ASSESSMENT OF HUMAN AND MACHINE PERFORMANCE IN ACOUSTIC SCENE CLASSIFICATION: DCASE 2016 CASE STUDY

Annamaria Mesaros, Toni Heittola, Tuomas Virtanen

Tampere University of Technology, Department of Signal Processing
Korkeakoulunkatu 10, 33720, Tampere, Finland
annamaria.mesaros@tut.fi, toni.heittola@tut.fi, tuomas.virtanen@tut.fi

ABSTRACT

Human and machine performance in acoustic scene classification is examined through a parallel experiment using TUT Acoustic Scenes 2016 dataset. The machine learning perspective is presented based on the systems submitted for the 2016 challenge on Detection and Classification of Acoustic Scenes and Events. The human performance, assessed through a listening experiment, was found to be significantly lower than machine performance. Test subjects exhibited different behavior throughout the experiment, leading to significant differences in performance between groups of subjects. An expert listener trained for the task obtained similar accuracy to the average of submitted systems, comparable also to previous studies of human abilities in recognizing everyday acoustic scenes.

Index Terms— acoustic scene classification, machine learning, human performance, listening experiment

1. INTRODUCTION

Acoustic scene classification has been recently receiving a lot of attention, mainly due to development of context-awareness applications for portable devices, and is commonly framed as a supervised classification problem in which input audio must be categorized into one of a number of predefined classes, which the system is trained to recognize. Acoustic scene classification task was present in evaluation campaigns on environmental sound classification and detection, namely DCASE 2013 [1] and DCASE 2016 [2]. The latter provided a dataset of sufficient size to facilitate methods based on deep learning, which resulted in significantly higher evaluated performance than in the former. In addition, the growing interest for the problem, manifested in the high number of participants to DCASE 2016, allows a wide comparison of state of the art methods.

A large amount of previous work on the topic is available, but the different datasets employed in each study makes comparison more difficult. For example in [3], a classification accuracy of 91% is reported on a dataset containing 3-second sound examples from 10 acoustic scene classes using mel-frequency cepstral coefficients (MFCCs) and hidden Markov models (HMM) with 11 states, while using 3 state HMMs only gives an accuracy of 78%. For the same dataset, authors of [4] reported mean average precision (MAP) of 0.99 using MFCCs and support vector machines (SVM), while the same configuration of MFCCs and SVMs on the DCASE 2013 dataset obtained only 0.52 MAP; however, a higher performance of 0.71 was obtained on DCASE 2013 data using histogram of gradients learned from time-frequency representations.

This work received funding from the European Research Council under the ERC Grant Agreement 637422 EVERYSOUND.

Human performance in recognizing audio scenes has not been the subject of many parallel studies involving both machine and human performance. As human performance must be assessed through listening experiments, obtaining a sufficient number of subjects is usually the main obstacle. For example the experiment in [3] used 14 subjects, each classifying a number of 30 randomly selected 3-second samples, with no initial training. Overall accuracy of human subjects was only 35%, in contrast to the 91% obtained with the MFCC-HMM system. In [5], a listening test was conducted using 19 subjects with 25 different scenes, to determine accuracy, reaction time and acoustic cues for recognition. The average recognition rate was 70%, with an average reaction time of 20 seconds, and most subjects reported recognition was based on prominent identified sound events. A follow-up study involving both human and machine performance used 24 categories, further grouped into 6 higher level ones [6]. Test subjects were required to answer as soon as possible, resulting in a performance of 69% for 24 classes and 88% for 6 classes, slightly higher than the performance of the automatic methods proposed in the same work (58%, and 82%, respectively). Average reaction time was found to be 13 seconds, ranging from 5 seconds (nature) to 21 seconds (library).

For the DCASE2013 acoustic scene dataset containing 10 categories, two independent listening experiments were performed, one in which test subjects had to listen to 50 samples each [7], another in which test subjects were allowed to classify as many test samples as they wanted [8]. In both cases, the subjects had to listen to the full audio. Both tests also considered human subjects as pre-trained, and did not provide any familiarization stage, therefore measure performance based on the personal experience of the subjects. Results of the two listening experiments are similar, with a performance of 72% [8] and 79% [7], both significantly superior to 55% average performance of the machine learning methods.

