INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Using Approximated Auditory Roughness as a Pre-filtering Feature for Human Screaming and Affective Speech AED

Di He[1], *Zuofu Cheng*[2], *Mark Hasegawa-Johnson*[3], *Deming Chen*[1]

[1]Coordinated Science Lab, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA 61801
[2]Inspirit IoT, Inc, Champaign, Illinois, USA 61822
[3]Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA 61801

dihe2@illinois.edu, zuofu.cheng@gmail.com, jhasegaw@illinois.edu, dchen@illinois.edu

Abstract

Detecting human screaming, shouting, and other verbal manifestations of fear and anger are of great interest to security Audio Event Detection (AED) systems. The Internet of Things (IoT) approach allows wide-covering, powerful AED systems to be distributed across the Internet. But a good *feature* to *pre-filter* the audio is critical to these systems. This work evaluates the potential of detecting screaming and affective speech using Auditory Roughness and proposes a very light-weight approximation method. Our approximation uses a similar amount of Multiple Add Accumulate (MAA) compared to short-term energy (STE), and at least 10× less MAA than MFCC. We evaluated the performance of our approximated roughness on the Mandarin Affective Speech corpus and a subset of the Youtube AudioSet for screaming against other low-complexity features. We show that our approximated roughness returns higher accuracy.

Index Terms: Audio Event Detection, pre-filtering, Auditory Roughness, computational complexity

1. Introduction

Internet of Things (IoT) has provided many applications a new approach. When powerful and reliable computational capability meets distributed wireless sensor networks, many tasks that used to suffer from cost, practicality and low-accuracy benefit [1]. Audio Event Detection (AED) based security or surveillance systems are one of these applications.

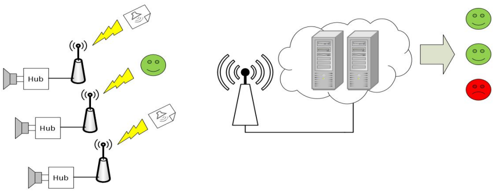

Figure 1: *Distributing AED across the network.*

AED for security purposes has been an interesting topic to both the research community [2, 3, 4, 5] and to commercial entities [6, 7]. Security events such as gunshots and explosions are relatively easy to detect, and commercial products based on AED have already been deployed in many US cities [1]. Security related human speech events, such as screaming, shouting, and other manifestations of fear and anger have proven to be more difficult to detect accurately; many works train dedicated models to detect these speech events [3, 4]. State-of-the-art methods include Deep Neural Networks (DNN) and Hidden Markov Models (HMM) [8] or Long-short Term Memory Recurrent Neural Networks (LSTM RNN) [9]. The computational complexity of these methods is high. The IoT approach, presented in Fig 1, does offer AED access to powerful computation capability. However, having enough power dedicated to each and every sensor all the time is inefficient and unpractical for large sensor networks. AED systems targeting human speech therefore use *pre-filtering* mechanisms: algorithms with low computational complexity that can detect events of interest with a low missed detection rate, and with a false alarm rate that may be relatively high, but that is low enough to limit the computation expended by the second-pass classifier [2, 4, 10]. Pre-filtering can reduce communication and computation costs by discarding audios when an event is likely absent.

Previously published pre-filtering algorithms fail to meet the three simultaneous requirements of high recall, acceptable precision, and extremely low complexity. The problem usually is the intrinsic limitation of the *features* used for *pre-filtering*. Some of these works [2, 10] rely on windowed Short-term Energy (STE). STE, although light in computation, as we will show later, fails to differentiate affective speech and neutral speech effectively. Other work [4] uses spectral features, which are much complex, on the order of over 10×, to extract than STE [11].

This paper considers the potential of a classical acoustic feature called the Auditory Roughness [12, 13, 14], as a prefilter feature. Auditory Roughness is a classic measure of ”harsh and unpleasant” sound with a long history. Although it used to be an acoustic concept, recent biological studies have found proof linking the concept with what fear is triggered in the mind by perceiving human screaming [14]. First, the standard Auditory Roughness feature is used to detect anger and fear in the Mandarin Affective Speech corpus [15], and screaming in the Youtube AudioSet [16], with recall and precision better than STE. Second, since the standard Roughness computation has complexity similar to that of spectral features, an approximate Roughness measure is proposed, with computational complexity similar to STE and at least 10× lighter than MFCC. The approximated Auditory Roughness feature is demonstrated to have recall and precision better than other pre-filtering features with similar computational complexity, including STE.

