# ACOUSTIC SCENE CLASSIFICATION:<br>AN OVERVIEW OF DCASE 2017 CHALLENGE ENTRIES

*Annamaria Mesaros, Toni Heittola, Tuomas Virtanen*

Laboratory of Signal Processing
Tampere University of Technology
PO Box 527, FI-33101 Tampere, FINLAND

## ABSTRACT

We present an overview of the challenge entries for the Acoustic Scene Classification task of DCASE 2017 Challenge. Being the most popular task of the challenge, acoustic scene classification entries provide a wide variety of approaches for comparison, with a wide performance gap from top to bottom. Analysis of the submissions confirms once more the popularity of deep-learning approaches and mel-frequency representations. Statistical analysis indicates that the top ranked system performed significantly better than the others, and that combinations of top systems are capable of reaching close to perfect performance on the given data.

***Index Terms***— acoustic scene classification, audio classification, DCASE challenge

## 1. INTRODUCTION

Acoustic scene classification is one major topic within the area of environmental sound classification and detection, as a generic classification problem setting the foundation for context awareness in devices, robots and many other applications. Partly, its popularity within the last few years is due to the international evaluation challenge on Detection and Classification of Acoustic Scenes and Events (DCASE), the task being present in each edition. The setup for acoustic scene classification in DCASE Challenge is as a supervised, multi-class, closed-set classification problem, representing therefore an entry level task that attracts new researchers to the field.

The problem of acoustic scene classification is not really novel, but it has been brought back to the spotlight within the last decade. During this time, machine learning approaches used to solve the problem have changed dramatically, with deep learning being currently the most popular. Plenty of work has been done before deep learning, using classical statistical models like Gaussian mixture models (GMMs) [1], hidden Markov models (HMMs) [2], and support vector machines (SVMs) [3]. Often the acoustic features used for representation were mel-frequency cepstral coefficients (MFCC), as they provide a compact and easy to calculate representation of the coarse spectrum of a signal, and have repeatedly proven to be successful in diverse audio classification problems including speech and speaker recognition, singer and instrument classification, and many others. Other low level spectral features used for acoustic scene classification include for example zero crossing rate, spectral centroid, spectral rolloff, spectral flux, and linear prediction coefficients [2].

Within DCASE Challenge, acoustic scene classification was a popular task from the beginning, with the highest number of participants in each of the three past editions. The development datasets used for it have gradually increased in size, from a modest dataset containing 10 scene classes each with 10 examples of 30 s in DCASE 2013 [4] to 15 scene classes each with 78 examples of 30 s in DCASE 2016 [5], to 15 scene classes each with 312 examples of 10 s in DCASE 2017 [6]. Given the higher amount of data available, the 2016 edition marks a clear transition to deep learning methods, with 22 of 48 submissions using some form of deep learning. Top performance systems were either ensemble classifiers [7, 8], or deep learning classification methods, in particular CNNs [9, 10], with the exception of one NMF-based approach that ranked second [11].

DCASE 2017 was the third edition of the challenge, and as such the third time an acoustic scene classification task was organized. The task was made more difficult by using 10 s audio segments, much shorter than the 30 s length used in the previous editions. In addition, a newly recorded evaluation dataset was used, creating an unexpected mismatch with the development data.

This paper presents an overview of the systems submitted to DCASE 2017 task 1, with statistical analysis including confidence intervals and comparison of classifiers using McNemar's test [12]. Combinations of submitted systems are also evaluated for a complete characterization of the problem and the systems' behavior. After this introduction, we continue by presenting shortly the task description in Section 2, including the dataset and provided baseline system. Section 3 presents the challenge results, an analysis of the submitted systems and the statistical analysis of their performance.

This work was supported by the European Research Council under the ERC Grant Agreement 637422 EVERYSOUND

978-1-5386-8151-0/18/$31.00 ©2018 IEEE

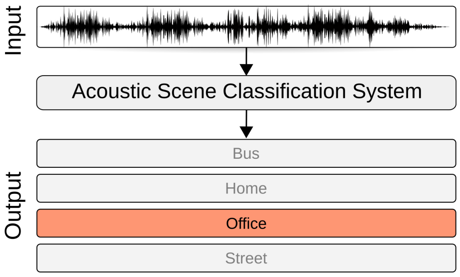


**Fig. 1**. Acoustic scene classification example

Finally, Section 4 presents conclusions and a preview of future directions for the acoustic scene classification task within DCASE Challenge.

