A BIOLOGICAL INVESTIGATION OF PERFORMATIVE SPEECH THROUGH SYNTHESIS

Emily Lau[1], Brechtje Post[1], Kate Knill[2]

[1]Theoretical and Applied Linguistics and [2]Dept. of Engineering,
University of Cambridge, Cambridge, UK
{ehynl2, bmbp2, kmk1001}@cam.ac.uk

ABSTRACT

Dramatic vocal performance is a layered and nuanced phenomenon. It is difficult to pin down the exact aspects of a performance that make it appealing to the human ear. This work seeks to investigate how the Bio-informational Dimensions (BIDs) impact the judgments of dramatic performance, in particular for the emotion "anger". Test participants listened to pairs of utterances that were resynthesized along the BIDs of size projection and dynamicity to varying degrees to simulate dramatic expressions of anger, and then rated the utterances' differences in dramatic expression. Increased size projection significantly improved listener ratings, while higher dynamicity lowered them. These results indicate that size projection is very important to simulating anger, while dynamicity had a more ambiguous effect that warrants further investigation.

Keywords: speech synthesis, emotional prosody, BIDs, performance, drama.

1. INTRODUCTION

Performative speech realizes itself in varied and dynamic ways. However, even with the subjective and nuanced nature of performance quality, there is no mistaking that there are some fundamental attributes and physical manipulations to dramatic speech that distinguish it from natural speech and make it appealing to the human ear [1, 2]. Researchers in both Linguistics and Artificial Intelligence have begun to probe expressive speech, which includes the specialized nature of dramatic attitudes, which would not only help us understand the nature of human expression, but also add range of expression to today's voice agents.

With regards to expressive prosody, much of the research has focused on directly mapping acoustic parameters and F0 contours to different emotions [3, 4, 5, 6, 7]. Alternately, other work has examined emotional expression in context of the evolution of human communicative functions [8, 9, 10]. Xu et al [11] has expanded on these works and proposed that affective speech is controlled by the so-called Bio-informational Dimensions (BIDs), which manipulate the vocal signal of the speaker to influence the behavior of the listener. The BIDs most investigated so far are:

- Size projection: body size projected by the speaker, associated with median pitch shift, formant shift, and voice quality
- Dynamicity: vigorousness of the speaker's speech stream, associated with pitch range and duration

These have both been found to be influential on the perception of vocal attractiveness, "poshness" and friendliness [12, 13, 14]. Therefore, this theory seems a promising framework in which to study dramatic speech.

This paper aims to observe the effects of size projection and dynamicity on listener judgment of dramatic speech, specifically dramatically angry speech. Because anger is associated with the high end of both dimensions [11], and performative speech is found to be more exaggerated than spontaneous speech [1], it was expected that speech resynthesized to exhibit high size projection and dynamicity would be perceived as most angry.

2. METHODOLOGY

A set of listening experiments was performed to assess how listener judgment of dramatic speech would change according to modulation of speech along the BIDs.

2.1. Stimuli

The stimuli for these experiments were made to target the emotion of "anger," since that is the most identifiable emotion. A 26 year old male speaker of Standard Southern British English (SSBE) recorded the base utterances. The recording was done in a soundproof recording booth in Cambridge

University's Phonetics Laboratory. A sentence taken from the ARCTIC Corpus - "Author of 'The Danger Trail,' Philip Steels, etc." - was used as the utterance content [15]. The speaker produced a set of recordings spoken with a neutral affect but with the varying voice qualities of "modal," "tense," and "more tense."

The utterances were then resynthesized in Praat [16], using a script adapted from stimuli generator code used in Xu et al [12]. Based on the literature survey from Xu et al [11], the emotion of anger is associated with high size projection and high dynamicity. Therefore, the utterances were manipulated upwards along each of these BIDs to different degrees and in varying combinations, in order to create different degrees of "anger." This resulted in three different levels along each dimension, resulting in 3 x 3 = 9 utterances, and therefore 36 possible utterances. Each pair of utterances was tested twice for statistical stability, resulting in 36 x 2 = 72 total judgments. The pairs were presented in reverse order when repeated, to prevent listener bias towards one particular stimulus. All pairs were presented to the listeners in random order. The manipulation parameters for each BID are shown in Tables 1 and 2.

SP Manipulation	Voice Quality	Formant Shift	Pitch Shift
+2	More tense	0.8	−0.5
+1	Tense	0.9	−0.25
0	Modal	1.0	0

Table 1: Size projection Praat resynthesis parameters

Dynamicity Manipulation	Pitch Range	Duration
+2	1.2	0.8
+1	1.1	0.9
0	1.0	1.0

Table 2: Dynamicity Praat resynthesis parameters

2.2. Experimental Setup

The listening experiment was conducted online on the platform Gorilla. There were a total of 30 participants, all speakers of SSBE. There were 10 males, 19 females and 1 non-binary participant, all of whom ranged between the ages of 18 and 50, and were resident in southeast United Kingdom.

