CHIME-HOME: A DATASET FOR SOUND SOURCE RECOGNITION IN A DOMESTIC ENVIRONMENT

Peter Foster [*], *Siddharth Sigtia* [*], *Sacha Krstulovic* [†], *Jon Barker* [‡], *and Mark D. Plumbley* [§]

[*] School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[†] Audio Analytic, Cambridge, UK
[‡] Department of Computer Science, University of Sheffield, UK
[§] Centre for Vision, Speech and Signal Processing, University of Surrey, UK

ABSTRACT

For the task of sound source recognition, we introduce a novel data set based on 6.8 hours of domestic environment audio recordings. We describe our approach of obtaining annotations for the recordings. Further, we quantify agreement between obtained annotations. Finally, we report baseline results for sound source recognition using the obtained dataset. Our annotation approach associates each 4-second excerpt from the audio recordings with multiple labels, based on a set of 7 labels associated with sound sources in the acoustic environment. With the aid of 3 human annotators, we obtain 3 sets of multi-label annotations, for 4378 4-second audio excerpts. We evaluate agreement between annotators by computing Jaccard indices between sets of label assignments. Observing varying levels of agreement across labels, with a view to obtaining a representation of 'ground truth' in annotations, we refine our dataset to obtain a set of multi-label annotations for 1946 audio excerpts. For the set of 1946 annotated audio excerpts, we predict binary label assignments using Gaussian mixture models estimated on MFCCs. Evaluated using the area under receiver operating characteristic curves, across considered labels we observe performance scores in the range 0.76 to 0.98.

***Dataset URL*: http://archive.org/details/chime-home**

***Index Terms*—** Computational Auditory Scene analysis, Sound Source Recognition, Datasets

1. INTRODUCTION

In the field of computational auditory scene analysis (CASA) [1], the problem of classifying and detecting sounds has received considerable interest in recent years. Related to audio classification and detection there exist tasks each of great practical concern, including robust speech recognition [2], audio source separation [3], and automatic music transcription [4].

This paper relates to the relatively unexplored CASA task of *sound source recognition* (SSR) [5], which we define as the task of assigning a set of semantic labels to a given audio recording, based on the sound sources contributing to the *acoustic scene* (cf. [6]). SSR closely relates to *acoustic scene classification* (ASC) [7], which aims at assigning a single semantic label to an audio recording, describing the environment which produced the acoustic scene. Further, SSR closely relates to *acoustic event detection* (AED) [8], which aims at identifying perceptually relevant segments in a recording and assigning a semantic label to each obtained segment.

In this paper, we consider the particular case of in-home SSR. For in-home SSR, potential applications include human activity monitoring, in particular for the purpose of safety and security: it may be useful to detect intrusions from the sounds associated with human presence. A further application is the monitoring of the elderly, where it may be useful to identify unusual activity patterns, or unusual sounds such as those produced by falling bodies or distress calls. For such applications, SSR might be used in audio-capturing devices such as surveillance cameras. A further possibility exists in connecting dedicated low-cost SSR devices with e.g. light control units or programmable thermostats, using the Internet of Things [9].

Despite its potential applications, SSR remains comparatively unexplored in the literature [7]; moreover there exist few investigations which consider the particular case of in-home SSR [10]. In the broader context of CASA, we note continuing efforts to improve the evaluation of machine listening systems, by way of evaluation methodologies [11], coordinated evaluations such as the classification of events, activities and relationships (CLEAR) evaluation [8, 12], the IEEE audio and acoustic signal processing (AASP) challenge on detection and classification of acoustic scenes and events (DCASE) [13], or publicly available datasets [14–16].

This paper aims to promote investigations on SSR, by providing a publicly available dataset based on existing recordings made in a domestic environment [17]. Notably, our dataset differs from [15, 16], being based on domestic environment recordings for the intended applications of human activity monitoring and voice detection; further our annotation protocol differs from [14] in that we obtain annotations with the aid of multiple annotators, among which we subsequently quantify agreement. Further, we introduce the CHiME-Home dataset, with the aim of proposing an IEEE AASP challenge on SSR as part of future work.

2. AUDIO RECORDINGS

We obtain annotations for approximately 6.8 hours of binaural audio recordings made in a domestic environment. The recordings comprise audio data previously made available for the Computational Hearing in Multisource Environments (CHiME) project [17, 18][1] . These recordings were obtained by positioning recording equipment inside an English Victorian semi-detached house.