We will briefly introduce the concept of distributed AED and Auditory Roughness in part 2. We will explain how and why we approximate roughness in part 3. Experiment setup and results are then presented in part 4. We conclude and discuss the work at the end.

[1] https://www.nytimes.com/2015/03/17/nyregion/shotspotter-detection-system-pinpoints-gunshot-locations-and-sends-data-to-the-police.html?_r=0

http://dx.doi.org/10.21437/Interspeech.2017-593

2. Background

2.1. Distributed AED

A distributed surveillance system designed under the framework of [1] would look like Fig 1. A large number of sensing nodes with wireless capability are deployed. Different kinds of sensors are grouped into each node and correlated by a hub controller. The controller must run non-stop to serve the sensors conducting minimum surveillance; it must also power-on and control more powerful sensors when needed. If pre-filter operations find signs of an event happening, features collected by the sensors will be uploaded to the cloud for further analysis.

In an AED scenario, if pre-filtering features don't exceed normal level, the audio is only buffered for a couple of seconds. Hubs will inform the cloud that everything is fine. If the pre-filtering feature exceeds a preset threshold, a request is initialized: the buffered audio is packed and uploaded for further analysis. Depending on the connection, further audio is either packed into chunks and uploaded, or streamed to the cloud as the event unfolds. On the cloud side, more powerful servers can be accessed on demand. If a request comes in, the servers will run much more sophisticated algorithms such as DNN+HMM or LSTM to conduct more complicated analysis. To build and run a system with a large number of sensors, both the sensors and the controllers must be cheap and low-power. Hardware targeting sensor hubs have limited resources, e.g., they typically have no dedicated multiplier.

2.2. Auditory Roughness

The term "Auditory Roughness" originated as a musical expression in the 19th century [12]. The term is now defined to be a psychophysical dimension, describing the human perception of harsh, raspy hoarse sounds [13]. The musical and perceptual concept was formalized as a sound quality measurement, with several standard definitions and published algorithms [2]. Recent studies have identified apparent neurobiological correlates of perceived Auditory Roughness [17].

Amplitude modulation frequency is one of the most important physical acoustic correlates of Auditory Roughness [18, 14]. In music and other non-vocal sounds, a modulation frequency of 30Hz or below is usually perceived as beats [18]. When modulation frequency exceeds 30Hz, it is considered rapid and the sensation of roughness appears. Though speech is complex, the same modulation frequency thresholds seem to apply: neutral speech has most of its modulation spectral energy at 1–10Hz [19], and modulation frequencies above 30Hz will trigger the brain's fear center [17]. The sense of roughness peaks at modulation frequencies of 70Hz [14] or 75Hz [18], but persists in response to modulations of up to 150Hz [18] or 300Hz [14].

Further studies [18] claim carrier frequency and the strength of amplitude modulation also affect the level of roughness. Some Auditory Roughness calculators [20, 18] consider each spectral peak as a carrier frequency, compute the modulation frequency and modulation strength of each carrier, and add the roughness effects from each carrier frequency together. Other algorithms compute Auditory Roughness by analyzing the distribution of energy within pre-determined frequency bands, e.g., in 24 bands uniformly distributed on a Bark scale [2] [21, 14].

3. Method

In our attempt to detect human screaming, fear, and anger, Auditory Roughness seems to be a good candidate feature. However, both of the standard methods used to compute Auditory Roughness are far too complicated for pre-filtering. We therefore propose a computationally much, much simpler approach to approximate roughness. Psycho-acoustic studies [17] agree that rapid and strong amplitude modulation, at around 30-150Hz, is the most important physical correlate of perceived roughness. Existing algorithms are computationally expensive because they assume that the speech signal contains many different carrier frequencies, and perform coherent demodulation and/or analysis of each. In this paper, we assume that there is only one instantaneous carrier frequency, whose modulation spectrum can be derived using fast non-coherent demodulation.

