Over-The-Air TV Detection using Mobile Devices

Mohamed Ibrahim*, Marco Gruteser*, Khaled A. Harras† and Moustafa Youssef‡
*WINLAB, Rutgers University
†Carnegie Mellon University
‡Egypt-Japan University of Science and Technology
Email: {mibrahim, gruteser}@winlab.rutgers.edu, kharras@cs.cmu.edu, moustafa.youssef@ejust.edu.eg

***Abstract*—We introduce a mobile sensing technique to detect a nearby active television, the channel it is tuned to, and whether it is receiving this channel over the air or not. This technique can find applications in tracking TV viewership, second screen services and advertising, as well as improving the efficiency of TV white space spectrum usage. The technique uses a three-stage detection process: It first uses a Gaussian mixture model on audio recordings from mobile phones to detect likely TV sounds in the area. It then correlates the recording with known TV channel audio to identify the channel and improve detection robustness. Finally, it applies a latency analysis to determine whether programming is received over-the-air or through alternate means such as cable or satellite TV. Our system is evaluated using diverse datasets that take into account different realistic scenarios of indoor environments for several users. The results show that the system can achieve an area under the curve (AUC) of 0.9979 and a false negative rate of 0.0132.**

I. Introduction

Understanding television (TV) viewership is of interest for advertising purposes and for supporting second screen services that display supplementary information about TV programs. It can also support TV white space networking, however. White space networks use licensed spectrum bands on a secondary basis, when they are not in use by the primary user that they have been licensed to. This technology is attractive because the average utilization of licensed spectrum bands can be quite low—it varies between 15–85% according to recent studies by the Federal Communications Commission (FCC) [11]. In the United States, white space networking is currently permitted in broadcast television spectrum [12].

Since unlicensed devices can only use the spectrum when it is not in use by licensed TV broadcasters, the system relies on a database lookup mechanisms to allow an unlicensed transmitter to determine whether the broadcast spectrum is available. The database indicates spectrum as available in locations where the received TV signal is less than -114 dbm, a value that is far below the sensitivity of typical TV receivers. The system therefore adopts a conservative approach and allows secondary use only where no active TV receiver can exist, which is often only in rural areas. A more aggressive approach would also allow secondary use in areas where no TV receivers are in use (a receivable signal is present, but no over-the-air TV is used). Recent work [30] experimentally shows that this would increase the total achievable rate for 4 watt secondary devices by a factor of 6 compared to current TV white space systems. Other simulations [24] show that 24 additional 6 MHz channels can be provided.

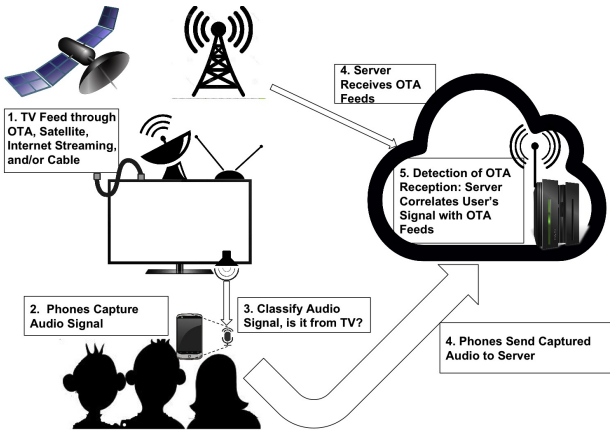


Fig. 1. Operation scenario of the proposed system.

More aggressive use of white space would require knowing the locations of TV receivers that operate over-the-air, however, not just the location of TV transmitters. Smart TVs could automatically report their position over an Internet connection [30]. This would ignore legacy TVs, however, which currently represent 71% of all U.S. TV homes [2]. Our earlier work-in-progress paper [15] also proposed using mobile audio sensing to detect TVs. This work, however, designed to classify limited number of audio sources (TVs, laptops and people talking) and used a single user data set in a controlled environment. Recent studies show that only 7% of US households receive TV content over the air (OTA) compared to 83% receiving TV through cable, satellite, or fiber connections [7]. Therefore just detecting TVs [15] does not provide a good estimate of where over the air TVs are used.

In this paper, we propose a passive TV detection system that leverages audio sensing on mobile phones. The system processes audio samples from phones or other mobile devices to detect if an active TV is in the vicinity, determine the channel it shows, and ascertain if it receives the programming over the air channel, cable, or satellite. We propose a TV detection system that first classifies the acoustic signals to determine whether they contain likely TV sounds. This filtering step is useful to reduce processing requirements and can also be used to reduce phone energy consumption. Once likely TV sounds have been detected, it applies standards correlation techniques [4], [18], [26], [5], [6], [3] to match the audio to possible channels.

