Abstract—Machine learning has recently drawn a lot of attention for a variety of applications, thanks to its good performance in identification, recognition, and regression problems. One such important application is vehicle-to-vehicle (V2V) communication propagation channels research. In this article, the challenges and opportunities of machine learning-based data processing techniques for evaluation of V2V channel measurements are presented. This paper reviews some state-of-the-art applications including identification of channel line-of-sight situations, tracking of multipath components (MPCs), and MPC clustering. The data obtained with these methods form, inter alia, the basis for accurate channel models. Furthermore, some challenges of machine learning-based data processing for V2V channel research are discussed as basis for future studies.

Index Terms—Vehicle to vehicle communications, channel modeling, machine learning, identification, clustering, tracking.

I. INTRODUCTION

Improving safety and efficiency of vehicle traffic is one of the main challenges of modern life. One of the most promising approaches to achieve this is coordination between vehicles. This would allow introduction of some new features, including i) active safety applications, e.g., emergency braking and intelligent actuators for seat belts and airbags, ii) intelligent traffic control applications, e.g., hazardous location notification and congestion/construction sites notification. In order to realize all of these applications, efficient and reliable communications between the vehicles are necessary. For this reason, vehicle-to-vehicle (V2V) communications have become a major research topic over the past decade.

A number of different standards have been developed for V2V communications, which mainly include: i) common local area networks like IEEE 802.11a/b/g, which are based on the transport protocol stack (TCP) and Internet protocol (IP), ii) specific V2V network like IEEE 802.11p, and iii) general cellular networks based on 4G/5G communication systems.

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

This work is supported by National Key R&D Program of China under Grant 2018YFF0212103 and 2016YFE020900, in part by the National Natural Science Foundation of China under Grant 61922012, 61771037, 61725101, and U1834210, in part by State Key Laboratory of Rail Traffic Control and Safety under Grant RCS2018ZT014 and RCS2018ZZ007; The work of AFM was supported by the California Transportation Authority under the METRANS program, and by NSF grant ECCS-1731694; (Corresponding author: Ruisi He.)
Data-processing: process the MPCs and their parameters for further modeling, which includes multiple sub-procedures such as line-of-sight/non-line-of-sight (LOS/NLOS) identification, MPC clustering, MPC tracking, and cluster parameter modeling; ii) Modeling and evaluation: model the channel and create channel realizations from the model, for the purpose of system simulations. The accuracy of the proposed models can be assessed by comparing the channel characteristics obtained from the model with those from different measurements.

Obviously, more detailed MPC parameters can help to characterize the channels more accurately; meanwhile, more wideband channels (as those used for 5G) require high sampling rate, and measuring time-varying channels needs many measurements over time to capture the evolution of channels. For these reasons, V2V channel measurements collect very large amounts of data. However, this massive amount of data brings difficulties in processing. For example, it has been shown that the propagation channel shows significantly different characteristics depending on the existence of a LOS, which is usually identified by human inspection or a parameter thresholding. However, it is hard and/or inaccurate to use these approaches for massive data. Similarly, human inspection has also been used for clustering, yet is inapplicable for large sets of data. Considering these limitations, more intelligent automated data processing techniques are required to improve efficiency and accuracy.

Machine learning has made obvious progress in recent years. As early as 1959, A. Samuel defined machine learning as “field of study that gives computers the ability to learn without being explicitly programmed” [2]. By now, machine learning has become a powerful tool for many applications of data processing, due to its great advantage for classification, recognition, and prediction tasks. Motivated by this, machine learning-based data processing for channel modeling is viewed as a promising solution for the data processing step of channel modeling. In this paper, we investigate three key applications of machine learning-based data processing for V2V channel research: LOS/NLOS identification, MPC tracking, and MPC clustering, which are red-colored in Fig. 1 and discuss the further development of these approaches.

