DCASE 2016 CHALLENGE: RANDOM SYSTEM PERFORMANCE IN TASK 3

Christian Kroos and Mark D. Plumbley

Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey
Guildford, Surrey, GU2 7XH, UK
{c.kroos, m.plumbley}@surrey.ac.uk

ABSTRACT

In this report we describe the creation of a random, data-blind systems to provide a random baseline for Task 3 (sound event detection in real life audio) in the DCASE 2016 challenge. Particular attention is paid to the results of two sound events occurring in the residential area scene, one very rare, the other very frequent. The relatively good performance of the random system in comparison to the results of the proper detection systems shows the difficulty of Task 3 given the current state-of-the-art sound detection methods.

Index Terms— sound detection, random baseline, rare events

1. INTRODUCTION

Chance level performance is one of the most important comparison levels in detection and identification experiments, be they with humans, other animals or machines. Unlike in the case of classification tasks with fixed numbers of classes, determining chance level in real-world sound event detection with multiple overlapping events is not analytically straightforward. More importantly, systems that act randomly on the test data, might be still able to improve their performance by using constraints of the training data, e.g., the relative frequency of the different event classes or the fact that is unlikely that more than n events occur simultaneously. The latter applies in particular if the ground truth consists of the labeling of human experimenters, who might not be able to distinguish a higher number of simultaneously occurring events and the actual number might vary and dependent on the types of events. Thus, chance level performance cannot always be easily determined and simulations are required.

As shown by [1], the two error measures used in the DCASE 2016 challenge, ER and F [2], identified in combination correctly systems that approach zero-output or all-active-output systems in the Office Synthetic task from DCASE 2013. The study also shows a clear difference between the employed random system and the proper detection systems on these metrics. However, the performance of the GMM-based DCASE 2016 baseline system in Task 3 (sound detection in real-life audio task) [3] hints at a more complex situation here, especially when looking at individual events. A number of infrequently occurring events were not detected at all by the baseline system and the overall ER value for the *Residential Area* scene were just below the critical ER boundary of 1 (performance of a zero-output system), while the ER value for the *Home* scene was even clearly above. We considered it therefore worthwhile to investigate the performance of a random (test) data-blind system. More specifically, our hypothesis was that a random system using training data constraints would, while not being able to approach performance values of any proper detection system, still fare relatively well. The distance between the proper systems and the random baseline would provide a non-trivial estimate of the overall state of the art in this task and gauge the degree of difficulty given current methods.

The relationship between very frequent and very rare events with regard to the overall performance evaluation of a system is often both complex and crucial. For instance, in surveillance and in health monitoring, critical events might be encountered extremely rarely, while events indicating a normal situation are likely to to be very frequent. But overall both event types need to be equally reliably detected and identified: The normal event to ensure the system is working, the critical event to trigger an action.

Table 1: Priors derived from the empirical distribution of the events in the training data set

	Event class	p
Home	(object) rustling	0.054
	(object) snapping	0.008
	cupboard	0.007
	cutlery	0.018
	dishes	0.049
	drawer	0.008
	glass jingling	0.009
	object impact	0.066
	people walking	0.024
	washing dishes	0.103
	water tap running	0.089
Residential area	(object) banging	0.003
	bird singing	0.292
	car passing by	0.148
	children shouting	0.013
	people speaking	0.087
	people walking	0.048
	wind blowing	0.036

The difference in frequency introduces a bias in the performance measure for proper detection systems when segment-based evaluation is used. Systems that tend to ignore rare events and are overly eager in detecting frequent events fare better than systems

The research leading to this submission was funded by EPRSC grant EP/N014111/1.

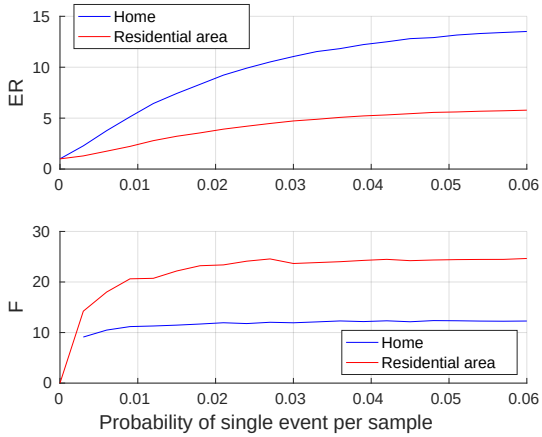

Figure 1: Segment-based evaluation measures averaged over all events in condition UNI.

