The Interdisciplinary Research of Big Data and Wireless Channel: A Cluster-Nuclei Based Channel Model

Jianhua Zhang
Beijing Haidian District, Xitucheng Road NO. 10, Mailbox NO. 92, 100876, China

Abstract: Recently, internet stimulates the explosive progress of knowledge discovery in big volume data resource, to dig the valuable and hidden rules by computing. Simultaneously, the wireless channel measurement data reveals big volume feature, considering the massive antennas, huge bandwidth and versatile application scenarios. This article firstly presents a comprehensive survey of channel measurement and modeling research for mobile communication, especially for 5th Generation (5G) and beyond. Considering the big data research progress, then a cluster-nuclei based model is proposed, which takes advantages of both the stochastical model and deterministic model. The novel model has low complexity with the limited number of cluster-nuclei while the cluster-nuclei has the physical mapping to real propagation objects. Combining the channel properties variation principles with antenna size, frequency, mobility and scenario dug from the channel data, the proposed model can be expanded in versatile application to support future mobile research.

Keywords: channel model; big data; 5G; massive MIMO; machine learning; cluster

I. Introduction

It is known that wireless communication conveys the information by electromagnetic waves between the transmitter antenna and receiver antenna, which is called as wireless channel. Obviously, wireless channel properties determine the performance bound of wireless communication systems [1]. Thus, the researches about the wireless channel propagation characteristics and modeling are fundamental in design, evaluation and deployment of wireless communication systems.

In last a few decades, the channel dimensions are continuously expanded from time domain to space, time and frequency three domains, which means that the channel properties are deeply dug. Another progress is that the finding channel characteristics are precisely described by different methodologies, i.e., channel modeling [2]. Above work and standard channel model also significantly promote the development of wireless communication from 2nd generation (2G) to 4th generation (4G). For example, ITU-R (International Telecommunication Union-Radio) released the tapped delay line model (ITU-R M.1225) for 3rd generation (3G) system in 1999 [3]. Then in the following ten years, this model is used to support 3G techniques and system. As earlier as 2007, ITU-R launched the draft of standard channel model for 4G. ITU-R M.2135 issued geometry based stochastical model (GBSM) in

2008, which was a double-directional model considering the delay, amplitude and angles stochastical properties in different test scenarios [4].

In October 2015 at Geneva, the 5th Generation (5G) mobile communication system was officially named as IMT-2020 [5]. Besides supporting the traditional requirement of large-area coverage and high-data rate transmission services, IMT-2020 is also expected to supply smart and reliable interconnection among humans and things. It is fantastic that the vision of IMT-2020 presents the convergence of wireless communication, Internet, Internet of Things (IoT) and Machine-Type Communication (MTC), which brings an explosive increase to traffic volume and stimulates wireless communication to the time of Big Data [6]. Obviously, such vision poses the big challenges to transmission techniques, network architecture and system energy efficiency. There are a lot of academic and industrial research of 5G and the promising solutions could be summarized as: (1) expanding the frequency bandwidth at the high frequency band (above 6 GHz) [7][8]; (2) utilizing the advanced radio techniques, for instance, 3D MIMO or massive MIMO [9][10], non-orthogonal multiple access [11], full-duplex multiplexing [12]; (3) employing the efficient network topology, like nano network [13] or software-defined network [14]. However, above technique research and system evaluation depends on the deep knowing of channel characteristics and reliable channel model, no matter in the time resolution (1ns or lower) or in the dimension of parameters, and the test scenario number. Those requirements also become the great challenges to the research of 5G channel model. Thus research institutes, colleges and industry all over the world started the 5G channel measurements and modeling researches from 2010. In Section II, the worldwide research progress of 5G channel will be summarized in detail.

While wireless communication is dramatically developed, computer science is penetrating acceleratedly, leading to the explosive data volume. “Big Data” is created to describe such phenomenon and now there is no a unified definition about it. On Wikipedia, it is explained that the data volume is very large and it cannot be truncated, managed, processed and transformed within a reasonable time by computer or manpower [14]. Data mining is the knowledge discovery in database, i.e., to dig the valuable and hidden rules from big volume of data by computing. It is expected that it has the powerful ability to predict the future. Data mining merges the knowledge of several subjects including computer science, statistics, extraction of information and image processing, etc. Particularly, machine learning theory achieves the obvious progress in recent years, which is hatched from pattern recognition and computable learning theory. As earlier as 1959, A. Samuel defined machine learning as “field of study that gives computers the ability to learn without being explicitly programmed” [16] and demolished the argument that machines cannot learn like humans by his chess program. At 2005, deep learning is implemented by G. E. Hinton etc., the machine is able to automatically acquire rules by analyzing the data and to predict the unknown future [17]. Machine leaning becomes more than hot and Science in Jul. 2015 is focusing on machine learning. Especially, the cover article of Science at Dec. 2015, “Human-level Concept Learning through Probabilistic Program Induction” written by B. Lake etc. attracts the researchers’ attention. In this article, with the three layers chars structure, i.e., char, stroke and sub-stroke, machine has the learning ability with a few data by Bayesian learning program, which means that machine can realize “learning with only one glance as human being” and it also passes the “Turing test” [18]. With such significant development, machine learning theory has been widely applied in different areas, including computer vision, natural language processing, biometrics recognition, Robot and medical diagnosing, etc.

Naturally, computer science is penetrated to communication from high efficient spectrum utilization to novel service type and traffic forecast [19][20]. [21] uses machine learning to predict channel state information in order to decease the pilot overhead. Especially for 5G, the "Big Data" appears and its related techniques are applied to conventional communication research in order to realize "5G vision". However, the wireless channel is physical electromagnetic waves in essence and currently 5G channel model researches still follow the conventional ways (analyzed in Section II). Considering that the channel measurement data of 5G already appears in big volume because of the increased antenna number, huge bandwidth and versatile application scenarios, thus this article tries to start the interdisciplinary research of Big Data and wireless channel, especially, a cluster-nuclei based channel model is proposed. The novel model takes advantages of both the stochastical model and deterministic model. The channel data is collected and channel parameters are estimated, and multipath components (MPCs) are clustered as the conventional stochastical channel model. In parallel, the scenario picture is recognized by computer and the environment is reconstructed by machine learning. Then the cluster-nuclei are found by matching the clusters with the real propagation objects, which are the key elements to bridge the deterministic environment and stochasitcal clusters. Thereby, a "wave, cluster-nuclei, channel" three-layer structure is formed naturally. Finally, the channel impulse response (CIR) predication in various scenarios and configurations can be realized by cluster-nuclei based channel model with the rules dug from the channel database of various scenarios, frequencies and antenna configurations. Thus this model is promising to support the 5G and beyond research for its low complexity of limited cluster-nuclei as well as physical mapping between clusters and objects.

