INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Robust Laughter Detection in Noisy Environments

Jon Gillick[⋆], *Wesley Deng*[†], *Kimiko Ryokai*[⋆], *David Bamman*[⋆]

[⋆]School of Information, University of California, Berkeley, CA, USA
[†]EECS, University of California, Berkeley, CA, USA
{jongillick,wesley1016,kimiko,dbamman} @berkeley.edu

Abstract

We investigate the problem of automatically identifying and extracting laughter from audio files in noisy environments. We conduct an empirical evaluation of several machine learning models using audio data of varying sound quality, finding that while previously published methods work relatively well in controlled environments, performance drops precipitously in real-world settings with background noise. In the process, we contribute a new dataset of laughter annotations on top of the existing AudioSet corpus, with precise segmentations for the start and end points of each laugh, and we present a new approach to laughter detection that performs comparatively well in uncontrolled environments. We discuss the utility of our approach as well as the importance of understanding the variability of model performance in a range of real-world testing environments.

Index Terms: Laughter, Annotation, Sound Event Detection, Paralinguistics, Nonverbal Communication

1. Introduction

Laughter is a fundamental human expression. It pervades our everyday auditory experiences, whether at a family dinner, a work meeting, a comedy club, or in conversation with a digital assistant like Alexa or Siri. Laughter can convey a range of emotions, but for the most part, it is associated with positive affect such as joy, happiness, amusement, and relief. While laughter is ubiquitous, we do not often know when, why, how much, or with whom we laugh.

Still, recognizing, measuring, and analyzing such laughter plays a role within a diverse range of research communities. In speech and language processing, laughter is employed in systems for speaker diarization [1], for affect recognition [2], and as a signal for studying social, cultural, linguistic, and rhetorical communication styles [3]. In developmental psychology, when and how laughter takes place can help us better understand the ways in which children with Autism Spectrum Disorder, Down syndrome, or Angelman syndrome relate to the people around them [4, 5]. In computer graphics, laughter provides a mechanism for editors or viewers to navigate through existing media in order to find climactic scenes or special moments [6]. In HCI, recent work argues for laughter as a rich source for reflecting on how we feel about or understand our everyday life activities [7, 8].

Despite the significance of laughter as an object of study in many different contexts across disciplines, recent research suggests that published approaches to automatically detecting laughter from audio perform poorly in real world acoustic environments. In particular, Ryokai et al. [7], who applied automatic laughter detection to mobile phone recordings made by research participants during their day-to-day activities, found that while the detection worked well in quiet environments, noisier recordings yielded many false positives that contained only background noise. In some recordings, these false positives outweighed the correctly identified laughter by as much as a factor of 10, calling into question the practical utility of existing methods for laughter detection.

This evidence suggests that previous research on automatic laughter detection may not account for the noisy, in-the-wild environments in which laughter often happens. Indeed, most previous work on laughter detection evaluates approaches using cleanly recorded audio in quiet settings, typically employing the ICSI meetings database [9], which was recorded in a conference room with a microphone for each speaker, or the Switchboard [10] or SSPNet [11] corpora of telephone conversations. More often than not, however, laughter takes place in relatively noisy places where multiple people gather and talk simultaneously among various other noises (e.g., being at a restaurant or having a gathering with others). It is rare that a single person laughs in a completely silent environment by themselves. Perhaps because of the nature of these datasets, previous work reports that traditional MFCC-based audio features may be sufficient for laughter detection [12]; the evidence from Ryokai et al. [7], however, which finds that existing models produce more false positives than expected when applied in real world scenarios, suggests that more investigation is needed.

This observation leads to the focus of this work: an empirical comparison of the quality of laughter detection methods in both controlled and uncontrolled environments, and, in the process, the development of a robust method based on ResNet that performs comparatively well in the presence of background noise. Our primary contributions in this paper include the following:

- We present a new dataset of laughter annotations on top of the existing AudioSet [13] corpus of in-the-wild YouTube videos, with precise segmentations for the start and end points of each laugh.
- We implement several different machine learning models for laughter detection and conduct an empirical evaluation using both clean and noisy audio data.
- We examine laughter as a case study of the challenges involved in recognizing variable-length events when we face the common yet challenging scenario in which strongly labeled in-domain training data is not available.

