

Running head: AUTOMATED MOOD QUANTIFICATION

Automated Mood Quantification of Contemporary Music

Chris Cooke

Capitol College

ABSTRACT

This study investigated the possibility of automatically quantifying mood according to two parameters, energy and tension, with values similar to those that would be assigned by humans. Humans completed an Activation-Deactivation Adjective Check List on fifty 20-second music clips with moods varying across samples but uniform within samples. Human measurement of energy and tension were calculated from the survey results. Thirty features relevant to mood were extracted from each of the clips using the Marsyas framework. The features were used as input to neural networks in a variety of configurations using the extracted features paired with the centroid of the human measurements to create training data. Most configurations performed slightly worse than human performance in the median, and with significantly higher dispersion. One configuration had better median performance than human evaluation with acceptable dispersion between the first and third quartile, but a large dispersion across the entire set of results. While these initial results are promising, further refinement will be needed to apply the techniques to real-world systems.

AUTOMATED MOOD QUANTIFICATION OF CONTEMPORARY MUSIC

Introduction

When someone sits down to listen to music, how does he decide what to play? In some cases one may have a particular musician or genre in mind, but sometimes one wants to hear music of a certain mood. Since mood varies within a genre or the work of a particular musician, play list selection in this case can be problematic unless the listener has extensive knowledge of the available library holdings. Consider the case of an online music site with hundreds of independent artists and thousands of songs, with more of each being added daily. The ability to build a play list based on mood under such conditions seems hopeless without automation.

Music information retrieval based on metadata such as artist name and song title has become routine. The chief problem is associating the metadata with a track in the first place, but with the advent of ID tags and databases such as the Gracenote database (Gracenote, 2006), this issue is solved even for the music of serious amateurs. Less work has been done in extracting information from the audio content of the song itself. Research into automatic genre classification shows promising results, but little has been done to detect mood.

Feng et al. (2003) used tempo features to classify mood into four basic categories. Liu et al. (2003) used a more sophisticated set of audio features to classify acoustic music clips into one of four moods derived from Thayer's model for mood and arousal. Thayer (1989) proposed a model for mood that consists, in its basic form, of two dimensions of arousal, energy and tension. Liu's team quantized a planar segment of these dimensions into four distinct moods by pairing high vs. low energy with high vs. low tension (Liu's paper uses the term "stress".) Liu's team had some success automatically classifying 20-second sound clips into one of these four

categories using spectral and time-based features as input to a hierarchical Gaussian Mixture Model, which used probabilistic methods to build a decision tree for classification.

While Liu's work represents a good first step, one might ask if a four-bin classification is too broad for practical play list construction based on mood. Rather than classifying a song into one of four moods, there might be some value in describing the degree of energy and tension arousal with some metric. Such a measure could then be associated with the audio file and be used to match a similar metric derived from the listener's mood, perhaps using a distance algorithm or a fuzzy expert system. The central question, then is given a set of extracted features, can a machine-learning algorithm create a measure of energy and tension arousal that is comparable to that of a human listening to the same clip?

Feature Selection

Musical features in human terms

Let us consider the qualitative characteristics of music (from a human perspective) that contribute to mood, and discuss which features would be useful in describing these characteristics. There are many characteristics that contribute to the overall mood of a piece. Some, such as the lyrics of a song or the mode of the music, require a degree of semantic interpretation that is beyond the current capabilities of the art. Others can be considered and represented with features that can be readily extracted. Pitch, harmonics, and beat contribute distinct effects to the mood, and can also be represented by readily extractable mathematical features. The Appendix has a technical overview of the Short-Time Fourier Transform (STFT), Mel-frequency Cepstral Coefficients (MFCC), Zero Crossings (ZC), Beat Histogram (BH), and Pitch Histogram (PH). Briefly, the STFT and MFCC provide information about the frequencies present in a sample, how much is concentrated in lower frequencies, and how much it varies

within a sample. The MFCC scales the sample in a manner that resembles the way humans hear sounds, making it a perceptual measure. Zero crossings provide a measure of noise. Beat histograms provide a measure of frequency with which sounds happen at different periodicities, whereas pitch histograms do a similar analysis over a shorter time period to measure the distribution of pitches.

Lower pitches tend to be interpreted as darker in mood, while higher pitches tend to be “lighter” and “brighter”. The human psyche tends to associate darkness with stillness, and light with activity. Because darkness impairs vision and makes it more difficult to perceive dangers, people also tend to associate darkness with anxiety. Hence, the overall frequency composition of a music sample can influence both the tension and the energy. An unfolded pitch histogram can capture an overall metric of high versus low pitches in a sample. Since STFT and MFCC features consider the frequency domain characteristics of a sample, they also provide a measure of this characteristic.

Contemporary music is rarely composed of pure tones. Most modern instruments generate harmonics, and the interaction of notes played simultaneously by several instruments introduces beat frequencies and additional harmonics. Certain combinations of harmonics are perceived as dissonance, adding both tension and movement (hence energy) to a sample. In addition to simultaneous harmonic dissonance, there is also dissonance associated by the sequence of notes played over time as they move from tension to resolution within a scale. One can measure simultaneous harmonic content with the same frequency domain measures already discussed, the STFT and MFCC, as well as spectral contrast and spectral flatness features. The folded pitch histogram can estimate the moving tension of a piece for Western music by

providing a measure of the degree to which notes in a piece stay in resolved versus tense portions of the scale.

Both the overall tempo of a music clip and the amount by which the tempo varies contribute to the mood of the clip. Faster music tends to be associated with more energetic moods, whereas one interprets slower tempos as having less energy. Very regular beats have less tension, whereas syncopated or unclear beats add tension. The beat histogram provides a measure of the overall tempo and its variability within a music sample (Gibson, 2005).

Audio features in mathematical terms

Selection of a good feature set is critical to successful computer audition, playing more of a role in the success of a system than the specific machine-learning algorithm employed. Machine-learning systems perform better with a smaller set of very informative features, rather than a larger set of generic features. Much of the work in feature selection has been applied to automatic genre classification, artist identification, or song identification. Tzanetakis (2002) proposed a 30-dimension feature set for general music description that included features based on STFT, MFCC, zero crossings, beat histogram, and pitch histogram. Carvalho and Chao (2004) evaluated STFT, Spectral Contrast Features (SCF), Spectral Flatness Measure (SFM), and MFCC features separately for mood classification into two categories, and found MFCC marginally outperformed STFT and SFM. Liu et al. (2003) used a similar feature set as that proposed by Tzanetakis, omitting the MFCC features and adding spectral contrast and intensity features.

