DATA-DRIVEN AUDITORY CONTRAST ENHANCEMENT FOR EVERYDAY SOUNDS AND SONIFICATIONS

Thomas Hermann

Ambient Intelligence Group
CITEC, Bielefeld University
Bielefeld, Germany
thermann@techfak.uni-bielefeld.de

Marian Weger

Institute for Electronic Music and Acoustics (IEM)
University of Music and Performing Arts
Graz, Austria
weger@iem.at

ABSTRACT

We introduce Auditory Contrast Enhancement (ACE) as a technique to enhance sounds at hand of a given collection of sound or sonification examples that belong to different classes, such as sounds of machines with and without a certain malfunction, or medical data sonifications for different pathologies/conditions. A frequent use case in inductive data mining is the discovery of patterns in which such groups can be discerned, to guide subsequent paths for modelling and feature extraction. ACE provides researchers with a set of methods to render focussed auditory perspectives that *accentuate inter-group differences* and in turn also enhance the *intra-group similarity*, i.e. it warps sounds so that our human built-in metrics for assessing differences between sounds is better aligned to systematic differences between sounds belonging to different classes. We unfold and detail the concept along three different lines: *temporal*, *spectral* and *spectrotemporal* auditory contrast enhancement and we demonstrate their performance at hand of given sound and sonification collections.

1. INTRODUCTION

The human auditory system is an amazing information data processor that has both phylogenetically and ontogenetically shaped to make sense out of the sounding world around us [1, 2]. Thus it is tuned for characteristics of sound as we encounter it in the world, be it music, language, soundscapes or interaction sounds, and it provides a mapping from sound space to meaning, i.e. it enables us to extract relevant information from the sounds. Sonification connects to these perceptual resources by providing a transformation of data into sound such that listening will in turn allow us to learn about the patterns in given data [3, 4]. Let's assume we are given a collection of data sets from patients under different conditions. Ideally, sonifications will represent the data so that meaningful differences in the data are perceivable. However, this requires that sonification designers know the pattern already before creating the sonification. While this may be true for sonifications that communicate information, it is not given in the case of exploratory data analysis [5, 6] where the goal lies in the discovery of hidden/unexpected patterns and which is inductive in nature. Hence by applying any given sonification method we will likely get sounds where a meaningful systematic difference may be not at all perceivable, or strongly masked by other less informative parts. Likewise in sound-based machine diagnostics or in auscultation, the overall sounds may include some features helpful for discrimination, yet they may be masked by acoustic elements that only aggravate discrimination. Luckily we are equipped with powerful perceptual skills for source separation and auditory focussing, yet these also have their limits. In summary, an inevitable problem both in sonification for inductive data mining and in real-world exploratory investigation is that relevant structures can be inaccessible as they are masked by irrelevant noise.

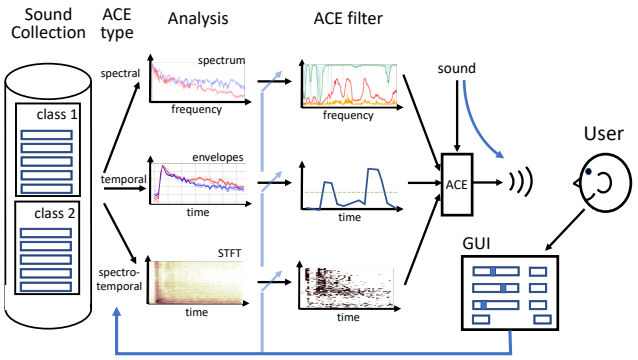

Figure 1: Auditory Contrast Enhancement improves inter-stimulus differences between sounds or groups of sounds.

Individual training, i.e. to rely on auditory learning alone, can empower listeners to better extract information in difficult situations, e.g. car mechanics become experts in associating sound patterns to engines condition, same as trained physicians learn what to attend to in auscultation to diagnose certain heart and chest problems. However, there is another issue: such implicit knowledge will be difficult to communicate to others, we lack a kind of *pointer into auditory structures* compared to the visual modality where we can more easily point our finger and thus share what we regard as relevant. The ACE presented in this paper, with its interactive controls will also serve as a novel kind of 'adjustable pointing device' that can direct novel listeners' attention to the relevant patterns in a complex sound/sonification, and thus help to better deal with the subjectivity of listening which still hinders scientific uses.

The trend in state-of-the-art modern diagnostics in our computer age, however, is to completely abandon the direct sensorial

contact with the raw data in favour of machine learning and AI to classify data and communicate the results in clear language. In a way, purely machine-learning based diagnosis 'throws the baby out with the bath water' by taking the humans and their domain expertise and broader knowledge out of the loop. This leaves the analyst out of touch with the details on which the classification is based, and it is prone to the risk of false positives and false negatives. However, it would also be suboptimal to leave the potential of machine learning unused. So our approach aims at an enhanced *human-machine cooperation*: (i) machine learning can provide a data-driven enhancement of sounds according to criteria derived from given (e.g. labeled) data. (ii) users can optimise the sounds interactively to further increase their contrast, (iii) this might trigger new ideas about relevant patterns and recursively lead to finer differential diagnosis, and result in suitable enhancement settings for practical applications.