Our annotations, code, and trained models are publicly available at: https://github.com/jrgillick/laughter-detection.

2. Related work

Researchers in the speech recognition community have explored methods for laughter detection in audio, often in support of systems for speaker diarization, which attempt to automatically track who spoke when in a multi-person conversation [1]. Published approaches to this task use traditional MFCC and pitch-based features, with the most recent work employing

http://dx.doi.org/10.21437/Interspeech.2021-353

fully-connected or convolutional neural networks on top of the MFCC's [12, 14]. In this body of work, laughter is treated as a variable length event, with performance metrics computed at the frame level. This is equivalent to segment-based metrics used in the Sound Event Detection literature with short segments that span a single 10-25 millisecond frame[15]. Though work over the last few years on many audio detection problems has demonstrated that features learned from spectrograms using deep convolutional networks, paired with data augmentation, significantly outperform previous approaches based on MFCC's [16, 17, 18], the audio classification methods established in recent years have yet to be systematically applied to or evaluated on the problem of laughter detection.

Although laughter is not a primary focus of work on audio event detection more broadly, some studies consider laughter as one of a larger list of categories [16, 19]. These works do employ the most recent machine learning methods, and they occasionally evaluate in noisy acoustic environments, but unlike the laughter detection literature, they do not attempt to identify laughter in a continuous stream of audio as is typical in a conversation. Instead, they perform multi-class classification on predefined temporal spans, with all events assumed to have a fixed duration [16, 19]. AudioSet [13], the dataset used most commonly for this purpose, provides annotations that only denote whether or not laughter occurs within a 10-second clip but do not specify more precisely when it occurs. For the applications that motivate this research, however, 10 seconds is not granular enough to be useful in practice, so more precisely annotated data is desirable.

Unlike previous work, we recognize laughter in noisy and uncontrolled real-life scenes, while at the same time detecting variable-length event boundaries with fine granularity. This setting for the detection problem is more challenging but is important for real-world use cases.

Research within HCI has also explored techniques for detecting laughter or related emotional signals through other modalities besides audio. Laughter detection might also be attempted, for example, through cameras trained on facial expressions [20, 21] or through wearable technology integrated into clothing in order to track body movements [22, 23, 24]. These kinds of sensing, however, have proved to be more intrusive and less practical than passive audio-based sensing using the microphones already embedded in our mobile phones [7].

3. Data

A primary contribution of this work is evaluating the accuracy of methods for automatic laughter detection in both controlled and uncontrolled environments; to do so, we assess accuracy on two datasets: existing data collected in a controlled environment (the Switchboard corpus), and naturalistic audio from YouTube, which we annotate with precise segmentations (AudioSet).

3.1. Switchboard

The Switchboard corpus of telephone conversations contains a total of about 260 hours of speech from 543 total speakers [10], captured in 2435 five-minute conversations. In Switchboard, conversations are finely segmented with annotations for the beginning and end time of each word, as well as for each occurrence of laughter. Laughter is annotated in two ways: either isolated between words, or combined with a word, as in the below examples:

- It's going to be really good [*laughter*] [*silence*]...
- Well [*laughter-i*] [*laughter-mean*] some of these guys...

The form of these annotations emphasizes the fundamental difficulty of recognizing laughter—naturally occurring laughter shows up not only as isolated events, but also intermixed with our speech (in preliminary experiments, we explored training models using both isolated and mixed laughter, finding that the isolated laughter alone performs best).

Of 2435 total conversations, we partition the dataset into 2159 for training, 119 for development, and 157 for testing, using the same splits as Ryokai et al. [7]. Aggregate statistics on the timing information across these partitions is summarized in Table 1.

As with most large annotated datasets for speech recognition, the data collection process focuses on quiet recording environments with close microphones, minimal background noise, and a known set of speakers. Additionally, the conversations here are limited to two people at a time. We treat this as an example of a highly controlled environment that allows us to determine the accuracy of laughter detection methods in idealized settings.