This work was funded by the UK Engineering and Physical Sciences Research Council (EPSRC) grants EP/M507088/1 and EP/K009559/1, and by Innovate UK grant 41311-293383.

[1] `http://spandh.dcs.shef.ac.uk/projects/chime/PCC/data/PCCdata48kHz_train_background.part{1,2,3}.tar.gz`

Label	Description
c	Child speech
m	Adult male speech
f	Adult female speech
v	Video game/TV
p	Percussive sounds, e.g. crash, bang, knock, footsteps
b	Broadband noise, e.g. household appliances
o	Other identifiable sounds
S	Silence / background noise only
U	Flag chunk (unidentifiable sounds, not sure how to label)

Table 1: Set of permitted labels and their descriptions, as provided to human annotators. An annotator may assign any subset of labels to an audio chunk, with the exception of labels S,U which respectively may only be assigned in isolation.

The recordings were selected from 22 sessions totalling 19.5 hours, with each session made between 7:30 in the morning and 20:00 in the evening [17]. In the considered recordings, the equipment was placed in the lounge (sitting room) near the door opening onto a hallway, with the hallway opening onto a kitchen with no door. With the lounge door typically open, prominent sounds thus may originate from sources both in the lounge and kitchen.

Prominent sound sources in the acoustic environment are two adults and two children, television and electronic gadgets, kitchen appliances, footsteps and knocks produced by human activity, in addition to sound originating from outside the house. The kitchen contains a fridge, washing machine, microwave oven, kettle, sink and tap, gas hob and a boiler. The lounge contains a television, gas fire and a radiator. The lounge has a carpeted floor, whereas the kitchen has a linoleum floor; the hallway is sparsely furnished with a wooden floor. Full details on the recording setup can be found in [17].

3. ANNOTATION PROTOCOL

To obtain annotations, we partitioned the audio recordings into non-overlapping 4-second chunks. We motivate such a chunk size following [15, 19]; in [19] it is further observed that a 4-second chunk size yields 82% accuracy for human listeners distinguishing among 14 acoustic scene classes. In addition, a 4-second chunk size represents a trade-off between temporal resolution of annotations and the time cost of obtaining annotations. For our considered recordings, we obtain 6138 chunks.

We recruited 3 paid postgraduate Engineering students with self-reported healthy hearing and asked them to each annotate the entire set of 6138 chunks. Thus, for each chunk and for each of the annotators, we obtain a set of labels. There are 9 permitted labels in total, listed in Table 1. Our choice of permitted labels is motivated by the sources present in the considered acoustic environment [17]: Human speakers (c,m,f); human activity (p); television (v); household appliances (b). Informal listening confirmed that there was an abundance of sounds from such sources. In our choice of labels, we further aim to balance utility for applications such as voice and human activity detection, while constraining the complexity of the annotation task. Further labels o,S,U respectively relate to any other identifiable sounds, silence, unidentifiable sounds. Labels S,U may respectively only be assigned in isolation. We require that annotators assign at least one label to a chunk, thus annotators may either assign one or more labels from the set {c,m,f,v,p,b,o}, or may alternatively 'flag' the chunk using a single label from the set {S,U}.

We aim at a representation of 'ground truth' in annotations, therefore a description of the acoustic environment based on [17] was provided to the annotators as additional information before the annotation process started. Further, annotators familiarised themselves with the sound sources in the recordings in an initial 'warm-up' phase: In the warm-up phase, annotators were presented with a balanced sample of 160 chunks, obtained using a preliminary set of annotations gathered for one hour of the recordings. Annotators were asked to distinguish between speech originating from human sources (labels c,m,f) and speech originating from the television (label v).

To control for any effects which might arise from the presentation order of chunks, we randomly shuffle chunks separately for each listener. However, rather than simply shuffle at the level of individual chunks, we first shuffle at the level of 5-minute recording segments, before shuffling chunks within each 5-minute segment. In this way, we retain an amount of recording context between successive chunks within the same 5-minute segment, which we observed reduced time cost in our preliminary annotations. By shuffling chunks within 5-minute segments, we aim to reduce the scope for memory effects, where perceptually salient events at chunk boundaries might influence the annotation of adjacent chunks. Shuffling of chunks avoids annotators being distracted by continuous speech content in recordings.

