SEMI-SUPERVISED LEARNING HELPS IN SOUND EVENT CLASSIFICATION

Zixing Zhang and Björn Schuller

Institute for Human-Machine Communication, Technische Universität München, Germany
(zixing.zhang|schuller)@tum.de

ABSTRACT

We investigate the suitability of semi-supervised learning in sound event classification on a large database of 17 k sound clips. Seven categories are chosen based on the findsounds.com schema: animals, people, nature, vehicles, noisemakers, office, and musical instruments. Our results show that adding unlabelled sound event data to the training set based on sufficient classifier confidence level after its automatic labelling level can significantly enhance classification performance. Furthermore, combined with optimal re-sampling of originally labelled instances and iteratively learning in semi-supervised manner, the expected gain can reach approximately half the one achieved by using the originally manually labelled data. Overall, maximum performance of 71.7 % can be reported for the automatic classification of sound in a large-scale archive.

Index Terms— Sound Event Classification, Semi-supervised Learning

1. INTRODUCTION

Sound event classification is attracting a growing attention recently in the field of acoustic signal analysis. Not only because it bears great interest for application in multimedia search based on sound with the rapid growth of multimedia data available on the web, but as it is also one of the most important key components to analyse environments, e. g., in surveillance [1, 2], monitoring of people in need of care, or detecting, localising, tracking and classifying sources of military interest in real time [3]. Obviously, there is also great benefit for humanoid and general robots, such as the one introduced in [4] for kitchen tasks, to better understand their acoustic environment. Finally, there is hope to better recognise and enhance speech and music, once the sound type of disturbance can be identified. Yet, most of the previous research focuses on comparatively small scale and often prototypical databases (e. g., as in [5]). In [4, 6], however, roughly 6 000 and 7 000 instances are investigated, respectively, for sound event classification. In table 1 a selective overview on related other work is given. – obviously, we cannot take into account all existing literature in this domain.

In this paper, we will focus on sound events classification in a large scale database, covering sound classes that reach from nature (i. e., nature, animals) over human beings (i. e., people) to artificial sounds (i. e., office, musical instruments, noisemakers, and vehicles). Furthermore, there is an ever-lasting belief in pattern recognition that 'there is no data like more data'. Compared to automatic speech recognition where many corpora comprise hundreds of hours of transcribed speech, databases annotated in sound event classification are still sparse as shown above. Semi-supervised and unsupervised learning can be a promising approach to remedy the issue of this data sparsity:

Table 1: Selective overview on previous research on sound event classification. Abbreviations are inst.: instance, I: isolated events, S: stream, F: frame level, C: per clip, and for feature types, E: Energy, FFBE: frequency-filtered band energies, ICA: Independent Component Analysis, MFCC: Mel-frequency cepstral coefficients, MP: matching pursuit, STE: subband temporal envelopes.

Work	***# Clips***	***# Classes***	***Type***	***Unit***	**Features**	**Domain**
[2]	134	5	S	F	MFCC+E	surveill.
[7]	3 000	14	I	C	MP+MFCC	environ.
[8]	115/ 10 500	7/ 105	I	F	MFCC	health care
[9]	918	12	I	F	MFCC	meeting
[10]	1 030/ 729/172	16	I/S	F	MFCC+E, FFBE	meeting
[4]	5 992	21	I	F	ICA	kitchen
[11]	732	8	I	F	MPEG-7	urban
[6]	705	10	I	C	STE	office & canteen

Assuming sufficiently robust automatic sound event classification engines, unlabelled data can be classified and integrated into an iterative re-training process. Such unlabelled data is practicably available in 'infinite' amount: One could not only profit from recording real life audio streams typically filled with various kinds and huge number of sound events [12], but add data from the web. Notably, studies dealing with semi-supervised adaptation of acoustic and language models in automatic speech recognition [13, 14] suggest that addition of unlabelled data in training is competitive with labelled data, even more so if one considers the enormous efforts usually required for manual annotation of speech data. Further, in speech recognition, recent real-life studies as the Google Voice Search show that semi-supervised learning has in fact already turned into common practice. As a rule of thumb, the need of roughly ten times the amount of unlabelled data is named there in comparison to labelled data in order to obtain the same gain as with labelled data. We thus investigate semi-supervised learning to improve a sound event classifier in continuation of our related efforts in semi-supervised emotion recognition [15]. The paper is structured as follows: in Section 2 we introduce the FINDSOUNDS sound event database that we use to classify real life sounds in seven categories. Then, we describe our brute force extraction of features and the classifier set-up in Section 3. In Section 4, we investigate the performance of semi-supervised learning before concluding in Section 5.

