ENHANCING TIMBRE MODEL USING MFCC AND ITS TIME DERIVATIVES FOR MUSIC SIMILARITY ESTIMATION

Franz de Leon, Kirk Martinez

Electronics and Computer Science, University of Southampton
University Road, Southampton, United Kingdom SO17 1BJ
email: {fadl1d09, km}@ecs.soton.ac.uk

ABSTRACT

One of the popular methods for content-based music similarity estimation is to model timbre with MFCC as a single multivariate Gaussian with full covariance matrix, then use symmetric Kullback-Leibler divergence. From the field of speech recognition, we propose to use the same approach on the MFCCs' time derivatives to enhance the timbre model. The Gaussian models for the delta and acceleration coefficients are used to create their respective distance matrix. The distance matrices are then combined linearly to form a full distance matrix for music similarity estimation. In our experiments on two datasets, our novel approach performs better than using MFCC alone. Moreover, performing genre classification using *k*-NN showed that the accuracies obtained are already close to the state-of-the-art.

Index Terms— MFCC, music similarity estimation

1. INTRODUCTION

Digital technology and the Internet have changed the music industry landscape. The increased accessibility of music has allowed consumers to store and share thousands of files on their computer's hard disk, portable media player, mobile phone and other devices. Given the large music collections available, there is a need for new applications for browsing, organising, discovering as well as generating playlists for users. The research field of Music Information Retrieval (MIR) aims to address these challenges by using content-based techniques for performing tasks such as audio music similarity estimation and genre classification.

Generally, the essential music dimension used in content-based approaches is timbre. Timbre can be defined as "the character or quality of a musical sound or voice as distinct from its pitch and intensity" [1]. It depends on the perception of the quality of sounds, which is related to the used musical instruments, with possible audio effects, and to the playing techniques [2]. In the field of speech recognition, the mel-frequency cepstral coefficients (MFCCs) have been widely used to model important characteristics in speech [3]. Since modelling speech characteristics and timbre are similar, the use of MFCCs has been extended with success in the field of music similarity [4]. In this paper, we propose to enhance MFCCs' performance in audio similarity tasks by using its time derivatives (e.g. delta and acceleration coefficients).

The paper is organized as follows. The following section presents some related works. Section 3 details how the MFCCs, delta and acceleration coefficients are computed and modelled. In section 4, we describe how the derived features are combined and used for audio music similarity estimation. The performance is evaluated with varied parameters and the results are explained in Section 5. Finally, we summarize our findings in Section 6.

2. RELATED WORK

One of the standard approaches to compute music similarity is to estimate a single multivariate Gaussian model on the MFCC vectors. In this way, the closed form solutions of the Kullback-Leibler (KL) divergence can be used to compute the similarity between two music models. Besides being fast, this method has been proven to outperform other more complex music similarity approaches [5]. Our approach considers the time derivatives of the MFCC vectors to enhance music similarity estimation performance. The time derivatives add dynamic information to the static cepstral features [6]. In other studies, the delta and acceleration coefficients were appended to the static cepstral features resulting in a three-fold increase in dimension (e.g. [19MFCC:19Δ:19ΔΔ]) [7]. These set of features can be used for genre classifiers. However, this would be impractical for quantifying music similarity estimation since the KL divergence involves numerically sensitive operations and computationally intensive matrix inversion. Our novel approach simplifies the problem by creating separate models for the time derivatives. Similar to the standard approach to MFCCs, the time derivative vectors are summarized with a single multivariate Gaussian model. The same distance computation is performed for the resulting Gaussian models. The results are then combined with the original MFCC distances.

3. MODELLING TIMBRE

This section details the computation of MFCCs and its time derivatives to model timbre.

3.1. Mel-frequency cepstral coefficients

The timbre component is represented by the MFCCs [3]. The normalized audio signals signal is divided into frames with a window size and hop size of 512 samples (~23 msec.). The length of the segment ensures that the segmented signal is pseudo-stationary while the hop size keeps the continuity of the segments. Next, a window function (e.g. Hanning window) is applied to each segment. This is necessary to reduce spectral leakage. The following steps are then performed to each segment:

1. Calculate the power spectrum using FFT.
2. Transform the power spectrum to Mel-scale using a filter bank consisting of triangular filters.
3. Get the sum of the frequency contents of each band.
4. Take the logarithm of each sum.
5. Compute the discrete cosine transform (DCT) of the logarithms.