## 2. TASK DESCRIPTION

The task of acoustic scene classification was set up as a straightforward multi-class supervised classification problem, with class labels describing the acoustic scene. Labeled audio examples were provided for training the systems, with each audio example having a single label. For each test example, a system was expected to provide a label from the set of known labels, as illustrated in Fig. 1.

### 2.1. Dataset

The task used the TUT Acoustic Scenes 2017 dataset, containing audio recorded in 15 different acoustic scenes; 3-5 minutes of audio was recorded in various locations for the following acoustic scenes: bus, cafe/restaurant, car, city center, forest path, grocery store, home, lakeside beach, library, metro station, office, residential area, train, tram, and park.

The development dataset has the same content as the complete TUT Acoustic Scenes 2016 dataset, but with original recordings being split into 10 s segments. The short audio segments provide less information for the decision making process in classification, thus increasing the task difficulty from the previous edition. This length is regarded as challenging for both human and machine recognition, based on the study in [2]. The development dataset contains 312 segments of 10 s per scene class (52 minutes). The evaluation dataset was recorded in similar locations approximately one year later than the development data, and contained 108 segments of 10 s per scene class (18 minutes). A detailed description of the data recording and annotation procedure is available in [13], while a more detailed description of the TUT Acoustic Scenes 2017 dataset can be found in [6].

### 2.2. Baseline system

The baseline system provided for this task uses a multilayer perceptron architecture (MLP) trained on log mel-band energies calculated in 40 ms frames with a 50% overlap and 40 mel bands. A 5-frame context was used, resulting in a feature vector length of 200. The MLP had two dense layers of 50 hidden units each, with 20% dropout, and an output layer of 15 softmax type neurons. Frame-based decisions from the network output were combined by majority voting to obtain the final class decision for one 10 s long test audio segment.

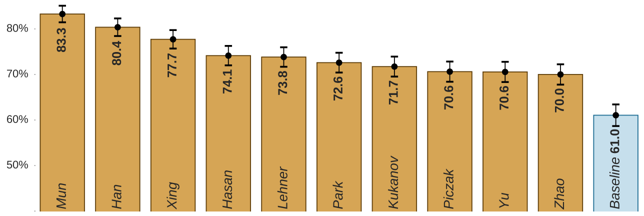


**Fig. 2**. Acoustic scene classification task accuracies based on the evaluation set with 95% confidence intervals; top systems, selected one per participating team.

The classification accuracy obtained by the system on the development set using the provided cross-validation setup is 73.8%, with class-wise performance ranging from 57% to 99.7%. Performance on the evaluation dataset is 61%. A detailed description of the baseline system and its class-wise performance can be found in [6].

## 3. CHALLENGE RESULTS

### 3.1. Submission statistics and ranking

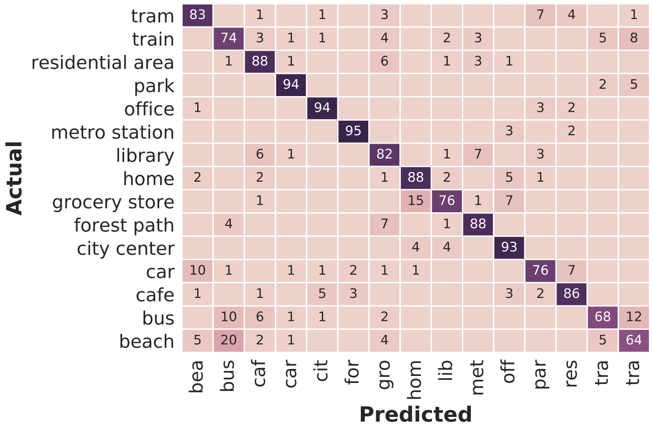
A number of 97 systems were submitted for this task, corresponding to 39 teams and 129 authors. The number of participating teams is similar to previous edition (34 teams in 2016), but the number of submissions was much higher because each team was allowed to submit a maximum of 4 systems, even though not all of them did so. Most of the submitted systems outperformed the baseline system. A selection of top systems performance and 95% confidence interval is presented in Fig. 2. Confidence intervals were calculated as a binomial proportion confidence interval for the classification output being correct or incorrect with respect to the ground truth. Based on Fig. 2, it can be seen that the confidence intervals for systems of neighboring ranks overlap significantly.

