Aalborg Universitet

Pitch Gestures in Generative Modeling of Music

Jensen, Kristoffer

Pitch Gestures in Generative Modeling of Music

Kristoffer Jensen[1]

[1] Aalborg University Esbjerg, Niels Bohr Vej 8,
6700 Esbjerg, Denmark
krist@create.aau.dk

Abstract. Generative models of music are in need of performance and gesture additions, i.e. inclusions of subtle temporal and dynamic alterations, and gestures so as to render the music musical. While much of the research regarding music generation is based on music theory, the work presented here is based on the temporal perception, which is divided into three parts, the immediate (subchunk), the short-term memory (chunk), and the superchunk. By review of the relevant temporal perception literature, the necessary performance elements to add in the metrical generative model, related to the chunk memory, are obtained. In particular, the pitch gestures are modeled as rising, falling, or as arches with positive or negative peaks.

Keywords: gesture; human cognition; perception; chunking; music generation.

1 Introduction

Music generation has more and more uses in today's media. Be it in computer games, interactive music performances, or in interactive films, the emotional effect of the music is primordial in the appreciation of the media. While traditionally, the music has been generated in pre-recorded loops that is mixed on-the-fly, or recorded in traditional orchestras, the better understanding and models of generative music is believed to push the interactive generative music into the multimedia. Papadopoulos and Wiggins (1999) gave an early overview of the methods of algorithmic composition, deploring "that the music that they produce is meaningless: the computers do not have feelings, moods or intentions". While vast progress has been made in the decade since this statement, there is still room for improvement.

The cognitive understanding of musical time perception is the basis of the work presented here. According to Kühl (2007), this memory can be separated into three time-scales, the short, microtemporal, related to microstructure, the mesotemporal, related to gesture, and the macrotemporal, related to form. These time-scales are named (Kühl and Jensen 2008) subchunk, chunk and superchunk, and subchunks extend from 30 ms to 300 ms; the conscious mesolevel of chunks from 300 ms to 3 sec; and the reflective macrolevel of superchunks from 3 sec to roughly 30–40 sec. The subchunk is related to individual notes, the chunk to meter and gesture, and the superchunk is related to form. The superchunk was analyzed and used for in a generative model in Kühl and Jensen (2008), and the chunks were analyzed in Jensen and Kühl (2009). Further analysis of the implications of how temporal perception is

related to durations and timing of existing music, and anatomic and perceptual finding from the literature is given in section 2 along with an overview of the previous work in rhythm. Section 3 presents the proposed model on the inclusion of pitch gestures in music generation using statistical methods, and the section 4 discusses the integration of the pitch gesture in the generative music model. Finally, section 5 offers a conclusion.

2   Cognitive and Perceptual aspects of rhythm

According to Snyder (2000), a beat is single point in time, while the pulse is recurring beats. Accent gives salience to beat, and meter is the organization of beats into a cyclical structure. This may or may not be different to the rhythmic grouping, which is generally seen as a phrase bounded by accented notes. Lerdahl and Jackendorff (1983) gives many examples of grouping and meter, and show how this is two independent elements; Grouping – segmentation on different levels is concerned with elements that has duration, and Meter – regular alternation of strong and weak beats is concerned with durationless elements. While grouping and meter are independent, the percept is more stable when they are congruent.
The accentuation of some of the beats gives perceptually salience to the beat (Patel and Peretz 1997). This accenting can be done (Handel 1989) by for instance an intensity rise, by increasing the duration or the interval between the beats, or by increasing the frequency difference between the notes.
   Samson *et al* (2000) shows that the left temporal lobe processes rapid auditory sequences, while there are also activities in front lobe. The specialized skills related to rhythm are developed in the early years, for instance Malbrán (2000) show how 8-year-old children can perform precise tapping. However, while the tapping is more precise for high tempo, drifting is ubiquitous. Gordon (1987) has determined that the perceptual attack time (PAT) is most often located at the point of the largest rise of the amplitude of the sound. However, in the experiment, the subjects had problems synchronizing many of the sounds, and Gordon concludes that the PAT is more vague for non-percussive sounds, and spectral cues may also interfere in the determination of the attack. Zwicker and Fastl (1999) introduced the notion of subjective duration, and showed that the subjective duration is longer than the objective durations for durations below 100ms. Even more subjective deviations are found, if pauses are compared to tones or noises. Zwicker and Fastl found that long sounds (above 1 second) has the same subjective durations than pauses, while shorter pauses has significantly longer subjective durations than sounds. Approximately 4 times longer for 3.2kHz tone, while 200Hz tone and white noise have approximately half the subjective duration, as compared to pauses. This is true for durations of around 100-200 ms, while the difference evens out to disappear at 1sec durations. Finally Zwicker and Fastl related the subjective duration to temporal masking, and give indications that musicians would play tones shorter than indicated in order to fulfill the subjective durations of the notated music. Fraisse (1982) give an overview of his important research in rhythm perception. He states the range in which synchronization is possible to be between 200 to 1800 msec (33-300 BPM). Fraisse furthermore has