3.2. Audioset Annotation

In order to evaluate models for laughter detection in more realistic recording conditions, we turn to AudioSet, a collection of YouTube videos recorded in a variety of in-the-wild settings that have been hand-labeled with more than 500 categories of sounds [13]. Among those clips, each of which lasts 10 seconds, 5696 are annotated as containing laughter. In this setting, the voices of the people laughing are unknown, and the laughter may take place far from the microphone or in the presence of background noise. The main drawback of AudioSet for laughter detection is that the data are weakly labeled: we know that laughter occurs within the 10-second window, but we do not know where. Switchboard, on the other hand, is referred to as strongly labeled data because of the precise segmentation of laughter events (to within a fraction of a second) in the annotations. Because of weak labels in AudioSet, much sound event detection work on AudioSet only operates at a granularity of 10 seconds. For this reason, AudioSet has not previously been used to evaluate systems for laughter detection.

To leverage the data in AudioSet for evaluation, we selected a random sample of 1000 clips from those that were tagged as containing laughter and manually annotated the start and end times of laughter events within those clips in a format similar to that of Switchboard (though without transcribing any speech). This new test set—a core contribution of our work—contains 148 minutes of audio, including 58 minutes of laughter and 1492 distinct laughter events.

To measure the accuracy of our annotations, a second contributor also annotated a sample of 10% of this data, finding a 95.2% per-frame inter-annotator agreement rate. Because of the ambiguity inherent in what sounds should be defined as laughter, the agreement rate is not perfect; still, it suggests a ceiling at which we can reasonably expect laughter detection models to perform.

3.3. Negative Examples and Class Balance

In order to assess the effects of background noise, we need to include sufficient negative examples, or distractors, in our test set; these negative examples take the form of audio that we know does not contain the sound of laughter. In Switchboard, the negative examples consist only of speech or silence, but in AudioSet, negative examples contain all sorts of environmental sounds.

The 10-second clips annotated in our AudioSet test data

consist of 39% laughter (by time), with the other 61% containing the sounds immediately before or after the sound of laughter. To better measure false positive rates in different noisy environments, we supplement these 1000 clips that do contain laughter with an additional 1000 randomly chosen clips from AudioSet that do not contain any laughter; we report metrics averaged over all 2000 of these clips. After including the negative examples, the test set contains about 20% laughter.

To fairly compare laughter detection performance between the environments represented in Switchboard and AudioSet, we subsample the Switchboard conversations in the test set so that the proportion of laughter (the class balance) in both test sets is the same. We do this by first choosing context windows around the annotated laughter events in Switchboard of appropriate length so as to yield a mix containing 39% laughter, after which we randomly choose 10 second clips from Switchboard that do not contain laughter until we have doubled the size of the Switchboard test set. Table 1 summarizes the annotated data included in Switchboard and AudioSet.

Table 1: *Laughter statistics annotated in Switchboard (SWB) and AudioSet (AS) datasets.*

Data Partition	Laugh Events	Laugh Minutes
SWB: Train	22,490	289
SWB: Validation	876	9
SWB: Test	1,119	12
AS: Test	1,492	58

3.4. Training Data

Although our newly annotated *test* data is sourced from AudioSet, we do not have access to similarly annotated *training* data from AudioSet; rather, we are interested in exploring the common scenario in which there is no strongly labeled in-domain training data available, which could apply to any other sound category in AudioSet. Given this limitation, we can train on either: **(1)** clean (out of domain) but strongly labeled data (SLD) from Switchboard, or **(2)** in-the-wild (in-domain) but weakly labeled data (WLD) from AudioSet. Setting aside the 1000 clips in the test data that we have annotated, we can extrapolate from Table 1 to estimate that the remaining 4696 clips in AudioSet contain laughter sounds totaling a comparable amount to that in the Switchboard training set; this means that we have roughly the same amount of training data in each scenario.

Working with weakly labeled training data is an open area of research that remains difficult in most settings, but largely because of the existence of weak labels in influential datasets like AudioSet, it has become common for sound event detection systems to still attempt predictions over shorter time-spans (e.g. 1 second rather than 10 seconds). Typically this is done by splitting the 10 seconds into smaller chunks and training with mean or max pooling, or simply by applying the weak labels directly to these shorter chunks and tolerating some noise in the training labels [25].

In experiments, we explore both options: first, training on Switchboard so that we can take advantage of strongly labeled (finely segmented) data, and second, training on AudioSet with noisy labels so that we can take advantage of in-domain data. Besides these tradeoffs regarding the data and labels, we keep the rest of training procedure the same, using the same model architectures, data augmentations, and pre-processing choices.

