Effectiveness of Signal Segmentation for Music Content Representation

Namunu C. Maddage Mohan S. Kankanhalli* and Haizhou Li

Institute for Infocomm Research, Heng Mui Keng Terrace, 119613 Singapore

*School of Computing National University of Singapore
{maddage, hli}@i2r.a-star.edu.sg mohan@comp.nus.edu.sg

Abstract. In this paper we compare the effectiveness of rhythm based signal segmentation technique with the traditional fixed length segmentation for music contents representation. We consider vocal regions, instrumental regions and chords which represent the harmony as different class of music contents to be represented. The effectiveness of segmentation for music content representation is measured based on intra class feature stability, inter class high feature deviation and class modeling accuracy. Experimental results reveal music content representation is improved with rhythm based signal segmentation than with fixed length segmentation. With rhythm based segmentation, vocal and instrumental modeling accuracy and chord modeling accuracy are improved by 12% and 8% respectively.

Keywords: Music segmentation, chord detection, vocal and instrumental regions.

1 Introduction

The fundamental step for audio content analysis is the signal segmentation where, within the segment, information can fairly be considered as quasi-stationary. After that, the feature extraction and other advanced processing steps, such as music segmentation, can follow in music content modeling. Higher accuracies of the above mentioned steps lead better performances in semantic music information processing such as music information indexing, retrieval, lyrics identification, polyphonic music transcription and music summarization. Fixed length segmentation has commonly been used in speech processing [2]. In the past, music research community has also employed fixed length signal segmentation technique for music content analysis research [3] [8] [9] [11] [12] [14]. However, compared to speech, music is wideband signal (> 10 kHz), structured and heterogeneous source in nature. Given the fact that music and speech have differences in terms of production and perception, it's not clear how suitable the fixed length segmentation is for music information modeling. Our focus in this paper is to analyze two signal segmentation techniques (frame level): fixed length segmentation and rhythm level segmentation for their effectiveness in modeling music chords, vocal and instrumental region information.

In our literature survey we found that importance of time domain multi resolution signal analysis for harmonic structure detection has been discussed in [9]. For similarity analysis, timbre level signal segmentation with the understanding of music structure has been discussed in [8]. Information carried by music signals can conceptually be represented or grouped as sound, tone, melody, harmony, composition performance, listening, understanding and ecstasy [10]. Similarly underlying information in music: time, harmony, acoustic events and music semantics can conceptually be represented bottom up in a pyramid [5]. From the music composition point of view, smallest note played in music can be considered as an information measuring unit. Recent studies in [4] [5] [13] have also suggested tempo based signal segmentation for music content analysis. It can be seen that when research advances, more efforts have been devoted to both understand the behaviors of music signals and incorporate music knowledge for music content modeling.

In this paper we compare the effectiveness of both fixed length (30ms frames) and rhythm based music segmentations for music information representation. Music region contents (vocal and instrumental class information) and music chords are chosen as different music contents to be represented. We consider intra class lower average feature distance which implies the feature stability within the music class, inter class higher average feature distance which implies higher feature deviation between two different music classes and higher accuracy of class content modeling as the parameters to measure segmentation effectiveness.

In section 2 we briefly explain music composition. Our rhythm based segmentation method is explained in section 3. Section 4 details modeling and analysis of different music contents. Experimental results are discussed in section 5 and we conclude the paper in section 6.

2 Music Composition and Signal Visualization

Duration of the song is measured as number of *Bars*. The steady throb to which one could clap while listening to a song is called the *Beat* and the *Accents* are the beats which are stronger than the others. In a song, the words or syllables in the sentence fall on beats in order to construct music phrase [7]. Figure 1, shows the time alignment between music notes and the words. Since accents are placed over the important syllables, the time signature of this musical phrase is 2/4, i.e. two quarter notes per bar. Approximately 90% of sound generated during singing is voiced [3]. In the perfect singing, these voiced sounds are held longer to align with the duration of a music note. This can even be seen in Figure 1, where the time duration of the word ‘Jack’ is equal to a quarter note and the length of a quarter note is defined according to the tempo (measured as number of beats per minute).