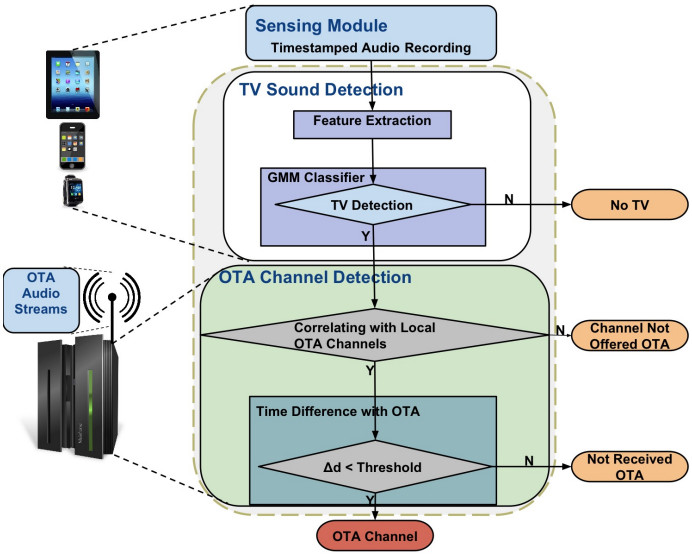


Fig. 2. *System overview of OTA TV detection.*

Finally, our system is able to determine the source of the currently watched channel, i.e. whether it is received over-the-air, cable, or satellite dish, by exploiting delay differences to a reference signal. Usage scenario is illustrated in Figure 1.

We evaluate the system using a diverse dataset from phones in six home environments. The results show zero false negatives in the dataset that belong to single users (easy cases) and a maximum 0.014 false negative rate among all datasets. Once a likely TV signal has been detected, the results show no errors in determining OTA channels.

The remainder of the paper is organized as follows. Section II, presents an overview of the system. We then present the proposed technique for detecting likely TV sounds in Section III. Section IV describes the over-the-air channel detection module. Section V details the data collection and presents the evaluation of our system. In Section VI, we discuss the main issues and challenges for our system. Section VII discusses the related work. We conclude and give directions for future work in Section VIII.

II. System Overview

In this section, we discuss the design of the TV detection system and provide an overview of how the system detects OTA TVs using only audio recordings from mobile phones.

Figure 2 shows the structure of the system. Our system consists of sensing module, TV sound detection module and OTA channel detection module. The sensing module collects timestamped audio recordings from mobile devices. The TV sound detection module extracts the features from the audio recordings and classifies the extracted feature vectors, determining whether they have been transmitted from TV or not, using Gaussian mixture model classifier. Finally, the OTA channel detection module decides whether the channel is offered OTA or not. If the channel is offered OTA, then this module checks if the channel received OTA or not by estimating the delay between the user signal and the OTA feed. The sensing and likely TV sound detection modules are designed to run on mobile devices, while the more computationally complex OTA channel detection is envisioned for the cloud. In this paper, we use the maximum achievable sampling rate in our current smart phones which is 44.1 kHz to maximize accuracy. Depending on application needs, this can occur periodically or be specifically triggered at a time of interest. For example, audio sensing could be triggered only when the phone is indoor and static using existing techniques that rely on light sensors, magnetometers, cell tower signals and accelerometer [31], [28], [14]. This would also reduce the power consumption compared to periodic triggering.

A. TV Sound Detection

TV sound detection seeks to identify audio samples that contain sound from a TV. This could potentially be achieved directly by continuously correlating these audio samples with OTA channels. However, this would impose several issues. First, the system needs to correlate with continuously received audio feeds of all OTA channels. This would most likely have to take place on a remote server. Therefore, mobile devices would need to upload all audio recordings which results in increased data and power consumption as well as privacy concerns. The TV sound detection module, therefore, first classifies and filters the audio samples on the mobile device. It determines which audio samples contain likely TV sounds and forwards only those to the external server for further processing. This is based on extracting Mel Frequency Cepstral Coefficients (MFCCs) and classifying them with a Gaussian mixture model. The classification seeks to distinguish samples that contain TV sound from many other everyday sounds and noise, like laptop audio, radio, conversations, etc. We assume that such a classifier can be trained offline over a sufficiently large dataset of sounds and then be installed on mobile devices.

B. OTA Channel Detection Module

The over-the-air channel detection module confirms that the audio samples include TV sounds and seeks to distinguish cable or satellite TV audio from over-the-air TV audio. We leverage existing real-time TV channel detection techniques [4], to confirm whether the audio sample matches audio from any TV channel. We can further limit the matching process to only local OTA channels, based on the user location. This requires that the server has access to real-time audio feeds from these channels. One implementation option is to have the server monitor all channels in an area with TV antennas and a set of TV tuners. If the result of the correlation is negative, then the system announces that the sample does not match local over-the-air channels.

If there is a match, however, it is still unclear whether the audio sample actually contains audio from a TV that receives over-the-air because several channels are broadcast both over-the-air and distributed over cable and satellite TV. The existing channel detection techniques therefore do not suffice for potential white space applications.

The system therefore determines whether the audio sample content was received OTA by analyzing the delay between the audio recording and the same channel received through OTA antenna. It exploits that satellite distribution and processing for cable distribution of the over-the-air channels usually adds significant delay. Delay can be computed as part of the correlation process. Only if the delay is consistent with over-the-air broadcasting, the system announces the detection of an over-the-air TV.

III. TV Sound Detection

In this section, we describe the feature extraction and classification algorithm underlying the likely TV sound detection process.

A. Feature Extraction

Feature extraction is an important step that affects both accuracy and computational complexity of the classifier/predictor. One heuristic is that TVs are usually larger in size and equipped with more capable speakers that reproduce high definition sounds compared to other audio devices like laptops or radios. As a result, one can expected to find more frequency components in audio signals coming from TVs than most other devices. Also, TVs are often tuned to louder volumes compared to laptops/tablets as these are more personal devices than TVs. More importantly, TV programming contains unique mixes of conversations and other sounds that differ from most natural conversations.

