designed to be as humanlike as possible using the latest technology for 'text-to-speech' synthesis [2]. For such devices, it is taken for granted that users prefer 'natural' voices over artificial or robotic voices. However, a human-sounding voice encourages users to overestimate the capabilities of the underlying technology, with negative consequences for subsequent interaction [3, 4, 5]. Nevertheless, consumer resistance and the lack of a suitable design methodology mitigate against the deployment of non-humanlike voices.

Based on research carried out by the first author as part of her MSc Dissertation [6], this paper attempts to bridge this gap by collating and analysing a large corpus of robot, alien and cartoon voices. The aim has been to gain some understanding of the relationship between particular vocal characteristics and the perceived 'persona' of the different characters portrayed [7]. It was hoped that this information could be used to better inform the design of future artificial voices in line with the principles espoused in [8]: "*It's better to be a good machine than a bad person*". Not only could this lead to the design of more appropriate voices for speech-enabled artefacts, but could also avoid entering the 'uncanny valley' [9] in which mismatched perceptual cues give rise to confusion and feelings of repulsion [10].

The paper is structured as follows: Section 2 reviews the ways in which voices may be manipulated, Section 3 describes the corpus of collected vocal samples, Section 4 presents an analysis of the data, and Section 5 summarises the results and concludes with suggestions for further work.

2. Robot, Alien and Cartoon Voices

The voices that were of interest in this study were not strictly limited to robots, aliens and cartoon characters; we were also concerned with talking machines, talking animals and indeed any real or imaginary artefact that might vocalise. In practice, there are three possible approaches to creating a desired vocal characterisation: (i) employ a skilled voice actor to adopt an unusual range, skill or voice quality, (ii) use a suitably configured speech synthesiser, or (iii) modify a voice in post-production by analogue manipulation or by digital signal processing [11]. The latter may be applied to natural or synthetic speech, hence it was of special interest to the study reported here.

2.1. Vocal Manipulation Techniques

There are many ways in which a voice (real or synthetic) may be manipulated in order to change some aspect of its characteristics, and several commercial products are available - particularly for use in professional music recording studios. One of the earliest devices was *Sonovox* (invented in 1939) which fed a sound source into a performer's throat so that they could use their tongue to shape the emitted sound. This arrangement enabled artefacts and musical instruments to articulate, and *Sonovox* was famously used in 1947 to make a piano talk in *Sparky's Magic Piano*. A more modern example is *Auto-Tune* [12] which was used by Cher in 1998 to create a unique pitch-jumping effect in her song "Believe". *Auto-Tune* was also used to create the voice of 'GLaDOS' (in *Portal 2*) and 'Brian' (for *Confused.Com*). In addition, there are many 'voice-changers' available on the internet: e.g., *Voxal* [13].

Techniques for vocal manipulation operate in either the time-domain or the frequency-domain [11]. Not only are these non-exclusive, but multiple techniques may be applied in any order. As a result, the number of potential effects is huge. Examples of specific manipulations are listed in Table 1.

Table 1: *Examples of vocal manipulation techniques (roughly in order of increasing complexity).*

Technique	Method
Time reversal	*delay line*
Speed change	*delay line*
Tremolo	*modulated amplitude*
Vibrato	*modulated pitch*
Ring modulation	*multiplication of two signals*
Comb filter	*short delayed version added to the original*
Echo	*long delayed version added to the original*
Flanger	*delay-modulated version added to the original*
Chorus	*multiple flangers with different delays*
Phaser	*phase-modulated version added to the original*
Reverberation	*convolution with room acoustic*
Pitch shift	*homomorphic filtering*
Harmony	*pitch-shifted version added to the original*
Filtering	f*requency shaping*
Formant shift	*altered vocal tract length*
Vocoding	l*inear prediction analysis-synthesis*

Many manipulations involve a 'low frequency oscillator' (LFO) that gives a time-varying character to the modified output. For example, vibrato and tremolo are achieved using an LFO to control amplitude or frequency respectively, and a "*wah-wah*" effect can be created by using an LFO to control the characteristics of a low-pass filter.