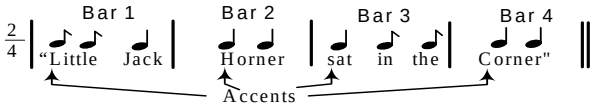


Figure 1: Rhythmic group of words

The chord knowledge has been applied to effectively detect the rhythm information [1], reveals that harmony changes are in discrete inter-beat time intervals. Common chord transitions are as follows.

- Chords are more likely to change on beat times than on other positions.
- Chords are more likely to change on half note times than on other positions of beat times.
- Chords are more likely to change at the beginning of the measures (bars) than at other positions of half note times.

Therefore, from the music composition point of view, progression of music chords (harmony event), and music regions i.e. pure instrumental, pure vocal, instrumental mixed vocal, and silence regions can be measured as integer multiples of music notes or inter-beat proportional units. Figure 2 depicts time domain and frequency domain visualization of music signal with their note alignments. Figure 2(middle) shows the normalized spectra differences. We can clearly see lower spectral difference within the music notes and higher spectral difference at the note boundaries. Therefore such time frequency visualization reveals quasi-stationary behavior of the temporal properties within music notes.

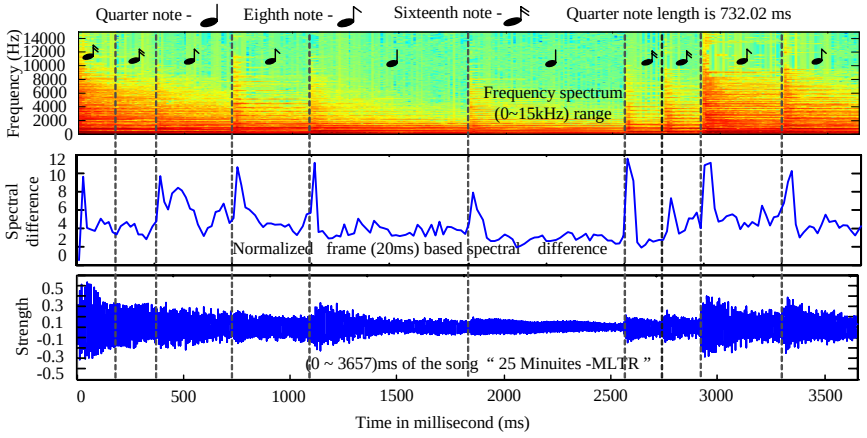


Figure 2: Note boundaries of a 3667ms long clip from the song "Paint My Love - MLTR". Quarter note length is 736.28ms

Thus there is a good reason for us to believe that inter-beat proportional segmentation is an effective method for music content modeling. In the next section we briefly discuss our inter-beat proportional segmentation technique which we further use for music modeling to compare the performances against fixed length signal segmentation.

3 Frame Level Music Segmentation

As mention earlier, we are interested in learning about the effectiveness of two signal segmentation techniques: fixed length segmentation and rhythm based segmentation for music content representation. In our previous work [4] [5], we detailed a rhythm extraction and an inter-beat proportional segmentation technique. In this rhythm based segmentation technique, we first decompose the signal into octave sub-band signals and detect onsets on each sub-band. Then final sub-band onsets are detected by taking weighted summation of sub-band onsets. For onset detection we followed similar approach in [14]. Figure 3 depicts the graphical user interface that we developed for onset detection. We then search equally spaced inter beat proportional intervals with the help of dynamic programming. Using this approach, we compute duration of smallest notes and then we segment the song into these inter-beat proportional signal frames. We called this segmentation as *Beat Space Segmentation*.

