Vowel Formant Profiles and Image Schemata in Auditory Display

Stephen Roddy
CONNECT Centre
Dunlop Oriel House
Trinity College Dublin
Dublin 2
roddyst@tcd.ie

Dermot Furlong
Music & Media Technologies
Dept. of Electronic & Electrical Engineering
Trinity College Dublin
Dublin 1
dfurlong@tcd.ie

This paper presents two evaluations intended to examine if listeners are more likely to associate certain vowel formant profiles with specific data types in an auditory display context. The data types and sounds chosen to reflect those data types are informed by findings from the field of cognitive science. The results of the evaluations suggest that to a limited degree, listeners associate certain vowel formant profiles with strength, largeness of size, darkness and tension. The results further suggest that the amount of noise present in the vocal gesture effects the listeners perception of the tension represented in the sound. These results have implications for the field of auditory display.

1. INTRODUCTION

Vowel sounds are proving increasingly effective for communicating information in the context of auditory display, the use of sound to present information to a listener (Roddy and Bridges 2016). Ramachandran and Hubbard (2001) explore links between vowel and visual information suggesting the link between auditory perception and might be mediated by cognitive structures discussed by Lakoff and Jonson (1999). Furthermore Feist (2013) and Nooteboom (1997) relate vowel sounds to pattern of tension and release while Zbikowski (2005) argue that tension and release patterns are cognised in terms of image schemata. The first experiment tests the likelihood of listeners relating vowel shapes to data-types. Those data-types are reflective of categories of image schemata, commonly shared fundamental gestalt patterns, which are critical components of conceptual metaphor (Lakoff and Johnson 1999). The second experiment aims to determine to what degree listeners associate vowel shape with patterns of tension and release patterns.

2. EXPERIMENTAL EVALUATIONS

2.1 Stimuli

The stimuli in these experiments were generated using formant synthesis techniques. Formant synthesis is a form of subtractive synthesis in which the frequency spectrum of signal is filtered to create formant areas that simulate the characteristics of specific resonating bodies. It is a common approach to the production of vowel sounds in speech synthesis. The stimuli for the two experiments presented here were created using the Reaktor 5 sound design platform because it is a powerful tool capable of creating high quality vowel sounds.

2.2 Evaluation and Recruitment

Participants were recruited through the online crowdsourcing platform Crowd Flower. Each evaluation was designed, hosted and delivered on the Survey Gizmo web-platform. Precautionary measures were taken to ensure that participants were using proper equipment. All participants were required to pass a validation test to prove that they were undertaking the evaluations using a 2-channel stereo setup with either a good set of headphones or a 2-speaker array. Potential participants who did not pass the validation test were not allowed to take part in the evaluations. Participants were recruited from a large international pool of 41 countries to ensure that the results obtained were not specific to a particular culture but could be generalised across a large and varied selection of people. Participants were financially compensated for their participation according to standard crowdflower rates. 139 participants took part in the evaluations. Of that number 26% were female and 74% were male. 20% of listeners had formal musical training and 27% played an instrument. Listeners undertook the each evaluation in a set

© 2018, *Stephen Roddy and Dermot Furlong. Published by BCS Learning and Development Ltd. Proceedings of British HCI 2018. Belfast, UK.*

order and as such it is possible that there were ordering effects. These effects would be unlikely to determine the results, due to the relative nature of the judgements investigated in this evaluation.

2.3 Evaluation 1: Attributes

The first evaluation was intended to determine how strongly listeners relate seven unique vowel formant profiles, A, U, O, I, E, Ü, Ä, to the embodied attribute schemas Big-Small, Dark-Bright, Heavy-Light, Strong-Weak, Rough-Smooth, Hot-Cold as discussed by Johnson (1987).

2.3.1. Design and Materials

Seven stimuli were used in this evaluation. The stimuli were synthesized using formant-filtering techniques in Reaktor 5. Each of the stimuli were 10 seconds long and featured a different vowel profile A, U, O, I, E, Ü, Ä. The stimuli have a clear vocal timbre with a central pitch. Listeners were presented with each of the stimuli and asked to choose which pole of each of the six attribute schemas best describes that sound. The options presented to the listeners were Big or Small, Dark or Bright, Heavy or Light, Strong or Weak, Rough or Smooth, Hot or Cold.

2.3.2. Results and Analysis

The results are presented in Figure 1 and 2 below. Listeners categorised A as the strongest sounding vowel, and U as the weakest but while 81% of listeners categorised A as strong only 42% categorise U as weak.