Participants were presented with stimuli pairs and asked to compare how "dramatically angry" the utterances were compared to each other by indicating their preference on this rating scale:

1. Clip A sounds much more dramatically angry.
2. Clip A sounds slightly more dramatically angry.
3. Both clips sound equally dramatically angry.
4. Clip B sounds slightly more dramatically angry.
5. Clip B sounds much more dramatically angry.

Therefore, if a listener felt the first utterance they heard was much more dramatically angry, they would select option 1, and would select option 5 if they thought that of the second sound clip.

2.3. Testing Alternate Synthesis Bases

To test how participants would react to vocal manipulations on voices that already had emotion applied, another experiment was carried out using emotional voices as a synthesis base. The same speaker recorded another set of recordings in an "angry" voice. The base utterance was similarly resynthesized in Praat, but instead of up each of the BIDs, it was manipulated one step up and one step down. These were similarly grouped into pairs and presented to participants to compare in the same manner. Voice quality was not included in the parameters included in these size projection manipulations, since it was difficult for the speaker to modulate his voice quality while speaking in an "angry" affectation. The manipulation parameters for each BID for this particular stimuli set are shown in Tables 3 and 4.

SP Manipulation	Formant Shift	Pitch Shift
+1	0.9	−0.5
0	1.0	0
−1	1.1	0.5

Table 3: Size projection resynthesis parameters in Praat for an emotional synthesis base

Dynamicity Manipulation	Pitch Range	Duration
+1	1.2	0.8
0	1.0	1.0
−1	0.8	1.2

Table 4: Dynamicity resynthesis parameters in Praat for an emotional synthesis base

3. RESULTS

A MANOVA was conducted on the results of the tests with both neutral and emotional synthesis bases. An ANOVA was also performed to observe size projection effects for specific pairs of stimuli at specific dynamicity levels, as well as an ANOVA to observe the reverse.

3.1. Neutral Synthesis Base

Figure 1 shows the effect of size projection and dynamicity on listener ratings. The listener trials were divided into three groups of size projection differences: -1 (1 size projection to 0 size projection), +1 (1 size projection to 2 size projection), and +2 (0 size projection to 2 size projection). The MANOVA showed a very significant upward effect of size projection ($F = 147.2$, $p < 2e\text{-}16$) on listener judgments, even between all three difference groups.

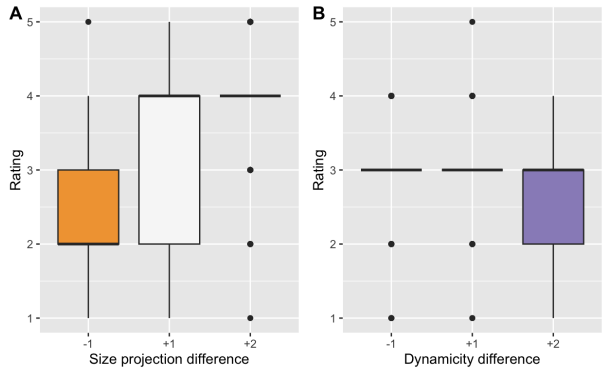

Figure 1: Ratings of dramatic anger across all trials, where A indicates the ratings for each size projection difference, while B indicates the ratings for each dynamicity difference.

The listener trials were similarly divided into three groups of dynamicity differences. This analysis showed a significant effect of dynamicity ($F = 9.715$, $p = 7.17e\text{-}05$). A closer examination using the Bonferroni correction shows no significant effect between the trials with a -1 difference versus a +1 difference ($p = 1.000$), while there was a definitely significant effect between those of a -1 difference versus a +2 difference ($p = 0.00011$). Contrary to the prediction that listener ratings would increase as dynamicity increased, the ratings decreased significantly in this case.

As expected, for the ANOVA of size projection effects at specific dynamicity levels, increased size projection did have a significant effect on listener ratings at all levels of dynamicity, and is quite apparent at the +2 dynamicity level ($F = 60.45$, $p < 2.16e\text{-}16$). As shown in Figure 2, there is a very pronounced effect between the -1 and +1 size projection levels ($p = 8.9e\text{-}14$), as well as between -1 and +2 ($p < 2.00e\text{-}16$), but interesting the effect is minimal between +1 and +2 ($p = 0.1$), which suggests there is a threshold to this effect.

