Noisemes: Manual Annotation of Environmental Noise in Audio Streams

Susanne Burger, Qin Jin, Peter F. Schulam, and Florian Metze

CMU-LTI-12-017

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Noisemes: Manual Annotation of Environmental Noise in Audio Streams

Susanne Burger, Qin Jin, Peter F. Schulam, and Florian Metze

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

{sburger,qjin,pschulam,fmetze}@cs.cmu.edu

Abstract

Audio information retrieval is a difficult problem due to the highly unstructured nature of the data. A general labeling system for identifying audio patterns could unite research efforts in the field. This paper introduces 42 distinct labels, the "noisemes", developed for the manual annotation of noise segments as they occur in audio streams of consumer captured and semi-professionally produced videos. The labels describe distinct noise units based on audio concepts, independent of visual concepts as much as possible. We trained a recognition system using 5.6 hours of manually labeled data, and present recognition results.

Index Terms: noise annotation, audio scene analysis, audio classification.

1. Introduction

Acoustic Scene Analysis (ASA) or Audio Event Detection (AED) is the task of identifying arbitrary acoustic events in a stream of audio data [1]. The retrieved information is already supporting multimedia information retrieval, robotic systems, and portable automatic speech recognition systems ([2], [3], [4]).

Research found that environmental noise enhances human comprehension and vision perception [5], suggesting that adding non-speech audio material promotes people's comprehension of visual languages and sceneries. Noise detection and filtering is a factor in increasing the robustness of automatic speech recognition systems [6]. More recent projects show that the detection of environmental noise provides helpful contributions in identifying and classifying events in consumer captured videos [2], and for summarizing what was detected [7].

The difficulties in general acoustic scene analysis can be broadly classified into two categories. The first difficulties are those related to technical complexity. High-quality, post-processed audio tracks from commercial videos where environmental noise often comprises a sequence of pre-recorded sounds, speech and music are relatively easy to analyze computationally [8]. Foreground sounds are usually artificially emphasized, and thus easier to segment from the background noise. In the middle of the field we find corpora collected in prepared scenarios and predefined events for scientific purposes [9]. These collections provide near-real samples of noise for initial research, but still avoid the multiplicity of what can be found in the real world. At the other end of the spectrum are consumer-captured home videos filmed with dated video cameras and low quality microphones, resulting in distorted and clipped sound tracks. Noises frequently overlap and mix, making source separation extremely difficult.

The second class of difficulties is related to the infrastructure and resources of the research community. ASA can be performed in a multitude of different environments, and system development can be driven by a wide array of tasks. Simple speech-music separation requires only two labels, and training data is relatively easy to obtain. As a task requires more detailed analysis, however, the number of labels can increase and segmentation and separation become extremely difficult. What is needed is a general way of describing the events within an audio stream that can be applied to all environments in tasks for which we would like to build acoustic scene analysis systems.

In this paper we propose a set of "noisemes", fundamental atomic units of sound that attempt to capture objective properties of the acoustic signal independent of any other information source that may be available. In science, acoustic signals are typically quantified and analyzed using Mel-frequency cepstral coefficients, pitch, fundamental frequencies, and energy levels. On the other hand, humans typically describe the qualities of a particular sound by largely relying on complementary sensory input, i.e. information that is "out of band" for a system that is performing ASA using solely a digital representation of the signal. We find associative descriptions where noises are linked to emotions, objects or locations. We are also often influenced by an image of which environment a noise has been heard before. The SoundNet [10] database provides a comprehensive association network of words and sound experiences.

Conceptually, our goal when developing the noisemes was to develop symbolic labels for quantitative properties of the signal. In this way, we hope to introduce acoustic concepts that are not biased by any information that cannot be directly obtained by observing the signal in both the time and frequency domains, but are still general enough to describe any acoustic pattern observed. We make the following contributions: We outline the noisemes, and review the methods used to develop the set of labels. We further motivate development of these labels in Section 2. We present occurrence statistics of the noisemes on a small collection of diverse videos extracted from the Internet. Finally, we demonstrate that the noisemes are recognizable by machine, and present classification results.