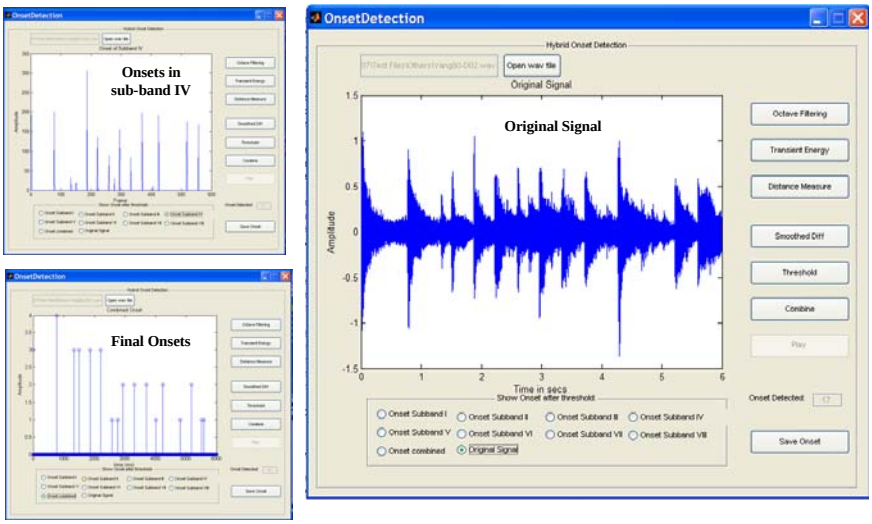
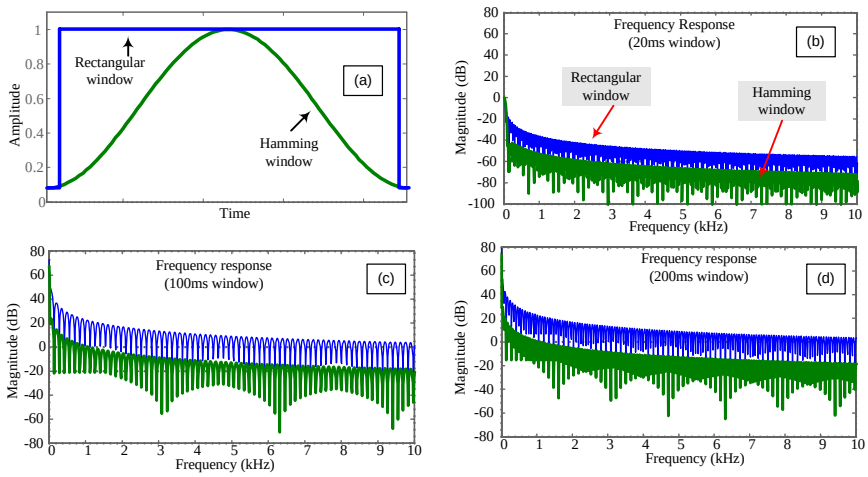


Figure 3: GUI for onset detection.

In speech processing, the sliding window technique is commonly used with two window type: hamming or rectangular. The window size varies from 20ms to 40ms. In Figure 4 we have shown both hamming and rectangular windows of 20ms, 100ms and 200ms durations in time and frequency domains. Both hamming and rectangular windows operate as low pass filters. Hamming window has very high stop band attenuation than rectangular filter. The bandwidth of hamming window is also higher than it is of a rectangular window. However, bandwidth decreases with the length of the window (see Table in Figure 4).

Unlike speech signals, music signals have wider bandwidth in nature. Therefore useful information in music spread beyond 10 kHz whereas in speech, useful information is well below 8 kHz. The bandwidths of both rectangular and hamming

windows are significantly smaller for music signal analysis. Compared to rectangular window, hamming window has sharp stop band attenuation and suppresses useful information in the higher frequencies nearly by 3 fold over the rectangular window. In view of this, the rectangular window is better for music analysis. Simple implication of using the rectangular window is that it analyzes the signal frame as it is. Thus rectangular window is considered in the feature extraction process.



	Rectangular window			Hamming window		
window size Hz	20	100	200	20	100	200
Mainlobe width (at -3dB) Hz	43	8.7	4.4	64.6	12.8	6.4
Relative sidelobe attenuation (dB)	-13.3	-13.3	-13.3	- 42.7	- 42.7	- 42.7

Figure 4: The frequency responses of hamming and rectangular windows when the window lengths are 20ms, 100ms and 200ms.

4 Feature Analysis

When the signal is quasi-stationary, the signal prediction and feature stability are improved [2]. From the statistical analysis point of view, the feature variance within the class (intra class) is very low. For different classes (inter class) feature variance is very high. In addition, extracted feature coefficients have high un-correlation when the signal section is quasi-stationary. This can be investigated by looking into the diagonal matrix at singular value decomposition (SVD). Diagonal matrix consists of singular values. Higher singular values imply high un-correlation among the coefficients of the feature vectors.

In order to analyze the features, first we extract the features from both fixed length (30ms) signal frames and proposed beat spaced signal frames using the data set described in section 5. Silence regions are removed from the music signals in the pre-processing.