Since the purpose in this paper is to quantify mood according to the dimensions in Thayer's model, the features of interest must convey the amount of energy and tension in the music sample. Tzanetakis' feature set provides a good starting point, and has the advantage of

being partially implemented in his Marsyas framework. This study used STFT, MFCC, zero crossing, and beat histogram features. Pitch histogram features might produce interesting results, but time constraints prohibited their inclusion. STFT, ZC, and MFCC features were evaluated over a 23-millisecond analysis window. The mean and standard deviation of each were taken over a 1 second texture window. The measures for all texture windows were accumulated over the duration of the clip, then averaged into a single mean and standard deviation measure for the entire clip. Here is a summary of the specific features extracted:

- Mean and standard deviation of spectral centroid, rolloff, and flux: (6 dimensions)
- Mean and standard deviation of zero crossings: (2 dimensions)
- Mel-frequency Cepstral Coefficients: only the first five coefficients used (5 dimensions)
- Beat histogram features: position, ratio, and relative amplitude of first two peaks, sum from 40 to 90 BPM, 90 to 140 BPM, and 40 to 250 BPM (8 dimensions)

Methods

In order to compare machine performance to human performance, we produced a measure of energy and tension arousal based on human judgment. These measures were then used as training data for a machine-learning system that took extracted features as input, and adjusted its parameters to minimize the error between its measure of arousal and the human input. Thus, there are two sides to this study to be considered, the human side and the machine side, with both sides using the same library of samples.

Deciding on the size of the sample library was not easy. It needed to be large enough to represent a good portion of the mood space and provide sufficient training examples to the neural network, yet still be small enough that human volunteers could assess each sample in the library. To build the sample library, we asked amateur and semi-professional musicians at an online

music site to submit 20-second clips of their work, with a consistent mood throughout a 20 second clip. We encouraged the musicians to provide clips with a variety of moods. The final sample library contained 50 samples, which was judged to be the maximum that participants could evaluate without undue fatigue, but still be large enough to provide reasonable training to the neural network.

Energy and tension arousal according to Thayer's model seem orthogonal, but are not actually independent. In fact, they are not even linearly dependent; for low values of tension energy and tension are positively correlated, but at higher values of tension they negatively correlate. (Thayer, 1989) Thus one would not expect good results if one were simply to play a clip and ask a human to rate the amount of each type of arousal. A common psychological test, the Activation-Deactivation Adjective Check List (AD ACL), exists for assessing the mood of a human subject within the dimensions of Thayer's model. The test consists of a series of 20 adjectives, 5 each for activation and deactivation of energy and tension. (See Table 1). When assessing human mood, the subject picks for each of these adjectives the degree to which the adjective describes their mood at that moment. Each degree is associated with a point value, and each adjective is associated with one of the arousal dimensions. The point values for each of the adjectives in each dimension are totaled to provide a measure of arousal for that dimension. Rather than apply this model directly to the human subject, this study played a music clip, and then asked the human subjects to describe the extent to which the adjective described the music. Since the practical environment of interest in this study is an online music site, we constructed an online survey to implement the AD ACL test. Survey participants were solicited from forum participants at the same online music site from which music clips were solicited. Each page of the survey provided a Macromedia Flash object for playing the music clip, then listed the AD

ACL adjectives in rows, with radio buttons in columns corresponding to the degree of description. (Figure 1) The AD ACL test gives results in a range of 10 to 40 for each type of arousal. These were shifted and scaled to a range of 0-100. For each clip, the centroid of the mood metric across the range of human survey participants was calculated to provide the measure for that clip.

For the machine portion of the experiment, the features described earlier were extracted from each clip using the Marsyas 0.2 package. The results from each clip were combined with the centroid human evaluation for that clip to provide one training sample to a machine-learning algorithm. Since the desired outputs were two continuous values in the range of 0-100, linear regression algorithms were used. The machine-learning portion was implemented as a neural network using EasyNN-Plus.

Evaluation and Discussion

Human mood evaluation

Six people responded completely to the survey. For each music clip, the centroid of human mood measures was calculated. Figure 2 shows a plot of the centroids for each clip within the Thayer mood space. (We will discuss the ten validation clips later.) The Euclidean distance between each participant's evaluation and the centroid was then calculated, and then these distances were averaged for each clip to provide a measure of overall human deviation from the centroid for that clip. Since the neural network required percent error thresholds to determine validation success, distances were also calculated for energy and tension separately, and then the percent error was determined by dividing this distance by the mean. The percent error for each participant was then averaged for each clip. Table 2 shows the results of the human evaluation of mood as captured by the online survey. These data support the psychological research that

indicates mood interpretation is highly variable between people. The median deviation was nearly 20 units, or a fifth of the distance across one dimension of the plane segment, with an overall spread from 8 to 28 units. Even if the automated system adequately estimates values with no more deviation than the humans, the ranges may be too broad for practical use without training the machine-learning algorithm to the tastes of a particular user. The percent deviation for energy and tension show that humans did not perform consistently better in estimating one dimension over the other.

Machine evaluation

The audio features specified earlier were extracted using Marsyas 0.2.2. These features were combined with the centroids of the human evaluations into a comma-separated values (CSV) file for input into the neural network package. Neural networks with various parameters were implemented using EasyNN-Plus version 7.0j. Single layer networks were attempted first, with both 4 and 12 hidden nodes. Results improved by adding a second hidden layer. Finally, a two-layer network was used with groups of features omitted.

Ten clips from the set of 50 were picked for validation purposes. Initially, these ten were selected at random by EasyNN. Some of these samples contained data outside the resulting training range of the other 40 clips, hence could not be used for validation. In these cases, the clip was swapped out of the training set for another clip, until an acceptable set of ten validation points was derived.

Comparative analysis

Except where otherwise noted, each network configuration was trained until the average error was less than 1% and all validation samples were within 50% of the expected result. The network's value of energy and tension for the validation clips was compared to the centroid of

human evaluation for the clip, and used to estimate the deviation of the neural network from human performance. Table 3 lists the minimum, first quartile, median, third quartile, and maximum deviations for each configuration. Figure 3 presents these statistics in graphical form with box-whisker plots.

Leave one out analysis

To verify that validation results were not specific to the particular clips chosen for the validation set, a “leave one out” analysis was performed for each configuration. The “leave one out” analysis gives an idea of the extent to which a network will generalize to new data (Schneider, 1997). For this analysis all 50 clips, without regard to prior designation, were divided into 5 pools of 10. The network was trained for a fixed number of cycles (chosen to be the number of cycles at which training stopped in the comparative analysis) with 4 of the pools used for training and one used for validation. After the fixed number of cycles, validation results were calculated, and then the pools were rotated. This process was repeated 4 times, so that each pool was used for training 4 times and validation once. The entire analysis was performed with two different methods of breaking the samples into pools: first clips 1-10 were pool 1, 11-20 were pool 2, etc., then the starting clip for each pool was picked randomly.

Single-layer network

In the comparative analysis, neural networks with one hidden layer quickly trained to reduce the average error to <1%. A network with 12 nodes on the hidden layer needed only 75 cycles, compared to the 378 required by the a 4 node network. In both cases no validation samples were correct with even 50% error in both energy and tension. Rather than stopping the training based on validation success, the comparative analysis for the single-layer networks stopped when average error was reduced to 1%. The validation clips were then queried against

the trained network and the deviation from the human centroid computed for comparison purposes.