Using a basic text interface which allowed replaying of chunks[2], annotators iteratively assigned a label string to each chunk. Thus, for a given annotator and chunk there may be multiple assigned labels. The audio playback was in stereo with a sampling rate of 48kHz; chunks were presented using headphones in a quiet office environment, with sound levels adjusted to a comfortable level. Annotators completed their task over the course of 5 sessions, each lasting in the range 2 to 3 hours. In addition to breaks between sessions, annotators were instructed to take a 10-minute break each hour. Annotators each completed their task within 14 hours, corresponding to a labelling time of approximately twice the total audio duration.

Having obtained annotations for 6138 chunks, we reserve 1760 chunks and their annotations for an evaluation dataset in a future IEEE AASP challenge on SSR. Subsequent analysis in this paper is based on the remaining 4378 annotated chunks.

4. ANALYSIS OF ANNOTATION DATA

Figure 1 displays annotator-wise histograms of label occurrences. We observe a relative abundance of labels c,f,v,p,o, compared to labels m,b,S,U. Denoting with σ and μ the standard deviation and mean, to quantify variation in agreement scores we compute the coefficient of variation σ/μ. Thus computed, we obtain a low amount of variation across annotators for labels c,m,f,v with respective values in the range $[0.019, 0.094]$. Contrastingly, we obtain a high amount of variation for labels p,b,o,S,U with respective coefficients of variation in the range $[0.455, 1.608]$. Disregarding label U, we observe a minimum count of 45 for label b; the maximum count is 2625 for label c. In the following, we disregard chunks labelled U due to its relative scarcity across annotators.

A necessary condition for any assumption of 'ground truth' in human-sourced annotations, is annotator agreement. Hence, we seek to quantify the amount of agreement between annotations. We denote with $\mathcal{L}$, $\mathcal{C}$ and N respectively the set of permitted labels, the set of audio chunks and the number of annotators. We asso-

[2]http://code.soundsoftware.ac.uk/projects/chime-home-dataset-annotation-and-baseline-evaluation-code/

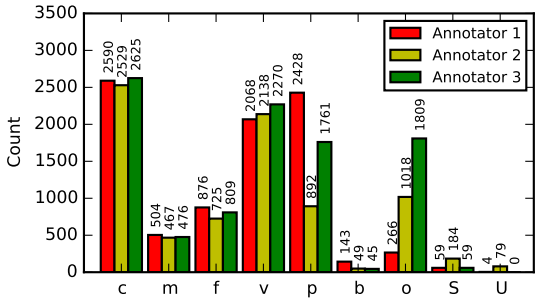

Figure 1: Annotator-wise histogram of label occurrences. See Table 1 for description of labels.

ciate a binary outcome variable $X_{n,l,c}$ with annotator n, label l, audio chunk c, where $n \in [1\,..\,N]$, $l \in \mathcal{L}$, $c \in \mathcal{C}$. We denote with $\mathcal{S}_{n,l}$ the set of audio chunks such that $X_{n,l,c} = \textit{True}$. As our measure of agreement, we adapt the Jaccard index, which has been widely applied as a measure of similarity between pairs of sets [20]. Given K sets $S_1, \ldots, S_K$, we define the generalised Jaccard index $J(S_1, \ldots, S_K)$ as

$$J(S_1, \ldots, S_K) = \frac{\left|\bigcap_{k=1}^{K} S_k\right|}{\left|\bigcup_{k=1}^{K} S_k\right|}. \quad (1)$$

Applied for the comparison of two sets, our definition is equivalent to the standard Jaccard index as given in [20]. Thus defined, as a measure of agreement between annotators $\{1, 2, 3\}$ about the presence of label l, we compute $J\left(\mathcal{S}_{1,l}, \mathcal{S}_{2,l}, \mathcal{S}_{3,l}\right)$.

Table 2 (a) displays agreement about label presence, across combinations of annotators. We observe strong agreement about labels `c`, `m`, `f`, `v` for all combinations of annotators (median 0.864). In contrast, we observe relatively low agreement about labels `p`, `b`, `o`, `S` (median 0.238). Comparing across pairs of annotators for labels `p`, `b`, `o`, we obtain relatively strong variation in agreement, with respective coefficients of variation in the range $[0.254, 0.747]$.

