Automatic recognition of accessible pedestrian signals

Sebastián Ruiz-Blais, Arturo Camacho-Lozano, and Juan Fonseca-Solís

Citation: Proc. Mtgs. Acoust. **30**, 055015 (2017); doi: 10.1121/2.0000675
View online: https://doi.org/10.1121/2.0000675
View Table of Contents: https://asa.scitation.org/toc/pma/30/1
Published by the Acoustical Society of America

Proceedings of Meetings on Acoustics

Volume 30 http://acousticalsociety.org/

Acoustics `17 Boston

173rd Meeting of Acoustical Society of America and 8th Forum Acusticum
Boston, Massachusetts
25-29 June 2017

Signal Processing in Acoustics: Paper 4pSPb2

Automatic recognition of accessible pedestrian signals

Sebastián Ruiz-Blais
Research Center in Information and Communication Technologies, University of Costa Rica, Montes de Oca, COSTA RICA; sebastian.ruizblais@ucr.ac.cr

Arturo Camacho-Lozano
Computer Science School, University of Costa Rica, Montes de Oca, COSTA RICA; arturo.camacho@ecci.ucr.ac.cr

Juan Fonseca-Solís
University of Costa Rica, Research Center in Information and Communication Technologies, San Pedro de Montes de Oca, San José, 115012060, COSTA RICA; juan.fonsecasolis@ucr.ac.cr

Accessible pedestrian signals (APS) enhance accessibility in streets around the world. Recent attempts to extend the use of APS to people with visual and audible impairments have emerged from the area of audio signal processing. Even though few authors have studied the detection of APS by sound, comprehensive literature in Biology has been published to detect other simple sounds like birds and frogs calls. Since these calls exhibit the same periodic and modulated nature as APS, many of these approaches can be adapted for this purpose. We present an algorithm that follows this approach. The algorithm was evaluated using a collection of 79 recordings gathered from streets in San Jose, Costa Rica, where an APS system will be implemented. Three types of sounds were available: low-pitch chirps, high-pitch chirps and cuckoo-like sounds. The results showed 91% precision, 80% recall, 83% F-measure, and 90% specificity.

Published by the Acoustical Society of America

 https://doi.org/10.1121/2.0000675

1. INTRODUCTION

Pedestrians often use multiple cues to make decisions concerning crossing streets. Young healthy people can easily take advantage of these cues, and develop a high degree of awareness of the traffic flow. However, people with vision or hearing disabilities (PWD) have much more difficulties using these cues, making them dependent on fewer sources of information and reducing their ability to tolerate noise. A special type of pedestrian unit, called *accessible pedestrian signal* (APS), has been designed since the 1970s to assist PWD in streets. Besides providing visible light signals, APSs produce highly audible sequence sounds: cuckoo like (for east-west crossings) and chirp-like (for north-south crossings). APSs vary depending on country, or even city, but in general the same sounds are used in Costa Rica and other countries, like the United States.[2] In these countries, there has been an increased interest in making mobile devices recognize the presence of APSs, and produce cues using voice commands and vibrations, enhancing PWD mobility. The voice commands include instructions, like rotations and shifts, that users follow to get aligned with respect to crosswalks. Some authors have accomplished this goal by detecting the location and orientation of zebra crossings using cameras and accelerometers (a computer vision approach).[1,6] Ivanchenko *et al.* reported 95% specificity and 72% recall for their algorithm called Crosswalk, and Ahmetovic *et al.* presented 100% precision and 77% recall for their solution called Zebralocalizer. However, to the best of our knowledge, there have been no efforts in using sound as the main source of information. To determine if APS can be detected automatically by acoustic methods, we proposed a technique based on the pitch estimation algorithm SWIPE-prime.[4] This algorithm uses raised-cosine shaped kernels in the frequency domain to accomplish the task, and has been further adapted to detect the calls of frogs, fish, and manatees in noisy environments.[3,5,8] Ideas from other authors were also used, particularly, the unsupervised similarity measurements employed to detect the vocalizations of a Hawaiian whale named *Balaenoptera acutorostrata* and the conceptualization of a feature vector space to detect Australian frogs.[7,9] This document is organized as follows: section 2 describes the frequency-temporal characteristics of the available APS, section 3 describes the algorithms designed to detect those sounds, section 4 discusses the proposed evaluation method and results, and section 5 presents the conclusions.