In the “leave one out” analysis for the single layer networks, the number of cycles allowed for training was set to roughly 100 times that needed in the comparative analysis, 3900 cycles for the 4 node network and 700 for the 12 node network. Under these conditions, the networks validated, with fairly consistent results between pools. Table 4 shows the results for the 4-node network, and Table 5 the 12 node results. Single-layer networks of both sizes performed reasonably consistently regardless of the chosen validation set.

Dual-layer network

For the network with two hidden layers, EasyNN chose an optimal network size of 12 nodes on the first hidden layer and 8 on the second hidden layer. In the comparative analysis, this network required 400 cycles for 100% of the validation samples to come within 50% of the expected result. Table 6 shows the results of the leave one out analysis, which again has reasonably consistent results.

Dual-layer network with feature subsets

The selection of 30 features was based on research attempting to create a general set of descriptive features for a variety of MIR applications. Since a small number of descriptive features tends to perform better than a large number of less descriptive features, one might reasonably ask if 30 features are really required for the specific problem of mood quantification. To address this consideration, the dual-layer network was rebuilt and retrained with categories of features removed from the input. This led to a lot of additional configurations, not all of which yielded interesting results, so not all configurations were included in the full comparison.

With the STFT and zero crossing features removed from the inputs, EasyNN calculated an optimal 9 nodes on the first hidden layer and 5 on the second. This network trained in based on validation error in 500 cycles. With the STFT and zero crossing features restored, and the beat histogram features removed, EasyNN calculated an optimal 10 nodes on the first hidden layer and 9 on the second. This network trained in based on average error in 100 cycles, but would not validate even after 6000 cycles. With just the MFCC features removed, results were similar to when the beat histogram features were removed. This configuration was not considered interesting enough to record the specific results.

Finally, a network was tried with just the MFCC features, removing the STFT, zero crossing, and beat histogram features. Given this set of input, EasyNN calculated an optimal 8 nodes on the first hidden layer and 6 on the second. This network trained in 100 cycles based on reducing validation error to less than 50%. The analyses were performed a second time with the tolerance for the validation error reduced to 40%. In this case, the network still required 100 cycles to train. Table 7 shows the results of the “leave one out” analysis for a 50% validation tolerance, and Table 8 for a 40% tolerance.

Discussion

Findings

The dual-layer neural networks generated answers of varying quality with reasonable computational effort for a real-world implementation. Most dual-layer configurations had median values a little higher than the human evaluation for the same set of clips, but the spread was significantly greater. Reasonable results were achieved when the set of features was reduced to just the MFCCs. In this case, the median lowered (indicating better estimation of the human centroid), with a comparable difference between the first and third quartiles. Unfortunately, the

difference between the minimum and maximum values was much greater than the spread among humans. The implication is that 50% of the time, a neural network with 2 hidden layers using only MFCC features will pick a value for energy and tension that is within the variation experienced by multiple humans evaluating the same clip. The rest of the time, though, the automated system results may vary widely from human evaluation. Half of these, though, vary in favor of the machine side, so only 25% of the samples might be wildly erroneous. Figure 4, Figure 5, and Figure 6 show plots in the mood space of individual human mood measures, the centroid human measure, and the neural network result for the best-case, median-case, and worst-case performance of the MFCC-only network at 50% tolerance. While these results are not good enough to recommend immediate practical use, they do indicate that the line of research is worth continued pursuit. Refinement to the human and machine side of the evaluation mechanisms may lead to better results in the future.

Possible Refinements

One possible way of reducing the human variation would be to remove either all survey responses for the participant that on the average deviates farthest from the human centroid, or remove the farthest participant on a per clip basis. Both of these alternatives run the risk of artificially inflating the precision of human judgment, since it is difficult to tell if the deviation was due to a legitimate difference in perceiving mood, or an experimental factor such as survey fatigue or not understanding the adjectives. With a sufficiently large pool of recipients, a better case could be made for an underlying normal distribution, and better statistical methods could be used for removing legitimate outliers.

Refinements might also be made on the human side by redesigning the survey, either to reduce the number of clips or the number of adjectives. Based on informal feedback from survey

participants, there is the possibility that fatigue had an influence in evaluation, despite the ability to save and return to the survey. It is also possible that the time commitment required to complete the survey discouraged participation, although no data were gathered to know for certain if this was the case.

On the machine evaluation portion, better results might be achieved with a better selection of validation data. Figure 2 shows that many of the validation points were close to the boundary of the overall region occupied by the data points. A more uniform distribution of the validation samples within this space might yield more consistent results.

Consideration of Figure 2 also shows few clips in this study covered the region of high energy and low tension or vice-versa. A more representative sample of clips, perhaps with specially composed samples when needed, might yield better results, since the results of neural networks are only valid when the query data lies within the bounds established by the training set.

Future Research

In addition to refining the current study, there are several additional research areas into which this study could be expanded.

The current study showed the best results when features were limited to MFCC. It is possible that pitch histogram features, either alone or with MFCC, may do a better job of characterizing mood, by finding dissonance and chord progressions. Pitch histogram implementation is still immature, but as the methods develop future research could look at implementing the pitch histogram and finding a set of features useful in extracting mood.

This study considered only 20-second clips of songs. To be useful in a real-world application, an entire song, possibly with mood variation within the song, would need to be

considered, and a mechanism devised for quantifying the variability of mood within a song. Some work has already been done in segmenting songs by apparent changes in various parameters. This work could be applied to the problem of automated mood quantification.

The variation of human emotion in evaluating mood leads to two possible areas of research. The first is to try training a system to the tastes of a specific person, rather than trying to create a general system that will evaluate for all people. While this might produce more consistent results, a practical implementation would be more complicated, since it would require storing the neural network weights for each individual user, and each user's mood evaluation for each song. There would also be extra effort required of the user initially in training the system. It is not necessarily the case that human variation needs to be reduced, however. Another avenue of research might take advantage of the apparent inability of humans to judge mood consistently. The expectation on the part of the human that the computer can adequately determine mood might be enough to cause the human to consider even mediocre computer results as favorable because of a placebo effect. Research in this area more properly belongs to the field of experimental psychology than computer science.

This study focused strictly on quantifying mood, with no classification whatsoever. Allowing a certain amount of classification might capture the best of both methods. This could be done by viewing the mood segment as a discrete space rather than continuous, and mapping either each human observation or the continuous centroid to the discrete points in this space. Another approach, particularly if fuzzy logic is to be applied to the problem after quantification, might be to fuzzify the human results before training the neural network. (Implementing the fuzziness directly into the human survey might also improve survey results.) The network would

then produce a fuzzy estimate of set membership for energy and tension, rather than distinct points.

Conclusion

In the introduction the question was posed, “Can a machine-learning algorithm create a measure of energy and tension arousal that is comparable to that of a human listening to the same clip?” This study shows that the answer is yes for about 75% of the samples, but the other 25% may have significantly more error than we would expect from humans. Refinements to the system may permit an unequivocal positive response to this central question.