The consequence of each of these manipulations is to alter the tone and timbre of a voice in various ways. Of course, the initial voice could be natural or synthetic and could already be imbued with a particular characterisation. For example, the voice actors for the 'Daleks' (from *Doctor Who*) speak in a stilted monotone prior to their voice being subjected to modification by ring modulation using a 30Hz LFO.

2.2. Example Voices

In general, there are some fairly standardised ways that have been found to produce acceptable imaginary voices. For example, an effective robot voice can be achieved by a small increase in pitch, followed by adding back the original (c.f. 'harmony') and introducing some echo. On the other hand, a reasonable alien sound may be created by decreasing the pitch and applying a chorus effect. Finally, a cartoon-like voice may be produced by applying a large pitch increase followed by a chorus effect and added tremolo. These, and many others, are often available as 'presets' in voice-changing products such as *Voxal* [13]. Specific examples of characters with voices created through the application of the techniques mentioned in Section 2.1 are listed in Table 2.

Table 2: *A selection of characters with manipulated voices.*

Character	Production	Technique
Aliens	*Toy Story*	chorus
Celestria	*Power Rangers*	phaser
Dalek	*Doctor Who*	ring modulation
Jinx	*Spacecamp*	pitch increase
King Laufey	*Thor*	pitch decrease
Klutzy	*Robot Holocaust*	comb filter
Marvin	*Hitchhikers Guide*	vibrato
Max	*Flight of the Navigator*	reverberation
Mechanoids	*Doctor Who*	tremolo
Proteus	*Demon Seed*	flanger
Tassadar	*Starcraft*	reverse reverb.
Ultron	*Ultimate Alliance*	echo

3. The 'RAC' Corpus

A corpus of relevant voices was collected by searching the internet for films and TV series with robot, alien and cartoon characters, online reviews, forums and YouTube's recommender side bar. Further suggestions were obtained by uploading a publicly editable document and providing anonymous social media users an opportunity to contribute suggestions. Voices were not limited to any accent, ethnicity or age range, nor were they required to be speaking a known human language. However, it was decided that there must be some human element to each voice, so voices made from animal sounds or beeps and whistles (such as 'Chewbacca' or 'R2-D2' from *Star Wars*) were excluded.

All characters were labelled as being either 'robot' or 'alien', as well as given an estimate of their size, gender, material (metal or organic) and good, evil or neutral 'persona'. Voices were also labelled with subjective impressions of delay, harmony, modulation or speed change, as well as objective vocal measurements such as pitch (mean and standard deviation), jitter, shimmer, harmonic-to-noise ratio (HNR) and number of voice breaks. The latter were computed using *Praat*, a standard open-source speech analysis tool [14]. Vocal features such as breathy, creaky or whispery voice quality were also labelled.

Cartoon voices were assigned as 'robots' or 'aliens' on the basis that the latter category includes anything that does not exist in the real world. So a talking chipmunk is an alien in the same way that a 'Dalek' (from *Doctor Who*) is an alien because, although chipmunks exist, they do not speak. So, for example, the cartoon character 'Stitch' (from *Lilo and Stitch*) was classed as an alien, whereas the 'Iron Giant' (from a cartoon series of the same name) was classed as a robot. In addition, the robot category was more specific; not only could it include characters that were made of metal, but it could also be subdivided into 'cyborgs' (human-robot combinations), computers (such as 'HAL 9000' from *2001: A Space Odyssey*) and automobile robots (such as 'Optimus Prime' from *Transformers*, 'KITT' (from *Knight Rider* and 'Crimebuster' from *Heart Beeps*).

In total, 93 voices were collected and annotated, with samples spanning a period from 1939 (*The Wizard of Oz*) to 2015 (*Chappie*) - see Table 3. We are not permitted to share the data.

