Music Genre Classification using an Auditory Memory Model

Jensen, Kristoffer

Published in:
Speech, sound and music processing: Embracing research in India

DOI (link to publication from Publisher):
10.1007/978-3-642-31980-8_7

Publication date:
2012

Document Version
Early version, also known as pre-print

Music Genre Classification using an Auditory Memory Model

Kristoffer Jensen[1]

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

Abstract. Audio feature estimation is potentially improved by including higher-level models. One such model is the Auditory Short Term Memory (STM) model. A new paradigm of audio feature estimation is obtained by adding the influence of notes in the STM. These notes are identified when the perceptual spectral flux has a peak, and the spectral content that is increased by the new note is added to the STM. The STM is exponentially fading with time span and number of elements, and each note only belongs to the STM for a limited time. Initial experiment regarding the behavior of the STM shows promising results, and an initial experiment with sensory dissonance has been undertaken with good results. The parameters obtained form the auditory memory model, along with the dissonance measure, are shown here to be of interest in genre classification.

Keywords: dissonance; activation; memory; decay; music classification.

1 Introduction

Audio feature extraction is useful in many situations, from digital musical instruments to music playback systems, from speech recognition to music information retrieval. This paper proposes to incorporate a high-level memory model in the feature extraction, in order to improve the estimation of the feature.

Psychologists consider memory to be the process by which we encode, store, and retrieve information. The understanding of the memory model was improved with the modal model of Atkinson and Shiffrin [1]. In this model, the stimuli first enters the sensory system, which contains memory of it's own, and then the Short Term Memory (STM), which has a limited time-span, and through rehearsal, it can then enter the Long Term Memory (LTM). According to Pashler and Carrier [2], stimuli reaches STM and LTM simultaneously, while according to Snyder [3], stimuli go through LTM to reach STM. As

only the STM is modeled here, it is not essential how stimuli reach the STM.

The STM paradigm was later replaced by the Working Memory by Baddeley and Hitch [4], to put more emphasis on the active behavior of the STM. This Baddeley and Hitsch model consists of a Central executive, and three slave systems, the phonological loop, the visuo/spatial sketchpad and the episodic buffer (which was suggested later). The capacity of the working memory was determined to be 7+-2 by Miller [5]. Most indicators show the major form of encoding in the STM is acoustic (Gross [6], p289), although this may be more a result of the process encountered, than of the property of the STM. The working memory model puts emphasis on several independent modules, and that the STM is associated with the attention processes.

The sensory store is approximately 250 ms long (Massaro and Loftus [7]). During the sensory store, the sound is subject to perceptual processing. It does not seem to be overwhelming evidence for the sensory store to be available for cognitive processing, and it is not modeled further here. It seems to be a reason for filtering, i.e. short sounds are not to be propagated into the STM.

With the increase of music on personal computers, the necessity of assisting users choosing among the songs has arisen. Such a choice can be random (Shuffle play), by automatic playlist generation, or relate to a degree of similarity between songs. Playlist generation can be done on audio features [8], for instance based on one song, or audio input, as in the query-by-humming systems [9,10,11]. Playlist generation can also be based on meta-data [12], and collaborative filtering.

This paper presents the Auditory Memory Model in chapter 2. The identification of auditory chunks, and the details of the calculation of the Auditory Memory Model content, along with the improved calculation of the sensory dissonance are presented in Chapter 3. The chapter 4 presents findings on how the Auditory Memory Model parameters and the improved sensory dissonance may improve music genre classification.

2 Models of Memory

Humans use memory to encode, store and retrieve information. Auditory information enters the brain through the auditory system and reaches the sensory store first. If the information is not reinforced, it is fading. Gross [6] gives an overview of the mechanisms of fading in the STM that include Decay (the mental representation breaks down over time), Displacement (STM has limited capacity, and old stimuli is thrown out by new stimuli), and Interference (learning is affected by context). Apparently, for all practical reasons, the limited capacity (7+-2 as reported by Miller [5]) is the main cause of memory purging in the STM. However, if no new stimulus is entered, the STM is here modeled to have a limited time span, as reported by Atkinson and Shiffrin [1].