4. Models

To explore the performance of different models on the datasets described above, we compare three different methods for automatic laughter detection: a baseline feed-forward neural network with engineered features, a ResNet model on spectogram data, and a ResNet model augmented with data transformations. Following previous work [7, 14], we make predictions centered at every frame using a sliding window that contains a total of 1 second of surrounding audio as context. We use the default frame rate of 43.1 fps as implemented in the Librosa library [26], which we found to perform slightly better than the 100 fps used in previous models for laughter detection [7]. Following the conventions in the Sound Event Detection literature, if a center frame falls within the annotated event boundaries, it is considered a positive example; if the center frame is outside the boundaries, it is considered a negative example [15]. Because Switchboard's audio is provided at a sample rate of 8000hz, we downsample the AudioSet data to 8000hz for consistency. All models are trained using Pytorch [27] for 100,000 steps with a batch size of 32.

4.1. Baseline

As a baseline, we use the feed-forward neural network laughter detection model from Ryokai et al. [7] This model is the most recently published example that is representative of existing approaches for laughter detection, which use neural networks on top of traditional audio features like MFCC's. We use the same features (39 MFCC and delta features) and the same 3-layer feedforward network architecture. For consistency with our other models proposed here, we use one second of context (rather than 0.75 seconds) of audio to make a prediction for a given point in time.

4.2. ResNet

The second model we examine is an adaptation of ResNet-18 [28] for binary audio classification, using 128-dimensional mel spectrograms as features. One hypothesis for why this model should perform better on noisy data is that the features learned through the many levels of representation in ResNet are more specific to the sound of laughter than the traditional features in the baseline model; models learning from MFCC's may rely on exploiting surface-level characteristics of sound, which can lead to errors when those same characteristics occur by chance in background noise. Using the Switchboard training data, we experimented with various hyperparameters and network sizes before choosing the settings that gave the best results on the Switchboard validation data to apply to our test datasets.

4.3. ResNet with Data Augmentation

For our third model, we keep the same ResNet architecture but add several forms of data augmentation during training, including mixing in different background ambiences with a varying signal to noise ratio, masking sections of the input spectrogram with SpecAugment [29], and applying pitch-shifting, time-stretching, and artificial reverberation. Pitch-shift and time-stretch augmentations are implemented using the Librosa library [26], and artificial reverberation is implemented following the convolution-based method from Ravanelli et al. [30]. These augmentations are applied on the fly during training to each 1-second window of audio, with settings for every augmentation chosen randomly from a range of possible values. By applying these augmentations, we increase the size of the training data by artificially

Table 2: *Laughter detection performance on Switchboard and AudioSet, with 95% bootstrap confidence intervals. Segment-based metrics are reported per frame for Precision, Recall, and F1 scores. SLD and WLD refer to strongly and weakly labeled data.*

Train on Switchboard (SLD)	Results on Switchboard Test Data			Results on AudioSet Test Data		
	PRECISION	RECALL	F1	PRECISION	RECALL	F1
Baseline	0.634 (±0.025)	0.752 (±0.023)	0.688 (±0.016)	0.224 (±0.016)	0.901 (±0.014)	0.359 (±0.021)
ResNet	0.677 (±0.022)	0.830 (±0.019)	0.747 (±0.017)	0.464 (±0.020)	0.748 (±0.018)	0.573 (±0.018)
ResNet + Augmentation	0.676 (±0.022)	0.847 (±0.018)	**0.752** (±0.016)	0.508 (±0.020)	0.759 (±0.017)	**0.608** (±0.015)
Train on AudioSet (WLD)						
Baseline	0.300 (±0.024)	0.765 (±0.026)	0.430 (±0.026)	0.372 (±0.019)	0.856 (±0.019)	0.519 (±0.019)
ResNet	0.439 (±0.036)	0.710 (±0.028)	0.542 (±0.030)	0.371 (±0.017)	0.928 (±0.012)	0.530 (±0.018)
ResNet + Augmentation	0.468 (±0.027)	0.700 (±0.025)	0.563 (±0.023)	0.385 (±0.018)	0.925 (±0.015)	0.545 (±0.018)

adjusting sounds as well as synthetically placing them into a variety of environments with different background noises and spatial characteristics. In addition to increasing the size of the training data and functioning as a regularizer, these kinds of augmentations have shown to be important for speech recognition in noisy environments [30].