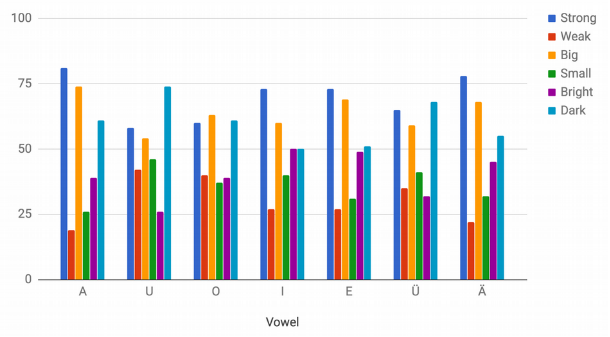

Figure 1: *Significant Vowel Attribute Results*

Listeners also categorised A to be the biggest sounding vowel and U to be the smallest but while 74% of listeners categorised A to be the biggest only 46% categorised U to be the smallest. Listeners categorised I as the brightest sounding vowel, and U as the darkest but while only 50% of listeners categorised I as bright 74% categorise U as dark. The results listed in Figure 2 are clustered more closely around the 50% mark of the scale suggesting that listeners had difficulty relating the vowel sounds to attribute schemas. The average values show that 70% of the time listeners tend to interpret all vowel sounds as strong. The results were analysed by performing a repeated measures logistic regression on each of the attribute ratings (see Kleinbaum and Klein, 2010), with vowel (A ,U, O , I, E, Ü Ä) as the predictor variable, and listener categorisation (strong, weak, big, small, bright, dark, heavy, light, rough, smooth, hot, cold) as the respective dependent variables. This was intended to determine whether the number of listeners selecting negative and positive poles of each attribute differed between vowels.

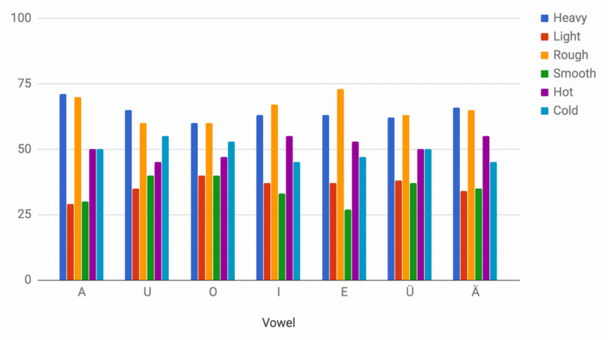

Figure 2: *Non-significant Vowel Attribute Results*

The results for weight Wald F(6, 133)=.931, p>.05, Nagelkerke r2 = .007, roughness Wald F(6, 133)=1.823, p>.05, Nagelkerke r2 = .013, and heat Wald F(6, 133)=1.039, p>.05, Nagelkerke r2 = . 008, were non-significant suggesting that listeners do not associate these attributes with vowel sounds. The results for size Wald F(6, 133)=3.53, p<.05, Nagelkerke r2 = .024 were strongly significant but only accounted for roughly 2% of the variance in listener response, indicating a small effect size. The results for brightness Wald F(6, 133)=3.95, p<.01, Nagelkerke r2 = .039 were significant but only accounted for roughly 4% of the variance in listener response, indicating a small effect size. The results for strength Wald F(6, 133)=5.17, p<.001, Nagelkerke r2 = .044 were strongly significant but only accounted for roughly 4% of the variance in listener response, indicating a small effect size. The results suggest that vowel formant profiles can not be used to control the listeners perception of weak strength, small size, increased brightness, heavy and light weight, rough and smooth texture, or hot and cold temperature in a vocal gesture. As such the results indicated vowel formant profiles cannot be used to control the listeners perception of strength along a scale from weak to strong, size along a scale from small to big, brightness along a scale from dark to bright, weight along a scale from heavy to light, texture along a scale from smooth to rough or temperature along a scale of cold to hot. These results do suggest that, to a limited extent, an A vowel formant profile can be used to lend a sense of strength or a sense of large size to a vocal gesture and a U vowel formant profile can be used to lend a sense of darkness to a vocal gesture.

2.4 Evaluation 2: Tension

The second evaluation is intended to determine how strongly listeners perceive dimensions of tension in vowel formant profiles, A, U, O, I, E, Ü, Ä s suggested by Johnson (1987).