### 3.2. Submissions analysis

A general analysis of the characteristics of the submitted systems reveals that the most popular classification approach was the convolutional neural network, with 55 of the 97 submission being based on a CNN architecture. In some cases the CNN was used as part of an ensemble, combined with a variety of techniques such as multilayer perceptron (MLP), recurrent neural networks (RNN), support vector machines (SVM), Gaussian mixture mdels (GMM), and i-vector. Recurrent network architectures were part of 18 systems, some being convolutional (CRNN), others LSTM and bi-LSTM. The CNNs

**Table 1**. Selected top ranked systems.

| Rank | System | Features | Classifier | Acc (95% CI) |
|---|---|---|---|---|
| 1 | **Mun** | log-mel energies, spectrogram | MLP, RNN, CNN, SVM | **83.3** (81.5 - 85.1) |
| 2 | **Han** | log-mel energies | CNN, ensemble | **80.4** (78.4 - 82.3) |
| 6 | **Xing** | spectrogram, CQT | CNN | **77.7** (75.7 - 79.7) |
| 8 | **Hasan** | MFCC, log-mel energies | GMM & CNN, ensemble | **74.1** (72.0 - 76.3) |
| 9 | **Lehner** | mel-scaled spectrograms, i-vectors | i-vector, CNN, ensemble | **73.8** (71.7 - 76.0) |
| 10 | **Park** | gammachirp energies | CNN | **72.6** (70.4 - 74.8) |
| 13 | **Kukanov** | log-mel energies | CRNN | **71.7** (69.5 - 73.9) |
| 14 | **Piczak** | spectrogram | CNN | **70.6** (68.4 - 72.8) |
| 14 | **Yu** | mel-filterbank features | MLP, ensemble | **70.6** (68.3 - 72.8) |
| 15 | **Zhao** | log-mel spectrogram | CNN | **70.0** (67.8 - 72.2) |
| 16 | **Bisot** | CQT | NMF, MLP | **69.8** (67.6 - 72.1) |



| Actual | bea | bus | caf | car | cit | for | gro | hom | lib | met | off | par | res | tra | tra |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| tram | 83 |  | 1 |  | 1 |  | 3 |  |  |  |  | 7 | 4 |  | 1 |
| train |  | 74 | 3 | 1 | 1 |  | 4 |  | 2 | 3 |  |  |  | 5 | 8 |
| residential area |  | 1 | 88 | 1 |  |  | 6 |  | 1 | 3 | 1 |  |  |  |  |
| park |  |  |  | 94 |  |  |  |  |  |  |  |  |  | 2 | 5 |
| office | 1 |  |  |  | 94 |  |  |  |  |  |  | 3 | 2 |  |  |
| metro station |  |  |  |  |  | 95 |  |  |  |  | 3 |  | 2 |  |  |
| library |  |  | 6 | 1 |  |  | 82 |  | 1 | 7 |  | 3 |  |  |  |
| home | 2 |  | 2 |  |  |  | 1 | 88 | 2 |  | 5 | 1 |  |  |  |
| grocery store |  |  | 1 |  |  |  |  | 15 | 76 | 1 | 7 |  |  |  |  |
| forest path |  | 4 |  |  |  |  | 7 |  | 1 | 88 |  |  |  |  |  |
| city center |  |  |  |  |  |  |  | 4 | 4 |  | 93 |  |  |  |  |
| car | 10 | 1 |  | 1 | 1 | 2 | 1 | 1 |  |  |  | 76 | 7 |  |  |
| cafe | 1 |  | 1 |  | 5 | 3 |  |  |  |  | 3 | 2 | 86 |  |  |
| bus |  | 10 | 6 | 1 | 1 |  | 2 |  |  |  |  |  |  | 68 | 12 |
| beach | 5 | 20 | 2 | 1 |  |  | 4 |  |  |  |  |  |  | 5 | 64 |

**Fig. 3**. Confusion matrix for top ranked system [14]

are used in acoustic scene and generally in audio classification as a form of image processing, with their connectivity patterns exploiting regions in the time-frequency representations of signals, therefore being capable of capturing both time and frequency evolution of signals. On the other hand, RNNs are much better at capturing the long-term temporal characteristics, with the LSTM variants having very good internal memory capabilities for processing of time-series. Also MLP and SVM were popular choices, with 11 systems each, most often as part of an ensemble of classifiers. All systems in top 10 make use of CNNs in some way, while first non-CNN-based, ranked 14 and 16, use MLP. Table 1 presents a selection of top systems and their characteristics, while Fig. 3 shows the confusion matrix of the top performing system.