This is modeled according to the activation model of Anderson and Lebiere [13], in which the decay is modeled as,

$$A_{decay} = 1 - d \ln(t+1), \tag{1}$$

where $d=0.5$, and $t>0$.

In order to ensure a homogenous model, the limited capacity of the STM is modeled in a similar fashion,

$$A_{displacement} = 1 - d \ln(N_c). \tag{2}$$

N_c is the number of chunks currently active in the STM. The total activation strength of an acoustic chunk is then,

$$A = A_{decay} + A_{displacement}, \tag{3}$$

and the chunk is propagated to the auditory processing, if $A>0$, or otherwise purged from the STM.

3 Encoding in Memory Models

The feature extraction is typically done using overlapping frames and extracting the relevant audio features from each frame. Here, this is done in a similar manner, with the addition of any interaction the current frame can have with the information in the auditory memory.

3.1 Chunks in the Auditory Memory

In order to encore the auditory chunks in the memory model and propagate them to the auditory processing, a method for separating auditory streams is necessary. Moore [14] gives a review of features useful for separating auditory streams that include fundamental frequency, onset times, contrast to previous sounds, correlated changes in amplitude or frequency and sound location. A useful algorithm for simulating most of these features is the perceptual spectral flux (Jensen [15]),

$$psf_+^t = \sum_{\left(a_k^t - a_k^{t-1}\right)>0} w_k \left(a_k^t - a_k^{t-1}\right), \tag{4}$$

where a_k are the magnitudes from an N point FFT and w_k is the frequency weight, in order to simulate the outer and middle ear frequency characteristics. t is the current time frame, and $t\text{-}1$ is the previous time frame. If k is the subset of all FFT bins that satisfies either $a^t\text{-}a^{t-1}>0$ or $a^t\text{-}a^{t-1}<0$, the directional spectral flux is obtained. The positive spectral flux (psf_+) is a measure of auditory onset, and the negative spectral flux, psf_-, is a measure of auditory offset. The chunk is activated when a significant level is found in psf_+. This allows the identification of the content of the auditory chunk within the sensory store time limit. By calculating the directional spectral flux, auditory events that are surrounded by concurrent auditory events can be encoded, assuming they do not start and end at the same time as the current auditory event. In order to identify the spectrum of a new note, it is calculated as the difference between the spectrums just after and just before the onset time,

$$a^n = a^{t+T} - a^{t-T}. \tag{5}$$

Here, T is set to 0.2 seconds. a^n is limited to non-negative values only. The peaks of the perceptual spectral flux is found by identifying peaks that are higher than the mean and the max in the surrounding time,

$$pk = psf > W_{mean} mean(psf(R_{mean})) + W_{\max} \max(psf(R_{\max})). \tag{6}$$

W_{mean} is here set to 0.1, and the mean is taken in the range $R_{mean}=1.5$ seconds, while W_{max} is 0.9, and $R_{max}=0.9$ seconds. The psf^+ for Stan Getz – First Song (for Ruth) is shown together with the spectrogram in figure 1.

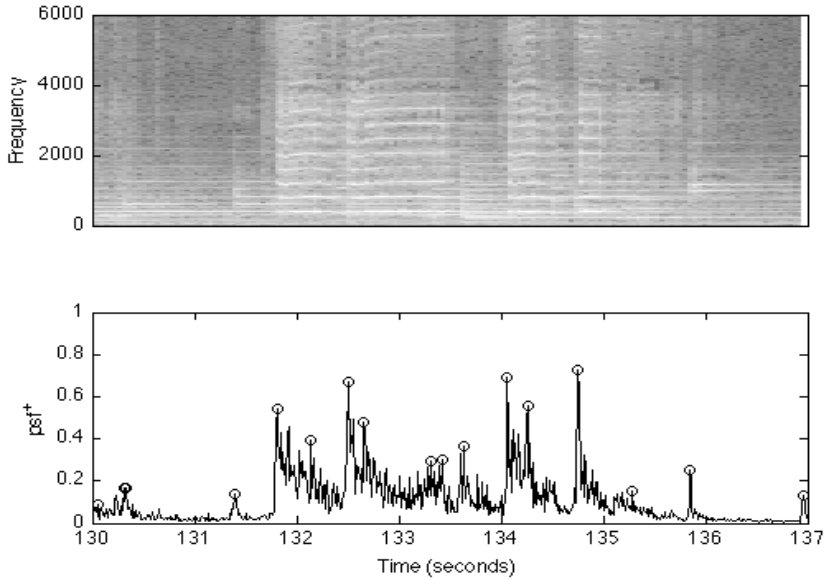

Figure 1. Spectralgram (top) and positive Perceptual Spectral Flux (bottom) for Stan Getz - First song (for Ruth) (excerpt). Found psf peaks are indicated with a ring.