analyzed classical music, and found two main durations that he calls temps longs (>400msec) & temps courts, and two to one ratios only found between temps longs and courts. As for natural tempo, when subjects are asked to reproduce temporal intervals, they tend to overestimate short intervals (making them longer) and under-estimate long intervals (making them shorter). At an interval of about 500 msec to 600 msec, there is little over- or under-estimation. However, there are large differences across individuals, the spontaneous tempo is found to be between 1.1 to 5 taps per second, with 1.7 taps per second most occurring. There are also many spontaneous motor movements that occur at the rate of approximately 2/sec, such as walking, sucking in the newborn, and rocking.

Friberg (1991), and Widmer (2002) give rules to how the dynamics and timing should be changed according to the musical position of the notes. Dynamic changes include 6db increase (doubling), and up to 100msec deviations to the duration, depending on the musical position of the notes. With these timing changes, Snyder (2000) indicate the categorical perception of beats, measures and patterns. The perception of deviations of the timing is examples of within-category distinctions. Even with large deviations from the nominal score, the notes are recognized as falling on the beats.

As for melodic perception, Thomassen (1982) investigated the role of interval as melodic accents. In a controlled experiment, he modeled the anticipation using an attention span of three notes, and found that the accent perception is described 'fairly well'. The first of two opposite frequency changes gives the strongest accentuation. Two changes in the same direction are equally effective. The larger of two changes are more powerful, as are frequency rises as compared to frequency falls.

3   Model of pitch gestures

Music is typically composed, giving intended and inherent emotions in the structural aspects, which is then enhanced and altered by the performers, who change the dynamics, articulations, vibrato, and timing to render the music enjoyable and musical. In this work, the gestures in the pitch contour are investigated. Jensen and Kühl (2009) investigated the gestures of music through a simple model, with positive or negative slope, and with positive or negative arches, as shown in figure 1. For the songs analyzed, Jensen and Kühl found more negative than positive slopes and slightly more positive than negative arches. Huron (1996) analyzed the Essen Folk music database, and found - by averaging all melodies - positive arches. Further analyses were done by comparing the first and last note to the mean of the intermediate notes, revealing more positive than negative arches (39% and 10% respectively), and more negative than positive slopes (29% and 19% respectively). According to Thomassen (1982) the falling slopes has less powerful accents, and they would thus require less attention.
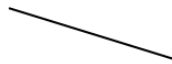
Figure 1. Different shapes of a chunk. Positive (a-c) or negative arches (g-i), rising (a,d,g) or falling slopes (c,f,i).