REFERENCES

- Bray, S. & Tzanetakis, G. (2005) Distributed audio feature extraction for music, *Proc. Int. Symp. Music Information Retrieval (ISMIR) 2005*.
- Carvalo, V., & Chao, C. (2004) Emotion detection in music. Retrieved April 4, 2006 from <http://penance.is.cs.cmu.edu/11-751/projects/Vitor-Chihyu.ppt>.
- Campbell, J. (1997). Speaker recognition: a tutorial. *Proc. IEEE, 85 (9)*, pp. 1437-1462.
- Cox, E. (1999). *The fuzzy systems handbook, 2nd ed.* San Diego: AP Professional.
- Feng, Y., Zhuang Y., & Pan, Y. (2003) Music information retrieval by detecting mood via computational media aesthetics, *IEEE/WIC Int'l Conf. Web Intelligence*, p. 235
- Gibson, D. (2005). *The art of mixing: a visual guide to recording, engineering, and production, 2nd ed.* Vallejo: MixBooks.
- Gracenote (2006). Gracenote: company info. Retrieved April 17, 2006 from <http://www.gracenote.com/music/corporate/>.
- Hamburg, M. & Young, P. (1994). *Statistical analysis for decision making.* Fort Worth: The Dryden Press.
- Hinn, D. (1996). *The effect of major and minor mode in music as a mood induction procedure.* Master Thesis, Virginia Polytechnic Institute.
- Liu, D., Lu, L., & Zhang, H. (2003) Automatic Mood Detection from Acoustic Music Data. *Proc. Int. Symp. Music Information Retrieval (ISMIR) 2003*.
- Russell, S. & Norvig, P. (2003). *Artificial intelligence: a modern approach, 2nd ed.* Prentice-Hall: Upper Saddle River
- Sarle, W. (ed.) (2002). *Comp.ai.neural-nets frequently asked questions.* Retrieved June 15, 2005 from <ftp://ftp.sas.com/pub/neural/FAQ.html>.

- Schneider, J. (1997). Cross Validation. Retrieved April 18, 2006 from
<http://www.cs.cmu.edu/~schneide/tut5/node42.html>.
- Thayer, R. (1989). *The biopsychology of mood and arousal*. Oxford University Press.
- Tzanetakis, G. & Cook, P. (1999). A framework for audio analysis based on classification and temporal segmentation. *Proc. 25th Euromicro Conf. Workshop Music Tech. Audio Proc.*
- Tzanetakis, G., Essi, G., & Cook, P. (2001). Audio analysis using the discrete wavelet transform. *Proc. WSES Intl. Conf. Acoustics Music (AMTA)*
- Tzanetakis, G., Essi, G. & Cook, P. (2001) Automatic Musical Genre Classification of Audio Signals. *Proc. Int. Symp. Music Information Retrieval (ISMIR)* Oct. 2001.
- Tzanetakis, G. & Cook, P. (2002) Musical Genre Classification of Audio Signals. *IEEE Trans. Speech Audio Processing, vol 10, pp. 293-302, July 2002.*
- Tzanetakis, G. (2002). Manipulation, Analysis, and Retrieval Systems for Audio Signals, Ph.D. Dissertation, Princeton University
- Witten, I. H. & Frank, E. (2005) Data mining: practical machine learning tools and techniques, 2nd ed., San Francisco: Morgan Kaufmann.
- Zhang, T. & Kuo, J. (1998). Hierarchical system for content-based audio classification and retrieval. *Proc. SPIE Conf. Multimedia Storage Archiving Sys. III, 3527, (pp.398-409)*

APPENDIX

Short-Time Fourier Transform (STFT)

The STFT calculates the frequency-domain over a short time segment, called the analysis window. There are three main features considered with the STFT:

- Spectral centroid: the center of gravity of the magnitude within the analysis segment. This provides a measure of the average frequency in the window.
- Spectral rolloff: the frequency below which 85% of the magnitude spectrum is distributed. This provides a measure of how much of the signal is concentrated in the lower frequencies.
- Spectral flux: the squared magnitude of the difference in spectral centroid between analysis segments. This measures how much the spectral centroid changes from segment to segment.

Mel-frequency Cepstral Coefficients (MFCC)

The MFCC is based on the STFT, but the frequency domain is scaled based on a logarithmic scale that approximates the frequency response of the human ear. There are 13 coefficients, but the first five are sufficient for describing music samples (Tzanetakis, 2002).

Time-domain Zero Crossings (ZC)

This feature measures the number of times in the time domain that the signal crosses from positive to negative within an analysis window. This provides an overall measure of musical noise. For example, distorted guitar and heavy drums will give higher ZC measures than classical music.

Beat Histogram (BH)

To calculate a beat histogram, the envelope is extracted from the signal over a sliding time window. (The signal may first be broken into frequency bands with the Discrete Wavelet Transform.) The envelopes are autocorrelated to find places where the signal is most similar to itself. Peaks are added to a histogram, and then features are extracted from the histogram. Figure 7 shows an example of a beat histogram for one of the clips in this study.

Pitch histogram (PH)

The pitch histogram is calculated similarly to the beat histogram, but over a much smaller window, so that the autocorrelated signals are of a higher frequency. There are two forms of pitch histogram. The unfolded pitch histogram simply shows the relative amplitude per pitch on the scale for all pitches. Thus the same note in two different octaves would have two bars on the unfolded pitch histogram. The folded pitch histogram takes the modulus of the pitch when adding to the amplitude of a pitch, so the occurrence of a note is added to the same bin regardless of octave. Furthermore, the folded pitch histogram is set up to have adjacent bins follow the circle of fifths for better comparison of related notes.

AUTHOR NOTE

The author would like to thank the musicians and staff at MacIDOL.com for their assistance with this study, and his wife and daughter for their support and understanding.

TABLES

	Activation	Deactivation
Energy	active energetic vigorous lively full-of-pep	sleepy tired drowsy wide-awake wakeful
Tension	jittery intense fearful clutched-up tense	placid calm at-rest still quiet

Table 1: AD ACL adjectives by dimension

Clip	Avg. deviation from centroid	Avg. % deviation - energy	Avg. % deviation - tension
1	23.45	44.05	37.04
2	14.17	13.89	13.73
3	26.04	28.40	48.39
4	25.69	35.92	32.31
5	20.78	23.81	17.65
6	12.70	11.11	9.01
7	27.06	65.38	50.00
8	8.75	6.18	9.52
9	13.21	7.92	13.02
10	16.23	23.42	46.06
11	21.39	30.23	22.22
12	11.31	14.04	7.91
13	17.84	48.94	41.67
14	18.97	21.67	15.09
15	12.22	5.031	13.73
16	23.61	28.46	56.74
17	18.55	16.18	16.26
18	21.57	42.50	26.53
19	17.44	35.85	36.16
20	14.59	5.94	17.33
21	22.10	36.71	23.10
22	24.68	20.43	21.41
23	19.27	25.70	22.94
24	16.22	10.50	18.70
25	20.15	25.00	46.77
26	10.81	9.00	7.69
27	13.61	36.36	38.33
28	15.93	42.85	13.73
29	23.18	33.33	59.32

