AUTOMATIC CONFLICT DETECTION IN POLICE BODY-WORN AUDIO

Alistair Letcher [1,†]   *Jelena Trišović* [2,†]   *Collin Cademartori* [3]   *Xi Chen* [4]   *Jason Xu* [5]

[†] Joint first authors   [1] University of Oxford   [2] University of Belgrade
[3] Brown University   [4] Carleton College   [5] University of California, Los Angeles

ABSTRACT

Automatic conflict detection has grown in relevance with the advent of body-worn technology, but existing metrics such as turn-taking and overlap are poor indicators of conflict in police-public interactions. Moreover, standard techniques to compute them fall short when applied to such diversified and noisy contexts. We develop a pipeline catered to this task combining adaptive noise removal, non-speech filtering and new measures of conflict based on the repetition and intensity of phrases in speech. We demonstrate the effectiveness of our approach on body-worn audio data collected by the Los Angeles Police Department.

*Index Terms*— Conflict detection, speech repetition, body-worn audio.

1. INTRODUCTION

Body-worn technology is beginning to play a crucial role in providing evidence for the actions of police officers and the public [1], but the quantity of data generated is far too large for manual review. In this paper we propose a novel method for automatic conflict detection in police body-worn audio (BWA). Methodologies from statistics, signal processing and machine learning play a burgeoning role in criminology and predictive policing [2], but such tools have not yet been explored for conflict detection in body-worn recordings. Moreover, we find that existing approaches are ineffective when applied to these data off-the-shelf.

Notable papers on conflict escalation investigate speech overlap (interruption) and conversational turn-taking as indicators of conflict in *political debates*. In [3], overlap statistics directly present in a hand-labelled dataset are used to predict conflict, while [4] detect overlap through a Support Vector Machine (SVM) with acoustic and prosodic features. The work in [5] compares variations on both methods. Using automatic overlap detection, their method achieves 62.3% unweighted conflict accuracy at best in political debate audio. This approach is all the less effective on BWA data, which is far noisier and more diverse. Besides being harder to detect, overlap serves as an unreliable proxy for conflicts between

police and public: these often involve little to no interruption, especially in scenarios where the officer is shouting or otherwise dominating the interaction.

We propose new metrics that successfully predict conflict in BWA along with speech processing and modeling techniques tailored to the characteristics of this data. Section 2 details adaptive pre-processing stages, feature extraction, and a SVM for non-speech discrimination. In Section 3, we develop metrics based on repetition using audio fingerprinting and auto-correlation techniques. This is based on the observation that conflict largely occurs in situations of non-compliance, where the officer repeats instructions loudly and clearly. Finally, performance is evaluated on a dataset of 105 BWA files provided by the Los Angeles Police Department (LAPD) in Section 5. The illustration below summarizes our conflict detection procedure.
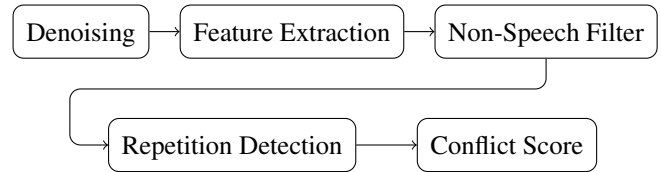
2. PRE-PROCESSING AND FILTERING

The success of our approach relies on pre-processing steps catered to the task at hand. We apply adaptive denoising procedures followed by feature extraction for supervised discrimination of non-speech, also called Voice Activity Detection.

2.1. Denoising

Persistent noise like traffic, wind and babble as well as short-term bursts of noise including sirens, closing doors and police radio are present along with speech in BWA audio. We filter persistent but non-stationary background noise based on optimally-modified log-spectral amplitude (OM-LSA) speech estimation, and apply minima controlled recursive averaging (MCRA) as described in [6]. Briefly, this approach computes the spectral gain while accounting for speech presence uncertainty, ensuring that noise removal best preserves speech components even when the signal-to-noise ratio (SNR) is low.

Let $x(n)$ and $d(n)$ denote speech and (uncorrelated, additive) noise signals respectively. Then the observed signal is $y(n) = x(n) + d(n)$, where n is a discrete-time index. The spectrum is obtained by windowing $y(n)$ and applying the short-term Fourier transform (STFT), denoted $Y(k,l)$ with frequency bin k and time frame l. The STFT of clean speech $X(k,l)$ can be estimated as $\hat{X}(k,l) = G(k,l)Y(k,l)$, where $G(k,l)$ is the spectral gain function. Via the LSA estimator, we apply the spectral gain function which minimizes

$$E\left[(log|X(k,l)| - log|\hat{X}(k,l)|)^2\right].$$

Let hypotheses $H_0(k,l)$ and $H_1(k,l)$ respectively indicate speech absence and presence in the kth frequency bin of the lth frame. Assuming independent spectral components and STFT coefficients to be complex Gaussian variates, the spectral gain for the optimally modified LSA is given by

$$G(k,l) = G_{H_1}(k,l)^{p(k,l)}G_{min}^{1-p(k,l)}.$$

Here $G_{H_1}(k,l)$ represents the spectral gain which should be applied in the case of speech presence and G_{min} is the lower threshold for the gain in case of speech absence, preserving noise naturalness. $p(k,l)$ is the *a posteriori* speech probability $P(H_1(k,l)|Y(k,l))$, computed using the estimates of noise and speech variance $\lambda_d(k,l)$ and $\lambda_x(k,l)$, the *a priori* SNR $\xi(k,l) = \frac{\lambda_x(k,l)}{\lambda_d(k,l)}$ and the *a priori* speech absence probability $q(k,l) = P(H_0(k,l))$.

To estimate the time-varying spectrum of non-stationary noise $\lambda_d(k,l)$, we employ temporal recursive smoothing during periods of speech absence using a time-varying smoothing parameter. The smoothing parameter depends on the estimate of the speech presence probability, obtained from its previous values and the ratio between the local energy of the noisy signal and its derived minimum. Given $\hat{\lambda}_d(k,l)$ we may immediately estimate the *a posteriori* SNR,

$$\gamma(k,l) = \frac{|Y(k,l)|^2}{\lambda_d(k,l)}.$$

This is used to estimate the *a priori* SNR given by

$$\hat{\xi}(k,l) = \alpha G_{H_1}(k,l-1)^2\gamma(k,l-1)$$
$$+ (1-\alpha)\max\{\gamma(k,l) - 1, 0\},$$

with weight α controlling the noise reduction and signal distortion. The estimate $\hat{\xi}(k,l)$ allows for computing the probability of *a priori* speech absence as described in [6], which finally enables computation of the spectral gain and in turn speech spectrum.

We perform this filtering method three times in sequence to reliably remove residual noise that may persist after one stage of filtering. Doing so produces excellent results, eliminating most persistent noise while crucially avoiding attenuation of weak speech components. Nevertheless, sudden bursts of noise are rarely eliminated because the filter cannot adapt in time. We apply the method below to remove them, which is equally crucial to reliable repetition detection.

2.2. Feature Extraction and Non-Speech Filter

The task of this section is to filter remaining non-speech. To begin, the audio signal is split into overlapping frames of size 0.06s and with 0.02s steps between start times. Over each frame, we compute 23 short-term features consisting of the first 13 Mel-Frequency Cepstral Coefficients (see [7]); zero-crossing rate; energy and energy entropy; spectral centroid, spread, entropy, flux and rolloff; fundamental frequency and harmonic ratio. Features which require taking the Discrete Fourier Transform are first re-weighted by the Hamming window. Since many meaningful speech characteristics occur in a longer time-scale, we additionally include the mid-term features obtained by averaging our short-term features across frames of size 0.3s and step 0.1s.

We apply a SVM with Radial Basis Function kernel [8, Chap. 12] to discriminate between speech and non-speech in this feature space. The SVM is trained on 38 minutes (22733 frames) of labeled speech and 47 minutes (28239 frames) of non-speech from BWA data. To evaluate predictive power we perform cross-validation (CV) with 10 folds [8, Chap. 7]. Our results are displayed in Table 1 and compare favourably with state-of-the-art papers in speech detection, which obtain error rates no lower than 5% in [9] and 12% in [10], on clean and noisy data (SNR at least 15 dB) respectively. Their learning algorithms include SVMs, Gaussian Mixture Models and Neural Networks.

False Positive	False Negative	Total Error
1.26%	3.61%	2.31%

Table 1: 10-fold CV error in speech/non-speech detection.

3. REPETITION DETECTION AND SCORING

Having eliminated most of the non-speech and noise, we turn to detecting repetitions as a measure of conflict. We split the audio into regions of interest and compare them using fingerprint and correlation methods based on [11] and [12].

3.1. Segmentation and Repetition

In order to reduce the time it takes to search for repetitions, we automatically break the signal into regions which contain entire syllables, words, or phrases. We begin by applying a band-pass filter between 300 and 3000Hz, which we found to carry the most information about speech in our recordings.

Let $E'(t)$ be the energy (squared amplitude) of the signal in a window of length 0.05s starting at time t, and define $E(t) = \mathbf{1}_{\{E'(t)>0.05\}}E'(t)$. This threshold filters windows with energy below 0.05, in which the signal-to-noise ratio is too small for reliable repetition detection. We define points $t_0 = 0$ and $t_1, \ldots, t_n$ by the following criteria:

1. $E(t_i)$ is a local minimum of $E(t)$.

2. There exists $t \in [t_{i-1}, t_i]$ such that $E(t) \neq E(t_i)$.

3. Each t_i is the earliest time satisfying (1) and (2).

This somewhat cumbersome definition deals with the possibility that E attains the same local minimum value at consecutive times, by taking the earliest such time. We define $t'_1, \ldots, t'_n$ analogously by taking the latest such times. This defines regions $[t'_{i-1}, t_i]$ delimited by local minima which are not trivially flat inside. We then let T_i be the earliest time in $[t'_{i-1}, t_i]$ such that $E(T_i)$ is a local maximum. Finally, let $s_0 = 0$ and define new endpoints s_j recursively by

$$s_j = \min \{t_i \mid t_i \geq s_{j-1} \text{ and } E(T_i) - E(t_i) > \sigma \} ,$$

where σ is the standard deviation of E. We define s'_j analogously with t, s replaced by t', s' everywhere. This isolates regions $R_j = \left[s'_{j-1}, s_j \right]$ which start at the bottom of an energy spike and finish at the other end, ignoring spikes that are too small to be meaningful. The definitions are illustrated in Figure 1 below, where one of our BWA spectrograms is overlaid with a depiction of its energy curve.
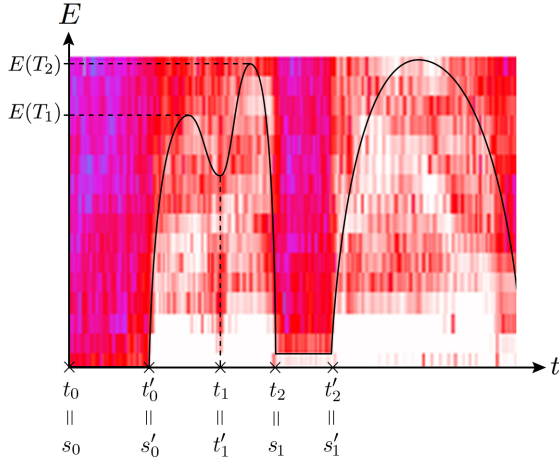

Fig. 1: Spectrogram overlayed with energy across time.

In this example, the local minimum $t_1 = t'_1$ is not equal to any s_j because the energy distance $E(T_1) - E(t_1)$ to the previous maximum is less than σ. The resulting regions usually contain syllables or short words. In order to form regions of longer words and short phrases, we concatenate these initial regions together. First we choose a cutoff distance $C = 0.02$s and let $k = 0$. For each region $R_j = [r_j^1, r_j^2]$, proceed as follows. If

$$\left| r_{j+k}^2 - r_{j+k+1}^1 \right| < C$$

then add a new region $[r_j^1, r_{j+k+1}^2]$, increment k, and repeat until the condition is false. Finally, segments shorter than 0.05s are discarded since any syllable takes longer to pronounce. These contain too little information to be reliably distinguished, and do not provide meaningful repetitions.

Fingerprinting: Following [11], our first measure associates a binary rectangular array called *fingerprint* to each region, and computes the percentage of entries at which two arrays differ. Regions are binned into N non-overlapping windows of length 0.1s in the time domain, which are then partitioned into $M = 32$ bands of equal length between 300 and 3000Hz in the frequency domain. We define $E_{n,m}$ to be the energy of window n within frequency band m for $1 \leq n \leq N, 1 \leq m \leq M$. We then take second-order finite differences

$$\Delta^2(n, m) = [E_{n,m} - E_{n,m+1}] - [E_{n-1,m} - E_{n-1,m+1}] ,$$

which provide a discretized measure of curvature in the spectral energy distribution over time. The value of the fingerprint at position (n, m) is now defined as $F(n, m) = \mathbf{1}_{\{\Delta^2(n,m)>0\}}$. Given a fingerprint pair, the percentage of positions at which arrays differ provides a measure E of dissimilarity between regions.

Correlation: The second metric, based on [12], makes use of the correlation between Fourier coefficients over short windows. Regions R_1 and R_2 are first split into N overlapping windows $w_1^i, \ldots, w_N^i$ for $i = 1, 2$. For each window w_n^i, let $\hat{w}_n^i(m)$ be the Fourier coefficients corresponding to frequencies between 300 and 3000Hz. For each m, we compute the correlation $C(m) = \frac{\sigma_{1,2}(m)}{\sigma_1(m)\sigma_2(m)}$ between the values of the mth coefficient of the two regions, where

$$\sigma_i(m) = \sqrt{\sum_{n=1}^{N} (\hat{w}_n^i(m) - \hat{w}^i(m))^2} ,$$

$$\sigma_{1,2}(m) = \sum_{n=1}^{N} (\hat{w}_n^1(m) - \hat{w}^1(m))(\hat{w}_n^2(m) - \hat{w}^2(m)) ,$$

and $\hat{w}^i(m) = \frac{1}{N} \sum_{n=1}^{N} \hat{w}_n^i(m)$. Finally, averaging $C(m)$ over $m = 1, \ldots, M$ yields an overall similarity measure C for R_1 and R_2. This measure is less sensitive and produces more false positives than fingerprints. On the other hand, correlation can pick up on noisy repetitions where fingerprints fail. Our approach is to combine these methods so as to balance their strengths and weaknesses.

3.2. Scoring

Combining the fingerprint and correlation metrics into a single score, define $S(E, C) = \sqrt{f_1(E)f_2(C)}$ where

$$f_1(E) = \mathbf{1}_{\{E<0.3\}} + \mathbf{1}_{\{0.3 \leq E \leq 0.45\}} \left[\frac{20}{3}(0.3 - E) + 1 \right] ,$$

$$f_2(C) = \mathbf{1}_{\{C>0.55\}} + \mathbf{1}_{\{0.25 \leq C \leq 0.55\}} \left[\frac{10}{3}(C - 0.25) \right] .$$

The functions f_1 and f_2 are designed to convert the outputs of each method to more meaningful levels of confidence that can

be compared and combined, taking into account our empirical observations about the behavior of each method. For example, both our experiments and the paper [11] suggest that a fingerprint difference above 45% corresponds to regions that are almost certainly not repetitions. Similarly, we are almost certain that a fingerprint difference below 30% corresponds to repeated regions. After evaluating segments, the measures are aggregated to score the entire audio file. This total score is computed as the average of non-zero scores among the top 5% of unique comparisons. As such, this score is higher for files that contain more or clearer repetitions, and lower for those with fewer or less distinguishable repetitions.

Though repetition tends to be more frequent in scenarios of conflict, significant disputes can further be distinguished from mild ones via a measure of intensity. High conflict scenarios often involve shouting or loud commands, producing higher energy. Accordingly, an intensity score is computed by averaging the energy among the same top 5% set of repetitions. The overall conflict score for an audio signal is the product of its repetition and intensity scores.

4. RESULTS AND DISCUSSION

We test our approach on a collection of 105 body-worn audio files provided by the LAPD, of lengths between 3 and 30 minutes each. The files are manually labeled according to level of conflict, where the classes and criteria are as follows:

2. **High conflict (3 files):** active resistance, escape, drawing of weapon, combative arguments.

1. **Mild conflict (15 files):** questioning of officer judgment, avoiding questions, avoiding to comply with commands, aggressive tone.

0. **Low conflict (87 files):** none of the above.

Figure 2 is a plot of files ranked in descending order of conflict score as determined by our method, illustrating that those labeled as high or mild conflict are concentrated toward the top. More specifically, all three files labeled as high conflict occur in the top 10 scores.

In general, the three classes are correctly prioritized by the scoring algorithm. Only 4 of the 18 files in classes 1 and 2 fell below rank 24. In other words, 78% of the files with any conflict would be found by reviewing only the top 23% of files in the list. The mean scores for each class, displayed in the figure, are clearly well-separated. Our method can thus be used to significantly reduce the time it takes to manually locate files containing conflict. Further, the algorithm automatically isolates the repetitions detected in a given file, which amount to very short audio portions relative to the entire signal. As such, we may quickly search through the high-rank audio files by listening to these portions.

Given a larger dataset, one could automatically determine the adequate scores to label a file as containing high/mild/low
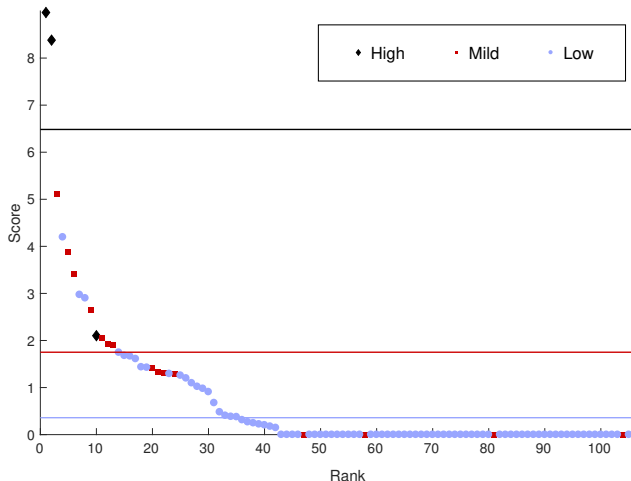

Fig. 2: Plot of conflict score against rank. Horizontal lines depict the mean score for the class of corresponding color.

conflict using a learning algorithm of choice. One may also input the fingerprinting, auto-correlation and intensity measures as features into the learning algorithm, producing a decision hyperplane in three dimensions.

In addition to their immediate use, our findings may also inform policy to better aid future work. We find that officer speech is vastly more informative than other voices, which are less comprehensible and contribute to false positives. To further improve performance, one may exclude all speech except that of the officer. This falls under the task of speaker diarization—see [13] for a recent review—and most studies in this area are based on relatively clean data (broadcast meetings, conference calls). State-of-the-art methods including [14] and [15] achieve no less than 18% diarization error rate on average, rising to 30% for some of the meetings, but perform much worse when applied to our BWA data. This obstacle may be overcome provided additional labeled data. Given a sample of the officer's voice that can be used to identify them elsewhere, our supervised learning task translates to speaker verification [16]. Such data could be provided by requiring officers to record a few minutes of clean speech once in their career; this sample could then be overlaid with non-speech extracted by our pipeline to render it comparable with BWA files featuring a range of noise environments.

5. CONCLUSION

To summarize, we offer a novel method for automatic conflict detection which is successful for applications in police body-worn audio. We are able to automatically select audio files which are very likely to contain conflict, despite a small number of high conflict files. We propose eliminating non-officer speech through speaker verification and using all three sub-scores as learning features to improve these results.

6. REFERENCES

[1] Barak Ariel, William A. Farrar, and Alex Sutherland, "The effect of police body-worn cameras on use of force and citizens' complaints against the police: A randomized controlled trial," *Journal of Quantitative Criminology*, vol. 31, no. 3, pp. 509–535, Sep 2015.

[2] George O. Mohler, Martin B. Short, Sean Malinowski, Mark Johnson, George E. Tita, Andrea L. Bertozzi, and P. Jeffrey Brantingham, "Randomized controlled field trials of predictive policing," *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1399–1411, 2015.

[3] Félix Grèzes, Justin Richards, and Andrew Rosenberg, "Let me finish: automatic conflict detection using speaker overlap," in *Interspeech*, 2013.

[4] Marie-José Caraty and Claude Montacié, *Detecting Speech Interruptions for Automatic Conflict Detection*, pp. 377–401, Springer International Publishing, 2015.

[5] Samuel Kim, Sree Harsha Yella, and Fabio Valente, "Automatic detection of conflict escalation in spoken conversations," in *Interspeech*. ISCA, 2012.

[6] Israel Cohen and Baruch Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[7] Kishore Prahallad, "Speech technology: A practical introduction, topic: Spectrogram, cepstrum, and mel-frequency analysis," Carnegie Mellon University, 2011.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*, Springer-Verlag New York, 2009.

[9] Benjamin Elizalde and Gerald Friedland, "Lost in segmentation: Three approaches for speech/non-speech detection in consumer-produced videos," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–6.

[10] Yuexian Zou, Weiqiao Zheng, Wei Shi, and Hong Liu, "Improved voice activity detection based on support vector machine with high separable speech feature vectors," in *2014 19th International Conference on Digital Signal Processing*, Aug 2014, pp. 763–767.

[11] Jaap Haitsma and Ton Kalker, "A highly robust audio fingerprinting system with an efficient search strategy," *Journal of New Music Research*, vol. 32, no. 2, pp. 211–221, 2003.

[12] Cormac Herley, "ARGOS: Automatically extracting repeating objects from multimedia streams," *Trans. Multi.*, vol. 8, no. 1, pp. 115–129, Sept. 2006.

[13] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[14] Gerald Friedland, Adam Janin, David Imseng, Xavier Anguera, Luke Gottlieb, Marijn Huijbregts, Mary Tai Knox, and Oriol Vinyals, "The ICSI RT-09 speaker diarization system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 371–381, 2012.

[15] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky, "A sticky HDP-HMM with application to speaker diarization," *The Annals of Applied Statistics*, vol. 5, no. 2A, pp. 1020–1056, 06 2011.

[16] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.