Zixing Zhang acknowledges funding from the Chinese Research Council.

Table 2: Quantitative description of the FINDSOUNDS database.

Category	# Subsets	# Clips	Duration [h]
People	45	2 540	2 h 9 min
Animals+Birds	85	2 841	2 h 42 m
Nature	19	937	1 h 17 min
Vehicles	34	2 166	2 h 47 min
Noisemakers	13	2 010	1 h 56 min
Office	18	1 769	1 h 01 min
Musical Instruments	62	4 674	3 h 49 min
Total	**276**	**16 937**	**15 h 41 min**

2. THE FINDSOUNDS DATABASE OF SOUND EVENTS

For the modelling and recognition of sound events, we collect audio data from the web via the FindSounds site[1] which provides a large amount and variety of sound events from real life. In addition, these sounds are readily categorised, and we stick with this schema. However, in order to better suit for our experiment, we disregard the categories without sufficient audio instances and cluster birds with animals, which leaves seven categories of common sound events in real life from sixteen:

People: produced by 45 different human behaviours, such as biting, baby's crying, coughing, laughing, moaning, kissing, etc.
Animals (including birds): come from 69 different non-bird animals, such as cat, frog, bear, lamb, blackbird, etc., and 16 kinds of birds.
Nature: includes 19 kinds of original sounds from nature environment, for instance, earthquake, ocean waves, flame, rain, wind, etc.
Vehicles: take 34 different types of vehicles and their behaviours into account, like motorcycling, braking, helicopter, closing door, etc.
Noisemakers: are composed of 13 various events in this domain such as alarm, bell, whistle, horn, etc.
Office: includes original office space sound events including keyboard typing, printing, telephoning, mouse clicking, etc.
Musical **Instruments**: stem from 62 various musical instruments like bass, drum, synthesiser, etc.

We converted all of the audio files into raw 16 bit encoding, mono-channel, and 16 kHz sampling rate since various formats and rates are used in the original versions as were retrieved from the web. Each of the sound clips generally lasts between 1 s to 10 s. Finally, roughly 15 hours of recording time and 16 937 instances were obtained in total, covering 276 aspects of real life sound events. Further details on the distribution of instances and total play time per category are summarised in Table 2. Owing to the origin of the sounds and classification scheme, this set will be referred to as FINDSOUNDS database.

3. ACOUSTIC FEATURES AND CLASSIFIER

Using our open-source openEAR [16] toolkit's feature extraction back-end openSMILE (Speech and Music Interpretation by Large space Extraction) [16], we extract the 'AVEC' set that consists of 1 941 features, composed of 25 energy and spectral related low-level descriptors (LLD) x 42 functionals, 6 voicing related LLD x 32 functionals, 25 delta coefficients of the energy/spectral LLD x 23 functionals, 6 delta coefficients of the voicing related LLD x 19 functionals, and 10 voiced/unvoiced duration features. Details for the functionals and LLDs are given in tables 3 and 4, respectively. The set of LLD covers a standard range of commonly used features in audio signal analysis and classification. The functional set has been based on similar sets, such as the one used for the INTERSPEECH 2011 Speaker State Challenge [17], but has been carefully reduced to avoid LLD/functional combinations that produce values which are constant, contain very little information and/or high amount of noise.

Table 3: Set of all 42 functionals. [1]not applied to delta coefficient contours. [2]for delta coefficients the mean of only positive values is applied, otherwise the arithmetic mean is applied. [3]not applied to voicing related LLD.

Statistical functionals (23)
(positive[2]) arithmetic mean, root quadratic mean standard deviation, flatness, skewness, kurtosis quartiles, inter-quartile ranges 1 %, 99 % percentile, percentile range 1 %–99 % percentage of frames contour is above: min + 25%, 50%, and 90 % of the range percentage of frames contour is rising max, mean, min segment length[3], std. dev. segment length[3]
Regression functionals[1] (4)
linear regression slope, and corresponding approximation error quadratic regression coefficient a, and approximation error
Local minima/maxima related functionals[1] (9)
mean and standard deviation of rising and falling slopes (minimum to maximum), mean and standard deviation of inter maxima distances amplitude mean of maxima, amplitude mean of minima, amplitude range of maxima
Other[1,3] (6)
LP gain, LPC 1-5

Table 4: Set of 31 low-level descriptors.