5. Results and Discussion

Because laughter is a variable-length event and many of our motivating applications are concerned with capturing the entirety of a laughter event rather than just counting the total number of events, we use segment-based metrics [15], with segments defined in this case as individual frames lasting 23 milliseconds, to calculate precision, recall, and F1 scores. Table 2 summarizes these results.

5.1. Evaluating in Controlled Environments: Switchboard

As the Switchboard section at the top of Table 2 illustrates, the baseline method of a featurized feed-forward neural network used in previous work achieves an F-Score of 0.69 on the clean environment of Switchboard; this model is substantially outperformed even in this environment by ResNet models that operate directly on the underlying spectogram, which achieve an F-Score of 0.75. In this environment, however, training with data augmentation does not lead to a significant increase in performance. Finally, as expected, training on the weakly labeled AudioSet data leads to much worse results here when evaluating in the clean Switchboard environment.

5.2. Evaluating in Uncontrolled Environments: AudioSet

Comparing the AudioSet and Switchboard sections of Table 2, we can see that results drop significantly across these two different recording environments: while our baseline model trained on Switchboard yields an F-score of 0.688 when evaluated on test data from that domain, this performance falls dramatically to 0.359 when evaluated on our newly annotated AudioSet data. This accords with the findings from Ryokai et al. [7]: a model trained on clean data (here, Switchboard) will generally perform worse when analyzing data in a noisy environment (here, AudioSet). This drastic difference highlights the value of investing annotation resources toward test data from uncontrolled and noisy environments. The model that has the best performance here is again ResNet augmented with data transformations; while this model does degrade in performance compared to evaluating on Switchboard, it suffers a less precipitous drop, achieving an overall F-score of 0.608, and yields a substantial absolute improvement of 0.249 points over the baseline model on this data. In this setting, the data augmentation does make a difference, providing a boost of more than 3 absolute points over the same ResNet model without augmentation.

5.3. Training on Weakly Labeled Data

By training separate versions of each model on both datasets, we can compare in this context the relative benefits of training on in-domain but weakly labeled data (AudioSet) against out-of-domain strongly labeled data (Switchboard). The results in Table 2 show that while the the weakly labeled data approach is reasonable, reaching an F-score of 0.545, within 6 points of the best performing model, training with strongly labeled data, even if it is out-of-domain, works best for laughter detection. This result is not evident here without our new annotations, again highlighting the need for continued investment in annotations for our test data that are *both* finely segmented *and* in-domain. While laughter is just one of hundreds of sound categories in AudioSet, our results suggest that when we want to study any of these sounds in depth, we should choose our evaluation scenarios carefully.

The more challenging evaluation dataset that we collect in this paper shows that much room for improvement still remains for detecting laughter in noisy real-world settings. In the absence of additional annotated data for training, leveraging the remaining AudioSet data or other weakly labeled corpora for semi-supervised learning offers one potential path forward.

6. Conclusion

Human laughter happens in noisy and lived environments. In an effort to build an automatic laughter detection system, we encountered the not-so-uncommon gap between theory and practice, a mismatch between clean training data and messy test data. To mitigate the disparity in performance when models trained in controlled environments are tested on real-world data, we have implemented a ResNet-based model for laughter detection, which is able to yield markedly better performance especially in noisy environments. Our work contributes a robust state-of-the-art machine learning method to detect human laughter and a newly annotated dataset for evaluation in noisy environments, while highlighting the importance of bridging the space between machine learning problems and their real-life uses in our noisy lives.

7. Acknowledgements

We thank the anonymous reviewers for their valuable feedback. The research reported in this article was supported by resources provided by NVIDIA.

8. References

[1] M. T. Knox, N. Morgan, and N. Mirghafori, "Getting the last laugh: Automatic laughter segmentation in meetings," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[2] M. T. Suarez, J. Cu, and M. Sta, "Building a multimodal laughter database for emotion recognition." in *LREC*, 2012, pp. 2347–2350.

[3] F. Haider, F. A. Salim, S. Luz, C. Vogel, O. Conlan, and N. Campbell, "Visual, laughter, applause and spoken expression features for predicting engagement within TED talks," *Feedback*, vol. 10, p. 20, 2017.