Table 3: *List of the 93 robot, alien and cartoon voices in the 'RAC' corpus.*

AlvinChipmunk	BigHero-Baymax	BicentMan-Galatea	BSG-Cylon	CaptainScarlet-mysterons
Chappie	Confused.com-Brian	Cyborgcop	DarkStar-Bomb20	DemonSeed-Proteus
DrWho-Icewar	DrWho-Cybermen	DrWho-Dalek	DrWho-Davros	DrWho-GreatIntelligence
DrWho-K9	DrWho-Mechaniod	DrWho-Silence	DrWho-Silence2	Dumbo-Casey
ET	Evolver	FlightoftheNav-Max	ForbiddenPlanet-Robby	GhostITShell-Proj2501
GIJ-CobraCommander	GOTG-Groot	GuyverDarkHero-Guyver	HarryPotter-Dobby	Heartbeeps-Crimecar
Heartbeeps-Val	HGTTG-Penguin	HGTTG-Vogon	HGTTG-Marvin	Hulk-Abomination
InspGadget-DrClaw	Intersteller-TARS1	Intersteller-TARS2	IronGiant-Giant	IronMan-Jarvis
JudgeDredd-ABCWar	KnightRider-Kitt	Lilo&Stitch-Stitch	LostInSpace-B9	LOTR-Gollum
LOTR-MouthSauron	LOTR-Treebeard	Marv-AlutAlianc-Ultron	Marv-SHSquad-Ultron	MenInBlack2-Zarthan
MichWeb-Cheesoid	Moon-Gerty	Portal2-GlaDos	PowerRangers-Alpha	PowerRang-Cestria
PowerRangers-Goldar	PowerRangers-Zordon	QuantumQuest-Fear	ReturnToOz-Ticktok	Robocop
RobotHolocaust-Klutzy	Rocky-Sico	ShortCircuit-Johnny5	SmashRobots	Spacecamp-Jinx
SpaceOdyssey-HAL	SparkyPiano	Starcraft-Tassadar	StarTrek-Borg	StarWars-C3PO
StarWars-DarthVador	StarWars-EmperorP	StarWars-EV-9D9	StarWars-JabbaTheHutt	StarWars-JarJarBinks
StarWars-Yoda	TheBlackCauldron-HornedKing	Tekken-Yoshi	TheBlackHole-Vincent	TheHobit-Smaug
Thor-KingLaufey	TMNT-Shredder	ToyStory-Aliens	Transformers-Deceptacon	Transformers-OptimusPrime2
Transformers-OptimusPrime-low	Tron1982-MCP	TronLegacy-Gem	Walle-Eve	Walle-Auto
Walle-Walle	WhatHappenedToRJ-RobotJones	WizardOfOz-Witch		

4. Data Analysis

4.1. General Observations

Of the 93 voices in the 'RAC' corpus, 50 were classed as 'robot' and 43 were classed as 'alien'. 64 were single recordings, 29 were concatenated samples and a large number (81) had audible background noise. Interestingly, 87 were categorised as 'male', but only 6 as 'female'. The most common effect in the corpus was echo or delay (66), followed by harmony (45), some form of modulation (40), slowing down (15) and speeding up (4). One of the more interesting effects was reverse reverberation in a character called 'Tassadar' (from *Starcraft*) which created an unusual inhalation sound prior to the speech. Pitch-changing effects were also found, such as quantised pitch shifts in 'Brian' (from *Confused.Com*) and a monotone in the 'Cylons' (from *Battlestar Galactica*). In terms of phonetic voice quality, 8 voices were creaky, 6 were whispery, 6 hoarse, 3 breathy and 3 tense/glottal.

In order to determine the relationship between the character voices in the 'RAC' corpus and normal unaltered human voices, 89 male and 42 female voices were selected from the TIMIT corpus [15] as 'controls' for comparison. The natural human voices were subjected to the same analysis techniques as the character voices, and various statistics were calculated across both sets.