3.2. Delta and acceleration coefficients

The performance of a speech recognition system can be greatly enhanced by adding time derivatives to the basic static parameters [8]. Delta Coefficients are computed using the following formula:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \quad (1)$$

where d_t is a delta coefficient at time t, c is the cepstral coefficient, computed using a time window Θ. The same equation can be applied to the delta coefficients to obtain the acceleration coefficients. Figure 1 visualizes the derived features for an audio clip.

3.3. Summarizing audio features

The features derived from each audio track must be summarized efficiently and take into consideration the similarity computation method that will be performed. In this work, the Mel-frequency cepstral coefficients are computed for each time segment or *frame*. These features are aggregated using the bag-of-frames approach to model global statistics. The bag-of-frames approach is more appropriate in this case since the tasks that will be performed are less selective, e.g. music similarity estimation. Previous works model the spectral information with a single Gaussian distribution with a diagonal covariance matrix [9]. Other studies used Gaussian Mixture

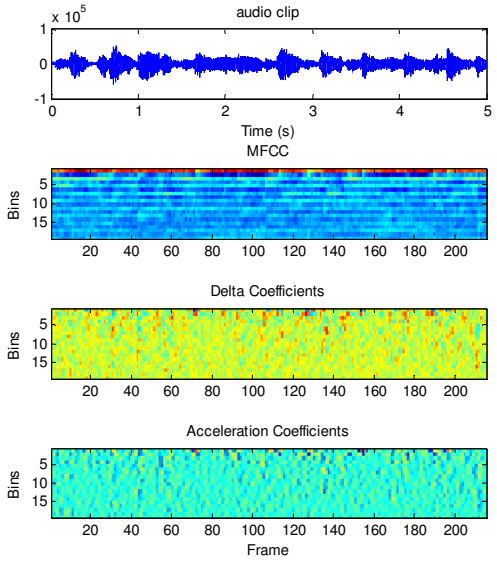

Fig. 1. MFCCs, delta and acceleration coefficients for a 5-second audio clip.

Models to model the distributions using the K-means algorithm and expectation-maximization algorithm [10], [11]. Subsequent works have shown that the same level of performance can be achieved using single Gaussian distribution with full covariance matrix [5], [12]. In this paper, the single Gaussian with full covariance approach is implemented to benefit from reduced computational complexity compared to the Gaussian Mixture Models. We extend this approach to delta and acceleration coefficients. Thus, each audio file is represented by three single multivariate Gaussian models; for MFCCs, delta and acceleration coefficients.

4. APPLICATION TO AUDIO MUSIC SIMILARITY ESTIMATION

In this section, we present a music similarity estimation method using the derived features. Each audio file is represented by three single multivariate Gaussian models. A single multivariate Gaussian probability density function is defined as:

$$\mathrm{N}(x \mid \mu, \Sigma) = \left(\frac{1}{2\pi}\right)^{n/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \quad (2)$$

where x is the observation (n-dimensional feature vector), μ is the mean, Σ is an n x n covariance matrix. Using a single Gaussian with full covariance matrix to model a music file, the similarity between two tracks can be computed using the Kullback-Leibler (KL) divergence. The KL divergence between two single Gaussians $p(x)=\mathrm{N}(x;\mu_p,\Sigma_p)$ and $q(x)=\mathrm{N}(x;\mu_q,\Sigma_q)$ is given by [13]:

