Words to Music Synthesis

Michael Krzyżaniak
RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion
Department of Informatics
University of Oslo
michakrz@uio.no

ABSTRACT

This paper discusses the design of a musical synthesizer that takes words as input, and attempts to generate music that somehow underscores those words. This is considered as a tool for sound designers who could, for example, enter dialogue from a film script and generate appropriate background music. The synthesizer uses emotional valence and arousal as a common representation between words and music. It draws on previous studies that relate words and musical features to valence and arousal. The synthesizer was evaluated with a user study. Participants listened to music generated by the synthesizer, and described the music with words. The arousal of the words they entered was highly correlated with the intended arousal of the music. The same was, surprisingly, not true for valence. The synthesizer is at `https://michaelkrzyzaniak.com/Ambisynth/emotion_synth/`.

Author Keywords

Music synthesis, emotion, sentiment, áffect, affect, valence and arousal, words, text, sound design

CCS Concepts

•Applied computing → Sound and music computing; Document management and text processing; •Human-centered computing → Interactive systems and tools;

1. INTRODUCTION

Sound design is a highly creative, time-consuming, and difficult-to-master art form that is ubiquitous in film, television, podcasting, radio; in grocery stores, train stations, architecture. There has been considerable effort over the past few decades towards using computers to aide or automate similar art forms such as composition of music, poetry, and short stories. However, comparatively little attention has been given to the automation of sound design. The field is wide-open not only from a research perspective, but also a economic one, as film-makers often cite sound-design as the most time-consuming and expensive part of their budget [15][5].

The work on automatic sound design in this paper builds on a previous study by other authors, which focuses on the automatic production of radio plays [3]. In that work, they start with a pre-existing script in the form of text, and they use it to automatically generate fully-produced audio. They first perform a semantic analysis of the text that identifies key features of the characters and the scenes. Then, using those features, they create an audio version of the text using appropriately selected speech synthesizers and background sounds. That study does not cover the inclusion of music to heighten the reading of the text. Consequently, this present study aims to extend the previous study to include background music. The goal here is to write a musical synthesizer whose input is a few lines of text, and whose output is background music that somehow underscores the meaning of the text as it is spoken. More generally, such a words-to-music synthesizer could have applications broadly across many situations where sound-design is used in conjunction with language.

NIME'20, July 21-25, 2020, Royal Birmingham Conservatoire, Birmingham City University, Birmingham, United Kingdom.

2. PREVIOUS WORK

Previous work has investigated the use of automatic sound design to create background music for video games [2][12], image-slideshows [10][4], and other visual media. These estimate the sentiment or impression of the input media, and then either synthesize music, retrieve music from a database, or produce variations on existing music that matches the estimated sentiment. One system [8] generates background music for *narrative* image sequences, and mentions, without elaboration, that it could be extended to text input. Regarding text input, a U.S. Patent [13] describes a system that literally translates words into music using a steganographic-like system, in which individual words are encoded and later decoded using a dictionary of pre-defined musical fragments. This system does not attempt to underscore the meaning of the text. Kanno et al. [9] present at system that generates music to summarize the mood of input text. Their architecture is quite similar to what is presented here. However, they assign mood values to a lexicon using purely computational techniques, and their final evaluation of the complete system is somewhat preliminary. My approach uses an externally-validated sentiment analysis for both words and musical features, and I perform a complete, blind, end-to-end evaluation of the completed system.

3. IMPLEMENTATION

'Valence and arousal' (v&a) provides a straightforward way of mapping words to music, and is the approach taken in this paper. V&a is a two-dimensional model of emotion that is frequently used to label both text and music, and can therefore serve as a common intermediate representation. Valence is sometimes called 'pleasantness', and represents how pleasurable or displeasurable a given emotion is. Arousal refers to how energetic or calm and emotion