As can be seen, the psf_+ peak detector captures many of the onsets, and it is therefore used as a note detector in this work. Each time the peak detector indicates a peak; a new note is inserted into the memory model. When the note has a weight (activation strength) below zero, it is purged from the STM.

3.2. Auditory Memory Content

In order to test the validity of the STM memory model and the note detection, a simulation was made on the Stan Getz – First Song (for Ruth) song. Note onsets were obtained according to eq. (6), and new spectral content according to eq. (5) is inserted into to STM for each new note. The note activation strength is calculated according to eq. (3), and notes are purged when $A<0$. Two measures were obtained, the number of elements in the STM, and the time span of the STM, taken as the time the first element has been in the STM. The results are shown in figure 2. The song gives 11.43 elements on average

(std=1.60) and an average duration of 3.02 seconds (std=0.50). The number of elements is above the 7±2 of Miller [5]. However, the elements (notes) that have been in the STM model for some time would have a low weight and very little influence. The time span of 3 seconds is a reasonable number, given that the STM has a span of 3-5 seconds according to Snyder [3].

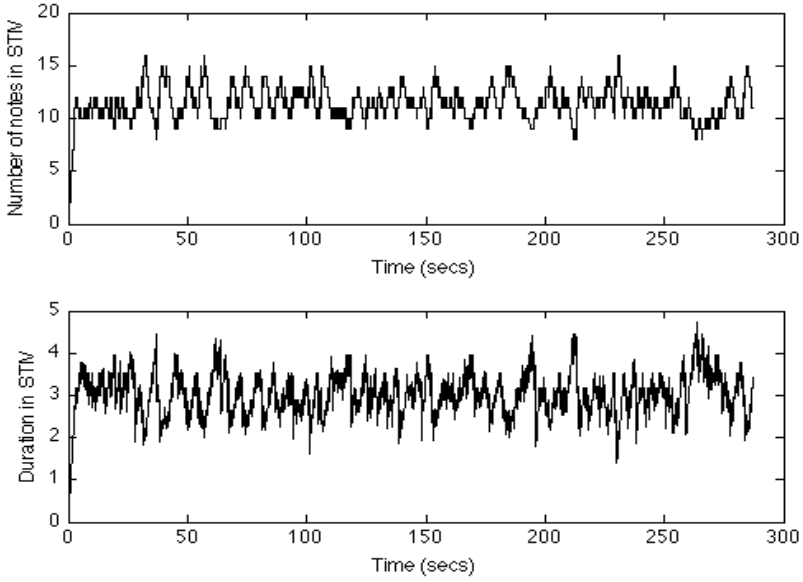

Figure 2. Number of elements in the STM (top) and duration of STM (bottom) for Stan Getz - First song (for Ruth).

3.3 Experiment

The sensory dissonance is a measure of the sum of the beatings over different auditory filters. If the beating is lower than one critical band, it adds to the total sensory dissonance, otherwise the beating disappears, and two tones appear instead. The sensory dissonance is additive (Plomp and Levelt [16]), meaning that if different partials are causing beating in different critical band, then each beating is added to the total sensory dissonance. The sensory dissonance is calculated, according to Sethares [17], as,

$$d_0 = a_1 a_2 \left(e^{\frac{f_1 - f_2}{0.0245 f_1 + 22.57}} - e^{\frac{f_1 - f_2}{0.015 f_1 + 13.74}} \right), \quad \textbf{(7)}$$

for two pure tones with frequencies f_1 and f_2 and amplitudes a_1 and a_2, and where $f_1 < f_2$.

