Robust Automatic Laughter Detection

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT

We present our iterative process of designing and implementing a novel machine learning algorithm that automatically identifies and extracts laughter from audio files. We present an empirical evaluation of several machine learning algorithms using three sets of audio data of varying sound quality. In the process, we contribute a new dataset of fine-grained laughter annotations on top of the existing AudioSet corpus of YouTube videos [9], with granular annotations for the start and end points of each laugh. We discuss the utility of our algorithm as well as the importance of understanding the variability of model performance in a range of real-world testing environments.

Author Keywords

Automatic laughter detection; Acoustic activity recognition

CCS Concepts

•**Human-centered computing** → **Human computer interaction (HCI);** Machine learning; Ubiquitous and mobile computing systems and tools.

INTRODUCTION

Laughter is a fundamental human expression. For the most part, laughter is associated with positive affect such as joy, happiness, amusement, and relief. While laughter is ubiquitous, we do not really know how much we laugh, when we laugh, or with whom we laugh. While many aspects of our lives are being quantified in commercially available health apps (e.g., daily amounts of our walking, sleeping, heart rate, etc. detected by wearable devices such as Fitbit, Apple Watch, and Oura Ring, just to name a few [3, 8, 25]), laughter has not yet been considered as a rich source for us to reflect on how we feel or understand our everyday life activities. What if we were able to celebrate perhaps even mundane but joyful moments of our lives, good or bad, through preserving and listening to our laughter?

Natural laughter, however, is difficult to anticipate. A family dinner, normally cheerful, may turn out to be without much laughter, while a serious work meeting with a colleague may end up producing much laughter. In a natural setting without any intervention (like tickling a loved one), by the time a good laugh has occurred, it would have been too late to turn on the recording device to capture that laugh.

This observation led us in previous work [30] to develop an automatic laughter detection algorithm to enable this kind of quantified engagement with laughter. For example, one application of automatic laughter identification may be that smart speakers, such as Amazon's Echo [2] or Google Home [12], may be listening only to our laughter, tallying the number of occurrences, and preserving its sound. Over time, we may end up with interesting cumulative data (e.g., who laughs the most and when in a household), the ability for us to recall some milestone laughter event (e.g., give me the laughter from my first day of school), or receive a surprise gift (e.g., here is a giggle of your daughter from this date last year), and potentially simple life-affirming signs which could even serve as an alternative to emotional biosensing systems [15]. With our automatic laughter detection classifier light enough to run on a smartphone, automatically identifying and collecting laughter could happen in real time and anywhere. Similar to how our photos are geotagged and could be organized spatially, over time, we may end up having a personal map that could show where we laugh the most in our daily lives. Some HCI researchers have already begun to incorporate our previous versions of laughter detection algorithms for various research purposes [5, 18, 36, 38].

However, our previous work uncovered one major limitation: our automatic laughter detection algorithm, trained on data from a quiet, controlled environment, worked well when there was minimum background noise; but, when our participants came back with their recordings from their daily lives, we found that their recordings were often filled with background noise. This, of course, makes sense as laughter happens 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. In addition, the quality of recording was poor, which added to noisiness.

This led to the development of our current work: an empirical comparison of the quality of laughter detection algorithms in a range of recording environments (ranging from very quiet to very noisy), and, more importantly, the development of a robust method that performs comparatively well in the presence of background noise. Our system uses modern convolutional neural network architectures based on ResNet [13] along with

UIST'20, October 20–23, 2020, Minneapolis, MN, USA

ISBN 978-1-4503-6708-0/20/04…$15.00

DOI: https://doi.org/10.1145/3313831.XXXXXXX

several forms of data augmentation that have proved effective for sound event detection [31] and speech recognition [27]. We predict laughter at a temporal resolution of 25 milliseconds, which allows us to find variable length events with precise boundaries. Our machine learning models are compact enough (less than 10MB) to be embedded in mobile devices and run fast enough to enable real time use. Code, data, and trained models are publicly available at http://anonymized.