We define *Auditory Contrast Enhancement* (ACE) as a system that transforms given input sound signals into enhanced output sound signals, which facilitates their perception and hence improves the conveyance of the underlying information. We differentiate between two types of ACE.

intra-stimulus contrast refers to the strengths of peculiarity of a single stimulus. It is conceptually similar to the visual domain where contrast refers to the degree to which areas of an image differ in luminance [7, p. 169]. Likewise spectrotemporal contrast can be enhanced in sound. This topic is extensively elaborated in our companion paper [8].

inter-stimulus contrast refers to systematically perceived differences between stimuli from two (or more) groups (A,B,...) accessible and assessed via their A-B comparison. Methods for this ACE will be introduced below.

The following differences further motivate to split the topic of ACE into two papers: as intra-stimulus ACE does not depend on any other data, it can enhance structure from an unfolding sound in real-time, and thus it can serve as a non-parametric post-processing plugin for interactive exploration practises such as percussion & auscultation, and be used in auditory augmentation and blended sonification [9, 10]. In contrast the inter-stimulus ACE depends on given samples at hand of which the detailed processing is crafted. The processing is then applied to either the given input samples, or could also be applied to other independent samples. Interactive uses of data-driven ACE for interactive applications is not so much a focus of this paper, but could be a promising continuation that merges both works.

For inter-stimulus ACE, we distinguish two special cases: (i) *supervised ACE learning* refers to a situation where a number of stimuli are given with their known attributes (e.g. class label), and (ii) *unsupervised ACE learning* has to base the ACE solely on a set of given sounds without knowledge of a ground truth interpretation. The paper mainly unfolds supervised ACE learning and only sketches concepts for unsupervised ACE learning.

Section 2 will formally introduced ACE followed by the presentation of ACE methods in Sec. 3. An implementation of the methods in python will be shown in Section 4. Section 5 will introduce a number of sound and sonification collections and demonstrate the ACE types at hand of these. This will lead to the discussion and conclusion.

Sound and sonification examples are provided as supplementary material via the following DOI: 10.4119/unibi/2935744.

2. DATA-DRIVEN AUDITORY CONTRAST ENHANCEMENT

We define Auditory Contrast Enhancement (ACE) as a system that transforms given input sound signals into enhanced output sound signals, which facilitates their perception and hence improves the conveyance of the underlying information.

We further define *inter-stimulus ACE* as a data-driven method that optimises the enhancement processor from a collection of sound recordings where sounds exhibit systematic differences. We distinguish the supervised learning situation where the correct label or attribute is known and the unsupervised learning situation where no such labels exist. We have the case of classification problems if the label is binary[1], or the case of regression, if a continuous variable describes the variation.

We first unfold ACE at hand of a collection of samples with a binary class label. With this focus, the two main goals of inter-stimulus ACE are (i) to enhance our ability to discriminate sounds that belong to different classes and (ii), to eliminate those parts (spectral, temporal or spectrotemporal) of the sounds that don't contribute to their discrimination. These goals should be reached under the following given conditions:

- the ACE should converge with increasing amount of training data.
- after a training phase, the ACE should be applicable to any previously unseen test data, and thus generalize beyond seen data.
- ACE application should yield a sound that still is an analogic (raw) representation of the underlying stimuli, according to Kramer's continuum [3].
- sound discrimination has priority over structural integrity: it is acceptable if resulting sounds are modified to become even not recognizable as the sounds before ACE application, as long as contrast is increased.

Practically, data-driven ACE is a software method to manipulate given sound signals consisting of parts: (i) a module to analyze a given collection of sound samples, either labeled or unlabeled resulting in a set of ACE features that allow to discriminate between relevant and irrelevant parts of the signal if it comes to perceive inter-stimuli differences. (ii) a module to apply the ACE features to a sound signal. (iii) a user interface that enables users to interactively select the ACE method and adjust any parameters involved in the transformation.

For (i) and (ii), we introduce Spectral ACE, Temporal ACE and the Spectrotemporal ACE in the subsequent sections. The relevant parameters for (iii) will be introduced along. The graphical user interface will be described in Sec. 4.

2.1. Problem Statement and Sound pre-processing

First let's formally introduce some nomenclature. We assume that a sound signal $s[n]$ is given which contains a sequence of sound events that clearly stand out from background noise. We assume that we have $m = m_1 + m_2$ sound events where m_i is the number of events belonging to class i. W.l.o.g. we can assume the sounds to be ordered. We limit our discussion first to binary classification settings, i.e. $i \in \{1, 2\}$. For example consider the case of judging

[1] more generally: discrete, yet in this paper we limit the treatment w.l.o.g. to binary classification problems

whether a wall is hollow or solid behind the wallpaper: we could have 10 impact sounds for knocking on the solid and hollow wall each as our collection.

As a first step we have to extract the individual sound events from the input signal. The onsets need to be properly aligned, at least for some of the ACE methods introduced below to work well.

To this end we compute the signal root mean square (RMS) of 1 ms analysis windows and accept it as event onset if a silence threshold, e.g. -20 dB, is exceeded. The end of an event is defined by the RMS staying below that threshold longer than a given silence time. Events are extended left and right with some milliseconds to make sure no transients are lost. The resulting events are $s_i^{(1)},\ \ i \in \{1, \dots m_1\}$ and $s_i^{(2)},\ \ i \in \{1, \dots m_2\}$. Furthermore, at this time we truncate all sounds to the smallest common duration. Alternatively it would be possible to use zero padding of shorter sounds. As another option, a set of sound files with proper alignment can be directly loaded.

Note that the unsupervised ACE learning will only have a single set of events to work with and no further label, but this is left for Sec. 6.

