Identification and Localization of Siren Signals

Hankyu Jang
Indiana University Bloomington
School of Informatics and Computing
hankjang@indiana.edu

Leonard Yulianus
Indiana University Bloomington
School of Informatics and Computing
lyulianu@iu.edu

Sun Woo Kim
Indiana University Bloomington
Department of Computer Science
kimsunw@iu.edu

Abstract

Identification and localization of ambulance sirens is an imperative feature especially for automated cars. Existing detection systems involve either training on highly specialized data or through expensive computation. Ambulance siren detection requires high precision and rapid classification such that the self-driving vehicles can abide to the rules of the road and accommodate emergency situations on the road. In this paper we explore a fast singular vector model classification method based on non-negative matrix factorization (NMF) and support vector machines (SVM). We train a dimensionality reduction model using our limited set of training ambulance signals that is sufficient to detect the presence of ambulance sirens in noisy signals. With the set of basis vectors, we perform localization with microphones in a triangle configuration. Our experiment on simulated data shows that our technique is capable of detecting and accurately estimating the location of the ambulance.

Keyword: non-negative matrix factorization, support vector machines, signal detection, localization

Introduction

With self-driving cars rising in popularity, it is crucial that the autonomous vehicles be able to discern an approaching ambulance and yield to make way. Due to the weight of the emergency situation, it is of utmost importance we detect siren signals at a rapid pace and deduce where it is coming from. In this paper, we introduce our approach of integrating signal detection with localization to address this problem.

Signal detection refers to the problem of identifying and discerning between the signals from noises. Localization aims to estimate the location from which a signal is originating from. These are indispensable features in sound processing systems including robust speech recognition (Karray and Martin 2003) and application of antennas in wireless communications (Godara 1997).

Research in recent years have focused on developing robust signal detection systems. Supervised and semi-supervised approaches train on mixtures of signals and noise that are matched or determined to be similar to that of the application and have been labeled with their corresponding activities (Guizhongz, Engineering, and Engineering 2002) (Sohn 1999). Since these methods require specialized training data that are difficult to obtain, unsupervised learning methods (at the users perspective) have been developed. One such method constructs a universal speech model (Germain, Sun, and Mysore 2013) through NMF and training on clean speech from a number of speakers.

Several methods have been proposed for localization (Su, Su, and Morf 1983) (Wang 1985). One such method is MUSIC (MUltiple SIgnal Classification) with Coherent Signal Subspace(CSS) (Wang 1985) which is an effective method with high spatial resolution. The disadvantages are that the *a priori* of an approximate localization estimate is required and the pre-estimation accuracy effects the final estimation result. Another method is array processing for estimating the location. The performance for this approach improves as the number of sensors are increased; however, in practical use this poses a huge limitation due to the physical size of the apparatus on which the equipments are placed.

A feasible proposed method utilizes a triangular microphone configuration. Through this configuration, a uniform resolution with respect to the location can be achieved. Each microphone pair faces to different angles, which allows us to expect improvement in the resolution through integration of the array data at these three pairs. Furthermore, this method will not require any *a priori* location by using the subspace analysis of the integrated array data (Hioka and Hamada 2004).

We explore a classification and localization method using a fast singular vector model classification method based on NMF and the triangular microphone configuration. The model of the ambulance siren is trained by retrieving a general basis vector of an ambulance signal through NMF. The robustness of our detection system will be tested using a set of test samples that mimic the real world situation. The microphones positioned at vertices of a triangle will pick up the same signals that are shifted in time depending on the distance of the ambulance to each vertex, from which we finally estimate the location.

Identification

Identification of siren signals and distinguishing them from noises is a non-trivial task. Not only are there numerous variations of siren signals, but there could also be infinitely different noises present in the scene of recording.