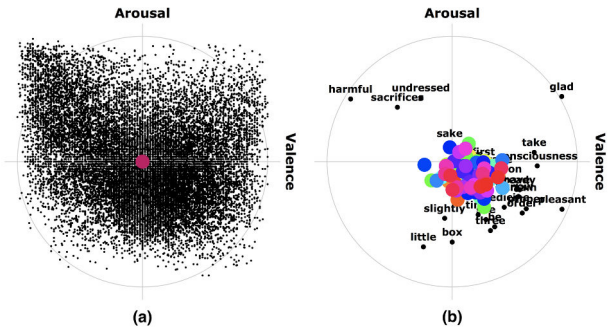

Figure 1: (a) All 20 000 words in the NRC VAD. The mean v&a are 0.000, represented by the pink dot. The distribution is, interestingly, skewed to the top left, with median valence 0.020 ('ambition'), and median arousal -0.040 ('ballet'). Random samples of words from this lexicon will have 0 mean. (b) Words taken from language are not randomly selected, and do not have 0 mean. The words plotted are from a randomly chosen passage of 'War and Peace' containing 30 valid words. One of the colored dots is the mean v&a of the plotted words. The other colored dots are the means of other similar passages of text chosen from a corpus of 5 books. According to the central limit theorem, the more valid words per sample, the more tightly clustered the colored dots will be. When scoring a passage of text I un-cluster the score by normalization such that the colored dots would have 0 mean and standard deviation of 1/3 so they cover the entire v&a space.

is. These dimensions taken together serve as a rough but adequate way of representing a wide variety of emotions. Because v&a is only a very approximate, emotions are often grouped by quadrant, defined by whether valence and arousal are positive or negative. To be clear, the goal of this study is not to induce felt emotions in listeners. Instead, v&a is only used here as a common labeling scheme; the goal is to extract v&a scores from a passage of text, and then to synthesize music that appropriately acculturated listeners would label with similar v&a scores.

3.1 Text To Valence and Arousal

As pre-processing for the desired music synthesizer, a method must be devised to extract v&a scores from a passage of text. There have been several studies that ask users to score the v&a of English words. The foundation of the present study is provided by the NRC VAD Lexicon [11], which contains ratings for twenty-thousand words compiled into handy, machine-readable text files [1]. Henceforth, words that have entires in the NRC VAD shall be called 'valid' words. The v&a of all valid words are plotted in Figure 1a.

This gives us v&a scores for individual words, but how should a whole passage of text be scored? Initially, it might seem reasonable to use the mean v&a of all of the individual words. Let such a mean be called an 'utterance'. The problem is that utterances tend to get clustered in the 'pleasant' quadrant, and utterances with more words are more tightly clustered. This is illustrated in Figure 1b. This is a consequence of the central limit theorem, because larger samples have smaller standard deviations, together with the fact that language is replete with neutral filler words that lie in this quadrant. It is therefore desirable to correct for this by normalizing the utterances based on the number of words, so that utterances with many words get spread back out to cover the entire v&a space. The lexicon as a whole has 0 mean v&a (and is slightly skewed towards the 'angry' quadrant), so a random sample of words should, on average, have 0 mean. However, a sample of real utterances taken from language do not have 0 mean, because in language some words are used more frequently than others. To find the true population statistics, I obtained a text file containing 5 complete books totaling over a million words[1]. I extracted 100 000 passages of text, containing exactly 30 valid words each, from random locations in the file. For each passage I calculated the utterance score, and averaging those gave me the *sample* mean and standard deviation for both v&a, for samples of size 30. Recalling that the *population* mean is equal to the sample mean, and the population standard deviation is equal to the sample standard deviation times the square root of the number of valid words in the utterance, I worked out that the population statistics are are $\mu_v = +0.152$, $\sigma_v = 0.707$, $\mu_a = -0.108$, and $\sigma_a = 0.498$ (the subscripts v and a indicate valence and arousal). Using these values, any utterance can be normalized by subtracting out the mean, and dividing by the *sample* standard deviation of that particular sample, which depends on the number of words in the sample. The equations are

$$v_{normal} = \sqrt{n} * (u_v - \mu_v)/(3 * \sigma_v) \quad (1)$$

$$a_{normal} = \sqrt{n} * (u_a - \mu_a)/(3 * \sigma_a) \quad (2)$$

where u is the utterance (mean v or a of the individual words), n is the number of valid words in the utterance, the factor of 3 keeps 99.7% of texts within the unit circle in v&a space, and v_{normal}, a_{normal} constitute the normalized utterance, which are taken to represent the v&a of the entered text, as a whole. This process is applied to all utterances with $n > 1$.