In this paper we present a detailed analysis of systems submitted to DCASE 2016 Acoustic scene classification task, and the comparison with the human performance on the same evaluation data, determined through a listening experiment. A detailed analysis of the listening experiment is also presented, taking into account the influence of familiarity with the acoustic scenes, user behavior, and characteristics of the acoustic scenes.

2. ACOUSTIC SCENE CLASSIFICATION IN DCASE 2016

Acoustic scene classification in DCASE 2016 was defined as the task of classifying a test recording into one of the 15 predefined classes characterizing the environment in which it was recorded – for example “bus”, “home”, “office”, as illustrated in Fig. 1.

2.1. Dataset and experimental setup

The task used the TUT Acoustic Scenes 2016 dataset [9] that consists of binaural recordings from 15 acoustic scenes: lakeside beach, bus, cafe/restaurant, car, city center, forest path, grocery store, home, library, metro station, office, urban park, residential area, train, and tram. Multiple recordings of 3-5 minutes length were performed for each scene class, and each recording was done in a different location to ensure high acoustic variability. All data was recorded in Finland. The data was distributed as 30-second segments, with clear indication of the original long recording that each segment belongs to. Complete details on data recording, annotation and audio postprocessing procedure can be found in [9].

The development set provided for the task contains approximately 70% of the total available data for each scene, while the other 30% was kept as evaluation set. A cross-validation setup was also provided with the development set, containing 4 folds. The partitioning into development and evaluation subsets, and furthermore into fold subsets, was done such that all 30-second segments from the same original recording were always included into the same subset. For each acoustic scene, there were 78 30-second segments in the development set and 26 in the evaluation set.

A baseline system was also provided, using MFCCs as features with a Gaussian mixture model (GMM) based classifier. MFCCs were calculated in 40 ms frames windowed with a Hamming window with 50% overlap and 40 mel bands; a feature vector for each frame was constructed from the first 20 MFCCs (including 0th MFCC), delta and acceleration coefficients calculated using a window length of 9 frames. In training, a GMM with 16 components was trained for each class using the expectation maximization algorithm. In testing, classification decision is based on maximum likelihood among all available models, with the likelihood accumulated over the entire test sample. Classification accuracy of the baseline system on the development data, obtained using the provided cross-validation setup, is 72.5%, with context-wise performance varying from 13.9% (park) to 98.6% (office). The baseline system classification accuracy on the later released evaluation set is 77.2%.

2.2. Challenge submissions

The task attracted a very high number of participants, receiving a number of 48 submissions from 34 different teams. Most submitted systems outperformed the baseline system; this was expected, given its simplicity. A large number of submissions used mel-frequency scale feature representations, namely MFCCs [10, 11] or log mel energies [12]. The choice is likely motivated by the fact that they provide a good characterization of the spectral properties of the signal, while also providing reasonably high inter-class variability that allows discrimination between the categories. Other feature representations included CQT-based time-frequency representations [13], combinations of various features (including mel-based) [14, 15, 16], and representations obtained in unsupervised way [17].

A large majority of the submitted systems were based on deep learning, signaling a shift towards pursuing highest possible performance with no concern for computational complexity, as usually neural networks have a large number of parameters to be optimized. Of the 48 systems, 22 used deep learning, 10 used SVMs [10], while ensemble classifiers account for other 10 [14, 16, 15]. Choices for neural networks include feed-forward, convolutional (CNN) [12, 17], recurrent (RNN, including LSTM), and combinations of neural networks with other techniques, specifically GMMs [11]. Factor analysis methods also perform well: even though they were not extensively used, i-vectors [14] and NMF based methods [13] are among top performing systems - exploiting the fact that each scene is composed of multiple sources whose joint variations can be explained using latent variables.