We approached the identification sub-task by employing feature extraction with non-negative matrix factorization (NMF) followed by classification via support vector machines (SVM). The assumption that samples lie on a lower dimensional subspace is made. NMF is performed for our dimensionality reduction procedure. It was deduced that NMF would be advantageous for our task compared to other matrix factorization methods such as principle component analysis (PCA) since NMF does not necessitate orthogonality in the basis vectors and still manage to find a new subspace regardless. For classification, we performed SVM. The dataset was gathered and labeled manually to pose this as a supervised learning problem. The SVM algorithm will build a model to assign the data samples into categories, ambulance or “others”, to perform the classification task.

Data

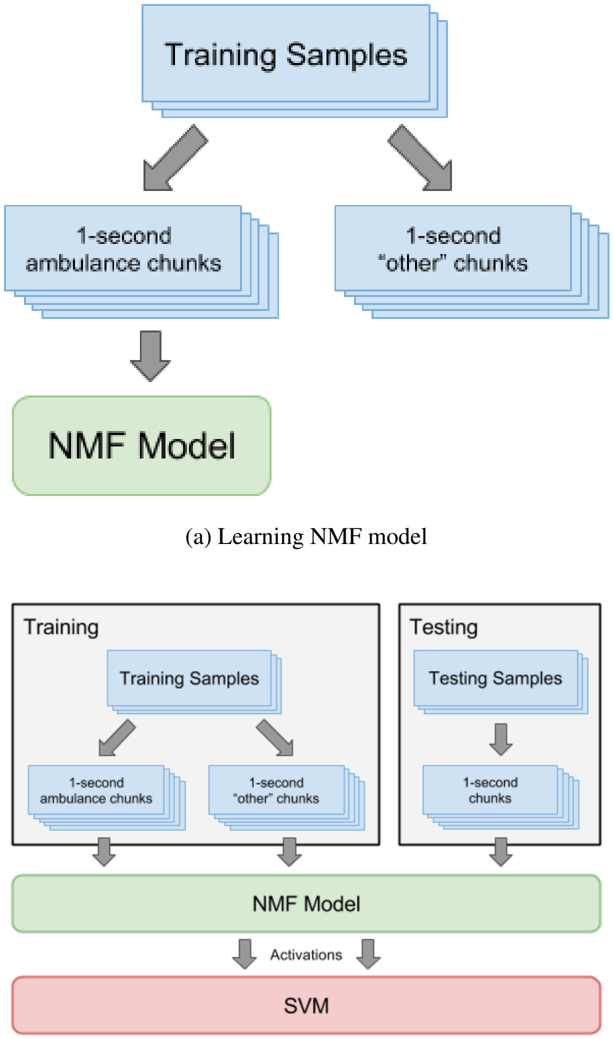
The dataset was split into two for the identification procedure; training and testing data. The audio samples were collected from online sources and real ambulances passing us by. Since most of the sirens were not recorded in a controlled environment, other noises were mixed, making our data “dirty”. This poses a difficulty in our procedure; therefore, we broke our collected data into 1-second chunks and labeled them manually. The manual labeling process was done by listening to each sample and labeling them into either ambulance or “others” classes using our subjective judgment. If a 1-second audio sample was judged to contain a portion of a siren signal, we labeled it as an ambulance.

Training Procedure

The training procedure is divided into two parts: learning the dimensionality reduction model and training the classifier. To learn the dimensionality reduction model, we took all the 1-second chunk samples that are labeled as an ambulance and concatenated them together into a long audio sample, **S**. We proceeded to train an NMF model **M** out of **S**, which was then used by the classifier to extract features from input signals (Figure 1 a). Instead of working on the raw 1-second signals, the classifier trains and tests on features extracted by **M** (Figure 1 b). We use SVM as our classifier with penalty parameter C of 1000, which we found works best in our training and testing dataset.

Testing Procedure

Upon completion of the training step, we proceed to testing (Figure 1 b). To simulate the process of real-time identification, we tested on 3 moderately long samples: only ambulance siren signals, only noises, and a mixture of both. The classifier works on 1-second segments of the long samples; thus, we slid the classifier over the samples at 250ms intervals. In order to smooth the detection, we introduced transition probability priors. We assigned a 0.5 transition probability for ambulance sirens to emerge from noise. On the other hand, given an ambulance siren has been detected and noise detection occurs, we assign a 0.1 transition probability from ambulance to noise. The results for the testing can be seen in Figure 2.