Table 4: *Correlations between measured vocal parameters and various simple 'personae'.*

	Character-Control	**Robot-Alien**	**Good-Evil**
Pitch (μ)	-0.1470	0.2225	0.0290
Pitch (σ)	-0.4732	0.1954	0.1439
Jitter	-0.5535	0.1865	0.3514
Shimmer	-0.6857	0.2470	0.3968
HNR	0.6646	-0.1868	-0.3568
Delay	-0.6905	0.0705	0.1871
Harmony	-0.5494	-0.1095	-0.0965
Breaks	-0.6550	-0.1133	-0.0307

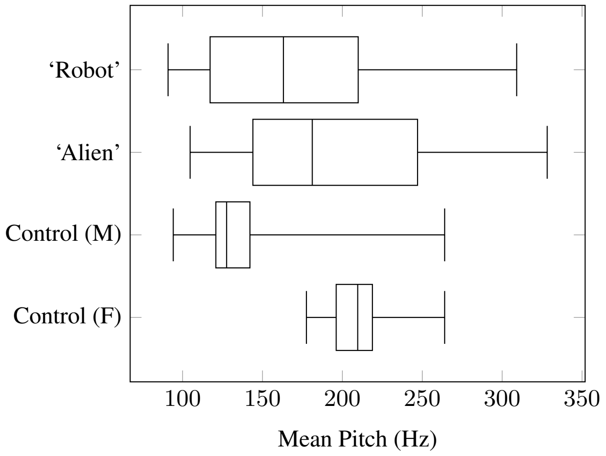

Figure 1: *Distribution of mean pitch for 'robot' and 'alien' character voices compared to unmanipulated male and female control voices.*

4.2. Summary Statistics

Correlations were computed between the various parameters and simple 'persona' characteristics (such as characters *vs.* controls, 'robots' *vs.* 'aliens', and 'good' *vs.* 'evil') - see Table 4. As might be expected, the results indicate that character voices differ from normal (control) voices on most of the measures, reflecting the manipulations that have taken place (especially in delay, voice quality and breaks). The difference between 'robot' voices and 'alien' voices not only shows up (to a modest extent) in the voice quality measures, but also in the mean pitch. It seems that the 'aliens' in the corpus had somewhat higher pitched voices than the 'robots' (unlike the *Voxal* pre-set mentioned in Section 2.2), but both have a much larger range than controls - see Fig. 1.

As mentioned, Table 4 suggests that voice quality plays a role in distinguishing the various 'personae'. For example, Fig. 2 shows that 'alien' voices have a slightly more unusual voice quality than 'robot' voices, both of which are quite different from unmanipulated control voices. Table 4 also indicates that voice quality plays a role in distinguishing the voices of 'good' characters from 'evil' characters - see Fig. 3.

It can also be seen from Table 4 that an important difference between the character voices and the controls is that the characters often contain an unusually large number of breaks (often caused by the use of a low frequency modulation effect). Fig. 4 illustrates an almost complete lack of overlap between the two groups for this parameter, with one character in particular - the 'Mechanoids' (from *Doctor Who*) - showing up as the most extreme example.

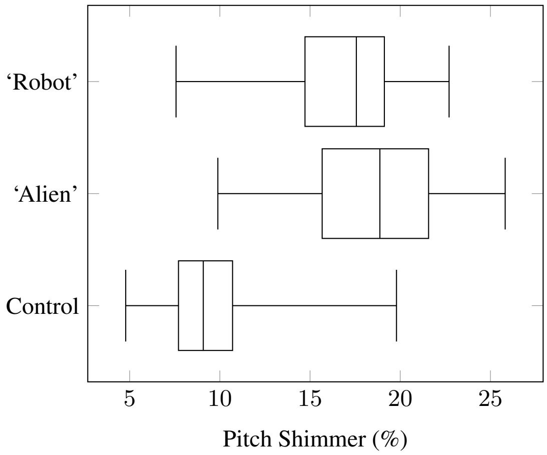

Figure 2: *Distribution of pitch shimmer for ‘robot’ and ‘alien’ character voices compared to unmanipulated control voices.*