The generative model is made through statistical models based on data from a musical database (The Digital Tradition 2010). From this model, note and interval occurrences are counted. These counts are then normalized, and used as probability density functions for note and intervals, respectively. This statistics are shown in figure 2. As can be seen, the intervals are not symmetrical. This corroborates the finding in Jensen and Kühl (2009) that more falling than rising slopes are found in the pitch of music. According to Vos and Troost (1989), the smaller intervals occur more often in descending form, while the larger ones occur mainly in ascending form. However, since the slope and arch are modelled in this work, the *pdf* of the intervals are mirrored and added around zero, and subsequently weighted and copied back to recreate the full interval *pdf*. It is later made possible to create a melodic contour with a given slope and arch characteristics, as detailed below.
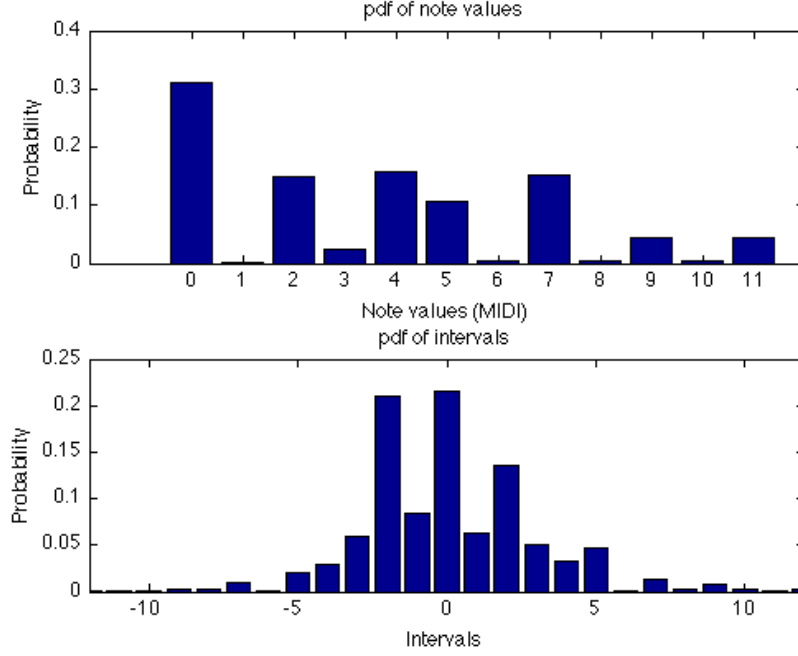
Figure 2. Note (top) and interval probability density function obtained from The Digital Tradition folk database.

In order to generative pitch contours with gestures, the model in figure 1 is used. For the pitch contour, only the neutral gesture (e) in figure 1, the falling and rising slope (d) and (f), and the positive and negative arches (b) and (h) are modeled here. The gestures are obtained by weighting the positive and negative slope of the interval probability density function with a weight,

$$pdf_i = \left[w \cdot pdf_i^+, (1-w) \cdot pdf_i^+ \right].$$ (1)

Here, pdf_i^+ is the mirrored/added positive interval *pdf*, and **w** is the weight. If **w=0.5**, a neutral gesture is obtained, and if **w<0.5**, a positive slope is obtained, and if **w>0.5**, a negative slope is obtained. In order to obtained an arch, the value of the weight is changed to **w=1- w**, in the middle of the gesture.

In order to obtain a musical scale, the probability density function for the intervals (**pdf_i**) is multiplied with a suitable **pdf_s** for the scale, such as the one illustrated in figure 2 (top),

$$pdf = shift(pdf_i, n_0) \cdot pdf_s \cdot w_r.$$ (2)

As **pdf_s** is only defined for one octave, it is circularly repeated. The interval probabilities, **pdf_i**, are shifted for each note **n_0**. This is done under the hypothesis that the intervals and scale notes are independent. So as to retain the register, approximately, a register weight **w_r** is further multiplied to the *pdf*. This weight is one

for one octave, and decreases exponentially on both sides, in order to lower the possibility of obtaining notes far from the original register.