(a) Learning NMF model

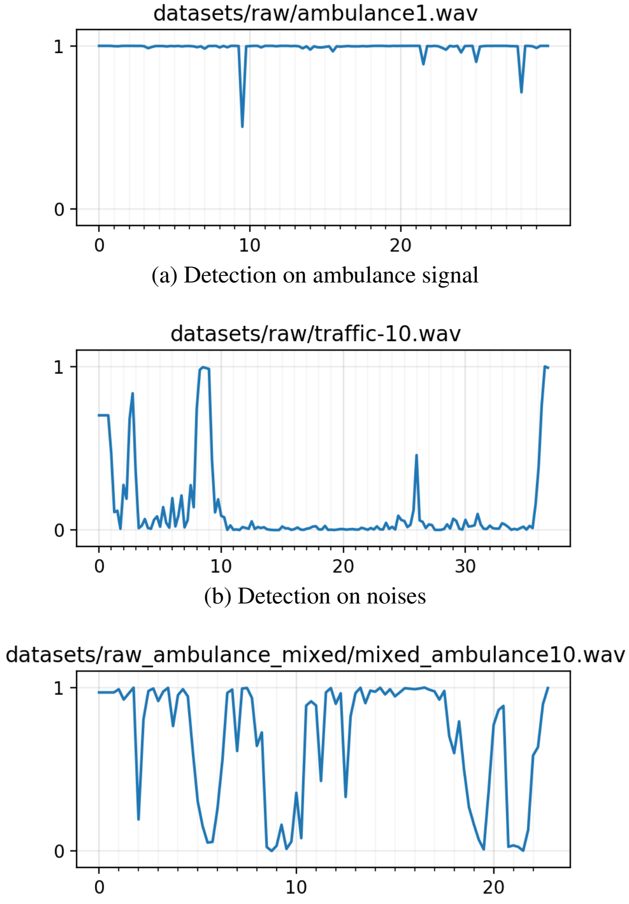
(b) Classifier

Figure 1: Training Testing

On only ambulance siren data, the detection is almost consistent except for a few time steps (Figure 2 a). This applies for the detection accuracy on only noises, which was represented by traffic sounds (Figure 2 b). The misclassification was expected to have occurred due to the presence of an actual ambulance siren in the background or a sound signal resembling that of a siren. On mixed data samples, the detection system performed relatively well (Figure 2 c). The audio file shown in the figure begins with an ambulance siren with traffic noises and the siren persists with additional loud honking noises. The “dips” in the explained by the loud honking noises interfering with the detection of the consistent ambulance siren signal.

Localization

The recording configuration contains three microphones **1**,**2**, and **3**, circumcenter **C**, ambulance **A**, distance d_c, and



(a) Detection on ambulance signal

(b) Detection on noises

(c) Detection on mixed signal

Figure 2: Testing results

the direction Θ_c (Figure 3). Our objective is to calculate d_c and Θ_c from the ambulance signals captured at the microphones at the vertices.

Notation

- R_{12}: distance between **1** and **2**
- R_{23}: distance between **2** and **3**
- R_{31}: distance between **3** and **1**
- r: radius of the circumcircle
- d_1: distance between **1** and **A**
- d_2: distance between **2** and **A**
- d_3: distance between **3** and **A**
- d_c: distance between **C** and **A**
- Δd_{12}, Δd_{23}, Δd_{31}: d_1 - d_2, d_2 - d_3, d_3 - d_1
- Θ_{12}, Θ_{23}, Θ_{31}: ∠ 1A2, ∠ 2A3, ∠ 3A1
- Θ_{C1}, Θ_{C2}: ∠ CA1, ∠ CA2
- Θ_c: direction from C to A