These kind of differences can be captured effectively using Mel Frequency Cepstral Coefficients (MFCCs). MFCCs are widely used features in the field of speech recognition and speaker identification. MFCCs give a representation of the short term power spectrum of the audio signal in the nonlinear mel scale. In other words, MFCCs are extracted by taking the discrete cosine transform of the logspectral energies of the audio signal. A more detailed discussion about the MFCC can be found in [8].

Using different microphones between training and testing may degrade the performance of the system. Therefore, a mean normalization step was proposed to overcome this problem of microphone heterogeneity [25]. Also, taking into account the correlation between the subsequent frames of a certain audio recording is an important aspect for extracting audio features. This temporal information can be captured by taking the first and second derivatives of the cepstral coefficients [25]. By applying these normalization techniques, we are able to align the testing data and the training data in the same feature space and as a result the same classifier can work across different phones.

B. Classification Algorithm

A major challenge for modeling a given data is choosing the family of the model (i.e. Gaussian, exponential or other distributions). Also, for a predictor like the Maximum A posteriori Probability (MAP), we need to determine the prior probabilities for the audio classes (e.g. the probability of the TV class). One of the solutions is to use recent statistics as an initial guess of these priors. The distribution of these priors can be updated independently for each user based on TV viewing/listening behavior at home. For example, we can use the classified samples with a probability of error less than a certain threshold to modify the distribution of the priors dynamically.

A probabilistic model is used in this paper as it belongs to the family of models that is able to provide an accurate confidence value for each classified sample. Therefore, we can use this confidence to reject the current result or request more data to be able to improve classification. For example, for detecting white spaces, there are more concerns about false negatives as it is more important to protect primary users, i.e. TV users. Therefore, we can conservatively classify results with low confidence values as TV sets to protect primary users.

A Gaussian Mixture Model (GMM) is a strong tool for approximating arbitrary shapes, therefore, it has been used widely in speaker identification. The acoustic space of a speaker can be characterized as a set of acoustic classes (vowels, nasals, fricatives) and each of these classes can be approximated using a Gaussian distribution. Therefore, we use a GMM to model the audio signature of the TVs as a speaker and model the non-TV audio sources in another GMM model. Figure 3 shows that the audio fingerprint of the non-TV classes cannot be modeled using a unimodal probabilistic model like maximum likelihood Gaussian model. This further motivates using GMM to model our classes.

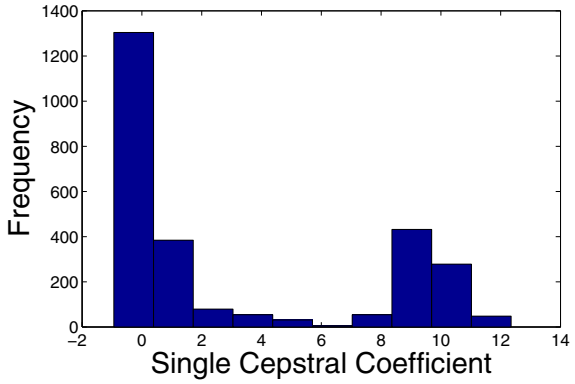
Given a set of feature vectors $X = \{x_1, x_2, ..., x_T\}$, we model the likelihood of a vector of features belonging to a certain speaker as a mixture of Gaussian distributions:

$$p(x_t|\lambda) = \sum_{i=1}^{M} w_i p_i(x_t)$$

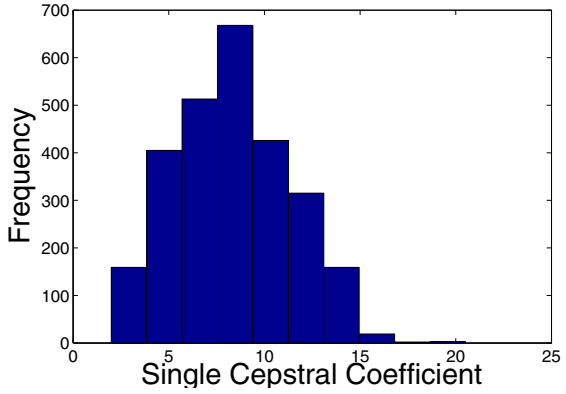
$$p_i(x_t) = \frac{1}{(2\pi)^{D/2}\left|\Sigma\right|^{1/2}} exp\left\{-\frac{1}{2}(x_t - \mu_i)'(\Sigma)^{-1}(x_t - \mu_i)\right\}$$

M is the number of mixtures, w_i is the weight of each mixture, $p_i(x_t)$ is the probability (Gaussian distribution) for such feature vector x_t, and λ is the class to be modeled. D is the length of the feature vector, μ_i is the mean vector and Σ_i is the covariance matrix.

In order to estimate the parameters of the model, we use maximum likelihood estimation, i.e. choose the set of parameters that maximizes the likelihood of the data given this set of parameters. This objective is, unfortunately, a nonlinear function of the parameters. This means we cannot just differentiate with respect to these parameters and obtain the optimal parameters. Instead, we use an iterative technique as Expectation Maximization (EM) [10]. In this technique, we start with an initial guess of the model parameters, then update these parameters while ensuring that $P(X|\lambda^{k+1}) \geq P(X|\lambda^k)$. After modeling each of the two classes as GMM, we classify a new feature vector by calculating the likelihood ratio test. A feature vector is classified as a TV if the following holds.



(a) Bimodal histogram.



(b) Unimodal histogram.

