INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

Synthesized Spoken Names: Biases Impacting Perception

Lucas Kessler, Cecilia Ovesdotter Alm, Reynold Bailey

Rochester Institute of Technology
lxk6020@rit.edu, coagla@rit.edu, rjb@cs.rit.edu

Abstract

Utilizing a existing neural text-to-speech synthesis architecture to generate person names and comparing them to reference names read aloud in a formal context, we explore how bias resulting from training data impacts the synthesis of person names, focusing on frequency and origin of names. Long-term, we aim to apply voice conversion of person names to aid the effective reading aloud of such names in celebratory ceremonies.
Index Terms: text-to-speech synthesis, person names, bias

1. Introduction

We analyze the perceptual characteristics of and potential biases present in synthesized person names, toward voice conversion for aiding readers in celebratory settings, such as graduation or award ceremonies, to render names aloud. We apply a pre-trained English text-to-speech (TTS) system [1], an open-source version of the *WaveNet* vocoder, which uses an available implementation of Tacotron2 [2] for mel-spectrogram-based generation of speech samples from text. The synthesized set of names vary by *frequency*, *origin*, and *length in number of names*. The model used in the system employed to generate names was trained on the LJ Speech Dataset [3], a single speaker corpus with approximately 24 hours of speech data.
Demo We will interactively play synthesized names to visitors and show results on a laptop or display, discussing findings outlined below on the perception of synthesized person names, and the long-term use case scenario. Some names selected by visitors can also be generated and played back in close to real-time.

2. Corpus Analysis of Spoken Names

We study a corpus of spoken person names comprising five years of public commencement ceremonies in computing disciplines from a US technical university. Thirteen speakers formally read person names and distinctions aloud, resulting in over 2,700 split wav samples of person names, each with a duration of 2-5 seconds, containing a name, if the graduate was given an honor, followed by silence. Occasionally, samples with awards were longer (*highest honors*, *summa cum laude*).

Table 1: *The top 5 person names are male domestic names.*

Top Names	Frequency
1. Michael	147
2. James	105
3. Joseph	96
4. David	74
5. Christopher	64

The corpus' name vocabulary follows a Zipfian distribution, as seen in Figure 1, which displays the linear qualities when frequency was plotted against frequency rank for names in the corpus. The top five most frequent names in the corpus are listed in Table 1. Names can occur as a first, middle, or last name and the frequency includes names where it is the first for one person and the last for another. The prominence of male names and limited number of Hispanic names highlights the well-known issues of uneven gender distribution and under-representation of minority groups in computing disciplines. From the corpus, sets of high, mid, and low frequency names were identified.

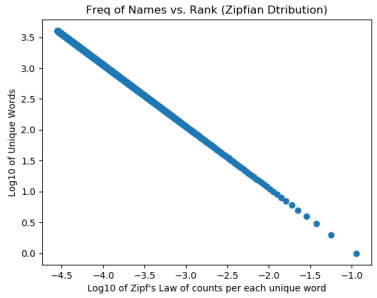

Figure 1: *Zipfian distribution for the person name data.*

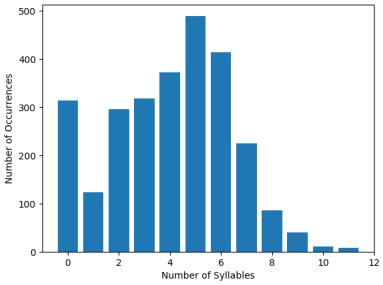

Figure 2: *Distribution of syllable counts in transcribed names* N = 2702*. Names not found were given a zero count. For pronunciation variants, the first alternative was selected.*

Using transcriptions in the CMU dictionary [4], Figure 2 displays the distribution of syllables in the names dataset. Around five syllables per name is most common.

3. Analysis Methods and Discussion

A well-performing TTS system [1] was used to synthesize 24 names corresponding to three factors: *name frequency* (high, mid, low, not in corpus), *name length* (two, three), and typical *name origin/source* (US domestic, international, blended). The synthesized names are in Table 2. None of these names appeared in full in the corpus; rather names were concatenated based on their frequency group.[1] The group of names not in the corpus was selected to analyze how underrepresented

[1]The first, middle, and last components were selected randomly from the distribution, setting aside name order.

Copyright © 2019 ISCA

names were rendered. Some names have a simple syllabic structure (*Hang Wani Patterson*, *Kumar Shah*), whereas others have more complex structure such as (*Lakotko Mahlerhba*, *Alotaibi Gandhi*).