Most submissions were based on mel-scale representations, with log mel energies and MFCCs being used in 27 and 19 systems, respectively. Mel-scale representations are often used and generally work well in sound classification problems, their modeling of human perception making them a comfortable choice when no better assumptions on the data can be made. Other spectral representation include spectrogram and CQT [15], [16] with CQT probably made popular by previous edition runner-up system. CQT is often used in music analysis for its exponential frequency resolution and for preserving the relative positions of harmonics, but its use for environmental sound analysis is not as clearly motivated. While in 2016 CQT was used in three systems, this time there were 13, of which 9 relied solely on CQT, and others used it in combination with spectrogram or MFCC. There was also one system based on low-level features that included spectral centroid, rolloff, zero-crossing rate and MFCCs and their derivatives, ranked only 54, at same level with the baseline.

Many participants made use of binaural audio, with one third using the two channels separately instead of the averaged audio provided as example in the baseline system. This was mostly used as a way to obtain more data for the deep-learning methods, with the different channels having slight variations in the captured audio. Another new element was the use of specific data augmentation techniques, unnoticed in 2016: there was much use of block mixing, pitch shifting, time stretching, mixing files of the same class, and adding Gaussian noise, in some cases all the techniques being used in the same system. A novel and unique method in the challenge was the augmentation of the dataset using generative adversarial networks (GAN), by the system that also achieved the best performance [14]. All data augmentation techniques are motivated by the use of deep learning, for creating more data and adding more acoustic variability to allow better learning and generalization.

A comparison of systems performance on the development and evaluation datasets reveals that most systems have a significant drop in performance for the evaluation dataset (10-20% in term of absolute accuracy). This is likely due to the mismatch in the data recording conditions, as the evaluation data was recorded one year later at similar or, in some cases, same locations. The situation was not intentional, being just a consequence of extending the previously available data with a new evaluation dataset, but it reveals the ease with which neural-network based systems overfit the data. As an observation, augmenting the dataset using GAN seemed to offer a more consistent performance in conjunction with the deep-learning methods, the corresponding system having only a 4% absolute drop in accuracy between development and evaluation sets. The Pearson correlation coefficient calculated between the development and evaluation performance for all systems is 0.42, which can be considered a medium strength of association between the two. This suggests that the performance of systems is somewhat consistent, and the gap in performance is due to data mismatch and not lack of generalization properties in the systems. Considering only the best system of each team, the correlation between the development and evaluation performance is 0.69, indicating very strong correlation. Based on this, we can assert that each team has produced at least one system that generalizes well for unseen data.

### 3.3. Statistical analysis of systems performance

The confidence intervals presented in Table 1 show that there is not a significant difference between performance of closely ranked systems, with only the top system being set apart from the others. To understand how much the different systems take similar or different decisions, the systems were compared in pairs using McNemar's test [12]. McNemar's test for comparing classifiers examines only the cases in which the prediction of one system is correct, and the prediction of the second system is wrong, therefore identifying if there is a difference in systems with respect to the test samples that are more difficult to classify. For systems with similar accuracy, this test indicates if the difference is statistically significant.

The null hypothesis for the statistical test is that the two classifiers being compared perform similarly, while the alternative hypothesis is that the difference is statistically significant. Figure 4 illustrates the results of this test using a significance level of 0.05. A red square in the illustration indicates a pair of systems for which the result does not allow rejecting the null hypothesis. For this comparison we considered only the best system of each team, plus the baseline, with the systems considered in order of their accuracy (team rank order).

As expected, we notice that many systems on neighboring ranks perform equivalently, with the indicators aligned close to the diagonal. The top four compared systems show statistically significant differences, while already between the fourth and the fifth the difference is not statistically significant. These are the same systems presented in Table 1, ranked 1, 2, 6, 8, and 9, with accuracies of 83.3%, 80.4%, 77.7%, 74.1%, and 73.8%, respectively. The second to fifth ranked submissions all belong to the same authors [17] and have accuracies from 80.4% to 79.6%, being based on the same method with very slight variations, with no significant difference detected using McNemar's test.