We evaluated the fidelity of the algorithms using three different sets of data—cleanly recorded data in the Switchboard corpus [11], audio data from publicly available YouTube videos aggregated in AudioSet [9], and audio data collected by our participants in their daily lives over a few weeks. The three sets offer unique insights into the usability and applicability of techniques for recognizing laughter not just in controlled environments, but also in the real world, relying only on the microphones built into today's smartphones.

Our contributions include the following:

- We present a new dataset of fine-grained laughter annotations on top of the existing AudioSet [9] corpus of YouTube videos, with granular annotations for the start and end points of each laugh.
- We design and implement a novel machine learning algorithm that robustly identifies and extracts laughter from audio files.
- We conduct an empirical evaluation of several machine learning algorithms using three sets of audio data of varying sound quality.

RELATED WORK

Acoustic activities recognition in HCI

Within the field of HCI focusing on ubiquitous computing and sensing, researchers have explored a variety of techniques to collect, classify, and make sense of the sounds around people [1, 4]. For example, SoundButton [33, 34] is a wearable device integrated into the user's clothing, watch, or other accessories to recognize the sound of daily objects like coffee makers or copy machines. More recent work such as BodyScope [41] and BodyBest [29] uses acoustic sensors embedded within a necklace, custom-made microphones attached to the throat area that try to detect non-speech body sounds (swallowing, breathing, etc.). These works primarily focus on creating the hardware sensor system for acoustic sound event detection, but lack a systematic evaluation of how well the software actually performs. In our work, we conduct thorough evaluations of the detection algorithms for a better understanding of their performance and robustness.

HCI communities also explore the potential of leveraging mobile phone-based sensing frameworks to identify sounds in various environments, e.g., in wood workshops to detect sawing and drilling sounds [23, 39], or at the dining table to distinguish between the sounds of eating soup and drinking tea [32]. However, these works are limited to specific indoor locations and do not consider a broad spectrum of acoustic environments. Our approach aims to build a detection algorithm that covers a wide range of scenarios where human activities naturally take place.

The proliferation of sound databases and machine learning techniques has encouraged researchers to increase the variety of sound categories a system can detect and to boost classification accuracy. SoundSense [22] attempts to classify three general sound types: speech, music, and ambient sound. Zhan et al. [42] use a hidden Markov model to classify 22 typical environmental sounds related to personal and social activities (talking, walking, etc). Ubicoustics [21] fine-tunes a pre-trained Convolutional Neural Network to recognize 30 sound classes of interest using data from professional sound effect libraries traditionally created for the entertainment industry. More recent work from Wu et al. [40] allows the detection algorithms to discover new sounds and learn to classify them as well.

Our work is different from this previous work in that we put laughter center stage, automatically detecting and extracting laughter within various social activities, and then evaluating the performance of the system. Importantly, unlike previous work, we both recognize events in noisy and uncontrolled real-life scenes, while at the same time finding with fine resolution event boundaries that vary in length. This setting for the detection task is more challenging but more ecological for HCI research.

Speech recognition

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 [17, 20]. Published approaches to this task use traditional MFCC and pitch-based features, with more recent work employing deep feed-forward or convolutional neural networks on top of these features [16, 19, 37]. Most previous work on benchmarking and comparing methods for laughter detection evaluates approaches using datasets of cleanly recorded audio in quiet settings, typically employing the ICSI meetings database [24] or the Switchboard corpus of phone conversations [11]. Perhaps because of the nature of these datasets, previous work reports that these traditional features may be sufficient for laughter detection [16], but our observation that our baseline model produces many false positives when applied in real world scenarios suggests that more investigation is needed.