Fig. 3. Histogram of two sample 30s recordings from non-TV audio sources for single cepstral coefficient, showing that unimodel distributions (e.g., maximum likelihood Gaussian) are unsuitable

$$\frac{p(x_t|\lambda_{TV})}{p(x_t|\lambda_{Non-TV})} > 1 \tag{1}$$

IV. Channel Detection

In order to detect the currently watched channel, given the audio recording from the mobile phone, we correlate the signal with OTA channels on our server. This server is connected to a TV audio feed from all over-the-air channels. In our implementation, the server is connected to TV tuners which are connected to antennas. Our system returns the most correlated channel from the OTA channels as the estimated channel. As in the work [4], we use the cross correlation as a measure of similarity between the two audio signals. This is also a standard way of finding the similarity between signals. Cross correlation between two signals x and y is defined as follows:

$$R_{xy}(m) = E(x_{n+m}y_n^*) \tag{2}$$

Where y^* denotes the complex conjugate of y and m is the lag.

Figure 4 shows how the correlation peaks for the currently watched channel compared to another channel.

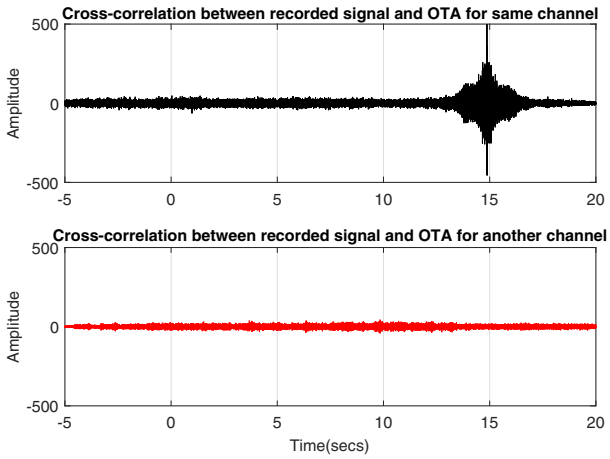


Fig. 4. Example for correlation between mobile recording and two OTA channels.

The time lag between two signals can be derived by finding the lag that maximizes the cross correlation:

$$T_{xy} = \frac{argmax_m(|R_{xy}(m))|}{sr} \tag{3}$$

Given that sr is the sampling rate.

After detecting the channel, we estimate the time lag between the recorded audio and the same channel from the audio feed. Since the delay difference in our measurements were large, threshold detection suffices. This threshold can be tuned based the lowest delay between OTA channels received via antenna with other sources including different satellite, cables and Internet streaming.

The system limits the complexity of finding the most correlated channel as follows (i.e. it reduces the search space for finding the most correlated channel). It builds on earlier work [4] that sorts the list of channels to be correlated according to the statistics of viewing this channel. These statistics can be fine grained by each user in order to be more accurate by continuously profiling the statistics for each user. In this paper, we add to the sorting technique another technique that limits the search space to the local channels based on the location of the user. Moreover, we limit the search space to only OTA channels as these are only the channels that we are interested in for the white space application.

V. Evaluation

We evaluate in this section our OTA TV detection system in several typical indoor TV environments.

A. Experimental Methodology

Our dataset is collected by 6 volunteers with different cell phones who take audio samples near TVs and in a variety of other settings. In order to evaluate our system effectively and realistically, we collected a dataset that includes many of the common scenarios in our daily lives in indoor environment. Therefore, we took all the possible sources of audio in indoor

Datasets	Num. of TV recordings	Num. of Non-TV recordings
1	855	71
2	10	20
3	135	59
4	453	568
5	607	193
6 (Challenging)	727	395

TABLE I
DATASET DETAILS FOR TV DETECTION MODULE.

Channel	Cable type	Number of recordings	Time difference (sec)
NBC	fiber optics cable TV	20	5.6
	satellite TV	20	12.07
CBS	satellite TV	20	8.04
NJTV	satellite TV	40	13.7
TXTV	satellite TV	40	11.3

TABLE II
DATASET DETAILS AND RESULTS FOR THE TIME DIFFERENCE MODULE.

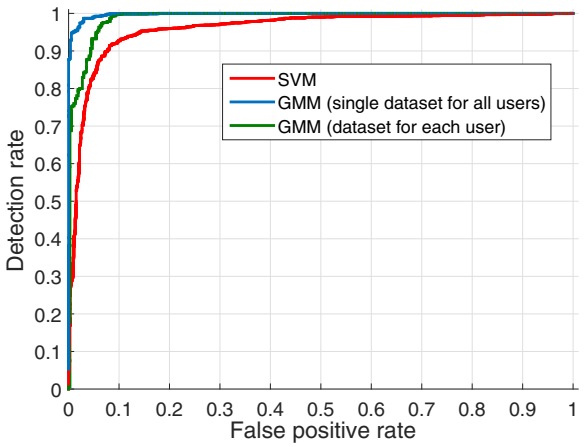
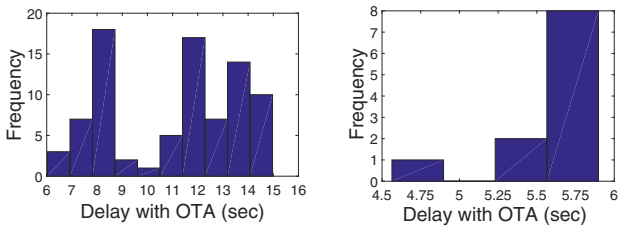


Fig. 5. Comparing our TV detection approach (GMM) to related work [15] (SVM).