2.4.1. Design and Materials

Fourteen stimuli were used in this evaluation. The stimuli were synthesized using formant filtering techniques in Reaktor 5. Each of the stimuli were 10 seconds long and featured a different vowel profile A, U, O, I, E, Ü, Ä. Two versions of each vowel sound, one with a distinct clear vocal timbre and one with a noisy vocal timbre, were created. The noisy timbres do not consist of pure noise with formant filters applied. They have a noisy sounding timbre but also have a central pitch. As such they can function as vowels. The clear stimuli have a clear vocal timbre with a central pitch. Listeners were presented with each of the 14 stimuli and asked to rate each one on a 5 point scale of Very Relaxed (1), Relaxed (2), Neutral (3), Tense (4) and Very Tense (5).

2.4.2. Results and Analysis

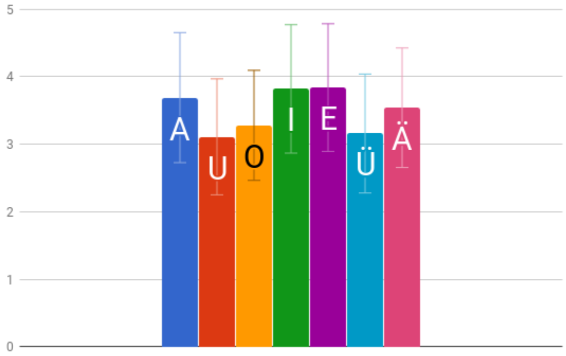

Figure 3: *Mean Likert Ratings and Standard Deviations for Clear Tension Stimuli*

The mean Likert ratings for the evaluation are presented in Figure 3 and 4. All of the results recorded fall within one standard deviation of the midpoint of the Likert scale. This suggests that vowel profile may be of only limited effect in modeling tension. On the basis of decreasing tension listeners rated clear vowels in the sequence E, I, A, Ä, O, Ü, U. The ratings for the clear I and E stimuli are roughly the same being differentiated by only .02 of a Likert category. The ratings for clear Ü and U stimuli are also roughly equivalent being differentiated by only .05 of a Likert category. On the basis of decreasing tension listeners rated noise based vowels in the sequence I, E, Ü, Ä, A, O, U. The noise based Ü and Ä stimuli were rated as roughly equivalent in tension being differentiated by only .02 of a Likert category. The noise based A and U stimuli were rated as roughly equivalent in tension being differentiated by only .05 of a Likert category, while U and O are only differentiated by .01 of a Likert category. The results, illustrated in Figures 3 and 4, were analysed using a repeated measures ANOVA with the design 2 (tone: clear vs. noise) x 7 (vowel: A vs. U vs. O vs. I vs. E vs. Ü vs. Ä) with repeated measures on both factors. There was a main effect of tone F(1, 135)=40.76, *p*<.001, η^2_p = .23, such that clear stimuli were judged more tense than noise stimuli. The highly significant *p* value and large effect size suggest that this is a practically significant result. There was a main effect of vowel F(6, 810)=19.71, *p*<.001, η^2_p = .13, such that the vowels ranked from most to least tense were I, E, A, Ä, Ü, O, U.

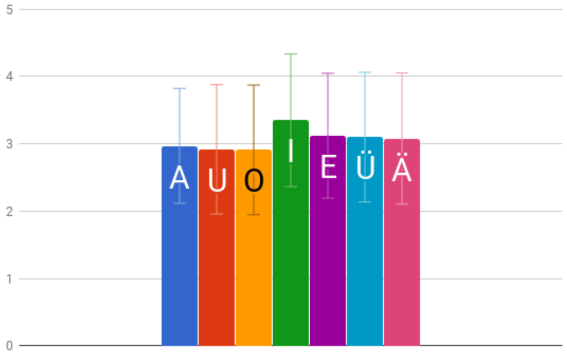

Figure 4: *Mean Likert Ratings and Standard Deviations for Noise Tension Stimuli*