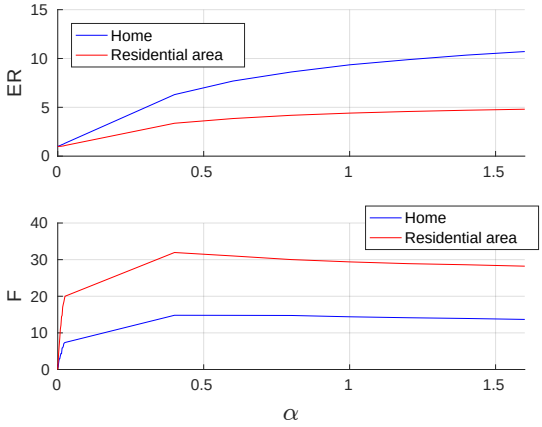

Figure 3: Segment-based evaluation measures averaged over all events in condition PRIOR_50.

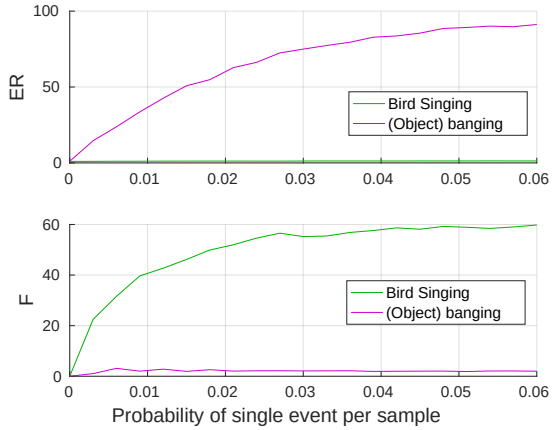

Figure 2: Segment-based evaluation measures of selected events in condition UNI.

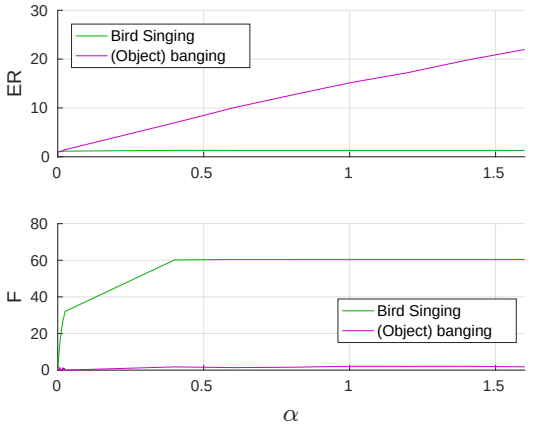

Figure 4: Segment-based evaluation measures of selected events in condition PRIOR_50.

on an comparable level of recognition success that make no distinctions. As mentioned above, in many scenarios it might not be a question of finding a suitable spot on the ROC (Receiver Operator Characteristics) curve since the preference for high rates of true positives or low rates of false positives needs to be decided differently based on the frequency and importance of individual event classes.

In the DCASE 2016 challenge Task 3 rare and frequent events are present and events occur simultaneously. It should be also noted that the relationship between the evaluation window and the sample rate used for the event detection plays a role. Most systems submitted to the challenge will presumably have a higher temporal resolution than the 1 Hz evaluation since they will use their feature extraction window length as their basic frame size. However, there should be no possible gains compared to the case where both resolutions are matched: The evaluation counts an event active no matter how much of the one-second evaluation segment it occupies if its onset or offset lies within the segment boundaries. Most detection algorithms will need a higher sample rate to adequately identify events, but in the output the detailed temporal information can be discarded. In this respect, the task - viewing it from the perspective of segment-based evaluation - resembles more an audio tagging than a detection task.

2. METHOD

The DCASE 2016 Task 3 data set and evaluation routines [3] in Matlab were employed for all evaluations. We will focus here only on the segment-based results since they are easier to interpret.

In the first approach we generated (pseudo-)randomly event entries on a per-sample basis for each individual event class using a fixed probability assuming an underlying uniform distribution (condition UNI). We increased the probability from $p = 0.003$ to $p = 0.06$ in steps of 0.003, after determining this range in pilot simulations as the most interesting. We used the same sample rate as in the DCASE base line, i.e., 50 Hz.

The second approach utilised frequency information available

in the training set, but again never considered the actual test data set. Priors were derived from the empirical distribution of the events in the training set of each fold for this evaluation and the full data set for the challenge, respectively. Table 1 shows the values based on 50 Hz sample rate. Note that we 'unwrapped' overlap by counting overlapping events as sequential in order for the individual values to sum up to 1 (this included an event 'other' that captured all segments that had no relevant event indicated).

We then tested a range of multiplier values applied to all event priors, i.e.,

$$p_e^{(m)} = \alpha \, p_e \tag{1}$$

We went from $\alpha = 0.001$ in steps of 0.002 to 0.025 and from 0.4 in steps of 0.2 to 1.6. These ranges were again chosen through manual inspection of pilot experiments. The sample rate was unchanged at 50 Hz (condition PRIOR_50).