3. METHODS FOR CONTRAST ASSESSMENT

In this section we introduce three approaches for measuring contrast between groups of sounds: *Spectral contrast* ignores temporal patterns and identifies frequencies at which the groups differ. *Temporal contrast* only evaluates the temporal evolution of a signal and identifies temporal segments at which the groups differ. Finally *spectrotemporal contrast* assesses systematic differences from the time-frequency analysis of signal collections using the Short-term Fourier transform (STFT).

3.1. Spectral Contrast

Yang et al. define spectral contrast as "the decibel difference between peaks and valleys in the [magnitude] spectrum" [11]. This definition, however, refers to contrast within a sample and is what we explore in the companion paper [8]. Here, we have to rethink the notion of contrast from the viewpoint of perceptual contrast between juxtaposed sounds $s_1^{(1)}, s_1^{(2)}, s_2^{(1)}, s_2^{(2)}, \dots$

Obviously we gain perceptual spectral contrast if we attenuate those frequencies at which the two collections of sound do not differ, and if we boost those frequencies at which they do. The spectral ACE thus simply becomes a filter.

This method will work well for instance with sounds whose spectral profile is rather constant. For instance, impact sounds such as hitting a kettle with a stick are characterized by a relatively stable spectrum determined by the physical invariance of the kettle shape. The initial excitation quickly excites a set of rather stable partial tones that decay with time. In contrast, if sounds exhibit substantial spectral changes over time, such as in a piece of music, spectral contrast will be a less usable.

Practically, we compute the one-dimensional discrete Fourier Transform for real input using the FFT algorithm. The complex-valued spectra $S_{j,k}^{(c)}$ for all given events j within class c at frequency cell k are stored for analysis as column vectors in a matrix $X^{(c)}$. The matrix $Y^{(c)} = |X^{(c)}|$ holds the spectral magnitude for all frequencies (in rows) and for all sounds (in columns).

For a given frequency (i.e. row k) the values of $Y_{*,k}^{(c)}$ represent the spectral energy in class c. We assume these values to be independent samples of an underlying unknown distribution. Under the null hypothesis H_0 that there is no difference between group $c = 1$ and $c = 2$ we can ask how likely it is that we observe the empirical means

$$\mu^{(c)} = \frac{1}{m_c} \sum_{i=1}^{m_c} Y_i^{(c)} \tag{1}$$

Under certain conditions, the normalized difference

$$t = \frac{|\mu^{(1)} - \mu^{(2)}|}{\sigma_{\text{err}}} \tag{2}$$

would be student-t distributed, allowing to compute the p-value, i.e. the probability of type-1 error of erroneously concluding a systematic difference while there is none. Hence statistical testing can help to identify if there is enough evidence to assume the spectral energy to be systematically different, or whether observed differences could be simply a product of random sampling.

Let's not engage in deeper statistical interpretation of the value of t and instead use the t-value simply as calibrated indicator for differences: t is simply the difference of the means in multiples of the joint samples' standard error. This is a useful criterion to adopt for spectral contrast. And regardless of any assumptions on the underlying distribution or statistical interpretation we can simply compute the vector $\vec{t}$ of t-values for all rows of $(Y^{(1)}, Y^{(2)})$.

The t-vector is the point of departure for defining the spectral ACE filter. Specifically we introduce two filters:

nonlinear spectral ACE Here we apply a nonlinear transfer function to $\vec{t}$ so that low values are drawn to 0 and large values soft clip to 1. We propose the transfer function

$$T_k = \tanh(g_{nl} \cdot |t_k|)^o \tag{3}$$

for all frequencies k using a user-adjustable nonlinear gain g_{nl} and order o as exponent for supressing frequencies that do not likely contribute to inter-group differences. Before filtering, T is normalized to maximum 1.

median-filtered t Here we first apply a median filter of user adjustable size r to the sequence $|t_k|$ and define the filter as

$$T_k = \text{median\_filter}(|t_k|, r)^o \tag{4}$$

for all frequencies k, again normalizing T to maximum 1. The median filter smoothes the spectral resonances which can be rather sharp, resulting in strong ringing after the inverse FFT. The median filter thus improves the temporal structure of the ACE-filtered sounds.

As T_k has the same dimension as the initial spectral vectors we can obtain the Spectral-ACE-filtered signal by

$$s_{\text{ace}} = \text{irfft}(\vec{T} \cdot \vec{S}) = \text{irfft}(\vec{T} \cdot \text{rfft}(s)) \tag{5}$$

where '$\cdot$' refers to the elementwise product of the two vectors. Note that S needs to be resampled if applied to signals s of different lengths. However, shorter input signals s can be zero-padded so that the available $\vec{T}$ works.

3.2. Temporal Contrast

Sound evolves in time and the temporal evolution of a sound's amplitude is a feature in which sounds can be different. For example two physical objects may differ in their internal damping and thus impact sounds with the objects may lead to different amplitude falloff over time, maybe so faint that we might overhear it.

Another example is cyclical machine sounds, e.g. from a printer or engine where wear and tear might change the friction and thus sound level over the cycle, which in turn would result in subtle difference compared to the sounds of new machines. With temporal contrast we aim at accentuating such moments in time.

