10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

Voice Puppetry: Exploring Dramatic Performance to Develop Speech Synthesis

Matthew P. Aylett[1], *David A. Braude*[1], *Christopher J. Pidcock*[1], *Blaise Potard*[1]

[1]CereProc Ltd., UK

{matthewa, dave, chris, blaise}@cereproc.com

Abstract

Technology and innovation is often inspired by nature. However, when technology enters the social domain, such as creating human-like voices or having human-like conversations, mimicry can become an objective rather than an inspiration. In this paper we argue that performance and acting can offer a radically different design agenda to the mimicry objective. We compare a human mimic's vocal performance (Alec Baldwin) of a target voice (Donald Trump) with the synthesis and copy resynthesis of a cloned synthetic voice. We show the conversational speaking style of natural performance is still a challenge to recreate with modern synthesis methods, and that resynthesis is hampered by current limitations in speech alignment approaches. We conclude by discussing how voice puppetry - where a human voice is used to drive a synthesis engine - could be used to advance the state-of-the-art and the challenges involved in developing a voice puppetry system.

1. Introduction

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

Recently Google demonstrated a conversational agent that booked hair dresser appointments and tables in a restaurant. The agent was capable of mimicking a human so closely over the phone that the person dealing with the call was not aware they were dealing with an artificial system[1]. Apart from raising significant ethical issues concerning deception, this highlighted the question *why mimicry?*.

Aylett et al [1] focused on the advantages and disadvantages of what they term the *mimicry objective* and presented an alternative approach which seeked to leverage the power of dramatic performance as a means of designing synthetic voices. In summary, Aylett et al argue the advantages of the mimicry objective are:

1. Evaluation metrics can be well specified.
2. Machine learning is easy to apply.
3. Research objectives are easy to describe.
4. Mimicry signals a metaphor for interaction to users.

Whereas the limitations are:

1. Mimicry is creepy (see [1] for a fuller discussion of how this relates to the so called "uncanny valley").
2. It avoids the responsibility of evaluating the system in the field or in an application environment.
3. It can lead to the user over-estimating a system's competence [2–4].
4. We lose all the benefits of **not being real**, for example, non-natural systems are perceived as non-judgmental [5–7].

In this paper, we extend Aylett et al [1], by exploring how we may use dramatic performance as a concrete method for improving and developing speech synthesis. We take a dramatic comic performance from a human mimic of Donald Trump (Alec Baldwin) and use this to guide synthesis from a synthetic voice cloned from Donald Trump's presidential addresses. We compare renditions of Alec Baldwin's performance in three conditions: 1. Default WaveNet style text to speech, 2. Copy resynthesis with Alec Baldwin's F0 and phone durations used to control synthesis, 3. A version of copy resynthesis with the F0 and phone durations heuristically modified to be closer to the source speaker (Donald Trump).

2. Copy Resynthesis and Voice Puppetry

Copy resynthesis is a technique where the parameterisation of human voice audio is used to directly control a speech synthesis rendition. It is a well established technique used to develop and evaluate speech synthesis systems (e.g. [8, 9]. For example, if we hypothesise that a system is perceived as unnatural because of poor prosody, we could use pitch and duration from a natural utterance, rather than from a generative model, and see if this still holds. Two assumptions undermine the copy resynthesis approach, first, the parameters are not normally completely independent, meaning errors can be caused by mismatching copy resynthesis parameters to synthesis parameters, second, there are often many acceptable natural renditions of a specific phrase and a single example of natural speech does not cover this space adequately. Furthermore, in modern DNN based WaveNet style speech synthesis systems (e.g. [10]), parameters are not always explicitly modelled or easily accessible, making copy resynthesis challenging.