Work over the last few years in audio event detection [14], speech recognition [43], keyword spotting [35], and other audio detection problems have demonstrated that learning features directly from spectrograms using deep convolutional networks significantly outperforms previous approaches based on MFCC's and other traditional features. In many cases, combining these networks with data augmentation techniques like synthetic background noise and artificial reverberation produces state-of-the-art results [27, 31]. To our knowledge, state-of-the-art 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 recent work on audio event detection more generally, some studies do consider laughter as one of a larger list of categories [14, 21]. These works do employ state-of-the-art machine learning methods, but they do not specifically evaluate on the task of extracting laughter. Instead, they perform multiclass classification to choose one event type from a fixed set of categories; these metrics are difficult to interpret from a usability perspective if we are interested in understanding how well we can capture laughter in real world situations. Additionally, this body of work does not evaluate detection at precise timing boundaries; all events are assumed to have a fixed duration of 10 seconds [14] or one second [21], a limitation necessitated by the datasets employed. For the applications that motivate this research, however, we are interested in maintaining control over how much surrounding context is captured along with the laughter event, so more precisely annotated data is desirable.

DATA

One primary contribution of this work is evaluating the accuracy of methods for automatic laughter detection; to do so, we assess accuracy on a range of datasets: existing data collected in a controlled environment (the Switchboard corpus); YouTube data with weakly supervised labels, which we annotate for fine-grained temporal spans (AudioSet); and in-the-wild participant data with a mix of controlled and noisy backgrounds.

Switchboard

The Switchboard corpus of telephone conversations contains a total of about 260 hours of speech from 543 total speakers [11]. Of 2435 total conversations, we partition the dataset into 2159 for training, 119 for development, and 157 for testing, with total timing information across these partitions delineated in table 3. In Switchboard, conversations are manually annotated with precise timings for each word as well as laughter and silence. Laughter is annotated in two ways: either isolated between words, as shown in table 1, or combined with a word, as in table 2. 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.

We use the Switchboard data for training because it contains both a large volume of laughter and precisely aligned annotations of when the laughter occurs. As with other large annotated datasets for speech recognition, however, the data collection process focuses on quiet recording environments with minimal background noise. Additionally, the conversations are limited to two people at a time. We treat this an example of a highly controlled environment that allows us to determine the accuracy of laughter detection methods in idealized settings.

Audioset

One available dataset that has more realistic recording conditions than Switchboard and also contains examples of laughter is AudioSet [9]. This data consists of over 2 million video clips of 10 seconds each, which have been hand-labeled with over 600 categories of sounds. Among those clips, 5696 have been annotated as containing laughter. One drawback of AudioSet, however, is that the annotations are weakly labeled; we know that laughter occurs somewhere within the 10-second window, but unlike Switchboard, the dataset does not provide associated timing information on where during the clip the laughter occurs. Because of the structure of this dataset, most sound event detection work on AudioSet only operates at a granularity of 10 seconds.

Start	End	token
73.154	73.304	it's
73.304	73.444	going
73.444	73.504	to
73.504	73.634	be
73.634	73.974	really
73.974	74.416	good
74.416	75.135	[laughter]
75.135	75.648	[silence]

Table 1. Laughter annotation from the Switchboard Dataset

Start	End	token
183.665	183.816	[laughter-i]
183.816	184.067	[laughter-mean]
184.067	184.227	some
184.227	184.287	of
184.287	184.417	these
184.417	184.717	guys

Table 2. Annotation of laughter while talking in Switchboard data

Data partition	Isolated Laughter (Mins:Secs)	Laughter While Talking (Mins:Secs)	Total Laughter Events
Train	289:33	76:29	22490
Validation	9:04	2:57	876
Test	12:02	3:38	1119

Table 3. Laughter statistics in Switchboard data

Annotation

In order to leverage the data in AudioSet for our evaluation, we selected a subset of 1000 clips from those that were tagged as containing laughter, listened to each of them, 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. While the 10-second clips that make up our AudioSet test data consist of 39% laughter, evaluating on only these short windows of time around the actual events presents an easier problem than detecting laughter in a longer surrounding context. In order to better measure each model's false positive rate in different noisy environments, we supplement the 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.

To fairly compare model performance between the telephone conversations in Switchboard and the YouTube videos in Au-

dioSet, we subsample the Switchboard conversations in the test set so that the proportion of laughter (the class balance) in both evaluation datasets is the same. We do this by first choosing context windows around the annotated laughter events in Switchboard that 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 evaluation set.