The highly significant *p* value and large effect size suggest that this ranking is reliable. Tone interacted with vowel F(6, 810)=10.96, *p*<.001, η^2_p = .08 and contrasts to decompose the interaction showed that there was a significant effect of vowel on both clear F(6, 130)=19.32, *p*<.001, η^2_p = .47 and noise F(6, 130)=4, *p*<.001, η^2_p = .16. These results confirm that the sequences of descending tension for both clear, E, I, A, Ä, O, Ü, U, and noise, I, E, Ü, Ä, A, O, U are of practical significance to auditory display design, due to highly significant *p* values and large effect sizes. The larger effect size for clear vowel stimuli also suggests that clear vowel sounds are more effective than noisy vowel sounds for creating a sense of tension for a listener. Clear stimuli were perceived by listeners to be reliably more tense than noise clips for the vowels A, Ä, I and E respectively. The result for A was F(1, 135)=52.56, *p*<.001, η^2_p = .28. The result for Ä was F(1, 135)=22.67, *p*<.001, η^2_p = .14, the result for I was

F(1, 135)=24.11, *p*<.001, η^2_p = .15 and the result for E was F(1, 135)=58.35, *p*<.001, η^2_p = .3. Less reliable were the results for U F(1, 135)=4.07, *p*<.05, η^2_p = .03, which showed a small effect size and were just over the threshold of statistical significance. The clear Ü stimulus was not reliably perceived to be more tense than its noise based counterpart F<1. The results indicate that vowels formants profiles have a practically significant effect on perceived tension, particularly for clear stimuli, however it must be noted that all of the stimuli were rated within one standard deviation of the mid-point of the Likert scale meaning that the effects, though reliably present, were not strong. The results suggest that the use of vowel formants to control the listener's perception tension in a vocal gesture is of limited effectiveness.

3. DISCUSSION

The results generated in these evaluations are preliminary in nature and many of the effects recorded translate to small perceptual results in terms of practical auditory display listening scenarios. The results of evaluation one suggest that vowel formant profiles cannot be used to effectively control a listeners perception of weight, strength, texture, heat, size and brightness in vocal gestures. They can be used however to add a limited sense of strength, large size, and darkness to a vocal gesture and to control the perceived sense of tension in a limited manner. This suggests that vowel formant profiles might help to represent these data types when coupled with other parameters like pitch and timbre. For example, when data representing the size of a phenomenon is mapped to a sonic parameter like pitch adding an A vowel formant area that becomes more pronounced as the size represented in the data increases might better represent increases in size to a listener. Further research is required to determine how effective auditory displays of this nature might prove. The results of evaluation two suggest that vowel formant profiles might be useful parameters for consideration in the design of auditory displays of data that represents the level of tension in some phenomenon. The results show that clear sounds are judged to be tenser than noisy sounds and that for clear sounds the sequence of descending tension is E, I, A, Ä, O, Ü, U, while for noise it is I, E, Ü, Ä, A, O, U. This suggests that listeners perceived the U vowel sound to have roughly the same level of tension when clear as it did when noisy. This suggests that both the noisiness of the timbre and the vowel formant profile might be useful parameters for auditory display designers to consider when developing mapping strategies to represent data related to tension. Based on these results designers of future systems might consider making more use of vowel formant profiles and exploring possible strategies for leveraging of the prosodic features of sound in their auditory display solutions. Additionally designers might consider and account for the possibility that the sounds they choose to use in an auditory display context will already come with certain associations for a listener that are not only determined by cultural factors and the traditional cognitive factors explored in HCI but are also influenced by factors described and researched in the field of embodied cognition.

ACKNOWLEDGEMENTS

The research presented in this publication has been supported by a Government of Ireland Postdoctoral Fellowship Award and a Postgraduate Research Studentship award from Trinity College Dublin. This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) and is co-funded under the European Regional Development Fund under Grant Number 13/RC/2077

4. REFERENCES

Feist, J. (2013). "Sound symbolism" in English. *Journal of Pragmatics*, *45*(1), 104-118.

Johnson, M. (1987). *The body in the mind: the bodily basis of meaning, imagination, and reason.* Chicago: University of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought.* New York: Basic books.

Nooteboom, S. (1997). The prosody of speech: melody and rhythm. The handbook of phonetic sciences, 5, 640-673.

Ramachandran, V., & Hubbard, E. (2001). Synaesthesia: a window into perception, thought and language. Journal of Consciousness Studies, 8(1), 3–34.

Roddy, S., & Bridges, B. (2016) Sounding Human with Data: The Role of Embodied Conceptual Metaphors and Aesthetics in Representing and Exploring Data Sets. In *Proceedings of the 1st Music Technology Workshop.* Dublin, Ireland.

Zbikowski, L.M., (2005). *Conceptualizing music: Cognitive Structure, Theory, and Analysis.* New York, US: Oxford University Press, 2005.