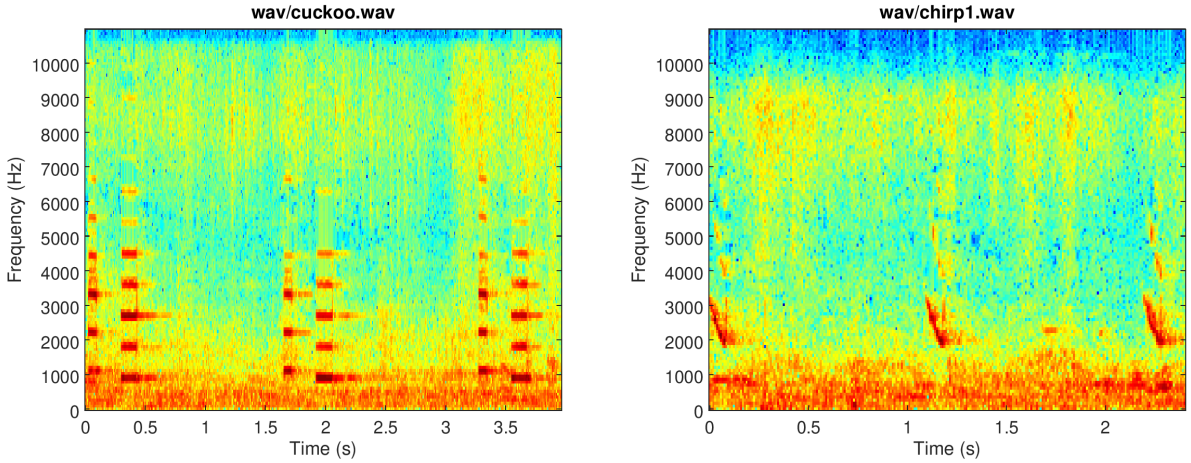
2. CHARACTERISTICS OF THE APS SOUNDS

Among the studied APS sounds, one is similar to the cuckoo bird call, and is formed by two harmonic tones with a silence in between. The other two signals correspond to chirps, which is a frequency modulation varying between two different frequency ranges, described more in detail in the next sections.[1] We determined the characteristics of the sounds using the clearest recordings, made on the same side of the sidewalk. Recordings made on the opposite sidewalk were ignored, since the signals were considerably modified due to propagation and noise, which impacted mainly the high frequency sounds.

A. CUCKOO-LIKE SOUNDS

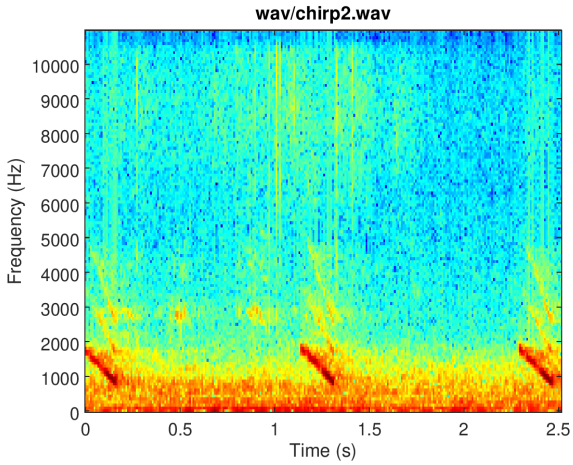
Figure 1a shows a cuckoo-like sound that consists of a 70 ms harmonic tone at 1100 Hz, followed by a 200 ms silence and a 130 ms harmonic tone at 900 Hz. The first tone has energy in the first six harmonics, while the second shows energy up to the seventh harmonic. However, in recordings made on the opposite sidewalk, high harmonics are much more blurry, and only the first three are clear for the 1100 Hz tone, and four for the 900 Hz tone. Opposite sidewalk recordings also have an echo, increasing the apparent duration of the tones. The time between event onsets is 1630 ms.

[1]The full description for this sounds can be found in the the *Uniform Traffic Control Devices* (MUTCD) standard: `http://mutcd.fhwa.dot.gov/kno-overview.htm` and the *APS Guide Overview*: `http://www.apsguide.org`



(a) Cuckoo sound of 3.9 *s with two tones of 1100 Hz and 900 Hz.*

(b) Chirp-1 sound of 2.4 s with a frequency modulation of 3 kHz to 2 kHz.



(c) Chirp-2 sound of 2.5 s with a frequency modulation of 1750 Hz to 950 Hz.

Figure 1: Spectrograms of the three studied sounds calculated using a Hamming window of 256 samples, a sampling rate of 22.050 ***kHz, and an overlapping factor of 50%.***

B. CHIRP SOUNDS

Figure 1b shows the first chirp sound (*chirp-1*), corresponding to a 100 ms frequency sweep from 3 kHz to 2 kHz. The event onsets are spaced at 1120 ms. Figure 1c shows the second type of chirp (*chirp-2*), a 160 ms sweep from 1750 Hz to 950 Hz. Most of the energy is confined to the first three odd harmonics (1, 3, and 5), although, in some cases some energy is present in harmonics 2 and 4. In addition, the energy decays as a function of harmonic number. For opposite sidewalk recordings, the third and fifth harmonics are practically absent, and the time between the event onsets is 1110 ms.

C. FINALIZATION PATTERN

Unlike cuckoo-like sounds, chirp sounds present a distinct finalization pattern to indicate to pedestrians that they should refrain from crossing. These patterns show an increased event frequency, shorter event duration, and a reduced frequency modulation range. For the first chirp, the events correspond to 30 ms

frequency modulations of a tone with frequencies ranging from 3 kHz to 2.6 kHz. In recordings made on the same sidewalk, a second harmonic with reduced energy is usually present, while on the opposite sidewalk, such harmonics are often missing from the recording. The event onsets are separated by 630 ms. For the second chirp, the events are either 60 ms, 90 ms or 140 ms long, and correspond to modulated tones with frequencies sweeping from 1600–1350 Hz, 1750–1400 Hz, and 1750–1050 Hz ranges, respectively. The first of these three sweeps usually presents energy in the first four harmonics, while the latter two have energy only in the first, third, and fifth harmonics. In this case, the events are unevenly and inconsistently spaced.

3. AUTOMATIC DETECTION SYSTEM

In this work, we propose an algorithm to automatically detect crosswalk sound signals, formed either by chirps or sets of two constant tones. For each window, the set of harmonics associated with the fundamental frequency are detected. For that purpose, the system produces a score for each of the three types of signals, and if one of them reaches a threshold, a detection is reported. To reduce computational cost, the signals are resampled to 22.05 kHz, since the relevant information is below 11 kHz (the Nyquist frequency for the fourth harmonic of the chirp-1 sound). A spectrogram of the signals is computed using a 9 ms window (corresponding to 8 periods of the lowest fundamental frequency at 900 Hz) and 50% overlapping.[4] The window size was rounded to the next power of two, to maximize the computation speed of the *fast Fourier transform* (FFT).