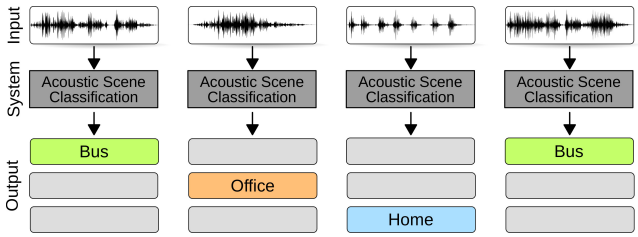

Figure 1: Acoustic scene classification: classifying a test example into one of the predefined acoustic scene categories.

2.3. Analysis of challenge results

The performance of submitted systems varies from 89.7% to 62.8%, with 10 top systems having over 85% accuracy, and 9 systems having accuracies lower than the provided baseline system. Figure 2 presents information on systems that performed better than the baseline. In the figure, the 95% confidence intervals are also presented, calculated as a binomial proportion confidence interval for the classification output being correct or incorrect with respect to the ground truth. It can be seen that confidence intervals of closely performing systems overlap significantly. A further analysis was performed using McNemar's test [18], by comparing the classification output in pairs, with a significance level of 0.05. The results show that 6 systems can not be considered as performing differently than the winner, and similarly, a number of systems can not be considered as performing differently from the baseline system.

Class-wise results show rather large difference in classification performance between the systems for different scene categories. For different systems, most difficult scenes are library, with lowest score obtained by at least one system 0%, and train, with lowest score 11.5%. Other relatively difficult scenes are cafe (lowest score 19%), residential area (23%), and home (34%). On the other hand, beach bus, car, and office all had a score of at least 69% in all systems. Detailed information on class-wise performance for all submissions is available on the challenge website [2].

The overall confusion matrix of the submitted systems is presented in Fig. 3, with values over 10 marked for the cells. Average performance across contexts for all systems is 80.9%, while the overall performance determined by a majority vote among all systems is 87.2%%. Office and tram are on average the easiest to recognize, with a 96% accuracy, while the average for the difficult classes stays low, with library having an average accuracy of only 43%, and train 52%. Most notable confusions are between urban park and residential area, and between library and home. These confusions are understandable from the human perspective, as the scenes have similar acoustic content, park having sounds of birds, children and human activity in the foreground and street sounds in the background, while residential area consists of low traffic street scenes having birds and human activity (gardens) as background. Similarly, library and home scenes are quiet and with rare sounds (some recordings in single person home, no conversation, reading or typing). The confusion between train and cafe is explained by recordings made in the restaurant car of the train, where the dominating acoustic characteristics reflect more a cafeteria than a train.

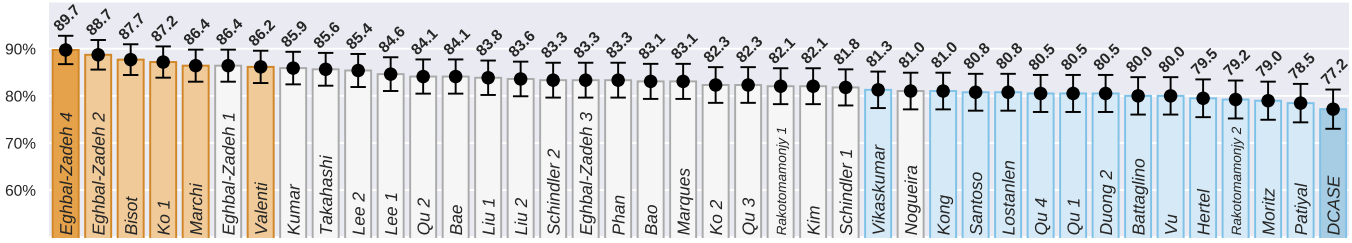

Figure 2: Acoustic scene classification task results on the evaluation set and 95% confidence intervals of their performance. Based on McNemar test with a significance level of 0.05, 5 systems cannot be judged to perform differently than the winner (orange), and 13 are not different from the baseline system (blue) under same statistical test conditions.

3. LISTENING EXPERIMENT

For a direct comparison with human perspective of the same data, a listening experiment containing the samples from the evaluation dataset was set up. Due to the size of the dataset, subsets containing 30 test samples were presented to each test subject, 2 samples for each scene category. The test samples per subject were randomly selected without replacement, resulting in the complete evaluation dataset being distributed among 13 test subjects.