To define temporal ACE for inter-stimulus contrast, let's quickly summarize the essential idea of spectral ACE, in order to define temporal ACE in analogy. In spectral ACE, we searched in *spectrum* for evidence that energies are different and accentuated those where a systematic difference was likely and removed the other frequencies. Likewise, for temporal ACE, we can search *along the time axis* for evidence that the energies are different and accentuate those times where a threshold evidence is exceeded, and remove all other times. This translates into two questions: how to define energy difference for a given time, and how 'to remove time'. For the first issue, we see that an instantaneous energy does not exist and is only defined in a short time span. The RMS of the signal is a good estimator. Practically, we use a triangle window of size 256 (i.e. 6 ms at 44100 Hz sampling rate) with a stride of 128. Figure 2 depicts the individual envelopes of 10 impact sounds (5 per group 'on wood' and 'on metal' each). It is visible that there are times at which the amplitudes differ systematically. The thick lines show the mean envelope of the two groups.

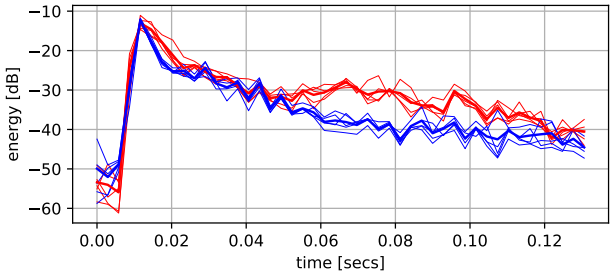

Figure 2: Energy envelopes of 10 impact sounds: 5x 'wood' (red), and 5x 'metal' (blue), the thick lines represent the mean energy over time. It can be seen that there are times at which the values differ systematically.

With these signals we can compute – in analogy to the frequency k in spectral ACE now for each time window n – the vector $\vec{t}$ of t-values

$$t_n = \frac{|\mu_n^{(2)} - \mu_n^{(1)}|}{\sigma_{\mathrm{err},n}} \tag{6}$$

which quantifies the class mean difference in multiples of their pooled standard error.

As to the second question, how to 'remove time', our first attempt was to suppress the energy for those times where t is low. However, this resulted in sparse sounds where much time was wasted with silence in unnecessarily long A-B-A-B comparison sequences. It makes sense to literally 'remove time' as the term indicates and cut and concatenate only those temporal segments in which differences stand out. This results in much shorter sounds, faster to compare and evaluate. Note, however, that this may make a signal completely different and even incomprehensible, for instance if the rhythm matters. However, it saves time if the interest is to discriminate groups.

As soft form of cutting time, instead of a binary decision, we considered a temporal warping that plays signal parts faster as they contribute little and slower as there is more difference. However, this has not been fully tested yet, and may even be irritating as it distorts the temporal structure further.

To decide which time segments to keep, we do not need a nonlinear transfer function as used in spectral ACE, as such a function should be monotonous anyway and thus wouldn't affect the result of a simple threshold operation apart from warping the threshold values as such. Thus we merely select times by taking all windows where $t_n > \theta_{\text{t-ACE}}$. As a rule of thumb, values around 3 would correspond to a 1% chance that the observed difference occurs randomly without significant differences in the means. However, take this with a grain of salt, as in this method a large number of t-tests are computed and no Bonferroni nor other correction is done, so a proper statistical interpretation is not possible.

Furthermore note that a proper temporal alignment and signal normalization is crucial for this method to give meaningful results: a slight shift of the signal in time would create large differences, which of course are not relevant, likewise would a set of louder or more quiet sounds between classes. We currently normalize sound events for peak amplitude 1, yet we see that this is not very robust to outliers. A normalization for the overall event energy as integral over time might be more meaningful in such situations. For stationary/cyclical sounds we recommend to establish temporal alignment from correlation analysis between signal energy amplitudes and choose the lag that yields maximum value.

Note furthermore that the t computation may yield NaN if means are exactly the same, which we replaced by zero for subsequent ACE computation and plotting.

3.3. Spectrotemporal Contrast

The previous two approaches have derived their evidence for systematic differences between the two groups of stimuli from spectral (resp. from temporal) energy alone. Spectrotemporal ACE combines both approaches, yet not in a sequential cascade-style fashion but by directly deriving the ACE criterion from a spectrotemporal analysis of the signals, commonly known as spectrograms. The Short-term Fourier transform (STFT) generates a representation where time windows of the signal are spectrally analyzed and thus a 2D-array of complex numbered activity within all time/frequency cells is computed. The magnitude of these values are the basis for the spectrogram. Fig. 3 depicts the mean arrays for all instances in the impact sounds on wood and metal. As these resolve both spectrum and time they are a more infor-

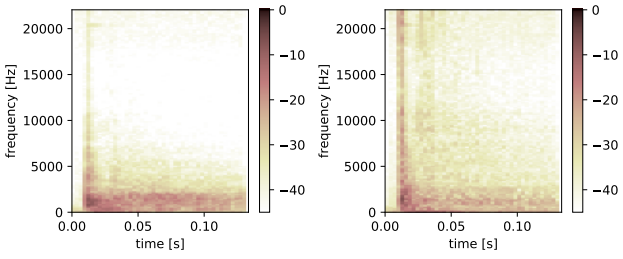

Figure 3: Mean magnitude-STFT-levels (in dB): plots for the impact sound classes wood (left) and metal (right)

mative source for evidence of systematic difference between two given sound collections.

For our signals at sampling rate 44100 Hz, an FFT size of 256 and a temporal stride of half the window size, i.e. 128, is used, using a cosine bell (Hann) window. The resulting spectrograms for example j in class c, named here $S_j^{(c)}[k,n]$ are functions of the frequency cell k and time segment n.