The rest of this article is organized as follow: the state of art of the channel research for 5G is summarized and the two main modeling methodologies are described in section II. In section III, the channel properties variation rules are proposed to be dug by "Big Data" techniques and two examples are given. Then a cluster-nuclei based channel model is presented by recent computer science progress in section IV. Finally, the conclusions are given and future works are also pointed in section V.

II. Channel Modeling Research Survey

The wireless channel modeling research can be traced back to 1940s when Bell Labs gave the Ricean stochastic model [22]. Over the past decades, channel model research has been obviously evolved, from the single path fading model, to narrow-band fading model, and then to broadband fading model based on tap delay line mechanism. To 4G channel model, it is expanded to double direction model, i.e., the space channel properties are included in order to support the MIMO research. Those changes are from narrowband to broadband, from time selective to time-frequency-space selectivity fading model. It is obvious that channel properties have been deeply dug and model's dimensions are continuously developed.

As pointed in section I, the requirements and technical trend of 5G bring the new challenges to the channel characteristics and modeling. In order to support the research of 5G and beyond, global academic institutes, universities and industrial companies have started the channel researches since 2010. Lund university did a massive MIMO channel measurement with 128 antennas with virtual antennas in indoor to outdoor and outdoor scenarios at 2.6 GHz with 50 MHz bandwidth [23]. Our team also completed channel measurements for macrocell, microcell, indoor hotspot and outdoor to indoor with 32 up to 256 antennas including omni-directional antenna and uniform linear array antenna at 3.5 GHz and 6 GHz with 200 MHz bandwidth in Beijing and Shanghai [24][25]. Those researches mainly concentrate on the massive MIMO which shows the impact of the increasing antenna number to the channel properties and modeling.

As for high frequency band channel research (>6 GHz), Professor T. S. Rappaport's team from New York University (NYU) completed indoor and outdoor scenarios' measurement at

28 GHz, 38 GHz and 73 GHz, with 800 MHz bandwidth, configured with horn antennas, and they proposed the close-in model to fit the pathloss [26]. Aalto University in Finland used sweeper as the transmitter and vector network analyzer as the receiver to collect the channel data of 81-86 GHz, whose scenarios were street canyon and the roof to the street with the longest measurement distance 685 m [27]. Tokyo Institute of Technology studied the transmission coefficient and the reflection coefficient of walls, floor, ceiling and windows at 60 GHz [28].

Channel research for 5G has also widely developed in China. Prof. Chengxiang Wang et.al proposed a geometry based massive MIMO channel modeling methodology [29]. The research team from Beijing Jiaotong University did a lot of research on high speed train scenario related channel properties and modeling research [30]. Xiang Cheng from Peking University was focusing on the mobile to mobile modeling methodology [31]. The research team from Nanjing University of Posts and Telecommunications was interested in the deterministic modeling methodology [32]. As for high frequency band, different universities and companies had different focus and they were summarized as Southeast University (45 GHz) [33], Tongji University (28 GHz) [34], North China Electric Power University (26 GHZ) [35], Shanghai Wireless Communication Research Center (14 GHZ), Chengdu University of Electronic Science and Technology (THz), recently 22 Radio Research Institute in Qingdao also bought expensive sounder to research high frequency band etc. Since 2014, our team also established a measurement platform using separated components, configured with horn antenna and omni-directional platform at 14 GHz and 28 GHz, now it is extending to 38 GHz with 800 MHz bandwidth. We have conducted indoor and outdoor measurement, and the channel properties of delay, angles and amplitudes have been studied and models are given [7][8][36][37]. Compared to 4G channel modeling work, many companies in China also devoted a great deal of manpower and material resources to support the channel research, including Huawei with multiple high frequency bands, ZTE with ray tracing model, and Datang at millimeter wave band.

The mainstream modeling methodologies for 5G channel can be classified into two types: one is GBSM model and another is deterministic model. Both of them are not new and their pros and cons are summarized in Table I. Based on GBSM model, we add the elevation angle at both arrival and departure sides and three dimensional (3D) channel model is achieved [2]. COST 2100 (Co - Operation in the field of Scientific and Technical research) model adds the twin cluster between the multiple scatters to describe the multipath components except for specular reflection waves and multiple links relation is also considered [38]. As for millimeter wave, NYU also adopts this model, and updates its statistical characteristics based on the channel parameters extracted from field measurements [39]. Even GBSM model has low complexity and smoothly evolving from 4G model, however, it cannot accurately predict the CIR on a specific location for its stochastical characteristic. In addition, the position of clusters is assumed to obey the random distribution, which has the deviation from the

Table I *The comparison of stochastical model and deterministic model*

	Stochastical Model	Deterministic Model
Method	Empirically predict the received signal fading from the probability distribution in a wide range and then reconstruct CIR by using simulation method in the sense of stochastics.	Based on electromagnetic field theory, solution or approximate solution of Maxwell's equations is used to determine the CIR of a location by detail geographic information.
Pros	Low complexity (limited clusters)	Having physical meaning
Cons	Lacking of physical meaning, a lot of channel phenomena is very difficult to explain.	Highly complex and relying on the precision of the geographic information and the accuracy of the approximate solution.

practical environment. Therefore, deterministic model gradually arises and it is mainly based on the theory of electromagnetic field.