Non-coherent demodulation begins with envelope detection, $|x[n]|$, where $x[n]$ is the audio signal, and $||$ denotes absolute value. A standard envelope detector extracts the envelope as a signal in its own right, by immediately lowpass filtering $|x[n]|$. The goal of Auditory Roughness detection is not, however, to perform an exhaustive analysis of the spectrum of $|x[n]|$; rather, we simply want to identify components of that spectrum in the neighborhood of 75Hz. To do so, we modulate several different frequency components down to baseband:

$$e_k[n] = |x[n]| \sin(\omega_k n) \tag{1}$$

where ω_K are a set of K different modulation frequencies in the neighborhood around 75Hz. Instantaneous roughness, $y[n]$, is then defined to be the weighted sum of demodulated envelopes:

$$y[n] = \sum_{k=1}^{K/2} w_k |x[n]| \left(\sin[\Omega_k n] + \sin[\Omega_{K+1-k} n]\right) \tag{2}$$

In Eq 2, Ω_k are frequencies near $75 * 2\pi/Fs$, chosen symmetrically so that $\Omega_{K+1-k} = 300\pi/Fs - \Omega_k$ and with symmetric combination weights $w_{K+1-k} = w_k$. To smooth out the signal, $y[n]$ is lowpass filtered to create the smoothed roughness signal $z[n]$:

$$z[n] = \sum_{m=0}^{255} b[m] y[n-m] \tag{3}$$

where $b[m]$ are the coefficients of a 256-tap FIR filter with a cutoff of 62.5Hz, and F_s is the sampling frequency. The smoothed roughness signal $z[n]$ is then downsampled by a factor of 128, to a sampling frequency of 62.5Hz. In order to match the output range of the original algorithms, its absolute value is computed as the Approximated Auditory Roughness $A[n]$:

$$A[n] = |z[n]| \tag{4}$$

In our approximation, Eq. 4 requires no or one additions and no multiplications. The FIR filter in Eq. 3 requires 256 real MAA. The weighted summation in Eq. 2 requires $K/2 = 2$ real MAA per sample for $K = 4$, which returns very similar results compared to larger K value. Since all factors in Eq 2 operate on the $|x[n]|$, we could pre-compute the result and store $\sum_{k=1}^{K/2} w_k \left(\sin[\Omega_k n] + \sin[\Omega_{K+1-k} n]\right)$ in our look-up-table, this reduces the complexity of this operation down to 1MAA/sample. When we implement Eq. 3 using a multiphase filter, the downsampling operations allow us to carry out both Eq. 2 and 3 once every 128 samples, thus requiring only $256 * (1 + 1)/128 \approx 4$MAA/sample. If the hardware specification allows, we can duplicate the filter coefficients for all possible multiplets in Eq 2, this brings the complexity down to 2MAA/sample, but at the cost of more memory. In practice, since our measure is not at all sensitive to phase, we can reduce memory usage at the cost of phase discontinuity. Goertzel algorithm [22] is a promising alternative for extracting the power

[2] http://www.ni.com/product-documentation/8169/en/\#toc4

around the frequency of interest. The Goertzel algorithm requires N real MAA and 1 CMAA to extract the power of a single discrete frequency, where N is the DCT window size. When we are only interested in 1 frequency component, the Goertzel algorithm has an advantage. But this advantage vanishes when the components of interests exceeds 2. The final computation has a complexity level similar to STE, which requires 1 MAA per sample. In comparison, Mel-frequency cepstral coefficients (MFCCs) require the computation of a full FFT once per frame; a 256-sample FFT computed once per 128 samples requires $256 \log_2(256)/128 = 16$ complex multiply-accumulate (CMAA) operations per sample, not including the filterbank accumulation, DCT, delta and second delta extractions. If we consider the remaining operation, the total complexity will not be less than 20 CMMA, the proposed approximated Roughness calculation is at least 10× less complex than MFCC, and is therefore well-suited to a sensor hub with no dedicated multiplier and with dozens of sensors to monitor.

4. Results

4.1. Mandarin Affective Speech

The Mandarin Affective Speech corpus [15] includes short phrases and sentences recorded from 68 speakers under different emotional states. Three of these emotions are neutral, anger and panic. One phrase is repeated by the same person under all emotion states. All audio files are normalized to have the same maximum amplitude.

In our experimental setup, we extracted a collection of features from every phrase under all emotional states. Features from every phrase pronounced under anger and panic were compared to the same phrase spoken neutrally. The five features include Auditory Roughness, which we extracted using open access Matlab toolbox MIRtoolbox [20], approximated roughness, STE, and the signal remainder after subtraction of its windowed media (SWM, [23]). Zero Crossing Rate (ZCR) was also included as it is a common feature for speech detection [24] and used in many previous papers to detect screaming and/or emotional speech [2, 3, 4, 5].