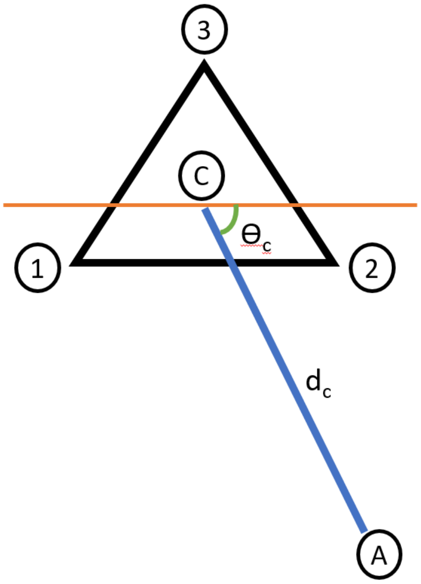


Figure 3: Diagram of the Experiment

Calculating d_1, d_2, and d_3

To get d_c, we calculate d_1, d_2, d_3. Then, we express $d_1 = d_2 + \Delta d_{12}$ and $d_3 = d_2 - \Delta d_{31}$ to reduce the number of unknown variables. Then, we focus on three triangles $\triangle A13$, $\triangle A23$, and $\triangle A12$. Using the fact that Θ_{12} = Θ_{23} + Θ_{31}, we derive the cosine of that angle.

$$\begin{aligned} cos(\Theta_{12}) &= cos(\Theta_{23} + \Theta_{31}) \\ &= cos(\Theta_{23}) * cos(\Theta_{31}) - sin(\Theta_{23}) * sin(\Theta_{31}) \end{aligned} \tag{1}$$

Since sine can be expressed in cosine, formula (1) can be expressed with only using the edges of the three triangles. Cosine rule is used here for each triangle. Hence we have one formula with one unknown variable d_2. We calculate d_2 using the quadratic formula. From d_2, we get d_1 and d_3.

Calculating d_c and Θ_c

We focus on three triangles $\triangle CA1$, $\triangle CA2$, and $\triangle A12$. Using the fact that Θ_{12} = Θ_{C1} + Θ_{C2}, we derive the cosine of that angle. Note that r can be calculate using the sine rule of the $\triangle 123$. Similar to the above formula (1), we have one formula and one unknown variable d_0. Hence we use the quadratic formula to calculate d_c from the following formula:

$$\begin{aligned} &\left[R_{12}^2\right] d_0^4 + \left[(-d_1^2 - d_2^2 - 2r^2 + R_{12}^2)R_{12}^2\right] d_0^2 + \\ &(R_{12}^2 - 2r^2)d_1^2 d_2^2 + (r^2 - d_1^2 - d_2^2)R_{12}^2 r^2 + (d_1^4 + d_2^4)r^2 = 0 \end{aligned} \tag{2}$$

Then we calculate Θ_c by focusing on $\triangle CA1$ and $\triangle C12$. We have all the edges, thus we use the cosine rule to calculate the angles in these two triangles. From them, we get Θ_c.

Experiment

Given the formula to calculate d_c and Θ_c from recordings of signals from three different microphones, we simulated the localization step of the ambulance siren signal. We used 44.1kHz sampling rate for the experiment. Thus, the time step for the recording is $\frac{1}{44100}$s. Sound travels 343m/s, which means for each time step for the recording, sound travels $\frac{343}{44100}$m. We used 441Hz beep sound as the replacement of the ambulance sound, which means the period is $\frac{1}{441}$s. Since the recording is using 44.1kHz sampling rate, there are 100 time steps in one period computed from period / time step.

From the above information, The sound is deduced to travel at approximately 0.85m in one period computed from 100 time steps * speed of sound in one time step. Hence, the location of the three microphones needs to be carefully chosen: R_{12}, R_{23}, and R_{31} should not exceed 0.78m. In the experiment, we used an equilateral triangle with R_{12}, R_{23}, and R_{31} equal to 0.75. In this setting, the distance differences Δd_{12}, Δd_{23}, and Δd_{31} is in one cycle.