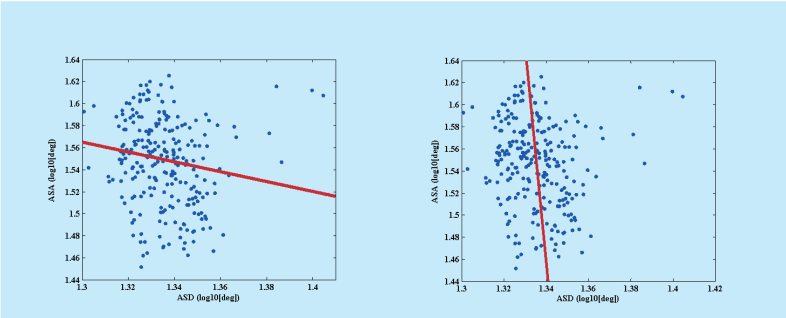
METIS (Mobile and Wireless Communications Enablers for twenty-three Information Society) program also proposes a map-based modeling method [41]. But the main problem of this modeling method is the very high computing complexity, and it heavily relies on the accuracy of the geographic information and the approximate algorithm. Both candidate 5G channel models are the improvement of the tradition models based on the stochastical or physical knowledge. Facing 5G and beyond with the increased antenna number, huge bandwidth and versatile application scenarios, the conventional models become extremely complex and the simulation is time consuming.

Moreover, the unknown channel properties need to be found from the huge volume channel data. Thus, this article tries to propose a modeling method combining statistical modeling and deterministic modeling with the channel properties dug from channel data.

III. Channel Properties Dig by Data Mining Techniques

One fact is that with 5G entering the age of "Big Data", data of channel measurement also appears the three of 4V properties of "Big Data" [42]. (1) Volume. Previously, channel measurement data in single-input single-output (SISO) system with 20 MHz bandwidth is lower than 1 Mbyte. Currently, the data size of massive MIMO configuration with 32×56 antennas and 100 MHz bandwidth increases to 32 Gbyte. Moreover, the channel bandwidth will be as wide as 1 GHz and the application scenarios includes the IOT and MTC for 5G, therefore, the volume of channel data will have an exponential increase. (2) Variety. Due to the physical properties of radio wave, wireless channel data will vary as frequencies, bandwidths and scenarios. For example, even in high speed train application, the propagation properties will be different in distinct sub-scenario like cutting, viaduct, tunnel, etc. (3) Value. One side is that the channel data is the basis of discovering new phenomenon, extracting new channel characteristics and supporting the accurate modeling of radio wave propagation. More importantly, the channel novel characteristics and hidden principles dug from the channel data, will promote the research of wireless communication system. Thus, the channel data have the definite and significant values. Because of the low precision and difficulty in decoupling the system responses, the real time collected data from existing networks cannot be used in the accurate modeling of wireless channel. So wireless channel data are often measured with high-precision instruments (sounder) and processed offline. Although channel data collection and processing lack "Velocity", obvious three "V" properties of channel data still encourage us to use data mining techniques to find more interesting channel properties, rules and give its more accurate model.

"Where is data, data mining is also there". In this section, the data mining technique is tried to be used in channel properties analysis and the novel channel model is proposed in next section. Channel fading is usually classified into large-scale and small scale fading. As for large scale fading, the measurement data is usually least square (LS) fitted, then pathloss and shading fading (SF) models are empirically given with the limited model parameters, including intercept value, fading exponent and SF variance, etc. More interesting while also complex channel properties are small scale fading, which are described by delay, amplitude, cross-correlation polarization ratio (XPR), azimuth angle of arrival (AAOA), azimuth angle of departure (AAOD), elevation angle of arrival (EAOA), elevation angle of departure (EAOD), Ricean (K) factor, Doppler frequency of totally L multipath component (MPC). All of above parameters are firstly estimated from the scenario specified channel data and their stochastical distributions with the cross-correlation are defined in a huge large-scale parameters (LSP) table. Following such LSP table, CIR will be produced based



(a) LS fitting (b) PCA fitting

Fig.1 *One example of data mining technique applied in channel property analysis*

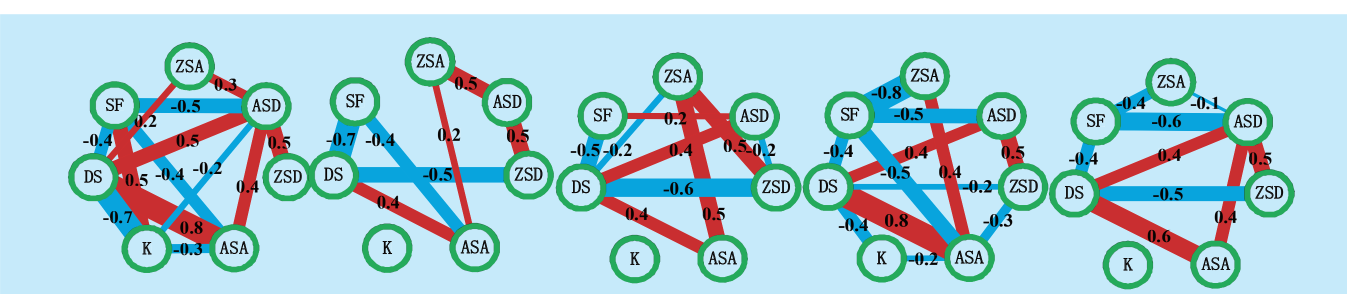
Table II *Part of LSP table from 3GPP TR 36.873*[43]

In the table below DS rms delay spread, ASD = rms azimuth spread of departure angles, ASA = rms azimuth spread of arrival angles, ZSD = rms zenith spread of departure angles, ZSA = rms zenith spread of arrival angles, SF = shadow fading, and K = Ricean K-factor.