The following Fig 2 presents an example of the features for a short phrase spoken by the same person under three different emotional status.

As we can see from Fig 2, STE and SWM fail to effectively separate affective speech, especially angry speech, from neutral. STE is closely related to the loudness of the original audio, as a result, natural speech recorded at close distance will have high STE as well. SWM worked mainly through detecting impulse in the audio waveform, however, affective speech shows limited difference in this measure. On the other hand, the Auditory Roughness value of angry speech is much larger than that of neutral speech. The maximum value of the angry speech is many times higher than the neutral speech. One can observe the shape of the Auditory Roughness is different from any other measure: it does not resemble the envelope of the audio input. Our approximated roughness cannot reach the same level of separation as the original Auditory Roughness, yet it still records observably higher peaks for angry and frightened speech. One could also observe, throughout the audio, Auditory Roughness and our approximated roughness do not maintain a high value, but rather peak out in the middle.

We ran through all short phrases in the corpus. Fig 3 presents overall statistics and confirms our finding above. The left-most boxplot, for example, represents the ratio between 2

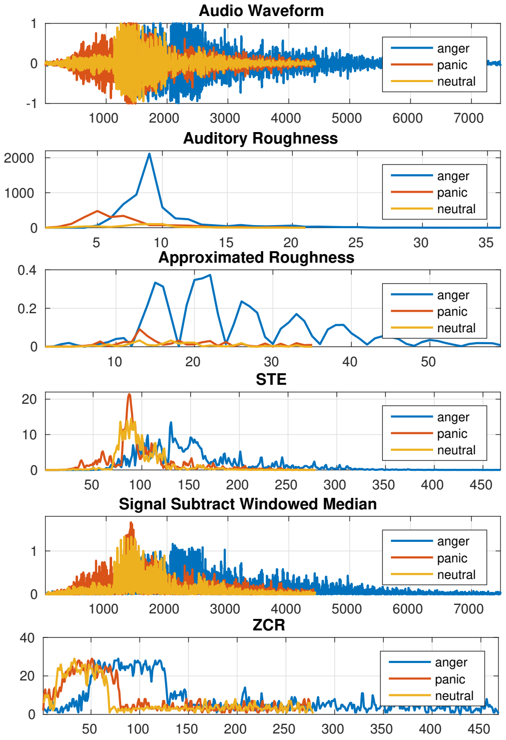

Figure 2: *Feature measures for the same short phrase by the same person under different emotional states.*

maximum values, one of which is for the feature extracted from an angry speech, the other being features extracted from the same speech under neutral state.

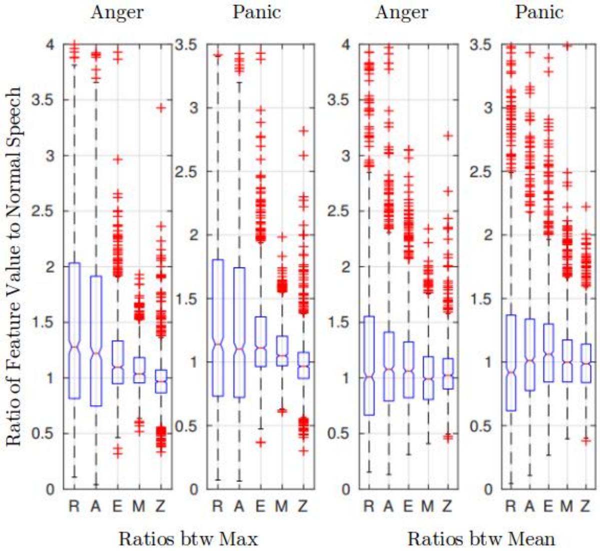

Figure 3: *Ratio between Maximum (left 2) Mean (right 2) values of angry to neutral (1, 3) and panic to neutral speech.*

The 5 letters in Fig 3 represent Auditory Roughness (R), Approximated Auditory Roughness (A), STE (E), SWM (M) and ZCR (Z). In all cases, a significant difference [25], with confidence level 0.005, can be found between the max value of roughness in any given audio file (both Auditory Roughness and our approximated roughness) in angry versus neutral speech. However, we can see roughness is less effective in separating panic and neutral speech. Also the average value of roughness, in any given waveform file, is less effective in separating affective versus neutral speech.