In the third and last approach we matched the sample rate to the evaluation interval, that is, it was set to 1 Hz. The range of α for multiplying the priors was adjusted based on pilot experiments, extending here from 0.1 to 2.5 in steps of 0.1 (condition PRIOR_1).

3. RESULTS AND DISCUSSION

3.1. Development data set

Figure 1 and Figure 2 show ER and F segment-based results for all events averaged and for the class *bird singing* and *(object) banging* separately in the UNI condition. The evaluation is based on a single run. As can be seen the two measures capture adequately the fact that no real detection happens, only insertion of events by chance. The relatively low ER values at low insertion rates are accompanied by low F values. With more events being inserted F rises, but so does ER.

Figure 3 and Figure 4 show ER and F segment-based results for all events averaged and for the class *bird singing* and *(object) banging* separately in the PRIOR_50 condition. The evaluation is based on 5 runs. F values are substantially higher, but at the expense of higher ER values (comparable to UNI condition). Only in the very low values for α a better balance between relatively F high values and low ER values is achieved as can be seen in Figure 5 and Figure 6 which give a detailed view of the range that was evaluated with a finer resolution (from $\alpha = 0.001$ to 0.025).

Finally, Figure 7 and Figure 8 show ER and F segment-based results for all events averaged and for the class *bird singing* and *(object) banging* separately in the PRIOR_1 condition. The evaluation is based on 10 runs. Not surprisingly, slightly higher F value are accomplished, but when seen in relation with the increase in ER, only a small gain is obtained when comparing to equivalent values in the PRIOR_50 condition, e.g., PRIOR_1 $\alpha = 1$ to PRIOR_50 $\alpha = 0.02$.

In all conditions, remarkable high F values for a single, very frequent event can be achieved, while maintaining a still relatively low ER value. The opposite is true regarding the single rare event we analysed. This suggests that any real detection system that is biased toward detecting frequent events while avoiding attempting to detect rare events, will gain a small advantage in an overall evaluation over systems with no such bias.

Based on our findings, we selected the PRIOR_1 condition with an $\alpha = 1$ (unmodified empirical priors with the virtual sample rate matched to the DCASE Task 3 evaluation segment length) as the final candidate to provide the best random baseline. This is of course in line with expectations.

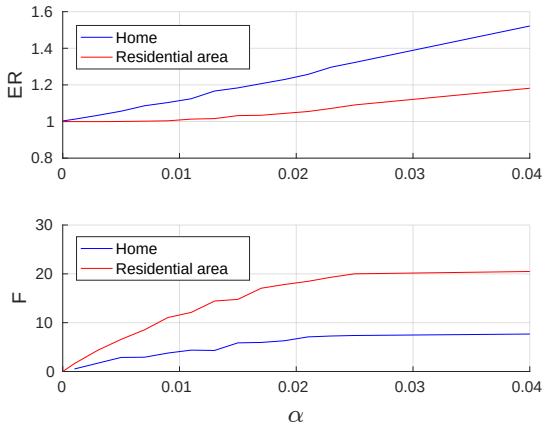

Figure 5: Segment-based evaluation measures averaged over all events in condition PRIOR_50 (detail).

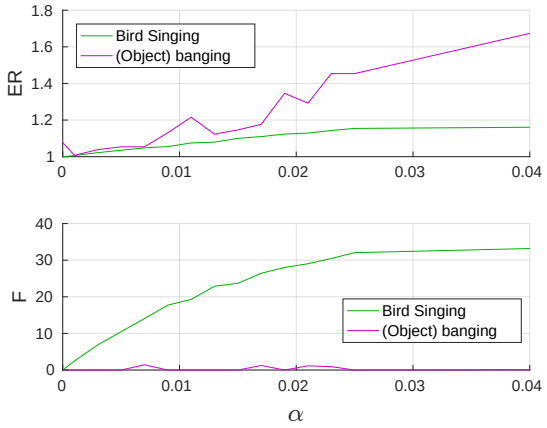

Figure 6: Segment-based evaluation measures of selected events in condition PRIOR_50 (detail).

The four-fold average segment-based performance values for the development data were: $ER = 1.4052$; $F = 7.7$ for the *Home* scene and $ER = 1.0609$; $F = 21.6$ for the *Residential Area* scene. For the selected events we found *bird singing* having $ER = 1.1479$ and $F = 33.6$ and *(object) banging* with having $ER = 1.4615$ and $F = 0.0$.