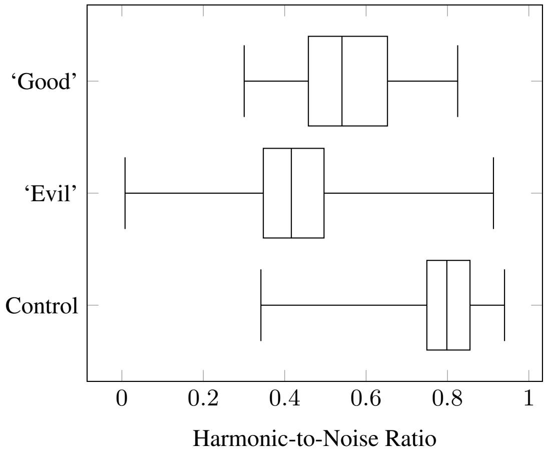

Figure 3: *Distribution of voice qualities (based on measured HNR) for ‘good’ and ‘evil’ character voices compared to unmanipulated control voices.*

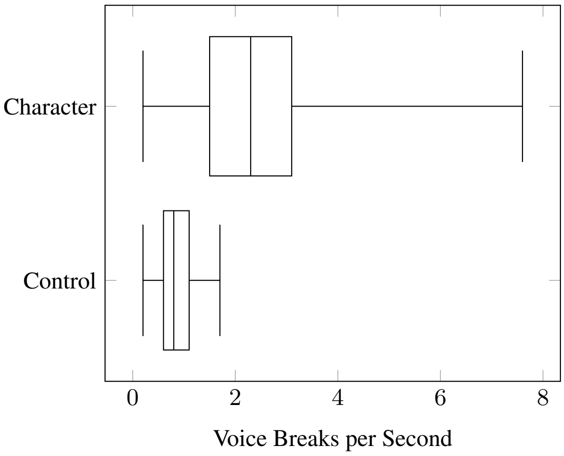

Figure 4: *Distribution of the number of voice breaks for character voices compared to unmanipulated control voices.*

4.3. Principal Component Analysis

In addition to computing the statistical correlations between various vocal parameters, the ‘RCA’ corpus was also analysed using principal component analysis (PCA) [16]. It was found that four components accounted for 67.4% of the total variance. The first component appeared to correspond to aspects of voice quality, serving to distinguish ‘good’ personae from ‘evil’ personae. The second component was linked to pitch which, with more female character data, could have related to gender. The third correlated to size, voice breaks and material, which could be regarded as aspects of ‘appearance’. The fourth component related to echo, delay and reverberation, which seemed to distinguish fictional from non-fictional characters.

As examples, the extreme characters for the first principal component were ‘GLaDOS’ (from *Portal 2*), ‘Galatea’ (from *Bicentennial Man*) and ‘Jar Jar Binks’ (from *Star Wars*) as the most ‘good’ characters, and ‘The Silence’ (from *Doctor Who*), ‘Abomination’ (from *The Incredible Hulk*) and the ‘Daleks’ (from *Doctor Who*) as the most ‘evil’ characters. Perceptually, the first three (the ‘goodies’) have near normal voice quality, whereas the final three (the ‘baddies’) are heavily manipulated. Extreme characters for the second principal component were ‘Gerty’ (from *Moon*) at the low-pitch end and ‘Robot Jones’ (from *Whatever Happened to Robot Jones?*) at the high-pitch end.

Overall, it is interesting to note that the PCA revealed that most of the variance in the data arises as a result of personality rather than appearance, thereby confirming the importance of a character’s voice as a key indicator of ‘persona’.