2. Development of semantic audio labels

We can motivate the development of the noisemes from two perspectives. From a machine learning perspective, we believe that establishing a general and objective collection of labels that can be applied to any type of audio data regardless of the source or content will better define the ASA task. The majority of machine learning tasks that have enjoyed considerable attention and developmental success over the past few years are those for which the problem is well defined, and for which hard metrics have been established by which researchers can objectively compare approaches and evaluate hypotheses. It is not immediately obvious how to concretely define the ASA task. Since settings vary widely it is difficult to drive progress forward because evaluation of different approaches is often done using very different criteria. For example, some projects may wish to simply segment speech and non-speech, while others may wish to classify the environment in which an audio track was recorded. It is not always immediately obvious how to apply techniques from one domain to another. Introduction of a consistent set of labels that can be applied to any type of data can help to remedy this situation by essentially establishing a *lingua franca* for researchers working in acoustic scene analysis.

The second advantage of developing a unified set of labels for ASA is from a data scientist's or annotator's perspective. It becomes much easier to develop high-quality ground truth resources for this task if all annotators speak a unified language, and can easily and objectively label segments of audio according to the quantifiable properties of the signal. E.g. rather than guessing that a certain noise is made from the engine of a car (a specific explanation that can only be confirmed or refuted by observing the source), an annotator would label the segment as "engine_light" or "engine_heavy".

2.1. Data

We picked a small subset of data from the TRECVID MED 2011 corpus [11]. This corpus consists of 1500 hours of video clips, mostly of the style consumer captured home video, but also including semi-professionally produced "How-to" videos. Parts of the videos were categorized as one of 18 events. An event is defined in [11] as "a complex activity occurring at a specific place and time, involving people interacting with other people and/or objects; it consists of a number of human actions, processes, and activities that are loosely or tightly organized and that have significant temporal and semantic relationships to the overarching activity; it is directly observable." Examples for the 18 events are: attempting a board trick, feeding an animal, changing a vehicle tire, making a sandwich, building a shelter, batting in a run, celebrating a birthday.

To ensure a broad variety of environmental noise, we picked at least 10 videos of each event, a total of 190 videos, plus 26 more video clips from a set of video clips that had not been categorized as one of the events. We only worked with the audio stream of the 216 videos, 5.6 hours of data, which we extracted as FLAC files with a sampling rate of 16 kHz from the provided MP4 AAC audio. We used PRAAT [12] as annotation tool.

2.2. Labeling process

Despite a large part of overlapping noises, we decided to label "monophone" noise, not polyphone noise mixes [13] by labeling noise on different tracks when it co-occurred with other noise. The human ear is fairly good in identifying sounds even if they are overlapped by other sounds.

We started by listening and annotating many different audio streams, using open labeling. We then structured, cleaned, and combined these open labels and sorted them under different aspects, such as source, similar features, similar concepts, effects, and prominence. This process was repeated several times. The aspects that seemed to serve best were the semantic aspect of possible sound source, and the feature based aspect of acoustic properties similar to speech features: voiced or showing formants, fricative-like or showing fuzziness, plosive-like or showing single pulses.

2.3. Noisemes

Eventually, a set of 42 noise units sufficiently covered noises heard from the audio streams. We call these units "noisemes", pronounced similar to /phoneme/, and also defined similar:
the smallest segmental unit of sound employed to form meaningful contrasts between noises.

Table 1 and Figure 1 show the two main aspects of the labels: Table 1 is organized in noise sources. Note that there is no confirmation of the actual source of a noise; the order serves as a meaningful way to memorize the noisemes. Figure 1 puts the labels in a space between formant-like, friction-like, and pulse-like acoustic features.

Using Table 1 to introduce the labels in more detail, we basically have four possible sources of noise: noise that sounds as if produced by vocal folds of humans or animals, noise that can be caused by direct human impact or activity, mechanical noise, and natural noise.

There is a general label for sounds produced by animals. For cases where the labeler recognizes a particular animal (e.g. a bird) a comment can be attached (e.g. "anim_bird"). The broad category *Human_noise_s* collects non-speech sounds produced by individual humans or animals. *Speech_s* contains speech produced by a single speaker. Speech is labeled as "speech" if it is understandable English, and is transcribed at the word level in an extra track. In other cases, speech is labeled as "ne_speech" if it is not English, and as "mumble" if it is not understandable. There is "singing" when one or several voices sing without accompanying music. *Human_m* describes vocal noise from multiple people. Specifically, we give the label "crowd" if it is overlapped and in disorder, and "cheer" if many voices repeat something in unison. A mixture of voice and music is labeled as "music_sing". There is also "music" without singing. The music labels are used when music is prominent. Similar to this is the background noise "radio", which refers to TV or radio, but is heard as weaker background noise.