(a) satellite TV delay. (b) fiber optics cable TV delay.

Fig. 6. Histogram of the delay for the collected samples of satellite TV and fiber cable TV with OTA content received via antenna.

environment into account like radios, tablets phones, laptops, microwave, washing machines, alarms. To test the limits, we also collected a special dataset with particularly challenging sound samples. This includes samples taken while a variety of loud background sounds occurred, including laptop music or crying babies. It also includes samples with very soft TV sound, where the TV is nearly muted. Our dataset for the TV detection module is summarized in Table I. Each audio recording is 30 seconds long.

We consider two training scenarios for evaluating our system: a centralized scenario and a distributed scenario. The centralized scenario uses the collected data from all phones as a single combined dataset for training and evaluation and we test the trained model using cross validation. The distributed scenario uses each phone's dataset separately to train and test. Here, we use 28% of each dataset for testing and the remainder for training. The intuition behind the distributed scenario is that the recording environment across phone users is likely to differ. Therefore, we ask to what extent training the classifier spearately for each phone improves the results. We used the MSR Identity Toolbox [23] to evaluate our proposed technique, which is a speaker recognition MATLAB toolbox using GMM.

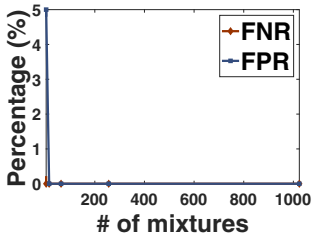
We compare our approach to an SVM classifier baseline as this is the only related work [15]. It uses MFCC along with other classic audio features like zero crossing rate, short time energy, spectral centroid and spectral spread.

To evaluate our system, we use the following metrics: False negative rate (FNR) which is the percentage of the TV samples that were classified as non-TV class. False positive rate (FPR) which is the percentage of the non-TV samples that were classified as TV class. For the Cognitive Radio application, false negatives (FNs) are more critical than false positives (FPs), as in the case of FN, a secondary user may interfere with a primary user (TV) if the TV is not correctly detected. False positives, however, are likely to be eliminated in subsequent processing steps.

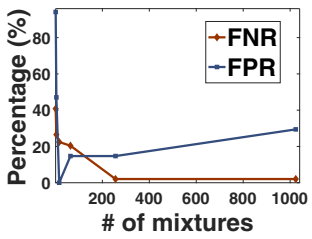
B. Overall System Evaluation

We present in this section the evaluation for the whole system focusing on the two novel modules: TV detection and channel detection module.

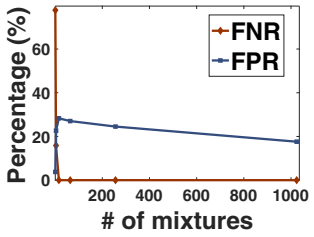
For the TV detection module, we compare our approach to the SVM classifier as related work [15]. Figure 5 shows that our proposed GMM classifier, in both centralized and distributed scenarios, is outperforming the SVM classifier. Our centralized GMM classifier achieves 0.9979 area under the curve (AUC) and 0.9847 AUC for the distributed GMM compared to 0.9597 AUC for SVM classifier. These results follow our intuition that GMM is better capable of modeling arbitrary shapes/density as long as enough training samples are available that capture most of the sound classes in the



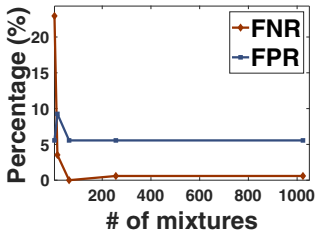
(a) User 1's dataset for training and testing.



(b) User 3's dataset for training and User 2's dataset for testing.

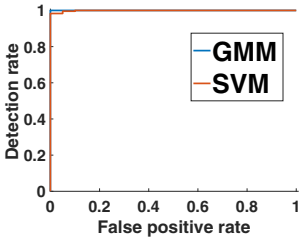


(c) User 4's dataset for training and testing.

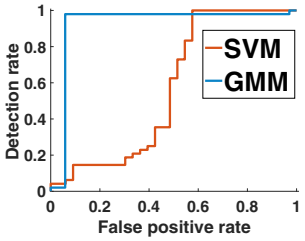


(d) User 5's dataset for training and testing.

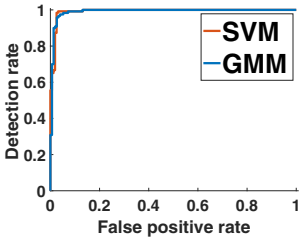
Fig. 7. Effect of varying the number of mixtures on the detection performance.



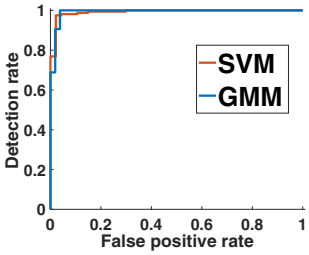
(a) User 1's dataset for training and testing.



(b) User 3's dataset for training and User 2's dataset for testing



(c) User 4's dataset for training and testing.



(d) User 5's dataset for training and testing.

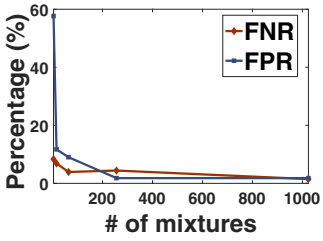
Fig. 8. Comparison between our technique (GMM) and related work [15] (SVM) using ROC curves.

real environment. Therefore, in our setting the centralized version of the GMM is able to achieve better results than the distributed one as it has more samples to best capture the real density. However, if enough data is collected for each user, that captures all the sound classes exist in her home, then the distributed configuration could outperform the centralized one.