30	26.53	37.59	27.10
31	19.84	49.09	28.57
32	21.80	28.37	22.69
33	20.52	37.93	32.52
34	21.80	20.29	24.19
35	16.39	15.38	17.65
36	17.26	25.16	31.22
37	28.94	39.76	44.00
38	20.90	31.43	30.71
39	16.23	6.75	20.65
40	26.09	47.95	27.27
41	22.97	20.31	32.96
42	17.82	10.13	17.72
43	16.07	25.14	36.78
44	20.50	28.89	15.92
45	18.20	16.16	33.33
46	26.57	20.99	39.22
47	18.80	21.43	8.95
48	20.28	41.94	30.67
49	28.93	28.67	38.78
50	17.87	11.48	18.91
Min	8.75	5.03	7.69
First Quartile	16.23	15.58	17.41
Median	19.56	25.15	25.36
Third Quartile	22.75	36.25	36.63
Max	28.94	65.38	59.32

Table 2: Variation in human evaluation of mood

	All - human	Validation set - human	Single layer 4 node	Single layer 21 node	Dual layer – all features	Dual layer – no STFT, ZC	Dual layer – MFCC only, 50%	Dual layer – MFCC only, 40%
Min	8.75	11.31	2.86	1.00	0.78	5.19	6.02	3.41
1 st Quartile	16.23	14.98	14.56	17.61	7.61	16.05	11.50	7.87
Median	19.56	20.70	20.70	22.33	24.11	22.04	14.34	13.32
3 rd Quartile	22.75	24.72	34.85	33.99	31.36	30.60	19.45	20.81
Max	28.94	26.57	42.71	47.65	42.88	42.05	45.58	37.48

Table 3: Comparative deviation statistics

1 layer, 4 nodes	Fixed starting rows		Random starting rows	
Pool	Validation %	Training error	Validation %	Training error
1	80.00	0.1420	80.00	0.2899
2	80.00	0.1429	70.00	0.1747
3	50.00	0.2623	60.00	0.0496
4	70.00	0.0276	50.00	0.2080
5	80.00	0.1489	70.00	0.1747

Table 4: Leave one out analysis for single-layer network with 4 nodes

1 layer, 12 nodes	Fixed starting rows		Random starting rows	
Pool	Validation %	Training error	Validation %	Training error
1	80.00	0.0873	70.00	0.2056
2	90.00	0.0279	80.00	0.1070
3	80.00	0.1976	90.00	0.0145
4	80.00	0.1070	80.00	0.0038
5	100.00	0.2023	90.00	0.1397

Table 5: Leave one out analysis for single-layer network with 12 nodes

2 layers	Fixed starting rows		Random starting rows	
Pool	Validation %	Training error	Validation %	Training error
1	60.00	0.2907	60.00	0.0030
2	90.00	0.0183	80.00	0.1276
3	60.00	0.2065	70.00	0.0143
4	70.00	0.0408	90.00	0.0750
5	80.00	0.0993	70.00	0.0603

Table 6: Leave one out analysis for dual-layer network

2 layers- MFCC only, 50% tolerance	Fixed starting rows		Random starting rows	
	Pool	Validation %	Training error	Validation %
1	80.00	0.1973	100.00	0.0256
2	80.00	0.0071	90.00	0.0267
3	70.00	0.0221	80.00	0.0127
4	90.00	0.0010	80.00	0.0006
5	100.00	0.1579	80.00	0.0138

Table 7: Leave one out analysis for dual-layer network with only MFCC features and 50% tolerance for validation

2 layers- MFCC only, 40% tolerance	Fixed starting rows		Random starting rows	
	Pool	Validation %	Training error	Validation %
1	70.00	0.1973	80.00	0.0145
2	80.00	0.0071	80.00	0.0154
3	60.00	0.0221	70.00	0.3627
4	90.00	0.0010	80.00	0.0211
5	80.00	0.1579	90.00	0.0041

Table 8: Leave one out analysis for dual-layer network with only MFCC features and 40% tolerance for validation