In analogy to the previously introduced ACE approaches we here compute as source for evidence of systematic inter-stimulus variations

$$t_j^{(c)}[k,n] = \frac{|\mu_{k,n}^{(2)} - \mu_{k,n}^{(1)}|}{\sigma_{\text{err},k,n}} \tag{7}$$

with help of the intra-class means of the STFT magnitudes at each given $[k,n]$. Figure 4 depicts the resulting t-array values for the two given vowel sounds in Fig. 3. Obviously the differences in

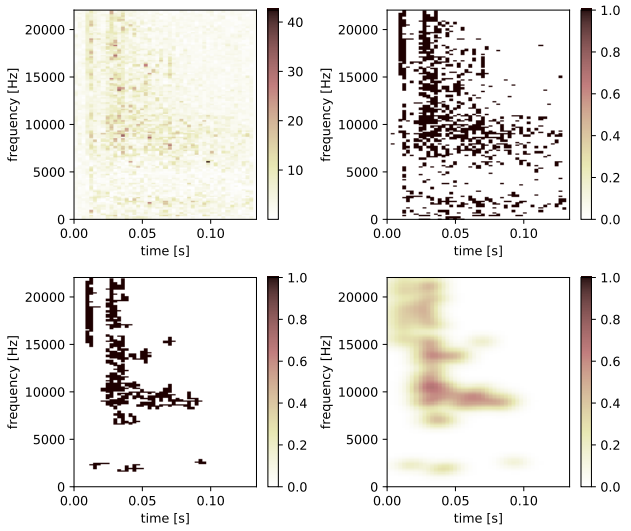

Figure 4: Plot of the 2D array of STFT cell-wise t analysis for wood vs metal impact sounds (whose means are depicted in Fig. 3. Differences in hf-signal stand out.

location and extent of formants are correctly analyzed.

Same as with temporal ACE, a good temporal alignment of any sounds in the two groups is required for the analysis to yield usable results. Such an alignment is easily obtained in the case of impact sounds due to the defining initial transient, and for sonifications where of course the sonification time is well controlled and known. It is less clear how to apply spectrotemporal ACE to stationary, patterned sounds such as cycling machine sounds, yet the same heuristics suggested for temporal ACE can be applied, to shift stimuli onsets in search of the least overall RMS of the t array.

As for the ACE, we proceed in analogy to spectral ACE. We derive a (now spectrotemporal) weighting array $\mathbf{w}$ for all time-frequency cells of the sounds and obtain the enhanced signal by applying the inverse STFT to the weighted STFT array

$$s_e[n] = \text{ISTFT}(\mathbf{w} \cdot \text{STFT}(s[n])) \tag{8}$$

For the weighting array we can take, in analogy to the spectral ACE, a nonlinearly warped t-array, for instance as

$$\mathbf{w}[k,n] = \begin{cases} 1 & \text{if } t[k,n] > \theta \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

which is shown Fig. 4 in the upper right for a threshold $t = 6$.

It turns out that due to the statistical nature of the many tests isolated pixels (cells in the STFT) are frequently supra-threshold. To remove these while retaining all larger blobs we can apply morphological operations from computer vision, namely a binary opening operation followed by a binary erosion and a binary propagation.

The result of these operations is shown in Fig. 4 in the lower left plot. To soften the weighting mask we furthermore apply a gaussian filter (scipy.ndimage.gaussian_filter) to blur with a user-controllable bandwidth σ resulting in a filter as depicted in Fig. 4 (lower left) for $\sigma = 2.5$. Note that σ is in units of pixels and equal bandwidth for spectral and temporal smoothing is currently taken.

Finally, the spectrotemporal ACEd signal is mixed with the original signal, so that any isolated enhancements are better contextualized through the original audio signal.

Note that different from temporal ACE, here no time is removed yet. This would be an additional and optional operation which would require a further criterion for excluding a time frame. The conservative approach would be to exclude only those time segments that do not have a single entry in the ACE mask after erosion. Instead of integrating this into the spectrotemporal ACE, however, an alternative procedure would be simply to cascade the temporal ACE described before.

4. IMPLEMENTATION

We implemented the ACE with python using numpy.ndarrays for audio signal representations and scipy functions to compute spectrum, STFT, t-values, and to apply morphological operations. As standard operations on audio signals is a bit tedious with plain python/numpy/scipy, a dedicated python audio coding package named pyA has been implemented by the first author. It will be made public on github and described elsewhere. With pyA, the necessary operations can be written in a very visible pythonic coding style.

For interactive testing, we developed a graphical user interface within the Jupyter ipython environment as shown in Figure 5. Basically, the user can specify audio files that include the sequence of sounds. Ideally these are separated by some silence or background noise so that the peak finder can identify them. The current code assumes q examples for class 1, followed by any number of examples for class 2. The GUI depicts in the upper row of plots the input signal, here showing 10 impact sounds on a table, five on wood, five on aluminum. The right panel depicts the results of the event finder, blue for class 1 and red for class 2. As the depicted GUI is for spectral ACE, the panel below shows the spectral mean of all q class 1 (red) and class 2 (blue) signals. Note that a truncation to common lengths is applied before this analysis. It can be seen that there are systematic differences. The plot below shows the spectral t analysis, which peaks at different frequencies. The following GUI elements allow to select ACE subtype and parameters, their changes causing an update of the bottom plot of the spectral ACE filter for weighting the original signal before re-synthesis. While much of that can probably be hidden from users in automatic ACE modes, it was helpful to inspect the details for development.