In order to obtain successive notes, the cumulative density function, *cdf*, is calculated from eq (2), and used to model the probability that r is less than or equal to the note intervals $cdf(n_0)$. If r is a random variable with uniform distribution in the interval (0,1), then n_0 can be found as the index of the first occurrence of $cdf>r$.

Examples of pitch contours obtained by setting $w=0$, and $w=1$, respectively, are shown in figure 3. The rising and falling pitches are reset after each gesture, in order to stay at the same register throughout the melody.
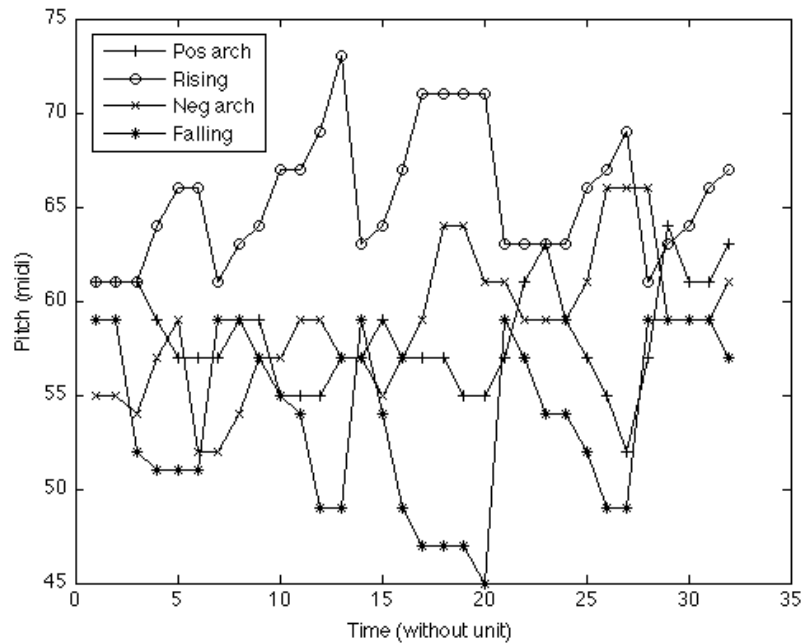
Figure 3. Pitch contours of four melodies with positive arch, rising slope, negative arch and falling slope.

The positive and negative slopes are easily recognized when listening to the resulting melodies, because of the abrupt pitch fall at the end of each gesture. The arches, in comparison, are more in need of loudness and/or brightness variations in order to make them perceptually recognized. Without this, a positive slope can be confused for a negative arch that is shifted in time, or a positive or negative slope, likewise shifted in time. Normally, an emphasis at the beginning of each gesture is sufficient for the slopes, while the arches may be in need of an emphasis at the *peak* of the arch as well.

4   Recreating pitch contours in meter

In previous work (Kühl and Jensen 2008), a generative model that produces tonal music with structures changes was presented. This model, that creates note values based on a statistical model, also introduces changes at the structural level (each 30 seconds, approximately). These changes are introduced, based on analysis of music using the musigram visualization tools (Kühl and Jensen 2008).
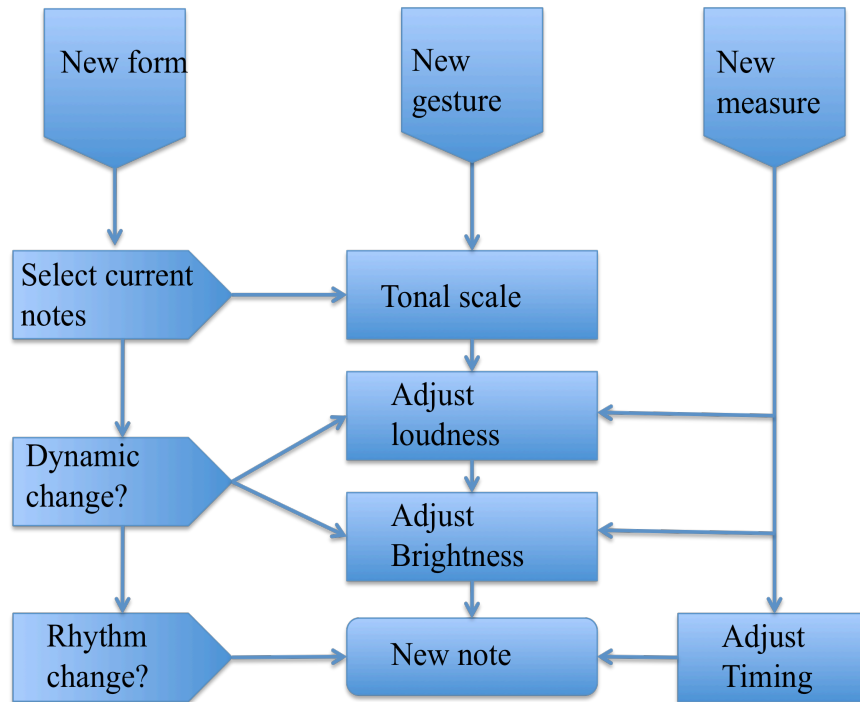