To facilitate the analysis of a passage of text, I built a user interface, depicted in the top half of Figure 2, that allows users to enter words. The interface plots each individual valid word in v&a space, and plots the normalized utterance score.

Note that this approach is not sensitive to grammar. How do you score a statement like 'I am not angry'? This approach supposes that people choose words for a reason; if a person were happy or neutral, they would have said so. When someone says 'I am not angry', they are asking you to imagine the entire universe that is their anger, telling you that they have cause to be angry, and are positioning themselves relative to that. The statement can only be interpreted within the framework of anger, and consequently, at least according to the method of analysis adopted here, the overall sentiment is still anger.

3.2 Valence and Arousal to Musical Features

Given the v&a scores for a passage of text, the goal is then to synthesize music that matches them. But what features should the music have? Other studies have implemented v&a-based synthesizers with bespoke feature sets [14][6]. For the purposes of this study, however, the answer is provided by the work described in [7]. In that paper, the researchers collected several short clips of classical music. They had two separate groups of people rate the clips. One group of listeners rated the v&a of each clip. The second group comprised musical experts, who scored the musical features for each clip. The features included,

[1]I used the file found at http://www.norvig.com/big.txt, and removed the word lists and code from the very bottom, after the last novel ends, but retained the Gutenberg legalese.

