Comprehensive “Puzzle Solving” Based on Simple Ideas in the Age of Media Information Overflow

Kunio Kashino
Senior Distinguished Researcher,
NTT Communication Science Laboratories

As the volume of music, photographs, and video on the Internet continues to increase, the need for accurate and high-speed searching of media information is growing rapidly. We asked Dr. Kunio Kashino, Senior Distinguished Researcher at NTT Communication Science Laboratories, to tell us about the current state of research on media search in today’s society and his thoughts on how researchers should view and approach their work.

Keywords: media search, media dictionary, matching

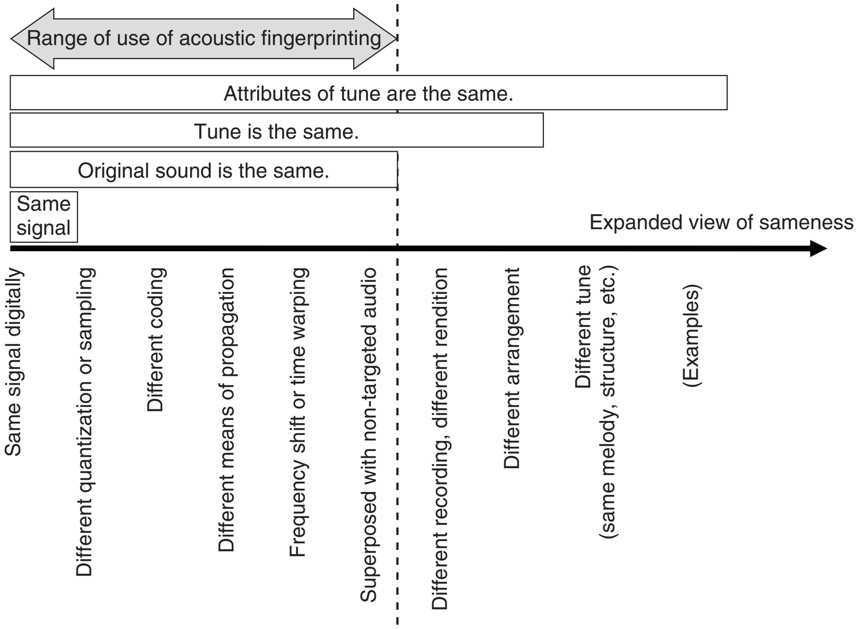
Aiming to create the first “media dictionary”

—Dr. Kashino, please tell us about your current research endeavors.

We call the physical means of conveying information in the form of sound, pictures, or video “media,” and I am researching techniques for analyzing such means of conveying information and identifying its content by computer. I am working in particular on deciphering the information conveyed by some sort of medium by creating and referencing a media dictionary.

We can compare the referencing of a media dictionary with the analysis of text. For example, when we come across a word that we don’t understand while reading something, we can look up that word in a dictionary. Furthermore, when analyzing character strings, we obtain a good result when we compare substrings with entries in a dictionary and find an item that matches. Here, the accuracy of analyzing character strings can be improved by preparing a dictionary with an extensive collection of entries. Similarly, we consider that preparing and referencing a media dictionary that assembles as many audio, picture, and video entries as possible will improve the accuracy of analysis. However, preparing a dictionary—even a conventional word dictionary—is not a trivial task, and in the case of media, even more difficult problems arise. This is why research in this area is needed.

One difficult problem is determining how to judge what is the same and what is different. In the case of words, a character sequence is the key to determining whether a match exists between two items. In the case of media, however, determining whether something is the same as something else is not that simple. For example, we can imagine the same song sung by different singers or the same tune with different arrangements, which means that we can treat two things as


Range of use of acoustic fingerprinting
Attributes of tune are the same.
Tune is the same.
Original sound is the same.
Same signal
Expanded view of sameness
Same signal digitally
Different quantization or sampling
Different coding
Different means of propagation
Frequency shift or time warping
Superposed with non-targeted audio
Different recording, different rendition
Different arrangement
Different tune (same melody, structure, etc.)
(Examples)

There are various viewpoints on *sameness* in media data. Among these, identifying original signals as the same data has come to be called acoustic fingerprinting and video fingerprinting technology.