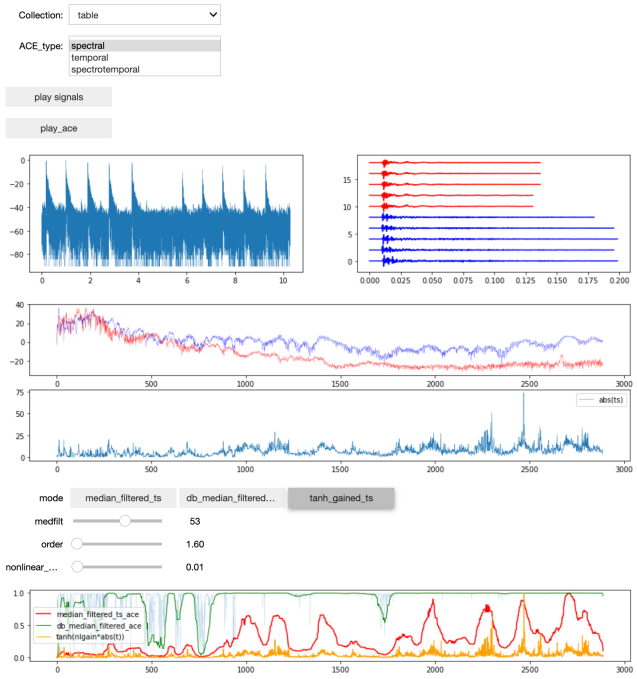

Figure 5: python GUI screenshot: users can select predefined sound collections and the ACE type, then play either original or enhanced sounds. Plots below provide some insight into differences between groups and allow to control ACE-type specific parameters.

5. ACE DEMONSTRATIONS

This section demonstrates the previously introduced ACE types at hand of some examples using different collections of sounds and sonifications.

5.1. Impact sounds

Impact sounds contain a lot of information usually in a very short time of few hundred milliseconds. They convey the material, the object size and resonances, friction and inner properties such as composition. For example, knocking on a half-filled bottle will sound different if filled with water or rice. We see further potential for ACE for auscultation via percussion on the human body.

Here we show ACE performance for the following sound collections (sound examples available on 10.4119/unibi/2935744):

table a collection of impact sounds from hitting a wooden table or on an aluminium laptop with a plastic ball pen used as mallet.

finger snaps a collection of finger snaps from one person using middle vs. ring finger

SC3 klank+noise a collection of noisy synthesized resonator banks using Supercollider's 'Klank' UGen with added Brownian noise, where the two classes differ in two detuned resonance frequencies.

5.2. Continuous sounds

Continuous and cyclical sounds are frequent in machines or rhythmic motor patterns such as sonification of swimming or repetitions of physiotherapeutic practices. For demonstration we here use

vowel a collection of vowel sounds with slightly different articulation place so that it is difficult to distinguish them

music a collection of modifications of a music signals where generally noise is added but temporally localized gains were applied as systematic difference.

5.3. Enhancing the discrimination of impact sounds

Let's start with the 'table' collection and listen to the original impact sound collection in S1.0 [2]. We can clearly perceive the differences without problem, so let's test how spectral ACE will accentuate them. With low order using the median-filtered-ts, we get sound example S1.1 which features very salient differences in response in high frequencies, while low frequencies are attenuated as there is no difference indication between the two groups. This aggravates generally as the exponent parameter 'order' is increased until only narrow ring resonances remain, see S1.2 and S1.3. The nonlinear spectral ACE which uses a $\tanh()$ warping of the absolute t-values is not as radical and leaves more of the low-impact parts intact, depending on the nonlinear gain parameter. However, because of the lacking median filtering, the resonances are very sharp, resulting in long ringing, so that the resonances have almost no damping and thus amplitude differences in these between the two groups stand particularly out at the expense of perceptibility of transients. Next example is S1.4, where a logarithmic (dB) mapping from the ACE feature to gain was applied, but on $(1-p)$-values. This maintains more of the original structure yet accentuates the differences quite well. Note that the same ACE filter is used identically for all 10 sounds, and is also capable to process new not-before-heard sounds and delivers an equally salient perceptual contrast.

Temporal ACE is not as good as expected – however, the physical merely exponential decay doesn't yield so pronounced differences. Sound example S1.5, however, shows at least how focussing on the segments of highest differences results in a notable shortening (and thus speeding up) of collection review.

The spectrotemporal ACE (examples S1.6– S1.9) combine advantages of spectral and temporal ACE in that they are capable to attenuate now spectrotemporal uninterestingness and thus render inter-stimulus differences more salient. S1.6 – S1.7 use the thresholded t-value array directly as filter, whereas S1.8 – S1.9 make use of larger σ to smooth the ace-filter.

The following sound examples are for the finger snap sound collection. Listen first to example S2.0 for the original collection. Examples S2.1 – S2.4 apply spectral ACE with different parameters, while example S2.5 uses the spectrotemporal ACE. The extreme transience of the sounds makes it hard for the STFT-based approach to resolve enough structure, for the same reason the temporal ACE delivers rather poor results.