Noise caused by possible human activity is organized in two groups: the short *noise_pulse* and the longer *noise_ongoing*. The short pulses are probably the hardest to identify – for humans as well as for a system. If heard without noise overlap and in fair quality, there is a perceivable difference between pulses that have a tonal aspect such as "bang", pulses that are dampened such as "thud", "clap" that sounds like an explosion, "click" which is very low energy, "knock" that has a light hollow tone, and "beep" which is tonal and longer than a typical pulse.

The ongoing noises are either a series of regular pulses and /or friction: "hammer", "scratch", and "washboard". "Washboard" seems like a visual object but nowadays is often used as a percussion instrument producing a very distinct staccato friction sound. "Applause", "rustle" and "clatter" are irregular sequences of pulses and friction. Tonal noise either caused by human activity, or by machine is "ring", "siren", "whistle", "squeak", or just "tone", differentiated by melody or frequency. For engine noise we use categories such as "engine_light", "engine_heavy", "engine_quiet" and "power-tool" to avoid the difficulty of finding thresholds between low, mid and high frequencies. Gusty, frictional noise is labeled as "wind", splashes are labeled as "water", and direct airflow into a microphone is labeled as "micro_blow". Any type of non-identifiable ongoing friction noise is called "white_noise". To account for unusual sounds that do not match any noiseme, we defined a "catch all" open label next to "animal", and called it "other". Since there is no visual during the labeling process, a labeler is instructed to use a label that is close to the sound. This blind labeling sometimes creates surprising discoveries, e.g. an audio track labeled with short segments of light engine noise turned out to be the snoring of three dogs.

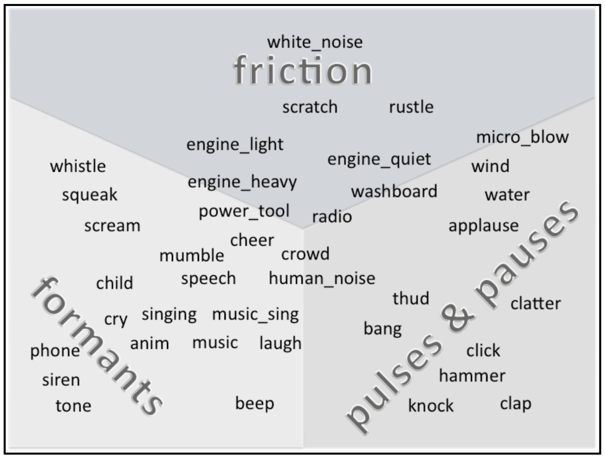

Figure 1: *Acoustic features: a noiseme is either formant-like, friction-like, pulse-like, or a mix of these features*.

Table 1: *Noisemes, grouped in broader categories, with description for each noiseme*.

broad	***Noiseme***	***sounds like …***
anim	animal	not identifiable animal
	anim_...	identified anim_dog, anim_bird ..
human_ noise_s	cry	crying
	human_noise	vocal noise: cough, sneeze, throat..
	laugh	laughter
	scream	screaming
	child	child/baby coos, animal coos
speech_ s	mumble	non-intelligible, single voice
	speech	intelligible speech English
	speech_ne	intelligible speech not English
singing	singing	only voice, a capella
human_ m	cheer	intelligible speech, multiple voices
	crowd	non-intelligible, multiple voices
music	music_sing	music with singing
	music	only music
noise_ pulse	knock	hits wood, cardboard, dry wall
	thud	hits floor, dirt, carpet, damped
	clap	hands, gun, shot-like, explosion
	click	quiet, mechanical click
	bang	hits metal, glass, tone-ish
	beep	very short beeps, computer
noise_ ongoing	clatter	bangs, knocks, pulses, irregular
	rustle	scratching, hiss, rustling, irregular
	scratch	short friction segments, regular
	hammer	bangs, knocks, pulses, regular
	washboard	fast pulses with rubbing, friction, regular
	applause	very fast claps comb. with friction,
engine	engine_quiet	rattle, sewing machine, video camera
	engine_light	high-freq. machine noise, drill-like
	power_tool	mid-freq. machine noise, race car
	engine_heavy	low-freq. machine noise, truck, tractor
noise_ tone	phone	classical telephone ring, ringing
	whistle	high-freq. tone
	squeak	tire squeak, friction squeak, high freq.
	tone	steady tone, horn, alarm
	sirene	oscillating sound waves
noise_ backgr_ nat	water	dubbing, splashing
	micro_blow	wind or breath hits microphone
	wind	gusts, flag clatter, pulses, and scratch
noise_ backgr	radio	radio/TV in background
	white_noise	fuzzy signal, air cond., waterfall, hum
other	other_creak	open for unseen noises