FIGURES

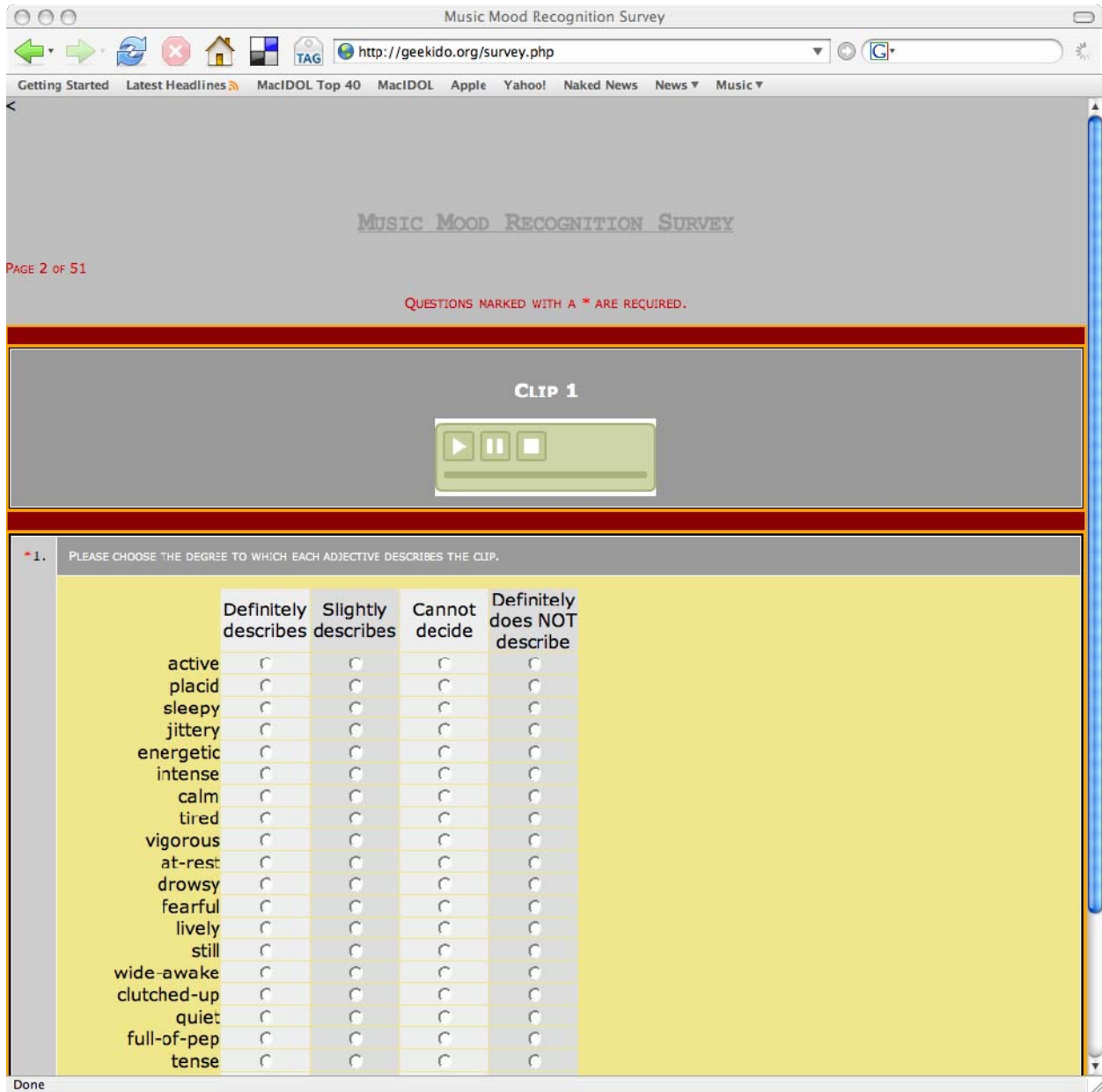


Figure 1: Sample survey page

Clip Mood Centroids from Human Evaluation

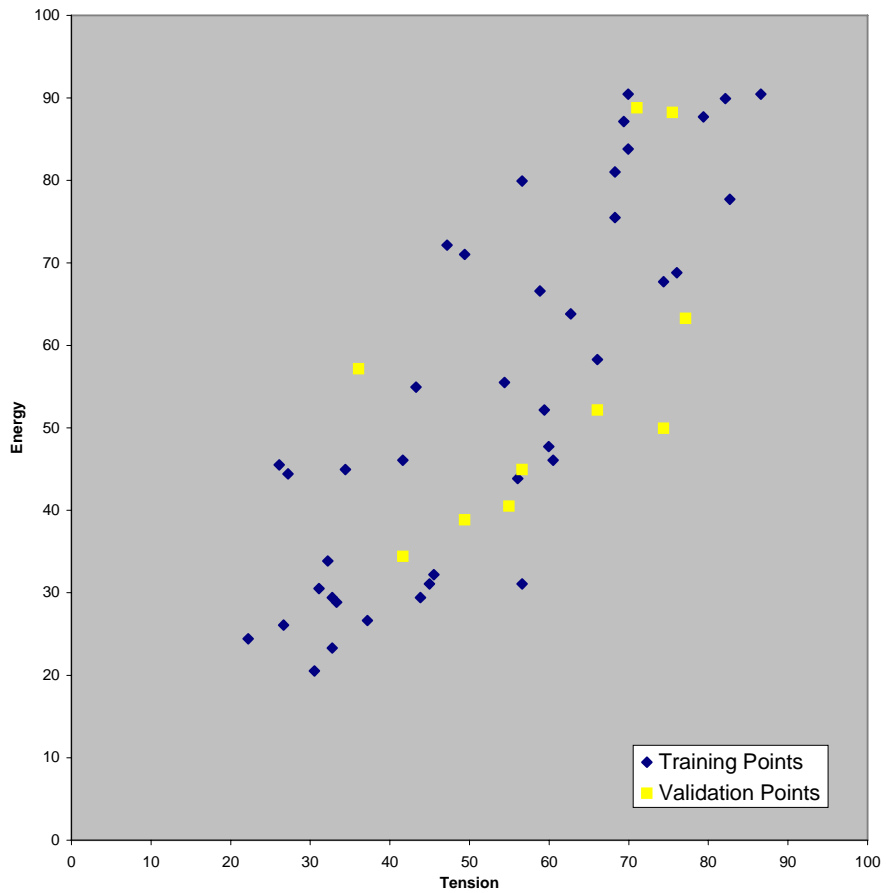


Figure 2: Centroids of human mood evaluation

Comparative Results of Mood Quantification Methods