4.2. Youtube AudioSet

To test our approximated roughness in a more realistic setup, we built a simple test on the Google AudioSet [16] for screaming. This dataset contains video recordings of screaming and shouting, many happening in real life and recorded through smart phones. The events are typically shorter than scream events in movies and other sources, but the recording quality varies from video to video. Annotation is at a relatively coarse temporal granulatiry: a window of 10s is marked in every video, which either contains some short affective speech events, separated by pauses, or contains non-stop screaming extending outside the 10s window. We found this difficult to use as the event boundaries are not comprehensive, and markings for non-screaming regions are not presented. We took some time to conduct more fine-grained annotation on the balanced training subset and extended the duration of the feature window from 10s to 30s to include non-screaming regions. A couple of the files were dropped as they clearly contain no human vocalizations. We annotated 55 files with a total of 0.5h duration. Since our features are extracted at relatively high rate, this transfers to $112k$ sample. About 33% of the 0.5h audio is screaming or affective speech; this is a larger dataset than any other open-access corpus of human scream audio we found. These annotations are available online [3], and will be released with a creative commons CC-BY license.

We designed a simple experiment to mimic the pre-filter operation explained in Sec 2. We marked out screaming as audio events of interests and labeled the begin and end of each event. For each event in the video, if the feature value exceeds a preselected threshold, we considered the event detected; during the time between events, feature values exceeding the threshold are considered false alarms. Any event longer than 2s is separated into multiple events. The ROC curves comparing different features are presented in Fig 4. We sweep through different threshold value to obtain this ROC curve. It is worth mentioning, in most videos, human vocal do not span the full duration, yet we did not run speech activity detection on top of the algorithm. Doing so would add more computation and therefore undermine the purpose of pre-filtering. As we can see in Tab 1, this made the detection task difficult even for relatively complex feature and classifiers.

We can see from Fig 4, mostly, our approximated roughness returns better value than all other features. To our surprise, our approximated roughness actually outperforms the original Auditory Roughness; the original Auditory Roughness measure has the advantage only in limited cases. We suspect this is because the majority of the events in our collection are screaming and shouting, making them more fitted to the category described in [17], and less similar to affective speech. The Equal Error Rate for our approximated roughness is around 30%.

4.3. Beyond Pre-filtering

We ran more experiments, in the interest of answering the following two more questions. First, how hard is our AudioSet corpus? The dataset was released very recently, and not much work has been published using it. Second, how well will our low complexity feature work in a classifier, say linear SVM; and how will it compare to high complexity features like MFCC?

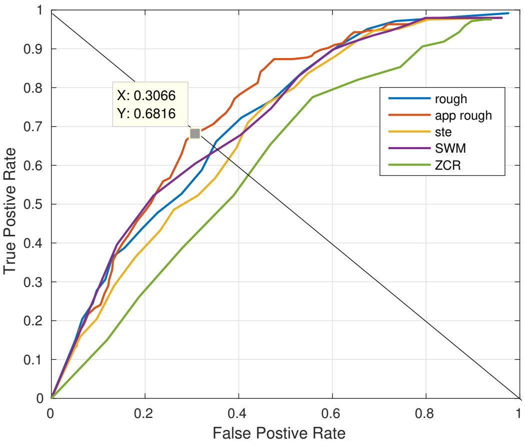

Figure 4: *ROC curves of testing different features on AudioSet.*

We stacked 5 frames, each containing approximated roughness, STE and ZCR and used a linear SVM to conduct a detection task. Our low complexity three-dimensional feature vector is compared to a standard MFCC vector extracted at 50ms per frame. As we can see in Tab 1, MFCC returns reasonable results compared to previous work, though with accuracy below that of most previous affective speech studies, suggesting that this corpus is difficult to classify. A feature vector with only STE and ZCR achieves an F1 score about 10% worse than that of MFCC; the same feature vector with Approximated Auditory Roughness included is only 8% worse than MFCC. The MFCC are extracted using the open source toolbox RASTA for Matlab [19] and the SVM is trained using SVMLight [26]. More could be compared between MFCC and our low-complexity features. But as the goal of this work is to study the potential of our approximated roughness, more detailed experiment on a complex classifier is beyond this work.