2.4. Observations

The segmenting and labeling of the audio tracks of the 216 videos resulted in a total of 6.8 hours of labeled noise duration in 1462 noise segments. 21% of the 5.6 hours of used audio streams contains label overlap. We additionally segmented and labeled if the original audio was substituted by post-processed audio, or if there was overlapping audio that was added later; 31% of the 5.6 hours of total duration was labeled as post-processed (complete new track or overlapping with original).

47% of the 5.6 hours contains speech (including mumble and non-English, articulated and crowd noise), 55% of this is understandable English speech, transcribed on word level. 33.8% of the 5.6 hours have music or music with singing. Most of it is part of the post-processed audio. Animal noise was only found at 1.7% of the total duration. However, animal noise turned out to be significant in categorizing two of the events, Animal_feeding and Animal_grooming. The short pulse-like noisemes take only 2.7% of the 5.6 hours of data, but they comprise 20% of the number of segments.

3. Automatic noiseme classification

We think the noiseme labels will be useful for detailed summaries of videos, providing evidences for a detection decision. For example, rather than only outputting a categorization decision as "this is a baseball game" event, we can provide justifying evidences by mentioning high occurrences of "crowd" "cheering", "applause", and "bang" labels. It is also important to have information regarding the presence, duration, and relative order of certain types of noisemes that can be used to infer the content of a particular video, for example, "scratch" followed by "clap", followed by "clatter" can be the typical temporal pattern for a particular skate board trick. Therefore, an actual goal of working with noiseme-annotated data is to be able to create event signatures or event fingerprints using particular patterns or significant co-occurrences of features from different modalities.

3.1. Experimental setup

Only one annotator annotated our data set. In order to verify the usefulness of the noiseme labels, we conducted an automatic noiseme classification task. We used 2/3 of the labeled noiseme data from 150 annotated videos for training noiseme models and the remaining 1/3 for testing. There are in total 2510 test segment trials. A few noisemes were not or only with one sample represented in the 150 videos. We excluded them from the automatic classification experiments and conducted noiseme classification experiments for 38 classes.

3.2. Experimental results

The Gaussian Mixture Model (GMM) is a popular statistical model for classification tasks [14]. A model based on a GMM consists of a finite number of Gaussian distributions parameterized by their a priori probability, mean vectors, and covariance matrices. The parameters of the model are typically estimated by maximum likelihood estimation, using the Expectation-Maximization (EM) algorithm. We extract 20-dimensional Mel-frequency Cepstral Coefficients (MFCC) with Cepstral Mean Normalization (CMN) applied. The first order derivative (delta MFCC) is appended to form a 40-dimensional feature in our experiments. We trained for each noiseme class a GMM model with 256 Gaussian mixtures.

Table 2 presents the noiseme classification accuracy with respect to the test trial durations. The general trend is that the classification accuracy increases with longer test trials. However, since we are only using MFCC features, for the shorter noisemes such as "beep", "clip", etc., other types of features such as prosodic features or temporal patterns might be considered.

We also looked at the top-5 hypotheses of the GMM-based noiseme classification system. It shows that the corresponding target noiseme class always appears in the top 5 hypotheses. This verifies that by using some acoustic features the defined noiseme classes can be distinguished by an automatic system. Some of the confusions we saw were intuitive, such as "mumble" was recognized as "speech", "speech_ne" was recognized as "speech", "wind" was recognized as "micro_blow" etc. However, there is some confusion that is harder to explain, for example "clap" recognized as "white_noise", and "hammer" recognized as "rustle". Noise overlap as well as the quality and the energy level of noise are possible reasons.