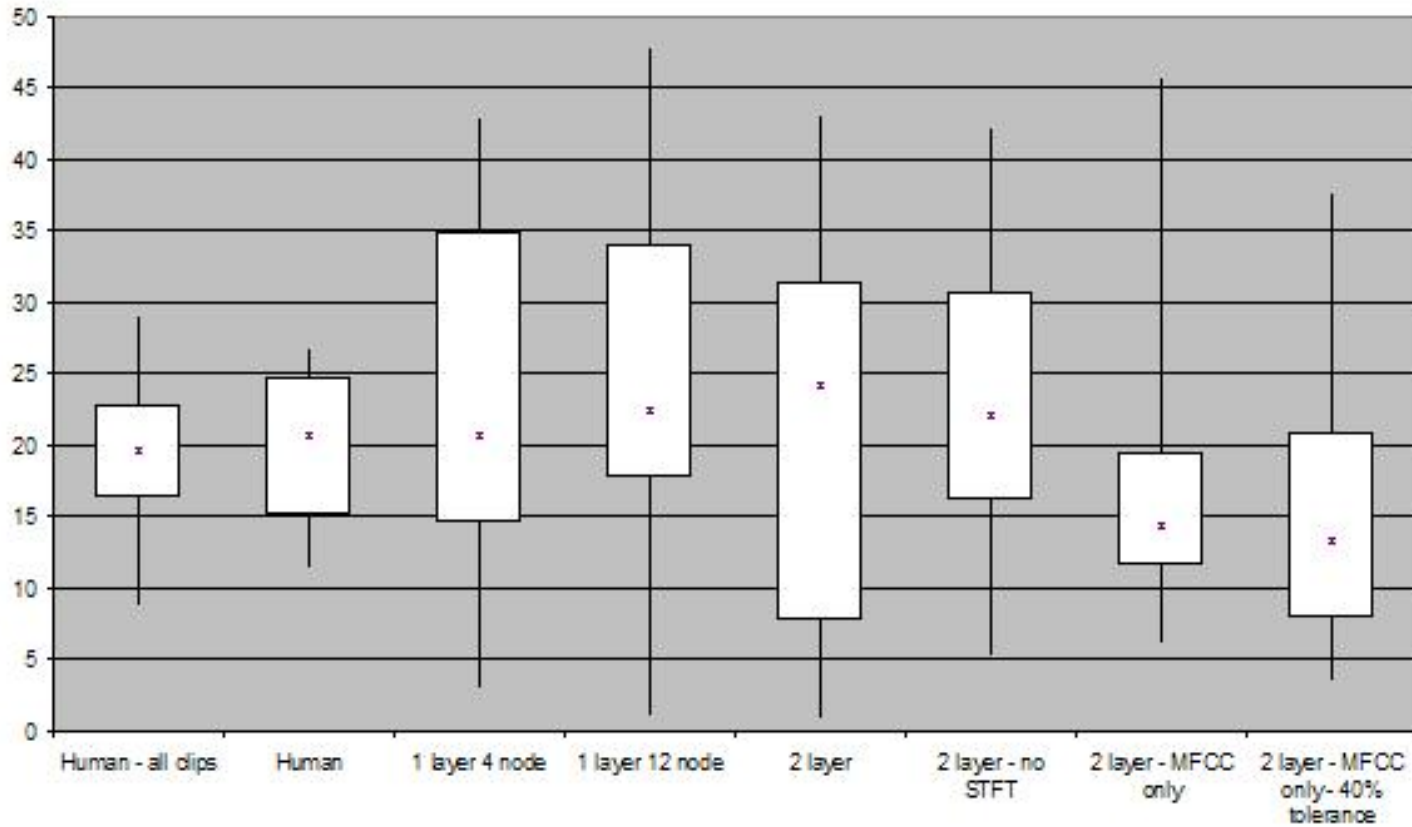


Figure 3: Comparative results

Best NN Performance

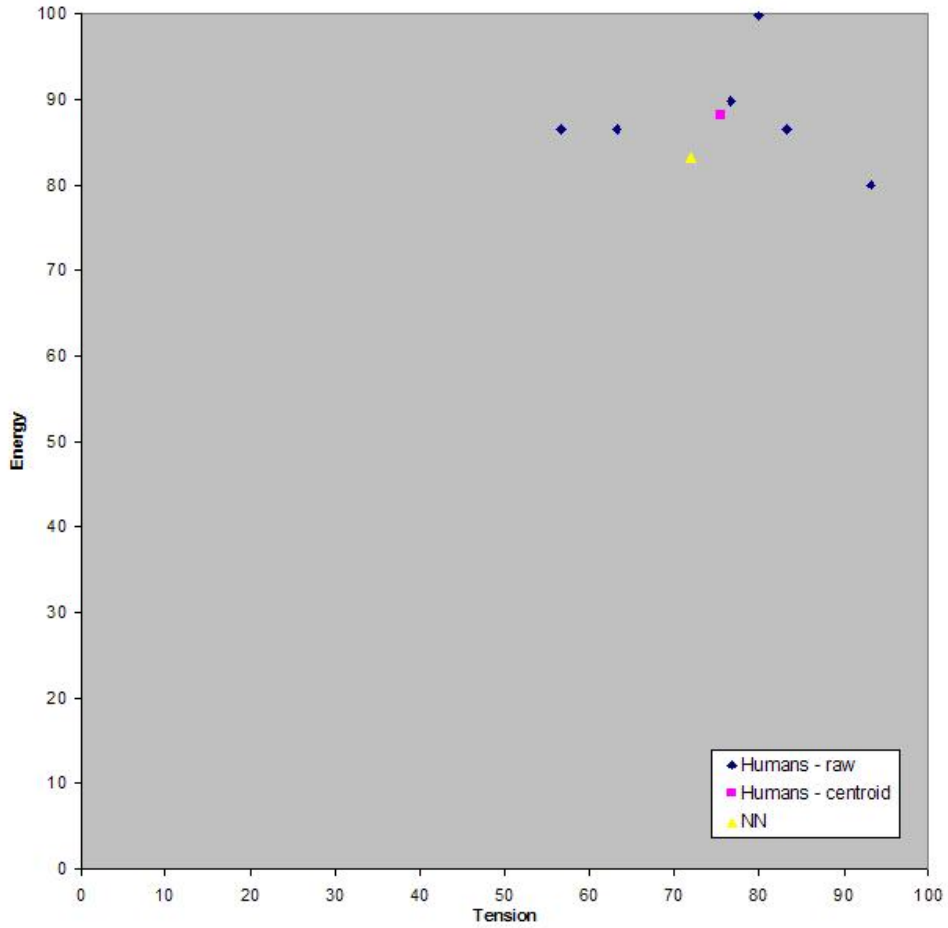


Figure 4: Best neural network performance

Median NN Performance

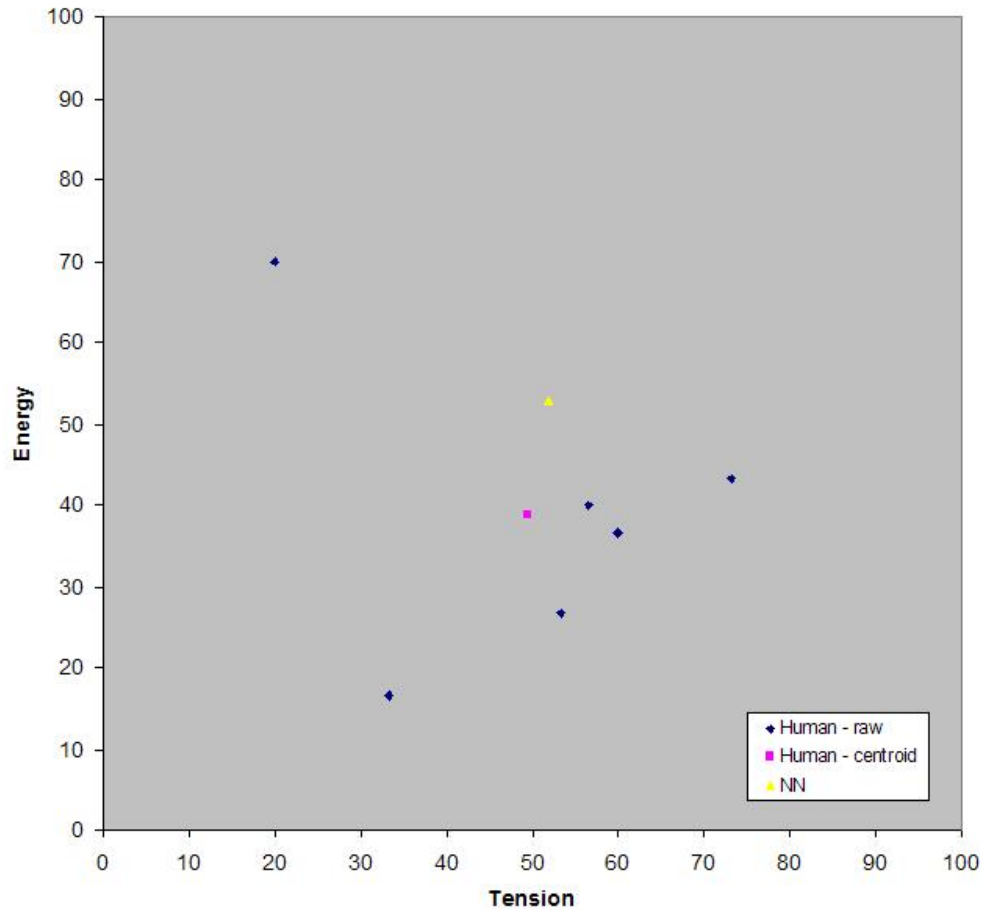