In-the-wild participant data

Our previous study [30] explored the novel concept of capturing, representing, and interacting with laughter, focusing on tangible UIs. In this previous work, we asked our participants to collect about 3 hours of audio at different occasions and places in their daily lives where they might encounter laughter of their loved ones. Instead of collecting one continuous recording over 3 hours, the participants were asked to record about 18 different files that are each about 10 minutes long. This invited the participants to think about and anticipate laughter in different occasions and events in their lives. Our previous study resulted in a total of 288 files from our 16 participants, each approximately 10 minutes long. For each audio file, we then ran a laughter detection algorithm (the baseline method described below) so that we ended up with a collection of extracted laughter sound files, which our research team could listen to. To protect participants' privacy, we committed through our IRB to only listening to the extracted laughter files but not the raw recordings of the participants.

While many of the extracted laughter files correctly contained laughter, to our surprise, we also encountered many false positives (i.e., a section of audio file identified by the algorithm to contain laughter but when our research team listened to it to confirm, there was no laughter). For example, while some files resulted in algorithmic precision of 1.0 (i.e., all the captured laughs by the algorithm actually contained laughs), some of the participants' files resulted in precision lower than 0.1 (i.e., less than 10% of captured laughs by the algorithm actually contained laughs). This low precision rate was largely due to background noise in participants' audio files.

In order to achieve the best results with our laughter detection algorithm, we originally asked our participants to collect audio files without any background noises. However, most files our participants ended up collecting included many noises. Our subsequent interviews with the participants revealed that participants made their audio recordings at birthday parties, field trips, in cars, etc. so that the recordings had a variety of noises (e.g., outdoor noises from wind, traffic, etc., overlapping conversations during gatherings, or general noises from a phone being carried around while recording).

While our previous research focused on human interaction with captured and preserved laughter, we did not deeply analyze how precise our laughter detection method was or could have been. In this paper, we are revisiting this audio data from our participants collected earlier, in order to assess the quality of several laughter detection methods when faced with the challenge of background noise in real-world audio data.

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.

Baseline

As a baseline, we use the feed-forward neural network laughter detection model from our previous work [30]. This model is the most recently published example that is representative of existing approaches for laughter detection that 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.

ResNet

The second model we examine is an adaptation of ResNet-18 [13] for binary audio classification. The model takes in one-second windows of audio and uses 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, which may rely on exploiting surface-level characteristics of sound that can occur by chance in background noise. 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 other evaluation datasets.

ResNet with Data Augmentation

For our third model, we keep the same ResNet architecture but add several forms of data augmentation during training, including randomly mixing in different background ambiences with a varying signal to noise ratio, using SpecAugment [27] to randomly mask sections of the input spectrogram, and randomly applying pitch-shifting, time-stretching, and artificial reverberation. These augmentations are applied on the fly duration training to each 1-second window of audio, with specific 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 adjusting the sounds of the voices recorded in Switchboard as well as synthetically placing those sounds into a variety of environments with different background noises and spatial characteristics.

Model Training

All models are trained on Switchboard data and tested on the different datasets described above. Because Switchboard is highly imbalanced in its proportion of laughter compared to speech (laughter makes up less than 2% of the total recording time), we train all models with negative sampling, randomly choosing an equal proportion of non-laughter in order to match with the laughter examples in every batch during training. All

	Switchboard			AudioSet		
	P	R	F	P	R	F
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)

Table 4. Laughter detection performance (timing overlap) on Switchboard and AudioSet, with 95% bootstrap confidence intervals.

models are trained using Pytorch [28] for 100,000 steps with a batch size of 32.

EVALUATIONS

The purpose of the quantitative evaluation of our algorithms is to determine the accuracy of the algorithms in correctly identifying laughter in any given audio file. More specifically, we are interested in investigating how different models perform in the presence of background noise. Since our goal is to use automatic laughter detection in real-world environments like restaurants, parks, or parties, we aim to directly assess how *robust* a method is at detecting laughter. In this section, we present three evaluations—one for each of the datasets described above, varying from the cleanest recording environment (Switchboard) to the noisiest (participant data).