The remainder of this article is organized as follows. In Section II, we introduce the LOS/NLOS identification process, and present several typical machine learning-based identification techniques. Section III presents widely used machine learning-based tracking algorithms for MPCs and analyzes the potential development of tracking algorithms for V2V channel modeling and other applications. Then, the clustering solutions for channel modeling are discussed in Section IV, including the common approaches and evolution for MPC clustering techniques. Section V discusses some future challenges and open issues for machine learning-based data processing for V2V channel modeling. Finally, conclusions are drawn in Section VI.

II. LOS/NLOS IDENTIFICATION

Intelligent identification of LOS and NLOS scenarios has drawn a lot of attention, mainly because:

Fig. 2. Probability distribution function (PDF) of two examples of extracted characteristics, including kurtosis and Rician-K-factor, as measured in a sample urban V2V environment.

i) LOS/NOS distinction helps localization of vehicles. Localization is always an important problem and an essential need for V2V communications and other applications. Localization is often obtained by measuring the range from one node, e.g., the TX, to multiple RXs, and combining those ranging results by triangulation. For range determination, a LOS situation is beneficial since the runlength of the LOS MPC can be directly mapped to the distance between Tx and Rx. Conversely, NLOS signals either need to be discarded, or require a complicated analysis about the reflection and scattering processes affecting the MPCs. Identification of LOS/NLOS allows to determine how to weight range measurements to different RXs when they are combined for triangulation. Thus, an accurate and efficient LOS/NLOS identification process can greatly improve the performance of localization.

ii) Due to different propagation channel characteristics, the V2V channels are usually separately modeled in LOS and N-LOS scenarios, respectively. For static channel measurements, the LOS and NLOS identification can be predetermined by the specific selection of measurement locations, and human inspection is straightforward. However, for dynamic V2V communications channels, LOS and NLOS conditions may occur randomly in the streets, influenced by the mobility of the vehicles. Therefore, V2V channel measurements cannot directly distinguish the LOS and NLOS conditions by selecting specific measurement locations. Manually checking with video taken during the measurements is one of the possible solutions for LOS/NLOS identification, but very time-consuming. Therefore, an intelligent automatic LOS/NLOS identification algorithm is required for V2V channel modeling.

A number of automatic LOS/NLOS identification algorithms are machine learning-based, due to the good performance of the machine learning tools for classification problems. In general, machine learning-based identification algorithms are able to learn the difference of the channel properties between the LOS and NLOS propagation channels based on labeled training data, and then, provided unidentified channel data, generate a classifier based on the learned inter-difference. In the following, we present the most widely used such algorithms.

Key parameters threshold-based identification: has been widely used due to its low computational complexity. The
The basic idea is to compute certain key channel parameters, e.g., Rician-K-factor, Kurtosis, or Skewness, which typically take on different values for LOS and NLOS channels. This enables to make a decision for LOS or NLOS depending on whether this parameter is above or below a certain threshold; the value of the threshold has to be obtained from training data that are labeled by human inspection. A drawback of this method is that for the time-varying V2V environment, the channel properties are dynamically changing with time, which makes it hard to determine a fixed threshold of any channel parameter and leads to unreliable identification.

Support vector machines (SVM)-based identification: is one of most widely used LOS/NLOS identification approaches considering the good performance of SVMs for classification as a supervised machine learning tool [3]. The basic idea is to build a training database from LOS and NLOS measurement data identified by human inspection, then use the SVM to learn the latent pattern between the LOS and NLOS measurement data, and to generate a classifier based on the learned characteristics. In the training phase, the SVM projects the training data into higher dimensions by using different kernel functions, e.g., Gaussian radial basis function (RBF), polynomial kernel function, or sigmoid kernel function, and seeks an optimal hyperplane to divide the LOS and NLOS training data. It is noteworthy that to achieve a good separability of training data, different channel features (parameters) are usually used for building the input training vector. In this case, how to choose trustworthy channel features for the input training vector is crucial for the accuracy of the LOS/NLOS identification. Commonly used channel features for SVM-based LOS/NLOS identification are presented in Table I.