A. CUCKOO-LIKE SOUNDS

For each window, harmonic tones with $f_0 = 900$ Hz or $f_0 = 1100$ Hz are detected computing a normalized product of the magnitude spectrum $A(t,f)$ and a harmonic kernel K_h, as follows:[2]

$$H(f_0,t) = \frac{\int \sqrt{A(t,f)} K_h(f_0,f) df}{\sqrt{\int A(t,f) df} \sqrt{\int [K_h^+(f_0,f)]^2 df}}, \tag{1}$$

with

$$K_h(f_0,f) = \begin{cases} \sin[2\pi(f/f_0 - 0.75)], & 0.75 < f/f_0 < 0.25 + N \\ \frac{1}{2}\sin[2\pi(f/f_0 - 0.75)], & (0.25 < f/f_0 < 0.75) \vee (0.25 + N < f/f_0 < 0.75 + N) \\ 0, & \text{otherwise,} \end{cases}$$

where $K_h^+(f_0,f) = \max\{0, K_h(f_0,f)\}$, and N is the number of harmonics ($N = 4$ for 900 Hz tones and $N = 3$ for 1100 Hz tones). Figure 2 shows an example of the kernel. This kernel is not a signal spectrum, but a mechanism for rewarding or penalizing energy distributions. The fundamental frequency with highest score is marked as the target, and is taken as the window pitch in case the prediction reaches a threshold of 0.14. If the predictions do not reach the threshold, then a zero value is assigned to that window. A two tone pattern is then estimated by the proportion of frequency predictions consistent with 70 ms of a 1100 Hz harmonic tone, and 140 ms of a 900 Hz harmonic tone, 200 ms apart. The score is $S_h(t) = T/w_t$, where T is the number of windows which tones were predicted correctly and, w_t is the total number of windows. The silent parts are not considered, since some recordings have echoes, potentially producing wrong scores. A cuckoo-type sound (the two tones, played in series, with a silence gap in between) will be identified when S_h exceeds a threshold of 0.45.

[2]By *harmonic tones* or *complex tones* we refer to those composed by the fundamental frequency and at least one harmonic, in contrast with *pure tones* which are formed just by the fundamental frequency.

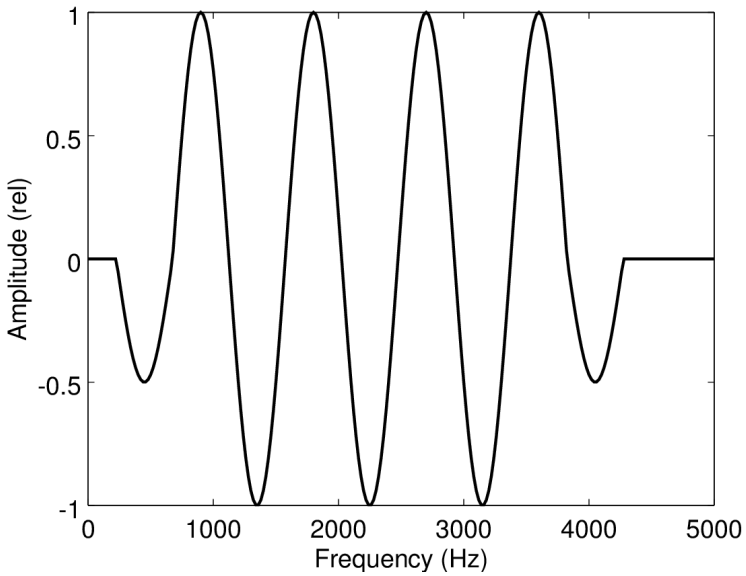


Figure 2: Kernel for a harmonic tone with 900 Hz fundamental frequency.

B. MODULATION PATTERN DETECTION

This section describes how to detect the modulation patterns. First, the presence of *chirp-1* and *chirp-2* tones is estimated for each time window. This stage produces a best matching frequency for each window. Then, the presence of a modulating pattern is determined using information about chirp duration. Subsequently, detections are made if the time interval between modulating events match the ones reported in section 2.2.

i. Tone estimation in windows

For each time window in the spectrogram, a similarity index is computed as the normalized product of the magnitude spectrum and a cosine kernel $K_1(f_0, f)$ with a frequency f_0 ranging from 2 kHz to 3 kHz, with 50 Hz steps. The normalized product is the one in section 3.1, differing only in the kernel. Figure 3 shows the kernel $K_1(f_0, f)$ described in the following equation:

$$K_1(f_0,f) = \begin{cases} \sin\left[\frac{\pi}{3}(10f/f_0 - 8.5)\right], & 0 < |f/f_0 - 1| < 0.15 \\ \frac{1}{2}\sin\left[\frac{\pi}{3}(10f/f_0 - 8.5)\right], & 0.15 < |f/f_0 - 1| < 0.45 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The best sine lobe width was empirically determined and corresponds to $0.3\,f_0$. The negative lobes are used to produce low scores for white noise, having a uniform spectrum.[3] The frequency yielding the maximal score is determined and assigned to the temporal window if the corresponding similarity index reaches an empirical threshold of 0.07. A value of zero is otherwise assigned. Similarly, a normalized product of the magnitude spectrum and an odd-harmonics kernel is computed for every time window, with fundamental frequencies f_0 in the range 950-1750 Hz with 50 Hz spacing. Again, the normalized product is the same as in section 3.1, but with a different kernel. Figure 4 shows kernel $K_{\text{odd}}(f_0, f)$, that was implemented as follows:

$$K_{\text{odd}}(f_0,f) = \sum_{i=1}^{N} K_2((2i-1)f_0, f)0.8^{f/f_0}, \tag{3}$$

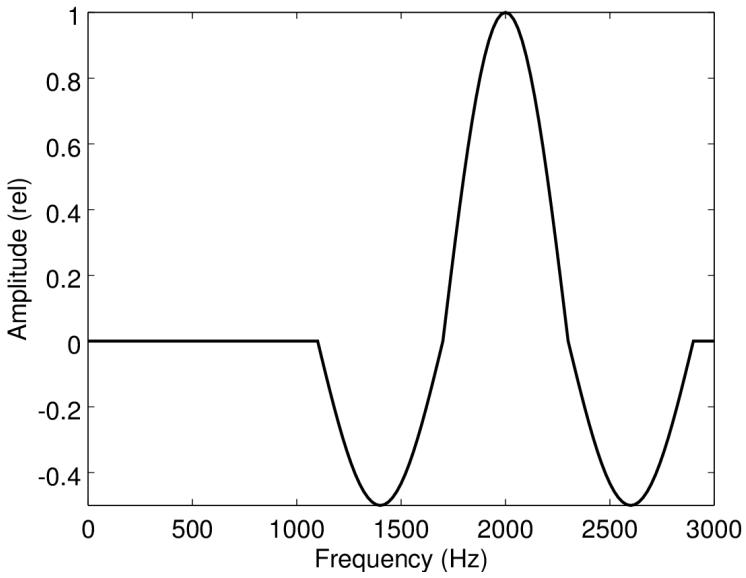


Figure 3: Sine kernel for a 2 kHz tone.

which describes the sum of N pure-tone kernels and a decaying factor that varies according to the frequency. The pure tones are described as follows:

$$K_2(f_0,f) = \begin{cases} \sin\left[2\pi(f/f_0 - 0.75)\right], & 0 < |f/f_0 - 1| < 0.25 \\ \frac{1}{2}\sin\left[2\pi(f/f_0 - 0.75)\right], & 0.25 < |f/f_0 - 1| < 0.75 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The decay is used to decrease the weight of the high order partials. The frequency with the highest score is determined and assigned to the time window in case it reaches a threshold of 0.07; otherwise, a value of zero is assigned.

ii. Modulation pattern estimation

To estimate the modulation patterns, we computed the difference between the predicted frequencies $P_f(t)$ and a kernel $M(t)$. The kernel describes the typical modulation pattern, and depends on the duration and frequency range of the events (see section 2.2). Figure 5 shows an example of the kernel. We used the Euclidean distance of the kernel and the non-zero frequencies $P_f^{'}(t)$, and then introduced a penalty for the null frequencies corresponding to failures of the previous stage of tone detection. The Euclidean distance d is inverted to produce scores in the unitary range as follows:

$$d(t) = 1 - \sqrt{\frac{\sum_{i=1}^{n-k}\left(P_f'(t+i) - M(i)\right)^2}{n\left(M(n) - M(1)\right)^2}} - p(k,n), \tag{5}$$

where n denotes the number of samples in the kernel M, k is the number of windows with zero values, t is the starting time of the predicted frequencies (which slides in time in order to detect modulations for the entire recording). The recursive penalty function $p(k,n)$ is specified like so:

$$p(k,n) = \begin{cases} 0, & \text{if } k = 0 \\ \left(p(k-1,n) + 0.02(k+2)\right) p(n,n)^{-1}, & \text{otherwise.} \end{cases} \tag{6}$$

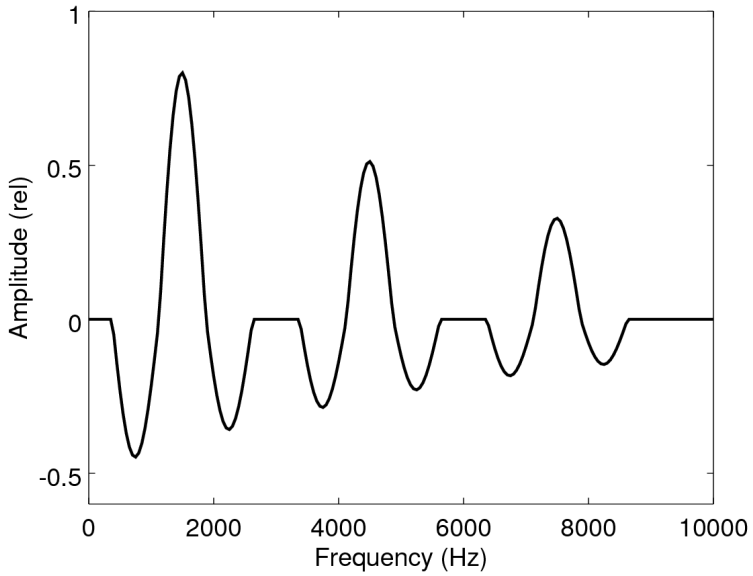


Figure 4: Odd harmonics kernel for a tone with 1500 Hz fundamental frequency.