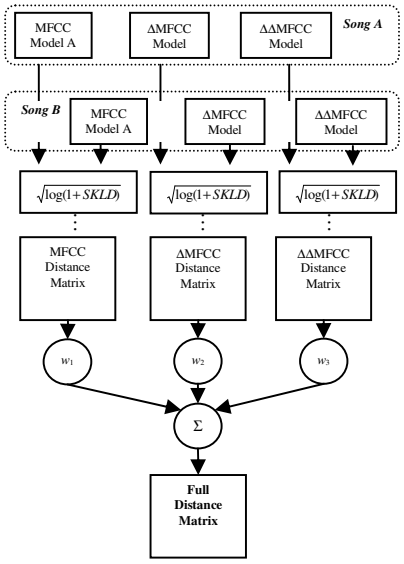

Fig. 2. Block diagram of our proposed music similarity estimation system.

$$2KL(p \| q) = \log\frac{|\Sigma_q|}{|\Sigma_p|} + Tr\left(\Sigma_q^{-1}\Sigma_p\right) + \\ + (\mu_p - \mu_q)^T \Sigma_q^{-1}(\mu_p - \mu_q) - d \quad (3)$$

where $|\Sigma|$ denotes the determinant of the matrix $|\Sigma|$, $Tr(\ )$ denotes the trace function of a matrix.

A common approach to compute acoustic timbre similarity is to use the symmetric version of the Kullback-Leibler Divergence (SKLD), defined between two single Gaussian distributions $x_1 \sim N(\mu_1,\Sigma_1)$ and $x_2 \sim N(\mu_2,\Sigma_2)$:

$$SKLD(x_1, x_2) = \frac{1}{2}KL(x_1, x_2) + \frac{1}{2}KL(x_2, x_1) \quad (4)$$

The main drawback of using these models is the fact that the SKLD does not hold the triangle inequality and consequently, is not a metric. It was shown that the transformation function $T{:}SKLD \rightarrow \{\log(1+SKLD)\}^{1/2}$ turns the symmetric Kullback-Leibler divergence into an exact metric when the statistical models compared are Gaussian [14].

To quantify music similarity based on the presented features, we compute pair wise similarities using the transformed symmetric KL divergence. This step produces three distance matrices. For each distance matrix, we apply distance space normalization [15]. Finally, the distance matrices are linearly combined into a full distance matrix. The weights of the linear combination must be optimized for the proposed system. For a given feature set, the weights are in the range of 0 to 1 with a step size of 0.1. The sum of the weights is 1. An intuitive approach is used instead of performing exhaustive comparisons by computing all possible permutations.

TABLE I
GENRE DISTRIBUTION OF TRACKS FOR THE ISMIR 2004 AND GTZAN DATASETS

ISMIR 2004	
Training Set	
Songs	729
Genres	classical (320), electronic (115), jazz_blues (26), metal_punk (45), rock_pop (101), world (122)
Full Set (Training & Development)	
Songs	1458
Genres	classical (640), electronic (229), jazz_blues (52), metal_punk (90), rock_pop (203), world (244)
GTZAN	
Songs	1000
Genres	country (100), rock (100), reggae (100), blues (100), disco (100), hiphop (100), jazz (100), pop (100), classical (100), metal (100)

5. EXPERIMENTS

5.1. Setup

Two datasets were used in our experiments. The first dataset is the training and development sets for the ISMIR 2004 genre classification contest [16]. Both the training and testing set are composed of tracks from six genres. The second dataset is the GTZAN genre collection [17]. The dataset consists of 1000 audio each 30 seconds long. It contains 10 genres, each represented by 100 tracks.

For each track, a 30-second clip was selected from the middle. For files that are less than 30 sec. long, the actual length was used. Each signal was normalized then divided into short overlapping segments (e.g. 23ms). The MFCCs were derived using 36 filter banks on Hanning-windowed segments. Twenty cepstral coefficients were obtained but only the last 19 were used. The delta and acceleration coefficients were then derived using Equation 1.

Objective statistics were derived from the full distance matrix. In music information retrieval, music similarity is taken in the context of genre, artist or album similarity. For our tests, the metric we used was the percentage of genre matches in the top 5, 10, 15 and 20 query (precision at R = 5, 10, 15, 20).

$$precision = \frac{|\{relevant\ items\} \cap \{retrieved\ items\}|}{|\{retrieved\ items\}|} \quad (5)$$

Artist filtering was applied for the ISMIR 2004 training set for comparison. This means that there is only one track per artist in the artist filtered dataset. The experiments were performed using different time windows for the delta coefficients, and weights for the individual distance matrices. Finally, we evaluated genre classification accuracy for the two datasets to compare the performance of combining MFCC with its time derivatives.