Evaluation 1: Switchboard

We first evaluate on the clean environment present in the Switchboard data. We report precision, recall, and F1 scores for measuring the *timing overlap* between predicted and annotated laughter times—the degree to which spans between the predicted laughter overlaps with the gold-annotated laughter. This metric computes the fine-grained overlap in continuous time between the annotated laughter and the predictions. We calculate it by dividing each sequence into 25ms chunks (with up to 0.5 seconds of context on each side available to the model at inference time) and evaluating precision, recall and F1 scores over these chunks.

As the Switchboard section of table 4 illustrates, the baseline method of a featurized feed-forward neural network we use in previous work achieves an F-Score of 0.688 in timing overlap on the clean environment of Switchboard; but this model is dramatically outperformed even in this environment by ResNet models that operate directly on the underlying spectogram, which achieve an F-Score of 0.752 when using data augmentation.

Evaluation 2: Audio from YouTube Videos

In previous work on laughter detection, we found that the method was not resilient to the types of background noise found in some of our participants' recordings. This experience suggested to us that in order to better evaluate the usability of laughter detection systems for the real-world applications that motivate this work, we need an evaluation dataset that better captures the variety of conditions in which we might record audio during everyday life.

The AudioSet section of table 4 presents the results of this analysis. Comparing these results to the Switchboard results, we can see 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 fine-grained AudioSet annotations. This accords with our experience in our previous work anecdotally examining performance on participant data: a model trained on clean data (here, Switchboard) will generally perform worse when analyzing data in a noisy environment (here, AudioSet). At the same time, the model that has the best performance 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.

Evaluation 3: Participants' Data

We are ultimately interested in investigating how the accuracy of laughter detection methods can be robust to noise in the user environment. In order to analyze this in finer detail, our first step was to revisit our original participants' files in terms of their relative noisiness or cleanness of background sounds. We created two buckets ("clean" and "noisy") based on the result of our previous algorithm's performance, without listening to participants' files at all. For each of the participant files in each bucket, we calculated the algorithm's precision (how many of its instances of classified laughter were indeed instances of laughter confirmed by our research team).

There are two considerations that make us focus on precision. First is respect for our participants' privacy: measuring recall in this dataset would require human listening to the entirety of a conversation, and we committed to only accessing the data that specifically pertained to laughter. Second, and more broadly, reflects the long-term goals of this work: given our scenarios of collecting audio files over much longer periods of time (e.g., days, weeks, months), precision is more important (since we seek examples of true laughs, not necessarily knowledge of how many laughs are correctly captured for a particular file).

The results are summarized in table 5. In comparing the perfomance across the different algorithms, we see substantial improvements on the "clean" group when moving from our original baseline algorithm (precision 0.520) to the ResNet model with data augmentation (0.672), an absolute increase of 0.152 points. However, the results for the "noisy" group demonstrate a dramatic improvement: while our original baseline method only yields a precision of 0.107 on this data, the ResNet model with data augmentation improves this to 0.643—a remarkable absolute improvement of 0.536 points.

	Clean	n	Noisy	n
Baseline	0.520 ($\pm$0.049)	392	0.107 ($\pm$0.014)	1959
ResNet	0.676 ($\pm$0.067)	185	0.610 ($\pm$0.051)	349
ResNet + Augmentation	0.672 ($\pm$0.059)	247	0.643 ($\pm$0.050)	356

Table 5. Precision of laughter detection on participant data, along with 95% confidence intervals.

DISCUSSION AND FUTURE WORK

The three evaluations carried out above demonstrate two things. One is the importance of understanding the variability of model performance in a range of testing environments; all three models described in this work were trained on Switchboard data and evaluated on three very different datasets and yield very different performance scores. While this mismatch between train and test data is a widely common one in machine learning, and drives much work in domain adaptation [6, 10, 26], it also reflects a common pattern found in real world practice, where a model is trained on one dataset (whether Switchboard for laughter detection or ImageNet [7] for image classification), and then deployed in another environment; in such cases, this work shows how that performance will degrade.