Cross-correlation	3D-UMi			3D-UMa	
	LOS	NLOS	O-to-I	LOS	NLOS
ASD vs DS	0.5	0	0.4	0.4	0.4
ASA vs DS	0.8	0.4	0.4	0.8	0.6
ASA vs SF	-0.4	-0.4	0	-0.5	0
ASD vs SF	-0.5	0	0.2	-0.5	-0.6
DS vs SF	-0.4	-0.7	-0.5	-0.4	-0.4
ASD vs ASA	0.4	0	0	0	0.4
ASD vs K	-0.2	N/A	N/A	0	N/A
ASA vs K	-0.3	N/A	N/A	-0.2	N/A
DS vs K	-0.7	N/A	N/A	-0.4	N/A
SF vs K	0.5	N/A	N/A	0	N/A
ZSD vs SF	0	0	0	0	0
ZSA vs SF	0	0	0	-0.8	-0.4
ZSD vs K	0	N/A	N/A	0	N/A
ZSA vs K	0	N/A	N/A	0	N/A
ZSD vs DS	0	-0.5	-0.6	-0.2	-0.5
ZSA vs DS	0.2	0	-0.2	0	0
ZSD vs ASD	0.5	0.5	-0.2	0.5	0.5
ZSA vs ASD	0.3	0.5	0	0	-0.1
ZSD vs ASA	0	0	0	-0.3	0
ZSA vs ASA	0	0.2	0.5	0.4	0
ZSD vs ZSA	0	0	0.5	0	0

on cluster-based channel methodology. Thus for the small scale fading, the accuracy of parameters distribution and cross-correlation is very important to determine the channel model performance. It is worthy trying to use data mining techniques to reduce its complexity and find some hidden rules from data. Here two examples are given to show the possibility of this idea. Here the simple principal component analysis (PCA) is used as one example. PCA uses an orthogonal transformation to convert a set of correlated data to a set of uncorrelated data, called as principal components. The direction of the principal component vectors can reflect the principal relation of the data. Thus, here it is used to get the main relation of ASA and ASD as Fig. 1(b). It is very clear that the fitted red line can better match the measured data, thus the relation between ASA and ASD can be described more exactly by PCA.

Another example is the GBSM model needed huge LSP table. As shown in Annex I, just a part of 3GPP TR 36.873 LSP table for the simulation of 3D MIMO channel model [43] is listed. The cross-correlation of seven parameters are listed, which are DS, ASD, ASA, ESD, ESA, SF and K. Their relations are defined by normalized values and such table is not explicit. In order to find their relation deeper and know what are the deterministic relation in different scenarios, here the data visualization techniques are tried to be used. In Fig. 2, the visualization correlation of those parameters given in Table II is illustrated in data relation point of view. As shown, each two of LSPs is linked with a line, and the width of the line is proportional to their correlation depth. The color of the line



(a) LOS of 3D-UMi (b) NLOS of 3D-UMi (c) O2I of 3D-UMi (d) LOS of 3D-UMa (e) NLOS of 3D-UMa

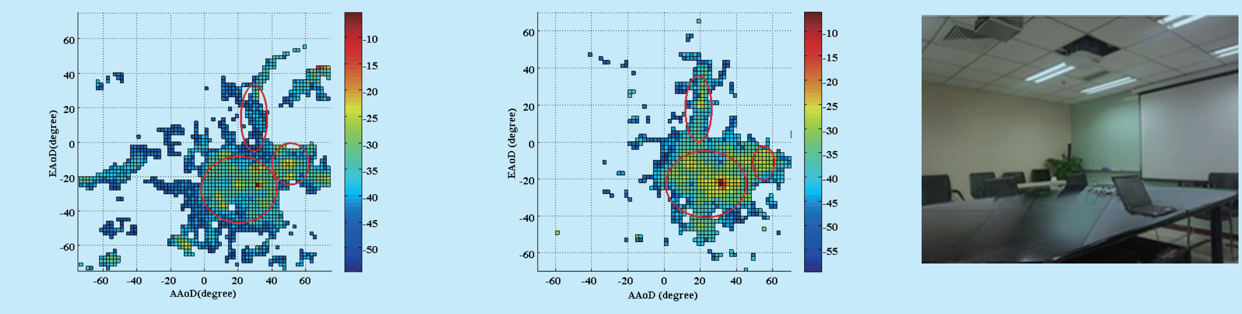
Fig.2 *Visualization of LSP parameters correlation* represents the polarity of the correlation, i.e., a red line indicates the correlation value is positive, while a blue line represents a negative correlation. The correlation values are also labeled over the lines. Now the relation between 7 parameters becomes more intuitive with the help of the visualization of data relation. For example, LOS cases of Fig.2 (a) and (d) lead to more and stronger relation between parameters than NLOS cases of Fig. 2 (b) and (e). It is reasonable that for LOS case, the dominant line of sight will lead to the more similarity among the arrived MPC and thus, high correlation appears between parameters from such MPC. In UMi NLOS case, we have the simplest cross-correlation between different parameters and K factor has no relation with other parameters of course. It is because the transmitter is below the average height of surrounding buildings and arrived MPC are usually multiple rounds reflected and diffused, leading to weak correlation. DS is a key parameter for it is related at least with other 3-6 parameters. Next, SF, ASD, ASA are apparent related with others. One strange phenomenon is that in NLOS of 3D-UMa, SF is strongly related with ASD while in other NLOS case, it is weak. It is worthy to be further researched and validated.

From the two above examples we can find that the data mining techniques can be used to improve the data analysis accuracy and find more new rules of channel data. Then the channel properties can be described more comprehensive and exact. More importantly, if we can set up a channel database, including different frequencies, scenarios, antenna configurations and numbers, bandwidths, etc., then the channel property rules, like variation with the frequency up to 100 GHz, with the antenna number and etc. can be found and input to the accurate channel modeling. That is also our future effort.

IV. A Cluster-Nuclei based Model

As presented in section III, data mining techniques can help us to know about the channel properties based on channel data. In this section, a novel channel model will be given considering the computer science progress.

Let us start from the life sciences in order to clearly explain the novel model. In the life sciences, the double spiral structure of 23 pairs chromosomes convey the versatile individual information. With the three-layer structure of “char, stroke, sub-stroke”, the computer has the learning ability of human being level by Bayesian learning program [18]. All of above tell us the truth that the basic and simple generating elements and structures hidden in the complicated, versatile wireless propagation phenomenon. Considering that the radio waves encounter the main scatters between transmitter and receiver, the received signal is their synthetic effects of reflection, scattering and diffraction, which will appear as MPC clusters. So there must be the relation between the MPC clusters defined in stochastical model and the scatters in the deterministic environment. In order to validate this idea, we conducted an experiment research. The measurement is implemented with 200 MHz bandwidth centered around 3.5 GHz and 6 GHz. The channel data was collected by 56×32 antenna and L=200 MPCs with channel parameters are estimated from collected CIR by spatial alternating generalized expectation (SAGE) algorithm [44] . As shown in Fig. 3 (a) and (b), 200 MPCs distribution in EAOD and AAOD for 3.5 GHz and 6 GHz are plotted, respectively. As expected, there exists obvious clustering phenomenon of 200 MPCs. The main factors to determine the radio propagation and clusters should be the objects in the environment. Such as smooth glass reflects the radio waves to another direction, the dense leaves scatters the radio waves in all directions, solid walls reflect and block most of the radio waves. So for a scatter, the radio wave must have some kinds of common features through this scatter. That is to say, there must have some mapping relation between the cluster and the main scatters in the channel. From Fig. 3(c), it can be visually observed that the largest MPC clusters are caused by the desktop and the wall reflection. So far, the judgment for the relation between