A possible explanation for these results is that annotators have different strategies for assigning labels to ambiguous sound classes: Relatively soft crashes e.g. those produced by cutlery might be associated with either label `o` or `p`, or even disregarded. We observe that if we create a single meta-label from labels `o`, `p`, Jaccard indices computed for the set of annotators $\{1, 2, 3\}$ increase to 0.490, suggesting relatively strong conflation of labels `o`, `p` across annotators. For meta-labels analogously created from label pairs {`o`, `b`}, {`o`, `S`}, we observe smaller gains in agreement and for annotator combination $\{2, 3\}$ alone, suggesting weaker label conflation in the latter cases.

That we observe relatively low agreement about label `S` suggests ambiguity about the perceptual salience of events in acoustic scenes. Similarly, an explanation for variation in average scores across labels `c`, `m`, `f` is that the speech sources in our dataset vary in their perceptual salience: Brief male utterances occurring in the background of the acoustic scene might be less perceptually salient and thus more difficult to identify, compared to long child utterances occurring in the foreground of the acoustic scene.

5. THE CHIME-HOME DATASET

To obtain greater confidence about 'ground truth' in annotations, we refine the dataset by retaining only those chunks where two or more annotators have assigned a given label, for all considered labels. Comprised of 44.5% of previously examined 4378 chunks, the resulting set of 1946 chunks $\mathcal{D} \subseteq \mathcal{C}$ is defined as

$$\mathcal{D} = \bigcap_{l \in \mathcal{L}} \left\{ c : \sum_{n=1}^{N} [X_{n,l,c} = \textit{True}] \geq 2, c \in \mathcal{C} \right\}. \quad (2)$$

Attempting to reduce ambiguity about label assignments by refining the dataset in this way, Table 2 (b) displays agreement about label presence, analogous to Table 2 (a). For the majority of labels, we observe that agreement scores increase, with relatively strong gains for the labels `p`, `b`, `S` across combinations of annotators. Examining the effect of merging labels `o`, `p` we observe gains in agreement for the combinations involving both annotators $1, 2$. The latter result suggests as a caveat that there remains an amount of conflation of labels `o`, `p` between annotators.

For each label l and for each chunk c in the refined dataset we subsequently perform a majority vote across annotators. In this way, we obtain a single multi-label annotation for each chunk, which we consider a 'ground-truth' assignment. Table 3 displays a matrix of label co-occurrences across chunks. As part of the CHiME-Home dataset, we make available the refined set of multi-label annotations (CHiME-Home-refined; 1946 associated chunks), in addition to raw multi-annotator data (CHiME-Home-raw; 4378 associated chunks)[3].

Label	Annotator combination			
	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
c	0.924	0.909	0.898	0.868
m	0.857	0.842	0.860	0.786
f	0.754	0.759	0.761	0.658
v	0.940	0.896	0.926	0.883
p	0.339	0.549	0.395	0.263
b	0.208	0.093	0.237	0.081
o	0.127	0.052	0.273	0.020
S	0.240	0.326	0.279	0.131
{o,p}	0.609	0.714	0.565	0.490
{o,b}	0.149	0.070	0.278	0.027
{o,S}	0.153	0.078	0.318	0.041

(a) Unfiltered dataset

Label	Annotator combination			
	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
c	0.964	0.966	0.953	0.942
m	0.937	0.920	0.891	0.874
f	0.883	0.917	0.858	0.829
v	0.973	0.976	0.983	0.966
p	0.514	0.910	0.476	0.450
b	0.947	0.474	0.526	0.474
o	0.102	0.163	0.850	0.058
S	0.963	0.963	1.000	0.963
{o,p}	0.744	0.901	0.761	0.703
{o,b}	0.145	0.179	0.834	0.079
{o,S}	0.162	0.219	0.861	0.121

(b) Refined dataset

Table 2: Agreement about label presence across combinations of annotators for (a) unfiltered dataset (4378 chunks) (b) refined dataset (1946 chunks). Symbols {`o`, `p`}, {`o`, `b`}, {`o`, `S`} denote meta-labels obtained by merging labels.

6. BASELINE RESULTS

To obtain a set of SSR baseline results, using the set of 1946 multi-label annotations we evaluate prediction performance for Gaussian mixture models (GMMs) estimated on sequences of MFCC features. For each label in the set {`c`, `m`, `f`, `v`, `p`, `b`, `o`}, we evaluate

[3] http://archive.org/details/chime-home

	c	m	f	v	p	b	o	S
c	**1214**	134	343	585	559	13	235	0
m		**174**	31	104	102	2	70	0
f			**409**	221	138	1	49	0
v				**1181**	225	1	145	0
p					**765**	7	295	0
b						**19**	0	0
o							**361**	0
S								**27**

Table 3: Matrix of label co-occurrences for refined dataset. See Table 1 for description of labels. Diagonal displays counts of individual label occurrences.

the performance of a binary classifier, considering as positive instances those chunks where the label of interest is present in the associated annotation string. We motivate our use of GMMs combined with MFCCs on the basis of previous work on ASC [21] and SSR [5]; further we note that GMMs with MFCCs have been proposed as an ASC baseline in the DCASE challenge [13]. We evaluate GMMs combined with MFCCs to obtain a set of baseline results using the CHiME-Home dataset.