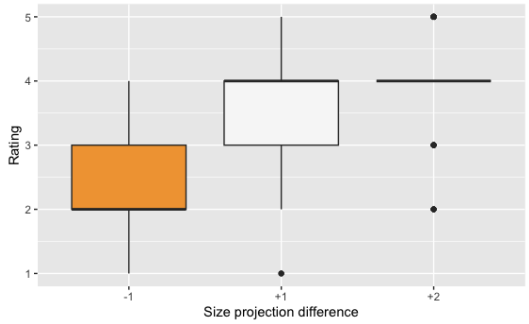

Figure 2: Ratings of dramatic anger for each size projection difference at the +2 dynamicity level.

A similar ANOVA for dynamicity showed this BID has a significant effect, but only at the +2 size projection level ($F = 7.939$, $p = 0.000499$), and is nowhere near as pronounced as the size projection effect at the same dynamicity level. As shown in Figure 3, the effect is most pronounced between the -1 and +2 size projection difference groups ($p = 0.0078$). As with the MANOVA analysis, the direction of the effect of dynamicity with regard to listener ratings was downwards rather than upwards.

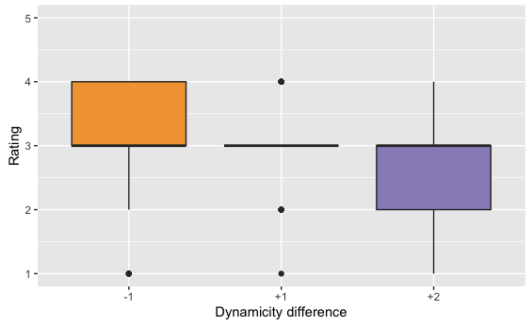

Figure 3: Ratings of dramatic anger for each dynamicity difference at the +2 size projection level.

3.2. Emotional synthesis base

Figure 4 shows the results of the listening test using stimuli resynthesized from an emotional synthesis base. As predicted, there was a substantially significant effect of size projection ($F = 147.2$, $p < 2e\text{-}16$), especially between -1 and +1 ($p < 2e\text{-}16$)

and between -1 and +2 ($p < 2e\text{-}16$) size projection differences.

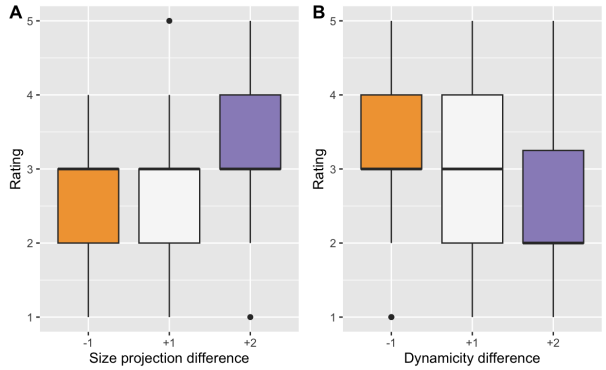

Figure 4: Ratings of dramatic anger across all trials for emotional synthesis base, where A shows the ratings for each size projection difference, while B shows ratings for each dynamicity difference.

The MANOVA for dynamicity also revealed a significant effect ($F = 9.715$, $p = 7.17e^{-5}$), most evident between trials of -1 and +2 dynamicity difference ($p = 1.1e^{-4}$), while quite minimal between trials of -1 and +1 dynamicity difference ($p = 1.000$).

4. DISCUSSION

Based on Xu et al [11] and other literature observing emotional acoustic correlates [17, 18, 19], these experiments utilized the assumption that anger is associated with higher size projection and dynamicity. Based on the study from Jurgens et al [1] observing acoustic differences between natural and acted speech, it was predicted that experimental participants would find utterances with more exaggerated features associated with anger to sound more "dramatic."

The MANOVA and individual ANOVAs for the main (neutral base) listening test indicate that size projection has a significant effect on what listeners consider to be "dramatic anger," which confirms findings in the literature [13]. Listeners consistently found utterances with higher size projection more dramatically angry. In the individual ANOVAs, however, this effect was not always linear, as listener ratings plateaued at higher dynamicity levels. This suggests that at a certain threshold, either the stimuli at higher dynamicity levels sounded similar, or listeners did not respond as well to more exaggerated acoustic features.

Dynamicity also had a significant, but not as pronounced, effect. Contrary to the prediction that listeners would rate utterances with higher dynamicity as more dramatically angry, ratings decreased when dynamicity increased. There are a few considerations for this effect. An inspection of the resynthesized stimuli, found that artificial manipulation of stimuli pitch range and duration in Praat did not consistently manipulate the pitch range, especially when it was meant to be increased. It is possible that while attempting to increase both size projection and dynamicity, lowering the pitch median cancelled out increases to pitch range. This consequence of the resynthesis process would explain why listeners seemed to find stimuli at higher dynamicity levels similar. However, it is noteworthy that when the pitch range increase did work, and the pitch range of the neutral control (508.9 Hz) is compared to that of the +2 size projection/+2 dynamicity stimulus (516.9 Hz), listener ratings still decreased, meaning that dynamicity may still have an effect on listener perception. The results of the MANOVA from the emotional synthesis base test could be interpreted to support this conclusion, as they show a significant effect of both size projection and dynamicity.