At the same time, this work also demonstrates a robust solution for the particular problem of laughter detection: while the baseline feed-forward neural network used in our previous work shows a severe degradation in performance when moving from the clean domain of Switchboard to the noisy domain of participant data, we present a method here that is more robust to such variation: ResNet with data augmentation is able to yield comparable performance across clean and noisy environments in our participant data, and shows markedly better performance across all three datasets tested here.

The two sets of audio data we produced in this work (our participant data and our fine-grained annotated AudioSet data) offer unique insights into the usability and applicability of techniques for recognizing laughter not just in controlled environments, but in the real world, relying only on the microphones built into today's smartphones. Our machine learning models are compact enough (less than 10MB) to be embedded in mobile devices and run fast enough to enable real time use. The ResNet-based algorithm with greater tolerance towards background noises opens up a variety of new application possibilities. As future work, we envision that a similar set of machine learning models could be applied to detecting and preserving non-verbal human expressions beyond laughter. For example, we might identify and preserve our exclamations and other interjections such as gasps, sighs, various types of cries (e.g., cries of joy, cries of trying times), and other types of non-verbal expressions that are not spotted by speech recognition systems.

This work presents a new dataset for evaluating automatic laughter detection algorithms and applies state-of-the-art machine learning methods to the problem. We find that applying modern deep learning models along with data augmentation allows us to make significant advances in the problem of automatic laughter detection, especially in noisy environments. Through our quantitative evaluations with data recorded in noisy conditions as well as our experience in applying these models to data collected from participants in a design study, we show that these improvements lead to increased usability in practice.

The more challenging evaluation datasets that we collect show that much room for improvement still remains for detecting laughter in noisy real-world settings. The most straightforward path toward bridging the gap in performance between controlled settings like Switchboard and real world settings like AudioSet likely involves annotating a dataset large enough that we can not only evaluate models realistically, but also reasonably train on data from a variety of real world settings. In the absence of more annotated data, leveraging the remaining AudioSet data or other weakly labeled corpora for unsupervised or semi-supervised learning offers one potential path forward.

We see several potential use cases of this technology. First, by focusing on the sound of laughter, we are being encouraged to listen to and pay attention to the details and enjoy the nuances of our loved ones' laughter without any visual distraction. One possible application may be to think about use cases based on a collection of personal laughter samples. For example, from a birthday party, a family may receive a soundtrack of laughter by all attendees (without revealing the identity of any individual). A musician may compose a symphony based on a large collection of laughter from their loved ones, or a sound designer may collect and curate a library of unique laughter sounds for creative reuse.

A more commercial application may use automatic laughter detection to understand service performance. For example, a company which offers its customer service over the phone may identify the presence and frequency of laughter as an additional perspective to reflect on their customer satisfaction. Similarly, physical performances in theaters or even restaurants where services are provided over a shared physical area might be able to track how much laughter is associated with a particular service that occurred in a particular location, or particular time of performances, without identifying individuals or analyzing the content of conversations at all.

Finally, automatically detected laughter may also be used as a way for editors or viewers to navigate through existing media in order to find climactic scenes or special moments [36].

By this, however, we are not advocating that our laughter or other types of non-verbal expressions should eventually be all identified and quantified in order for us to keep track of our health, happiness, and other feelings (as we sometime see in quantified-self technologies). Rather, we have merely demonstrated technical possibilities to automatically identify naturally occurring human laughter, and have hinted at how this

might be extended to identifying other non-verbal expressions. We have described some possible applications of automatic laughter detection algorithms to elicit questions around future use case scenarios of what we might do with captured laughter as an additional point of reflection of our daily lives. While focusing on the sound of laughter and identifying it as variable length events, we are going beyond detecting presence, absence, or frequency of laughter. We want to emphasize that this work fundamentally pertains to recognizing laughter as the sound of our lived expressions shared with our loved ones, something that could be preserved, revived, revisited, and even cherished. Each instance of laughter carries personal meaning and creates an opportunity for us to listen to it again later, and to get reconnected with our past experiences. Our automatic laughter detection system enables such novel encounters and reflection, and can help drive the appreciation and celebration of everyday moments in our lives.