Energy & Spectral (25)
loudness (auditory model based), zero crossing rate, energy in bands from 250 – 650 Hz, 1 kHz – 4 kHz, 25 %, 50 %, 75 %, and 90 % spectral roll-off points, MFCC 1–10, spectral flux, entropy, spectral variance, skewness, kurtosis, psychoacoustic sharpness, harmonicity
Voicing related (6)
F_0 (Sub-harmonic summation (SHS) followed by Viterbi smoothing), probability of voicing, shimmer (local), jitter, 'jitter of jitter', logarithmic Harmonics-to-Noise Ratio (logHNR)

As classifier, we use Random Forests which can provide very good generalisation properties and are able to well cope with large feature spaces, as each tree in a forest fulfils an implicit information gain based feature ranking. In addition, feature sub-spaces are randomly assigned to the trees of the forests. Thus, for representative results in our experiments, we chose Random Forests with 30 trees, and 150 random features for each tree in a forest. For further reproducibility besides using our open source feature extractor and the FINDSOUNDS database that can be retrieved from the web, we use the classifier implementation provided by the Weka toolkit [18].

[1] http://www.findsounds.com/types.html – accessed on 25 July 2011.

Table 5: Confusion matrix for seven categories of sound event classification using original labels for both training folds F1 and F2 (cf. line 'supervised' in table 8). FINDSOUNDS database.

	Classified as						
Truth [#]	***People***	***Animals***	***Nature***	***Vehicles***	***Noisem.***	***Office***	***Instrum.***
People	**564**	153	11	26	17	25	50
Animals	126	**717**	7	35	23	20	18
Nature	18	35	**157**	42	44	10	6
Vehicles	37	37	26	**476**	86	15	45
Noisem.	22	43	36	77	**372**	72	48
Office	29	37	1	16	111	**364**	31
Instrum.	32	33	6	31	47	16	**1395**
Confusions	264	338	87	227	328	158	198

4. EXPERIMENTS

4.1. Sound Event Classification

As preferred evaluation measure we employ average recall among classes without weighting by the number of instances, i. e., unweighted accuracy (UA) as was suggested in [17]. The reason for this choice is the imbalance among the classes. Further, we partly additionally provide 'normal' (weighted) accuracy (WA), precision, and F_1-measure in the summary of optimal configuration. There, we also provide the average area under the receiver operating characteristic (ROC), which plots the true positive rate over the false positive rate achieved by a binary classifier for each class vs. the remaining ones. The area under curve (AUC) generally reaches from 0.5 (random guessing) to 1.0, equal to the whole graph area.

The experiments were performed based on random partitioning of the FINDSOUNDS database into three stratified folds to allow for a partitioning into two training and one completely disjunct testing set. This is needed, as the first fold (F1, 5 646 instances) is used with its original manually assigned labels for training, each. The second fold (F2, 5 646 instances) is used either without its original labels ($F2_U$, 'unlabelled') or with these labels (F2) in order to be able to compare using it in semi-supervised or supervised manner for training. Finally, the third folder (5 645 instances) is always used for testing.

Table 5 first gives the confusion matrix for seven categories of sound event classification using the original labels from fold 1 and 2 in the training. This resembles the highest accuracy in our experiments and shall give a good overview on arising confusions in 'best case'. It can be seen that the sounds from people and animals tend to get confused just as sounds from vehicles, noisemakers and in the office environment do – this is well in line with expectation by common sense.

4.2. Semi-supervised Learning in Sound Event Classification

From now on, to next investigate the potential of semi-supervised learning for sound event classification, we first take fold 1 with its original labels for training (F1, as shown in Table 8) and fold 3 for testing as a baseline reference, i. e., in this baseline fold 2 is ignored. Second, we agglomerate for training fold 1 with its original manually assigned labels and fold 2 without its original labels but labelled by the system trained on fold 1 for semi-supervised adaptation in diverse strategies (F1 + $F2_U$, as shown, for example, in Table 8), and again test on fold 3. Finally, as a reference for supervised learning, we consider agglomerating fold 1 and fold 2 using the original labels of both for training (F1 + F2, as shown in Table 8), and also use fold 3 for testing.