However, copy resynthesis is still a powerful and effective tool for exploring how speech may be decomposed, controlled and mimic a human voice. In addition, copy resynthesis can be used as a basis for a *voice puppetry* system. Such a system allows an input natural speech to control the output speech for a target voice. This contrasts, but is related to voice morphing, where a source speakers voice is converted directly into a target speakers voice without the requirement of a speech synthesis system. So called *Puppetry* systems are commonly used to control the rendering of graphics characters and lip-syncing, where a human subjects movements are mapped onto a potentially very different body form. The ability to extend this control to the vocal performance of an artificial characters voice, despite ethical dangers, is an important area of research.

[1]https://www.theguardian.com/technology/2018/may/08/google-duplex-assistant-phone-calls-robot-human

10.21437/SSW.2019-21

3. The Difference between Mimicry and Dramatic Performance

To understand the differences between mimicry and dramatic performance it is important to consider context. Mimicry is, and always will be, an important element of performance. However, performance is more concerned with communicating a sense of character and narrative than being an undetectable copy of the source voice. Impersonation, where the objective is to deceive a listener, is a context that requires pure mimicry. The human mimic must create a completely natural version of the target voice to accomplish a communicative goal, for example retrieving confidential information. For a performance, where the objective is satire or story telling, there is no requirement to deceive the listener, rather creating a caricature, where the target speaker's voice style and delivery is exaggerated, is a common approach. For example, `https://tinyurl.com/y5ebpe22` an example of Donald Trump giving a presidential address, whereas `https://tinyurl.com/yyg6retc` is an example of Alec Baldwin satirising Donald Trump giving a presidential address. Alec Baldwin is not impersonating Donald Trump, we can tell it is a satire because it is a performance, we can tell it is not the original speaker but that doesn't matter. If our overall goal is to perform, just mimicking a natural voice will never be enough.

From a practical perspective, in terms of speech synthesis, performance can be regarded as mimicry overlapping a subset of expressive speech (see [11] for a review of speech sythesis work in epxressive and emotional speech). Identifying the expressive speech techniques used in human vocal performance and intergating them with mimicked speech synthesis is a big research challenge. In this paper we start by focusing on pitch, timing and speech rate and test the hypothesis that if we use Baldwin's performance based on these parameters we can improve the quality and performance of a WaveNet style synthetic voice cloned from Donald Trump. Furthermore we can potentially improve Baldwin's performance by using voice puppetry to make it sound more like Donald Trump without losing the performance elements.

4. Method

For this pilot study we selected 8 utterances from Alec Baldwin's performance in SNL's Cold Open (Alec Baldwin Returns to Mock President Trump's Emergency Declaration) taken from YouTube `https://www.youtube.com/watch?v=8vQlhWBvAwY`. Utterances were chosen to avoid applause, laughter and minimise background noise and disfluency. We down sampled to 16kHz and aligned the audio based on a text transcription using a proprietary aligner.

Two baseline synthesis examples were generated from the transcription of Baldwin's utterances using a synthetic voice cloned from Donald Trump's presidential addresses.

tacotron Tacotron/WaveNet based system. Audio was resampled to 22khz. A Tacotron front end producing 80-dimension mel spectrograms was trained using an open-source toolkit by Rayhane Mama[2]. The front-end was modified to produce synthesis from phonetic sequences rather than English text. The WaveNet model was trained to produce 16-bit (mixture of logistic distributions) 22khz output over 430,000 steps using Ryuichi Yamamoto's WaveNet Vocoder[3], initialised using weights pretrained on the LJ Speech dataset [12]. Exponential moving average (EMA) weights were used at synthesis time.

cprcdnnv2 CereProc real time WaveNet style system. This system uses the CereProc language front end, a DNN system for parameter generation, and a proprietary neural vocoder. The system runs at many times real time and includes prosody targeting.

In section 5 we show the differences between the aligned Baldwin data and the Trump synthesis data in terms of duration, pitch and pausing. Following this analysis we then created two copy resynthesis versions of each section.

auto Pause length, pitch and phone duration targets were used to generate synthesis using the *cprcdnnv2* system. Pause and phone durations were used explicitly, pitch was used as an advisory target to the DNN vocoder and depending on other parameters would potentially be altered in the generated speech.

modified Short phrase breaks were removed, speech rate was slowed to match Donald Trump's synthetic voice more closely, and phrase finality was heavily enforced by extending the duration of phrase final phones. These heuristic modifications were made in response to issues in the enforced alignment identified in the section 5.