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 this issue, we have implemented a ResNet-based algorithm with data augmentation, which is able to yield comparable performances across clean and noisy environments in our participant data, and shows markedly better performance across all three datasets we have tested. Our work contributes a robust state-of-the-art machine learning method to detect human laughter, while highlighting the importance of bridging the space between machine learning problems and their real-life uses in our noisy lives.

ACKNOWLEDGMENTS

Removed for anonymous review.

REFERENCES

[1] Gregory D Abowd, Anind K Dey, Peter J Brown, Nigel Davies, Mark Smith, and Pete Steggles. 1999. Towards a better understanding of context and context-awareness. In *International symposium on handheld and ubiquitous computing*. Springer, 304–307.

[2] Amazon. 2020. Amazon, Inc. Echo. (2020). `https://www.amazon.com/b?&node=9818047011&ref=ODS_v2_FS_AUCC_category` Retrieved on May 4, 2020.

[3] Apple. 2020. Apple, Inc. Watch. (2020). `https://www.apple.com/watch/` Retrieved on May 4, 2020.

[4] Matthias Baldauf, Schahram Dustdar, and Florian Rosenberg. 2007. A survey on context-aware systems. *International Journal of Ad Hoc and Ubiquitous Computing* 2, 4 (2007), 263–277.

[5] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper). *arXiv preprint arXiv:1906.01815* (2019).

[6] Hal Daumé and Daniel Marcu. 2006. Domain Adaptation for Statistical Classifiers. *Journal of Artificial Intelligence Research* 26 (2006), 101–126.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[8] Fitbit. 2020. Fitbit, Inc. (2020). `https://www.fitbit.com/us/home` Retrieved on May 4, 2020.

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

[10] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 513–520.

[11] John J Godfrey and Edward Holliman. 1997. Switchboard-1 Release 2. *Linguistic Data Consortium, Philadelphia* 926 (1997), 927.

[12] Google. 2020. Google, Inc. Home. (2020). `https://store.google.com/us/product/google_home` Retrieved on May 4, 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[14] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, and others. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[15] Noura Howell, Greg Niemeyer, and Kimiko Ryokai. 2019. Life-Affirming Biosensing in Public: Sounding Heartbeats on a Red Bench. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

[16] Lakshmish Kaushik, Abhijeet Sangwan, and John HL Hansen. 2015. Laughter and filler detection in naturalistic audio. (2015).

[17] Lyndon S Kennedy and Daniel PW Ellis. 2004. Laughter detection in meetings. (2004).

[18] Joshua Y Kim, Greyson Y Kim, and Kalina Yacef. 2019. Detecting depression in dyadic conversations with multimodal narratives and visualizations. In *Australasian Joint Conference on Artificial Intelligence*. Springer, 303–314.

[19] Mary Tai Knox and Nikki Mirghafori. 2007. Automatic laughter detection using neural networks. In *Eighth Annual Conference of the International Speech Communication Association*.

[20] Mary Tai Knox, Nelson Morgan, and Nikki Mirghafori. 2008. Getting the last laugh: Automatic laughter segmentation in meetings. In *Ninth Annual Conference of the International Speech Communication Association*.

[21] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-Play Acoustic Activity Recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. Association for Computing Machinery, New York, NY, USA, 213–224. DOI: http://dx.doi.org/10.1145/3242587.3242609

[22] Hong Lu, Wei Pan, Nicholas D Lane, Tanzeem Choudhury, and Andrew T Campbell. 2009. SoundSense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*. 165–178.