These results call into the question the assumptions around anger and the BIDs espoused at the outset of the experiment. While these may not have necessarily been wrongful, it is possible they did not correlate with listener expectations of what "dramatic anger" sounds like. It is also worth considering whether specific parameters always strictly correlate with a specific BID.

5. CONCLUSIONS

This paper attempted to discover if the BIDs of size projection and dynamicity contribute to listener judgment of performative anger when used to artificially manipulate utterances. It was predicted that listeners would find utterances with higher size projection and dynamicity more dramatically angry. While size projection and dynamicity both had significant effects, the effect of the former was larger. Higher size projection improved listener ratings showing it is quite impactful for listeners. Unexpectedly higher dynamicity lowered ratings.

The study also has interesting implications about the impact of dynamicity and how it correlates with the emotion of anger. While the resynthesis process performed this manipulation imperfectly, this calls into question listener expectations of dramatic anger and whether this was reflected in the stimuli. Further research should investigate these correlations and the nuances of different emotional subsets.

6. REFERENCES

[1] R. Jurgens, K. Hammerschmidt, and J. Fischer, "Authentic and play-acted vocal emotion expressions reveal acoustic differences," *Frontiers in Psychology*, vol. 2, pp. 1–11, 2011.

[2] R. Jurgens, A. Grass, M. Drolet, and J. Fischer, "Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected," *Journal of Nonverbal Behavior*, vol. 39, pp. 195–214, 2015.

[3] K. R. Scherer, "Voice quality analysis of american and german speakers," *Journal of Psycholinguistie Research*, vol. 3, 1974.

[4] K. Scherer, "Vocal affect expression: a review and a model for future research." *Psychological Bulletin*, vol. 99, no. 2, p. 143, 1986.

[5] K. R. Scherer, "Vocal correlates of emotional arousal and affective disturbance." *Handbook of Social Psychophysiology*, pp. 165–197, 1989. [Online]. Available: https://psycnet.apa.org/record/1989-97735-007

[6] K. R. Scherer, R. Banse, H. G. Wallbott, and T. Goldbeck, "Vocal cues in emotion encoding and decoding," *Motivation and Emotion*, vol. 15, pp. 123–148, 1991.

[7] S. Mozziconacci, "Prosody and Emotions," in *Proc. First Int'l Conf. Speech Prosody (Speech Prosody '02)*, 2002, pp. 1–9.

[8] E. S. Morton, "On the occurrence and significance of motivation-structural rules in some bird and mammal sounds," *The American Naturalist*, vol. 111, no. 981, pp. 855–869, 1977.

[9] J. J. Ohala, "An ethological perspective on common cross-language utilization of F0 of voice," *Phonetica*, vol. 41, pp. 1–16, 1 1984. [Online]. Available: https://www.degruyter.com/document/doi/10.1159/000261706/html

[10] C. Gussenhoven, "Intonation and interpretation: Phonetics and phonology," in *Proc. Speech Prosody 2002, International Conference*, 2002.

[11] Y. Xu, A. Kelly, and C. Smillie, "Emotional expressions as communicative signals," *Prosody & Iconicity*, pp. 33–60, 2013.

[12] Y. Xu and L. Noble, "Friendly speech and happy speech-are they the same?" in *Proc. 17th International Congress of Phonetic Sciences*, 2011. [Online]. Available: https://www.researchgate.net/publication/266605583

[13] Y. Xu, A. Lee, W.-L. Wu, X. Liu, and P. Birkholz, "Human vocal attractiveness as signaled by body size projection," *PLoS ONE*, vol. 8, p. e62397, 4 2013. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0062397

[14] L. Jiao, C. Wang, C. Hsu, P. Birkholz, and Y. Xu, "Does posh English sound attractive?" in *Proc. Interspeech 2017*, 2017. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1691

[15] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. Fifth ISCA workshop on speech synthesis*, 2004.

[16] P. Boersma, "Praat: doing phonetics by computer," 2006. [Online]. Available: http://www.praat.org/

[17] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[18] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[19] S. Chuenwattanapranithi, Y. Xu, B. Thipakorn, and S. Maneewongvatana, "Encoding emotions in speech with the size code," *Phonetica*, vol. 65, no. 4, pp. 210–230, 2008.