4.1 Feature analysis for vocal instrumental boundary detection

It is found in our previous experiments that the spectral domain features are better for characterizing both the instrumental and vocal music [5]. Sung vocal lines always follow the instrumental line such that both pitch and harmonic structure variations are also in octave scale. Thus we used "Octave Scale" instead of "Mel Scale" to calculate Cepstral coefficients to represent the music content. These coefficients are called Octave Scale Cepstral Coefficients (OSCC) as detailed in [5]. For feature analysis, we extract 20 OSCCs from both 30ms frames and beat space length frames. The average distances (*Avg Dis*) between feature vectors are computed according to Eq. (1), where n, V_i, and L are number of feature vectors, i^{th} feature vector and dimension of the feature vector respectively.

$$Avg\ Dis = \frac{2}{n*(n-1)} \sum_{i=1}^{n-1} \sum_{j=i}^{n} \left[norm \left| V_i - V_j \right| / L \right] \qquad (1)$$

We model vocal and instrumental region using 64 Gaussian mixtures for each region class. We consider both pure vocal and instrumental mixed vocal regions as vocal region and pure instrumental region as instrumental region.

4.2 Feature analysis for chord detection

Chord detection is important to identify the harmony event in the music. Here we extract the pitch class profile (PCP) features which are highly sensitive to F0s (fundamental frequencies) of the music note and less sensitive to timbre of the source of the note [11]. Calculation of PCP features for each chord is similar to the method described in [4]. Then we compare the distance between feature vectors extracted from same chord using Eq. (1)

For chord modeling, we consider 48 chords, each of 12 music chords belong to Major, Minor, Diminish and Augmented chord types. Each single layer chord model [5] is consisted of 128 Gaussian mixtures.

5. Experiments

For the vocal and instrumental class feature analysis experiments (section 4.1), we use 120 popular English and Chinese songs sung by 12 artists; MLTR, Bryan Adams, Westlife, Shania Twain, Mariah Carey, Celine Dion, Huang Pingyuan, Wen Zheng, ADu, Liu Ruoying (Rane), Leung (Jasmine), and Li Qi . For music chord feature analysis, we constructed a chord data base which includes recorded chord samples from original instruments, synthetic instruments and chord extracted from English songs with the help of music sheets and listening tests. We have about 10 minutes of each chord sample spanning from C2 to B8.

All experimental data are sampled at 44.1 kHz with 16 bits per sample and mono format. Listening tests have been carried out to annotate vocal instrumental boundaries. Music signals are then framed into both beat space segments and fixed length segments to extract OSCC features. We carry out experiments; inter class and intra class feature distance measure, *Avg Dis(.)* in Eq. (1), to examine effectiveness of both beat space and fixed length segmentation for content representation. It can be seen in Figure 5, the average intra class feature distance of vocal and instrumental regions are lower when OSCC features computed from beat space segments than they are computed from 30ms segments. *Avg Dis(.)* is higher for inter class (Vocal - Inst) when features are extracted from beat space segmented frames.

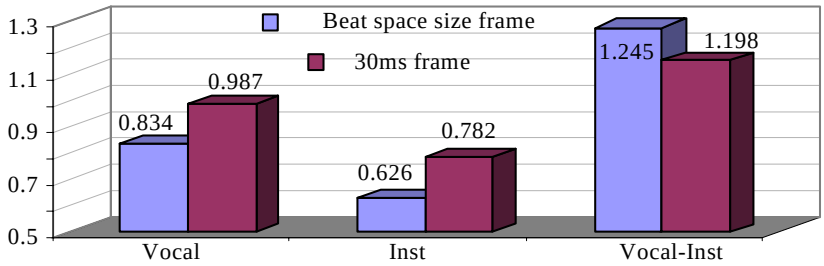


Figure 5: Distance between both inter and intra class feature vectors

Singular value decomposition (SVD) is then applied to vocal and instrumental feature sets to find the correlation between octave scale cepstral coefficients. The plot of singular values in the diagonal matrices of SVD for both vocal and inst class feature vectors is shown in Figure 6. Singular values are higher for features calculated from beat space frames than from fixed length frames. Thus we can conclude coefficients of features calculated from BSS are more de-correlated than coefficients of features calculated from fixed length.