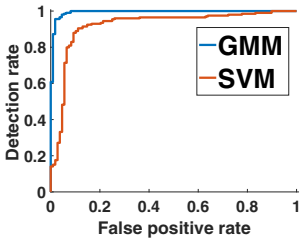
For the time difference module, we show using our experiments that the latency difference between the OTA service and other TV services is large enough to detect the OTA service. More specifically, the latency difference between the OTA service and the other two services that we examined (satellite and fiber optic cable) is at least 4 seconds. Figure 6 shows the histograms for the latency difference between the two cable services and the OTA service. As we see in the figure, there is a more notable delay for satellite TV. However, there is also a clear latency difference of at least 4 seconds for the fiber optic cable service. These large differences mean that we can distinguish over-the-air TVs from cable and satellite in this case study.

C. Finding Optimal Mixtures' Number

In this section, we explore the space of the number of mixtures and see how this affects the detection performance in terms of FPs and FNs. Figures 7 and 9(a) show that the optimal number of mixtures is different for each dataset, therefore we need to use different number of mixtures for each dataset. Different mixture number is needed for each user, as each user has his own viewing behavior and consequently has different density than the other users. One way of quantizing the number of mixtures, is to derive optimal value for each class of users like singles and family.



(a) Number of mixtures' effect on TV detection performance.

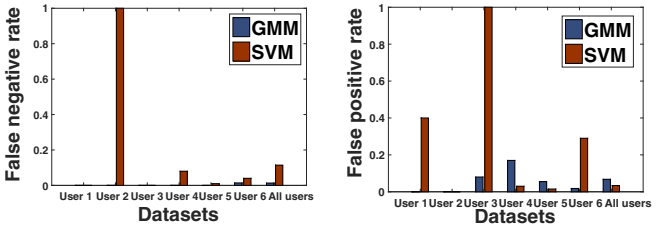


(b) ROC curves compared to related work [15].

Fig. 9. Evaluation of TV sound detection using challenging dataset (User 6).

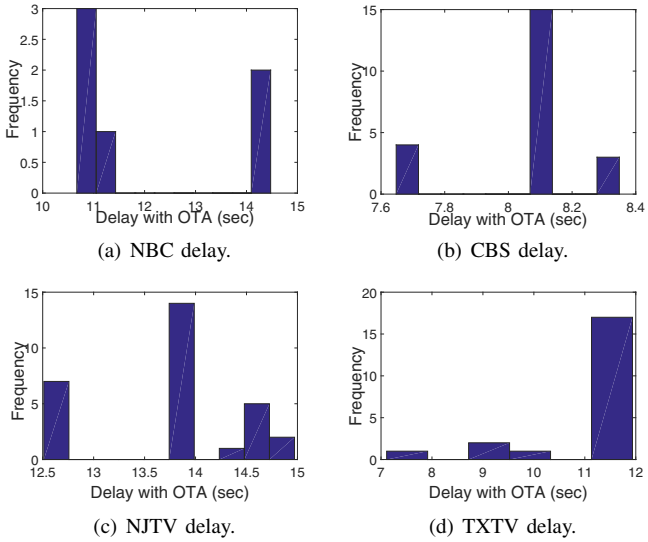
D. Evaluating Distributed TV Detection

Table III and Figure 10 summarize our results for the TV sound detection module. Also, Figure 8 and Figure 9(b) show the ROC curves for all the datasets. Using GMM, our system outperforms the SVM and proves that GMM is capable of capturing the TV and Non-TV audio fingerprints effectively. Moreover, the results show that using GMM we achieved 0 false negatives in the datasets collected in single occupant homes. These datasets appear to be easy for both techniques, likely the datasets are more homogeneous and contain less noise. However, for phones collecting data in a family setting, the results show that they are more challenging to classify. In particular, dataset 6 which is was collected in the most challenging setting shows higher errors. Even in this case, however, the false negative rate remains near zero. Also, note that the results show high false positives in the fourth dataset because the testing part included classes of sounds that did not occur in the training data (for example, the sound of ringing phone). Performance could be further improved in two ways.



(a) FN rate. (b) FP rate.

Fig. 10. Comparing our TV detection approach (GMM) to SVM [15] in terms of FNs and FPs.



(a) NBC delay. (b) CBS delay.

(c) NJTV delay. (d) TXTV delay.

Fig. 11. Histograms for each channel received through satellite TV.

First, we could augment this new data set with an universal model that includes a wide range of sound classes. Second, we could leverage the advantage of using probabilistic techniques and exploit the confidence of each detection. If the confidence is low, we could delay making a decision until more data is available. Our results shows that for the misclassified cases, confidence values are indeed low (i.e., almost 1 likelihood ratio).

E. Evaluating the Time Difference Module

Table II summarizes the dataset and time difference results for the channel detection module. Figure 11 shows a latency difference histogram for each channel with respect to OTA content received via antenna. The satellite TV shows an average of 11.05 seconds delay compared to signal received over the antenna. Even for fiber optics cable TV, the delay is higher than the signal received through antennas by an average of 5.6 seconds. While the results may differ across locations based on the details of the TV distribution system, these results show significant differences in delay across providers. Moreover, satellite and cable TV broadcast centers usually receive broadcast channel programming signals through antennas and then retransmit the signals through their networks [1], [16]. Therefore, they add propagation and processing delay compared to direct OTA reception. These delays can be due to longer distances, particularly for round trip propagation to satellites and due to amplifiers and filter components in the distribution network [16].