**Random forest-based identification:** shows a superior classification performance for a moderate complexity with a relatively large feature size. It can be deployed as supervised classification. Therefore, random forest has also been adopted for LOS/NLOS identification and shows good classification accuracy [4]. The basic idea is similar to SVM-based identification where the training database is built by using labeled LOS and NLOS measurement data; then the random forest method generates a classifier based on the learning data. The main difference between the random forest and SVM is that random forest contains individual decision trees for different sets of channel features, and then the final decision is made by synthesizing the classification results of each decision tree. Since the random forest automatically determines the weight of the training features based on the correlation of the features and classes, it can incorporate a large number of features and has good robustness. In practice, the channel features used in the training process of a random forest are usually similar to those of SVM-based identification (Table I).

**Artificial neural network (ANN)-based identification:** As an important and widely used machine learning tool, ANNs were developed to solve classification, recognition and regression problems. Considering the good robustness of learning ability and flexibility of network design, ANNs have been used for LOS/NLOS identification in many studies [5], [6]. Similar to other machine learning-based identification algorithms, ANNs usually learn inter-difference between the LOS and NLOS data, then find the mapping relationships between the input training data (measurement data) and the output label (LOS or NLOS label). It is noteworthy that the performance of ANN-based LOS/NLOS identification highly depends on the network architecture, e.g., the number of hidden layers, and the number of nodes in each layer. These parameters are usually empirically determined based on the scales of the database (number of training data and number of selected features for each training sample) and performance validation.

At the current state of research, machine learning-based LOS/NLOS identification solutions generally outperform the static key parameters threshold-based identification solutions, since the former solutions are able to distinguish the data by considering different channel features. Most of the current machine learning-based LOS/NLOS identification solutions use extracted features, like the ones in Table I, as the training input, which makes the identification accuracy sensitive to the selection of the features. With the development of MIMO, the angular characteristics of propagation channels can be observed and also show noteworthy differences between the LOS and NLOS scenarios. Consequently, for MIMO-based V2V communications, the angular information is an important referential feature for LOS/NLOS identification which is able to increase the identification accuracy [7]. Fig. 2 gives the probability distribution of two examples of channel characteristics, i.e., the kurtosis and Rician-K-Factor, in Table I.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum received power over delay samples</td>
<td>Since the LOS MPC generally contains more power than the NLOS MPCs, the maximum received power of each snapshot is usually referential information for distinguishing LOS/NLOS.</td>
</tr>
<tr>
<td>Kurtosis of the received power</td>
<td>Kurtosis of the received power measures the peakedness of the probability distribution, which is defined by the ratio between the fourth and second order moments of the received signal’s amplitude.</td>
</tr>
<tr>
<td>Skewness of the received power</td>
<td>Skewness of the received power measures the asymmetry of the probability distribution, thus the skewness of a Rayleigh distribution is generally larger than that of a Rician distribution.</td>
</tr>
<tr>
<td>Rising time</td>
<td>Rising time measures the time interval between the moment receiving highest amplitude to the end of the signal.</td>
</tr>
<tr>
<td>RMS-delay spread</td>
<td>RMS-delay spread measures the second central moment of the power delay profile, in the current snapshots.</td>
</tr>
<tr>
<td>Rician-K-Factor</td>
<td>Rician-K-Factor is defined as the ratio between the power of a (possible) dominant MPC (typically the LOS component) and the power in the remaining MPCs.</td>
</tr>
</tbody>
</table>
extracted in V2V channel data measured with an 8 × 8 MIMO system with 15 MHz bandwidth at 5.9 GHz. From Fig. 2 it is found that there is always some overlapping area for both two examples of channel features, which causes difficulties to determine a threshold of any channel feature to precisely distinguish the LOS and NLOS conditions. However, the machine learning-based approaches do not classify LOS/NLOS on any single feature, but instead projects the data into multiple feature-dimensions and seeks the most suitable hyperplane to classify the LOS/NLOS situation [7].