Figure 4. The generative model including meter, gesture and form. Structural changes on the note values, the intensity and the rhythm is made every 30 seconds, approximately, and gesture changes are made on average every seven notes.

With respect to chroma, an observation was made that only a subset of the full scale notes were used at each structural element. This subset was modified, by removing and inserting new notes from the list of possible notes, at each structural boundary. The timbre changes include varying the loudness and brightness between loud/bright and soft/dark structural elements. The main rhythm changes were based on the identification of short elements (10 seconds) with no discernable rhythm. A tempo drift of up-to 10% and insertion of faster rhythmic elements (Tatum) at structural boundaries were also identified. These structural changes were implemented in a generative model, which flowchart can be seen in figure 4. While the structural elements certainly were beneficial for the long-term interest of the

music, the lack of short-term changes (chunks) and a rhythm model impeded on the quality of the music. The meter, that improves the resulting quality, is included in this generative model by adjusting the loudness and brightness of each tone according to its accent. The pitch contour is made through the model introduced in the previous section.

The notes are created using a simple envelope model and the synthesis method dubbed brightness creation function (**bcf**, Jensen 1999) that creates a sound with exponentially decreasing amplitudes that allows the continuous control of the brightness. The accent affects the note, so that the loudness brightness is doubled, and the duration is increased by 25 %, with 75% of the elongation made by advancing the start of the note, as found in Jensen (2010).

These findings are put into a generative model of tonal music. A subset of notes (3-5) is chosen at each new form (superchunk), together with a new dynamic level. At the chunk level, new notes are created in a metrical loop, and the gestures are added to the pitch contour and used for additional gesture emphasis. Finally, at the microtemporal (subchunk) level, expressive deviations are added in order to render the loops musical. The interaction of the rigid meter with the more loose pitch gesture renders the generated notes a more musical sense, by the incertitude and the double stream that results. The pure rising and falling pitch gestures are still clearly perceptible, while the arches are less present. By setting *w* in eq(1) to something in between (0,1), e.g. 0.2, or 0.8, a more realistic, agreeable rising and falling gestures are resulting. Still, the arches are more natural to the ear, while the rising and falling demand more attention, in particular perhaps the rising gestures.

5 Conclusion

The automatic generation of music is in need of model to render the music expressive. This model is found using knowledge from time perception of music studies, and further studies of the cognitive and perceptual aspects of rhythm. Indeed, the generative model consists of three sources, corresponding to the immediate microtemporal, the present mesotemporal and the long-term memory macroterminal. This corresponds to the note, the gesture and the form in music. While a single stream in each of the source may not be sufficient, so far the model incorporates the macrotemporal superchunk, the metrical mesotemporal chunk and the microtemporal expressive enhancements. The work presented here has introduced gestures in the pitch contour, corresponding to the rising and falling slopes, and to the positive and negative arches, which adds a perceptual stream to the more rigid meter stream.