VI. Discussion

In this section, we discuss the limitations of this technique. First, the current technique cannot detect TVs that are muted. If it were necessary to detect such TVs as well, the system would need to be complemented with other sensing techniques such as smart phones' remote controls, infrared sensors, cameras, or smart TVs; which directly report their position and OTA reception.

Second, one may choose to use channel detection directly without detecting the source of the audio. As alluded to earlier, there are several disadvantages to such approach: A user may watch a TV channel on a laptop or tablet using a streaming service. Correlation would give us no information on whether this channel is being watched through TV or not. Furthermore, continuously uploading the signal from a mobile device would create both privacy and energy consumption issues. Therefore, we apply TV sound detection to filter audio samples and use the channel detection part only when the user watches TV. Third, using both parts gives our decision higher confidence than using only the channel detection module.

Third, the system can raise incentive questions and privacy concerns since it might capture sensitive information as part of audio samples. These can be alleviated by minimizing the audio samples that are transferred to an external server and more importantly, by ensuring that such samples are never stored. They could also be reduced, if future work would allow the complete OTA channel detection to run on the mobile device. In this case, the recorded information never needs to leave the device. Moreover, these privacy concerns might be further minimized by employing privacy-preserving signal processing techniques involving homomorphic encryption, locality sensitive hashing (LSH), or secure binary embeddings (SBE) [17]. The homomorphic encryption approach uses public-key cryptography which is a computational demanding approach compared to the other approaches. On the other hand, the other two approaches (LSH and SBE) trades correlation accuracy for computational complexity. Incentives for users could be simply usage of white spaces and more network capacity that can become available through this system. Users that contribute more data could also be rewarded with priority access to the white space database or, perhaps to some extent, the white space channels themselves.

Fourth, the delay estimation requires time synchronization between the mobile phone and the servers. Fortunately, the delay differences we observed were on the order of seconds, which can be detected with a level of time synchronization that can be easily achieved by current cellular network or Internet time synchronization protocols.

Datasets	Technique	FN rate	FP rate
1	SVM [15]	0	0.4
	GMM	0	0
2	SVM [15]	1	0
	GMM	0	0
3	SVM [15]	0	1
	GMM	0	0.08
4	SVM [15]	0.08	0.03
	GMM	0	0.17
5	SVM [15]	0.01	0.15
	GMM	0	0.055
6 (Challenging)	SVM [15]	0.04	0.29
	GMM	0.014	0.018
All	SVM [15]	0.1144	0.0678
	GMM	0.0132	0.0337

TABLE III
RESULTS SUMMARY FOR THE TV SOUND DETECTION MODULE.

VII. RELATED WORK

Detecting TV sets has been proposed before in [27] which relied on special devices for sensing. This work required the usage of special sensor in the vicinity of the TV set to detect the power leakage of a receiver's local oscillator as a way to detect TV set. These systems are not easy to deploy and so they don't scale.

On the other hand, several approaches were proposed for detecting TV shows, commercials, music and channel identification [4], [18], [26], [5], [6], [3]. Also, this work includes scene boundary detection [19] and TV shows recognition [13]. However, all this work focus on detecting the contents regardless of its source and hence cannot identify the source of the heard audio.

Our preliminary work [15] proposed passive TV detection technique using mobile phones to detect TV. However, in this work-in-progress paper, we evaluated the proposed system using limited dataset in a controlled environment. Also, we showed on this limited dataset, that the three classes for audio detection are linearly separable and used SVM based on this assumption. In this paper, we target probabilistic modeling for audio classes as we need confidence for each detection, while not limiting the audio classes to three classes. Moreover, we propose a novel latency analysis approach for determining whether programming is received over-the-air or through cable or satellite. Therefore, our system is complete and well evaluated compared to our short published paper.

Extensive work has been done in speaker identification [25]. Most of this work using Gaussian Mixture Models for speaker identification [21], [20], as it is able to approximate smoothly different shapes of audio models. Also, they proposed an adaptation mechanism [20] using maximum a posteriori estimation, to estimate a model for new speaker by adapting a universal model representing the space of different speakers. Another line of work focused on modeling the speakers using Hidden Markov Models [22], [29], but assuming text-dependent applications. Finally, recent work proposed the i-vector framework where the speaker models are estimated through a procedure called Eigenvoice adaptation [9]. In this work, they represent speakers with the identity vector (i-vector) which are bigger than the underlying cepstral feature vector but much smaller than the GMM supervector.

Work in [4] is the most similar to our channel detection module, but in our system we use the mobile phone location to limit the number of channels to correlate with. Moreover, in our system, we can find the time lag between the channels and so, being able to determine the source of the OTA channel.

VIII. CONCLUSION

In this paper, we proposed a passive over-the-air TV detection system that leverages audio sensing on mobile devices. Our system relies on two main components: TV sound detection and OTA channel detection. The former detects audio samples with likely TV sounds based on Gaussian Mixture modeling. If likely TV sounds are present, the samples are sent to a server where the OTA channel detection module correlates the audio signal with live OTA channels to determine the channel and audio delay. Finally, the system determines whether audio sample contains sounds from TV programming received over-the-air based on a delay analysis. In our experiments, the system achieved an area under the curve (AUC) of 0.9979 and false negative rate of 0.0132.