Table 2. *Noiseme Classification Accuracy wrt trial length*

Trial Duration (seconds)	Classification Accuracy (%)	Num of trials
0-1	**52.9%**	855
1-5	**55.1%**	1326
5-10	**63.1%**	198
10-15	**83.3%**	36
15-20	**87.5%**	16
20-25	**93.8%**	16
25-30	**85.7%**	14
>30	**95.9%**	49

4. Conclusions

In this paper, we introduced the concept of "noisemes" as a generalization for segmenting audio into speech and non-speech, which we not only use as a pre-processing step for speech recognition, but also to extract information from the audio track itself. We describe a process to extract and name salient features from the audio signal and from audible characteristics only, without being influenced by the context given by visual information. This allows us to extract robust labels, which can be detected automatically to create a sound signature of a video, which then can be used for the classification into events. We will continue to refine our automatic classifiers, and work towards linking these audio labels with semantic information that can be derived from the video, or from text that describes the event: i.e. while the "engine_heavy" noiseme is an abstract, sound-defined class, what context is needed to determine that a particular instance is a truck versus a fixed machine in a factory.

5. Acknowledgements

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

6. References

[1] Bregman, A., "Auditory Scene Analysis," MIT Press, Cambridge, 1990.

[2] Ellis, D., "Scene Analysis for Speech & Audio Recognition", http://www.ee.columbia.edu/~dpwe/talks/MIT-2003-04.pdf 2003-04-16.

[3] Ma, L., Smith, D. J., Milner, B. P., "Environmental noise classification for context-aware applications," In Proceedings of the International Conference on Database and Expert Systems Applications (DEXA), Lecture Notes in Computer Science, vol. 2736, 360—370, 2003.

[4] Temko, A., Nadeu, C., Macho, D., Malkin, R., Zieger, Ch., Omologo, M., Editor: Waibel, Al., Stiefelhagen, R., "Acoustic Event Detection and Classification," in Computers in the Human Interaction Loop, Human–Computer Interaction Series, 2009 Springer London, Isbn: 978-1-84882-054-8, p.61-p.73, 2009.

[5] Watanabe K., Shimojo, S., "When sound affects vision: effects of auditory grouping on visual motion perception," Psychological Science, vol. 12, no. 2, pp. 109–116, 2001.

[6] Temko, A., Nadeu, C., "Classification of acoustic events using SVM-based clustering schemes, Pattern Recognition, 39 (4) (2006), pp. 682–694, 2006.

[7] Andersson, T., "Audio Classification and content description," Lulea University of Technology, Multimedia Technology, Ericsson Research, Corporate unit, Lulea, Sweden, March, 2004.

[8] Cano P., Koppenberger, L., "Automatic sound annotation," Proc. IEEE Workshop Mach. Learn. Signal Process., p.391 , 2004.

[9] Burger, S., "The CHIL RT07 Evaluation Data, " in Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007, LNCS4625, p. 390-400, Baltimore, 2007.

[10] Ma, X., Fellbaum, Ch., Cook, P.R., "SoundNet: investigating a language composed of environmental sounds," In Proceedings of the 28th international conference on Human factors in computing systems (CHI '10). ACM, New York, NY, USA, 1945-1954, 2010.

[11] TRECVID: http://www-nlpir.nist.gov/projects/tv2011/tv2011.html. http://www-nlpir.nist.gov/projects/tvpubs/tv11.slides/tv11.med.slides.pdf

[12] Boersma, P., Weenink D., "Praat: doing phonetics by computer" (Version 5.3) [Computer program]. 2009. Retrieved March, 2011, from http://www.praat.org/

[13] Mesaros, A., Heittola, T., Eronen, A., Virtanen, T., "Acoustic event detection in real life recordings," 2010 European Signal Processing Conference (EUSIPCO-2010), 2010.

[14] Reynolds, D., Rose, R., "Robust Text-independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, Jan. 1995, pp. 72-83, 1995.