The normal beat as is given by different researchers to be approximately 100 BPM, and Fraisse (1982) furthermore shows the existence of two main note durations, one above and one below 0.4 secs, with a ratio of two. Indications as to subjective time, given by Zwicker and Fastl (1999) are yet to be investigated, but this may well be creating uneven temporal intervals in conflict with the pulse.

The inclusion of the pitch gesture model certainly, in the author's opinion, renders the music more enjoyable, but more work remains before the generative model is ready for general-purpose uses.

References

1. Fraisse, P. (1982). *Rhythm and Tempo*. *In* D. Deutsch (ed.) The Psychology of Music, first edition. New York: Academic Press, pp. 149-180,
2. Friberg, A. (1991) *Performance Rules for Computer-Controlled Contemporary Keyboard Music*. Computer Music Journal 15(2). pp 49-55
3. Gordon, J. W. (1987) *The perceptual attack time of musical tones*, Journal of the Acoustical Society of America, pp 88-105
4. Handel, S. (1989) *Listening*. Mit Press
5. Huron, D. (1996). The Melodic Arch in Western Folk songs. *Computing in Musicology*, 10, 3-23.
6. Jensen, K. (1999) *Timbre Models of Musical Sounds*, PhD Dissertation, DIKU Report 99/7.
7. Jensen, K. (2010) *Investigation on Meter in Generative Modeling of Music*, Proceedings of the CMMR, June 21-24, Malaga.
8. Jensen, K., O. Kühl, (2009) *Towards a model of musical chunks,* Lectures Notes in Computer Science, Springer-Verlag, LNCS5493 pp. 81-92.
9. Kühl, O., K. Jensen (2008) *Retrieving and recreating Musical Form,* Lectures Notes in Computer Science, Springer-Verlag, LNCS 4969. pp. 270-282.
10. Kühl, O. (2007) *Musical Semantics*, Bern: Peter Lang.
11. Lerdahl, F. and Jackendoff, R. (1983) *A Generative Theory of Tonal Music*. Cambridge, Mass.: The MIT Press.
12. Malbrán S. (2000) *Phases in Children's Rhythmic Development*. *In* R. Zatorre and I. Peretz (eds.) The biological foundations of music. Annals of the New York Academy of Sciences.
13. Papadopoulos G., and G. Wiggins (1999) *AI methods for algorithmic composition: a survey, a critical view and future prospects*. AISB Symposium on Musical Creativity, pp 110-117
14. Patel, A. and I. Peretz (1997) *Is music autonomous from language? A neuropsychological appraisal*. *In* I. Deliège and J. Sloboda (eds.) Perception and cognition of music. Hove: Psychology Press, pp. 191-215
15. Samson S., N. Ehrlé, M. Baulac (2000). *Cerebral Substrates for Musical Temporal Processes*. *In* R. Zatorre and I. Peretz (eds.) The biological foundations of music. Annals of the New York Academy of Sciences.
16. Snyder, B. (2000) *Music and Memory*. *An Introduction*. Cambridge, Mass.: The MIT Press.
17. The Digital Tradition (2010), http://www.mudcat.org/AboutDigiTrad.cfm, visited 1/12-2010.
18. Thomassen, J., M. (1982). Melodic accent: Experiments and a tentative model, J. Acoust.Soc. Am. 71(6), 1596-1605
19. Vos P.G. and J. M. Troost (1089). Ascending and Descending Melodic Intervals: Statistical Findings and Their Perceptual Relevance. Music Perception, 6(4), pp. 383-396
20. Widmer, G. (2002). *Machine discoveries: A few simple*, *robust local expression principles*. Journal of New Music Research, 31, 37-50.
21. Zwicker, E. and Fastl, H. (1999). *Psychoacoustics: facts and models*. Springer series in information sciences. Berlin, 2nd updated edition.