This technique could potentially be used as part of bootstrapping process or a maintenance process for an over-the-air TV database to enable more aggressive white space usage. It would likely have to be complemented with additional reporting methods to ensure completeness, however. The information may also be useful for simply gathering statistics of over-the-air TV distribution, to understand in which regions the TV spectrum could be used more aggressively with minimal

impact on viewers. Moreover, parts of this system might also find use in more traditional advertising applications that collect TV viewership information.

REFERENCES

[1] An Overview of Satellite TV: Learn more about satellite television and how it compares to cable television service. https://www.xfinity.com/satellite-tv.

[2] TV-CONNECTED DEVICES PAVE THE WAY FOR NEW WAYS TO WATCH CONTENT. http://www.nielsen.com/us/en/insights/news/2017/tv-connected-devices-pave-the-way-for-new-ways-to-watch-content.html.

[3] I. Bisio, A. Delfino, F. Lavagetto, and M. Marchese. A Television Channel Real-Time Detector using Smartphones. *IEEE Transactions on Mobile Computing*, 99(PrePrints):1, 2013.

[4] I. Bisio, A. Delfino, G. Luzzati, F. Lavagetto, M. Marchese, C. Fra, and M. Valla. Opportunistic estimation of television audience through smartphones. In *Performance Evaluation of Computer and Telecommunication Systems (SPECTS), 2012 International Symposium on*, pages 1–5, 2012.

[5] M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668 –696, april 2008.

[6] C.-Y. Chiu, D. Bountouridis, J.-C. Wang, and H.-M. Wang. Background music identification through content filtering and min-hash matching. In *ICASSP 2010*, pages 2414 –2417, 2010.

[7] Consumer Electronics Association. U.S. Household Television Usage. https://www.ce.org/News/News-Releases/Press-Releases/2013-Press-Releases/Only-Seven-Percent-of-TV-Households-Rely-on-Over-t.aspx, 2013.

[8] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, 1980.

[9] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):788–798, 2011.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[11] FCC Spectrum Policy Task Force. Report of the spectrum efficiency working group. Technical report, FCC, November 2002.

[12] Federal Register. Unlicensed operation in the TV broadcast bands, December 2010.

[13] M. Fink, M. Covell, and S. Baluja. Social-and interactive-television applications based on real-time ambient-audio identification. In *Proceedings of EuroITV*, pages 138–146. Citeseer, 2006.

[14] S. Hemminki, P. Nurmi, and S. Tarkoma. Accelerometer-based transportation mode detection on smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, page 13. ACM, 2013.

[15] M. Ibrahim, A. Saeed, K. Harras, and M. Youssef. Unconventional TV detection using mobile devices. In *International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, 2013. UBICOMM'13.*, 2013.

[16] V. Jones. An Introduction to Basic CATV. http://people.seas.harvard.edu/~jones/cscie129/nu_lectures/lecture13/CATV/CATV.html.

[17] M. Pathak, J. Portelo, B. Raj, and I. Trancoso. Privacy-preserving speaker authentication. In *International Conference on Information Security*, pages 1–22. Springer, 2012.

[18] M. Ramona and G. Peeters. Audio identification based on spectral modeling of bark-bands energy and synchronization through onset detection. In *ICASSP 2011*, pages 477 –480, may 2011.

[19] Z. Rasheed and M. Shah. Scene detection in Hollywood movies and TV shows. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II – 343–8 vol.2, 2003.

[20] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1):19–41, 2000.

[21] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on*, 3(1):72–83, 1995.

[22] R. C. Rose and D. B. Paul. A hidden markov model based keyword recognition system. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 129–132. IEEE, 1990.

[23] S. O. Sadjadi, M. Slaney, and L. Heck. Msr identity toolbox v1. 0: A matlab toolbox for speaker recognition research. *Speech and Language Processing Technical Committee Newsletter*, 2013.

[24] A. Saeed, M. Ibrahim, K. Harras, and M. Youssef. Towards dynamic real-time geo-location databases for tv white spaces. *IEE Network Magazine*, 2015.

[25] R. Togneri and D. Pullella. An overview of speaker identification: Accuracy and robustness issues. *Circuits and Systems Magazine, IEEE*, 11(2):23–61, 2011.

[26] A. L. Wang. An industrial-strength audio search algorithm. In *ISMIR 2003*, 2003.

[27] B. Wild and K. Ramchandran. Detecting primary receivers for cognitive radio applications. In *DySPAN 2005*, pages 124 –130, 2005.

[28] Z. Yan, V. Subbaraju, D. Chakraborty, A. Misra, and K. Aberer. Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach. In *Wearable Computers (ISWC), 2012 16th International Symposium on*, pages 17–24. Ieee, 2012.

[29] K. Yu, J. Mason, and J. Oglesby. Speaker recognition using hidden markov models, dynamic time warping and vector quantisation. *IEE Proceedings-Vision, Image and Signal Processing*, 142(5):313–318, 1995.

[30] X. Zhang and E. W. Knightly. Watch: Wifi in active tv channels. In *Mobihoc 2015*. ACM, 2015.

[31] P. Zhou, Y. Zheng, Z. Li, M. Li, and G. Shen. Iodetector: A generic service for indoor outdoor detection. In *Proceedings of the 10th acm conference on embedded network sensor systems*, pages 113–126. ACM, 2012.