Features: Based on a sampling rate of 48kHz, we compute magnitude spectra using a 1024-sample window and a 512-sample hop size. Subsequently, using the Librosa library [22] we obtain MFCCs based on 40 Mel-spaced filters centred between 0Hz and 24kHz. After discarding the energy coefficient, we retain the 13 first MFCCs for estimating GMMs. We standardise (zero mean, unit standard deviation) MFCCs with respect to training data.

Model estimation and prediction: To predict presence versus absence of a given label, we estimate a pair of GMMs, with a each GMM respectively pertaining to positive and negative instances in training data. We estimate full-covariance Gaussians using Expectation-Maximisation. We vary the number of Gaussians k using values in the set $\{1, 2, 4, 8\}$. For estimating GMMs, we use Scikit-learn[4] version 0.15.2. To predict presence versus absence of the label of interest for a given audio chunk, we compute the log-likelihood ratio of the associated feature vector sequence, with respect to estimated GMMs.

Evaluation: For each considered label, we use 10-fold cross validation to estimate GMMs and to score chunks. To account for any imbalance in the ratio of positive instances to negative instances, following [16] we quantify predictive accuracy by computing receiver operating characteristic (ROC) curves (cf. [23]) from obtained log-likelihood ratios. As a summary statistic, for each ROC curve we subsequently estimate the area under the curve; larger values indicate better classification.

Results: Figure 2 displays prediction accuracies across considered labels and in response to the number of GMM components k. Maximising performance with respect to k, we observe AUC values ranging from 0.80 (label f) to 0.98 (label v). Notably, maximally attained performance for predicting the presence of male speech or female speech labels is lower, compared to child speech. As suggested in Section 4, a possible explanation for such variation is that child speech is relatively perceptually salient in the recordings, compared to male and female speech: The evaluated combination of GMMs and MFCCs might fail to adequately represent less perceptually salient events such as those occurring the background of acoustic scenes, or events of limited duration. Our own informal listening suggests that male and female speech utterances are indeed of shorter duration, compared to child speech utterances. Similarly, we conjecture that a relatively high perceptual salience associated with video game/TV sources (label v) compared to human activity,

[4]http://scikit-learn.org/0.15/

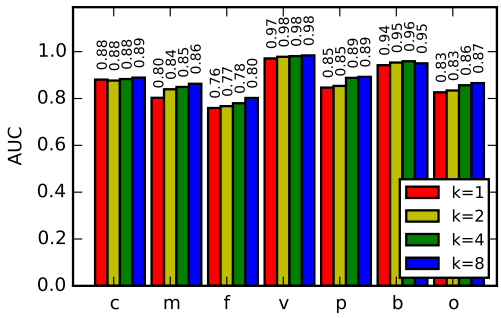

Figure 2: Label prediction accuracy quantified as area under curve (AUC) of receiver operating characteristic (ROC) curves, in response to number of GMM components k.

other identifiable sounds (respective labels p, o) gives rise to further performance differences between labels.

7. CONCLUSIONS AND FURTHER WORK

We have introduced the CHiME-Home dataset for investigating the task of in-home sound source recognition (SSR). The dataset comprises multi-label annotations of audio recordings, obtained with the aid of multiple annotators. Our results on annotator agreement inform our decision to refine the set of obtained annotations: Thus we provide a set of refined multi-label annotations, in addition to raw multi-annotator data.

The evaluated GMM baseline represents a starting point for more detailed investigations in SSR. Among approaches, we consider unsupervised feature learning techniques such as non-negative matrix factorisation [24] or sparse autoencoders [25] as potential means of separating sources contributing to the acoustic scene; source separation may prove advantageous for discriminating between events produced by the source of interest, versus 'background' events or noise. Further, we believe it will be instructive to explore the use of representing sequential structure, alternatively using convolutional feature learning techniques [26], or by learning sequential models of acoustic events produced by the sources which we seek to identify.