III. MULTIPATH COMPONENT TRACKING

For V2V communications system, MPC tracking is another important topic, mainly because: i) Modeling the time-varying V2V communication channels requires to capture the time evolution characteristics. To achieve this, the MPCs need to be tracked over time.

ii) Many V2V reactive adaptation techniques are often too slow comparing the dynamic changing of the environment. The impact of the processing delay can be reduced by a good prediction of channels, which is facilitated by MPC tracking.

iii) For time-varying channels, the channel properties in the time dimension can be utilized to identify clusters. We will discuss this in detail in Section IV.

In the past, MPC tracking has been achieved by a variety of algorithms, which can be roughly placed into two categories:

Threshold-based MPC tracking solution: Since the channel measurement data are collected in individual snapshots, we cannot obtain the ground truth of MPC moving trajectories. In this case, the most intuitive solution for tracking is to measure and compare the distance between all MPC pairs in consecutive snapshots, where the MPC pairs with small distance are usually considered as the same MPC moved during the snapshots. In other words, if two MPCs observations in consecutive snapshots have the smallest relative distance, they will be associated and become part of the “track” of a temporarily evolving MPC. For the threshold-based MPC tracking algorithm, the value of the threshold and the method used for measuring the difference between the MPCs are crucial to the tracking performance. Based on the propagation physics, the multipath component distance (MCD) is proposed to measure the distance among MPCs, which shows good accuracy [8]. To properly describe the MPCs’ behavior, e.g., birth/death, split, and merge, a threshold of the measured distance is usually set to select the MPC sets during the snapshot. If there are multiple MPCs with relative distances between snapshots smaller than the threshold, all of them will be tracked and associated [9]. The threshold-based MPC tracking algorithm usually has a low computational complexity, but also low accuracy.

Machine learning-based tracking solution: On the other hand, there are many studies focusing on using machine learning-based tracking algorithm, which can be further divided into two categories:

i) Classical tracking method-based. There are multiple tracking algorithms in the machine learning field. For tracking of MPCs, the most widely used include Kalman filters, where the temporal evolution of each MPC is modeled as a linear function with Gaussian noise, and the tracking results in each snapshot are modified by considering the observed and estimated tracking results; Particle filters, where the temporal evolutions of multiple MPCs are modeled as a joint probability distribution to estimate the moving positions; and Extended Kalman filters, which are similar to Kalman filters but model the temporal evolution of each MPC as a non-linear function. Since the tracking performance of the these methods has been validated in many other applications, they are promising solutions for MPC tracking [10].

ii) Matching method-based. For the classical tracking methods, the temporal evolution is often estimated by using recursion functions, which causes considerable computational complexity. Thus there are other studies considering the MPC tracking in consecutive snapshots as a weighted matching problem, and tracking the MPC by identifying the optimal matching relationships between consecutive snapshots. For these studies, the Hungarian method [11] and the Kuhn-Munkres method [12] are commonly used to identify the optimal matching relationship due to their low computational complexity.

Based on the current study, the machine learning-based tracking performance is generally better than the simple thresholding solutions. For example, [12] shows, based on synthetic channel data, that the Kuhn-Munkres-based tracking method shows distinctly superior performance compared to distance-threshold based tracking method. Besides, [13] gives the comparison between the different filter-based tracking. However, it is noteworthy that various methods may perform differently for different measurement data. Therefore, we cannot conclude which method is absolutely best.

IV. MULTIPATH COMPONENTS CLUSTERING

MPC clustering processing has become an essential part of V2V channel modeling, mainly due to:

i) The purpose of the channel modeling is to accurately model the MPCs in wireless channels, meanwhile, the MPCs are usually found to have a cluster structure, which means a certain group of MPCs often has similar characteristics, e.g., angle, delay, and power. In this case, instead of modeling each MPC, modeling the clusters can significantly simplify the channel models.

ii) Clustering results for V2V channel measurement data
can well reflect the objects and obstacles in the physical environment, e.g., cars, trees, and buildings. Such a mapping can be used for a variety of purposes, including validation of the data by cross-checking the measurement videos.