5.2. Results

In Table II, we tabulate the *precision* after returning *R* items using the optimum weights. Note that these *precisions* are presented as the average across all the genres for the particular dataset. Moreover, only the best combinations are presented. Based on the results, there is a significant improvement in the precision in using MFCC in conjunction with the delta coefficients than using MFCC alone. This proves the importance of time derivatives as the static MFCCs alone don't have temporal information. On the average, the best precisions were obtained when the delta coefficients were given more weight than MFCC (w_1=0.4, w_2=0.6). However, this does not imply that delta coefficients could replace MFCCs to model timbre. For example, using delta coefficients alone on GTZAN dataset we obtained *5-precision* of 0.7308; while for MFCC alone, 0.7448.

There were no significant improvements or degradation in the precisions using three features (0.9MFCC*+0.1ΔΔMFCC) as compared to using only two (MFCC*≅0.4MFCC+0.6ΔMFCC). This means that the acceleration coefficients can be disregarded in modeling timbre. Thus, the experiments showed that a good model for timbre involves the MFCCs and its delta coefficients. The experiments also determined if the window size used to compute the time derivatives can affect the system's performance. Table II shows inconsistency in the results. However in most cases, using Θ=3 frames is better than Θ=5 frames.

Using the ISMIR 2004 dataset, we compared the precision with and without artist filtering. With artist filtering, the number of returned items was limited to 5 since there are only 5 tracks in the *jazz_blues* genre. Without artist filtering, the precision is around 0.73; as compared to with artist filtering that resulted to around 0.55. There is a difference in the performance of around 0.20 which is already consistent with other studies that used the same dataset [18].

Our final experiments used *k*-nearest neighbors to perform genre classification. This method, while being simple, is already established for performing music similarity measures [19][15]. We want to determine the improvement in the classification accuracy using our proposed timbre model. Figure 3 shows that the combination of MFCC and its time derivatives consistently perform better than MFCC alone. For the ISMIR 2004 and GTZAN datasets, the best accuracies at *k*=1 are 0.811 and 0.816 respectively (0.4MFCC+0.6ΔMFCC). As previously observed, there is no significant difference in the accuracies using three features. Using only these simple features that model timbre, it is worthy to note that the performance of the system is not far from the state-of-the-art. Thus, the new timbre model can serve as a foundation that can be enhanced by other features (e.g. fluctuation patterns [20], onset patterns [19], tempo [21]).

TABLE II
R-PRECISION USING MFCC, DELTA AND ACCELERATION COEFFICIENTS

Collection	Features	Artist Filter	Returned Items, *R*			
			5	10	15	20
ISMIR2004 Train, Θ=3)	MFCC	no	0.7006	0.6088	0.5538	0.5175
	0.4MFCC+0.6ΔMFCC	no	0.7296	0.6414	0.5776	0.5411
	0.9MFCC*+0.1ΔΔMFCC	no	0.7309	0.6373	0.5769	0.5418
	MFCC	yes	0.5466	na	na	na
	0.4MFCC+0.6ΔMFCC	yes	0.5509	na	na	na
	0.9MFCC*+0.1ΔΔMFCC	yes	0.5375	na	na	na
ISMIR2004 Train, Θ=5)	0.4MFCC+0.6ΔMFCC	no	0.7323	0.6414	0.5741	0.5347
	0.9MFCC*+0.1ΔΔMFCC	no	0.7403	0.6377	0.5742	0.5317
	0.4MFCC+0.6ΔMFCC	yes	0.5543	na	na	na
	0.9MFCC*+0.1ΔΔMFCC	yes	0.5507	na	na	na
ISMIR2004 Full, (Θ=3)	MFCC	no	0.7633	0.6857	0.6450	0.6138
	0.4MFCC+0.6ΔMFCC	no	0.7939	0.7178	0.6780	0.6463
	0.9MFCC*+0.1ΔΔMFCC	no	0.7923	0.7187	0.6778	0.6449
ISMIR2004 Full, (Θ=5)	0.4MFCC+0.6ΔMFCC	no	0.7990	0.7258	0.6797	0.6462
	0.9MFCC*+0.1ΔΔMFCC	no	0.7985	0.7246	0.6797	0.6459
GTZAN (Θ=3)	MFCC	no	0.7448	0.6455	0.5871	0.5454
	0.4MFCC+0.6ΔMFCC	no	0.7932	0.7034	0.6495	0.6101
	0.9MFCC*+0.1ΔΔMFCC	no	0.7930	0.7066	0.6519	0.6111
GTZAN (Θ=5)	0.4MFCC+0.6ΔMFCC	no	0.7852	0.6984	0.6432	0.6021
	0.9MFCC*+0.1ΔΔMFCC	no	0.7866	0.6958	0.6409	0.6034