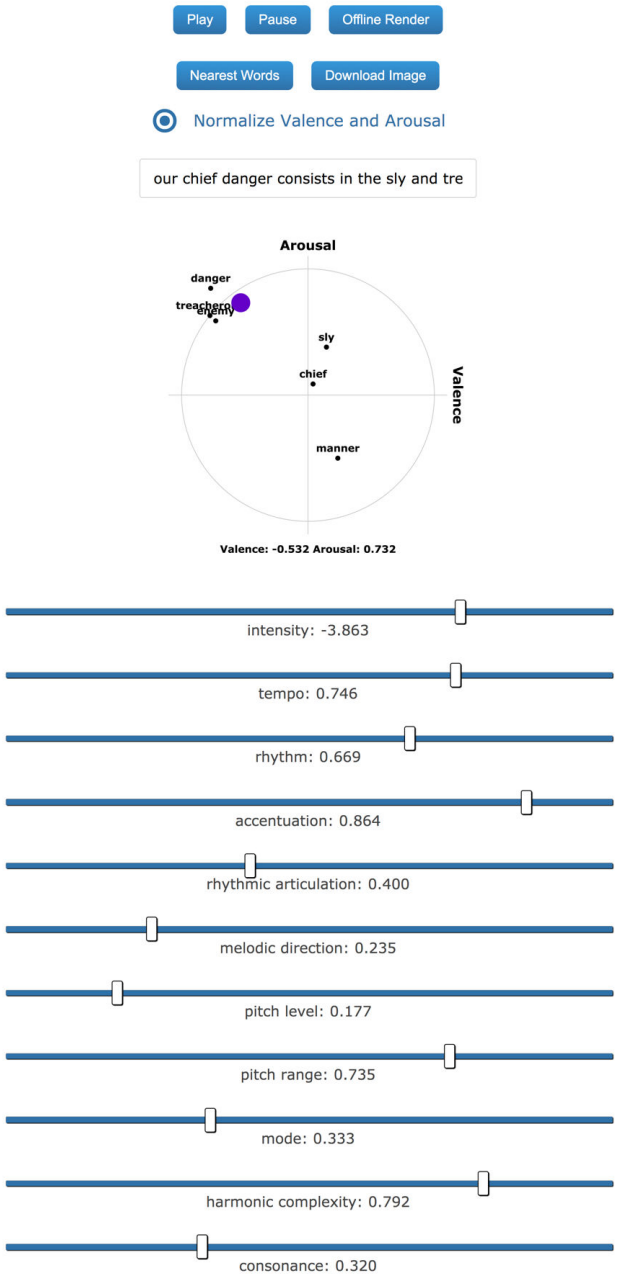

Figure 2: The user interface for the text-to-music synthesizer. The top half shows the analysis of the sentence "our chief danger consists in the sly and treacherous manner in which the enemy approaches us", taken from an Aesop fable. Each valid word is plotted individually, and the purple dot shows the normalized utterance score, which represents the passage as a whole. The normalization process helps correct for neutral words like 'chief' and 'manner' (possibly mis-interpreted here), which tend to drag the utterance towards the mean. In the NRC VAD, the closest words to the purple dot are 'persecution', 'bombardier, and 'stampede', which reasonably agree with the sentiment of the original sentence. The bottom half of the figure shows the settings of all of the musical features that correspond to this text. Entering text automatically sets the features. Alternatively, users can drag the purple dot around, and the musical features will automatically update to the correct values, or users can drag the sliders to set individual features to arbitrary values.

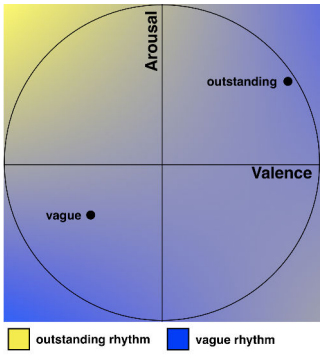

Figure 3: An example of the mapping between v&a and musical features. The background gradient shows how the 'rhythm' feature maps to v&a, with yellow representing outstanding rhythms and blue representing vague ones. Overlaid for comparison are the v&a scores for the words 'outstanding' and 'vague', according to the NRC VAD. Note that there is no reason to expect that the word 'vague' would embody the same v&a as a vague rhythm.

for example, melodic direction, and mode, and the experts rated them subjectively on a Likert scale, even for features such as tempo which could have been measured objectively. (The complete list of features is given in Section 3.3 below.) The researchers also measured the sound intensity in dBA for each clip. By combining the results from the two different groups, they were able to map v&a to individual musical features. For clarity, the the map between v&a and the 'rhythm' feature is shown in Figure 3. The bottom half of Figure 2 shows the interface I built where selecting particular v&a values causes all of the musical feature sliders to be set to the correct values.

3.3 Musical Features to Music

The upshot of all the work described above is that users can enter text into a web interface and find out, for instance, how fast, consonant, and accentuated the corresponding music should be. However, knowing that the harmonic complexity should be 6 on a 9-point Likert scale, et cetera, clearly does not fully specify what the music should sound like, and leaves a lot open to interpretation. Filling in the gaps is a compositional task, and what follows is a description of the compositional decisions that I, the author, made. I have a background in music composition, and am able to fill in the missing parts by drawing on shared fuzzy cultural knowledge about how music relates to emotional labels.

The overview is that the music that I wrote consists of several cello-like synthesizers (henceforth referred to a 'cellos') that each plays a monophonic sequence of notes using a drunkard's walk, moving randomly up and down a scale, either stepwise or by small random leaps. Each note has an ADSR envelope applied to its amplitude. In addition, the music contains a low-pitched drone that sustains alternately the first and fifth scale degree of the current mode. More details about this are provided in the following subsections for the individual musical features, or in the source code itself, available on my website, linked in the Abstract.

3.3.1 Sound Intensity (dBA)

In the reference paper [7], they measured the A-weighted sound intensity for each clip. That is not reproducible in a synthesizer because it is not easy to control the spectral dis-

tribution of energy when synthesizing music with so many other constraints, and in any event, listeners could turn the volume up and down, thereby changing the sound intensity in dBA. Nonetheless, in that study, the entire v&a space has a range of 15 dBA. Therefore, here, I approximated that by implementing a simple gain control that ranges from -15 to 0 dB (un-weighted), and maps correctly to v&a space. This makes the music relatively louder or softer for different v&a scores, in approximately the prescribed way.