Table 6: Unweighted accuracy of semi-supervised learning with different confidence levels combined with/-out re-sampled manually labelled data. 2·F1: re-sampling (doubling up) fold 1 instances; $F2_U^1$: from fold 2 select the instances with a confidence level above 0.45, 0.5, 0.6, 0.7, 0.8, separately after classification based on fold 1. FINDSOUNDS database.

	Confidence Level				
UA [%]	$> \mathbf{0.45}$	$> \mathbf{0.5}$	$> \mathbf{0.6}$	$> \mathbf{0.7}$	$> \mathbf{0.8}$
F1+F2^1_U	61.3	61.5	61.5	61.6	62.1
2·F1+F2^1_U	62.0	62.1	61.5	**63.1**	62.5

Table 7: Unweighted accuracy of iterating semi-supervised learning with confidence values above 0.7 and 0.8 combined with re-sampling or non-re-sampling manually labelled data. 2·F1: re-sampling (doubling up) fold 1 instances; $F2_U^1$, $F2_U^2$, $F2_U^3$: the first, second, and third time of iterating semi-supervised learning. FINDSOUNDS database.

	Confidence Level			
	$> \mathbf{0.7}$		$> \mathbf{0.8}$	
UA [%]	**F1**	**2·F1**	**F1**	**2·F1**
F2^1_U	61.6	63.1	62.1	62.5
F2^2_U	62.0	62.2	63.0	62.6
F2^3_U	62.0	61.7	62.6	**63.2**

As to the semi-supervised learning, we take the confidence of the classifier in five levels into account (> 0.45, > 0.5, > 0.6, > 0.7, and > 0.8) to enhance the influence of correctly labelled data in the semi-supervised labelling process and suppress falsely labelled data. Furthermore, we consider two additional strategies: re-sampling of the originally labelled data and repeatedly iterating the semi-supervised learning process. In Table 6, it can be seen that, with increasing confidence level, the unweighted accuracy of training on agglomerated non-re-sampled labelled data (1·F1) and conditionally selected instances from semi-supervised learning ($F2_U^1$) gradually increases from 61.3 % to 62.1 %. To emphasise more on the manually labelled data for the classifier, the table next shows re-sampling by copying (2·F1). Compared to the former non-re-sampled strategy, the performance is improved. The most impressive gain is seen when the confidence level is higher than 0.7, achieving UA of 63.1 %. The details for this accuracy are also shown in Table 8. With the iterating strategy, i. e., repeatedly re-labelling the unlabelled data fold 2 using fold 2 in training with labels from the last iteration, we only took into acount confidence levels higher than 0.7 according to the former experiment. Table 7 shows the UA of up to three times iterating the semi-supervised learning process. When non-re-sampling (1·F1), a gain is also obtained (62.0 % vs. 61.6 % UA for confidence level > 0.7, and 63.0 % vs. 62.1 % UA for confidence level > 0.8). However, we notice that iterating benefit is limited, as UA partly begins to decrease after the third iteration. Even a larger iteration number

Table 8: Classification evaluation on seven sound categories with un-/supervised learning. Rec.: recall, U/WA: un-/weighted accuracy, Prec.: precision, AUC: Area Under receiver operating Characteristic. FINDSOUNDS database.

	[%]	UA	WA	Rec. People	Rec. Animals	Rec. Nature	Rec. Vehicles	Rec. Noisem.	Rec. Office	Rec. Instrum.	Prec.	F_1	AUC
supervised (baseline):	F1	61.1	67.0	61.7	68.2	39.7	60.2	52.7	57.9	87.2	66.9	66.7	91.8
semi-supervised:	2·F1+F2$_U^1$	**63.1**	**68.5**	61.7	72.5	47.4	61.8	51.9	58.4	87.9	68.3	68.1	91.9
supervised (all):	F1+F2	**66.5**	**71.7**	66.7	75.8	50.3	65.9	55.5	61.8	89.4	71.6	71.5	92.7

will not lead to better results (not shown in the table). Finally, we combined the re-sampling and iterating strategies striving to exploit the advantages of both. From the line 2·F1 in Table 7, it can be seen that re-sampling outperforms the baseline setting in 4 out of 6 cases.