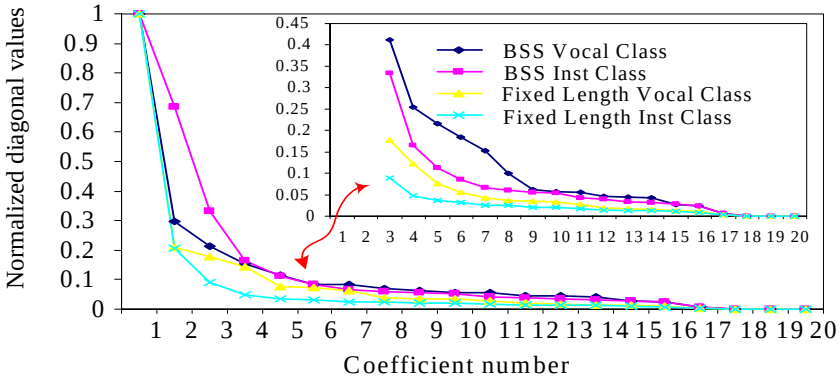


Figure 6: Singular values in diagonal matrix of SVD.

Then we model vocal and instrumental region information using OSCC features and 64 Gaussian mixtures. In this experiment all the songs of each artist are used by cross validation where 5 songs of each artist are trained and test at each turn. We then compare modeling effectiveness of vocal and instrumental classes with different frame sizes, beat space frames (X1.0), half beat space frame (X0.5) and fixed length frames (FIX Length). Figure 7 depicts the correct vocal and instrumental class feature classification results. Compared to the average vocal and instrumental classification accuracy with fixed length signal frames, we managed to achieve 12% and 7% higher average classification accuracies with X1.0 and X0.5 beat space segmentations respectively.

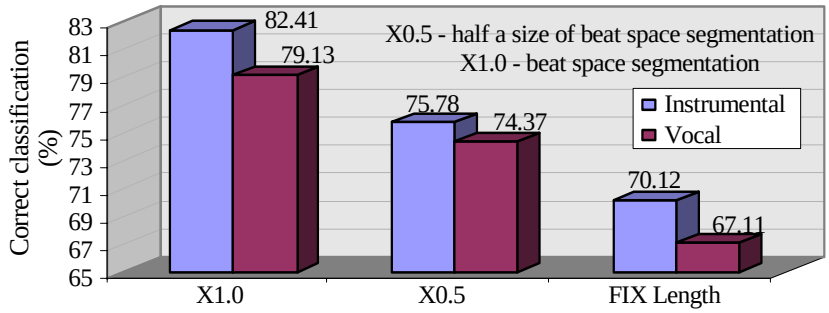


Figure 7: Effect of classification accuracy with frame size

Thus results of the above experiments; inter class and intra class feature distance measure, singular value analysis and vocal instrumental content modeling accuracies, imply higher improvement of feature stability with beat space proportional frame size than with fixed length signal frames. It can be concluded that within beat space segment, signal section can be considered more stationary than the signal section of fixed length.

Following experiments are focused on the analysis of signal frame size for music chord representation. Figure 8(a), details the intra chord class average feature distance, *Avg Dis(.).* It can be seen that *Avg Dis(.)* is lower when the frame size is beat space than it is fixed length (30ms). The average distance between chords (inter class) is 26.67% higher when feature vectors are extracted from beat space signal frames than they are extracted from fixed length frames. Figure 8(b) shows the average distance between chord C and other chords.

We also compare the correct chord detection accuracies when they are modeled with Gaussian mixtures. It's found that 75.28% and 67.13% frame based correct chord classification accuracies are obtained with beat space frames and fixed length frames respectively.

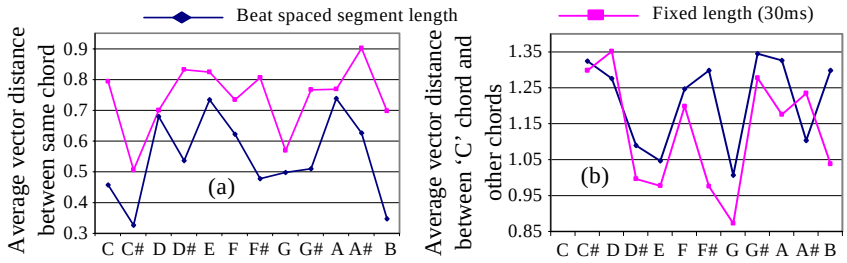


Figure 8 (a): Average feature vector distance between the same chords. **(b):** Average feature vector distance between 'C' chords and other chords.