3.3.2 Tempo (Slow to Fast)

All cellos are driven by a single underlying metronome. At each tick of the metronome, each cello selects a new note (or rests, or sustains, or re-articulates the current note) at regular intervals. The tempo adjusts the frequency of the metronome. Additionally, the parameters of the ADSR are adjusted to be broader at slower tempi, to mimic the natural effects of slower bowing on an acoustic cello.

3.3.3 Rhythm (Outstanding to Vague)

Making the rhythm more vague makes it, at each metronome tick, less likely that a new pitch will be chosen, meaning that the current note will be sustained instead. Additionally, when a new note *is* chosen, it is less likely to be accented. Overall, this has the effect that vague rhythms have sparse note onsets occurring at random times, and without strong accent, while outstanding rhythms have many notes coming at equal intervals with many accents applied to randomly selected notes.

3.3.4 Accentuation (Light to Marcato)

Raising the accentuation level raises the peak of the ADSR envelope for accented notes. At the lightest level of accentuation, accented notes will sound identical to unaccented ones, i.e. there will be no accents. At the most marcato level of accentuation, accented notes will have a much louder onset than unaccented notes.

3.3.5 Rhythmic Articulation (Staccato to Legato)

Staccato notes have shorter ASDR attack and decay durations, and a lower sustain level, so that in the extreme case there is a very sharp attack and decay with no sustain. This interacts with the tempo parameter, and these durations are set as a percentage of the note length.

3.3.6 Melodic Direction (Ascending to Descending)

Notes are chosen via a Brownian walk; at each time step, a new note will be randomly chosen within a certain range above or below the current note. For ascending melodies, it is more likely that a note above the current note will be chosen, but it will be within only a narrow interval above. However, if on the off chance a note below the current note is selected, it will be selected from a larger range. This has the effect that the melody will be characterized by several consecutive notes ascending by step, interspersed by leaps down.

The reference paper [7] found melodic direction to be entirely determined by emotional valence, with *positive* valence associated with *descending* melodies. This association is highly suspect, as it contradicts common sense. Throughout Western music history, descending melodies have very frequently been associated with *negative* valence emotions. The Baroque Lament, for example, is defined both by a sad character and a descending chromatic bassline. Lyrics describing an ascent into heaven or descent into hell are frequently painted with melodic fragments in the corresponding directions. At the apotheosis of Bach's mass in b minor, the Crucifixus, depicting the moment when Christ is murdered and buried, opens with a descending melody, repeated by the various voice parts in descending order of pitch. Immediately after that is the Resurrexit, the moment where Christ rises from the dead, which opens with an ascending melodic run repeated by the various voice parts in ascending order of pitch. The words 'ascending' and 'descending' have positive and negative valence, respectively, according to the NRC VAD.

There are at least two possible reasons that the reference study had inverted results. The first is that because they had a small number of experts rating a small number of clips, the observation could easily be a statistical anomaly that would disappear with a larger sample size. Second is that they evidently put 'ascending' on the left of the Likert scale, and 'descending' on the right, which seems somewhat counterintuitive given the layout of the piano and the level of musical training of the participants, so they may have been confused by this. Either way, I took the liberty of re-inverting the findings of the paper and associated ascending melodies with positive valence.

3.3.7 Pitch Level (Low to High)

The Brownian walk is bounded; if a cello reaches the top or bottom of the note range, it will then leap to a random note somewhere else within the range. Adjusting the pitch level raises or lowers the top and bottom notes of the range.

3.3.8 Pitch Range (Narrow to Wide)

This adjusts the top note of the note range, while leaving the bottom of the range untouched.