Using the information that the first three systems in our comparison are significantly different, we calculate the performance of their combined outputs with a majority vote rule. The obtained performance is only 84.69%, which is not much higher than the 83.3% accuracy of the top system, meaning that in many cases two of the three systems still mis-classify the data. If we investigate the best case scenario between the three systems, by considering a correct item if *at least one* of the systems has classified it correctly, we obtain an accuracy of 96.05% - this indicates that most test items are indeed correctly classified by at least one of the three considered systems, and the possibility of improving performance by classifiers fusion exists, if suitable rules for fusion can be found. The average performance of all 97 systems is 64.33%, while a majority vote fusion of all systems obtains a performance of 73.52%. We contrast this with the human performance obtained on similar data [18], in which average human performance was 54.4% (87 participants), with participants from Finland, familiar with the recorded soundscape, scoring a better accuracy of 60%.

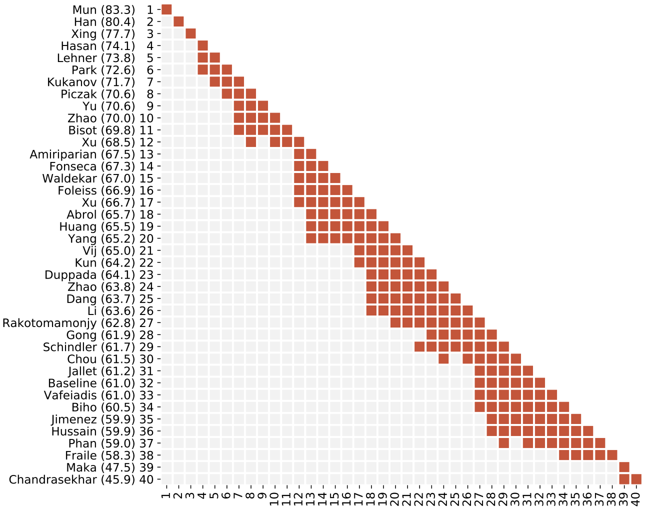


**Fig. 4**. Output of McNemar's test comparing classifiers; red squares mark the pairs for which the null hypothesis that classifiers perform similarly could not be rejected

## 4. CONCLUSIONS

The topic of acoustic scene classification attracts a lot of interest within the DCASE challenge, and this provides an interesting perspective on the current trends for its solutions. Each year, a large number of submissions are available for comparison and statistical analysis, often setting the next popular feature representation and machine learning technique. In the 2017 challenge, convolutional networks have dominated the methods, while the mel representations stayed favorite from previous editions. In contrast to 2016, there was significant difference in performance between first few top systems, and most test audio segments were correctly classified by at least one of them - suggesting that fusion of different features and methods may achieve close to perfect classification accuracy. The upcoming challenge raises the difficulty of the acoustic scene classification task further by employing a more diverse and much larger dataset, in combination to the short audio segment duration, opening the way for new approaches.

## 5. REFERENCES

[1] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[2] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, Jan 2006.

[3] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 23, no. 1, pp. 142–153, Jan. 2015.

[4] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. on Multimedia*, vol. 17, no. 10, pp. 1733–1746, October 2015.

[5] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, Feb 2018.

[6] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 85–92.

[7] H. Eghbal-zadeh, B. Lehner, M. Dorfer, and G. Widmer, "A hybrid approach with multi-channel i-vectors and convolutional neural networks for acoustic scene classification," in *2017 25th European Signal Processing Conference (EUSIPCO)*, Aug 2017, pp. 2749–2753.

[8] E. Marchi, D. Tonelli, X. Xu, F. Ringeval, J. Deng, S. Squartini, and B. Schuller, "Pairwise decomposition with deep neural networks and multiscale kernel subspace learning for acoustic scene classification," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 65–69.

[9] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen, "A convolutional neural network approach for acoustic scene classification," in *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 1547–1554.

[10] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 11–15.

[11] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature learning with matrix factorization applied to acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1216–1229, June 2017.

[12] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[13] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference 2016 (EUSIPCO 2016)*, 2016.

[14] S. Mun, S. Park, D. K. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyperplane," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 93–102.

[15] W. Zheng, J. Yi, X. Xing, X. Liu, and S. Peng, "Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 133–137.

[16] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Nonnegative feature learning methods for acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 22–26.

[17] Y. Han and J. Park, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 46–50.

[18] A. Mesaros, T. Heittola, and T. Virtanen, "Assessment of human and machine performance in acoustic scene classification: DCASE 2016 case study," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE Computer Society, 2017, pp. 319–323.