(a) Spatial distribution of MPCs at 3.5 GHz (b) Spatial distribution of MPCs at 6.0 GHz (c) The actual scene graph

Fig.3 *The experiment in beijing university of posts and telecommunications*

the multi-path cluster and the scatter in the propagation environment is artificial. If we can take advantage of the computer vision and figure processing achievement to realize the environment intelligently recognition and find the automatically mapping relation between clusters and objects, then we can propose a new channel model to combine the stochastical and deterministic models together.

Firstly, a three-layer structure (as Fig. 4) is given in order to propose the new channel model. In this structure, a cluster-nuclei is defined as one of clusters which is aggregated by a large number of waves (MPC). There are three important features for cluster-nuclei. (1) it has a certain shape. (2) it has the mapping relation between scatters in the real propagation environment and clusters. (3) It dominates the channel impulse response generation in various scenarios and configurations. With the introduction of cluster-nuclei, the three-layer structure is formed, that is, "the wave, cluster-nuclei and channel". The reason to introduce cluster-nuclei is decreasing the complexity to model the channel from numerous MPCs directly. Another reason is that cluster-nuclei is mapped with the environment objects, rather conventionally clusters only stochastically gotten from delay, angles and power.

With the proposed three-layer structure, the mapping rules between scatters and cluster-nuclei can be dug and the produce method of channel impulse response can be learned by using data mining techniques and machine learning algorithms. Fig.5 presents the principle of the proposed cluster-nuclei based channel modeling. Firstly, the channel data in one scenario with the specific frequency is collected, stored and preprocessed and thus the big database by gathering data from various scenarios and frequency is constructed. Then the channel parameters of each wave such as amplitude, delay, Doppler, EAOD, EAOA, AAOD, and AAOA, are estimated

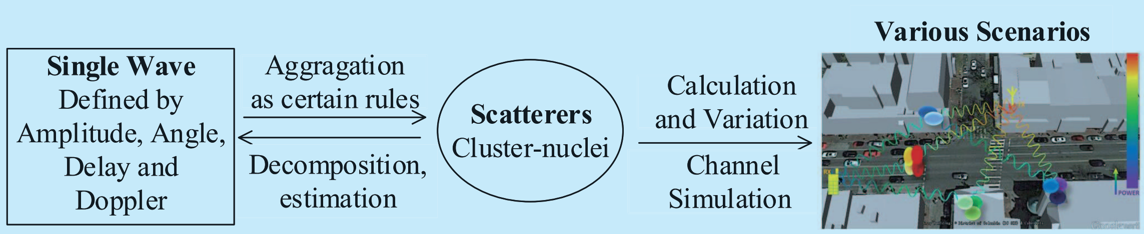


Fig.4 *Wave-cluster nuclei-channel three-layer structure*

by the accurate channel parameter estimation algorithm, like SAGE algorithm, etc. The MPCs are clustered with cluster algorithms such as K-means, K-Nearest Neighbor (KNN) [45]. In parallel, the computer vision algorithms and figure processing can be used to identify the texture from the measurement scenario picture in order to reconstruct 3 dimensional (3D) propagation environment and find the main deterministic objects [46][47]. The key step is to search the mapping and matching relation between clusters and scatters based on the cluster characteristics and objects properties. Then the mechanism and methods to form the cluster-nuclei will be clear. With the limited number of cluster-nuclei, the channel impulse response can be produced by machine learning, i.e., decision tree [48], neural network [49], etc. Based on the database from various scenarios, frequency and antenna configurations, channel changing rules could be dug and found, then input to the cluster-nuclei based modeling. Finally, the channel impulse response predication in various scenarios and configuration can be realized.

Stochastical channel models like GBSM lack physical meaning and deterministic channel models like ray tracing and the map-based modeling method are highly complex and rely on the precision of the geographic information. In contrast, the proposed model takes advantages of both the stochastical and deterministic models, i.e., low complexity with the limited number of cluster-nuclei while cluster-nuclei has the physical mapping to real propagation objects. In addition, channel properties are dug from database, including the variation with the antenna number, frequency, mobility and etc. In order to support massive MIMO, high frequency band and high mobility required by 5G and beyond. Such properties will be input to the channel realization to achieve the channel impulse response predication in versatile environments.

The scatters in Fig. 3(c) are recognized using SIFT algorithm and the radio waves are clustered using the K-Means algorithm. It is obviously that these scatters and clusters can be matched with each other. The sketch map of the mapping relationship is shown in Fig. 6. Three main clusters (C1 C2 C3) between the transmitter and receiver can be matched with