While manual clustering has been done in the past, an intelligent clustering algorithm not only can operate on much larger data sets, but also can identify the clusters in terms of the MPCs’ behavior in multiple domains, e.g., angle, delay, and time domain, which is hard to achieve by inspection.

Clustering is one of the most fundamental research areas in machine learning, so that many MPC clustering algorithms are machine learning-based, and show good performance. Current MPC clustering algorithm can be divided into three categories: **Clustering based on power delay profiles (PDP):** the clusters are identified in the delay domain, especially when the measurement bandwidth is large. In this case, the PDP of each cluster is usually modeled as a single-sided exponential function, as shown in Fig. 2 (a), where different clusters have different initial delays and usually have the same decay coefficient [14]. Clusters in the PDP have been observed in many measurements and given rise, e.g., to the Saleh-Valenzuela model.

**Clustering based on extracted MPC:** the angular information can be observed in MIMO communication systems, providing important referential information for clustering. Thus many machine learning-based MPC clustering algorithms are developed by considering the relationships among MPCs in multiple domains, usually including angle and delay domain. To achieve this, the MPCs need to be extracted first by using high-resolution channel parameter estimation algorithms, e.g., space-alternating generalized expectation-maximization (SAGE) or joint maximum likelihood estimation (RiMAX), then clustered by using different clustering methods based on the inter-relationship between the MPCs among all domains [15].

**Clustering based on spectrum estimation result:** By using spatial filters, e.g., Bartlett beamformer or Capon beamformer, the power angle spectrum (PAS) estimation result can be obtained with low computation complexity, though relatively low accuracy. The PAS can provide the angular distribution of MPCs but without requiring high-resolution extraction of MPC parameters. In this case, identifying clusters in the PAS is considered as target recognition process, where the clusters are separated from the background noise based on the distribution of elements in the PAS [12].

In the past, the clusters has been identified using PDP [14], using the angle-delay domain (MPCs with high-resolution extraction) [15], as shown in Fig. 3, and using the power spectrum [12]. In practice, the angle domain may consist of multiple dimensions, e.g., azimuth and elevation. Based on a recent study, evaluations based on Fourier techniques, i.e., PDP-based and PAS-based clustering solutions generally have lower computational complexity, while clustering solutions based on high-resolution estimation of the MPCs generally have better accuracy; detailed comparison between different clustering algorithms is given in [15].

As mentioned before, considering the time-variations of V2V channels, the tracking results of MPCs may contribute to the clustering process. In time-varying channels, some clusters may be hard to identify in a single snapshot, but can be easily identified based on the temporal evolution of the MPCs during consecutive snapshots. Consequently time-varying characteristics can be utilized for identifying the physical objects in the environment; details are discussed in the next section.
V. OPEN ISSUES

The challenges and open issues discussed in this section can be considered as guidelines for developing future machine learning-based data processing techniques for more precise and efficient V2V communication channel modeling.

A. Intelligent scenario recognition

So far, V2V channels are usually modeled in different representative scenarios, e.g., freeway, urban, and suburban, which are predetermined while designing the measurement campaigns. Note that, the scenario-based channel model is also a general model but parameterized for a particular environment. However, simply defining the scenario by the location of the vehicle is a rough way and brings difficulties to characterize the channels, since even in single scenario there are different typical aspects, e.g., traffic jam areas and truck blocking areas, which cannot be determined in advance. In this case, an intelligent scenario identification can provide an accurate classification to more precisely model the V2V channels. Work in this area might follow the LOS/NLOS identification, since it is a fundamental scenario identification, extended identification techniques can be developed for further specific scenarios identification.