The partials to take into account in eq (5) are the partials in the current frame, and the partials of the auditory chunks in the STM. Thus, in order to calculate the total dissonance, first the dissonance of the current frame is calculated as,

$$d_{tot} = \sum_k \sum_{l>k+1} d_0(f_k, a_k, f_l, a_l). \quad \textbf{(8)}$$

In practice, only the partials pairs within one critical band needs to be taken into account, as the influence of two partials with a distance greater than one critical band is weak. When the influence of the notes currently present in the STM is to be added, this is done in the same way, however the influence is weighted with the total activation strength,

$$d_{stm} = d_{tot} + \sum_n A^n \sum_k \sum_l d_0(f_k, a_k, f_l^n, a_l^n). \quad \textbf{(9)}$$

This is done for the spectrum of all notes n in the STM, and for the spectrum of the notes as identified in eq. (5).

The sensory dissonance (eq. 8), and the total sensory dissonance (eq. 9) are now calculated for Stan Getz – First song (for Ruth). An excerpt of the result is shown in figure 3.

The dissonance increases as can be expected when the influence of the notes in the STM is added to the dissonance. The total dissonance is approximately doubled.

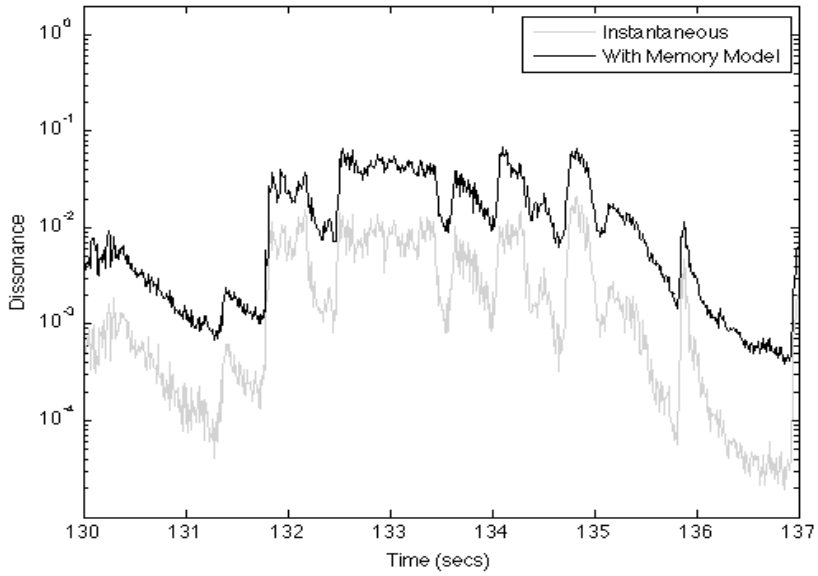

Figure 3. Instantaneous and total sensory dissonance for Stan Getz - First song (for Ruth) (excerpt).

4 Experiment

Music genre classification is an important area of research today. Much research in done in this field, using audio features [8], meta-data [12], or collaborative filtering. The general idea is that automatic genre classification may assist users in selecting the music. While music audio information (timbre, rhythm, melody/chords, etc) may be of interest in the automatic genre classification, often, genres are defined by other information, meta-data, etc. Nonetheless, audio features may still be of assistance in this field.

This work presents initial findings in the use of the features of the auditory memory model (Number of elements Duration, Dissonance) on a medium size music database. This database, that consists of 1320 songs in 11 different genres, was first used in [18]. The Auditory memory model features have been calculated for all songs, and the resulting values are shown in figure 4.