In terms of computational complexity, there is not much overhead added in computing the symmetric Kullback-Leibler divergence twice (e.g. MFCC and delta coefficients). The system works on *stored* matrices such as the covariance and inverse covariance matrices.

The results show that using timbre models is important for content-based music similarity estimation. Timbre may contain salient information that roughly describes music genre. Research had shown that humans have the ability to distinguish and classify music after listening to short clips of audio. This implies the viability of using timbre as this feature can be easily extracted from short clips. The major limitation is that humans do not compute a weighted sum of similarities with respect to different aspects of music. In fact, the concept of audio similarity is subjective to listeners and a single aspect which is similar can be considered to judge similarity. Nevertheless, the computational model presented in this paper hopes to contribute on the improvement of content-based systems.

6. CONCLUSION

We have investigated a method for enhancing MFCC features for music similarity estimation using its time derivatives. Our novel approach applies the standard single multivariate Gaussian with full covariance matrix to model MFCCs' delta and acceleration coefficients. Using the Kullback-Leibler divergence to calculate music similarity, experiments have shown a consistent improvement in the performance in using the delta coefficients in conjunction with the MFCCs. Performing genre classification using *k*-NN showed that the accuracies obtained are already close to the state-of-the-art. In addition, recent studies have proved that our approach has a potential to be applied on larger databases [14][22].

We will further investigate our method and determine how other low-level features can be used to improve these initial results. We will also explore alternative ways of integrating these low-level features with our timbre model.

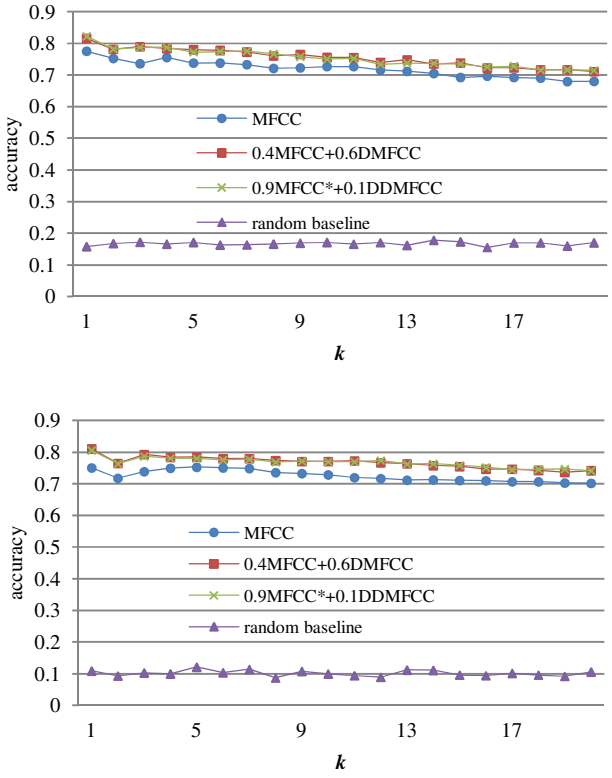

Fig. 3. Genre classification accuracy for ISMIR 2004 (top) and GTZAN (bottom) datasets.