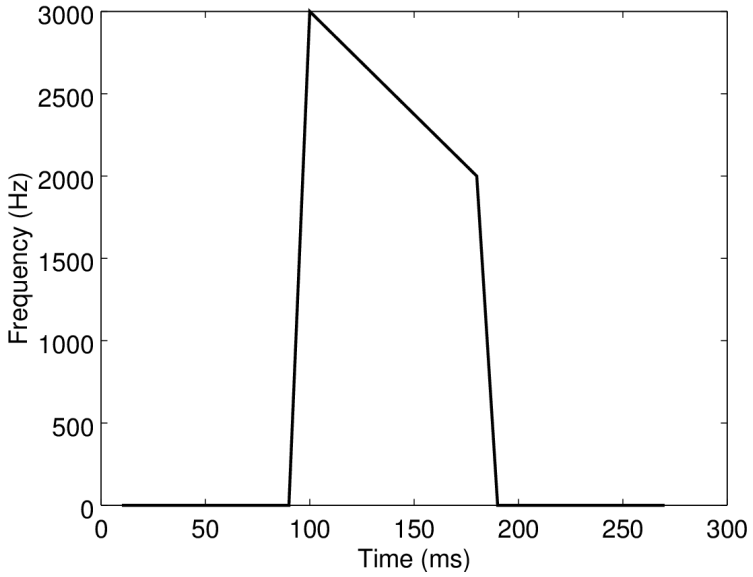
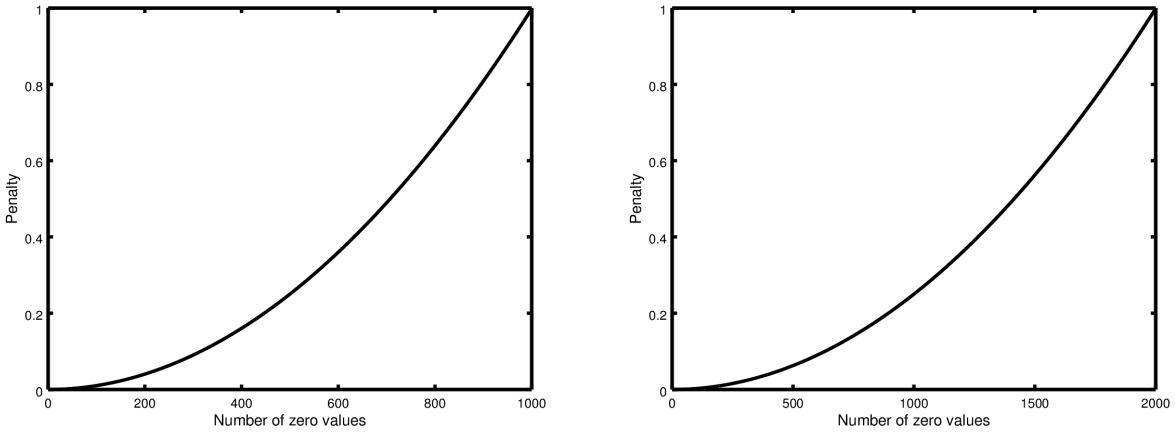


Figure 5: Kernel used for modulation detection for frequencies between 3 kHz and 2 kHz.



(a) Penalization score for $n = 1000$***.*** ***(b) Penalization score for*** $n = 2000$***.***

Figure 6: Penalization scores for two sizes of kernel.

Figure 6 shows that this function penalizes predictions containing mostly zeros. The penalty function was empirically tuned, such that some omissions in the tone estimation stage are allowed, but as more samples are omitted the penalties grow faster. As a consequence, it is possible that some tones are not detected because of noise conditions, but will still be detected at the present step. A few omissions usually produce low penalties while a considerable number of omissions will produce high penalties.

iii. Intervals between events

To determine if the spacing between events is correct, a product of the frequency modulation scores with different lags is computed as follows:

$$P_s(t) = \prod_{i=0}^{q} d(t + il + \varepsilon),$$

where q is the number of events considered, l is the lag duration, and ε is an error that corresponds to the range $|\varepsilon| \leq 0.025l$. If q successive events have high frequency modulation pattern scores, the product approaches unity; otherwise, the product approaches zero. This product is computed for *chirp-1* with $q = 3$ and for *chirp-2* with $q = 2$, with the corresponding spacings described in section 2.2. The products exceeding a threshold of 0.3 are considered as detections.

C. FINALIZATION PATTERN FILTERING

Events in the finalization sequence of the chirp-like sounds provided shorter frequency modulations with lower amplitude, making them more susceptible to noise pollution. To separate the main sequence from the finalization pattern, we developed a filtering algorithm that verifies that the distance between consecutive peaks and the amplitude between consecutive peaks remain constant. This is called the *restrictive approach* and works as follows. First, the algorithm iterates over the peaks of the alert signal and recursively calculates the mean and variance of the distance between peaks and the distance between amplitudes. For example, the recursive mean, m, and variance, s^2, for amplitude between peaks is defined like so:

$$m[i] = \frac{m[i-1](i-1) + a[i]}{i}$$

and

$$s^2[i] = \frac{(a[i] - m[i])^2}{i-1} + \frac{i-1}{i} s^2[i-1],$$

where $a[i]$ is the amplitude of the i-th peak. Within each iteration, the peaks that exceed 4σ or more (either in time separation or amplitude) are removed. The variance is initialized using one quarter of the initial mean. Knowing that APS sounds in the main sequence do not vary over time, the mean and variance are not adjusted if the first four consecutive peaks preserve the same distribution. Figures 8, 9, and 10 show how the restrictive approach eliminates false positives providing a cleaner alert signal. However Fig. 10 also shows that true positive peaks could be erroneously eliminated, which is a disadvantage.

4. EVALUATION

A. COLLECTED DATA

A total of 79 recordings were collected from 11 different Novax DS100 units in the Costa Rican great metropolitan area (GAM). *Cuckoo* and *chirp-1* were the most common sounds, with 36 and 30 recordings respectively, while *chirp-2* was the less common sound with 13 recordings. The recordings were acquired using three smartphones of medium-gamma (Samsung Galaxy Ace S5830, Samsung Galaxy S Duos S7562, LG G2 mini D618), employing a sampling rate of 44.1 kHz, and a quantization of 32 bits. Audio was recorded either during the morning, when less traffic was present in streets, as well as after midday when roads were more populated. The weather conditions were mostly sunny and cloudly, the last one with the presence of light rain. The duration of the recordings ranged from 13 s to 35 s, depending of the activity period of the Novax units, and the signal-to-noise ratio (SNR) corresponded to 1.1 dB (Novax units were designed to maintain an SPL level of +5dB above the ambient noise, with a limit on 90 dB).[3] As mentioned in section 3, the recordings were downsampled to 22.05 kHz, and requantized to 16 bits. A lowpass filter with $f_c = f_s/2$ was applied before downsampling, to avoid aliasing.