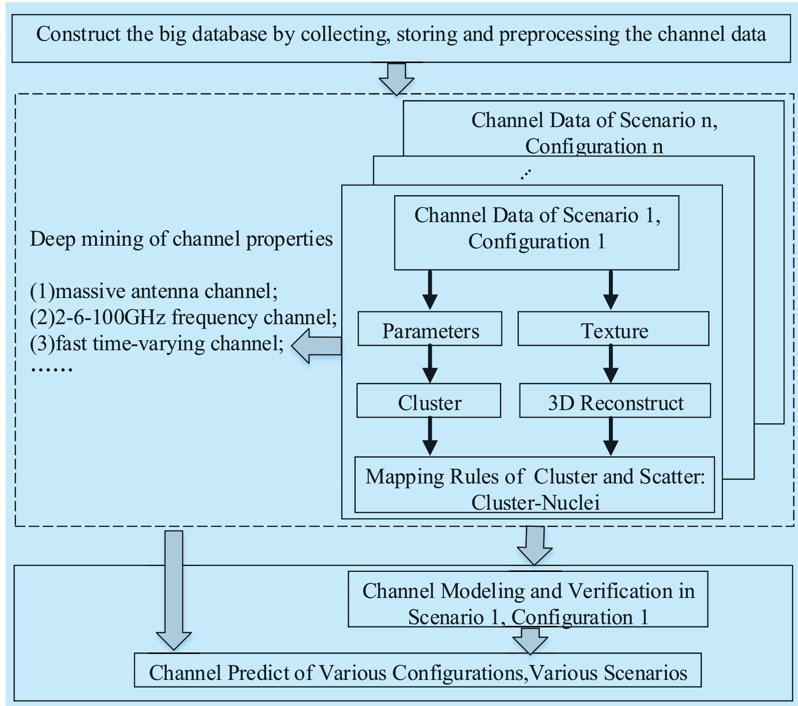


Fig.5 *The principle of the cluster-nuclei based channel modeling*

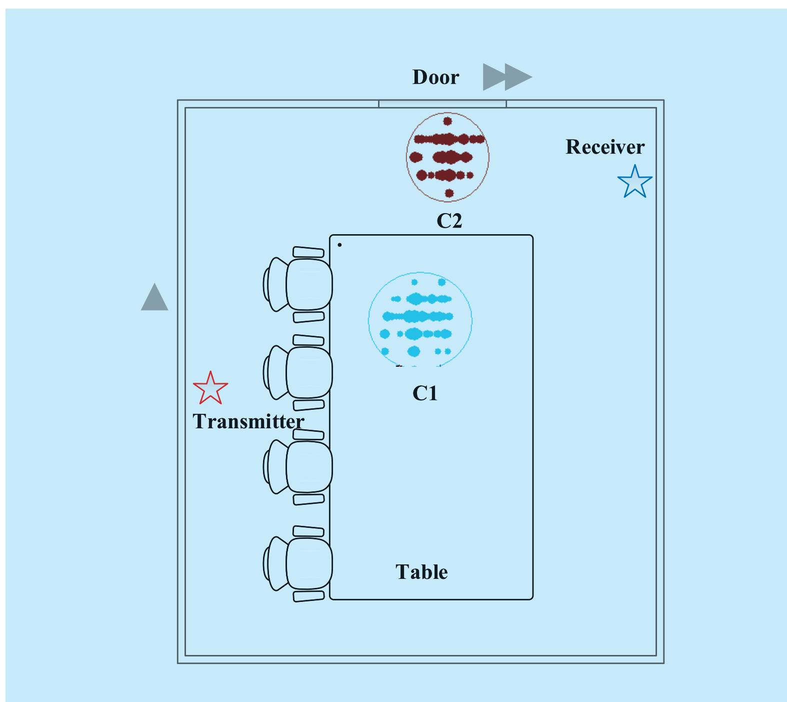


Fig.6 *The mapping relationship between clusters and scatters*

three main scatters, the table, the door and the wall, respectively.

The first step to generate CIR using the method of machine learning is realized. This new method replaces the method of generating CIR by the correlations of large scale parameters. In the new method, the most stable parameter is generated firstly and then is loaded into a neutral network which is trained by the data of measurement. The other parameters can be generated by the neutral network and the CIR is gotten. The simulated CIR is presented in Fig. 7. The accuracy of this new method will be tested in the later research.

V. CONCLUSIONS

This article introduces the 5G channel model requirements firstly and summarizes its recent research progress with the observation that the current 5G channel model is the 4G channel model extension facing the challenges to meet 5G high requirements. As 5G reveals big data feature, the channel measurement data also has the properties of big data which are big volume, form variety and scientific value. Meanwhile, with the development of computer vision and figure processing, 3D propagation environment could be preliminary reproduced from photograph by identifying the texture. Motivated by those, this article tries to do the interdisciplinary research of big data and wireless channel, specifically a cluster-nuclei based channel model is proposed. Here cluster-nuclei is defined and "the wave, cluster-nuclei and channel" three-layer structure is formed in order to match the propagation objects with clusters, thus taking advantage of the deterministic model and the stochastical model. Based on the database from various scenarios, frequencies and antenna configurations, channel changing rules could be dug and found, and then input to the cluster-nuclei based modeling. Finally, the channel impulse response predication in various scenarios and configurations can be realized by machine learning.

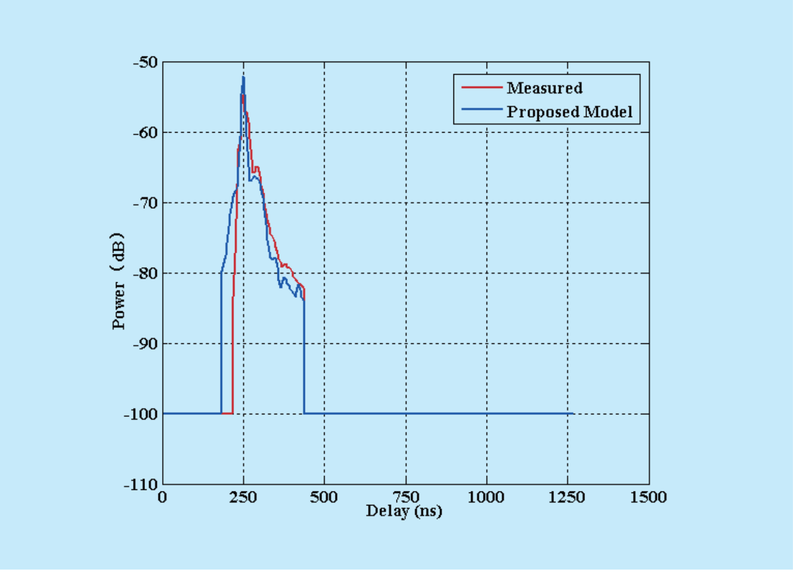


Fig.7 *The generated CIR by machine learning*

In order to promote this research, some open issues are worthy to be noted as: (1) How to construct a trusted database? For different universities and companies have their own data collected by their platform, such data is labelled with different frequencies, bandwidths, antenna configurations and scenarios. Then we need to evaluate such data form and merge the data into database. So it is a complex and time-consuming work to construct this database. But in order to start this step, we began to share our collected data by www.zjhlab.net and expect that more companies or universities can input their data. (2) How to map clusters with scatters and describe the cluster-nuclei, like from the cluster's gravity center and geometry size or other dimensions. Which dimension is more conform to the actual channel data among space, time and frequency dimension? (3) How to use machine learning to generate the channel impulse response (channel modeling) from the cluster-nuclei? Supervised deep learning or Bayesian learning? With the dug rules of the channel in various scenarios, various frequencies, multi-antennas from database, extending the nuclei-based channel model has the ability to predict the channel response under various configurations? Thus this article is just an early start of the interdisciplinary research of big data and wireless channel, and it is promisingly found that data mining and machine learning techniques really give us another new view of conventionally channel research, not matter channel properties discovery or channel modeling methodology.