B. Deep learning-based clustering

The current clustering algorithms mainly identify the cluster in each snapshot individually, where the time-varying characteristics of the MPCs have not been well utilized. The reflected signals from multiple static objects may occasionally arrive at the receiver at a similar time and angle, making them difficult to be distinguished and clustered, e.g., the clusters 1 and 2 reflected by the tree and the building at time 1 shown in Fig. 4. However, with the movement of the receiving vehicle, these clusters usually have different evolution patterns and eventually separate, as shown as the clusters 1 and 2 at time 2 in Fig. 4. In this case, we can identify these clusters based on their unique evolution patterns. We can then use the identified clusters to build a training database for deep learning networks to generate a cluster recognition network, such that, for each individual snapshot where the evolution pattern is not available, these clusters can be directly identified by the trained network. For this application, deep learning-based clustering solutions are anticipated to perform well.

C. Deep learning-based channel modeling

Most of the current channel models are developed based on estimated channel parameters, which extract MPCs first then characterize the MPCs using different data processing techniques, i.e. a clustering process for cluster-based channel modeling and a tracking process for time-varying channel modeling. In this sense, the accuracy of the clustering and tracking will impact the accuracy of the final models. In fact, some machine learning methods, e.g., deep learning, are able to be used for modeling the channels, considering their good performance for regression. However, deep learning is sensitive to the variety of training data, thus simply using common system settings, as input is not precise enough to accurately reconstruct the channels. Improving the performance of deep learning methods can be achieved by using more detailed training data, which can be obtained by other processing progresses, as shown in Fig. 5. With the further development of machine learning methods, the performance and efficiency of data processing has been dramatically improved, which may
contribute to the deep learning-based V2V channel modeling research.

VI. CONCLUSION

This article has provided an overview of the applications of machine learning for V2V communications channel research. We have analyzed results on the data processing that improves the accuracy and efficiency of channel characterization, mainly including the LOS/NLOS identification, MPC clustering and tracking. Finally, we have discussed some challenges and opportunities of machine learning-based V2V channel modeling, and several interesting research problems were suggested to stimulate future data processing for V2V channel modeling innovations.

REFERENCES


Chen Huang (ch_628@usc.edu) received his B.S. and M.S. degree from the Chongqing University of Posts and Telecommunications in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree with the Beijing Jiaotong University. He is also a visiting scholar at University of Southern California. His research interests are in machine learning based channel modeling.

Andreas F. Molisch (molisch@usc.edu) is the Solomon Golomb Andrew and Erna Viterbi Chair Professor at the University of Southern California. He was previously at TU Vienna, AT&T (Bell) Labs, Lund University, and Mitsubishi Electric Research Labs. His research interests are in wireless communications, with emphasis on propagation channels, multiantenna systems, ultrawideband systems, and localization. He is a Fellow of NAI, AAAS, IEEE, IET, Member of the Austrian Academy of Sciences, and recipient of numerous awards.

Rui He (ruisi.he@bjtu.edu.cn) received the B.E. and Ph.D. degrees from Beijing Jiaotong University (BJTU) in 2009 and 2015, respectively. He is currently a full professor with the State Key Laboratory of Rail Traffic Control and Safety, BJTU. His research interests include radio wave propagation, machine learning and clustering analysis, vehicular communications, LTE-Railways, and 5G communications.

Rui Wang (wang78@usc.edu) received the M.S. and Ph.D. degree from University of Southern California, in 2012 and 2018, respectively. His research interests include wireless channel measurements and modeling for vehicle-to-vehicle, mm-wave communication systems.

Pan Tang (tangpan27@bupt.edu.cn) received the B.S. degree in electrical information engineering from the South China University of Technology. he is currently working toward the Ph.D. degree in the Beijing University of Posts and Telecommunications. His current research interests include mmWave channel modelling, V2V channel modeling.

Zhangdui Zhong (zhdzhong@bjtu.edu.cn) is a full professor at Beijing Jiaotong University (BJTU) and a chief scientist with the State Key Laboratory of Rail Traffic Control and Safety, BJTU. His research interests include wireless communications for railways and control theory and techniques for railways. He received the Mao Yisheng Scientific Award of China and the Zhan Tianyou Railway Honorary Award of China.