[23] Paul Lukowicz, Jamie A Ward, Holger Junker, Mathias Stäger, Gerhard Tröster, Amin Atrash, and Thad Starner. 2004. Recognizing workshop activity using body worn microphones and accelerometers. In *International conference on pervasive computing*. Springer, 18–32.

[24] Nelson Morgan, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Adam Janin, Thilo Pfau, Elizabeth Shriberg, and Andreas Stolcke. 2001. The meeting project at ICSI. In *Proceedings of the first international conference on Human language technology research*. Association for Computational Linguistics, 1–7.

[25] Oura. 2020. Oura, Inc. (2020). https://ouraring.com/ Retrieved on May 4, 2020.

[26] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.

[27] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*. 2613–2617. DOI: http://dx.doi.org/10.21437/Interspeech.2019-2680

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. 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., 8024–8035.

[29] Tauhidur Rahman, Alexander Travis Adams, Mi Zhang, Erin Cherry, Bobby Zhou, Huaishu Peng, and Tanzeem Choudhury. 2014. BodyBeat: a mobile system for sensing non-speech body sounds.. In *MobiSys*, Vol. 14. Citeseer, 2594368–2594386.

[30] [Authors removed for anonymous review]. 2018. Capturing, Representing, and Interacting with Laughter. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article Paper 358, 12 pages.

[31] J. Salamon and J. P. Bello. 2017. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Processing Letters* 24, 3 (2017), 279–283.

[32] Jae Mun Sim, Yonnim Lee, and Ohbyung Kwon. 2015. Acoustic sensor based recognition of human activity in everyday life for smart home services. *International Journal of Distributed Sensor Networks* 11, 9 (2015), 679123.

[33] Mathias Stäger, Paul Lukowicz, Niroshan Perera, Thomas Von Büren, Gerhard Tröster, and Thad Starner. 2003. 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. 17–00.

[34] Mathias Stager, Paul Lukowicz, and Gerhard Troster. 2004. Implementation and evaluation of a low-power sound-based user activity recognition system. In *Eighth International Symposium on Wearable Computers*, Vol. 1. IEEE, 138–141.

[35] Raphael Tang and Jimmy Lin. 2018. Deep residual learning for small-footprint keyword spotting. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5484–5488.

[36] Anh Truong and Maneesh Agrawala. 2019. A Tool for Navigating and Editing 360 Video of Social Conversations into Shareable Highlights. In *Proceedings of the 45th Graphics Interface Conference on Proceedings of Graphics Interface 2019*. Canadian Human-Computer Communications Society, 1–9.

[37] Hitomi Tsujita and Jun Rekimoto. 2011. HappinessCounter: Smile-Encouraging Appliance to Increase Positive Mood. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11)*. Association for Computing Machinery, New York, NY, USA, 117–126. DOI: http://dx.doi.org/10.1145/1979742.1979608

[38] Ryoko Ueoka. 2019. Laugh Log: E-Textile Bellyband Interface for Laugh Detection and Logging. In *International Conference on Human-Computer Interaction*. Springer, 426–439.

[39] Jamie A Ward, Paul Lukowicz, Gerhard Troster, and Thad E Starner. 2006. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE transactions on pattern analysis and machine intelligence* 28, 10 (2006), 1553–1567.

[40] Jason Wu, Chris Harrison, Jeffrey P Bigham, and Gierad Laput. 2020. Automated Class Discovery and One-Shot Interactions for Acoustic Activity Recognition. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[41] Koji Yatani and Khai N Truong. 2012. BodyScope: a wearable acoustic sensor for activity recognition. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 341–350.

[42] Yi Zhan and Tadahiro Kuroda. 2014. Wearable sensor-based human activity recognition from environmental background sounds. *Journal of Ambient Intelligence and Humanized Computing* 5, 1 (2014), 77–89.

[43] Ying Zhang, Mohammad Pezeshki, Philémon Brakel, Saizheng Zhang, Cesar Laurent Yoshua Bengio, and Aaron Courville. 2017. Towards end-to-end speech recognition with deep convolutional neural networks. *arXiv preprint arXiv:1701.02720* (2017).