B. EVALUATION METHOD

To evaluate the method, the numbers of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), were calculated using the manual and estimated annotations corresponding to the onset of each event. The manual annotations were obtained by looking at the spectrograms and establishing truth onsets, as a basis for comparison with the automatically annotated ones. Figure 7 shows that to estimate the metrics we defined two sets of window's indexes, called A and B, that contain the indexes of the activity sequence bounded by the first and the last annotated onset. Set A contains the indexes of the activity sequence determined by the human expert, and set B contains the indexes of the activity sequence determined by the algorithm. These sets are defined as

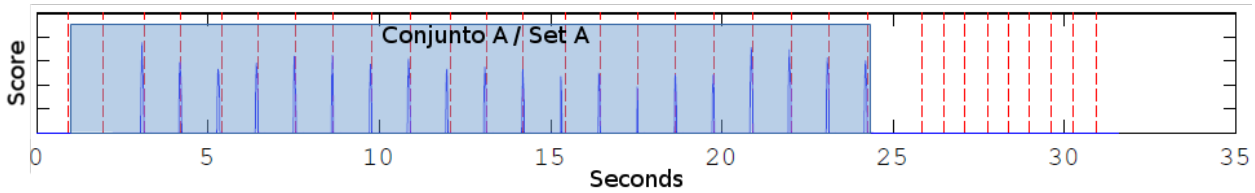
$$A = \{i \in \mathbb{N} / t_{A_1} \leq i\gamma \leq t_{A_N}\}$$

and

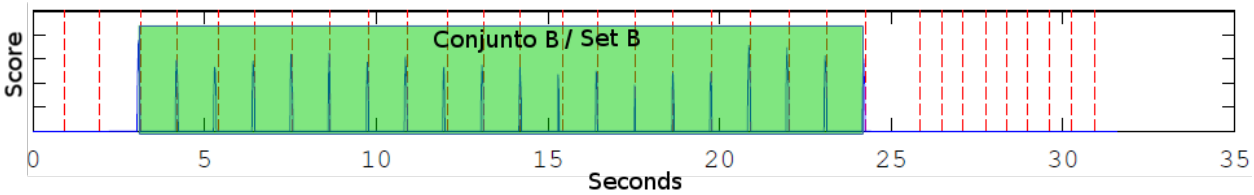
$$B = \{j \in \mathbb{N} / t_{B_1} \leq j\gamma \leq t_{B_M}\},$$

where $\{t_{A_1}, t_{A_N}\}$ are the times of the first and last manually annotated onsets (in seconds), $\{t_{B_1}, t_{B_M}\}$ are the times of the first and last automatically annotated onsets, and γ is the period of time captured by each window in the *short time Fourier transform* (STFT). The number of windows of the activity sequence that both sets have in common represents the true positives, in other words, $\text{TP} = \text{card}(A \cap B)$, where card stands for *cardinality*. The number of windows where an alert was emitted without the presence of an APS sound

[3]Taken from Novax specifications: `http://www.novax.com/ds100series-aps-specifications`



(a) Manually-annotated set (truth-basis).



(b) Automatically detected set.

Figure 7: Manually obtained and automatically detected sets. The red-dashed vertical lines corresponds to the manually-annotated onsets, and the blue-solid peaks of the signal (shaded by the sets) correspond to the onsets detected by our algorithm. In this example, the last nine onsets were ignored by the system as they belong to the finalization sequence.

are the false positives, that is, $\mathrm{FP} = \mathrm{card}(B - A)$. The number of windows where an alert should be emitted, but was not, are the false negatives, in other words, $\mathrm{FN} = \mathrm{card}(A - B)$. The number of windows where an alert was not emitted when an APS was not activated are the true negatives, that is, $\mathrm{TN} = \mathrm{card}(B^c \cap A^c)$. Due to simplifications, we did not penalize the missing onsets inside the activity sequences. However, this is unlikely to happen, according to what we observed in the non-filtered estimations produced by the model.

i. Detection rates

From the measurements of TP, FP, TN and FN, we calculated the metrics of precision, recall, F-measure, and specificity, defined below. These metrics were introduced for the sake of comparison with the results mentioned in section 1, and because they provide a convenient way of visualizing performance in the unitary range. Briefly, precision and specificity show how well the method avoids false positives, recall shows how well is done the detection of APS, and F-measure combines precision and recall. They are defined more precisely as follows:[4]

Precision (p): How often an emitted alert is real, in other words, the percentage of times users would not be at risk of suffering an accident when obeying an alert. Also known as *predictive value for a positive result*. Intended to be very high, as it is an indicator of safety. Calculated as $p = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$.

Recall (r): How often will an alert be emitted if the APS was activated (probability of detection), equivalently, how useful is the system to identify moments when crosswalks can be used. Also known as *sensitivity* or *exhaustivity*. Intended to be high, but not critical. Calculated as $r = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$.

F-Measure (F): Harmonic mean between the safety and usefulness of the system, in other words, how close are measurements to their true type: positive or negative. Substitute for *accuracy* when data is asymmetric (i.e. $(\mathrm{TP} + \mathrm{FP})/(\mathrm{TN} + \mathrm{FN}) \ll 1$), which is the case for APS, as the main sequence represents the positives and it spans almost the entire recording. Calculated as $f = 2pr/(p + r)$.