Amazon Mechanical Turk was then used to present a MUSHRA-style test to 36 subjects. The first 18 compared all four synthetic utterances together with the original Baldwin rendition, the second group of 18 subjects listened to the four synthetic utterances alone. The test was performed with and without the Baldwin natural examples because: 1. Baldwin's voice is different from Trump's and 2. The resynthesis examples noticeably recreate Baldwin's performance when you hear them side-by-side. Thus, by including the Baldwin examples the test implicitly suggests the objective is to sound like this natural reference, whereas without them the test is a traditional MOS style naturalness test without a reference. To listen to stimuli from the experiment go to `https://tinyurl.com/yyw38ou7`.

5. Comparison of Baldwin data with Trump synthesis data

Figure 1 shows the differences between synthesis generated using the voice clone of Donald Trump vs Alec Baldwin's performance of Donald Trump. The first observation is that Alec Baldwin's style of speech is conversational whereas the presidential addresses used to build the Trump voice are citation speech. The average phone rate for Baldwin is 90ms vs 110ms for the synthesis output. The pitch range is higher and Baldwin includes many short pauses. Some of these are caused by disfluencies and do not have typical phrase final F0 shifts.

The assumption that a performance, because it is prescripted, will be closer to citation speech is not shown for this data. The Trump performance Baldwin creates is generated very quickly to keep up with current affairs and a formal script may not be used, more likely, it is a semi-scripted improvisation.

This presents a major challenge for alignment. Accurate forced alignment of conversational speech with included disfluencies is problematic. Elision, hypo-articulation and rapid

[2] `https://github.com/Rayhane-mamah/Tacotron-2`

[3] `https://github.com/r9y9/WaveNet_vocoder`

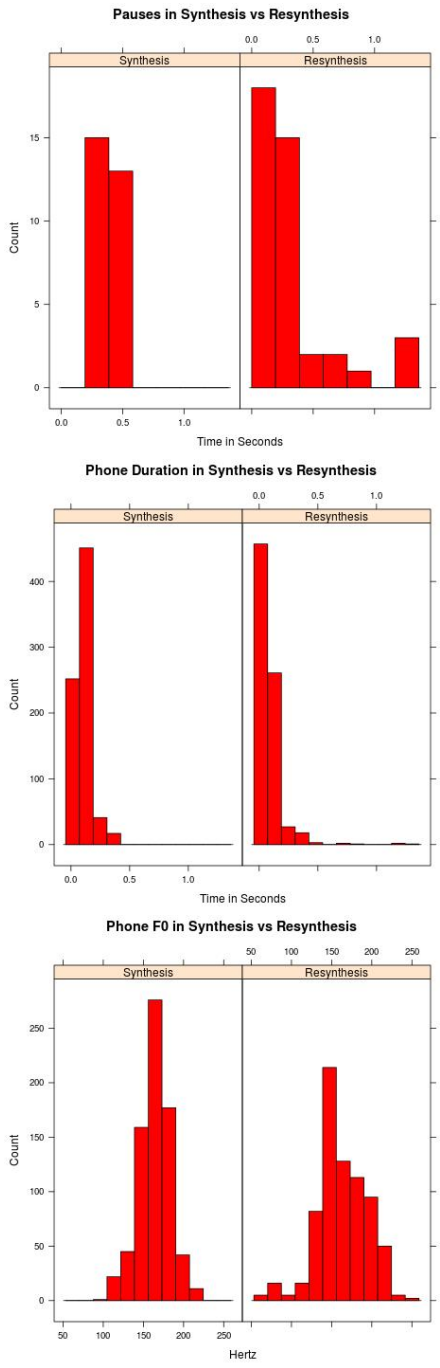

Figure 1: *Top: histogram of mid-utterance pause lengths in synthesis (controlled directly by punctuation) and resynthesis (controlled by alignment of Baldwin data), Middle: Phone duration modelled by synthesis system vs alignment of Baldwin data, Bottom: Pitch modelled by synthesis system vs pitch analysis of Baldwin data.*