Table 1: *Linear SVM on AudioSet.*

	stack5(wo Rough)	stack5(w Rough)	MFCC
Prec	65.12%	67.25%	72.13%
F1	63.34%	65.47%	73.73%

5. Conclusions

In this work, we evaluated the Auditory Roughness as a feature for pre-filtering the audio for AED targeting human screaming and affective speech. Detecting these events is of interest to distributed security and surveillance AED systems. In order to be useful in a large distributed system, detection must have extremely low computational cost; MFCC is too expensive. We proposed a method to approximate roughness using a combination of frequency modulation and multiphase filtering. This allow us to extract a feature with computational cost similar to STE. We proved through experiments on the Mandarin Affective corpus, and a subset of the Google AudioSet, that our approximated roughness also has accuracy out-performing other low-complexity features.

6. Acknowledgements

We would like to thank Hui Jiang and Zizhen Liu for contributing to the annotation. Also we appreciate Dushyant Bhan and Xuesong Yang for joining the discussion.

[3] https://github.com/dihe2/Audioset-Balanced-Training-Annotation

7. References

[1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (iot): A vision, architectural elements, and future directions," *Future generation computer systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[2] T. Ahmed, M. Uppal, and A. Muhammad, "Improving efficiency and reliability of gunshot detection systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 513–517.

[3] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*. IEEE, 2007, pp. 21–26.

[4] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection in noisy environments," in *Signal Processing Conference, 2007 15th European*. IEEE, 2007, pp. 1216–1220.

[5] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 2005, pp. 1306–1309.

[6] J. J. Donovan and D. Hussain, "Audio-video tip analysis, storage, and alerting system for safety, security, and business productivity," Aug. 16 2011, uS Patent 7,999,847.

[7] S. Moroz, M. Pauli, W. Seisler, D. Burchick, M. Ertern, and E. Heidhausen, "Optical muzzle blast detection and counterfire targeting system and method," Feb. 9 2005, uS Patent App. 11/052,921.

[8] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.

[9] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," *arXiv preprint arXiv:1609.09430*, 2016.

[10] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell, "Soundsense: scalable sound sensing for people-centric applications on mobile phones," in *Proceedings of the 7th international conference on Mobile systems, applications, and services*. ACM, 2009, pp. 165–178.

[11] J.-C. Wang, J.-F. Wang, and Y.-S. Weng, "Chip design of mfcc extraction for speech recognition," *INTEGRATION, the VLSI journal*, vol. 32, no. 1, pp. 111–131, 2002.

[12] H. Von Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Longmans, Green, 1912.

[13] P. N. Vassilakis, "Auditory roughness as means of musical expression," *Selected Reports in Ethnomusicology*, vol. 12, 2005.

[14] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and models*. Springer Science & Business Media, 2013, vol. 22.

[15] T. Wu, Y. Yang, Z. Wu, and D. Li, "Masc: a speech corpus in mandarin for emotion analysis and affective speaker recognition," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*. IEEE, 2006, pp. 1–5.

[16] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dartaset for audio events," in *IEEE ICASSP*, 2017.

[17] L. H. Arnal, A. Flinker, A. Kleinschmidt, A.-L. Giraud, and D. Poeppel, "Human screams occupy a privileged niche in the communication soundscape," *Current Biology*, vol. 25, no. 15, pp. 2051–2056, 2015.

[18] P. N. Vassilakis and K. Fitz, "Sra: A web-based research tool for spectral and roughness analysis of sound signals," in *Proceedings of the 4th Sound and Music Computing (SMC) Conference*, 2007, pp. 319–325.

[19] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.

[20] O. Lartillot, P. Toiviainen, and T. Eerola, "Mirtoolbox," 2008.

[21] v. W. Aures, "A procedure for calculating auditory roughness," *Acustica*, vol. 58, no. 5, pp. 268–281, 1985.

[22] K. Banks, "The goertzel algorithm," *Embedded Systems Programming*, vol. 15, no. 9, pp. 34–42, 2002.

[23] A. Dufaux, "Detection and recognition of impulsive sounds signals," *Institute de Microtechnique Neuchatel, Switzerland*, 2001.

[24] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in *American Society for Engineering Education (ASEE) Zone Conference Proceedings*, 2008, pp. 1–7.

[25] N. Cressie and H. Whitford, "How to use the two sample t-Test," *Biometrical Journal*, vol. 28, no. 2, pp. 131–148, 1986.

[26] T. Joachims, "Svmlight: Support vector machine," *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund*, vol. 19, no. 4, 1999.