3.3.9 Mode (Minor to Major)

In the reference paper [7], they consider only major and minor modes. Here, all seven neo-Renaissance church modes are used, arranged from the most sharp to the most flat; Lydian, Ionian, Mixolydian, Dorian, Aeolian, Phrygian, Locrian; with Lydian being the most 'major' and Locrian the most 'minor'. The low-pitched drone that underlies the cellos gives definition to the different modes.

3.3.10 Harmonic Complexity (Simple to Complex)

This is amongst the least clearly defined of the musical features; it is not clear what harmonic complexity is. Because the cellos proceed melodically and independently of one-another, it is not obvious how to construct chords, let alone sequences of chords, using this model. So this feature controls how many cellos are playing. The most simple music will have just a single cello and a drone, resulting in monophonic music. The most complex music will have several cellos plus the drone, and naturally most of the chords will be rich with diatonic clusters and many complex chords with 7ths and 9ths.

3.3.11 Consonance (Dissonant to Consonant)

Consonance adjusts the probabilities of the various scale degrees of being selected during the Brownian walk. For the most dissonant melodies, the walk proceeds in the usual fashion with equal probability of selecting the various notes within the given range, thus the music will contain all notes in equal proportion. For more consonant melodies, the Brownian walk will tend to skip over more dissonant scale degrees like the second and seventh, and will tend to land on more consonant scale degrees like the root and fifth. The most consonant melodies will contain only the root, and the Brownian walk will jump between octaves. This interacts with the 'melodic direction' feature, as both impose fuzzy constraints on the note selection process.

4. EVALUATION

A user study was conducted to evaluate the synthesizer's ability to generate music that matches the v&a of its text input. The overview is that, in the study, participants listened to musical clips generated by the synthesizer with a variety of v&a values, and they were asked to describe the music with words. The hypothesis was that, if the underlying models are correct, the v&a scores for words entered by participants should match the v&a of the music they were listening to at the time.

4.1 Experimental Design

In advance, I prepared eight 45-second audio clips that roughly evenly cover v&a space. The clips had all combinations of ± 0.5 valence and arousal, and all combinations of ± 0.85 valence and arousal. I also prepared a 2-minute clip in which the v&a slowly progressed around the unit circle in v&a space, so that the music continually changed, and progressed through all four quadrants. The participants were agnostic to the fact that the clips relate to v&a. The participants first listened to the 2-minute clip, which served to calibrate the their expectations about the range of different musics that would be heard in the remainder of the experiment. Then the participants were presented one of the prepared 45-second clips, along with a text-input field on a computer user-interface, and asked to respond to the following

> Listen to the audio and describe it with at least four or five words in English. Your description can be literal or metaphorical. What does the sound make you think of? How does it make you feel? What is the character of the music?

This was repeated 16 times total, such that each participant rated each clip twice. Because a person's impression of a piece of music is likely to be coloured by what they heard immediately before it, each participant rated the clips once in a random order, and then again a second time in a different random order, and the orders were different for each participant. Participants also completed a brief entrance and exit questionnaire, and a consent form. The entire listening test was conceived as a series of web forms and put on the internet, so that participants could complete it at the time and location of their choosing.

4.2 Participants

We recruited 22 people to participate in the study, primarily from academic mailing lists. They ranged from 25 to 57 years of age ($\mu = 35$). Fourteen of them responded to all 16 clips that were presented to them, and the remainder listened to eight or fewer clips before quitting prematurely. The participants submitted a cumulative total of 997 non-unique, 'valid' words that have entries in the NRC VAD ($\mu = 125, \sigma = 16.7$ words per audio file, or about 3.9 valid words per response). Most of the responses were very descriptive and evocative, as desired, for example "retro, cyberpunk, classical, walking in the city at night while it rains". This is exactly the type of description that might be found, for example, in a radio play or film, for which a sound-designer might want to synthesize background music. A couple of participants also referred to previous clips in their responses, such as "reminiscent of #5. different, but too subtly to differentiate." The only valid words here are 'reminiscent' and 'different', and the context, about whatever clip number five was, is lost.