5. Summary and Conclusion

The research reported in this paper has attempted to bridge the gap between voice-enabled artefacts in the fictional and non-fictional worlds by collating a large corpus of robot, alien and cartoon voices and comparing them with normal control voices. The aim has been to gain some understanding of the relationship between particular vocal characteristics and the perceived ‘persona’ of the different characters portrayed. It was hoped that this information could be used to better inform the design of future artificial voices in line with the principle that “*It’s better to be a good machine than a bad person*” [8].

The study has confirmed that the majority of robot, alien and cartoon voices are manipulated to fit the narrative context, and that such manipulations are correlated with different ‘personae’ in predictable ways. In particular, it has been shown that voice quality, delay/echo/reverberation and voice breaks are major factors that influence the perceived character. These results, coupled with existing evidence that it is possible to infer a speaker’s physical attributes such as age, weight and height from their voice [17], lend support to the view that future voice-enabled artefacts should not be designed to be as humanlike as possible, but should adopt vocal characteristics that are *appropriate* to their physical makeup and cognitive capabilities.

Ultimately, what is required is a set of guidelines (and associated tools) that would allow the designers of voice-enabled artefacts to ‘dial-up’ appropriate vocal characteristics in line with the visual and behavioural affordances of the target ‘persona’. In order to achieve this, a more in-depth understanding of the relevant dependencies is required than the preliminary results reported here, and this is the subject of ongoing research.

6. References

[1] J. Ludwig, "Animation Sound Design: Ben Burtt Creates the Sounds for Wall-E (Part 2 of 2)," 2009. [Online]. Available: https://www.youtube.com/watch?v=eySh8FOUphM

[2] P. Taylor, *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press, 2009.

[3] C. Nass and S. Brave, *Wired for Speech: How Voice Activates and Advances the Human-computer Relationship*. Cambridge, MA: MIT Press, 2005.

[4] R. K. Moore, "Spoken language processing: Where do we go from here?" in *Your Virtual Butler, LNAI*, R. Trappl, Ed. Heidelberg: Springer, 2013, vol. 7407, pp. 111–125.

[5] ——, "From talking and listening robots to intelligent communicative machines," in *Robots That Talk and Listen*, J. Markowitz, Ed. Boston, MA: De Gruyter, 2015, ch. 12, pp. 317–335.

[6] S. Wilson, "Characteristics of Robot, Alien and Cartoon Voices," MSc in Computer Science with Speech and Language Processing, University of Sheffield, 2015.

[7] R. K. Moore, R. Marxer, and S. Thill, "Vocal interactivity in-and-between humans, animals and robots," *Frontiers in Robotics and AI*, vol. 3, no. 61, 2016.

[8] B. Balentine, *It's Better to Be a Good Machine Than a Bad Person: Speech Recognition and Other Exotic User Interfaces at the Twilight of the Jetsonian Age*. Annapolis: ICMI Press, 2007.

[9] M. Mori, "Bukimi no tani (the uncanny valley)," *Energy*, vol. 7, pp. 33–35, 1970.

[10] R. K. Moore, "A Bayesian explanation of the Uncanny Valley' effect and related psychological phenomena," *Scientific Reports*, vol. 2, no. 864, 2012.

[11] J. Rose, *Audio Postproduction for Film and Video*. Taylor & Francis, 2012.

[12] *Auto-Tune (by Antares Audio Technologies)*. [Online]. Available: http://www.antarestech.com

[13] *Voxal Voice Changer (by NCH Software)*. [Online]. Available: http://www.nchsoftware.com/voicechanger/

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." [Online]. Available: http://www.praat.org/

[15] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Darpa TIMIT acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, pp. 385–403, 1993.

[16] I. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer-Verlag, 2002.

[17] R. M. Krauss, R. Freyberg, and E. Morsella, "Inferring speakers' physical attributes from their voices," *Journal of Experimental Social Psychology*, vol. 38, no. 6, pp. 618–625, 2002.