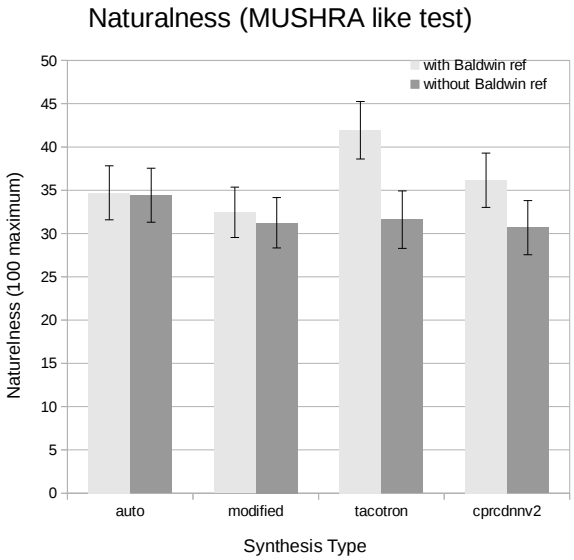

Figure 2: *Means of naturalness scores for all four synthesis types (The Baldwin natural reference had a mean of 68) with and without the Baldwin natural reference. Error bars show ±1 standard error.*

speech rate can cause significant issues. Furthermore pitch extraction from conversational speech can also be problematic with semi-voiced regions leading to pitch analysis errors. In addition, requesting pitch and duration values from a DNN synthesis system that are not sufficiently represented in the training data can also cause synthesis errors.

In order to try and reduce alignment errors in the resynthesis, and in order to make the copy resynthesis more closely match the original Trump data used to build the voices, the following heuristic modifications were carried out to produce a modified copy resynthesis **modified**:

1. All breaks less than 100ms were removed.
2. Phone durations were forced to be 1.2 times the original aligned duration.
3. At phrase final position a minimum word average phone length of 130ms was enforced to a maximum of 1.5 times the original duration.
4. All phone durations over 400ms were cut to 400ms. Pitch targets below 100Hz were set to 100Hz and above 220Hz were kept at 220Hz.

6. Results from the listening test

A repeated measures ANOVA by-subjects analysis of listening test was carried out with **synthesis type** (tacotron, cprcdnnv2, auto, modified) as a within subject factor, and the presence of a natural Baldwin reference (**natref**) as a grouping factor. Mauchly's Test of Sphericity showed sphericity could not be assumed. Synthesis type was a significant factor ($F(3, 32)=4.016$, $p<0.025$) with Greenhouse-Geisser correction, and there was a significant interaction between natref and synthesis type (($F(3, 32)=4.955$, $p<0.01$). **Natref** alone did not have a significant effect.

We would have expected removing the natural reference would have artificially inflated the overall naturalness score of all the synthesis types as well as increasing the differences between them. In fact no overall significant effect of removing the natural reference was observed (see Figure 2). Removing the Baldwin reference only inflated the perceived naturalness of the pure synthesis conditions - tacotron and cprcdnnv2. Hence the strong interaction between **natref** and **synthesis type**. The overall **synthesis type** effect was caused by the modified copy synthesis being regarded as worse in terms of naturalness compared to both the automatic copy resynthesis and the pure synthesis conditions ($p<0.025$ with a Bonferroni correction for multiple comparisons). However given the small number of materials (only 8 utterances) it would be dangerous to infer too much from these results.

7. Discussion

This is an early stage study that aimed to explore the use of copy resynthesis for determining dramatic performance techniques of a human mimic. We underestimated the conversational nature of the performed speech we analysed and this has strongly effected our results.