[4]Calculated according Microsoft's Azure's performance formulas, Juhola's lessons at Tampereen Yliopisto (`http://www.uta.fi/sis/tie/tl/index/Rates.pdf`) and the Emergency Medicine-Ambulatory Care (EMAC) guidelines in Emory University (`https://www.med.emory.edu/EMACancurriculum/diagnosis/sensand.htm`).

Table 1: Performance of the original and restrictive proposed approaches, and the Ivanchenko and Ahmetovic ones.

	Performance (%)			
Metric	Original	Restrictive	Ivanchenko	Ahmetovic
Precision	91	95	–	100
Recall	80	72	72	77
F-Measure	83	77	–	–
Specificity	90	96	95	–

Specificity (s): How often will the system prevents itself from sending a false alert when the APS is not activated (robustness to noise). Intended to be very high; otherwise, the system would not be able to differentiate environmental noise from APS. Calculated as $s = \mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$.

C. RESULTS AND ANALYSIS

Table 1 shows the results obtained using the original and restrictive approaches. The original approach showed a balanced configuration with a precision and specificity 10% higher than recall and F-Measure, while the restrictive approach increased the difference even more, to 21% (in average). In both cases this is interpreted like so: the chance that a pedestrian might cross when cars are in movement, is less likely than not crossing when the APS is in green. This behavior is desired to accomplish the *primum non nocere* principle (first, do not harm). In comparison with the results reported by Ivanchenko *et al.* our method showed -5% specificity and $+8\%$ recall in the original version, and $+1\%$ specificity and the same recall in the restrictive approach. In comparison with the results reported by Ahmetovic *et al.* our method showed -9% precision and $+2\%$ recall in the original version, and -5% precision and -5% recall in the restrictive approach. Doing a naive judgment by omitting the differences between video and audio processing, and supposing that all crosswalks have an APS installed, we could say that the restrictive version of our algorithm shows higher metrics than Ivanchenko's Crosswalk, but is not better than Zebralocalizer.

5. CONCLUSIONS

We developed a method to automatically detect three kind of APS in Costa Rica. We did so by recognizing the pitch of individual sounds using musical kernels. One kernel with all the harmonics (*cuckoo-like* sounds), and the other with only odd harmonics (*chirp-like* sounds). The main sequence patterns were also detected differently in both cases. For *cuckoo* sequences, we clustered the tonal windows together, penalized missing occurrences, and measured silence intervals between events. In contrast, for *chirps*, we created frequency-contour templates and used the Euclidean distance to find correspondences. The manual-annotated recordings took in the field allowed us to calculate the performance metrics using an *adhoc* methodology. The positive and negative predictions calculated using a set approach let us to find promising results: 91% precision, 80% recall, 83% F-measure, and 90% specificity. An additional temporal-amplitude filtering stage was proposed to restrict metrics in terms of precision and specificity, with the inconvenience of diminishing recall and F-measure. This restrictive approached showed 95% precision, 72% recall, 77% F-measure, and 96% specificity, which are better metrics than obtained by Crosswalk, a video based solution. We identified that more work is needed to improve the recall, and therefore increase the usefulness of the system. The method might be implemented as a mobile application, however some (if not most) of those

technologies are not capable of providing real-time processing with the existent general purpose operative systems.

6. ACKNOWLEDGMENTS

We thank Mario Monge and Sharon Bejarano for proposing the idea of studying APS sounds in the Costa Rican streets. They provided invaluable insights and shared many of the recordings used here. We also thank Paul Hursky for his thoughtful revision of this paper.

REFERENCES

[1] Ahmetovic, D., Bernareggi, C., and Mascetti, S. "Zebralocalizer: identification and localization of pedestrian crossings". In: Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services. ACM. 2011, pp. 275–284. DOI: 10.1145/2037373.2037415

[2] Ashmead, D. H. et al. "Which crosswalk? Effects of accessible pedestrian signal characteristics". In: Institute of Transportation Engineers. ITE Journal 74.9 (2004), p. 26.

[3] Camacho, A., García-Rodriguez, A., and Bolaños, F. "Automatic detection of vocalizations of the frog Diasporus hylaeformis in audio recordings". In: The Journal of the Acoustical Society of America 130.4 (2011), p. 2500. ISSN : 00014966. DOI: 10.1121/1.3654948.

[4] Camacho, A. and Harris, J. G. "A sawtooth waveform inspired pitch estimator for speech and music". In: The Journal of the Acoustical Society of America 124.3 (2008). DOI: 10.1121/1.2951592

[5] Castro, J. M., Rivera-Chavarría, M., and Camacho, A. "Automatic manatee count using passive acoustics". In: The Journal of the Acoustical Society of America 137.2220 (2015). DOI: 10.1121/1.4920090.

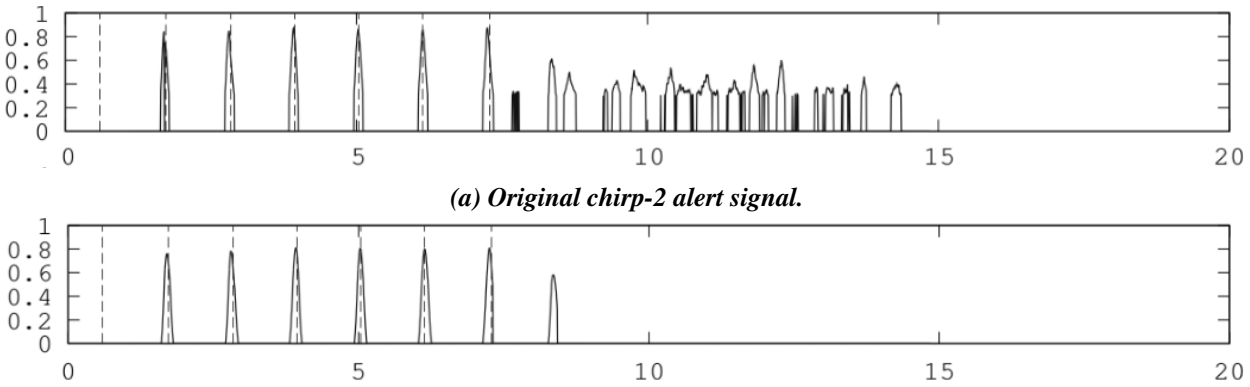
[6] Ivanchenko, V., Coughlan, J., and Shen, H. "Crosswatch: a camera phone system for orienting visually impaired pedestrians at traffic intersections". In: International Conference on Computers for Handicapped Persons. Springer. 2008, pp. 1122–1128. DOI: 10.1007/978-3-540-70540-6_168