4.3 Analysis

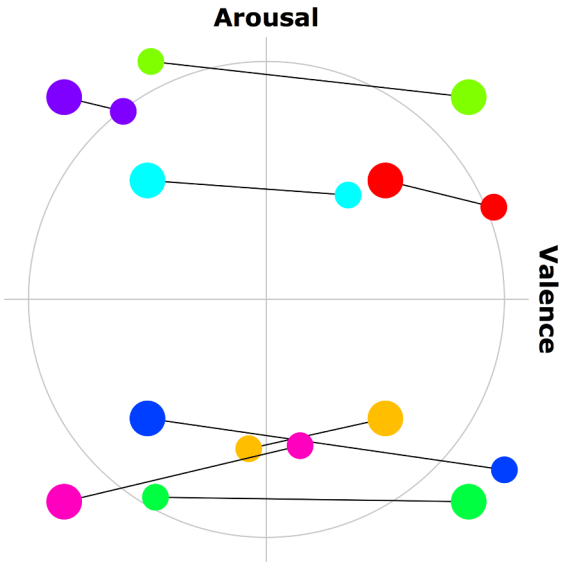

Figure 4: Results of the user study. In each pair of dots connected by a line, the larger dot represents the intended v&a of a clip of music, and the smaller dot represents the normalized utterance score for all of the words entered cumulatively by all of the participants for that clip. The words correlate strongly with the arousal of the music, but not the valence.

I did a very small amount of pre-processing on the user-submitted words, such as changing the word 'sci-fi', which is not valid, to the word 'scifi', which is valid. Additionally, I removed a small number of responses where users had used the input field to report that the audio clip was not loading. Then, for each audio file, I concatenated all of the words entered by all of the participants for that file, and calculated the normalized utterance score (described in Section 3.1) for all of the words taken together. Figure 4 plots the intended v&a of each clip of music, and the normalized utterance score corresponding to the words entered for the respective clip. As can be seen, the arousal of the music is strongly linearly correlated to the arousal of the user submitted words with $m = 1.04, b = -0.02, r^2 = 0.96$. This is very significant, with $p = 0.00002$. By contrast, the valence of the music and words are uncorrelated. If anything, the valence appears to be negatively correlated, with some exceptions, although not significantly so, even when removing the exceptions. While there is a reasonable degree of variability between participants, these results are robust within the participants, there are no extreme outliers, and excluding any small number of participants does not change the overall results.

4.4 Interpretation

These results are surprising in two key ways. First, it is surprising that arousal is as strongly correlated as it is. This may in part be because tempo is such an obvious feature and strong predictor of arousal, that listeners are more likely to enter words like 'calm' for slow tempi and 'energetic' for faster tempi. Additionally, literal descriptions of the tempo, like 'slow' and 'rapid' ('fast' is not in the NRC VAD) have corresponding arousals. It would be interesting to repeat the experiment while holding the other features constant to see if arousal remains as strongly correlated. Secondly, it is

interesting how poorly valence is correlated. According to [14], "mode is a harmonic structure that has a strong relationship with musical valence", and according to [7], "mode, harmonic complexity, and rhythmic articulation best differentiated between negative and positive valence". It would have been reasonable to predict that mode alone would have given rise to a reasonable degree of correlation. It could be that, in the synthesizer, the mode is too obscured by other features. For example, when consonance is very high, scale degrees that distinguish one mode from another might sound infrequently or not at all, thereby obfuscating the mode. Melodic direction, which is also associated with valence, is also somewhat obscured by consonance, because when dissonant notes are excluded the melody necessarily proceeds by leaps in both directions. Moreover, in the reference study [7], most of the v&a space is covered with the mode parameter approximately equal to 5 on a 9 point Likert scale, i.e. slightly major, with only the top left corner being extremely minor. One might naturally question whether there should be a greater range of mode values, particularly in the lower half of the v&a space. Finally, v&a are often taken together with a third dimension, called 'dominance', and the flattening of this dimension here may have caused some confusion in the results.