3.1. Experiment setup

The experiment was implemented to be used online, such that the DCASE challenge participants can take part. Subjects were advised to do the experiment in a noiseless environment and to use good quality headphones. Subjects were asked if they reside in Finland, for analyzing influence of soundscape familiarity on the classification accuracy. First, the participants were offered a familiarization stage, in which 3 examples of 10 seconds for each of the 15 scene classes were presented, along with their label. The familiarization samples were randomly selected from the development set for each subject. Test subjects were instructed to listen to as many samples and as many times as they want, and proceed to the test when confident enough of their abilities. A record of the samples played by the user on the familiarization page was kept, to analyze recognition performance with respect to the use of training.

The test material was presented as separate task pages, with one single audio sample per page and radio buttons for the 15 acoustic scene classes. Test subjects were informed that the audio sample can be listened to multiple times, and instructed to select an answer when confident in their choice. It was not necessary to listen fully the audio sample. A record of the time spent on each task page was kept, to analyze recognition performance with respect to reaction time.

3.2. Analysis of human performance

A number of 87 participants provided 2610 individual task answers. For evaluation, each audio sample is considered as a separate test item and compared to the corresponding ground truth. The overall performance of the human subjects calculated over all answers was 54.4 %, which is surprisingly low when compared with the performance of machine learning methods. Previous parallel performance studies resulted in human performance similar or higher than machine learning performance, with the notable exception of very short audio samples in [3]. We hypothesize that the results are dominated by test subjects whose own experience of soundscapes does not correspond with the characteristics of the data recorded in Finland. Indeed, by grouping test subjects based on location, we obtained a recognition accuracy of the group familiar with Finnish soundscape of 60.4% (with a 95% CI 55.9-64.8), while for the other group the accuracy is only 53% (with a 95% CI 50.9-55.1), indicating a clear relationship between familiarity and performance. These results are included in the "Familiarity" panel in Fig. 5.

The confusion matrix obtained from the listening experiment is presented in Fig. 4, showing that confusions between scene categories is more distributed than for the machine learning methods. Similarity in the two cases is observed for confusion between park and residential area, tram to train and train to cafe, likely due to the reasons described in Section 2.3, but human performance is notably lower than machine performance. Other confusions tend to be grouped within subjectively similar acoustic scenes, such as nature scenes (beach, park, forest path, residential area), street scenes (city center and residential area) or quiet space (office, library, home).

To study the effect of the training phase, subjects were grouped based on their behavior in the familiarization stage into three groups: subjects that listened to at most half of the provided examples, subjects that listened to between half and all, and subjects that listened to all provided examples at least once. Performance of each group was assessed separately. Based on the results presented in Fig. 5 in the "Training" panel, we observe that subjects that spent more time in the familiarization stage and listened to all data had higher scores, indicating a clear relationship between training and recognition performance.

To study user behavior, subjects were grouped based on the average time they spent on the task pages, considering that a conscientious test subject will listen to the entire audio sample before selecting an answer. Based on conclusions from [6] (13 s) and [5](20 s), we created groups for subjects that listened on average to less than half the audio sample (spending under 15 s per task page), between half and full length (15 to 30 s), and subjects that listened to the audio sample fully at least once (over 30 s). Performance of the three groups is presented in Fig. 5 in the "Test subject" panel. Results indicate that subjects that listened to more of the test sample had higher scores, even though there was not a very high difference between users that tend to listen to the entire sample or over half of it; however, performance of users that tended to rush through the listening experiment was significantly lower.

To study the influence of the acoustic characteristics, the relative ease of recognizing samples from different acoustic scenes was analyzed by taking into account the average time spent per task page, irrespective of the test subject. We consider that the samples that are easy to recognize require only a short time, and the test subject is therefore confident in selecting an answer early. The same temporal limits were used for grouping the obtained answers