ACKNOWLEDGEMENT

The author would like thank to all colleagues who took part in the 2016 NSFC major project application proposal drafting from Beijing Jiaotong University, Shandong University and Beijing University of Posts and Telecommunications. Also thanks to Dr. Chenwei now in Cambridge University for the discussion in the machine learning. This research is supported in part by National Natural Science Foundation of China (61322110, 6141101115), and Doctoral Fund of Ministry of Education (201300051100013).

References

[1] MOLISCH A. Wireless communications[M]. UK: John Wiley & Sons, 2011:1-43.

[2] ZHANG Jianhua. Review of Wideband MIMO Channel Measurement and Modeling for IMT-Advanced Systems[J]. Chinese Science Bulletin, 2012, 57(19):2387-2400.

[3] ITU-R M. 1225, Guidelines for Evaluation of Radio Transmission Technology for IMT-2000[R]. 1997.

[4] ITU-R M. 2135: Guidelines for Evaluation of Radio Interface Technologies for IMT-Advanced[R]. 2008.

[5] ITU-R WP5D TD-0591, Naming for International Mobile Telecommunications[R], 2015.

[6] ITU-R WP5D TD-0625, IMT Vision – Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond[R], 2015.

[7] LEI Mingyang, ZHANG Jianhua, TIAN Lei, et al. 28-GHz Indoor Channel Measurements and Analysis of Propagation Characteristics[C]// IEEE, International Symposium on Personal, Indoor, and Mobile Radio Communication. Beijing, China: IEEE press, 2015208-212.

[8] MIAO Runquan, TIAN Lei, ZHENG Yi, et al. Indoor Office Channel Measurements and Analysis of Propagation Characteristics at 14 GHz[C]// IEEE, International Symposium on Personal, Indoor, and Mobile Radio Communications. Hong Kong, IEEE press, 2015:2199-2203.

[9] LIU Guangyi, HOU Xueying, WANG Fei, et al. Achieving 3D-MIMO with Massive Antennas from Theory to Practice with Evaluation and Field Trial Results[J]. IEEE Systems Journal, 2016:1-10.

[10] JIN Shi, WANG Xiaoyu, LI Zheng, et al. On Massive MIMO Zero-Forcing Transceiver using Time-Shifted Pilots[J]. IEEE Transactions on Vehicular Technology, 2015:1-1.

[11] SAITO Y, KISHIYAMA Y, BENJEBBOUR A, et al. Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access[C]// Vehicular Technology Conference (VTC Spring), 2013 IEEE 77th. Dresden Germany: IEEE press, 2013:1-5.

[12] LIAO Yun, WANG Tianyu, SONG Lingyang, et al. Listen-and-Talk: Protocol Design and Analysis for Full-duplex Cognitive Radio Networks[J]. IEEE Transactions on Vehicular Technology, 2016:1-1.

[13] PENG Mugen, YIN Bonan, WANG Wenbo. Bacteria-Based Molecular Communication and Body-Area Nano-Networks[J]. Journal of Beijing University of Posts and telecom, 2016, 39(1): 1-11.

[14] FENG Zhiyong, QIU Chen, FENG Zebing, et al. An Effective Approach to 5G: Wireless Network Virtualization[J]. IEEE Communications Magazine, 2015, 53(12):53-59.

[15] https://zh.wikipedia.org/wiki/ big data

[16] SIMON P. Too Big to Ignore: The Business Case for Big Data[M], UK: John Wiley & Sons, 2013:1-47.

[17] HINTON G, OSINDERO S, TEH Y. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[18] LAKE B, SALAKHUTDINOV R, TENENBAUM J. Human-level Concept Learning through Probabilistic Program Induction[J]. Science, 2015, 350(6266): 1332-1338.

[19] ZHANG Ping, LIU Yang, FENG Zhiyong, et al. Intelligent and Efficient Development of Wireless Networks: A review of Cognitive Radio Networks[J]. Chinese Science Bulletin, 2012, 57(Z2):3662-3676.

[20] LIU Jingchu, DENG Ruichen, ZHOU Sheng, et al. Seeing the Unobservable: Channel Learning for Wireless Communication Networks[C]// 2015 IEEE

Global Communications Conference (GLOBECOM). California, USA: IEEE Press, 2015:1-6.

[21] LIANG Chen, LI Haihan, LI Yunzhou, et al. A Learning-Based Channel Model for Synergetic Transmission Technology[J]. China Communications, 2015, 12(9):83-92.

[22] RICE S. Mathematical Analysis of Random Noise[J]. Bell System Technical Journal, 1944, 23(3): 282-332.

[23] GAO Xiang, EDFORS O, RUSEK F, et al. Massive MIMO Performance Evaluation Based on Measured Propagation Data[J]. IEEE Transactions on Wireless Communications, 2015, 14(7): 3899-3911.

[24] ZHANG Jianhua, PAN Chun, PEI Feng, et al. Three-Dimensional Fading Channel Models: A survey of Elevation Angle research[J]. IEEE Communications Magazine, 2014, 52(6):218-226.

[25] http://www.c114.com.cn/news/41/a927539.html

[26] MACCARTNEY G, RAPPAPORT T. 73 Ghz Millimeter Wave Propagation Measurements for Outdoor Urban Mobile and Backhaul Communications in New York City[C]// International Conference on Communications (ICC). Sydney, Australia: IEEE Press, 2014: 4862-4867.

[27] KYRÖ M, RANVIER S, KOLMONEN V, et al. Long Range Wideband Channel Measurements at 81–86 GHz Frequency Range[C]// Proceedings of the Fourth European Conference on Antennas and Propagation. IEEE, 2010:1-5.