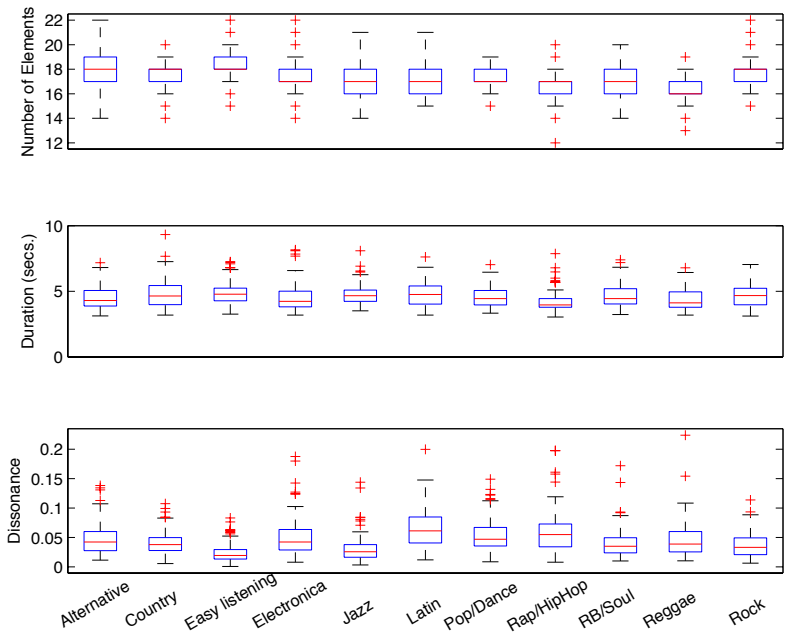

Figure 4. Boxplot of Number of element, duration and dissonance of 11 genres of music.

It is clear that there are variations in these features between the different genres. For instance, *Easy Listening* has relatively many events, while *Rap/HipHop* and *Reggae* has fewer. *Country*, *Easy Listening* and *Latin* has longer durations in the auditory memory. *Easy Listening* and *Jazz* has low dissonance while *Latin*, and *Rap/HipHop* has high dissonance.

An initial experiment has been performed to evaluate the capacity of these features to classify the music into these 11 genres. The classification is done on the mean, max and standard deviation of the Number of Elements, Duration, and Dissonance. It is done using discriminant analysis assuming normal data by fitting multivariate densities with covariance estimates.

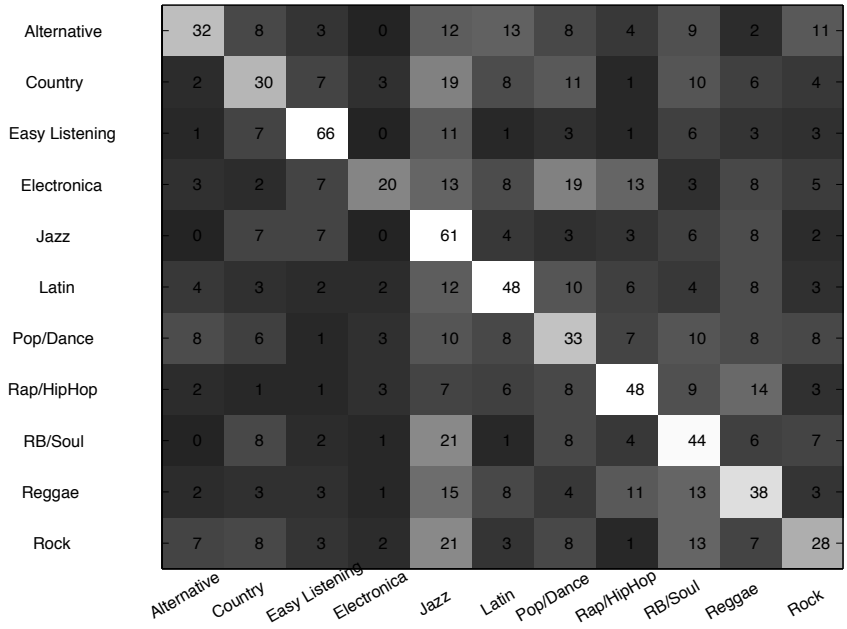

Figure 5. Confusion matrix of the classification of the 11 genres. The percentages is shown in each classification.

The resulting confusion matrix is shown in figure 5. Some genres gives promising results, this includes in particular genres with low dissonance, *Easy Listening* with 66% and the *Jazz* with 61%, but also the genres with high dissonance, *Latin* with 48% and *Rap/HipHop* with 48%. *Electronica*, *Rock*, *Country*, and *Pop/Dance* have low scores. Perhaps these genres contain too much variation within the genre to being easily classified with audio features.

All in all, the classification using the features of the auditory memory model classifies approximately 40% of the songs correctly. [12] found that humans classify approximately 57% correctly in an experiment using a subset of the same music collection. It is therefore seen as a very promising result to obtain 40% correct song identification using only the Auditory Memory Model features.