Table 2: *Person names synthesized for perceptual analysis.*

Name	Frequency	# Names	Origin
1. Michael Joseph Martin	High	3	Domestic
2. David Douglas	High	2	Domestic
3. Ramesh Mohammed Singh	High	3	International
4. Kumar Shah	High	2	International
5. Matthew Brian Rodriguez	High	3	Blended
6. James Mehta	High	2	Blended
7. Kelly Sarah Lawrence	Middle	3	Domestic
8. Jared Logan	Middle	2	Domestic
9. Pankaj Mayur Abdulaziz	Middle	3	International
10. Alotaibi Gandhi	Middle	2	International
11. Siddharth Carlos Reed	Middle	3	Blended
12. Parker Aniket	Middle	2	Blended
13. Abigail Sue McCartney	Low	3	Domestic
14. Myles Herman	Low	2	Domestic
15. Alshammari Oluyinka Omari	Low	3	International
16. Lakotko Mahlerhba	Low	2	International
17. Hang Wani Patterson	Low	3	Blended
18. Brianna Poon	Low	2	Blended
19. Joslyn Darby Reno	Not in corpus	3	Domestic
20. Deavon Barth	Not in corpus	2	Domestic
21. Emmanuel Davos Simonetta	Not in corpus	3	International
22. Hoo Huan	Not in corpus	2	International
23. Amelia Malakai Brigham	Not in corpus	3	Blended
24. Borys Wells	Not in corpus	2	Blended

Next, the synthesized names were rated perceptually on four criteria: intelligibility, naturalness, ability to reproduce the name in written form (*What name do you hear?*), and for names from the corpus, similarity to the original spoken name in a side-by-side comparison of the synthesized and the original rendering. Three questions were rated on a 5-point scale, with five being the best rating[2]: (1) *How intelligible is the following clip?*; (2) *How natural does the clip sound?*; and (3) *How similar does the clip sound compared to the original commencement audio?*

Table 3: *Results for 5-point rating: Mean (std.). Sim. = Similarity, Nat. = Naturalness, Int. = Intelligibility, N = # names.*

Category	Sim.	Nat.	Int.	*N*
High fr.	4.2 (1.2)	3.6 (.9)	4.1 (.7)	18
Mid fr.	3.5 (1.6)	3.3 (.8)	3.4 (1.2)	18
Low fr.	3.3 (1.7)	3.5 (1.0)	3.5 (1.2)	18
Not in set	NA	3.4 (1.0)	3.4 (1.0)	18
Domestic	4.8 (.4)	3.8 (.7)	4.3 (.8)	24
Int'l.	2.3 (1.4)	2.9 (.9)	2.8 (1.0)	24
Blended	3.5 (1.5)	3.4 (1.0)	3.4 (1.1)	24
3 names	3.7 (1.4)	3.4 (1.0)	3.4 (1.2)	36
2 names	3.7 (1.6)	3.4 (.9)	3.7 (1.1)	36

Results on the perceptual Likert ratings in Table 3 show that high-frequency names are characterized by improved perception for intelligibility over low-frequency names. In addition, the intelligibility and naturalness ratings for domestic US names are substantially higher than for international names, with blended names hovering between them. Both of these findings point to a bias in name synthesis, affirming both a frequency effect and that monolingual bias in training data has an unfortunate impact on synthesized rendering of person names. It also suggests how challenging reading names aloud can be.

[2]Ratings: 5 = strongly agree, 4 = agree, 3 = neutral, 2 = disagree, 1 = strongly disagree

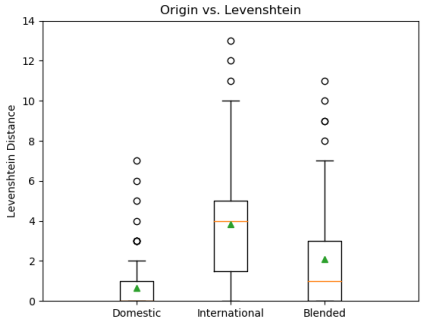

Figure 3: *Origin of names and their Levenshtein distance.*

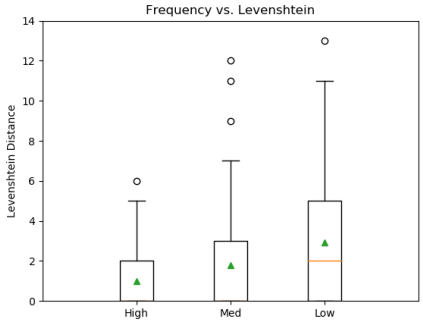

Figure 4: *Frequency of names and their Levenshtein distance.*

In addition, for each name component, the Levenshtein distance was calculated by comparing the original written reference name in the commencement corpus and the written perceived interpretation of the heard name. The orthographic distance between the written forms reflects how intelligible and natural the synthesized name was. If the name was intelligible and natural the distance tended to be lower and, if not, higher. Again, as seen in Figure 3, the bias in synthesized rendering is noticeable for international names, which tend to be characterized by larger Levenshtein distance. Moreover, low-frequency names were characterized by more distance as seen in Figure 4.

4. Conclusion

The show and tell demonstration will discuss the important issue of bias in text-to-speech synthesis on the basis of an applied use case–the rendering of person names in high-stakes celebratory settings. Long-term, we aim to apply name voice conversion to aid improved reading by international, low frequency, and underrepresented person names in these important contexts.

5. References

[1] "WaveNet TTS vocoder." [Online]. Available: https://github.com/r9y9/wavenet_vocoder

[2] J. S. et al., "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," *CoRR*, vol. abs/1712.05884, 2017. [Online]. Available: http://arxiv.org/abs/1712.05884

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

[4] "The CMU pronouncing dictionary." [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict