AUTOMATIC CONFLICT DETECTION IN POLICE BODY-WORN AUDIO

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

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

Introduction

Body-worn technology is starting to play a crucial role in providing evidence for the actions of police officers and the public, but the quantity of data generated is far too large for manual review. Moreover, existing metrics for automatic conflict detection such as speech overlap and conversational turn-taking are ineffective when applied to this data. Besides being extremely difficult to detect in such noisy and diverse environments, overlap is a poor indicator of conflict in police-public interactions. The latter involve little to no interruption, particularly in scenarios where the officer is shouting or otherwise dominating the interaction. Instead we observe that conflict occurs in situations of noncompliance, where the officer often repeats instructions loudly and clearly. As such, we develop a pipeline combining adaptive noise removal, non-speech filtering and new measures of conflict based on the repetition of phrases in speech, using audio fingerprinting and correlation techniques. We demonstrate the effectiveness of our approach on body worn audio data collected by the Los Angeles Police Department (LAPD).

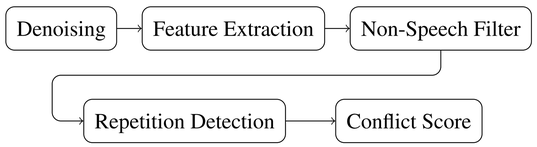

Fig. 1: Summary of the conflict detection procedure.

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 police body worn 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 [1]. 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 is low.

Denote by $Y(k,l)$ the spectrum of noisy speech, obtained by windowing and applying a short-term Fourier transform (STFT) with frequency bin k and time frame l to the audio. The clean speech spectrum $X(k,l)$ 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 gain function that minimizes the following expression:

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

Acknowledgements

This research was supported by the LAPD and the UCLA Institute for Pure and Applied Mathematics, NSF Grant DMS-0931852.

Assuming independent spectral components and STFT coefficients to be complex Gaussian variates, the spectral gain is given by

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

- G_{H_1} is the gain applied in the case of speech presence;
- G_{min} is the lower threshold for the gain applied in the case of speech absence, preserving noise naturalness;
- $p(k,l)$ is the a posteriori speech probability.

These parameters are computed using statistical estimates of noise and speech variance, as well as the a priori speech absence probability.

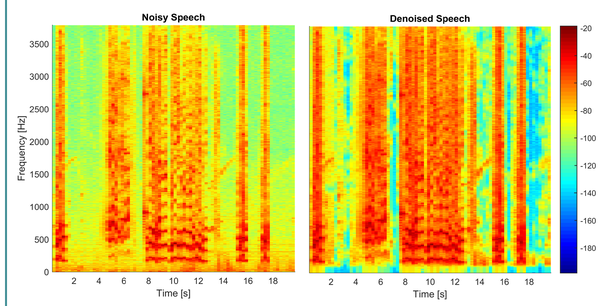

Fig. 2: Sample spectrograms of noisy (left) and filtered speech (right).

Non-Speech Filter

This step is performed in order to filter the non-speech remaining after the denoising stage. To begin, the audio signal is split into overlapping frames. Over each frame, 23 short-term features are computed, consisting of the first 13 Mel-Frequency Cepstral Coefficients, zero-crossing rate, energy and energy entropy, spectral centroid, spread, entropy, flux, roll-off, fundamental frequency and harmonic ratio. To account for meaningful speech characteristics occurring on a longer time-scale, we additionally include the mid-term features obtained by averaging these features across 15 consecutive short-term frames.

We apply a Support Vector Machine (SVM) with Radial Basis Function kernel to discriminate between speech and non-speech in this feature space. To evaluate predictive power we perform cross-validation (CV) with 10 folds based on 38 minutes of labeled speech and 47 minutes of unlabeled speech. Our results are displayed in Table 1 and compare favourably with state-of-the-art papers in speech detection.

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

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

Repetition Detection

In the presence of residual noise, detecting speech repetitions requires a robust method capturing only the general trends of a waveform. We develop a segmentation technique to automatically split the audio into segments containing entire syllables/words/phrases, and apply a band-pass filter between 300 and 3000 Hz. We then apply fingerprint [2] and correlation methods for repetition detection.

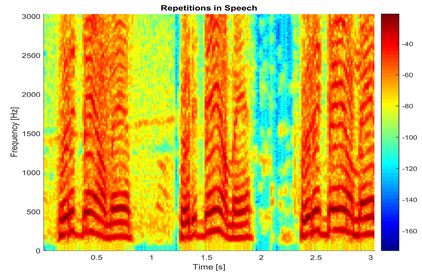

Fig. 3: Sample spectrograms for three instances of a single phrase.

Fingerprint Method

Each segment is divided into N windows of length 0.1s and 32 frequency bands. The energy in window n and band m is denoted by $E(n,m)$, while the second-order finite difference

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

is computed. For each window n and band m of the audio segment, the fingerprint $F(n,m)$ equals to 1 if $\Delta^2_{n,m} > 0$ and 0 otherwise. Given a fingerprint pair, the percentage of positions at which arrays differ provides a measure E of dissimilarity between regions.

Correlation Method

This method makes use of the correlation between Fourier coefficients over short windows. Regions R_1 and R_2 that are being compared are first split into overlapping windows. For every window, we compute Fourier coefficients corresponding to frequency bands between 300 and 3000 Hz. For each coefficient m, we compute the normalized correlation

$$C(m) = \frac{\sigma_{1,2}(m)}{\sigma_1(m)\sigma_2(m)} \ ,$$

where $\sigma_i(m)$, for $i \in \{1,2\}$, represents the variance of the mth coefficient over all windows for each of the regions and $\sigma_{1,2}(m)$ equals to the covariance of the mth coefficients corresponding to R_1 and R_2. Averaging $C(m)$ over each band m yields an overall similarity measure 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 thus to combine both methods so as to balance their strengths and weaknesses.

Conflict Scoring

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

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

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

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. 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% comparisons.

Results

We test our approach on a collection of 105 body worn audio files provided by the LAPD. The files are manually labeled according to the level of conflict: high (3), mild (15) or low (87).

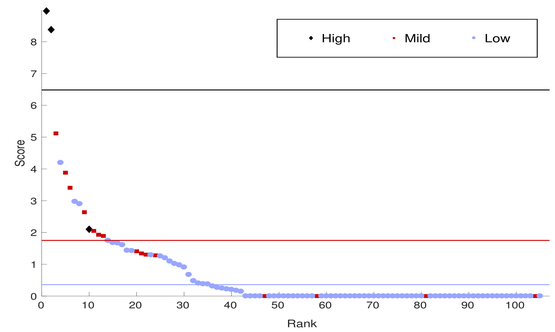

Fig. 4: Plot of conflict score against rank. Horizontal lines depict the mean score for each class.

Figure 4 is a plot of files ranked in descending order of conflict score as determined by our method illustrating the following:

- The videos labeled as high or mild conflict are concentrated toward the top;
- The mean scores for each class are well-separated.

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, it is possible to quickly search through the high-rank audio files by listening to these portions and drastically reduce the time needed for manual review.

References

[1] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," Signal Processing, vol. 81, no. 11, pp. 2403–2418, 2001.
[2] J. Haitsma and T. 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.