A limitation of the dataset is its restriction to a single domestic environment. Thus, the dataset does not permit evaluation of how models generalise to other environments. A possible approach to overcoming this limitation might involve transforming the existing recordings, similar to the approach proposed in [11]. While we may view the restriction to domestic environment recordings as a further limitation, we believe it is offset by the relevance of in-home SSR. Finally, while our choice of labels aims at voice and activity detection tasks, we concede that a more fine-grained annotation might, for example, allow distinction between human footsteps and falls.

A further potential application of our dataset involves using the obtained annotations to validate large-scale, crowdsourced datasets [27]. For future work, we aim to investigate in greater detail the amount of agreement between annotators, in particular examining possible label conflation between annotators. Finally, we note that our annotations might further be refined to obtain a taxonomy-based annotation of source activity [15], or segmentation at the level of individual acoustic events [14].

8. REFERENCES

[1] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, IEEE Press, 2006.

[2] S. Srinivasan and D. Wang, "Transforming binary uncertainties for robust speech recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2130–2140, 2007.

[3] G.-J. Jang and T.-W. Lee, "A maximum likelihood approach to single-channel source separation," *The Journal of Machine Learning Research*, vol. 4, pp. 1365–1392, 2003.

[4] M. D. Plumbley, S. A. Abdallah, J. P. Bello, M. E. Davies, G. Monti, and M. B. Sandler, "Automatic music transcription and audio source separation," *Cybernetics & Systems*, vol. 33, no. 6, pp. 603–627, 2002.

[5] B. Defréville, F. Pachet, C. Rosin, and P. Roy, "Automatic recognition of urban sound sources," in *Proc. 120th Audio Engineering Society Convention*, 2006.

[6] Albert S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, 1994.

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

[8] A. Temko, C. Nadeu, D. Macho, R. Malkin, C. Zieger, and M. Omologo, "Acoustic event detection and classification," in *Computers in the Human Interaction Loop*, pp. 61–73. Springer, 2009.

[9] R. H. Weber and R. Weber, *Internet of Things*, Springer, 2010.

[10] L. C. De Silva, C. Morikawa, and I. M. Petra, "State of the art of smart homes," *Engineering Applications of Artificial Intelligence*, vol. 25, no. 7, pp. 1313–1321, 2012.

[11] G. Lagrange, M. Lafay, M. Rossignol, E. Benetos, and A. Roebel, "An evaluation framework for event detection using a morphological model of acoustic scenes," *arXiv preprint arXiv:1502.00141*, 2015.

[12] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The CLEAR 2006 evaluation," in *Multimodal Technologies for Perception of Humans*, pp. 1–44. Springer, 2007.

[13] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: an IEEE AASP challenge," in *Proc. 10th IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[14] M. W. W. van Grootel, T. C. Andringa, and J. D. Krijnders, "DARES-G1: Database of annotated real-world everyday sounds," in *Proc. NAG/DAGA Intern. Conf. Acoustics*, 2009.

[15] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 24th ACM Intern. Conf. Multimedia*, 2014, pp. 1041–1044.

[16] D. Stowell and M. D. Plumbley, "An open dataset for research on audio field recording archives: freefield1010," in *Proc. Audio Engineering Society 53rd Intern. Conf.: Semantic Audio*, 2014.

[17] H. Christensen, J. Barker, N. Ma, and P. D. Green, "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," in *Proc. 11th INTERSPEECH Conf.*, 2010, pp. 1918–1921.

[18] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. D. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013.

[19] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[20] B. Clarke, E. Fokoue, and H. H. Zhang, *Principles and Theory for Data Mining and Machine Learning*, Springer Science & Business Media, 2009.

[21] 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," *Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[22] B. McFee, M. McVicar, C. Raffel, D. Liang, and D. Repetto, "librosa: v0.3.1," 2014, DOI 10.5281/zenodo.12714.

[23] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[24] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[25] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. 14th Intern. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2011, pp. 215–223.

[26] P. D. O'Grady and B. A. Pearlmutter, "Convolutive non-negative matrix factorisation with a sparseness constraint," in *Proc. 16th IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, 2006, pp. 427–432.

[27] M. Cooke, J. Barker, M. L. G. Lecumberri, and K. Wasilewski, "Crowdsourcing for word recognition in noise," in *Proc. 12th INTERSPEECH Conf.*, 2011, pp. 3049–3052.