On the Use of Memory Models in Audio Features

Jensen, Karl Kristoffer

Published in:
Proceedings of the International Symposium of Frontiers of Research on Speech and Music and Computer Music Modeling and Retrieval

Publication date:
2011

Document Version
Tidlig version også kaldet pre-print

On the Use of Memory Models in Audio Features

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 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.

Keywords: features; activation; memory; decay; encoding.

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.

2 Models of Memory

The information in the STM is fading, if not reinforced. 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 [8], in which the decay is modeled as,

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

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,

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

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

$$E = C + D, \tag{3}$$

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

3 Encoding in Memory Models

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 [9] 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 [10]),

$$psf^t = \sum_k w_k \left(a_k^t - a_k^{t-1} \right), \tag{4}$$

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,

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

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.

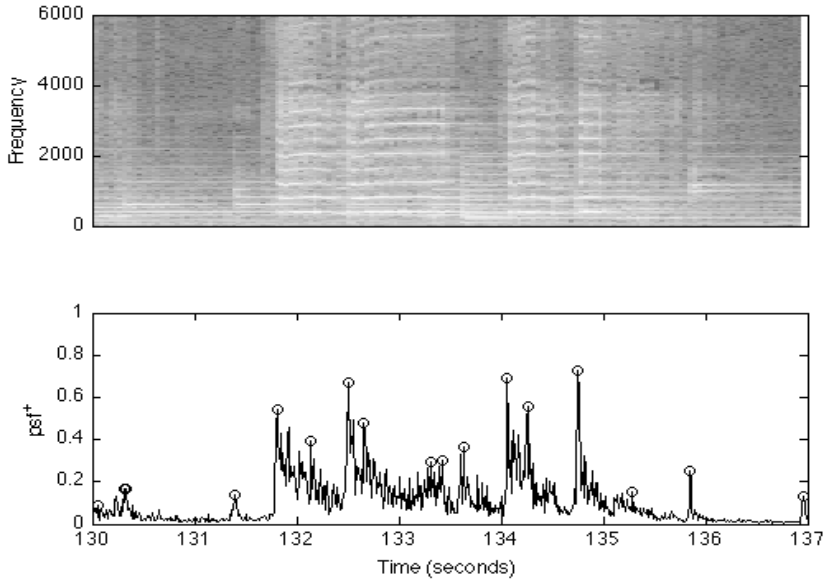

Figure 1. Spectralgram (top) and positive Perceptual Spectral Flux (bottom) for Stan Getz - First song (for Ruth) (excerpt). Found psf peaks are indicated with a ring.

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.

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 $E<0$. 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].

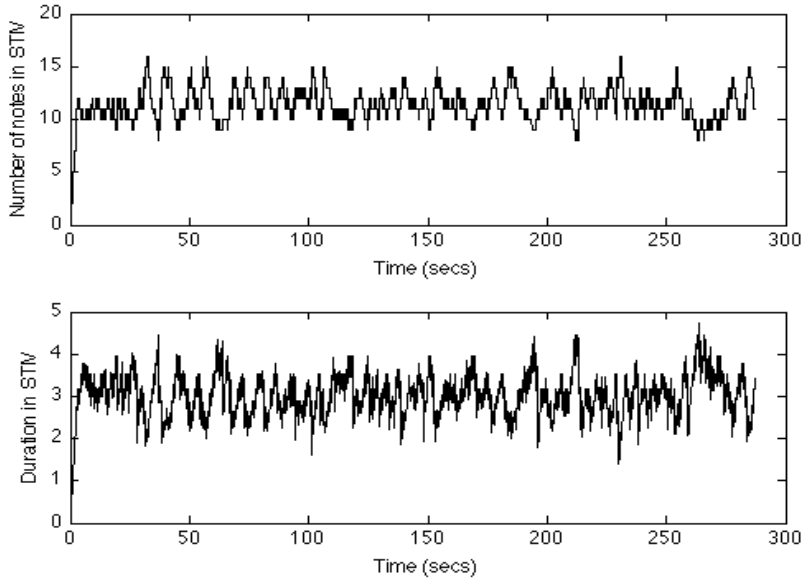

Figure 2. Number of elements in the STM (top) and duration of STM (bottom) for Stan Getz - First song (for Ruth).

4 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 [11]), 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 [12], as,

$$d_0 = a_1 a_2 \left(e^{-0.84 \frac{(f_2 - f_1)}{0.0207 f_1 + 18.96}} - e^{-1.38 \frac{(f_2 - f_1)}{0.0207 f_1 + 18.96}} \right), \quad \textbf{(7)}$$

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,

$$d^0 = \sum_k \sum_{l=k+1} d_0(f_k^0, a_k^0, f_l^0, a_l^0). \quad \textbf{(8)}$$

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,

$$d^0 = d^0 + \sum_n E^n \sum_k \sum_l d_0(f_k^0, a_k^0, f_l^n, a_l^n). \quad \textbf{(9)}$$

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. However, the total dissonance seems smoother than the instantaneous dissonance. This is verified in the standard deviation of the measures, where the std of the total dissonance is approximately five times lower than the instantaneous sensory dissonance.

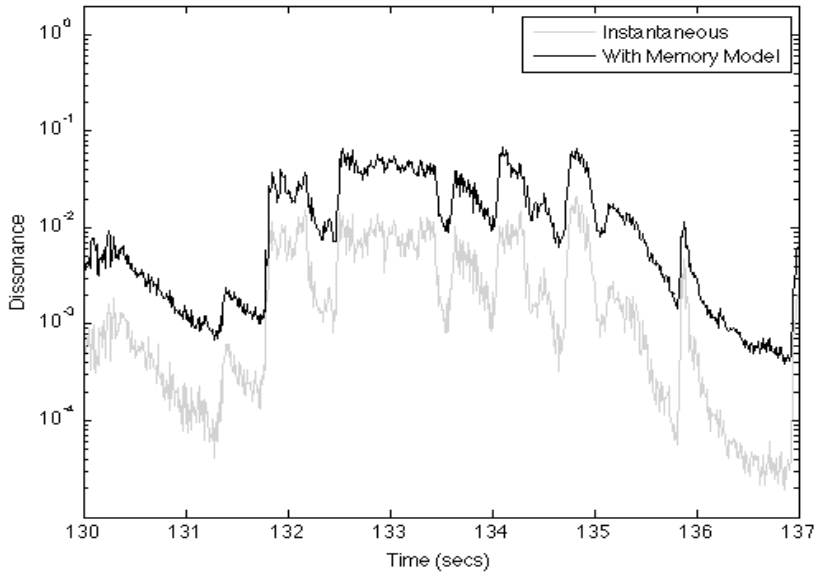

Figure 3. Instantaneous and total sensory dissonance for Stan Getz - First song (for Ruth) (excerpt).

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 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 a 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 one experiment involving dissonance measure. 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. Furthermore, it is smoother, with a significantly lower standard deviation.

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.

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. Anderson, J. R. & Lebiere, C. (1998). Atomic components of thought. Hillsdale, NJ: LEA.
9. Moore, B, C., J. (1997). *Psychology of Hearing*. Academy Press
10. Jensen, K., (2007). 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. 11 pages.
11. Plomp R. and W. J. M. Levelt (1965). Tonal Consonance and Critical Bandwidth. J. Acoust. Soc. Am. 38(4), 548-560.
12. Sethares, W. (1993). Local consonance and the relationship between timbre and scale. Journal of the Acoustical Society of America 94 (3): 1218–1228.