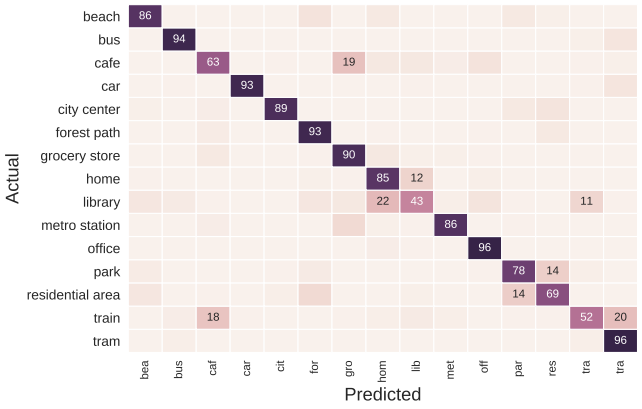

Figure 3: Confusion matrix of all submitted systems

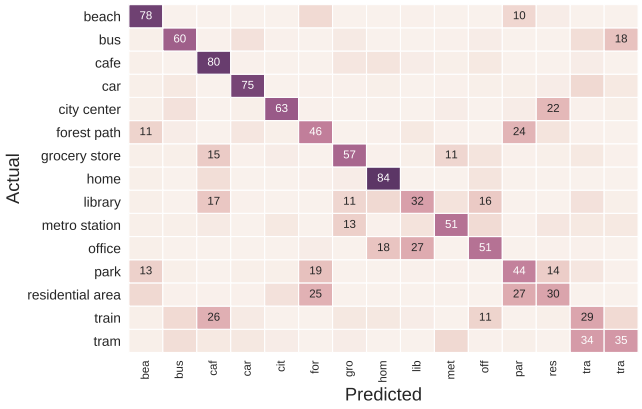

Figure 4: Confusion matrix of human classification

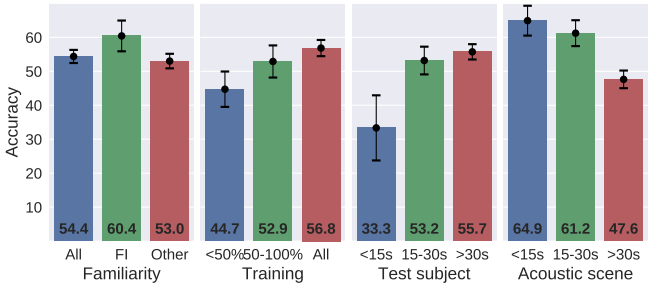

Figure 5: Human performance analyzed with respect to familiarity with the scene, use of training, test subject behavior and relative ease of recognition

as previously (<15 s, 15-30 s, >30 s), but this time grouping individual answers, not test subjects. Results presented in Fig. 5 show that indeed the test samples that prompted a quick reaction from the users had the highest average recognition accuracy, even though the difference between the first two groups is not very large. On the other hand, samples that were listened to fully or possibly multiple times had a significantly lower accuracy, suggesting that if the cues for recognition are not found early enough, listening to the sample multiple times does not help with recognition. This result correlates with the class-based accuracy, as acoustic scenes that needed smaller time per task page on average had the highest recognition accuracy (beach, cafe, car have average time per task page 14, 18, 22 s and accuracy 78, 80, 75 %, while library, park, train and tram have average time per task page 35, 39, 33, 35 s and accuracy 32, 44, 29, 35 %).

3.3. Expert listener

In order to investigate the reason for the large gap between human and machine performance, an additional listening experiment was prepared, containing the complete evaluation dataset to be classified by one research assistant who had recorded and annotated data in TUT Acoustic Scenes 2016 dataset. We consider that he was exposed to much higher amount of training data than the other subjects, and was familiar with the definition of the classes used in the data collection. The experiment was set up the same way, including a familiarization stage, as the data recording process was done 8-12 months earlier. Performance of the expert listener was calculated only for the test samples he has not recorded himself (262 of 390 samples). For this set, his recognition accuracy was 77.1%.

Class-wise performance of the expert listener was better than the overall machine performance for six of the 15 classes, with notable 100% accuracy on bus, home and park, while being significantly lower than machine performance for others, for example, forest path 57% vs 93%, library 25% vs 43%, and residential area 50% vs 69%. The gap in performance between human and machine performance observed for some classes by comparing figures 3 and 4 is present to a lesser degree in the expert listener results, e.g. office (machine 96%, human 51%, expert 70%), or tram (machine 96%, human 35%, expert 70%). The 77.1% average is comparable to the human performance obtained in the other mentioned studies with comparable number of categories, and is also close to the average performance of all systems, which is 80%. This indicates that with training, human performance in recognizing 10-25 acoustic scenes can reach 80%.