In conclusion, the comparison of the baseline reference (F1), most efficient semi-supervised learning (2·F1+F2$_U^1$), and supervised learning (F1+F2) are shown in detail in Table 8. As expected, the best average result is obtained when using the original labels of the data partitions fold 1 and fold 2 for training (66.5 % UA). Yet, this also clearly shows that the semi-supervised learning significantly (one-sided z-test, $p < 0.05$) improves the performance of sound event classification with boost in UA of 2 % absolute over not using fold 2 at all which is almost half the gain achieved by supervised training (5.4 %). In our case, the nature class being the most sparse, benefited most from semi-supervised learning.

5. CONCLUSION

We investigated the potential of semi-supervised learning in a large scale sound event classification task. The results show that adding unlabelled data with high confidence level to the training data can enhance recognition performance. Furthermore, re-sampling originally labelled data and iterating the semi-supervised learning process both boosted classification accuracy in our experiments by strengthening the weight of the originally labelled data, while the latter strategy gradually increases the semi-supervised learning advantage. However, the results are – as one would expect – below the gain that can be expected when adding labelled data. Yet, the fact that manual labelling of sound event data is highly costly while large amounts of sound event data per se are publicly available makes consideration of semi-supervised learning a promising approach in future machine-based sound analysis.

Our future efforts will continue to focus on agglomerating huge amounts of unlabelled sound event data and its application in analysis of real-life sound streams in combination with source separation. Further, active learning can be involved to decide on which new instances are promising before classifying them and adding them to the learning material. Finally, newly labelled data could be associated with multiple classes in a fuzzy manner based on the confidence measure, rather than with only one as was done here.

6. REFERENCES

[1] A. Temko, R. Malkin, C. Zieger, D. Macho, and C. Nadeu, "Acoustic event detection and classification in smart-room environments: Evaluation of chil project systems," in *Proc. IV Biennial Workshop on Speech Technology*, Zaragoza, 2006, pp. 1–6.

[2] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Proc. ICME*, Amsterdam, 2005, pp. 1306–1309.

[3] B. G. Ferguson and K. W. Lo, "Acoustic cueing for surveillance and security applications," in *Proc. SPIE*, Orlando, FL, 2006.

[4] F. Kraft, R. Malkin, T. Schaaf, and A. Waibel, "Temporal ICA for Classification of Acoustic Events in a Kitchen Environment," in *Proc. INTERSPEECH*, Lisbon, 2005, pp. 2689–2692.

[5] C. Zieger and M. Omologo, "Acoustic event classification using a distributed microphone network with a GMM/SVM combined algorithm," in *Proc. INTERSPEECH*, Brisbane, 2008, pp. 115–118.

[6] T. H. Dat and H. Li, "Probabilistic distance svm with hellinger-exponential kernel for sound event classification," in *Proc. ICASSP*, Prague, 2011, pp. 2272–2275.

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

[8] Y. Peng, C. Lin, M. Sun, and K. Tsai, "Healthcare audio event classification using hidden markov models and hierarchical hidden markov models," in *Proc. ICME*, Piscataway, NJ, 2009, pp. 1218–1221.

[9] T. Heittola and A. Klapuri, "TUT acoustic event detection system 2007," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007*, Berlin, Heidelberg, 2008, pp. 364–370, Springer.

[10] A. Temko and C. Nadeu, "Classification of acoustic events using SVM-based clustering schemes," *Pattern Recognition*, vol. 39, pp. 682–694, 2006.

[11] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Automatic recognition of urban environmental sound events," in *Proc. CIP 2008, Eurasip*, Santorini, 2008, pp. 110–113.

[12] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *Proc. EUSIPCO*, Aalborg, 2010.

[13] D. Hakkani-Tur, G. Tur, M. Rahim, and G. Riccardi, "Unsupervised and active learning in automatic speech recognition for call classification," in *Proc. ICASSP*, Montreal, 2004, pp. 429–432.

[14] G. Tur and A. Stolcke, "Unsupervised Language Model Adaptation for Meeting Recognition," in *Proc. of ICASSP*, Honolulu, HY, 2007, pp. 173–176.

[15] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised Learning in Cross-Corpus Acoustic Emotion Recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Big Island, HY, 2011.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia (MM)*, Florence, 2010, pp. 1459–1462.

[17] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Proc. INTERSPEECH 2011*, Florence, 2011, pp. 3201–3204, ISCA.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, pp. 10–18, 2009.