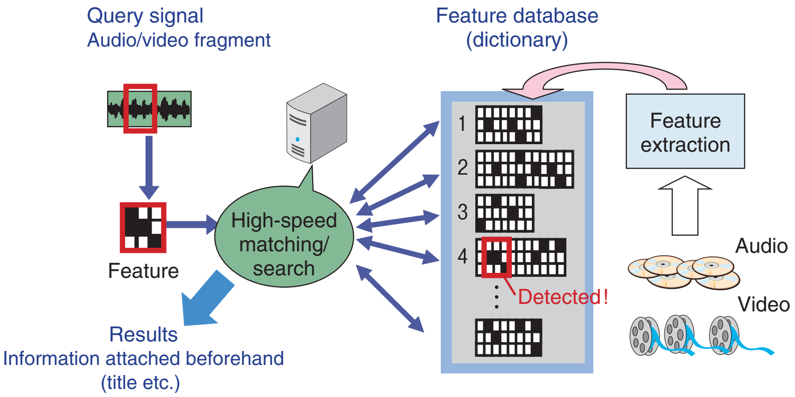
Fig. 1. Range of "same" media data (using music as an example).

being the same in one sense and as being different in another. Furthermore, for a person who does not regularly listen to music in a certain genre, all tunes in that genre may sound the same. In contrast, a person who is very familiar with music of that genre may be able to recognize a certain recording and tell which parts of that recording are especially moving. To that person, such fine differences may be very significant.

These are important issues, but in our research, we began by searching for media data that include audio or video that "sounds the same" or "looks the same" as a fragment of audio or video used as input. This type of technology eventually came to be called acoustic fingerprinting and video fingerprinting (**Fig. 1**). Even by narrowing down the problem in this way, media data itself can undergo major changes for a variety of reasons, such as sound being superposed with other sound at high volume or video being post-processed, so performing accurate searches for target media data is not that easy.

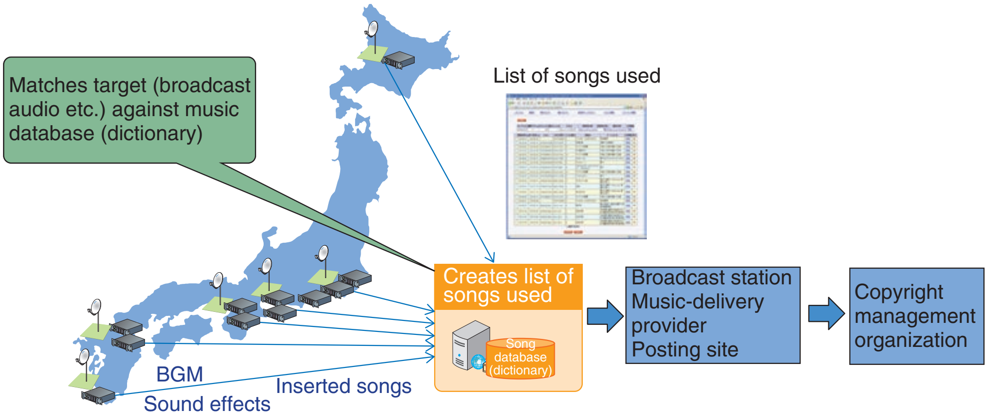
Now, if we assume that the sameness of media data can be determined by some method, there is still another problem that we must deal with, namely, search speed. Nowadays, individuals can easily create media data; on top of that, there is an increasing amount of media data being automatically generated. To search such media data, it has become a common practice to use metadata attached to the main body of data, that is, auxiliary information related to content. However, the effectiveness of using only external metadata attached without human intervention has its limits, and the amount of data is increasing so dramatically that it would take too long to process such data by manual means. There is therefore a need for technology that can automatically analyze and search the actual content of media data in an efficient manner (**Fig. 2**).