[4] V. Reddy, E. Williams, and A. Vaughan, "Sharing laughter: The humour of pre-school children with down syndrome," *Downs Syndrome Research and Practice*, vol. 7, no. 3, pp. 125–128, 2001.

[5] D. M. Richman, E. Gernat, and H. Teichman, "Effects of social stimuli on laughing and smiling in young children with angelman syndrome," *American journal on mental retardation*, vol. 111, no. 6, pp. 442–446, 2006.

[6] A. Truong and M. Agrawala, "A tool for navigating and editing 360 video of social conversations into shareable highlights," in *Proceedings of the 45th Graphics Interface Conference*. Canadian Human-Computer Communications Society, 2019, pp. 1–9.

[7] K. Ryokai, E. Durán López, N. Howell, J. Gillick, and D. Bamman, "Capturing, representing, and interacting with laughter," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1–12. [Online]. Available: https://doi.org/10.1145/3173574.3173932

[8] K. Ryokai, J. Park, and W. Deng, "Personal laughter archives: Reflection through visualization and interaction," in *Proceedings of the 2020 ACM International Symposium on Wearable Computers*, ser. UbiComp-ISWC '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 115–118. [Online]. Available: https://doi-org.libproxy.berkeley.edu/10.1145/3410530.3414419

[9] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at icsi," in *Proceedings of the first international conference on Human language technology research*. Association for Computational Linguistics, 2001, pp. 1–7.

[10] J. J. Godfrey and E. Holliman, "Switchboard-1 release 2," *Linguistic Data Consortium, Philadelphia*, vol. 926, p. 927, 1997.

[11] H. Salamin, A. Polychroniou, and A. Vinciarelli, "Automatic detection of laughter and fillers in spontaneous mobile phone conversations," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2013, pp. 4282–4287.

[12] L. Kaushik, A. Sangwan, and J. H. Hansen, "Laughter and filler detection in naturalistic audio," in *Proc. Interspeech*. International Speech and Communication Association, 2015.

[13] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[14] M. T. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

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

[16] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (icassp)*. IEEE, 2017, pp. 131–135.

[17] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," *arXiv preprint arXiv:1701.02720*, 2017.

[18] R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5484–5488.

[19] G. Laput, K. Ahuja, M. Goel, and C. Harrison, "Ubicoustics: Plug-and-play acoustic activity recognition," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 213–224. [Online]. Available: https://doi.org/10.1145/3242587.3242609

[20] J. Hernandez, M. E. Hoque, W. Drevo, and R. W. Picard, "Mood meter: Counting smiles in the wild," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ser. UbiComp '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 301–310. [Online]. Available: https://doi.org/10.1145/2370216.2370264

[21] H. Tsujita and J. Rekimoto, "Happinesscounter: Smile-encouraging appliance to increase positive mood," in *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 117–126. [Online]. Available: https://doi.org/10.1145/1979742.1979608

[22] M. Stäger, P. Lukowicz, N. Perera, T. Von Büren, G. Tröster, and T. Starner, "Soundbutton: Design of a low power wearable audio classification system," in *Proceedings of the Seventh IEEE International Symposium on Wearable Computers (ISWC'03)*, vol. 1530, no. 0811/03, 2003, pp. 17–00.

[23] M. Stager, P. Lukowicz, and G. Troster, "Implementation and evaluation of a low-power sound-based user activity recognition system," in *Eighth International Symposium on Wearable Computers*, vol. 1. IEEE, 2004, pp. 138–141.

[24] T. Rahman, A. T. Adams, M. Zhang, E. Cherry, B. Zhou, H. Peng, and T. Choudhury, "Bodybeat: a mobile system for sensing non-speech body sounds." in *MobiSys*, vol. 14, no. 10.1145. Citeseer, 2014, pp. 2 594 368–2 594 386.

[25] Q. Kong, C. Yu, Y. Xu, T. Iqbal, W. Wang, and M. D. Plumbley, "Weakly labelled audioset tagging with attention neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1791–1802, 2019.

[26] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680

[30] M. Ravanelli, P. Svaizer, and M. Omologo, "Realistic multi-microphone data simulation for distant speech recognition," *arXiv preprint arXiv:1711.09470*, 2017.