6. Conclusion

In this paper we compare the effectiveness of both beat space segmentation and fixed length segmentation for music content representation and modeling. Music contents, vocal and instrumental region contents and music chords are selected for effective analysis.

For vocal and instrumental contents, it is found that intra class average feature distances are lower when features are extracted from beat space signal segments than from fixed length frames. However inter class average distance is higher for the features extracted from beat space frames than feature extracted from fixed length frames. Lower average distance in intra class feature vectors indicates the compactness of the feature vectors for a particular class. Higher distance in inter classes implies a higher separation of features which belong to different classes. Thus features extracted from beat space signal frames we are able to effectively represent vocal and instrumental contents. Singular values of SVD study has also indicated that features extracted from beat space segments are more uncorrelated than they are extracted from fixed length segments. Vocal and instrumental content modeling results highlight 12% accuracy improvement with beat space frames than with fixed length frames.

When we compare features extracted from beat space signal frames and fixed length signal frames for music chord representation and modeling, we also observed that more stable feature can be obtained with beat space signal frames than with fixed length signal frames. Average chord modeling accuracy has also been improved by 8% with beat space frames compared with fixed length frames.

It can be concluded that music contents; vocal and instrumental contents and chords can more accurately be modeled using rhythm level signal segmentation (beat space segmentation) than using fixed length signal segmentation. We will continue our research in this direction to improve the accuracies of music content representation and modeling.

Acknowledgement

Authors thank Mr. Ardy Salim for his support in the development of onset detection GUI.

References

1. M. Goto, "An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds," *Journal of new Music Research,* Vol.30, No.2, (June. 2001), 159-171.
2. R. D. John, H. L. John, and G. P. John, *Discrete-Time Processing of Speech Signals.* IEEE Press, 1999.
3. Y. E. Kim, *Singing Voice Analysis / Synthesis.* PhD. Thesis, Massachusetts institute of Technology, Sept 2003.
4. N. C. Maddage, C.S. Xu, M. S. Kankanhalli, X. Shao, "Content-based Music Structure Analysis with the Applications to Music Semantic Understanding," *In ACM Multimedia Conference*, New York, 2004.
5. N.C. Maddage, H. Li and M. S. Kankanhalli. "Music Structure based Vector Space Retrieval," *Proc. ACM SIGIR Conference,* August, 2006.
6. T.D. Rossing, F. R. Moore, and P. A. Wheeler, *Science of Sound.* Addison Wesley, 3rd edition 2001.
7. *Rudiments and Theory of Music.* (1949), The associated board of the royal schools of music, 14 Bedford Square, London, WC1B 3JG.
8. J.-J. Aucouturier and M. Sandler. "Finding Repeated Patterns in Acoustic Musical Signals: Applications for Audio Thumbnailing," *AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio,* Finland, 2002.
9. J.C. Brown. "Calculation of a Constant Q Spectral Transform," Journal of Acoustic Society of America. Vol. 89. No. 1, 1991.
10. R. Jourdain. Music, *The Brain, and Ecstasy: How Music Captures Our Imagination.* HarperCollins press, 1997.
11. T. Yoshioka et, al. "Automatic Chord Transcription with Concurrent Recognition of Chord Symbols and Boundaries," *In Proc. of 5th International Conference of Music Information Retrieval (ISMIR),* 2004.
12. T. L. New, and Y. Wang. "Automatic Detection of Vocal Segments in Popular Songs," *In Proc. of 5th International Conference of Music Information Retrieval (ISMIR),* 2004.
13. D.P.W. Ellis, and G. E. Poliner. "Identifying 'cover songs' with Chroma Features and Dynamic Programming Beat Tracking," In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* 2006.
14. C. Duxburg, M. Sandler, and M. Davies. "A Hybrid Approach to Musical Note Onset Detection," In *Proceedings of International Conference of Digital Audio Effects (DAFx).* Hamburg, Germany, Sept, 2002.