5.4. Enhancing and Isolating structure from noise

The synthetic Klank sound collection (original sounds in S3.0) features a large amount of noise, capable of masking subtle dif-

[2] All sound examples are provided with description on `https://doi.org/10.4119/unibi/2935744`

ferences between the groups. Apparently, spectral ACE is well capable to attenuate most noise channels, more radical as the order parameter as exponent is increased, audible in examples S3.1 to S3.4. At extreme setting, only the two frequencies which were actually changed between the class 1 and 2 in this collection remain, showing clearly that spectral ACE was successful in finding and highlighting these. Example S3.5 shows that spectrotemporal ACE performs less well here as noise level fluctuations over time result in equal energy over time at relevant frequencies, giving raise to significant temporal structure.

So far spectral ACE seems to have some advantages, but unfortunately it strongly affects the temporal structure due to long resonances.

5.5. Formant changes in continuous signals

The next set of examples test ACE with speechlike sounds. The sound collection 'vowel' features articulatory sounds with slightly different articulation place (listen to example S4.0 for the unmodified sounds). The first 5 are an 'a' as in 'bar' followed by an 'ä' like in 'bear' but tried to articulate more similar to the first vowel. Perceptually they can be quite easily discerned, so let's see how ACE is able to boost the differences. Example S4.1 is the spectral ACE using the $\tanh()$ weighting with low order and low nonlinear gain. It has already shown to give long ringing for resonant frequencies. We hear that the temporal structure is largely lost, but the contrast between class 1 and 2 increases clearly. The same happens with the median-filtered-t-values mode (example S4.2). Here, the differences at high frequencies are strongly enhanced, resulting in audible differences. However the low frequency content did not pass through and thus the original formant structure is barely perceivable. Yet we argue that this doesn't matter if the focus is on classifying sounds as belonging to either the one or other class. In comparison, spectrotemporal ACE (Sound example S4.3) is rather useless and only creates a rougher and noisier version of the sound. The reason for that might be that the STFT number of samples per segment is low with only 256, thus resulting in a poor spectral resolution of 22050 Hz / 256 = 100 Hz, which is perhaps not high enough to distinguish formant differences between 'a' and 'ä'. Temporal ACE did not help at all, so we skip a sound example.

5.6. Enhance Multivariate Time-Series sonifications

Finally we test the ACE on sonifications. A frequent data type are multivariate time-series, such as EEG, ECG, EMG or motion capture sensor streams. For the example here we created a parameter mapping sonification of the building dataset [12] of hourly consumption of electrical energy, hot water, and cold water, time of day, outside temperature, outside air humidity, solar radiation and wind speed for 175 days, all variables scaled to arbitrary units in $[0, 1]$.

Since the purpose of this example is to test how ACE would enhance differences between groups, we created a rather straightforward parameter-mapping sonification of all variables as amplitudes of oscillators tuned to quart-spaced fixed frequencies. We chose 5 sunny days in summer for class 1 (days 13-17 in the building1 dataset) and 5 sunny days in late autumn, starting from midnight on day 141 in the dataset, skipping cloudy days as seen in the measurements of solar radiation, so that the groups are more homogenous. Each day is sonified in about 250 ms from midnight to midnight. The raw sonifications can be heard in sound example S5.0. As the frequencies are constant for each stream, spectral ACE can be expected to provide a good enhancer for systematic activation differences. In fact, examples S5.1 and S5.2 show that the differences in energy in the highest frequency oscillators are significant enough to constitute difference and are thus accentuated. We expected a better contrast in the solar radiation profiles as these are quite different in the different seasons. However, spectral ACE can't see their variation over time and only uses time-free spectral energies, so any differences here do not stand out. That is different in temporal ACE which accentuates certain parts of the signals, see examples S5.3. However, different from intuition which would rather expect differences over daytime to be accentuated, the temporal ACE pronounces differences before sunset and after dawn. The reason is likely that with 0 solar radiation, one variance source within those time windows is reduced, making those times appear more different than those times where solar radiation contributes to variance within the samples of each class. So temporal ACE does work, yet not necessarily as expected. It is a starting point but needs more research to design it to be more sensitive to changes in relevant structures occurring in sonifications. Finally a spectrotemporal ACE example is provided as examples S5.4

5.7. Detecting modifications in longer sound clips

The last example is a set of sounds where a snippet of music was systematically amplified or attenuated at different locations in time in the two classes. Listening to S6.0, the original collection of 2×3 events per class makes clear that it takes a long time to review many sounds, 2.5 s per sound each. Temporal ACE reduces these sounds by removing all those time segments where no systematic differences in amplitude between the two groups can be found, as evaluated by the t-value of the two samples. In turn, the resulting sound is shortened to about 250 ms, depending on the t-threshold, allowing a much faster review of the sound examples S6.1 and S6.2. Also, the differences between the groups becomes clearer: a boosting of the second chunk while attenuating the first between class 1 and class 2.

6. DISCUSSION

We have introduced *data-driven ACE* as a method to automatically modify audio signals so that a contrast between given classes becomes more salient. As first step, we presented spectral, temporal and spectrotemporal ACE and gave exampled for a number of sound collections with systematic differences between two groups. While most sound examples gain perceptual contrast, particularly the spectral ACE has subjectively proven most useful to help discriminating sounds into the two classes. One problem with the current approach is that still the method depends on a number of parameters that cannot be automatically chosen easily, so it requires the human in the loop to tune parameters for good results. Yet this might be acceptable as this would only be required once, e.g. for the designer of a tool to enable machine diagnostics-by-listening. However, it would be certainly nice to integrate some good heuristics for automatic parameter selection, e.g. from testing all parameter combinations on a grid and applying a machine listening based contrast assessment to choose useful initial settings.