5 Conclusion

Audio features are difficult to evaluate objectively, in that the comparison to human perception is made more difficult by the necessary interpretation of the sensory perception by human subjects. One possible solution towards this mismatch is to improve the audio feature estimation in a way, so that it is closer to human perception. This is attempted in this work by the inclusion of a Auditory Short Term Memory module. New notes are inserted in the STM, if they have novelty, as measured by perceptual spectral flux, and the notes have an activation strength that is exponentially decreasing with time and the number of elements in the STM. When it is below zero, the corresponding note is purged from the STM.

An initial experiment shows that the STM behaves in a plausible way, i.e. it has an appropriate number of notes and time span, as compared to the literature.

The STM model has been used in the calculation of the sensory dissonance. When comparing the instantaneous dissonance with the total dissonance obtained by adding the dissonance between the current frame and the notes in the STM, the total dissonance has a higher mean (approximately the double) as could be expected.

The inclusion of a memory model in the estimation of audio features certainly makes sense from a theoretical point-of-view and it also gives plausible values in an initial experiment.

An experiment using the features of the Auditory Memory Model (Number of Elements, Duration in Memory, and Dissonance) shows promising results, and should be an inclusion in music information retrieval systems based on auditory features.

References

1. Atkinson, R.C.; Shiffrin, R.M. (1968). Human memory: A proposed system and its control processes. In Spence, K.W.; Spence, J.T. *The psychology of learning and motivation* (Volume 2). New York: Academic Press. 89–195.
2. Pashler H., and M. Carrier (1996). Structures, Processes, and the Flow of Information. in *Memory: Handbook of Perception and Cognition*, ed Bjork and Bjork, Academic Press. 3-29
3. Snyder, B. (2000) *Music and Memory. An Introduction*. Cambridge, Mass.: The MIT Press.

4. Baddeley, A.D., & Hitch, G. (1974). Working memory. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, 47--89). New York: Academic Press.
5. Miller G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. Psychological Review 63 (2): 81–97
6. Gross R. (2005). *Psychology: The Science of Mind and Behaviour*. Hodder Arnold Publication.
7. Massaro, D., and G. R. Loftus (1996). Sensory and Perceptual Storage. In Elizabeth Ligon Bjork and Robert A. Bjork (Eds.). *Memory*. San Diego: Academic Press. 86-99.
8. Foote, J. A similarity measure for automatic audio classification. *Proc. AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora*, Stanford, Palo Alto, California, USA. (1997)
9. McNab, R.J., Smith, L. A., Witten, I.H., Henderson, C.L. and Cunningham, S.J. Towards the digital music library: Tune retrieval from acoustic input, *Proceeding DL'96*, 11-18. (1996)
10. Rolland, P. Y., Raskinis, G., Ganascia, J. G. *Musical content-based retrieval: an overview of the Melodiscov approach and system*, ACM Multimedia (1), 81-84. (1999)
11. Ghias, A., Logan, J., Chamberlin, D. and Smith, B. C. Query by humming - musical information retrieval in an audio database, *Proceedings Multimedia*. 231-236. (2001)
12. Pauws, S. and Eggen, B. *PATS: Realization and user evaluation of an automatic playlist generator*, Proceedings of the 3rd ISMIR ,Ircam, France. 222-230. (2002)
13. Anderson, J. R. & Lebiere, C. Atomic components of thought. Hillsdale, NJ. (1998)
14. Moore, B, C., J. *Psychology of Hearing*. Academy Press. (1997)
15. Jensen, K., Multiple scale music segmentation using rhythm, timbre and harmony, EURASIP Journal on Applied Signal Processing, Special issue on Music Information Retrieval Based on Signal Processing. (2007)
16. Plomp R. and W. J. M. Levelt. Tonal Consonance and Critical Bandwidth. J. Acoust. Soc. Am. 38(4), 548-560. (1965)
17. Sethares, W. Local consonance and the relationship between timbre and scale. Journal of the Acoustical Society of America 94 (3): 1218–1228. (1993)
18. A.Meng, Temporal feature integration for music organization. Ph.D. dissertation, IMM, Denmark Technical University. (2006)