This dictionary referencing of media has already begun to be used in areas familiar to all of us. For example, copyright management of music and sound effects used in TV and radio programs is accomplished through such dictionary referencing (**Fig. 3**). More than five million sound sources are registered in the dictionary currently in use. In this method, a portion of each entry in the dictionary is compared with the actual broadcast audio from moment to moment, which makes it possible to instantaneously identify which sound source is being played at what hour, minute, and second on what channel. The basis for this technology was born about 20 years ago, but today, the speed of processing far exceeds what was



Processing is broadly divided into *feature extraction* and *high-speed matching*. Feature data are extracted from audio and video and stored beforehand. These data are then compared with partial features of the query signal at high speed. Quantification of features and high-speed search methods are current research topics.

Fig. 2. Mechanism for identifying "same" audio or video.


Matches target (broadcast audio etc.) against music database (dictionary)
List of songs used
Creates list of songs used
Song database (dictionary)
Broadcast station
Music-delivery provider
Posting site
Copyright management organization
BGM
Sound effects
Inserted songs

An NTT Group company has been providing a song-list creation service for music copyright management since 2007. This service can identify and list all songs used in target broadcasts and music-delivery services and can detect the use of songs in background music (BGM) and short-time use of music.

Fig. 3. Copyright management in song use.

possible at that time; in fact, it's several thousand times faster. In addition, it's not uncommon for multiple sounds to be superposed in the middle of a broadcast program, and good progress is being made on technology that can identify audio under such conditions without any problems.

—What idea led you to begin your research?

As a graduate student, I was interested in developing a way to recognize various types of sounds and noises, which is a very difficult problem. In any case, it is exceptionally hard to recognize individual sounds from a mixture of sounds by computer. However, the overlapping of sounds is not normally a problem to human beings in everyday life; indeed, it can be rather enjoyable, as in the case of musical chords. Furthermore, when listening to lyrics together with some sort of accompaniment, people can even make out phonemes that scarcely appear on the waveform, as in the case of unvoiced consonants. I found this to be a fascinating puzzle, as to what kind of mechanism made this possible. After giving this much thought, I concluded that the mechanism would have to be some sort of matching, as done when checking a dictionary.

This is a very simple concept. The general perception of matching has been that it cannot be useful for processing complex information, or that it is not a viable approach in research. Regardless, I moved forward in my research one step at a time, having various thoughts such as "What would happen if all sounds and things in the world were described in a dictionary" and "Couldn't entries in such a dictionary be derived from actual data?" My idea of creating a media dictionary began around this time.

As an experiment, I tried to record the sounds of traffic during my commute to work. This recording included a variety of sounds such as wind and automobile noise. Since matching can be a robust process in the case of overlapping sounds, I thought that creating a dictionary of traffic sounds would enable me to analyze the soundscape in much the same way that text can be analyzed by referencing an ordinary dictionary.

Composite and comprehensive approach based on simple ideas

—Has your research progressed smoothly? Have you had any difficult experiences?

In 1998, more than two years after I began my research, I learned by chance that there was a need in the world for technology that could identify the appearance of previously registered audio and video content. Specifically, there was a need for detecting TV and radio commercials. At that time, the checking of broadcast commercials was carried out for the purpose of verifying the broadcast of certain commercials or as part of marketing surveys, but this was accomplished by visual means using human labor. However, it was clear that the technique that my colleagues and I were studying could do this checking much faster and more accurately than people could do manually. I lost no time in turning the core part of this technology into a software library and even wrote up an extensive programming guide book so that other people could make use of this technology. Much to my surprise, survey companies that target commercials came to adopt this technology one after another. In this way, I experienced how a simple idea could create value.