1. The high speech rate, disfluencies, and pitch variation are hard for synthesis systems to model and align. Thus our copy resynthesis suffered from problems caused by alignment errors and pitch analysis errors.
2. Furthermore the conversational nature of the text being synthesised reduced the perceived naturalness we would expect to gain from a good WaveNet style speech synthesis system. It was difficult to cut utterances from the source data which avoided background noise and were also textually coherent. For example losing the initial text in the phrase *"[I'm not going], to do the voice"* due to background laughter.
3. The lack of context, very important for resolving disfluencies, is lacking in this sort of utterance by utterance naturalness test. This is shown by the low perceived naturalness of the Baldwin natural utterances (mean of 68 when we might expect 80/90). It raises the issue that utterance by utterance evaluation will not tease out dramatic performance techniques that may extend over many paragraphs.

It appears that having the source of the copy synthesis as a natural reference for puppetry voices biases naturalness against pure synthesis renditions that have an alternative prosodic realisation. A much better comparison would have been to used natural speech from Donald Trump. However, because we cannot control the utterances produced by our performer in this study such a comparison was impossible.

8. Conclusion

The initial idea, using Alec Baldwin's performance to control a Trump cloned voice synthesiser is a simple one. In reality it is significant challenge. In future work we will instead use a dramatic performance that can be controlled to match a source speakers natural speech, firstly to explore acted conversational style speech rather than actual a conversational style speech, and to consider a more effective evaluation process for assessing the success or failure of applying copy resynthesis to guiding the dramatic performance of a speech synthesis system.

The recent developments of WaveNet style speech synthesis have opened a large range of potential use cases. Voice puppetry using parametric systems in the past were severely limited by the quality of the parametric systems. In many cases WaveNet style synthesis is hard to differentiate from natural speech, thus commercial voice puppetry systems have become a real possibility. This comes with advantages, in terms of ways of exploring speech synthesis and expressive speech synthesis, and also dangers in terms of unethical use of cloned speech synthesis voices.

9. Acknowledgements

This work was supported by the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No 780890 (Grassroot Wavelengths).

10. References

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

[2] E. Luger and A. Sellen, "Like having a really bad pa: the gulf between user expectation and experience of conversational agents," in *CHI '16*. ACM, 2016, pp. 5286–5297.

[3] R. K. Moore, H. Li, and S.-H. Liao, "Progress and prospects for spoken language technology: What ordinary people think." in *INTERSPEECH*. San Francisco, CA, 2016, pp. 3007–3011.

[4] B. R. Cowan, N. Pantidi, D. Coyle, K. Morrissey, P. Clarke, S. Al-Shehri, D. Earley, and N. Bandeira, "What can i help you with?: infrequent users' experiences of intelligent personal assistants," in *MobileHCI '17*. ACM, 2017, p. 43.

[5] G. M. Lucas, J. Gratch, A. King, and L.-P. Morency, "It's only a computer: Virtual humans increase willingness to disclose," *Computers in Human Behavior*, vol. 37, pp. 94–100, 2014.

[6] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *AAMAS '14*, 2014, pp. 1061–1068.

[7] M. Chollet, T. Wörtwein, L.-P. Morency, A. Shapiro, and S. Scherer, "Exploring feedback strategies to improve public speaking: an interactive virtual audience framework," in *UbiCopm '15*. ACM, 2015, pp. 1143–1154.

[8] W. J. Holmes, "Copy synthesis of female speech using the jsru parallel formant synthesiser," in *First European Conference on Speech Communication and Technology*, 1989.

[9] O. Turk and M. Schroder, "Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 965–973, 2010.

[10] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Advances in Neural Information Processing Systems*, 2017, pp. 2966–2974.

[11] M. Schröder, "Expressive speech synthesis: Past, present, and possible futures," in *Affective information processing*. Springer, 2009, pp. 111–126.

[12] K. Ito, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.