4. CONCLUSIONS

The growing amount of data available for training supervised machine learning methods to classify acoustic scenes brings a clear improvement of performance and generalization properties to algorithms, observable between the consecutive editions of DCASE. At the same time, human subjects tested on large amount of data with high variability perform noticeably worse than automatic methods when not sufficiently trained for recognizing the acoustic scenes. Confusions between classes on both human and machine sides originate from common characteristics of the scenes, such as similar sounds (urban park vs residential area, forest path vs park) or quiet background not providing sufficient acoustic cues for recognition (home vs library), or from the very definition of the acoustic scene (restaurant vs restaurant car in the train).

Follow-up work includes extension of the database, recording in more diverse geographical locations for increased acoustic variability. To increase the complexity of the problem, audio segments of shorter length will be provided, reducing the amount of information available in the decision making process, resulting in added difficulty for both machine and human decisions. The extension of the database will possibly benefit DNN-based methods, but this remains to be confirmed after the completion of the next challenge [19].

5. REFERENCES

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

[2] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. Plumbley, "Detection and classification of acoustic scenes and events DCASE 2016," http://www.cs.tut.fi/sgn/arg/dcase2016/, 2016, [Online; accessed 4-Apr-2017].

[3] L. Ma, D. J. Smith, and B. P. Milner, "Context awareness using environmental noise classification." in *INTERSPEECH*. ISCA, 2003.

[4] 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.

[5] V. Peltonen, A. Eronen, M. Parviainen, and A. Klapuri, "Recognition of everyday auditory scenes: Potentials, latencies and cues," in *Audio Engineering Society Convention 110*, May 2001.

[6] 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 Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, Jan 2006.

[7] J. Krijnders and G. t Holt, "Tone-fit and mfcc scene classification compared to human recognition," http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/abstracts/SC/KH.pdf, [Online; accessed 4-Apr-2017].

[8] D. Barchiesi, D. Giannoulis, D. Stowell, and M. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *Signal Processing Magazine, IEEE*, vol. 32, no. 3, pp. 16–34, May 2015.

[9] 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)*, Budapest, Hungary, 2016.

[10] B. Elizalde, A. Kumar, A. Shah, R. Badlani, E. Vincent, B. Raj, and I. Lane, "Experiments on the DCASE challenge 2016: Acoustic scene classification and sound event detection in real life recording," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 20–24.

[11] G. Takahashi, T. Yamada, S. Makino, and N. Ono, "Acoustic scene classification using deep neural network and frame-concatenated acoustic feature," http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-acoustic-scene-classification, DCASE2016 Challenge, Tech. Rep., September 2016, [Online; accessed 15-Mar-2017].

[12] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "DCASE 2016 acoustic scene classification using convolutional neural networks," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 95–99.

[13] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Supervised nonnegative matrix factorization for acoustic scene classification," http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-acoustic-scene-classification, DCASE2016 Challenge, Tech. Rep., September 2016, [Online; accessed 15-Mar-2017].

[14] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "CP-JKU submissions for DCASE-2016: a hybrid approach using binaural i-vectors and deep convolutional neural networks," http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-acoustic-scene-classification, DCASE2016 Challenge, Tech. Rep., September 2016, [Online; accessed 15-Mar-2017].

[15] 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 *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 65–69.

[16] S. Park, S. Mun, Y. Lee, and H. Ko, "Score fusion of classification systems for acoustic scene classification," http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-acoustic-scene-classification, DCASE2016 Challenge, Tech. Rep., September 2016, [Online; accessed 15-Mar-2017].

[17] J. Kim and K. Lee, "Empirical study on ensemble method of deep neural networks for acoustic scene classification," http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-acoustic-scene-classification, DCASE2016 Challenge, Tech. Rep., September 2016, [Online; accessed 15-Mar-2017].

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

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