[28] SATO K, MANABE T, IHARA T, et al. Measurements of Reflection and Transmission Characteristics of Interior Structures of Office Building in the 60 GHz Band[J]. IEEE Transactions on Antennas and Propagation. 1996, 45(12):1783-1792.

[29] SALIH O, WANG Chengxiang, AI Bo, et al.Adaptive General Models for Digital Wireless Channels[J]. IEEE Transactions on Wireless Communications, 2014, 13(9):5173-5182.

[30] AI Bo, CHENG Xiang, KÜRNER T, et al. Challenges toward Wireless Communications for High-speed Railway[J]. IEEE Transactions on Intelligent Transportation Systems, 2014, 15(5):2143-2158.

[31] CHENG Xiang, WANG Chengxiang, AI Bo, et al. Envelope Level Crossing Rate and Average Fade Duration of Nonisotropic Vehicle-to-Vehicle Ricean Fading Channels[J]. IEEE Transactions on Intelligent Transportation Systems, 2014, 15(1):62-72.

[32] LIU Yuanjian, ZHU Hongbo, and WEI Cao. A Novel Diffracted Tubes Technology Based on the Method of Ray tracing[C]// 2011 China-Japan Joint Microwave Conference. Hangzhou, China: IEEE Press, 2011: 1-3.

[33] WANG Haiming, HONG Wei, CHEN Jixin, et al. IEEE 802.11aj (45GHz): A New Very High Throughput Millimeter-Wave WLAN System[J]. China Communications, 2014, 11(6):51-62.

[34] YIN Xuefeng, CEN Ling, KIM M. Experimental Multipath-Cluster Characteristics of 28-GHz Propagation Channel[J]. IEEE Access, 2016, 3:3138-3150.

[35] GENG Suiyan, and ZHAO Xiongwen. Feasibility Study of E-band mm-Wave for Gigabit point-to-point Wireless Communications[J]. Microwave & Optical Technology Letters, 2013, 55(8):1969-1972.

[36] ZHANG Jianhua, TANG Pan, TIAN Lei. The current and future of high frequency channel modeling[J]. Telecommunication Network Technology, vol.2016, no.3, pp.10-17, 2016.

[37] HUANG Fusheng, TIAN Lei, ZHENG Yi, ZHANG Jianhua. Propagation Characteristics of Mobile Channel in Urban Micro-cells at 3.5 GHz and 6 GHz[C]// In Proc. 83th Veh. Tech. Conf (VTC2016-Spring), Nanjing, China, May, 2016, accepted.

[38] LIU Lingfeng, OESTGES C, POUTANEN J, et al. The COST 2100 MIMO Channel Model[J]. IEEE Wireless Communications, 2012, 19(6):92-99.

[39] RAPPAPORT T, MACCARTNEY G, SAMIMI M, et al. Wideband Millimeter-wave Propagation Measurements and Channel Models for Future Wireless Communication System Design[J]. IEEE Transactions on Communications, 2015, 63(9):3029-3056.

[40] WANG Xiyu, LI Bin, YUAN Xi, et al. Elevation Angle Research in Three-Dimension Channel Model Using Ray-tracing[C]//General Assembly and Scientific Symposium (URSI GASS), 2014 XXXIth URSI.Beijing, China: IEEE Press, 2014:1-4.

[41] METIS M. Wireless Communications Enablers for the Twentytwenty Information Society, EU 7th Framework Programme project[R]. ICT-317669-METIS.

[42] HOLLEY K, SVAKUMARL G, KANNAN K. Enrichment Patterns for Big Data[C]// International Congress

on Big Data (BigData Congress). Alaska, USA: IEEE Press, 2014: 796-799.
[43] 3GPP TR36.873, Study on 3D channel model for LTE [R]. 2014.
[44] FLEYRY B, TSCHUDIN M, HEDDERGOTT R, et al. Channel parameter estimation in mobile radio environments using the SAGE algorithm[J]. Selected Areas in Communications IEEE Journal on, 1999, 17(3):434-450.
[45] CZINK N, CERA P, SALO J, et al. A Framework for Automatic Clustering of Parametric MIMO Channel Data Including Path Powers[C]// IEEE Vehicular Technology Conference. Montreal, Canada: IEEE Press, 2006:1 - 5.
[46] CHEN L, PAPANDREOU G, KOKKINOS I, et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[J]. Eprint Arxiv, 2014(4):357-361.
[47] PAPANDREOU G, CHEN L, MURPHY K, et al. Weakly- and Semi-Supervised Learning of a DCNN for Semantic Image Segmentation[C]// 2015 IEEE International Conference on Computer Vision. Santiago, USA: IEEE Press, 2015:1742-1750.
[48] FAYYAD, USAMA M. On the induction of decision trees for multiple concept learning[M]. University of Michigan, 1992.
[49] CYBENKO G. Approximation by Superpositions of a sigmoidal function[J]. Mathematics of Control Signals & Systems, 1989, 2(4):303-314.

Biographies

Jianhua Zhang is the Drafting Group (DG) Chairwomen of ITU-R IMT-2020 (5G) channel model and vice-Chairwomen of China IMT-2020 spectrum group. She received her Ph.D. degree in circuit and system from Beijing University of Posts and Telecommunication (BUPT) in 2003 and now is professor of BUPT. She has published more than 100 articles in referred journals and conferences and 40 patents (one received 2016 government Patent Excellence Award). She was awarded "2008 Best Paper" of Journal of Communication and Network. In 2007 and 2013, she received two national novelty awards for her contribution to the research and development of Beyond 3G TDD demo system with 100Mbps@20MHz and 1Gbps@100MHz respectively. In 2009, she received the "second prize for science novelty" from Chinese Communication Standards Association for her contributions to ITU-R 4G (ITU-R M.2135) and 3GPP Relay channel model (3GPP 36.814). From 2012 to 2014, she did the 3 dimensional (3D) channel modeling work and contributed to 3GPP 36.873 and also the member of 3GPP "5G channel model for bands up to 100 GHz", now issued as 3GPP 38.900. Her current research interests include 5G, artificial intelligence, data mining, especially in massive MIMO and millimetre wave channel modelling, channel emulator, OTA test and etc.