Figure 5: Median neural network performance

Worst NN Performance

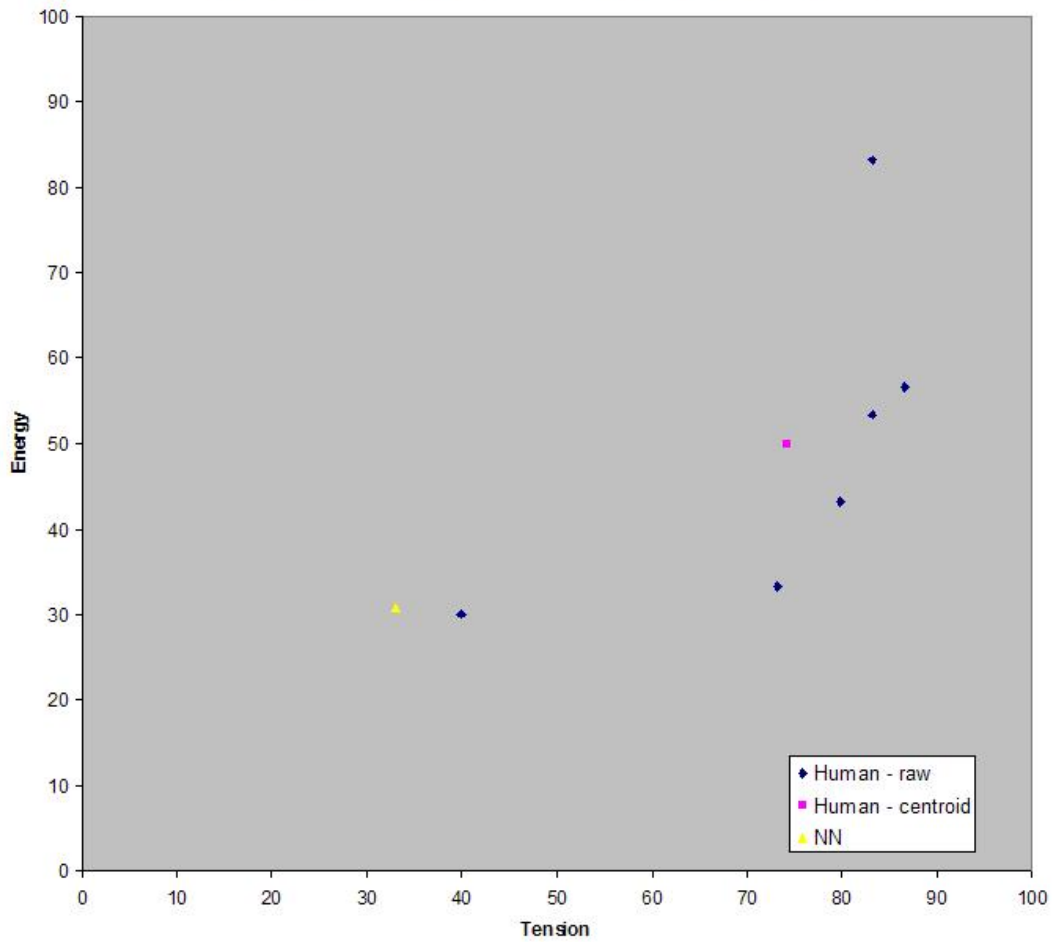


Figure 6: Worst neural network performance

Clip 4 Beat Histogram

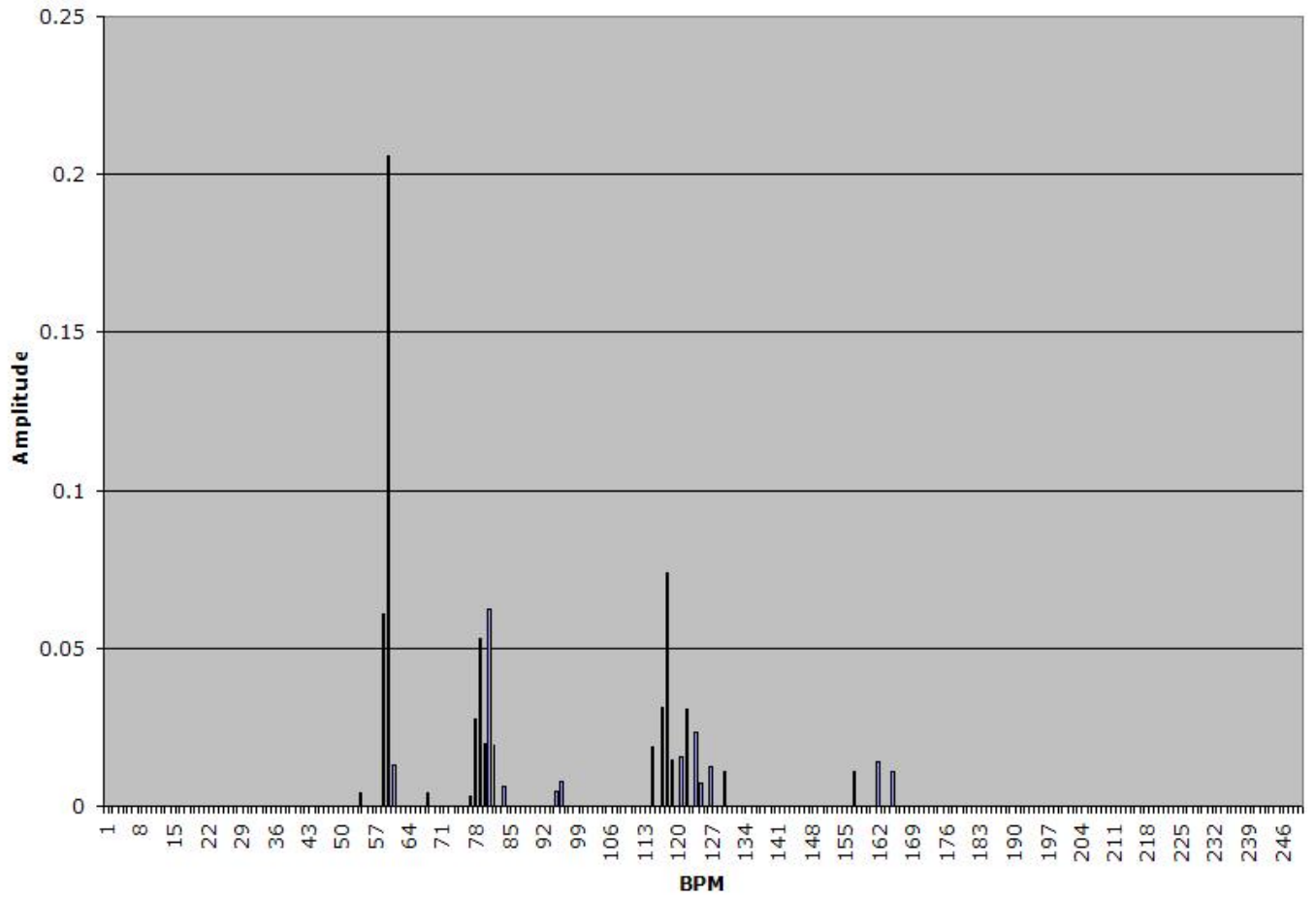


Figure 7: Example beat histogram