3.2. Challenge data set

The challenge submissions were evaluated by the DCASE 2016 organisers and the results published on-line[1]. Attesting to the difficulty of Task 3, the random data-blind system performed relatively well. In both error metrics it was found on the lower edge of the distribution of the proper systems, performing better as one other system in each metric though different ones with respect to ER and F. Its overall challenge result values were $ER = 1.1488$

[1] http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-sound-event-detection-in-real-life-audio

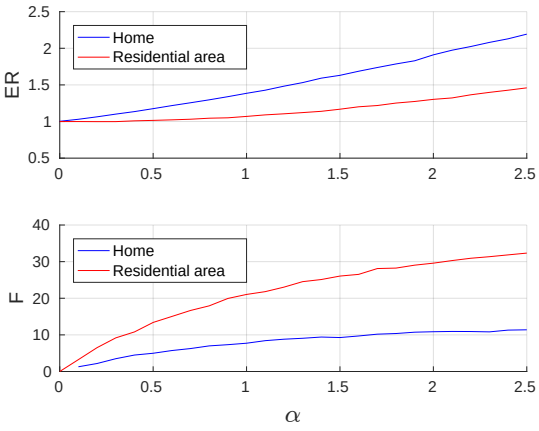

Figure 7: Segment-based evaluation measures averaged over all events in condition PRIOR_1.

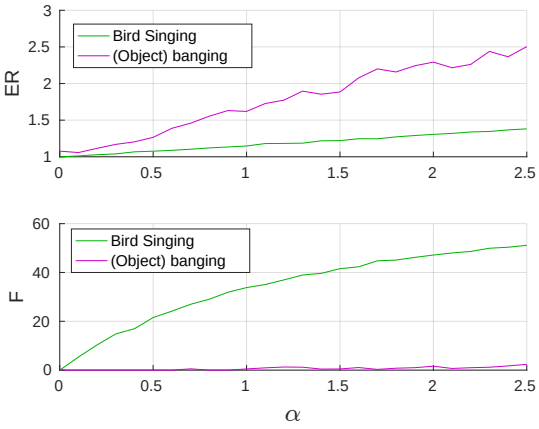

Figure 8: Segment-based evaluation measures of selected events in condition PRIOR_1.

and $F = 16.8$. In the *Home* scene it achieved $ER = 1.6394$ and $F = 6.7$, corresponding to the 14. and 13. place, respectively, in the ranking of the 17 submitted systems. In the *Residential Area* scene its application resulted in $ER = 1.6154$ and $F = 12.5$, corresponding to the 14 and 15 place. As was expected, its performance is consistent across all events.

While in the *Home* scene results from the development evaluation and the challenge closely match, there is substantial difference in the *Residential Area* scene with the challenge values being less favourable. This is not surprising as random system would vary greatly in their output by nature and only the output of a single system run was submitted to the challenge. Usually it would be, of course, advisable to produce a sizable number of runs and subsequently use means and standard deviations of these runs. This, however, was not a viable option here as it would have inundated the challenge with random system submissions. The difference between development and challenge results should serve as a reminder that random systems could perform even better: if the difference would have gone in the other direction with a 'lucky' run the F value in *Residential Area* might have reached 30 % and the ER might have even crossed the crucial boundary of 1. As consequence, one would require for any proper system a pronounced difference from the evaluation results of the random system. Once the challenge ground truth labeling will be published, we will be able to provide confidence intervals.

The error assessment of two selected events resulted in the following: *bird singing* had $ER = 1.1695$ and $F = 35.0$ and *(object) banging* $ER = 1.0909$ and $F = 0.0$. Here only the ER of event *(object) banging* differs markedly from the development results, being much lower in the challenge and with this emphasising the point made above about 'lucky' systems. Note that for these two events the random system competes as well as proper detection systems. The very rare event (*(object) banging*) was not detected by any system (F is always 0). Only 9 systems had no false alarms either (ER = 1), resulting in rank 10 for the random system with the second best ER value (presumably based on a single segment's false positive). The very frequent event *bird singing* exhibited high F values signaling high detection rates in all systems except for one. The random system scored only rank 15 here, but was still above the DCASE challenge official baseline. However, the high detection rates were generally achieved at the expense of high false alarm rates: Only two systems accomplished an ER value below 1. The random system achieved rank 8 when the *bird singing* results were ranked according to their ER values.

4. CONCLUSION

We provided a random baseline for Task 3 of the DCASE 2016 challenge. Using only distribution information contained in the training data, 'detected' events were generated with the probability of the empirical priors. No test data were used in the making of the result files. The evaluation results point to the difficulty of this task as the random (test) data-blind system was in the range of the less well performing proper systems.

The analysis of the performance of two selected events, one very rare, the other very frequent, showed a surprising good performance relative to the proper detection systems. In applications, in which very frequent events indicate normal status and the absence of these events a critical situation and simultaneously vice versa the occurrence of very rare events a critical situation and their absence normal status, a different weighting of the different classes of events might be necessary in the evaluation of sound detection systems.

5. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[2] G. E. Poliner and D. P. W. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 154–154, 2007.

[3] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *24rd European Signal Processing Conference 2016 (EUSIPCO 2016)*, Budapest, Hungary, 2016.