Temporal contrast was demonstrated to work, and it is useful to reduce long sound signals into short 'difference thumbnails'

which only present those parts in which differences may lurk.

Interestingly the ACE modifications can be applied to any input signal: spectral ACE independent on signal duration, temporal ACE independent on sampling rate, only spectrotemporal ACE requires matching duration and sampling rate. That means that an ACE trained to enhance contrast between two extremes such as different pathologies reflecting in chest tones, or different materials behind the surface while knocking on a wall, can also be applied to any signal with unknown label. It will be a useful experiment to measure how ACE will reshape the accuracy of classification particularly for sounds that are on the continuum between the two extremes used for ACE training. As a preliminary first test for this, we applied the ACE on a continuous transition of vocal sounds continuously varying from 'a' to 'ä'. Sound examples S4.4 and S4.5 show the original and the enhanced version. We see that more research is needed and more experience needs to be gained with ACE. One possible study could be to ask subjects to assign ACEd sound examples to classes. The mean of ratings for stimuli along the connecting line between class 1 and 2 would probably be somewhat sigmoidal without ACE, yet it should move towards a steeper sigmoidal function with proper ACE.

The data-driven approach to ACE can be extended to work in *unsupervised learning* settings. Consider we have a collection of sounds yet no class label. Assume further that there are systematic differences in the sound features. If intra-class variation is lower than inter-class variation, the first eigenvector of the feature data set covariance matrix (i.e., the first principal axis) should be aligned to the line connecting the centroids of the two clusters. A useful ACE could be derived from that information alone. For instance assume that the features would be the magnitude spectrum components. Then the PCA vector $\vec{u}$ would show certain positive or negative elements. If we take the ACE to be 0 for those frequencies where $|u_i| < \theta$, i.e. is smaller than a threshold θ and 1 else, we would filter out those frequencies that do not change much along the main variance axis of the data. In turn, the remaining frequencies will become more salient. This and more refined approaches for extending data-driven ACE to unsupervised learning remain subject of future research.

7. CONCLUSION

Auditory Contrast Enhancement has been introduced in this paper as a method to process sound in general, and sonifications in particular with the goal to facilitate the perception of relevant sonic differences between selected groups of sound, e.g. created under a different condition. We have focussed on the three special cases of spectral, temporal and spectrotemporal contrast and introduced data-driven enhancements that transform sound in a systematic and reproducible way. The presented ACE processors provide a supervised-learning method yet instead of merely reporting the results as text, they provide an interactive sound manipulation method to better use the human-built-in listening skills to distinguish patterns in data. ACE is capable of removing signal components that apparently do not contribute to any differences between selected groups, and of actively boosting signal parts where differences between groups are likely. In consequence, the resulting signal is less prone to masking. We believe that ACE can serve as a widely applicable sound post-processor for many situations where sounds are perceived in the listening mode of diagnostics or exploration. A thorough psychophysical validation of the method will be required to optimize the methods further, yielding suitable control parameters and interfaces that establish ACE as a standard plug&play post-processing component for sonification tool chains.

8. ACKNOWLEDGMENT

This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG). Thanks to the developers of python/numpy/sciepy/matplotlib/jupyter for their amazing tools.

9. REFERENCES

[1] R. Fay and A. Popper, "Evolution of hearing in vertebrates: The inner ears and processing," *Hearing research*, vol. 149, pp. 1–10, 2000.

[2] E. Hester, "The evolution of the auditory system: A tutorial," *Contemporary Issues in Communication Science and Disorders*, vol. 32, pp. 5–10, 2005.

[3] G. Kramer, Ed., *Auditory Display - Sonification, Audification, and Auditory Interfaces*. Addison-Wesley, 1994.

[4] T. Hermann, "Taxonomy and Definitions for Sonification and Auditory Display," in *Proceedings of the 14th International Conference on Auditory Display (ICAD 2008)*, P. Susini and O. Warusfel, Eds. Paris, France: IRCAM, 2008.

[5] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[6] T. Hermann, "Sonification for Exploratory Data Analysis," PhD Thesis, Bielefeld University, Bielefeld, Germany, Feb. 2002.

[7] A. M. Colman, *Oxford Dictionary of Psychology*, 3rd ed. Oxford University Press, 2009.

[8] M. Weger, T. Hermann, and R. Höldrich, "Real-time Auditory Contrast Enhancement," in *Proceedings of the 25th International Conference on Auditory Display*, Newcastle, U.K., 2019.

[9] T. Bovermann, R. Tünnermann, and T. Hermann, "Auditory Augmentation," *International Journal on Ambient Computing and Intelligence (IJACI)*, vol. 2, no. 2, pp. 27–41, 2010.

[10] R. Tünnermann, J. Hammerschmidt, and T. Hermann, "Blended Sonification: Sonification for Casual Interaction," in *ICAD 2013 - Proceedings of the International Conference on Auditory Display*, 2013, pp. 119–126.

[11] J. Yang, F.-L. Luo, and A. Nehorai, "Spectral contrast enhancement: Algorithms and comparisons," *Speech Communication*, vol. 39, no. 1-2, pp. 33–46, 2003.

[12] L. Prechelt, "PROBEN1 - a set of neural network benchmark problems and benchmarking rules," Universität Karlsruhe, Karlsruhe, Tech. Rep. 21/94, 1994.