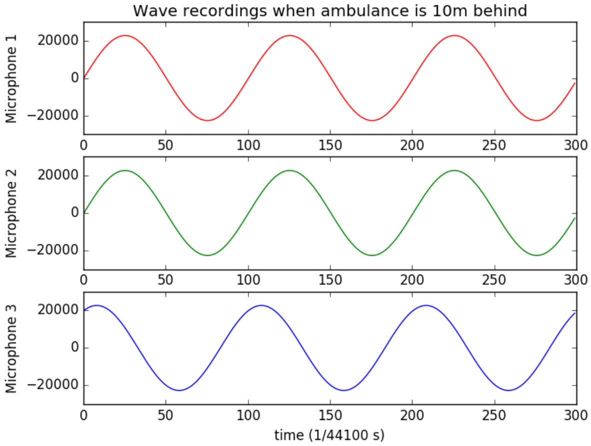


Figure 4: Three microphone recordings

The ambulance was assumed to be far behind the car. Then, $\Delta d_{12} \approx 0$ and $\Delta d_{31} \approx 0.75 * \frac{\sqrt{3}}{2}$. This Δd_{31} was approximated by 83 * timestep * speed of sound difference of the two input waves in microphone 3 and microphone 1. In the experiment, we created various shifted versions of the input signal, and fed them to the model we created above to get the distance and the direction from the ambulance to the circumcenter. The three input waves are depicted in Figure 4. In this setting the model found the position of the ambulance from distance of 10m and the direction of 90 degrees.

Conclusion

We have presented a method based on non-negative matrix factorization for ambulance siren detection that trains a model by retrieving a general basis vector of the ambulance signal. Upon detection, localization is performed through the triangular placement of microphones.

Physical experiments were conducted and exposed limitations in the realistic domain. Each experiment instance was initiated with an initial beep sound in the center of the triangle. The objective was to coordinate the beginning of the recordings of the three microphones due to the lack of a device that could record simultaneously from three sources. Theoretically sound, however practically it posed difficult to place the beep sound directly in the center. When the beginning of each recording (from each vertex) was shifted to the unequal detection of the beep, the siren signals were shifted equivalently which distorted their corresponding time of siren signal arrivals. For testing purposes, we thus ran the simulated experiments.

Our experiments on simulated data show that our approach performs as expected. It is important to note, however, that the training data is not comprehensive and further simulations with higher sampling rate is needed. For 44.1kHz, the signals travel fast making it hard to get diverse Δd distances. Also, additional physical experiments need to be conducted. The reason being real-time siren detection and localization will require more analysis and filtering. We believe that more precise equipment and averaging over repeated instances will make a physical detection and localization system possible. However, we defer this for future work.

References

Germain, F. G.; Sun, D. L.; and Mysore, G. J. 2013. Speaker and noise independent voice activity detection. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 732–736.

Godara, L. C. 1997. Application of antenna arrays to mobile communications, part II: Beam-forming and direction-of-arrival considerations. *Proceedings of the IEEE* 85(8):1195–1245.

Guizhongz, L.; Engineering, E.; and Engineering, I. 2002. T C. *Energy* 2(1):1124–1127.

Hioka, Y., and Hamada, N. 2004. DOA Estimation of Speech Signal Using Microphones Located at Vertices of Equilateral Triangle. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* E87-A(3):559–566.

Karray, L., and Martin, A. 2003. Towards improving speech detection robustness for speech recognition in adverse conditions. *Speech Communication* 40(3):261–276.

Sohn, J. 1999. A statistical model-based voice activity detection. *IEEE Signal Processing Letters* 6(1):1–3.

Su, G.; Su, G.; and Morf, M. 1983. The Signal Subspace Approach for Multiple Wide-Band Emitter Location. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 31(6):1502–1522.

Wang, H. 1985. . Coherent Signal-Subspace Processing for the Detection Wide-Band Sources. 823–831.