7. ACKNOWLEDGMENTS

Mr. Franz de Leon is supported by the Engineering Research and Development for Technology Faculty Development Program of the University of the Philippines and DOST.

8. REFERENCES

[1] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Academic Press, 2003, p. 413.

[2] N. Orio, "Music Retrieval: A Tutorial and Review," *Foundations and Trends in Information Retrieval*, vol. 1, no. 1, pp. 1-96, 2006.

[3] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, Aug. 1980.

[4] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-Based Music Information Retrieval: Current Directions and Future Challenges," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668-696, Apr. 2008.

[5] M. I. Mandel and D. P. W. Ellis, "Song-level Features and Support Vector Machines for Music Classification," in *Submission to MIREX 2005*, 2005, pp. 594-599.

[6] S. Furui, "Speaker-independent isolated word recognition based on emphasized spectral dynamics," in *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1991, vol. 11, pp. 1991-1994.

[7] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic Annotation and Retrieval of Music and Sound Effects," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467-476, Feb. 2008.

[8] J.-julien Aucouturier and F. Pachet, "Improving Timbre Similarity : How high's the sky?," *J. Negative Results Speech Audio Sci.*, vol. 1, 2004.

[9] G. Tzanetakis, G. Essl, and P. Cook, "Automatic Musical Genre Classification Of Audio Signals," in *Proceedings of ISMIR 2001: The International Conference on Music Information Retrieval and Related Activities*, 2001.

[10] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.*, 2001, pp. 745-748.

[11] J.-julien Aucouturier and F. Pachet, "Music Similarity Measures : What's the Use?," in *3rd International Conference on Music Information Retrieval*, 2002, pp. 157-163.

[12] E. Pampalk, "Audio-Based Music Similarity and Retrieval : Combining a Spectral Similarity Model with Information Extracted from Fluctuation Patterns," in *Submission to MIREX 2006*, 2006.

[13] W. Penny, "Kullback-Liebler Divergences of Normal, Gamma, Dirichlet and Wishart Densities," 2001.

[14] C. Charbuillet, G. Peeters, S. Barton, and V. Gouet-Brunet, "A fast algorithm for music search by similarity in large databases based on modified Symetrized Kullback Leibler Divergence," in *2010 International Workshop on Content Based Multimedia Indexing (CBMI)*, 2010, pp. 1-6.

[15] K. Seyerlehner, M. Schedl, T. Pohle, and P. Knees, "Using Block-Level Features for Genre Classification , Tag Classification and Music Similarity Estimation," in *Submission to Audio Music Similarity and Retrieval Task of MIREX 2010*, 2010.

[16] "ISMIR 2004 Audio Description Contest-Genre/Artist ID Classification and Artist Similarity." [Online]. Available: http://ismir2004.ismir.net/genre_contest/index.htm.

[17] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, Jul. 2002.

[18] J. H. Jensen, M. G. Christensen, M. N. Murthi, and S. H. Jensen, "Evaluation of MFCC Estimation Techniques for Music Similarity," in *Proceedings of the 14th European Signal Processing Conference*, 2006, no. Eusipco.

[19] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer, "On rhythm and general music similarity," in *Submission to Audio Music Similarity and Retrieval Task of MIREX 2009*, 2009, no. Ismir, pp. 525-530.

[20] E. Pampalk, "Computational Models of Music Similarity and their Application in Music Information Retrieval," Vienna University of Technology, 2006.

[21] F. D. Leon and K. Martinez, "Submission to MIREX 2011 Genre Classification and and Audio Similarity Tasks," in *Submission to Audio Music Similarity and Retrieval Task of MIREX 2011*, 2011.

[22] D. Schnitzer, A. Flexer, and G. Widmer, "A Filter-and-Refine Method for Fast Similarity Search in Millions of Tracks," in *ISMIR 2009*, 2009, no. April, pp. 537-542.