At the same time, there were many difficult things to deal with. The target applications naturally expanded, and the types and scale of dictionaries increased rapidly. In the early 2000s, we took up the problem of finding a way to send a song title by email to mobile phone users who would have their phones "listen" to some music that was being played out loud. To do this, we registered sound sources in a dictionary on a scale of several hundred thousand tunes to begin with. At this scale, we noticed interesting phenomena involving errors in recognition, which appeared similar to how two unrelated people may look similar by chance. Then, on increasing this scale by an order of magnitude, we encountered still other types of phenomena. Next, in 2008, we tried identifying known content, such as movies or music pieces, included in the movies posted on the Internet, but to target a scale corresponding to all posted movies, which were increasing on a daily basis, we saw that there would be no other way but to increase the processing speed by about 100 times. At first, I didn't think this was feasible, but on brainstorming with my colleagues, we came up with some ideas and somehow overcame this problem.

—Now that you have overcome some difficulties and achieved some results, are you close to achieving your objectives?

No, not yet! My awareness of the problem when I began my research 20 years ago was centered on the ingenious mechanisms that human beings use to hear sounds and see things. Since then, processing that, in a sense, far exceeds human capabilities has come to be realized, but I cannot say as yet that the original problem has been solved. Moreover, as I mentioned earlier, finding a way of determining what exactly is the same and what exactly is different is still an issue to be addressed. In recent years, it has become relatively easier to store and process huge amounts of media data, so we can consider that even newer approaches may be applicable to this problem.

Taking a puzzle-solving approach from diverse viewpoints

—From here on, how do you plan to approach your research?

I believe it is important that media data be analyzed in a composite and comprehensive manner. In daily information processing as performed by human beings, a person unconsciously mobilizes a variety of senses starting with sight and hearing to understand the surrounding environment. From the beginning, I have had an interest in solving problems in a concrete manner. Real-world problems are often complex and appear in all sorts of forms. Starting with the manifestation of some kind of phenomenon, the task is to work backwards to identify and analyze the problem. Here, however, being able to solve the problem from only one viewpoint is quite rare; a researcher needs to think in a composite and comprehensive manner. This can be compared to the way in which a general physician treats the "whole" patient instead of just focusing on a particular organ.

I would also like to look at problems from a broader field of view, not just from the viewpoint of a researcher. To give an example of what I mean here, let me take you back to the time when it became possible for mobile phones to take videos. At that time, we performed a demonstration of how information obtained from a TV screen captured with a mobile phone could be used to take the user to a website that would provide related information on the program being shown. Technically speaking, the ability to identify screen content from a small and faint image was quite interesting at that time. However, our project was actually a failure. In actuality, viewers are not interested in pressing a button on their mobile phone and capturing video while watching TV. This is a perfect example of how researchers can become self-righteous in their application of technology.

From this experience, I realized that I would like to be motivated by "solving puzzles" in my research while taking to heart the need to approach things from a variety of standpoints and views.

—Dr. Kashino, can you leave us with some advice for young researchers?

I feel that establishing a research theme is very important. I want young researchers to pursue a theme that they believe to be important—not simply a theme that is currently trendy in society. However, at that time, you should not become self-righteous about the theme that you want to choose; you must examine closely why you feel that theme is important and whether anyone else thinks it to be important. If no one at all is working on that problem, perhaps it could be that it is simply not important. However, the best scenario is to find a problem that almost no one else thinks is important but is, in reality, extremely important. Such a research theme will likely grow and develop like the trunk of a tree. Additionally, when you find yourself up against an obstacle, it's a good policy to reconsider what, in the end, is really of importance in your research.

■ Interviewee profile

Kunio Kashino

Senior Distinguished Researcher and Head of Media Information Laboratory, NTT Communication Science Laboratories.

He received a Ph.D. from the University of Tokyo for his pioneering work on music scene analysis in 1995. Since joining NTT in 1995, he has been working on audio and video analysis, search, retrieval, and recognition algorithms and their implementation. He has received several awards including the Maejima Award in 2010, the Young Scientists' Prize for Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2007, and the IEEE (Institute of Electrical and Electronics Engineers) Transactions on Multimedia Paper Award in 2004. He is a senior member of IEEE. He is also a Visiting Professor at the National Institute of Informatics, Tokyo, and at the Graduate School of Information Science and Technology, the University of Tokyo.