5. LIMITATIONS AND FUTURE WORK

One obvious limitation of this work this is that the music is likely to be received as a single musical composition, with a single instrumentation and overall style. In the future, to be a more useful tool for sound designers, the synthesizer could be made to include more instruments and other styles. At the very least there could be a few synthesis algorithms for users to select between. Another limitation is that the synthesizer is static, in the sense that if a user enters a passage of text that changes emotion halfway through, the synthesizer will average the emotions to produce a single v&a score, rather than moving over time from one to the other. In the future it might be interesting to use a sliding window over the text to produce dynamic music that changes v&a over time. Related to this is that the synthesizer does not know when to produce silence, which would be important in a fully-automated setting. One idea is that the synthesizer should only produce music only when the v&a score is sufficiently extreme. Either way, as it stands, the synthesizer is not meant to be fully automated, and the expectation is that a sound-designer would identify where music is needed, enter text, and then tweak the parameters to their liking.

6. ACKNOWLEDGEMENTS

This work was partially supported by the UK Engineering and Physical Sciences Research Council IAA grant. This work was partially supported by the Research Council of Norway through its Centres of Excellence scheme, project number 262762. Thanks to Philip B Jackson at CVSSP, University of Surrey, and Kyrre Glette at RITMO University of Oslo, for their feedback. Thanks to Will Buchanan and Russ Bradbury at RPPtv Ltd for their support and feedback. This project was partially supported by, and carried out in collaboration with RPPtv Ltd.

7. REFERENCES

[1] Canada National Research Council. The sentiment and emotion lexicons download. `http://sentiment.nrc.ca/lexicons-for-research/`, 2016. Accessed: 2020-01-08.

[2] P. Casella and A. Paiva. Magenta: An architecture for real time automatic composition of background music. In *International Workshop on Intelligent Virtual Agents*, pages 224–232. Springer, 2001.

[3] E. T. Chourdakis and J. D. Reiss. From my pen to your ears: automatic production of radio plays from unstructured story text. In *15th Sound and Music Computing Conference*, pages 142–149, 2018.

[4] P. Dunker, P. Popp, and R. Cook. Content-aware auto-soundtracks for personal photo music slideshows. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–5. IEEE, 2011.

[5] S. Follows. Full costs and income of a $1m independent feature film. `https://stephenfollows.com/full-costs-and-income-of-1m-independent-feature-film/`, 2015. The user Robert Niessner comments "The biggest single cost was the music ... and audio post production." Accessed: 2020-01-08.

[6] A. Godbout, I. A. Popa, and J. E. Boyd. Emotional musification. In *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, page 6. ACM, 2018.

[7] P. Gomez and B. Danuser. Relationships between musical structure and psychophysiological measures of emotion. *Emotion*, 7(2):377, 2007.

[8] K. Ishizuka and T. Onisawa. Generation of variations on theme music based on impressions of story scenes considering human's feeling of music and stories. *International Journal of Computer Games Technology*, 2008, 2008.

[9] S. Kanno, T. Itoh, and H. Takamura. Music synthesis based on impression and emotion of input narratives. In *Sound and Music Computing Conference (SMC2015)*, pages 55–60, 2015.

[10] C.-T. Li and M.-K. Shan. Emotion-based impressionism slideshow with automatic music accompaniment. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 839–842, 2007.

[11] S. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, 2018.

[12] A. Naushad. Condition driven adaptive music generation for computer games. *arXiv preprint arXiv:1306.1746*, 2013.

[13] N. Ruetz and D. Warhol. Systems, methods and automated technologies for translating words into music and creating music pieces, May 20 2014. US Patent 8,731,943.

[14] I. Wallis, T. Ingalls, E. Campana, and J. Goodman. A rule-based generative music system controlled by desired valence and arousal. In *Proceedings of 8th international sound and music computing conference (SMC)*, pages 156–157, 2011.

[15] WIRED. The slow mo guys answer slow motion questions from twitter. `https://youtu.be/vS5z3tR6xTI?t=150`, 2018. Gavin Free says "I would say that sound design is by far the most time consuming part of making these videos". Accessed: 2020-01-08.