[7] Mellinger, D. K. et al. "A method for detecting whistles, moans, and other frequency contour sounds". In: The Journal of the Acoustical Society of America 129.6 (2011). DOI: 10.1121/1.3531926

[8] Ruiz-Blais, S., Rivera-Chavarría, M., and Camacho, A. "Autonomous detection of neotropical sciaenid fishes". In: The Journal of the Acoustical Society of America 18.010001 (2012). DOI: 10.1121/1.4792734.

[9] Taylor, A. et al. "Monitoring Frog Communities: An Application of Machine Learning". In: Proceedings of the Eighth Annual Conference on Innovative Applications of Artificial Intelligence. IAAI96. AAAI Press, 1996, pp. 1564–1569. ISBN: 978-0-262-51091-2.

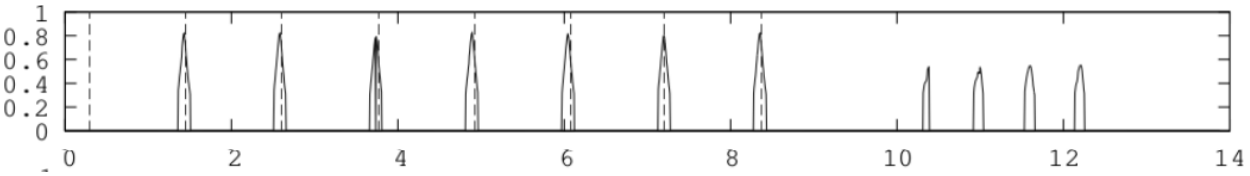
A. EXAMPLES OF THE TIME-AMPLITUDE POST-PROCESSING



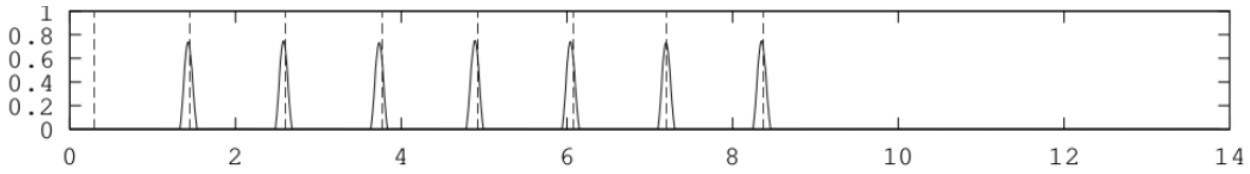
(a) Original chirp-2 alert signal.

(b) Restrictive chirp-2 alert signal.

Figure 8: Original and filtered chirp-2 alert signal. The horizontal axis represents the time in seconds and the vertical axis corresponds to the amplitude of the alert. The filtering stage improved noticeably the output signal in the interval $8 - 15$ ***s.***

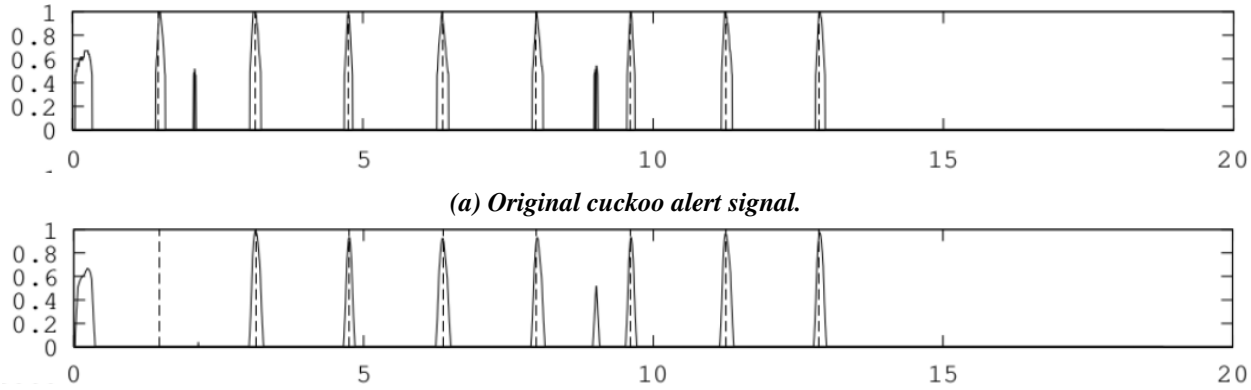


(a) Original chirp-2 alert signal.



(b) Restrictive chirp-2 alert signal.

Figure 9: Original and filtered chirp-2 alert signal. The filtering stage also improved noticeably the output signal eliminating the last four peaks corresponding to the finalization sequence.



(a) Original cuckoo alert signal.

(b) Restrictive cuckoo alert signal.

Figure 10: Original and filtered cuckoo alert signal. In this case the filtering stage produced undesired results eliminating the 2nd peak, and leaving untouched the 8th one.